BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

OBJECTIFICATION OF FACIOKINESIS ASSESSMENT

OBJEKTIVIZACE HODNOCENÍ FACIOKINEZE

MASTER'S THESIS

DIPLOMOVÁ PRÁCE

AUTHOR BSc Anna Vančová

AUTOR PRÁCE

SUPERVISOR doc. Ing. Jiří Mekyska, Ph.D.

VEDOUCÍ PRÁCE

BRNO 2025

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Master's Thesis

Master's study program **Bioengineering**

Department of Biomedical Engineering

Student: BSc Anna Vančová **ID:** 222703

Year of study: 2 ***Academic year:*** 2024/25

TITLE OF THESIS:

Objectification of faciokinesis assessment

INSTRUCTION:

1) Conduct a literature review on faciokinesis and current approaches to assessing faciokinesis in patients with dysarthria, with particular focus on the 3F Test – Dysarthric Profile. 2) From the provided video database of patients with Parkinson's disease, segment the sections where faciokinesis was assessed, as well as other sections where the patient performs various speech tasks. 3) Select one of the available deep neural network models for detecting key facial landmarks and apply it to the segmented recordings. 4) Based on the available literature, implement parameters that utilize these landmarks (and their temporal dynamics) to quantify hypomimia and, more generally, facial mimicry disorders. 5) Perform a statistical analysis to evaluate the correlation between the values of these parameters and the scores of the 3F Test. 6) Propose a method to automatically estimate faciokinesis scores from the 3F Test based on facial recordings.

RECOMMENDED LITERATURE:

[1] NOVOTNY, Michal; TYKALOVA, Tereza; RUZICKOVA, Hana; RUZICKA, Evzen; DUSEK, Petr et al. Automated video-based assessment of facial bradykinesia in de-novo Parkinson's disease. Online. Npj Digital Medicine. 2022, roč. 5, č. 1. ISSN 2398-6352, doi: 10.1038/s41746-022-00642-5.

[2] ROUBÍČKOVÁ, Jaroslava. Test 3F: dysartrický profil. 3., dopl. a přeprac. vyd., (V nakl. Galén 1.). Praha: Galén, c2011. ISBN 978-80-726-2714-1.

Date of project specification: 10.2.2025 ***Deadline for submission:*** 28.5.2025

Supervisor: doc. Ing. Jiří Mekyska, Ph.D.

doc. Ing. Radim Kolář, Ph.D.
Chair of study program board

ABSTRACT

Assessing faciokinesis offers insight into the extent of damage to corresponding brain areas in certain disorders, much like dysarthria assessment, which evaluates impaired speech-related motor control. While reliable clinical tests exist, they require professional supervision. Despite progress in video-based facial movement analysis, automated quantification of faciokinesis remains limited. The goal of this work is to create a tool capable of objectively assessing faciokinesis. Test 3F used in the Czech Republic, is a dysarthria assessment tool, including a Faciokinesis subtest.

The database used consists of video recordings of 99 individuals with PD and 50 healthy controls performing Test 3F. Facial landmarks were extracted from video segments using MediaPipe FaceMesh, and relevant features were derived to train XGBoost regression models. Statistical analyses, including Spearman's rank correlation, examine biomarker-score relationships while accounting for confounders like age, sex, and medication.

The constructed pipeline was able to achieve an MAE of up to 0.205 and an RMSE of 0.301 in predicting the scores of individual tasks, and 2.249 MAE and 3.002 RMSE for the total Faciokinesis score prediction. SHAP analysis provided insight into model decisions, aligning well with clinical patterns such as reduced or asymmetric facial movements. The results highlight the potential of such models in supporting clinical diagnostics, though limitations in data diversity and quantity restrict generalizability.

KEYWORDS

Faciokinesis, Dysarthria, Facial analysis, Machine learning, Biomarkers, XGBoost

ABSTRAKT

Hodnocení faciokinézy poskytuje vhled do míry poškození odpovídajících oblastí mozku u některých onemocnění, podobně jako hodnocení dysartrie, které se zaměřuje na poruchy motorického řízení řečových svalů. Ačkoli existují spolehlivé klinické testy, jejich provedení vyžaduje odborný dohled. Navzdory pokroku v analýze mimiky na základě videozáznamů zůstává automatizované kvantitativní hodnocení faciokinézy omezené.

Cílem této práce je vytvořit nástroj schopný objektivně hodnotit faciokinézu. Test 3F, používaný v České republice, je klinický nástroj pro hodnocení dysartrie, který zahrnuje i subtest faciokinézy.

Použitá databáze obsahuje videozáznamy 99 pacientů s Parkinsonovou nemocí a 50 zdravých kontrol provádějících Test 3F. Z video segmentů byly pomocí MediaPipe FaceMesh extrahovány obličejové referenční body, ze kterých byly následně odvozeny příznaky pro trénink regresních modelů XGBoost. Statistické analýzy, včetně Spearmanovy korelace, zkoumaly vztahy mezi biomarkery a skóre s ohledem na možné zkreslující faktory, jako je věk, pohlaví a medikace.

Vybudovaný pipeline dosáhl MAE až 0,205 a RMSE 0,301 při predikci jednotlivých úloh, a MAE 2,249 a RMSE 3,002 při predikci celkového skóre faciokinézy. Analýza pomocí SHAP poskytla vhled do rozhodování modelu, které dobře korespondovalo s klinickými vzorci, jako je omezený nebo asymetrický pohyb obličeje.

Výsledky poukazují na potenciál těchto modelů jako podpory klinické diagnostiky, přesto však omezení v diverzitě a množství dat snižují jejich zobecnitelnost.

KLÍČOVÁ SLOVA

Faciokineze, Dysartrie, Analýza obličeje, Strojové učení, Biomarkery, XGBoost

Rozšířený abstrakt

Hodnocení faciokineze poskytuje vhled do rozsahu poškození odpovídajících oblastí mozku u některých neurologických poruch, obdobně jako hodnocení dysartrie, které se zaměřuje na narušenou motorickou kontrolu svalů odpovědných za produkci řeči. Ačkoli existují spolehlivé klinické testy, jejich provedení vyžaduje přítomnost kvalifikovaného odborníka. Přestože v posledních letech došlo k významnému pokroku v analýze obličejových pohybů pomocí videozáznamů a strojovému zpracování získaných dat, automatická kvantifikace faciokineze zůstává dosud omezená.

Pochopení problematiky bylo prohloubeno prostřednictvím rešerše literatury zaměřené na poruchy faciokineze, dysartrii a metody jejich hodnocení. Faciokineze se ukázala být častým klinickým ukazatelem, jehož abnormality často slouží jako indikátory neurologického poškození. Je úzce spojena s dysartrií, která je charakterizována narušením svalové kontroly nutné k tvorbě řeči. Poruchy faciokineze mohou mít různý původ, avšak nejčastěji jsou v literatuře zmiňovány hypokinetická dysartrie, typ parézy mimických svalů typické pro Parkinsonovu nemoc, a Bellova obrna. Pro tyto diagnózy je běžnou praxí kvantifikovat míru postižení pomocí škálových testů. Širším způsobem existuje řada slibných studií, které využívají počítačovou analýzu k detekci abnormalit jak u zdravých jedinců, tak u osob s poruchami. Pro klasifikaci byly použity různé modely strojového učení, včetně SVM, XGBoost, statistické analýzy a neuronových sítí.

Cílem této práce je vytvořit nástroj schopný objektivního hodnocení faciokineze. V České republice je pro hodnocení dysartrie využíván Test 3F, který obsahuje i subtest Faciokineze. Tento subtest se dále člení na tři podskupiny zaměřené na pohyby rtů, čelisti a jazyka, všechny skupiny obsahují pět úkolů. Za každý úkol lze získat 0, 1 nebo 2 body, takže maximální bodové hodnocení za tuto podskupinu je 10 a celkové skóre za faciokinezi může činit maximálně 30 bodů.

Na základě videozáznamů výkonu jednotlivých úkolů v rámci tohoto subtestu a jejich následného zpracování výpočetními metodami je možné automaticky určit odpovídající skóre každého úkolu. Databáze použitá v této práci obsahuje záznamy 99 osob s Parkinsonovou nemocí a 50 zdravých kontrol při provádění Testu 3F. Bylo nutné sjednotit formát videozáznamů a poté vybrat ty video segmenty, které obsahují provedení úkolů patřících k podtestu faciokineze, a další, které mohou být relevantní pro hodnocení pohybu obličeje. V důsledku toho vzniklo za normálních okolností u každého testovaného subjektu 13 video segmentů. Z nich byly pomocí modelu FaceMesh (MediaPipe) extrahovány obličejové landmarky. Celkem jich je 468 a pokrývají celý obličej, pro každý snímek videa, přičemž každý bod je popsán souřadnicemi x, y a z.

Získané landmarky slouží jako základ pro výpočet příznaků pokrývajících celý obličej, které byly následně využity pro trénování modelů strojového učení. Vybrané biomarkery odpovídají směrodatným odchylkám v čase různých parametrů, jako je euklidovská vzdálenost, entropie obrazu v definovaných oblastech a úhly mezi specifickými body. Bylo definováno 13 takových příznaků. Při přípravě dat byly zvoleny moderní a efektivní nástroje s cílem zachovat kvalitu a robustnost zpracování.

Po extrakci příznaků následovalo ošetření odlehlých hodnot a odstranění konfondujících faktorů, jako je věk, pohlaví a dávkování medikace, které byly odstraněny metodou regresního očištění (regression-out). Pro testování statistické významnosti korelací mezi příznaky a skóre byla použita Spearmanova korelace s korekcí FDR za účelem výběru skutečných korelací. Ačkoli celková korelace mezi příznaky odvozenými z obličejových landmarků a skóre souvisejícími s úkoly byla obecně nízká a pouze několik z nich zůstalo statisticky významných po aplikaci FDR korekce, v datech lze stále pozorovat několik intuitivně smysluplných vztahů.

Modelem strojového učení použitým v této práci je XGBoost regrese, která byla optimalizována pomocí RandomizedSearchCV s desetiskladovou křížovou validací. Tento model je vhodný pro úlohy s vysokou dimenzionalitou a zahrnuje regularizační techniky pro minimalizaci přeučení, což je klíčové při práci s klinickými daty, která jsou často limitovaná svým rozsahem i variabilitou.

Pro každé skóre (úkol, podskupina, celková faciokineze) byl vytvořen samostatný regresní model. Vzhledem k nedostatečnému zastoupení některých cílových hodnot bylo nutné věnovat zvláštní pozornost tomu, aby tyto hodnoty nezískaly nedostatečnou pozornost během učení. S ohledem na nevyváženou distribuci skóre byla aplikována vlastní váhovací strategie, terá upravuje příspěvek každého vzorku k celkové ztrátové funkci na základě vzácnosti jeho cílové hodnoty. Dále byla testována technika SMOTE pro oversampling minoritních tříd. Výsledky modelů trénovaných s i bez oversamplingu byly vzájemně porovnány. Pro ověření správnosti dosažených výsledků a vhodnosti zvoleného modelu, byl pro daný úkol vytvořen také jednoduchý jednorozměrný konvoluční neuronový model (1D CNN).

Sestavený pipeline dosáhl hodnoty MAE až 0,205 a RMSE 0,301 při predikci skóre jednotlivých úkolů. Model predikující celkové skóre faciokineze dosáhl MAE 2,249 a RMSE 3,002. V případech s použitím oversamplingu byly metriky pro celý model nižší, avšak bylo možné zaznamenat zlepšení v chybovosti predikce některých jednotlivých skóre. Výhody této techniky se rovněž projevily při analýze interpretovatelnosti příznaků, kde byla u některých příznaků pozorována vyšší míra přispění k predikci. Výsledky ukazují, že XGBoost překonal CNN ve většině úkolů, zejména při predikci individuálních skóre obličeje, což zdůrazňuje robustnost metod založených na stromových modelech v situacích s malým množstvím dat.

Důležitost jednotlivých příznaků byla hodnocena pomocí SHAP analýzy, která poskytla interpretovatelné výstupy. Hodnota SHAP je veličina vypočítaná na základě teorie her, která určuje, do jaké míry daný příznak přispěl k rozhodnutí modelu. Z toho lze snadno vyvodit závěry o tom, jaký vliv mohly mít jednotlivé příznaky na různé predikce. Například snížený pohyb úst či asymetrická aktivita obličeje korelovaly s nižším klinickým skóre. U jedinců s nižším skóre byl častěji pozorován větší rozsah pohybu v jiných oblastech obličeje než v té, na kterou se úkol zaměřuje, což naznačuje, že provedení úkolu je doprovázeno kompenzačními pohyby a zvýšenou mírou soustředění. Některé vztahy mezi příznaky jsou přímé a podporované korelačními analýzami, zatímco jiné odrážejí složité souvislosti objevené algoritmem XGBoost, které by bylo obtížné zachytit tradičními statistickými metodami nebo pouze lidským úsudkem.

Tato zjištění zdůrazňují potenciál přístupů strojového učení, zejména modelu XGBoost, při podpoře objektivního a interpretovatelného hodnocení motorických funkcí obličeje, čímž nabízejí cenný nástroj pro zlepšení klinické diagnostiky neurologických poruch. Zvolený přístup preferuje nástroje, které vyvažují efektivitu a spolehlivost, čímž se metoda stává praktickou, přímou a reprodukovatelnou. Výsledky ukazují potenciál tohoto typu modelu při podpoře klinické diagnostiky, byť obecná použitelnost je omezena rozmanitostí a velikostí datové sady.

VANČOVÁ, Anna. *Objectification of faciokinesis assessment*. Master's Thesis. Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering, 2025. Advised by doc. Ing. Jiří Mekyska, Ph.D.

Author's Declaration

Author:	BSc. Anna Vančová
Author's ID:	222703
Paper type:	Master's Thesis
Academic year:	2024/25
Topic:	Objectification of faciokinesis assessment

I declare that I have written this paper independently, under the guidance of the advisor and using exclusively the technical references and other sources of information cited in the paper and listed in the comprehensive bibliography at the end of the paper.

As the author, I furthermore declare that, with respect to the creation of this paper, I have not infringed any copyright or violated anyone's personal and/or ownership rights. In this context, I am fully aware of the consequences of breaking Regulation § 11 of the Copyright Act No. 121/2000 Coll. of the Czech Republic, as amended, and of any breach of rights related to intellectual property or introduced within amendments to relevant Acts such as the Intellectual Property Act or the Criminal Code, Act No. 40/2009 Coll. of the Czech Republic, Section 2, Head VI, Part 4.

I have used ChatGPT-4o by OpenAI to translate and proofread English texts, enhancing clarity and fluency. I take full responsibility for the final content and confirm that these tools were used in accordance with the guidelines for generative AI tools issued by the Brno University of Technology.

Brno ....

author's signature*

*The author signs only in the printed version.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the advisor of my diploma thesis, doc. Ing. Jiří Mekyska, Ph.D. for his professional guidance and consistently prompt responses to all my questions. His supportive approach greatly contributed to the success of this work.

Contents

List of Figures

List of Tables

Introduction

The accelerated pace of technological advancement of today's world offers many opportunities for enhancing overall quality of life. The role of a bioengineer is to identify these opportunities, develop them, and make them available to as many people as possible. Of the numerous diseases that afflict humanity, there is no known whose cure or diagnosis does not require better, faster and more accessible solutions. It is encouraging to consider the prospect of patients obtaining information about their condition from the comfort of their own homes. This, however requires a device capable of collecting the data essential for evaluation and a diagnostic program that can draw the appropriate conclusions. While innovative solutions exist for the former, the latter represents a significant challenge.

The objective assessment of faciokinesis may be regarded as a potential alternative diagnostic tool in a number of disease areas. Among other conditions, neurological impairment such as Parkinson's disease, head or facial trauma, stroke, Bell's palsy, depression and other psychiatric conditions are associated with symptoms that can affect the functioning of a person's facial muscles, including the ability to express emotions, articulation and voice formation during speech. Such neurological impairments can result in patients suffering from dysarthria, a speech disorder, characterised by impaired articulation, slowness of speech and an overall reduction in communicative ability.

The severity of dysarthria can range from mild to severe, and a variety of tests are available to assess this. These tests requires patients to perform different tasks while a competent individual scores the execution of the task. Although these tests are effective, they are time-consuming and depend heavily on the scoring of the professional, resulting in subjective results. To circumvent this subjectivity, and to facilitate reasons of reproducibility and convenience, a number of recent studies have concentrated on the assessment of faciokinesis disorders using digital recordings. Nevertheless, the majority of this research is focused on diagnostic procedures rather than the extent of damage.

Test 3F, mainly used in Czech Republic, consists of three main parts depending on the type of task: faciokinesis, phonorespiration, and phonetics. **The aim of the thesis is to objectively assess subtest faciokinesis, by predicting a score for individual tasks with a selected machine learning model.** This involves addressing the following subgoals:

1. Getting to know the problematics of faciokinesis assessment in patients with dysarthria in detail.
2. Cleaning and processing the database of video recordings of Parkinson's patients who are undertaking the Test 3F.

3. Extracting features that adequately quantify pathologies associated with faciokinesis.
4. Making statistical analysis of the features.
5. Designing and implementing an Artificial Intelligence model that will objectively assess faciokinesis.

Structure of the diploma thesis

The initial chapter presents the findings of a comprehensive literature search on the subject of faciokinesis disorders. This includes a description of the general features of faciokinesis, its alterations in the neurological disorder dysarthria, methods of evaluating the extent of facial movements, and the findings of recent research in the field of computerised analysis of faciokinesis. Furthermore, the Test 3F is discussed in relation to its utility as a dysarthric profile assessment test, as well as an analysis of its constituent subtests.

The second chapter initially presents the data set employed in the study, delineating its attributes and providing an overview of its statistical characteristics. The Methods section also details the cleaning of the dataset, the tools used to process the videos, and the selection of video segments for further processing. Moreover, a readily available deep neural network model is employed for the purpose of detecting key facial landmarks. Then follows the methodological part of the work that is carried out within the framework of the thesis. This section contains the description of the markers that are extracted from the landmarks, the statistical characteristics of the correlations between the markers and the scores of the faciokinesis tasks from the test. Finally, the methodology for designing the machine learning model is presented, outlining the data preparation steps, model architecture, parameter optimization, and strategies applied to address class imbalance.

The third chapter presents the results of the study, including the Spearman correlation values, the performance metrics of the trained models, and an analysis of feature contributions to model learning.

The last chapter is dedicated to the discussion of these results, offering possible interpretations, limitations, and directions for future research.

1 Theory

1.1 Faciokinesis and its disorders

The word faciokinesis refers to the action of facial muscle movement, typically used in anatomical or physiological discussions to describe the dynamics of how the facial muscles move during expressions, speech, or other facial activities. Many movements are included here, which can be voluntary, but also spontaneous and emotion-induced. This latter is a great difference between the limb movements, and a further difference is that the muscles do not have fixed insertion points to the bone or connective tissue, but form a unique structure on their own. [3]

The facial muscles are responsible for carrying out these activities, such as blinking, movement of the eyebrows and forehead, grimacing, lip movement, smiling, orofacial motor functions such as chewing, swallowing, vocalization and speech. These and the spontaneous, emotion-related facial expressions are performed by overlapping, but different brain areas. [4]

A total of 42 muscles are responsible for facial movement, of which the mimic muscles are innervated by facial nerve, and the masseter and temporalis are innervated by the trigeminal nerve [5]. Their coordinated functioning is necessary for mimicry, speech production and emotion expression.

There are a number of diseases, of which one of the symptoms is a breakdown in the proper functioning of the facial muscles. This may result in hypomimia (loss of spontaneous facial movements expressing emotions [4]), or in many cases dysarthria, a speech disorder.

If communication skills deteriorate, so does the individual's relationship with the outside world, including family, friends or work. The person may be perceived as less intelligent, and the frequency of communication may decrease as the enjoyment of communication decreases, and family members may become distant. However, sometimes this symptom is treated as a secondary symptom and is not adequately addressed. [6]

In the absence of other symptoms of the disease, or in cases where the disease cannot be clearly diagnosed from them, a reduction in the natural range of facial movement could be a possible biomarker, thus allowing a diagnosis to be made. It can be used as a complement to other diagnostic tools or, in certain cases, on its own.

1.1.1 Dysarthria

Dysarthria is a neurological disorder that develops when the peripheral or central nervous system is damaged, resulting in damage to the motor system responsible for speech, causing problems with voicing. The person is able to interpret speech, to formulate what one wants to say, nevertheless has problems with communication. This differs from aphasia, where the problem arises from the patient's inability to understand and formulate both spoken and written speech and language structures. [6]

If dysarthria is accompanied by non-motor symptoms such as cognitive impairment, mental illnesses such as depression, anxiety, which are common in Parkinson's disease (PD) for example [7], they can affect facial mobility. These types of cases are called cognitive-communication disorders. [6]

The disorder can be divided into subtypes. There can be slight differences in the classifications of various studies, however the divisions depend on the origin of the neuroanatomical lesions. According to [1] can be distinguished six subtypes. The types are described in the Table 1.1.

Dysarthria can be improved by therapies, exercises prescribed by a speech therapist or by treating the underlying cause. For this it is important to assess the extent of the damage. Different approaches are used to assess the severity of dysarthria. In general, the factors that make up speech are observed separately, such as respiration, phonation, resonance, prosody, articulation. Facial movements are also observed during the assessment, focusing on the accuracy, range, force and speed of movement of the lips, tongue and jaw, as well as the tone of the neck and facial muscles. Additional parameters can be determined from the speed of movement. [8]

1.2 Possible assessment of face movements

Determining the degree of facial movement is an important factor in certain diseases and disorders. It can reveal the type and extent of neural damage. An important tool for testing the movement of muscles is electromyography (EMG), facial movement testing can be done by surface electromyography (sEMG) or needle EMG [9]. However it is unwieldy and inconvenient to use and requires a high level of expertise, and thus alternative solutions are usually resorted to.

Commonly used assessment tests are done by professionals, with the patient present in person, tasks carried out on instructions of professional, and scored on a scale. The evaluation is subjective and depends on the competence of the examiner. The tasks vary depending on the suspected disease for which the test is being carried out, so there are many variations used worldwide. To illustrate, the following examples are provided.

Table 1.1: Classification of dysarthria [1]

Subtypes	**Origin**	**Disorders**	**Characteristics**
flaccid	bilateral, lower motor neuron lesions or bulbar palsy	brainstem stroke, traumatic brain injury, or neuromuscular disorders	breathy voice quality, hypernasality, imprecise consonants
spastic	bilateral upper motor neuron lesions/pseudobulbar palsy	bilateral strokes, tumors, and degenerative diseases such as primary lateral sclerosis	strain-strangle, harsh voice quality, slow rate, imprecisely articulated consonants, often with hypernasality
ataxic	damage in the cerebellum or its connections	cerebellar degenerations, strokes, and tumors	irregular articulatory breakdowns, excessive and equal stress, also referred to as scanning speech
hypokinetic	extrapyramidal or basal ganglia diseases	Parkinson's disease	rapid rate, reduced loudness, monopitch (unvarying pitch level), monoloudness (unvarying loudness level)
hyperkinetic	extrapyramidal disorders that have increased rather than decreased movement	dystonia, Joseph disease, and Huntington disease	prolonged phonemes, variable rate (sometimes too fast, sometimes too slow), harsh voice quality, inappropriate pauses or silences, voice stoppages
mixed, example: spastic-flaccid	both the upper and lower motor neuron systems	amyotrophic lateral sclerosis (ALS), multiple strokes	hypernasality, strain-strangle, liquid sounding voice quality, extremely slow rate, severe consonant imprecision

In the assessment of Bell's palsy, one of the most common cause of facial palsy (FP), much research uses the House-Brackmann facial nerve grading scale (HBS), due to its simplicity. However, as it is designed to assess facial nerve functionality in post-operative recovery, it is therefore not the most appropriate technique for general rating [10]. A study has shown that the Sunnybrook Facial Grading Scale and the Facial Nerve Grading Scale 2.0 may be an appropriate grading system for multiple scene scores [11]. Electronic assessment can be done with the eFACE scale, even so it is still a subjective test, where clinicians use an application to view videos and apply scoring [12]. The two most frequently used measures for Bell's palsy are the Facial Clinimetric Evaluation (FaCE) Scale and the Facial Disability Index (FDI), which have been translated into several languages. [10]

The Unified Parkinson's Disease Rating Scale (UPDRS) test or the Hoehn-Yahr scale are used to screen for PD, and these tests also include sections that examine facial movement and speech formation. [13]

Most tests are able to assess adequately, but there are difficulties, such as the subjective nature mentioned earlier, or the fact that dysarthria tests are only available in certain languages and there is limited number of multi-lingual research. This may also be the reason why a uniformly proven assessment system has not been developed to date. However, more recently, with the development of machine learning techniques and multimedia tools, attempts have emerged to objectively assess the degree of facial muscle function.

1.3 Test 3F – determination of dysarthric profile

The dysarthric profile can be determined from the speech, there are various studies that use a number of typical features, such as changes in the intensity of the speech in PD cases. Test 3F is an assessment tool specialised in the Czech language to determine the severity of dysarthria. The patient performs various tasks related to speech, articulation and facial movement, from which the tester, usually a speech language pathologist (SLP), evaluates the degree of dysarthria. [14]

The test itself is preceded by a pretest, the purpose of which is to distinguish the cause of the speech impairment that is cognitive in origin and not related to motor impairment, such as dementia. The main part is divided into three units: F1 – faciokinesis, F2 – phonorespiration, F3 – phonetics, thus ensuring the broadness and complexity of the investigation. The test can take 30-60 minutes to complete. Each section is further divided into 3 parts, each of which consists of five tasks. The performance of the tasks is scored, with points being awarded on a scale of 0 to 2:

- 0 points: the patient does not perform the task at all or shows only a small sign of doing so,

- 1 point: the patient has moderate or mild difficulty in performing the task,
- 2 points: the patient performs the task completely correctly.[14]

However, scoring can be more relaxed, the test allows half points if the SLP cannot decide which of two points to use. This makes it more comparable to other foreign tests where a 5-point scale is used, even so the 3-point scale has the advantages of being quick, simple and reasonably stable. By summing the points, we get the dysarthric index. The maximum score available is 90, which indicates a perfectly healthy condition. The distribution of the points in each category is the following:

- 85-90 points – without malfunction,
- 74-85 points – very slight dysarthria, or simple, reduction in motor skills without pathologic cause,
- 57-73 points – mild dysarthria,
- 36-56 points – moderate dysarthria,
- 17-35 points – severe dysarthria,
- 0-17 points – anartria.[14]

The assessment of the test is subjective and depends on the judgement of the person taking the test. This scoring system is able to evaluate dysarthria in patients. When properly performed, a comprehensive picture of the symptom can be obtained and appropriate therapy can be offered to reduce it. [14]

1.3.1 Faciokinesis in Test 3F

The faciokinesis part of the test focuses on tasks that can be used to determine the range of facial movements. The movements to be performed are usually repeated a few times, in most cases three or five times, and the average of these is used to determine the score.

The test targets 3 different parts of face:

1. **Lips**: First, the movement of pulling the lips between the teeth is assessed. Next is the squeezing of a spatula with the lips, which the tester tries to pull out of the mouth, testing the strength of the grip. Further, the ability to inflate the face and hold air in facial cavity is tested. In addition, the two edges of the mouth are pulled into a smile and the mouth is alternated between a puckering and a smile in a rapid manner.
2. **Jaw**: This part involves opening and closing the mouth freely and then against resistance, which the tester performs with finger and spatula. The next task is to move the mandible left and right and then circle the mandible. This is followed by tightening of the masseter muscle, which is determined by the tester by palpation.

3. **Tongue**: In this part, the patient performs tasks such as sticking the tongue out of the mouth, raising and lowering the tip of the tongue inside and outside the mouth. In addition to the horizontal movement, lateral movement is also tested, moving the tongue from one corner of the mouth to the other and then, as a final task, moving the tongue around the lips in both directions. [14]

Beyond this part of the test, however, it is also worth observing the facial movements, even specifically during speech, as there may be visible signs of dysarthria. However, it is a natural reaction for a muscle group to try to compensate for a lack of range of motion, so that the movement of the observed area may appear to be fine, but this does not always mean that the muscle is working well. For example, the lips and the jaw may help the movements due to a lack of control of the tongue. [14]

1.4 Computerized analysis of faciokinesis

The work focuses on the objective assessment of facial movement, which seems to be in demand, as evidenced by the fact that there are many papers and research on the subject. There are several approaches, attempts have been made in relation to different diseases, such as FP [15], but the most widespread research on the topic is on PD hypomimia and facial bradykinesia.

Novotny et al.'s research aimed to develop a fully automatic video-based hypomimia assessment tool and to determine the prevalence of hypomimia in de novo PD patients and to summarise its characteristics. The assessment was based on a large database of video recordings of spontaneous speech from 91 PD subjects and 75 healthy controls. Spontaneous speech is a good basis for such an evaluation, in the sense that it contains the most natural and accessible facial movements. In the videos, 12 facial markers were observed with help of a computer vision system: areas of forehead, nose root, eyebrows, eyes, lateral canthal areas, cheeks, mouth, and jaw. The markers were derived from facial landmarks using two definitions: euclidean distance between two landmarks and a surface description of predefined areas of interest. The variation of these values was then used to calculate standard deviations. The final result is promising, the discrimination between the groups with area under the curve of 0.87 was achieved. The most frequently appearing signs of hypomimia in the PD cohort were related to the movement of the mouth and jaw, and the variability in the movement of the forehead and the nasal bridge folds. [2]

In the study of Skibinska et al., 73 PD and 46 control subjects participated while performing various speech-related tasks. This was audio and video recorded. These recordings were used for speech analysis and in terms of facial landmarks movement. As a result, acoustic analysis achieved 77% balanced accuracy and the facial analysis 81%. Together they achieved 83%. The most efficient discriminative task was the tongue twister. The XGBoost model was used for the evaluation. The research is promising in that it is multimodal, and the results suggest that better results are obtained when audio and video are tested together than when they are tested separately. [16]

An objective assessment of hypomimia from video recordings was implemented in the study of Bandini et al. from 2017. 17 PD and 17 control subjects were asked to perform given general facial expressions in front of the camera. Euclidean distance from a neutral baseline was determined. As a result, it was found that the observed distances were larger in healthy controls. In addition, a facial expression recognition algorithm was also trained, a multi-class support vector machine (SVM) for facial expression recognition was used with the help of external databases. The classifier achieved 88% accuracy. The anger and disgust facial expressions showed the largest difference between the two groups. [17]

In a very recent paper, an automated model for detecting PD at an early stage was tested on a database including voice and video. The size of the dataset is large, 130 PD and 90 helathy control (HC) subjects were recorded while performing specific speech tasks. A novel audio-visual fusion model have been used. In order to complement the two datasets, the features were integrated by a Transformer-based cross-attention mechanism. With this method, PD subjects could be discriminated with an accuracy of 92.68%. For the facial features, images were extracted from the video, and a face recognition deep learning model (Multi-task Cascaded Convolutional Networks) was run on these images. On the resulting data, a specially designed local feature extractor was used, which consists of two parts: it encodes each video frame in embedding sequences, thus finding the fine details in the images. In the second part, it can describe the temporal dynamics across frames, including the temporal factor. This data was fed into the fusion model. [13]

One of the aims of the research of Oliveira et al., was to try to solve the problem of small data sets. Often a complicating factor is the difficulty of finding PD persons willing to participate in such research and also making the recordings is time-consuming. However, it is difficult to obtain good results from a small sample size using machine learning, as it is often under-representative of the variety of cases that arise in society. Therefore, in addition to the existing database, of which only 7% was PD, Oliveira et al. generated synthetic data with the help of Conditional Generative Adversarial Network (CGAN). Then Test-Time Augmentation was used, whereby Gaussian noise was added to the test data to create modified instances. The algorithm predicted some class for each of them, but the final prediction for the original test entry was accomplished by voting. The original test set was used for the classification. This method resulted in an accuracy of 83% for the discrimination of non-PD instances. The study was using features set from videos of three facial expressions. Facial action units (AUs) were provided with OpenFace software, and the variances of AUs were calculated. AU represents individual facial muscle components. [18]

In the following study [19], the database included 35 PD and 26 age- and sex-balanced HC individuals. Video recordings were made in a simple way using a laptop camera, and the participants had to pronounce different sounds and syllables during the test. To determine facial landmarks, FaceReader software was used (version 7.0; Noldus Information Technology) to monitor facial regions such as eyes, lips and eyebrows. A two-tailed Student's test was used for statistical analysis of the data, which showed that 49 facial landmarks had significantly less movement in PD cases compared to the control group. In comparison, it was an interesting finding that for some syllables, however, the range of movement could be larger, suggesting that different pronunciations have different effects on facial muscles.

The research of Knoedler et al. aims to determine the presence of facial palsy using machine learning tools, which it combines with the HBS. Nine facial expressions were imaged in 51 patients and 10 HC. The tasks performed included facial expressions such as smiling, eyebrows raising, closing the eyes, among others. A multiclass neural network was trained, where the final result of the classification is to classify the patients into one of the groups based on the aggregation of the images taken of them. The number of groups is defined by HBS as the FP severity group I-VI. The input was the preprocessed facial images of the patients. The accuracy of the algorithm was 98% when all the patient data were used for training. With this method, the classification of a patient's data is performed in 112 ms. [15]

A case with a larger database is [20], where images of 200 PD subjects and 10 healthy controls were captured during 8 facial expressions. Facial landmarks in the images were detected manually by 3 trained clinicians. The algorithm was used to localize these landmarks and the accuracy of the final result was determined by root mean square error normalized by the interocular distance (NRMSE). It was calculated that the landmark's localization was better in this case, than publicly available algorithms. The accuracy could be improved, when the database contain more photographs.

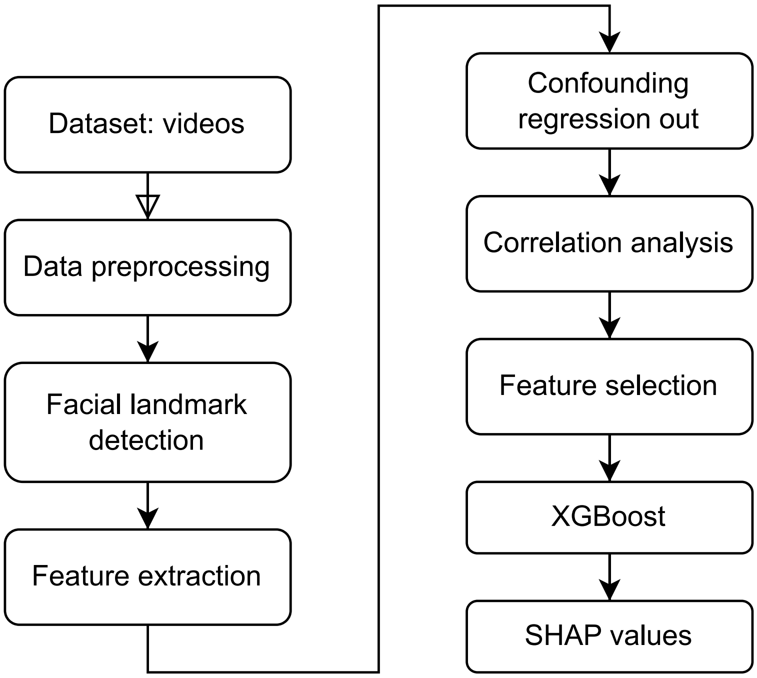


Fig. 2.1: Flowchart of the algorithm

2 Methods

In this chapter, the methods used to find a solution for the described problem are explained. It contains the steps of data preparation for feature extraction, statistical analysis of features and machine learning approaches. The individual steps are indicated on the Figure 2.1.

2.1 Data preparation

To evaluate a dataset using either statistical or machine learning methods, it is essential to understand, standardize, and preprocess the database. This section is dedicated to these tasks, providing information about the dataset's origin and detailing the steps taken to extract features that effectively describe the problem to be addressed.

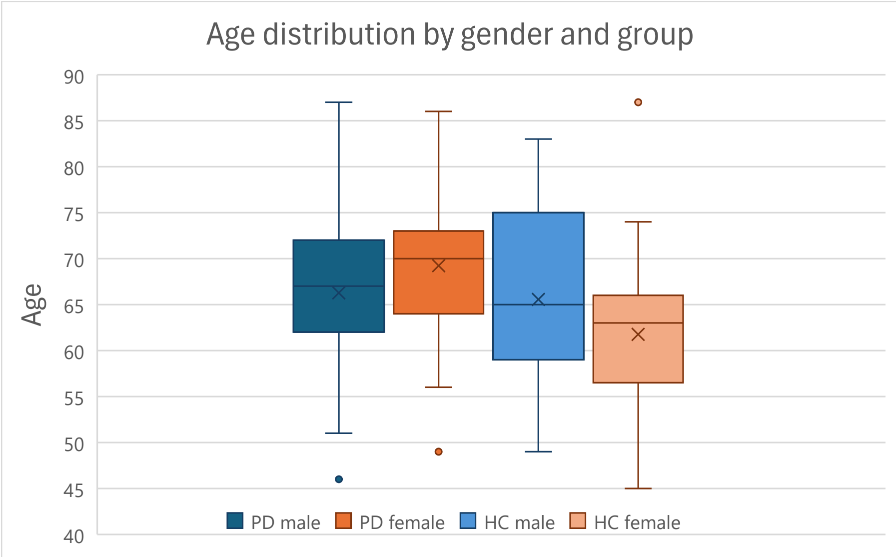


Fig. 2.2: Age distribution by gender and group

2.1.1 Description of the raw data

The database processed in this work consists of 149 recordings of 99 persons with PD and 50 HC. The recordings of the study participants were made within the framework of the project of the Czech Republic Ministry of Health no. NT13499 (Speech, its disorders, and cognitive functions in Parkinson's disease), at First Neurological Clinic of St. Anne's University Hospital in Brno. All subjects with PD at the time of the recording were in the ON state, approximately one hour after taking the drug L-dopa. All study participants signed informed consent, and the study was approved by the local ethics committee. The participants were 50 HC (25 females with mean age 61,76 ± 9,05 years, and 25 males with mean age 65,56 ± 8,90 years) and 99 PD (40 females, with mean age 68,98 ± 7,65 and 59 males with mean age 66,27 ± 8,63). The age distribution of the participants is illustrated in Figure 2.2.

In the PD cohort, the duration of PD is 7,85 ± 4,39 years, and the mean UPDRS III. score is 24,96 with standard deviation 11,96. The mean levodopa equivalent dose (LED) is equal to 1009,02 ± 544,78 mg. The mean dysarthria index at PD patients is 74,39 ± 8,91.

In the videos, participants are instructed externally by a professional to perform the Test 3F tasks, of which the Faciokinesis part is the most relevant for this work. The recordings also include a monologue where participants introduce themselves, this 1–2 minute spontaneous speech may be suitable for further task-independent

Table 2.1: Statistical description of scores of PD in examined tasks

	F1 score	1.1	1.4	1.5	2.1	2.3	2.4	3.1	3.2	3.3	3.4	3.5	Lips	Chin	Tongue
Min	9	0	1	0	1	0	0	1	0	0	1	0	2	2	4
Q1	23	1	2	1	2	1	0	2	2	1	2	1	8	6	7,5
Median	25	2	2	2	2	1	1	2	2	1	2	1	9	8	8
Q3	26	2	2	2	2	2	1	2	2	2	2	2	10	9	9
Max	30	2	2	2	2	2	2	2	2	2	2	2	10	10	10
IQR	3	1	0	1	0	1	1	0	0	1	0	1	2	3	1,5
MAD	2	0	0	0	0	1	1	0	0	1	0	0	1	1	1

Table 2.2: Statistical description of scores of HC in examined tasks

	F1 score	1.1	1.4	1.5	2.1	2.3	2.4	3.1	3.2	3.3	3.4	3.5	Lips	Chin	Tongue
Min	24	1	1	1	1	0	0	1	1	1	1	1	8	6	7
Q1	27	2	2	2	2	1	1	2	2	2	2	2	10	8	9
Median	28	2	2	2	2	2	1	2	2	2	2	2	10	9	10
Q3	29,75	2	2	2	2	2	2	2	2	2	2	2	10	10	10
Max	30	2	2	2	2	2	2	2	2	2	2	2	10	10	10
IQR	2,75	0	0	0	0	1	1	0	0	0	0	0	0	2	1
MAD	2	0	0	0	0	0	1	0	0	0	0	0	0	1	0

analysis. In addition further information can be obtained from the sub-task in which the subjects repeat specific sentences.

The F1 part of the test also includes tasks that require physical intervention by the test-taker. Assessment of these tasks can be problematic for the reason that in these cases the video does not show the person's face well, some parts are covered. Based on this knowledge, segmentation of the videos into shorter segments containing the performance of each task can be done. The following video segments were thus produced: monologue, repeating words after the examiner (task 9.2), and tasks 1.1, 1.4, 1.5, 2.1, 2.3, 2.4, 3.1, 3.2, 3.3, 3.4, 3.5. All tasks have been scored by an expert and this information is available. Statistical description of the data is seen in the Table 2.1 in PD and Table 2.2 in HC group.

The distribution of scores for each task is not even across the dataset, with most data in the 2–point category, fewer in the 1–point category, and the 0–point category being heavily underrepresented. This is a problem for classification that needs to be addressed in machine learning tasks.

2.1.2 Preprocessing

In the original database, the video files are in MOD (a module file format) or Windows Media Video format (with extensions ".MOD" and ".WMV"). They have been converted to MP4 files using the ffmpeg-python module to facilitate further processing. The resolution of the videos after conversion is 640x480 and the sampling frequency is 26.29 frames per second.

The segmentation of the videos was made manually by application LosslessCut [21], an open-source video and audio editor, where first the timestamps were determined for individual tasks and the segments labeled, then the segments, as well as the timestamps exported. This program is designed to make it easy to label segments or even change them later. The parameters of the video segments are not changed from the previous video, as the video is not re-encoded during export. The Smart Cut option can be configured in the program, which allows cutting the video at any time, in addition to the keyframes.

2.2 Facial landmarks detection

In order to extract features, a facial landmark detector can be used. Today, there are numerous methods for face recognition and facial landmark detection, which are also exploited in everyday techniques, such as biometric identification through face recognition or motion capture for filmmaking.

There are several ways to deal with the problem, however neural networks achieve significantly more accurate results than the others. It can be categorized as a computer vision problem. The first attempts included fitting deformable face meshes, this is a holistic method [22], but Random Forest and Gradient Boosting methods such as Ensemble of Regression Trees (ERT) showed better results under uncontrolled conditions, i.e., under inconsistent illumination or from different angles. However, even these cannot outperform neural networks in in-the-wild multivariate datasets. According to the output representation they can be divided into two types: direct (coordinate-based) or heatmap-based, which gives the probability of a landmark's position. [23]

In terms of facial landmark detectors, efforts have been made to date to make detection as accurate and fast as possible, and to work well in cases where the face is not fully visible or visible from different angles. For the purpose of objective assessment of faciokinesis, it is important that the facial landmarks detector is reliable and as accurate as possible. While a real-time solution is not required, the time it takes to process a video frame should be a parameter considered in respect of the large data set.

One possible solution is using a model of MediaPipe. MediaPipe is an open-source framework developed by Google to support developers in building perception pipelines [24]. It has a model developed for face recognition and 3D surface facial landmark detection, being a freely available pre-trained solution named FaceMesh[25]. This model is using machine learning techniques and has been trained

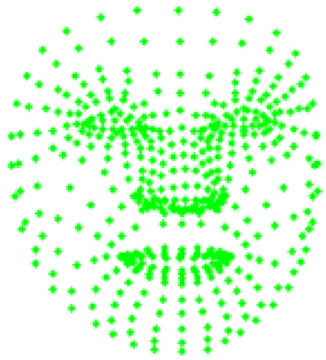


Fig. 2.3: Plotted 2D output of MediaPipe facial landmark detector

on the 300W-LP dataset, which is a large-scale benchmark dataset for facial landmark localization. This dataset includes over 60,000 images of faces with variations in pose, expression, and lighting conditions, suggesting a diverse range of facial features and potentially diverse ethnicities. [26] The model appears in several articles, such as talking face generation with high resolution [27], dementia detection [28], or facial electromyography intensity map generating [29], where they successfully detected important points on the face and used these points for further purposes.

The model can be installed as a library in Python and used immediately. In order to make the detection in the areas of interest as accurate as possible, the Attention mesh model can be used to find the facial mesh, simply by setting the refine_landmarks variable to True at initialization. This model performs more accurate prediction around the lips and eyes. [30]

The output of the detection code consists of 468 facial landmark across each frame with x, y, z coordinates. The z dimension is corresponding to assumption of in depth coordinates obtained from synthetic data. An example is shown on Figure 2.3. The order of the detected points remains consistent across frames, enables straightforward tracking how much each point has moved relative to the previous frame.

In cases where the person is wearing glasses there is no problem with determining the position of the landmarks, as the model was trained on a diverse data set, furthermore uses a regression-based approach to predict landmark coordinates.

2.3 Feature extraction

In order to objectively assess the faciokinesis, the facial landmarks obtained can be used to quantify the facial movements during the given tasks. These extracted features can then be used as input to the machine learning model. The task is to define appropriately descriptive parameters about the pathology.

The landmarks provide excellent coverage of the whole face, with a particular emphasis on the points around the mouth and eyes, the position of which is most variable during voluntary movement. Therefore, the focus is primarily on the parameters arising from these landmarks. It is possible to quantify the relationship of defined points to each other, for example by calculating the Euclidean distance between their positions, the area defined by several points, or the angles between them. Information can also be obtained from the image covered by the area specified by landmarks, such as entropy. By defining this data in each frame, a time series can be obtained and the final parameter can be calculated from it. For this there are also several possibilities: mean, standard deviation, range, variance, entropy. The distance values must be normalized to ensure that they are not affected by the displacement of the head. [16]

Most of the markers are defined based on the article by Novotny et al.[2] These markers cover not only the speech-forming muscles, but the whole face, what is especially important in the spontaneous speech task, where the articulation and emotion based movements are supposed to have the highest range. The standard deviation of the time series of all parameters is taken into account. The markers are summarized in the Table 2.3 and the exact description of the markers can be found as follows.

From the forehead area the entropy of the image within a quadrilateral defined by four landmarks is extracted. Its essence is to describe the variability of the motion of the frontal folds. The marker of the nose root is the entropy of a defined area covering the bridge of the nose. The variability of the galberral wrinkles can be observed by this marker.

There are more markers to describe the area of the eyebrow. The first is the eyebrow elevation, the distance between the eyebrows and the tip of the nose, normalized by the distance between the medial eye corners. This is calculated for both sides of the face. It is defined as a parameter for vertical eyebrow movement. The next is eyebrow tilt, this parameter describes the change in the angle enclosed by the line connecting the medial eye corners and the line that fits the eyebrow line. A bilateral parameter, it describes the change in positioning of the eyebrows. At last, the shape of the eyebrow, which is the obtuse angle of the triangle defined by the two ends and the centre of the eyebrow. A parameter defined on each side, describing the deformation of the eyebrow.

From the eye area the size of the opened eye is observed by the palpebral fissure area size defined by the corresponding landmarks around the eye. It is calculated separately for each eye. Also, the latheral canthal areas are observed, there is extracted the entropy of the rectangular image imposed on the latheral canthal area. Bilateral parameter, able to detect the appearance of wrinkles around the eye. Cheek areas are determined by a triangle, and the entropy is calculated from the difference between two frames on this area. Cheek area motion is described by it.

From the area of the mouth the elevation and depression of upper lip is described by the distance of the upper lip and the nose tip. Normalization is done by the distance of the medial eye corners. It gives information about the movement of the upper lip. The elevation and depression of lower lip is calculated from the distance between the lower lip and the nose tip. The normalization is done by the distance between the medial eye corners. It describes the movement of the upper lip. Mouth corner adduction and abduction is defined by the distance between the corner of the mouth and the tip of the nose, normalized with medial eye corners distance. It represents the change in shape of the mouth. Elevation and depression of the jaw is calculated from the distance between chin and the nose tip. The normalization is done with the medial eye corners distance. This marker describes the movement of the mandible.

Another feature has been added to give more emphasis to the chin area, as there are more tasks associated with this area. This feature is represented by standard deviation of the entropy of the jaw area calculated from each frame.

The conversion of the video frames' images to greyscale is advantageous in the calculation of surface markers, thereby reducing the data volume for processing, and to normalize the pixel values to a range between 0 and 1.

Following the completion of the feature calculation, the dataset comprised 149 rows, with each row corresponding to a patient, and contained 261 columns, containing the patient's ID, the features and scores for all tasks, the score for the entire F1 section of the test, and the scores for the subtests. Additionally, the values for the confounders, including age, sex, and levodopa equivalent dose, were incorporated into the dataset.

The feature values were visualised with histograms, where some outliers could be clearly distinguished. Following a thorough examination of the data, it was determined that the outliers were related to cases where part of the subject's face was outside the image. In such cases, the facial landmark detector was unable to assign coordinates to the facial landmarks, which in turn affected the calculated values. To address this issue and other potential outliers, a threshold has been established for each feature. Values above this threshold are considered unreliable,

Table 2.3: Markers describing the whole face [2]

Marker	Type	Description
Forehead area	surface	The standard deviation of the entropy of the image within a quadrilateral defined by four landmarks.
Nose root	surface	The standard deviation of the entropy of a defined area covering the bridge of the nose.
Eyebrow elevation	Euclidean distance	The standard deviation of the distance between the eyebrows and the tip of the nose, normalized by the distance between the medial eye corners.
Eyebrow tilt	angle	The standard deviation of the change in the angle enclosed by the line connecting the medial eye corners and the line that fits the eyebrow line.
Eyebrow shape	angle	The standard deviation of the obtuse angle of the triangle defined by the two ends and the centre of the eyebrow.
Eye	area size	The standard deviation of the palpebral fissure area size defined by the corresponding landmarks around the eye.
Latheral canthal areas	surface	The standard deviation of the entropy of the rectangular image imposed on the latheral canthal area.
Cheeks	surface	The standard deviation of the entropy from the difference between two frames from the cheek area determined by a triangle.
Upper lip elevation/depression	Euclidean distance	The standard deviation of the distance of the upper lip and the nose tip.
Lower lip elevation/depression	Euclidean distance	The standard deviation of the distance between the lower lip and the nose tip.
Mouth corner adduction/abduction	Euclidean distance	The standard deviation of the distance between the corner of the mouth and the tip of the nose, normalized with medial eye corners distance.
Jaw elevation/ depression	Euclidean distance	The standard deviation of distance between chin and the nose tip normalized by the medial eye corners distance.
Jaw entropy	surface	The standard deviation of the entropy of a rectangular area under the mouth.

which could distort the learning pattern of the final model. To prevent information loss and avoid excluding all data from a given patient in the analysis, outlier values were replaced with the median value for the corresponding feature. This approach ensures that the substituted values do not bias the prediction in either direction. The threshold value was empirically determined as five times the value of the interquartile range.

In certain cases, the video recording of task performance was missing. These missing data points were imputed using the median value of the corresponding features. The proportion of missing data accounted for 1.24% of the entire dataset.

2.4 Statistical analysis

A dataset can be effectively characterized through statistical analysis. By examining it, insights can be gained into the nature of the data, the relationships between features, and their impact on the outcome of interest. In order to find the most ideal set of parameters and to gain insight into how the markers relate to the predicted values, correlation needs to be investigated.

Initially, a correlation matrix was formulated utilising the *pandas* library's *corr* function. The matrices for all tasks were then plotted, as illustrated in Figure 2.4. The correlation matrices demonstrate that no features exhibit a strong correlation with the scores; however, correlations between them can be identified. Features that are bilateral are predominantly correlated with each other, which is unsurprising given that during articulation, the two sides of a person's face move symmetrically to each other. Additionally, a correlation between the position of the lower lip and the chin can be detected, with the lower lip also tracking movement when the lower jaw is opened. In some cases, a slight correlation between certain confounder factors and the score is observed, predominantly for levodopa equivalent dose. To ensure that this correlation does not affect learning from the data set, regression-out correction was applied. This process allows for the filtering of correlations introduced by sampling that may influence the outcome. The correction was executed through a methodology that eliminates confounders from the designated features by employing linear projection. The implementation of linear regression residualisation was achieved through the execution of matrix operations.

In the next processing step, Spearman's rank correlations was calculated between each biomarker and the scores. This is a non-parametric test that describes the monotonicity of the relationship between two variables. It works by ranking the variables in order, making it possible to assess the strength of non-linear relationships. In Python, Spearman's rank correlation is implemented in the *stats* module within the *scipy* library. The function returns a number between -1 and 1, where 0

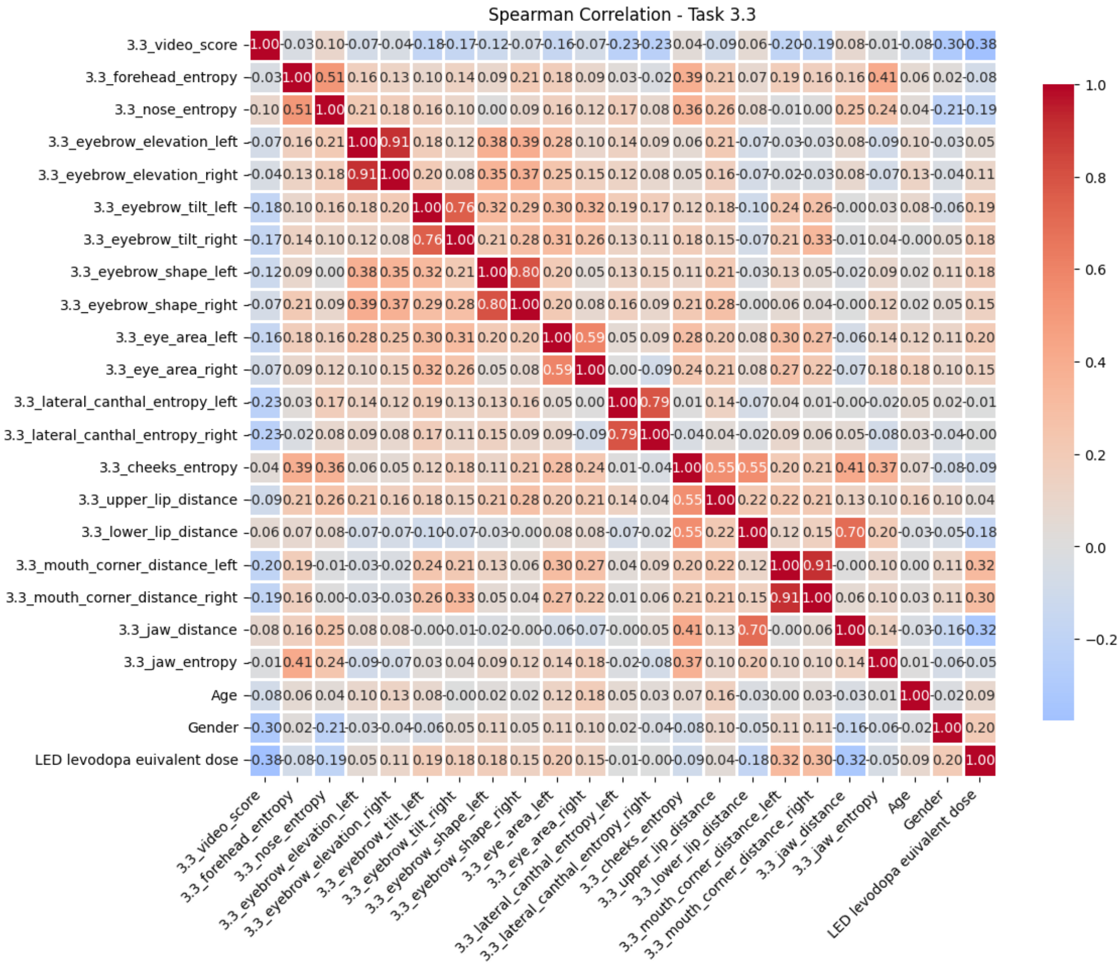


Fig. 2.4: Correlation matrix for task 3.3

indicates that there is no monotonic relationship between the two variables, -1 signifies a perfect negative monotonic relationship, and 1 represents a perfect positive monotonic relationship.

For the *p*-values obtained for correlation, it is beneficial to calculate a False Discovery Rate (FDR), which helps to prevent the rejection of a true null hypothesis, thereby reducing the proportion of false positives. Details of this approach can be found, for example, in the book by [31]. The FDR method establishes a threshold for determining the significance of results with a given probability of error, identifying correlations strong enough to have a low likelihood of occurring by chance. The function used for this is called *multipletest* in the *statsmodels* library with the Benjamini/Hochberg method.

Based on this correlation analysis, it is possible to gain insight into the relationship between the calculated biomarkers and the scores that are desired to be predicted. In some cases, it is possible to determine which markers can be true predictors of the outcome and can be used as strong candidates for predictors in the machine learning model.

2.5 Machine learning

This chapter describes potential methods for using machine learning tools to accurately predict the scores of Test 3F. Determining the score for each task can be interpreted as a classification task, however estimating the score for each subgroup or for the entire F1 part of the dysarthria test is more effectively treated as a regression task.

Given the characteristics of these datasets, applying regression to the individual tasks is expected to yield more favorable outcomes. This is primarily due to the misinterpretation of the potential score range, which is limited to the values 0, 1, and 2, with the value of 0 often being underrepresented. As a result, the model struggles to distinguish the class associated with 0 points, particularly in cases where data points are absent. Regression, however, allows for the generation of continuous scores, which can fall below 1, overcoming this limitation. While the predicted scores may not be integer values, they can be effectively thresholded during the final evaluation to facilitate meaningful classification.

2.5.1 XGBoost model

XGBoost, a scalable machine learning system for tree boosting, may be a suitable implementation. The model constructed using this method has the capacity to address numerous issues and is highly regarded in the field of machine learning. One of its key advantages is the ability to achieve strong results efficiently while also providing explanations for the importance of features in decision-making, making it a popular choice over deep neural networks [32]. XGBoost employs a gradient tree boosting algorithm, to iteratively build a strong model by combining the predictions of weak learners, such as decision trees. At each step, a new tree is added to minimize the loss function, which quantifies the error of predictions. Furthermore, XGBoost is using regularization techniques to prevent overfitting. It parallelizes tree construction by optimizing split calculations, making it faster than many alternatives.

Originally implemented in Python programming language, and it can be installed as a package under the name *xgboost*. Here an XGBoost regressor (XGBRegressor) was used with the help of Scikit-Learn wrapper interface. Three different model configurations were implemented, corresponding to the three levels of prediction targets: per-task scores, per-group scores, and the overall F1 score. While the general pipeline remained consistent across all configurations, the fine-tuning process was adapted individually. In order to optimally build the model, certain parameters need to be specified. In addition, as it employs decision trees, it is necessary to specify the depth of the trees. As the depth of the tree increases, the model is able to detect a greater number of patterns; however, this also increases the probability of overfitting. It is necessary to define the learning rate, which sets step size for each update. Decreasing the value improves accuracy but require more boosting rounds. The number of boosting rounds (trees) can be set using the *n_estimators* parameter. Other parameters which are considered to be set are *colsample_bytree* and objective. By *colsample_bytree* the fraction of features to consider for each tree is determined. Lower values increase randomness and help prevent overfitting. With *objective* the Loss function for optimization can be chosen.

Identifying these values can be challenging. However, for hyperparameter tuning, optimisation techniques, such as *RandomizedSearchCV*, can be employed. These methods facilitate the selection of the most optimal parameter from a number of parameters, for which the model provides the best result. In each model configuration, hyperparameter optimization was performed using RandomizedSearchCV with 10-fold cross-validation. This method selects random combinations of specified hyperparameters and evaluates them via cross-validation. The number of parameter combinations to be tested is set using the *n_iter* argument – higher values are advisable to ensure a more thorough search. Based on the number of parameters and computational constraints, 1000 configurations were tested in this study.

The hyperparameter grid included the following:

- *n_estimators*: Number of gradient-boosted trees.
- *max_depth*: Maximum depth of each tree; deeper trees allow for more complex

models but increase the risk of overfitting.

- *learning_rate*: Controls the step size at each boosting iteration.
- *subsample*: Fraction of the training set used to grow each tree. Lower values can reduce overfitting by introducing randomness. For instance, a value of 0.5 means that half the data is randomly sampled for each tree.
- *colsample_bytree*: The subsample ratio of features (columns) for constructing each tree, also helping to prevent overfitting.
- *min_child_weight*: A node is split only if the sum of instance weights in the child is above this threshold. This acts as a regularization mechanism.
- *reg_alpha*: L1 regularization term on weights.
- *reg_lambda*: L2 regularization term on weights; both help to constrain model complexity.
- *gamma*: Minimum loss reduction required to make a further partition on a leaf node. Higher values make the algorithm more conservative, aiding in overfitting prevention. [33]

Each model explicitly defined its learning task and objective, which determines the loss function to be minimized during training. The appropriate choice depends on the nature of the prediction problem. Empirical analysis of learning curves – which track error reduction over successive boosting rounds – also informed this selection. These curves help evaluate whether the model is learning effectively and whether overfitting occurs.

For per-task score predictions, the objective *reg:squaredlogerror* was applied. This loss function emphasizes relative error by applying a logarithmic transformation to the residuals, meaning discrepancies in higher score values are penalized less severely than errors at lower values. Since the score range (1–10) is relatively narrow, this behavior is not problematic. In fact, it is beneficial due to the lower frequency of low-score samples in the data, making their accurate prediction more critical.

For predicting the overall faciokinesis score, which exhibits a wider range and higher variance, the *reg:tweedie* objective was used. This loss function is well-suited to targets with high dispersion and proved effective under these circumstances.

In order to guarantee the reliability of the performance evaluation, k-fold cross-validation (CV) can be used. This approach is particularly as it uses all data for training and is less prone to overfitting compared to traditional training-validation splits. The aforementioned method can be implemented via the *sklearn* package, which is available from the *scikit-learn* repository. The value of K should be selected according to the size of the dataset. To illustrate, when k equals 10, the dataset is partitioned into 10 subsets, of which 9 are employed for training and 1 for testing. In practice, k = 10 is often the most balanced choice, as evidenced by its frequent usage in numerous studies. This approach reduces the likelihood of overfitting while requiring manageable computational resources. [34] [35]

Due to the underrepresentation of certain target values, special attention was required to ensure these did not receive insufficient learning focus. Given that the dataset is skewed, it is a prudent choice to use stratified CV. This technique ensures that the class distribution in each fold matches the original dataset, which can be critical when the scores (0, 1, 2) are imbalanced. This would maintain the proportional representation of target variable distributions within the subsets. However, when there are only a few samples in one of the group, it is impossible to partition the dataset this way. An other solution involved a custom weighting strategy, which adjusts the contribution of each sample to the overall loss function based on the rarity of its target value. This effectively encourages the model to reduce error more aggressively on underrepresented scores, increasing their importance during training [36]. For this purpose the equation of 2.1 was used. With this formula the weight for each class can be calculated with the fraction of number of every sample in the training data and the number of different classes multiplied by the number of samples in the actual class.

$$weight_i = \frac{n_{\text{samples}}}{n_{\text{classes}} \cdot n_i} \tag{2.1}$$

Another technique considered was SMOTE (Synthetic Minority Oversampling Technique), a widely used method for balancing class distribution by synthesizing new data points through interpolation between existing minority samples [37]. However, the application of SMOTE in regression tasks remains debated. Some studies, as [38] do not recommend using synthetic oversampling in structured medical datasets unless the representativeness of synthetic samples can be guaranteed. Others, as in [39] suggest oversampling might benefit weak classifiers but yields little advantage for state-of-the-art models, noting that even basic random oversampling can be as effective as more complex SMOTE variants.

In this study, an attempt was made to oversample the minority classes. Because the task is regression-based, a regression-compatible SMOTE implementation from the *smogn* library was initially explored. However, this approach failed due to the limited number of distinct target values (0.0, 1.0, 2.0), which caused the algorithm's interpolation to break down — it expects a broader range of continuous values. As such, classification-style SMOTE was instead used. Synthetic samples were generated using 3 nearest neighbors for interpolation, and only those score groups were oversampled where the number of original samples fell below a set ratio compared to the most populous group. The number of synthetic samples was computed using the formula in 2.2, where majority count is the number of samples in the majority group and count is the number of samples of whole training data. This approach helps prevent overfitting by avoiding excessive duplication of minority samples.

$$synthetic\,count = \left(\frac{majority\,count}{count}\right)^{\frac{1}{2}} \tag{2.2}$$

In extremely low-sample cases (e.g., where only 1–3 samples exist), even SMOTE fails to function effectively, particularly when *k_neighbors*=3. For these edge cases, manual oversampling was employed. Specifically, 20 new synthetic samples were generated using *pandas.DataFrame.sample()* based on the available minority-class data, and Gaussian noise was added to each to preserve natural variability. This aimed to mimic realistic fluctuations observed in empirical data.

Although the model is designed for regression, its final use case is to return discrete score values. A straightforward approach was implemented: rounding the predictions to the nearest valid score, with boundaries enforced to prevent predictions beyond the known extremes. In practice, the model never predicted beyond the training data range. Based on this rounding strategy, accuracy can be computed as the percentage of cases where the predicted score exactly matches the true label. This approach may be applicable if the model is employed to predict a specific score.

2.5.2 Metrics

The results of the models predictions need to be evaluated. The following regression metrics can be used. Mean Absolute Error (MAE) (2.3) measures average error magnitude without penalizing large errors excessively. Mean Squared Error (MSE) (2.4) penalizes larger errors more than smaller ones, useful for tasks sensitive to outliers. Root Mean Squared Error (RMSE) (2.5) is the square root of MSE, interpretable in the same units as the target variable. Estimation Error Rate (EER) (2.6) normalizes MAE by the target's value range, providing a scale-invariant metric. These metrics are calculated as follows, where y is the original score, $\hat{y}$ the predicted score, and N is the total number of predictions.

Mean Absolute Error

$$MAE(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N} \tag{2.3}$$

Mean Squared Error

$$MSE(y, \hat{y}) = \frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{N} \tag{2.4}$$

Root Mean Squared Error

$$RMSE(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{N}} \tag{2.5}$$

Estimation Error Rate

$$EER(y, \hat{y}) = \frac{MAE(y, \hat{y})}{\text{target range}} \tag{2.6}$$

2.5.3 SHAP method

In some cases, it is not evident what the model has decided upon when attempting to ascertain the correct prediction. Consequently, as a final measure, to determine the significance of features, the SHAP (SHapley Additive exPlanations) method can be employed. This is a game theory approach, installable as "shap" in Python [40], which employs a distinct high-speed algorithm for tree models. This enables the contribution of each feature to the prediction to be visualised for better understanding. By understanding these contributions, it becomes easier to interpret the model's decisions and refine the predictive performance.

2.5.4 Deep neural networks

The primary methodological focus of this thesis lies in leveraging the XGBoost algorithm, which is particularly well-suited to small datasets – a common scenario in behavioral and psychological research. XGBoost demonstrates strong performance under such conditions [41], while also offering computational efficiency and significantly faster training compared to deep learning alternatives.

Nonetheless, to contextualize and validate the performance of XGBoost models, it is valuable to compare them against alternative machine learning paradigms. As such, this thesis also implements a simple deep neural network, designed with minimal tuning, to serve as a comparative baseline. The chosen architecture is a one-dimensional convolutional neural network (CNN), specifically adapted for regression tasks.

Data preparation for CNN

This model utilizes raw facial landmark data extracted from video recordings. These landmarks form three-dimensional time series. For consistency across samples, each sequence was trimmed or padded to exactly 300 frames. In cases where a video contained more than 300 frames, the video was trimmed after reaching the predefined number of frames. For shorter videos, zero-padding was used after the last frame to maintain the sequence length. This preprocessing step was necessary because CNNs require fixed-size inputs for each minibatch. Standardizing the number of frames simplifies the model architecture and enables efficient batch processing on GPUs.

The average number of frames per video was approximately 300, which motivated the choice of this specific frame count. Additionally, the selection of this number was made considering the balance between temporal information richness and computational efficiency, ensuring that the model would not be slowed down by excessively large inputs.

This approach preserves the temporal dynamics of the videos, and due to the task-oriented trimming of each video – starting from the moment the participant begins the instructed action – it is guaranteed that the relevant motion is captured from the first frame. Furthermore, the nature of the task instructions often leads participants to repeat the same movement multiple times within a video. This ensures that each sample contains at least one full instance of the target motion.

Separate samples were generated for each evaluated task. For models that predict group-level scores or the global F1 score, relevant task-based sequences were concatenated to form the input sample.

To address class imbalance in the target score distributions, data augmentation was applied selectively. Underrepresented classes were synthetically expanded using Gaussian jitter, adding small, random noise drawn from a specified standard deviation to the landmark coordinates. This approach simulates natural facial movement variations, encouraging the model to learn more robust, invariant patterns.

In addition, landmark dropout was used, randomly zeroing 5–15% of landmark positions within a sequence. Importantly, augmentations were applied only to minority class samples, to avoid skewing the majority class representations. All inputs were normalized across samples to ensure stable training dynamics.

Model architecture and training

The 1D CNN was implemented using PyTorch. It comprises three stacked 1D convolutional layers (kernel size = 3) with increasing channel dimensions. Each convolution is followed by Batch Normalization and LeakyReLU activation, which together enhance training stability and allow the model to capture subtle temporal variations. Max pooling layers reduce the temporal resolution between convolutional blocks. After global pooling, a fully connected layer with dropout transforms the temporal output into a latent representation. Dropout serves as a regularization technique to prevent overfitting.

The model employs Huber loss, a robust loss function well-suited to ordinal regression tasks with noisy labels. Compared to MSE, Huber loss is less sensitive to outliers while preserving differentiability near zero, it is an essential property when handling ordered, non-uniform score distributions. The optimization strategy uses AdamW with a weight decay of 0.005 and a learning rate of 0.001. After tuning, the best-performing training configuration was found to be 30 epochs with a batch size of 16. Model performance was evaluated using 10-fold cross-validation, consistent with the evaluation strategy used for the XGBoost models.

3 Results

This chapter presents the outcomes of the methodologies described in the Methods chapter. It includes the results of the correlation analysis between the computed features and the F1 test scores, the predictive performance of the trained XGBoost models based on selected metrics, as well as the evaluation of the CNN models.

3.1 Results of correlation analysis

From the cleaned dataset, the precise correlation values between features and task scores were calculated using Spearman's rank correlation method, with p-values adjusted using the FDR correction. The results of this correlation analysis are presented in Tables 3.1, 3.2, and 3.3 highlighting the features that show statistically significant correlations with task scores under either the uncorrected or FDR-adjusted conditions. Overall, only a limited number of features exhibit significant correlations, and these correlations tend to be weak in magnitude. Notably, some task scores do not display any monotonic relationship with the features. According to the Spearman correlation the higher biomarker values are associated with higher task performance when positive, and lower task performance, when negative.

After FDR correction, only four correlations remain significant in individual task scoring, namely 1.4_eyebrow_elevation_left, 2.3_cheeks_entropy, 3.3_lateral_canthal_entropy_right, and 3.4_eyebrow_shape_right.

In addition to the feature 1.4_eyebrow_elevation_left, the task 1.4 also shows a positive correlation with its counterpart, 1.4_eyebrow_elevation_right, as well as with 1.4_forehead_entropy. Further positive correlations were observed with 2.4_jaw_entropy and 2.1_cheeks_entropy. All other features listed in the table show negative correlations with their corresponding tasks.

Furthermore, among the correlations observed with the task groups, only one remained significant after applying FDR correction. Additionally, eyebrow movements and the entropy of the nasal bridge also showed correlations with the scoring of chin movements. For the tasks related to the lips, overall, the movement of the mouth showed correlations, all in a negative direction.

The correlation analysis of the total F1 score included features from two tasks that are not part of the Faciokinesis examination: the monologue and Task 9.2. Thus, only one feature showed a significant correlation after FDR correction, the right-sided mouth corner distance during Task 9.2, where the task was to repeat words. Other features that showed correlations without FDR correction were negative, what is associated with lower scores on the overall Faciokinesis test.

Table 3.1: Spearman correlation analysis of features with task scores, presented both before and after applying FDR correction. Only correlations yielding statistical significance ($p < 0.05$) are displayed.

Biomarker	**Spearman correlation**	***p* value**	***p* value corrected**	**Significant after FDR**	**Task**
1.4_eyebrow_elevation_left	0.23	0.0044	0.0436	True	1.4
1.4_eyebrow_elevation_right	0.22	0.0088	0.0584	False	1.4
2.4_jaw_entropy	0.20	0.0149	0.1486	False	2.4
1.4_forehead_entropy	0.18	0.0319	0.1595	False	1.4
2.1_cheeks_entropy	0.17	0.0414	0.2759	False	2.1
1.1_mouth_corner_distance_left	-0.16	0.0495	0.2297	False	1.1
1.1_mouth_corner_distance_right	-0.17	0.0449	0.2297	False	1.1
3.1_lower_lip_distance	-0.17	0.0445	0.2970	False	3.1
3.4_eyebrow_tilt_right	-0.18	0.0281	0.1407	False	3.4
3.5_eyebrow_shape_left	-0.19	0.0221	0.1473	False	3.5
3.5_eyebrow_shape_right	-0.19	0.0209	0.1473	False	3.5
3.1_eyebrow_elevation_left	-0.19	0.0201	0.2009	False	3.1
1.5_mouth_corner_distance_left	-0.19	0.0200	0.1334	False	1.5
1.1_eyebrow_shape_left	-0.20	0.0171	0.1714	False	1.1
2.1_jaw_entropy	-0.20	0.0152	0.1525	False	2.1
3.4_eyebrow_shape_left	-0.20	0.0121	0.0805	False	3.4
3.3_lateral_canthal_entropy_left	-0.21	0.0119	0.0794	False	3.3
1.5_mouth_corner_distance_right	-0.23	0.0053	0.0533	False	1.5
2.3_cheeks_entropy	-0.23	0.0042	0.0422	True	2.3
3.3_lateral_canthal_entropy_right	-0.25	0.0022	0.0223	True	3.3
3.4_eyebrow_shape_right	-0.26	0.0015	0.0150	True	3.4

Table 3.2: Spearman correlation analysis of features with group scores, presented both before and after applying FDR correction. Only correlations yielding statistical significance ($p < 0.05$) are displayed.

Biomarker	**Spearman correlation**	***p*-value**	***p*-value corrected**	**Significant after FDR**	**Group**
1.1_jaw_distance	-0.16	0.0483	0.4315	False	Lips
1.1_mouth_corner_distance_left	-0.17	0.0381	0.4315	False	Lips
1.4_mouth_corner_distance_left	-0.18	0.0329	0.4315	False	Lips
1.5_mouth_corner_distance_right	-0.18	0.0298	0.4315	False	Lips
1.1_mouth_corner_distance_right	-0.24	0.0029	0.1665	False	Lips
2.4_eyebrow_elevation_left	0.21	0.0106	0.2993	False	Chin
2.4_eyebrow_elevation_right	0.17	0.0378	0.5386	False	Chin
2.3_nose_entropy	-0.20	0.0158	0.2993	False	Chin
2.3_cheeks_entropy	-0.28	0.0006	0.0326	True	Chin
3.3_lateral_canthal_entropy_left	-0.17	0.0455	0.3916	False	Tongue
3.5_eyebrow_elevation_left	-0.17	0.0420	0.3916	False	Tongue
3.3_eyebrow_shape_left	-0.17	0.0390	0.3916	False	Tongue
3.4_eyebrow_tilt_right	-0.18	0.0302	0.3916	False	Tongue
3.1_eyebrow_elevation_left	-0.18	0.0273	0.3916	False	Tongue
3.5_eyebrow_shape_left	-0.19	0.0243	0.3916	False	Tongue
3.5_eyebrow_shape_right	-0.19	0.0219	0.3916	False	Tongue
3.3_forehead_entropy	-0.21	0.0126	0.3916	False	Tongue
3.3_lateral_canthal_entropy_right	-0.21	0.0123	0.3916	False	Tongue
3.2_eyebrow_shape_left	-0.23	0.0052	0.3916	False	Tongue

Table 3.3: Spearman correlation analysis of features with Faciokinesis total scores, presented both before and after applying False Discovery Rate (FDR) correction. Only correlations yielding statistical significance ($p < 0.05$) are displayed.

Biomarker	**Spearman Correlation**	***p*-value**	***p*-value corrected**	**Significant after FDR**
1.5_eyebrow_tilt_right	-0.17	0.0362	0.6276	False
3.2_mouth_corner_distance_right	-0.17	0.0347	0.6276	False
3.1_upper_lip_distance	-0.18	0.0316	0.6276	False
1.1_mouth_corner_distance_right	-0.18	0.0291	0.6276	False
1.5_lateral_canthal_entropy_right	-0.18	0.0292	0.6276	False
monolog_eye_area_right	-0.18	0.0309	0.6276	False
1.4_mouth_corner_distance_left	-0.19	0.0248	0.6276	False
1.4_mouth_corner_distance_right	-0.19	0.0185	0.6276	False
2.3_upper_lip_distance	-0.20	0.0176	0.6276	False
2.3_nose_entropy	-0.23	0.0055	0.4510	False
2.3_cheeks_entropy	-0.25	0.0023	0.2864	False
9.2_mouth_corner_distance_right	-0.31	0.0001	0.0359	True

Overall, considering all correlation tests no consistent linear or monotonic correlation is observed between the extracted features and the task scores; where such correlations exist, they tend to be weak and sporadic.

3.2 XGBoost model performance

The performance of the XGBoost models is summarized in Table 3.4, which includes results for models trained on individual tasks, group-level scores, and the overall F1 subtest score. The labels refer to the task numbers for which the XGBoost model is intended to predict the scores. Additionally, results are shown for a model trained specifically to predict the faciokinesis test score using only the top 20 most relevant features identified via SHAP values. All models were trained under two settings: using the original dataset and using an oversampled version to address class imbalance.

Among the individual task models, the best-performing one was task 3.4, which achieved a MAE of 0.205 and an RMSE of 0.301 without oversampling. The oversampled version of this model also performed comparably well with MAE of 0.213 and RMSE of 0.304. This suggests that task 3.4 contained patterns that were particularly learnable by the model. The second-best model was task 2.1, which also showed stronger results without oversampling, the MAE in this case was 0.253. On the other hand, the weakest performing models were those for tasks 2.3 and 2.4, with MAE of 0.662 and 0.625 and RMSE of 0.771 and 0.738 respectively. The errors indicate that the XGBoost algorithm struggled to capture meaningful patterns for accurate regression in these cases.

Models predicting task group scores yielded relatively similar performance levels. Among them, the best was group 3, corresponding to tasks associated with tongue movements, achieving 1.115 MAE and 1.318 RMSE. Regarding the total F1 score prediction models, the version trained without oversampling using selected features outperformed the full-feature model, the error was 2.249 in contrast to 2.418. However, the trend reversed in the oversampled setting, where the model trained on all features achieved slightly better results.

When comparing models, Equal Error Rate (EER) serves as the most informative metric due to its normalization of MAE across target ranges. The EER was found to depend on both the size of the feature set and the range of values the model needed to predict. Based on this, the model labeled as F1* achieved the lowest error, with a value of 0.075.

In general, oversampling slightly degraded overall model performance, although exceptions existed. For instance, in two out of three group score models, oversampling led to improvements in evaluation metrics.

Furthermore, per-class evaluation metrics were computed to better illustrate the average deviation between true and predicted values within each score class. These results are reported in Tables 3.5, 3.6, 3.7, 3.8, 3.9 and 3.10, where the number of samples within each target class is illustrated, as well as the corresponding prediction errors produced by the model for each class. It is evident that performance varies across target classes, certain scores are predicted with greater accuracy than others. This discrepancy can be attributed to imbalances in the target distribution, which biased the model toward overrepresented classes during training.

Interestingly, in some cases, oversampling improved prediction accuracy for minority classes. For example, comparing the results in Tables 3.5 and 3.6 for the 1.1 model, the 0-class metrics showed clear improvement in the oversampled version, the MAE decreased from 1.527 to 1.374. This improvement is also noticeable when examining individual predictions: oversampling enabled the model to more confidently output rare values, whereas the original data often led the model to predict values too close to the class average. This effect highlights how oversampling can mitigate bias introduced by class imbalance.

However, this benefit came at a cost: improved minority class performance sometimes coincided with reduced performance on majority classes, likely due to added noise and occasional overfitting introduced by the synthetic examples.

The results of the CNN models used for comparison are presented in Table 3.11. Overall, the error values were consistently higher across nearly all tasks compared to the XGBoost models. Nevertheless, the best-performing CNN model was again associated with Task 3.4, with MAE of 0.341 and RMSE of 0.197, mirroring the trend observed with the XGBoost approach. Furthermore, evaluation over broader scoring ranges appeared to benefit the CNN models, as evidenced by the relatively low EER values achieved in models predicting group scores, particularly those for the chin, lips, and tongue task groups, as well as the composite faciokinesis F1 score.

3.3 Feature interpretability

Feature interpretability was assessed across all model variants, both with and without the application of oversampling techniques. In the absence of oversampling, the calculated feature importances were generally low and equally distributed across all features. For models predicting individual task scores, importance values typically remained around 0.1. Slightly higher error values were observed for models 2.3 and 2.4, which correspond to tasks involving chin movements. In these cases,

Table 3.4: Final model metrics for all trained XGBoost models evaluated with 10-fold CV. F1* indicates the model for predicting total F1 scores with choosing the first 20 most relevant feature.

	Metrics without oversampling				**Metrics with oversampling**			
Model	**MAE**	**MSE**	**RMSE**	**EER**	**MAE**	**MSE**	**RMSE**	**EER**
1.1	0.417	0.257	0.505	0.208	0.410	0.247	0.494	0.205
1.4	0.326	0.162	0.398	0.163	0.330	0.175	0.414	0.165
1.5	0.391	0.232	0.478	0.195	0.407	0.264	0.509	0.204
2.1	0.253	0.136	0.351	0.126	0.269	0.142	0.368	0.134
2.3	0.662	0.599	0.771	0.331	0.666	0.628	0.788	0.333
2.4	0.625	0.554	0.738	0.312	0.628	0.567	0.747	0.314
3.1	0.341	0.199	0.436	0.171	0.374	0.220	0.460	0.187
3.2	0.325	0.221	0.461	0.162	0.393	0.302	0.536	0.197
3.3	0.509	0.358	0.590	0.255	0.510	0.383	0.611	0.255
3.4	0.205	0.096	0.301	0.103	0.213	0.098	0.304	0.107
3.5	0.440	0.263	0.509	0.220	0.449	0.273	0.517	0.225
1	1.345	2.907	1.658	0.134	1.195	2.543	1.549	0.120
2	1.263	2.349	1.521	0.126	1.317	2.517	1.566	0.132
3	1.115	1.776	1.318	0.112	1.097	1.751	1.301	0.110
F1	2.418	9.684	3.024	0.081	2.507	10.685	3.182	0.084
F1*	2.249	9.498	3.002	0.075	2.734	12.391	3.461	0.091

Table 3.5: Metrics per score classes in task-scoring model without oversampling, evaluated with 10-fold CV

Target class	**MAE**	**MSE**	**RMSE**	**EER**	**Sample count**	**Model**
0	1.527	2.331	1.527	0.763	1	1.1
1	0.659	0.472	0.683	0.330	37	
2	0.320	0.161	0.393	0.160	110	
1	0.686	0.491	0.692	0.343	25	1.4
2	0.249	0.091	0.297	0.124	122	
0	1.332	1.808	1.332	0.666	2	1.5
1	0.553	0.385	0.605	0.277	31	
2	0.314	0.147	0.370	0.157	114	
1	0.838	0.748	0.841	0.419	17	2.1
2	0.172	0.050	0.220	0.086	131	
0	1.195	1.568	1.233	0.597	25	2.3
1	0.309	0.130	0.347	0.154	52	
2	0.745	0.625	0.782	0.372	71	
0	0.852	0.818	0.889	0.426	46	2.4
1	0.311	0.133	0.362	0.156	69	
2	0.982	1.089	1.020	0.491	34	
1	0.824	0.727	0.833	0.412	26	3.1
2	0.232	0.082	0.279	0.116	121	
0	1.608	2.596	1.608	0.804	3	3.2
1	0.760	0.630	0.788	0.380	20	
2	0.227	0.101	0.304	0.113	124	
0	1.470	2.235	1.470	0.735	5	3.3
1	0.580	0.401	0.627	0.290	53	
2	0.429	0.245	0.486	0.215	89	
1	0.756	0.601	0.762	0.378	15	3.4
2	0.138	0.033	0.175	0.069	132	
0	1.248	1.562	1.248	0.624	2	3.5
1	0.436	0.265	0.510	0.218	63	
2	0.429	0.237	0.485	0.214	83	

Table 3.6: Metrics per score classes in task-scoring model with oversampling, evaluated with 10-fold CV

Target class	MAE	MSE	RMSE	EER	Sample count	Model
0	1.374	1.887	1.374	0.687	21	1.1
1	0.646	0.473	0.678	0.323	63	
2	0.324	0.155	0.389	0.162	110	
1	0.705	0.538	0.723	0.353	55	1.4
2	0.247	0.093	0.298	0.123	122	
0	1.291	1.668	1.291	0.645	22	1.5
1	0.643	0.472	0.676	0.322	59	
2	0.313	0.164	0.390	0.157	114	
1	0.803	0.700	0.809	0.401	47	2.1
2	0.194	0.064	0.245	0.097	131	
0	1.228	1.605	1.255	0.614	42	2.3
1	0.322	0.156	0.366	0.161	52	
2	0.727	0.651	0.794	0.364	71	
0	0.840	0.786	0.879	0.420	46	2.4
1	0.303	0.132	0.358	0.152	69	
2	1.065	1.272	1.103	0.533	34	
1	0.807	0.705	0.814	0.404	56	3.1
2	0.281	0.122	0.337	0.141	121	
0	1.874	3.521	1.874	0.937	23	3.2
1	0.697	0.577	0.745	0.348	49	
2	0.313	0.184	0.420	0.156	124	
0	1.609	2.779	1.618	0.805	21	3.3
1	0.505	0.348	0.581	0.252	53	
2	0.458	0.278	0.518	0.229	89	
1	0.656	0.478	0.676	0.328	44	3.4
2	0.157	0.047	0.207	0.079	132	
0	1.285	1.653	1.285	0.642	22	3.5
1	0.448	0.268	0.511	0.224	63	
2	0.435	0.248	0.492	0.218	83	

Table 3.7: Metrics per score classes in group-scoring models without oversampling, evaluated with 10-fold CV

Target class	MAE	MSE	RMSE	EER	Sample count	Model
2	5.962	35.543	5.962	0.596	1	1
5	4.641	21.543	4.641	0.464	1	
6	2.415	6.087	2.430	0.241	8	
7	1.146	1.939	1.209	0.115	5	
8	0.714	0.707	0.805	0.071	21	
9	0.977	2.235	1.069	0.098	29	
10	1.528	3.134	1.736	0.153	84	
2	4.698	22.068	4.698	0.470	1	2
4	3.684	13.573	3.684	0.368	1	
5	2.183	4.935	2.195	0.218	6	
6	1.407	2.451	1.520	0.141	25	
7	0.737	0.771	0.776	0.074	17	
8	0.601	0.599	0.720	0.060	40	
9	1.286	2.015	1.369	0.129	29	
10	2.164	5.077	2.204	0.216	30	
4	2.323	5.397	2.323	0.232	1	3
5	2.321	5.560	2.356	0.232	3	
6	1.967	4.105	2.002	0.197	7	
7	0.956	1.139	1.022	0.096	15	
8	0.644	0.626	0.734	0.064	34	
9	0.836	1.072	0.979	0.084	38	
10	1.510	2.665	1.593	0.151	51	

Table 3.8: Metrics per score classes in group-scoring models with oversampling, evaluated with 10-fold CV

Target class	MAE	MSE	RMSE	EER	Sample count	Model
2	7.550	57.001	7.550	0.755	21	1
5	4.810	23.135	4.810	0.481	21	
6	2.380	6.668	2.384	0.238	25	
7	1.884	3.752	1.922	0.188	20	
8	1.023	1.335	1.130	0.102	42	
9	0.632	0.557	0.687	0.063	29	
10	1.264	2.492	1.527	0.126	84	
2	4.513	20.367	4.513	0.451	21	2
4	3.567	12.721	3.567	0.357	21	
5	2.360	5.852	2.362	0.236	15	
6	1.588	3.184	1.686	0.159	25	
7	1.065	1.713	1.107	0.106	17	
8	0.580	0.564	0.701	0.058	40	
9	1.183	1.811	1.292	0.118	29	
10	2.026	4.548	2.082	0.203	30	
4	2.466	6.080	2.466	0.247	21	3
5	2.378	5.729	2.393	0.238	23	
6	2.035	4.867	2.151	0.204	18	
7	1.087	1.558	1.162	0.109	27	
8	0.693	0.722	0.782	0.069	34	
9	0.791	0.963	0.905	0.079	38	
10	1.354	2.200	1.433	0.135	51	

Table 3.9: Metrics per score classes in F1-scoring model without oversampling, evaluated with 10-fold CV

Target class	**MAE**	**MSE**	**RMSE**	**EER**	**Sample count**
9	13.652	186.364	13.652	0.455	1
17	7.797	60.800	7.797	0.260	1
18	7.743	61.384	7.743	0.258	2
19	5.116	26.944	5.116	0.171	3
20	5.157	26.595	5.157	0.172	1
21	4.001	20.128	4.001	0.133	6
22	2.058	5.356	2.095	0.069	8
23	1.918	5.327	2.079	0.064	12
24	1.515	2.954	1.635	0.051	15
25	1.857	5.684	1.938	0.062	12
26	1.023	1.496	1.116	0.034	26
27	1.103	1.579	1.130	0.037	8
28	2.330	6.626	2.403	0.078	20
29	2.214	5.665	2.306	0.074	14
30	3.927	17.684	3.988	0.131	20

Table 3.10: Metrics per score classes in F1-scoring model with oversampling, evaluated with 10-fold CV

Target class	**MAE**	**MSE**	**RMSE**	**EER**	**Sample count**
9	13.470	181.428	13.470	0.449	21
17	9.005	81.085	9.005	0.300	21
18	6.625	47.534	6.625	0.221	22
19	4.903	25.677	4.903	0.163	23
20	5.012	25.120	5.012	0.167	21
21	3.863	19.170	3.863	0.129	12
22	2.475	6.633	2.501	0.082	8
23	2.217	6.516	2.252	0.074	12
24	1.133	1.970	1.283	0.038	15
25	2.105	13.452	2.143	0.070	12
26	1.098	1.583	1.182	0.037	26
27	1.217	2.711	1.295	0.041	8
28	2.499	8.267	2.587	0.083	20
29	2.433	7.093	2.479	0.081	14
30	4.368	20.170	4.436	0.146	20

Table 3.11: Resulting metrics for the CNN model

Model	MAE	MSE	RMSE	EER
1.1	0.450	0.301	0.545	0.225
1.4	0.435	0.294	0.524	0.218
1.5	0.521	0.402	0.622	0.261
2.1	0.383	0.259	0.497	0.192
2.3	0.738	0.821	0.901	0.369
2.4	0.695	0.777	0.861	0.347
3.1	0.435	0.271	0.511	0.218
3.2	0.446	0.358	0.582	0.223
3.3	0.592	0.501	0.699	0.296
3.4	0.341	0.197	0.436	0.171
3.5	0.502	0.346	0.585	0.251
1	1.724	5.172	2.205	0.172
2	1.667	4.415	2.070	0.167
3	1.643	4.278	2.025	0.164
F1	3.755	23.675	4.657	0.125

features such as cheek and jaw entropy, as well as jaw distance and nose entropy, contributed more significantly to the model's learning process. In the case of models targeting group-level score predictions, interpretability improved modestly: the highest individual feature importances reached values of approximately 0.4, with several additional features achieving values around 0.2. However, for the model predicting the overall faciokinesis score, feature importance values again declined, and even after performing feature selection based on SHAP values, the magnitudes of importance remained limited. In contrast, models trained with oversampled data consistently demonstrated higher feature importances. This suggests that the oversampling process contributed not only to improved class balance but also to a clearer differentiation of the most relevant predictive features.

In several models, it is not possible to clearly determine whether one specific feature is universally more informative than others in the context of model evaluation. However, there are multiple instances where one or two features received notably higher SHAP values than the rest. For example, in model 1.1, jaw distance and mouth corner distance right stood out; in model 1.5, it was mouth corner distance left; in model 2.3, cheeks entropy was dominant. In model 3.1, eyebrow elevation left and lower lip distance appeared as most relevant; in 3.2, eye area left, lateral canthal entropy right, and eyebrow shape left were prominent. For model 3.3, lateral canthal entropy left and jaw distance showed the strongest contributions, while in

3.4 and 3.5, eyebrow shape right and eyebrow shape left were the most influential, respectively.

Notably, the most important features remained largely consistent between models trained with and without oversampling, even if there were slight shifts in their ranking. One exception was found in model 1.1, where jaw distance rose to the top position in the oversampled version, whereas it had not even ranked among the top ten in the version trained on the original dataset.

The beeswarm plots presented in Appendix A summarize the SHAP value distributions for each feature. These visualizations, much like the earlier bar plots, provide insights into how each feature contributes to the model's output – not only in terms of average importance but also in terms of directionality. Unlike the bar plots, the beeswarm plots allow one to see whether a high or low value of a feature contributes positively or negatively to the prediction. While in many cases the color distribution does not show clear separability (indicating that output prediction cannot be precisely determined based on feature values alone), certain patterns and tendencies are still discernible.

4 Discussion

This chapter serves to evaluate and interpret the results obtained through the objective scoring tools developed for the Faciokinesis part of Test 3F. While the outcomes of the correlation analysis yielded relatively low values overall, a number of interesting and consistent associations emerged between the facial movement–based biomarkers and the F1 facial scores. These findings suggest that, despite weak global correlations, certain features may still carry relevant predictive signals.

The performance metrics of the XGBoost regression models indicate that, in general, these models performed well across most tasks. This discussion also addresses the advantages and limitations of applying minority class oversampling in these models. The results show that XGBoost outperformed CNNs in the majority of tasks, particularly when predicting individual facial scores, highlighting the robustness of tree-based methods in small-data scenarios.

Furthermore, feature importance values calculated using the SHAP technique were notably higher in models trained with oversampling. This suggests that oversampling not only impacted performance but also made feature contributions more distinguishable. This chapter aims to explore these relationships in greater detail, providing insight into the internal decision-making processes of the models and their implications for facial function assessment.

4.1 Interpretation of correlations

Although the overall correlation between the features derived from facial landmarks and the task-related scores was generally low, and only a few remained statistically significant after applying FDR correction, several intuitively meaningful relationships can still be observed in the data.

In Task 1.4, where the objective was to pull the corners of the mouth into a smile, participants who showed greater changes in the position of their right and left eyebrows during the movement received higher scores. Additionally, the degree of forehead entropy may also play a role, where a higher standard deviation of entropy could indicate changes in skin texture and the formation of wrinkle lines during the task. These movements are involuntary and are typically observed in healthy individuals when smiling.

In Task 1.1, where the task involved pulling the lips inward between the teeth, a greater change in the distance between the mouth corners showed a negative correlation with the scores. Physiologically, this can be explained by the fact that individuals who perform the task easily can complete it quickly and with a smaller range of motion, whereas those who received lower scores may have attempted the movement multiple times and performed it with less fluidity.

In Task 2.4, where the task was to circle with the chin in both directions, a positive correlation was found with jaw entropy, a relationship that is intuitively understandable. A similar pattern appeared in Task 2.3, which involved moving the chin side-to-side. Here, cheek entropy was negatively correlated with the task scores, suggesting that individuals who performed the task incorrectly may have attempted to compensate for missing movements with the activation of other facial muscles, which can be observed through increased cheek movement.

In Task 3.1, where participants had to stick out and retract their tongue, greater movement in the lower lip distance and left eyebrow elevation showed a negative correlation with the scores. The more these facial areas moved during the task, the lower the expected performance score. This likely reflects multiple attempts and compensatory movements. A similar pattern may explain the findings in Tasks 3.3, 3.4, and 3.5, where only negative correlations with certain features were observed. Interestingly, although these tasks are associated with tongue movements, the correlated features mainly involve the eye region. For example, for Tasks 3.3 and 3.4, lower lateral canthal entropy and lower right eyebrow shape deformation correlated with better performance. The increased movement of unrelated facial areas may reflect poor neuromotor control, where participants cannot isolate specific movements efficiently.

The only feature showing significance after FDR correction with the group scores was the cheek entropy in Task 2.3, which shows a negative correlation with the scores of chin-related tasks. In other words, greater variability in chin texture is associated with lower task scores.

Interestingly in the tongue movement tasks, variability in the movement of the eyebrows, eyes, and forehead also showed negative correlations with performance scores. This may suggest that individuals who struggle with task execution engage in compensatory or global facial movements, possibly reflecting difficulties with fine motor control or increased cognitive effort during the tasks.

Features that showed correlations with the scores of F1 without FDR correction were negative, suggesting that, in general, though not consistently, greater variability in the movements of certain facial areas is associated with lower scores on the overall Faciokinesis test. It indicates that higher variability usually suggests more variable or uncontrolled movements, and more stable and coordinated facial movement is characteristic of better task execution.

The absence of strong correlations in some tasks may indicate either redundancy in features, low variability in the scoring, or subtle facial movements not captured effectively by the current biomarker set. It is also important to note, the facial movements in these videos are related to task performance, however they are likely only one of several contributing factors. Cognitive load, emotional responses, or individual anatomical differences may affect these relationships. This makes it challenging to identify potential biomarkers that could reliably predict task scores based on simple statistical associations. However, this limitation does not hinder the performance of the XGBoost model, as the algorithm is capable of simultaneously considering multiple features and capturing complex, non-linear relationships that are often beyond the scope of traditional statistical methods. Therefore, this analysis serves as an exploratory step prior to more advanced modeling techniques.

4.2 Effectiveness of the models

The XGBoost regression models proved to be a highly suitable choice for the challenges posed by this task. The dataset was significantly skewed and featured a relatively high number of input variables, what are characteristics that are typically problematic for many machine learning techniques. XGBoost, however, is particularly well-equipped to handle such conditions. In fact, in the case of predicting overall F1 subtest scores and subgroup scores (where features from multiple tasks were utilized simultaneously), the model capitalized on the feature richness, resulting in notably lower Equal Error Rate (EER) values. It is worth noting, however, that this improvement might also be partially attributed to the broader prediction range in these cases.

As highlighted in the results, XGBoost consistently outperformed the convolutional neural network models, particularly in individual score prediction tasks, where error metrics such as MAE and RMSE were markedly lower. This demonstrates XGBoost's robustness in handling structured tabular data and its ability to extract relevant patterns even from relatively small and imbalanced datasets.

Interestingly, the oversampling techniques applied to mitigate class imbalance did not result in overall error reduction. This suggests that the original distribution of features already carried strong predictive signals, particularly for the majority classes. However, a notable improvement was observed in the per-class prediction errors, especially for underrepresented classes. Through oversampling, the model was better able to attend to these classes, achieving a level of sensitivity that would have been more difficult to reach using simple class weighting strategies. This highlights XGBoost's flexibility in adapting to data augmentation strategies while preserving its predictive capacity.

The findings of this work that XGBoost regression models are highly effective for faciokinesis test scoring are consistent with recent literature examining machine learning approaches for structured facial movement data. Recent studies have used machine learning, including XGBoost, to assess facial expressivity in Parkinson's disease patients by analyzing facial action units. For instance, models were trained to detect hypomimia, which is conceptually similar to faciokinesis scoring. These studies found that tree-based models like XGBoost can robustly classify and quantify subtle changes in facial muscle activity, often outperforming simple neural networks in structured AU data and small, imbalanced datasets. [42]

Hybrid approaches combining deep learning for feature extraction and XGBoost for classification have shown strong performance in other studies too. For instance, the Face-GPS study demonstrated that XGBoost could effectively classify emotions from AU-derived features, achieving high accuracy and interpretability, with qualities valuable for clinical scoring as well. [43]

Furthermore, XGBoost has been applied to predict depression severity from facial movement data, particularly using AU features. These models have achieved competitive error rates and have been noted for their ability to handle imbalanced data, similar to the challenges in faciokinesis scoring. [44] Taken together, these findings support the conclusion that XGBoost is not only well-suited for the prediction tasks presented in this work, but is also a model of choice in the broader context of clinical facial movement analysis.

4.3 Feature importances

SHAP values provide valuable insight into the decision-making process of the XGBoost models. The corresponding figures (A.5, A.6, A.7, A.8, A.11, A.12) reveal how each feature influenced the prediction in a positive or negative direction. When comparing models trained on the oversampled datasets to those trained on the original data, SHAP values tend to be higher overall. Nevertheless, the most important features maintained their top rankings in terms of contribution, suggesting that both model types rely on similar underlying patterns.

In the Model 1.1, jaw distance and right mouth corner distance pushed the predictions in a negative direction. Given the negative correlation between mouth corner distance and the score, this result is intuitive. However, the negative contribution of jaw distance may indicate that participants who performed poorly in this task tended to compensate for limited lip movement by moving their jaw instead.

In Model 1.4, the most influential features were around the eye region, which is consistent with the correlation analysis. Greater eyebrow movement likely reflects better performance in smile-related tasks, while limited motion corresponds to lower scores.

Model 1.5, designed to assess lip pursing and spreading, was most influenced by the distance of the mouth corners from the center. Interestingly, lower values of this feature were associated with higher scores, implying that smaller variability in mouth corner movement was a better indicator of successful task performance.

Entropy-based features often served as effective biomarkers. In Model 2.1, greater jaw entropy was associated with lower scores, while higher cheeks entropy contributed positively to the model's output. This aligns with the nature of the task (mouth opening and closing), where texture variation around these areas is expected.

Despite the similarity of the tasks, Models 2.3 and 2.4 displayed different sets of key features. For Model 2.3, cheeks entropy was significant, with lower values contributing to higher scores. In contrast, Model 2.4 highlighted nose entropy, where higher values led to higher predictions. These patterns were also evident in the correlation analyses.

In Model 3.1, eyebrow elevation left was the most important feature, where lower values predicted higher scores. This relationship mirrored that observed in the correlation data. For Model 3.2, eye area left had the greatest impact. Generally, lower values pushed predictions toward lower scores, though some exceptions appeared at higher score levels. The next most impactful features were lateral canthal entropy right and eyebrow shape left, both inversely related to the predicted outcome. In Model 3.3, lateral canthal entropy left emerged as the most influential, followed closely by jaw distance and lateral canthal entropy right, all of which had a negative relationship with the predicted scores.

Model 3.4 was most influenced by features around the eyes, with eyebrow shape left leading the list. Lower values were associated with better performance, possibly indicating that poor neuromotor control in underperforming participants limited eyebrow shape variability, a finding supported by correlation results. Since this task involved tongue motion, it's notable that eye-related features remained informative. Similarly, Model 3.5 highlighted eyebrow shape left as its most important feature, where lower values again predicted higher scores. Upper lip distance followed closely, suggesting that lower variability in upper lip position helped the model associate the action (tracing the tongue along the lips) with better control and execution.

In the overall Lips composite model, top features included eye area right and cheeks entropy from Task 1.4, and mouth corner distance left and jaw distance from Task 1.5. Particularly interesting was the eye area right, where extremely low values led to lower scores, medium-high values contributed to higher scores, and very high values again slightly decreased predictions. This complex relationship highlights XGBoost's ability to detect nuanced patterns that are not easily explained by simple intuition.

For the Chin composite model, the most impactful features were cheeks entropy from Task 2.3, which was inversely related to the score, and eyebrow elevation on both sides from Task 2.4, where higher values corresponded with better outcomes.

In the Tongue composite model, eye area left (from Task 3.2) and jaw distance (from Task 3.1) were most influential. Both showed an inverse relationship with the prediction, though eye area left again followed the earlier described non-linear trend, similar to that observed in the Lips model. Finally, the composite model for the total Faciokinesis score produced relatively low SHAP values overall. However, the top contributing feature was forehead entropy from Task 3.1, where higher values led to lower predicted scores, again indicating involuntary effort.

In the case of bilateral features, it is interesting to note that while in some instances both sides hold similar importance, it is often the case that only one side significantly influences the model, or they even shift the prediction in opposite directions. These observations may indicate asymmetrical facial movement, which is common in cases of facial palsy.

In summary, the SHAP analysis reveals both intuitive and non-obvious patterns in model decision-making. Some feature relationships are straightforward and supported by correlation analyses, while others reflect complex associations discovered by the XGBoost algorithm that would be difficult to capture through traditional statistical methods or human reasoning alone.

4.4 Future directions

The models developed in this study provide a promising foundation for initiating objective video-based evaluation of faciokinesis, which could ultimately support clinical decision-making. However, the research is not without limitations, most notably those related to the dataset. This is a common issue in similar studies, as real-world clinical datasets often suffer from underrepresentation of certain patient groups. While XGBoost is known for its ability to perform well with smaller datasets, the models would undoubtedly benefit from increased data volume, particularly from those cases that are currently sparsely represented, or entirely missing from the training set. Increasing sample diversity would enable the models to better capture subtle variations in facial movement and improve their robustness for practical deployment.

A larger dataset would also be essential if deep neural networks were to be leveraged for this task. CNNs are inherently limited in modeling the temporal dynamics of videos and may not be well-suited for capturing the sequential patterns of facial motion. Future work could explore the use of architectures specifically designed for temporal modeling, such as recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks, or Transformer-based models, which are more adept at learning time-dependent representations.

In addition, it would be beneficial to test the system using a different but related facial assessment dataset. This could aid in fine-tuning the models and improving generalizability. While there are still some options available within the current dataset, its potential for continued performance improvement may soon be exhausted. Further experimentation with data augmentation techniques could prove valuable, but care must be taken to ensure that the generated samples are sufficiently diverse and realistic.

For the system to be clinically useful, it is crucial to determine whether the prediction error margins for the generated scores are within an acceptable range. To strengthen the real-world utility of the framework, future research should involve clinical validation, including expert-rated comparisons and tracking of patient outcomes. Aligning model outputs with clinician feedback would enhance both trust and interpretability, and could potentially lead to the development of a hybrid human-AI scoring interface. In this case, the processing pipeline should be adapted for real-time use. This would require minimizing latency in both data preprocessing and prediction steps. While the tools used in this work are relatively efficient compared to alternatives, additional optimization – or the integration of faster software and hardware solutions – will be necessary to meet the demands of real-time clinical environments.

Conclusion

This diploma thesis' aim was to create a method developed to automatically determine the score of the Faciokinesis component of Test 3F based on video recordings of facial expressions. A comprehensive understanding of the problem was obtained through a literature review covering topics such as disorders of faciokinesis, dysarthria, and faciokinesis assessment. This revealed that faciokinesis is a widely observed clinical feature with abnormalities often serving as indicators of neural damage. Moreover, it is closely associated to dysarthria, which is characterised by damage of the muscles responsible for sound production. While faciokinesis disorders can have various origins, the most frequently reported in the literature are hypokinetic dysarthria, a type of facial muscle paralysis associated with Parkinson's disease, and Bell's palsy. For these diseases is a common approach to quantify the degree of impairment by a test, thus different scales are used.

In the Czech Republic, the Test 3F tool is used to determine the dysarthric profile, it has 3 subsections, of which one is Faciokinesis. It is further divided into 3 subregions, which address the lips, jaw and tongue. More broadly, there are a number of promising studies that use computerized analysis to detect abnormalities in both healthy individuals and those with impairments. A variety of machine learning models have been used for classification, including SVM, XGBoost, statistical analysis, and neural networks.

Basic statistical descriptors were defined to describe the database. This work also includes steps for processing the raw data, resulting in shorter video segments extracted from the videos of each individual undergoing testing, which include the performance of a single task from the F1 test. After selecting the relevant tasks, there are thus 13 video segments for each test subject. In order to make the videos quantifiable, facial landmark detection was run on them, resulting in 468 landmarks. During data preparation, the most recent tools were selected from a range of implementations to provide the most efficient and robust processing while maintaining the quality of the information.

The biomarkers used as features for training the machine learning models were selected and computed based on the methodology proposed by Novotny et al. [2]. Upon examining the extracted features, outlier handling and the removal of potential confounding factors were necessary. Subsequently, a statistical analysis was conducted on these biomarkers, including a correlation analysis between feature values and task scores. For this purpose, Spearman's rank correlation was employed, revealing generally weak associations.

The machine learning pipeline built on these features followed a largely consistent structure across all models. Hyperparameter optimization leveraged the practical advantages of RandomizedSearchCV, while final model performance was evaluated using cross-validation across four different metrics. Due to the underrepresentation of certain score groups within the training data, corrective measures were implemented, including the construction of a custom weighting function, as well as experiments with SMOTE-based and manual oversampling. The impact of these approaches on model performance was also systematically assessed.

The results demonstrated that the proposed XGBoost-based approach was capable of achieving a MAE of 0.205 and a RMSE of 0.301 when predicting individual task scores. For the total Faciokinesis score, the best model reached an MAE of 2.249 and RMSE of 3.002. These results confirm the feasibility of automatic scoring using machine learning methods. Additionally, SHAP value analysis provided interpretable insights into model behavior, revealing clinically consistent patterns such as reduced or asymmetrical facial movements being associated with lower scores.

These findings highlight the potential of machine learning approaches, particularly XGBoost, to support objective and interpretable assessment of facial motor function, offering a valuable tool for enhancing clinical diagnostics in neurological disorders.

Bibliography

[1] Howard S. Kirshner and Martin A. Samuels. Speech and language disorders. In *Neurologic Localization and Diagnosis*, pages 177–189. Elsevier, 2023. doi:10.1016/B978-0-323-81280-1.00031-7.

[2] Michal Novotny, Tereza Tykalova, Hana Ruzickova, Evzen Ruzicka, Petr Dusek, and Jan Rusz. Automated video-based assessment of facial bradykinesia in de-novo Parkinson's disease. *Npj Digital Medicine*, 5(1):Article number: 98, 2022. doi:10.1038/s41746-022-00642-5.

[3] Luigi Cattaneo and Giovanni Pavesi. The facial motor system. *Neuroscience & Biobehavioral Reviews*, 38:135–159, 2014. doi:10.1016/j.neubiorev.2013.11.002.

[4] M. Bologna, G. Fabbrini, L. Marsili, G. Defazio, P. D. Thompson, and A. Berardelli. Facial bradykinesia. *Journal of Neurology, Neurosurgery & Psychiatry*, 84(6):681–685, 2013. doi:10.1136/jnnp-2012-303993.

[5] Jacqueline C. Junn and Peter M. Som. Maxillofacial Skeleton and Facial Anatomy. *Neuroimaging Clinics of North America*, 32(4):735–748, 2022. doi:10.1016/j.nic.2022.07.008.

[6] Gabriela Zamišková, Pavel Ressner, Jana Dlouhá, and Dana Šigutová. Poruchy řeči u Parkinsonovy nemoci. *Neurologie pro praxi*, 11(2):112–116, 2010.

[7] Melissa J. Armstrong and Michael S. Okun. Diagnosis and Treatment of Parkinson Disease. *JAMA*, 323(6):548–560, 2020. doi:10.1001/jama.2019.22360.

[8] Abeer Muneer Altaher, Shin Ying Chu, Rahayu binti Mustaffa Kam, and Rogayah A Razak. A Report of Assessment Tools for Individuals with Dysarthria. *The Open Public Health Journal*, 12(1):384–386, 2019. doi:10.2174/1874944501912010384.

[9] Nikolaus P. Schumann, Kevin Bongers, Hans C. Scholle, Orlando Guntinas-Lichius, and Yingchun Zhang. Atlas of voluntary facial muscle activation. *PLOS ONE*, 16(7):e0254932, 2021. doi:10.1371/journal.pone.0254932.

[10] J. Thielker, K. Geißler, T. Granitzka, C. M. Klingner, G. F. Volk, and O. Guntinas-Lichius. Acute Management of Bell's Palsy. *Current Otorhinolaryngology Reports*, 6(2):161–170, 2018. doi:10.1007/s40136-018-0198-0.

[11] Adel Y. Fattah, Anthony D. R. Gurusinghe, Javier Gavilan, Tessa A. Hadlock, Jeff R. Marcus, Henri Marres, Charles C. Nduka, William H. Slattery,

and Alison K. Snyder-Warwick. Facial Nerve Grading Instruments. *Plastic and Reconstructive Surgery*, 135(2):569–579, 2015. doi:10.1097/PRS.0000000000000905.

[12] Caroline A. Banks, Nathan Jowett, Babak Azizzadeh, Carien Beurskens, Prabhat Bhama, Gregory Borschel, Christopher Coombs, Susan Coulson, Glen Croxon, Jaqueline Diels, Adel Fattah, Manfred Frey, Javier Gavilan, Douglas Henstrom, Marc Hohman, Jennifer Kim, Henri Marres, Richard Redett, Alison Snyder-Warwick, and Tessa Hadlock. Worldwide Testing of the eFACE Facial Nerve Clinician-Graded Scale. *Plastic & Reconstructive Surgery*, 139(2):491e–498e, 2017. doi:10.1097/PRS.0000000000002954.

[13] Cuihua Lv, Lizhou Fan, Haiyun Li, Jun Ma, Wenjing Jiang, and Xin Ma. Leveraging multimodal deep learning framework and a comprehensive audio-visual dataset to advance Parkinson's detection. *Biomedical Signal Processing and Control*, 95:106480, 2024. doi:10.1016/j.bspc.2024.106480.

[14] Jaroslava Roubíčková. *Test 3F*. Galén, Praha, 3., dopl. a přeprac. vyd., (v nakl. galén 1.) edition, 2011.

[15] Leonard Knoedler, Maximilian Miragall, Martin Kauke-Navarro, Doha Obed, Maximilian Bauer, Patrick Tißler, Lukas Prantl, Hans-Guenther Machens, Peter Niclas Broer, Helena Baecher, Adriana C. Panayi, Samuel Knoedler, and Andreas Kehrer. A Ready-to-use Grading Tool for Facial Palsy Examiners—Automated Grading System in Facial Palsy Patients Made Easy. *Journal of Personalized Medicine*, 12(10):1739, 2022. doi:10.3390/jpm12101739.

[16] Justyna Skibińska and Jiri Hosek. Computerized analysis of hypomimia and hypokinetic dysarthria for improved diagnosis of Parkinson's disease. *Heliyon*, 9(11):e21175, 2023. doi:10.1016/j.heliyon.2023.e21175.

[17] Andrea Bandini, Silvia Orlandi, Hugo Jair Escalante, Fabio Giovannelli, Massimo Cincotta, Carlos A. Reyes-Garcia, Paola Vanni, Gaetano Zaccara, and Claudia Manfredi. Analysis of facial expressions in Parkinson's disease through video-based automatic methods. *Journal of Neuroscience Methods*, 281:7–20, 2017. doi:10.1016/j.jneumeth.2017.02.006.

[18] Guilherme C. Oliveira, Quoc C. Ngo, Leandro A. Passos, João P. Papa, Danilo S. Jodas, and Dinesh Kumar. Tabular data augmentation for video-based detection of hypomimia in Parkinson's disease. *Computer Methods and Programs in Biomedicine*, 240:107713, 2023. doi:10.1016/j.cmpb.2023.107713.

[19] Fan Xu, Xian wei Zou, Li qiong Yang, Shi cong Mo, Quan hao Guo, Jing Zhang, Xiechuan Weng, and Guo gang Xing. Facial muscle movements in patients with Parkinson's disease undergoing phonation tests. *Frontiers in Neurology*, 13:1018362, 2022. doi:10.3389/fneur.2022.1018362.

[20] Diego L. Guarin, Yana Yunusova, Babak Taati, Joseph R. Dusseldorp, Suresh Mohan, Joana Tavares, Martinus M. van Veen, Emily Fortier, Tessa A. Hadlock, and Nate Jowett. Toward an Automatic System for Computer-Aided Assessment in Facial Palsy. *Facial Plastic Surgery & Aesthetic Medicine*, 22(1):42–49, 2020. doi:10.1089/fpsam.2019.29000.gua.

[21] Mikael Finstad. LosslessCut, 2024. URL: https://github.com/mifi/lossless-cut.

[22] Constantino Álvarez Casado and Miguel Bordallo López. Real-time face alignment. *Journal of Real-Time Image Processing*, 18(6):2239–2267, 2021. doi:10.1007/s11554-021-01107-w.

[23] Kostiantyn Khabarlak and Larysa Koriashkina. Fast facial landmark detection and applications. *Journal of Computer Science and Technology*, 22(1):12–41, 2022. doi:10.24215/16666038.22.e02.

[24] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. doi:10.48550/arxiv.1906.08172.

[25] Google LLC. Face landmark detection guide. URL: https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker.

[26] Zieb Rabie Alqahtani, Mohd Shahrizal Sunar, and Abdelmonim M. Artoli. Comparative analysis of pre-trained deep learning models for facial landmark localization on enhanced dataset of heavily occluded face images. *Journal of Advances in Information Technology*, 15(11):1252–1263, 2024. doi:10.12720/jait.15.11.1252-1263.

[27] Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P. Namboodiri, and C. V. Jawahar. Towards generating ultra-high resolution talking-face videos with lip synchronization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5198–5207. IEEE, 2023. doi:10.1109/WACV56688.2023.00518.

[28] Chuheng Zheng, Mondher Bouazizi, Tomoaki Ohtsuki, Momoko Kitazawa, Toshiro Horigome, and Taishiro Kishimoto. Detecting Dementia from face-related features with automated computational methods. *Bioengineering*, 10(7):862, 2023. doi:10.3390/bioengineering10070862.

[29] Tim Büchner, Sven Sickert, Roland Graßme, Christoph Anders, Orlando Guntinas-Lichius, and Joachim Denzler. Using 2D and 3D face representations to generate comprehensive facial electromyography intensity maps. In *Advances in Visual Computing*, pages 136–147. Springer Nature Switzerland, Cham, 2023. doi:10.1007/978-3-031-47966-3_11.

[30] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time. *CoRR*, abs/2006.10962, 2020. doi:10.48550/arXiv.2006.10962.

[31] R.H. Riffenburgh. Managing results of analysis. In *Statistics in Medicine*, pages 325–343. Elsevier, third edition edition, 2012. doi:10.1016/B978-0-12-384864-2.00015-9.

[32] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, 2016. ACM. doi:10.1145/2939672.2939785.

[33] XGBoost parameters, 2022. URL: https://xgboost.readthedocs.io/en/stable/parameter.html.

[34] Victor Lumumba, Dennis Kiprotich, Mary Mpaine, Njoka Makena, and Musyimi Kavita. Comparative analysis of cross-validation techniques. *American Journal of Theoretical and Applied Statistics*, 13(5):127–137, 2024. doi:10.11648/j.ajtas.20241305.13.

[35] Johannes Allgaier and Rüdiger Pryss. Cross-validation visualized: A narrative guide to advanced methods. *Machine Learning and Knowledge Extraction*, 6(2):1378–1388, 2024. doi:10.3390/make6020065.

[36] XGBoost regression with sample weight – MLflow, 2024. URL: https://www.restack.io/docs/mlflow-knowledge-xgboost-regression-sample-weight-mlflow.

[37] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi:10.1613/jair.953.

[38] Ahmad Hassanat, Ghada Altarawneh, Ibraheem Alkhawaldeh, Yasmeen Al-abdallat, Amir Atiya, Ahmad Abujaber, and Ahmad Tarawneh. The jeopardy of learning from over-sampled class-imbalanced medical datasets. *IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7, 07 2023. doi:10.1109/ISCC58397.2023.10218211.

[39] Yotam Elor and Hadar Averbuch-Elor. To SMOTE, or not to SMOTE?, 2022. arXiv:2201.08528.

[40] SHAP, 2024. URL: https://github.com/shap/shap.

[41] Muhamad Arief Liman and Gede Putra Kusuma. Facial expression recognition using deep learning and neural embeddings. *Revue d'Intelligence Artificielle*, 38(4):1201–1209, 2024. doi:10.18280/ria.380414.

[42] Anas Filali Razzouki, Laetitia Jeancolas, Sara Sambin, Graziella Mangone, Alizé Chalançon, Manon Gomes, Stéphane Lehéricy, Marie Vidailhet, Isabelle Arnulf, Jean-Christophe Corvol, Dijana Petrovska-Delacrétaz, and Mounim A. El-Yacoubi. Explaining facial action units' correlation with hypomimia and clinical scores in Parkinson's disease. *NPJ Parkinson's Disease*, 11(1):11–53, 2025. doi:10.1038/s41531-025-00895-3.

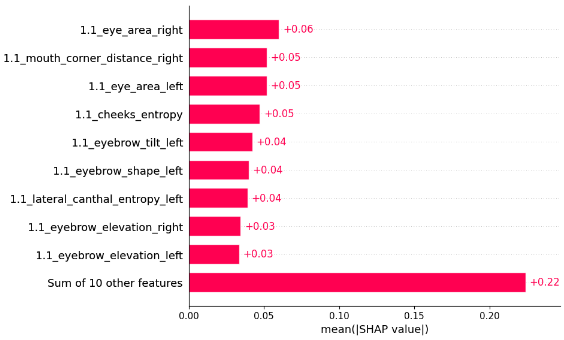
[43] Juni Kim, Zhikang Dong, and Pawel Polak. Face-GPS: A comprehensive technique for quantifying facial muscle dynamics in videos, 2024. arXiv:2401.05625.

[44] Gregorius Natanael Elwirehardja Bens Pardamean Brilyan Nathanael Rumahorbo, Kenjovan Nanggala. Analyzing important statistical features from facial behavior in human depression using XGBoost. *Communications in Mathematical Biology and Neuroscience*, 35:1–24, 2023. doi:10.28919/cmbn/7916.

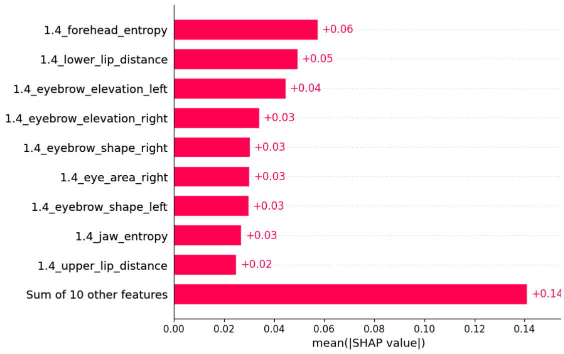
Symbols and abbreviations

AU	(Facial) Action Unit
CNN	Convolutional Neural Network
CV	Cross-validation
EER	Estiamtion Error Ratio
EMG	Electromyography
FDR	False Discovery Rate
FP	Facial Palsy
HBS	House-Brackmann scale
HC	Healthy Control
MAE	Mean Absolute Error
MSE	Mean Squared Error
PD	Parkinson's disease
RMSE	Root Mean Squared Error
SHAP	SHapley Additive exPlanations
SLP	Speech Language Pathologist
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
UPDRS	Unified Parkinson's Disease Rating Scale

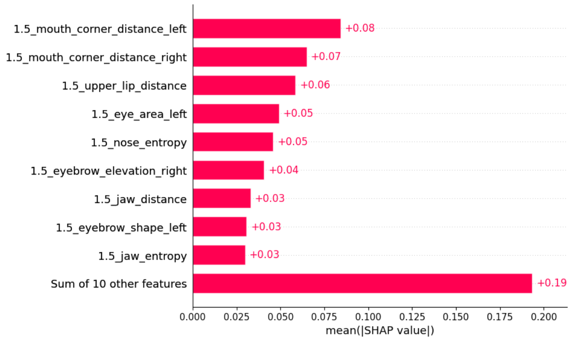
A SHAP values of XGBoost models



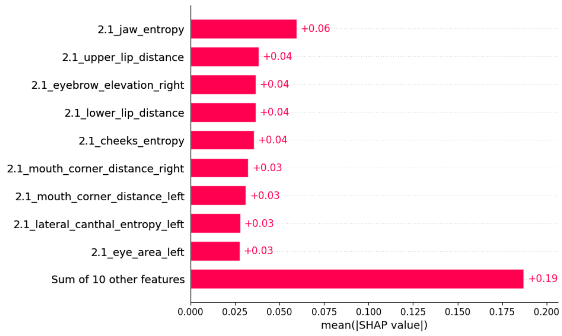
(a) SHAP values of model 1.1



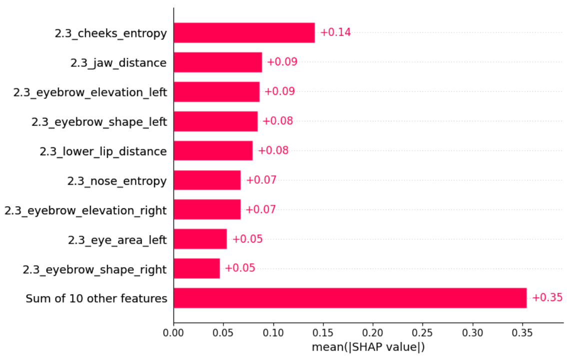
(b) SHAP values of model 1.4



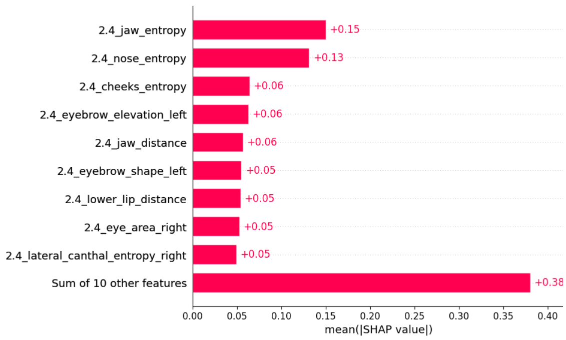
(c) SHAP values of model 1.5



(d) SHAP values of model 2.1

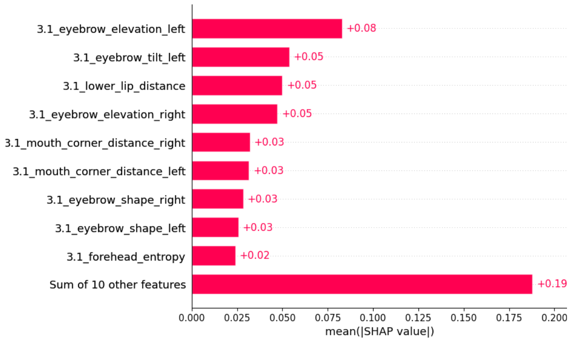


(e) SHAP values of model 2.3

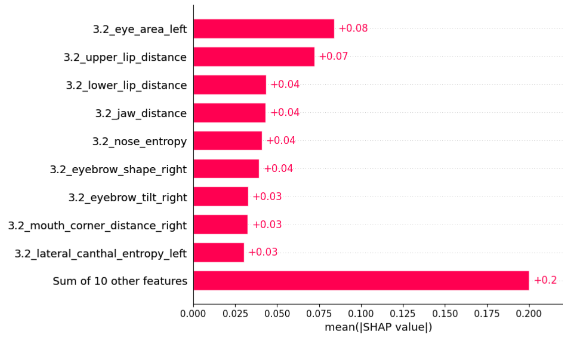


(f) SHAP values of model 2.4

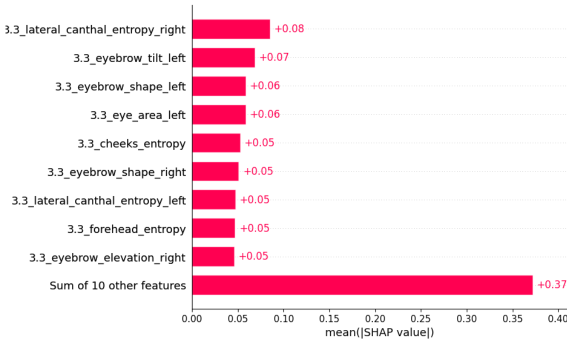
Fig. A.1: Mean absolute SHAP values of features in models without oversampling, for models 1.1, 1.4, 1.5, 2.1, 2.3, 2.4



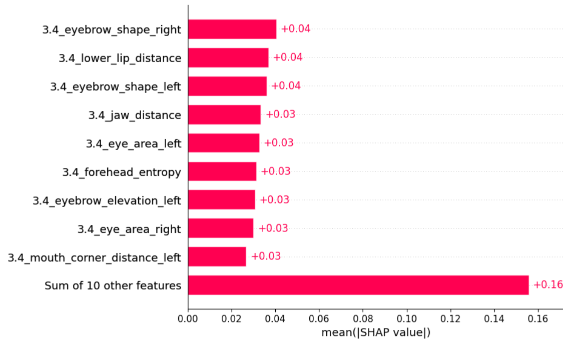
(a) SHAP values of model 3.1



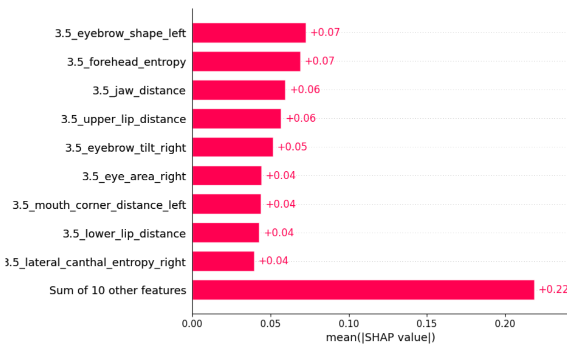
(b) SHAP values of model 3.2



(c) SHAP values of model 3.3

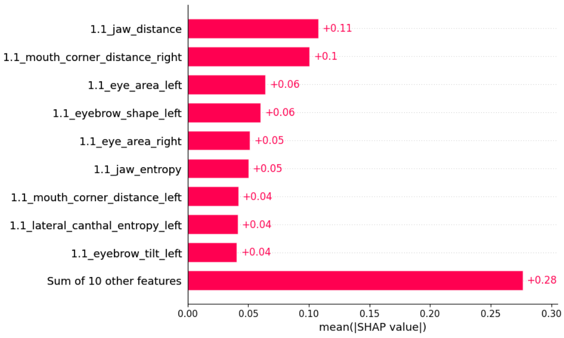


(d) SHAP values of model 3.4

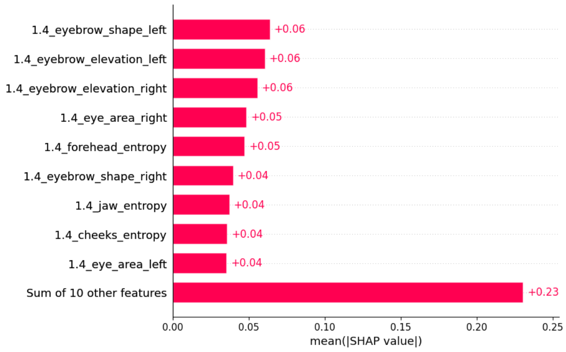


(e) SHAP values of model 3.5

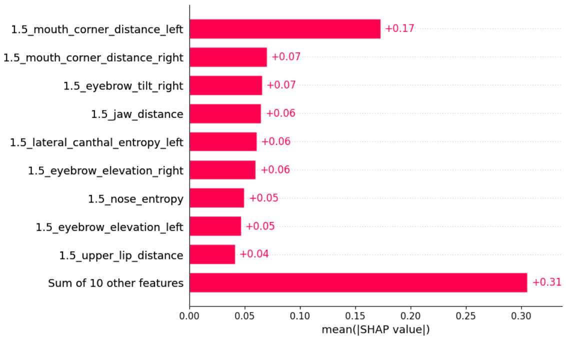
Fig. A.2: Mean absolute SHAP values of features in models without oversampling, for models 3.1, 3.2, 3.3, 3.4, 3.5



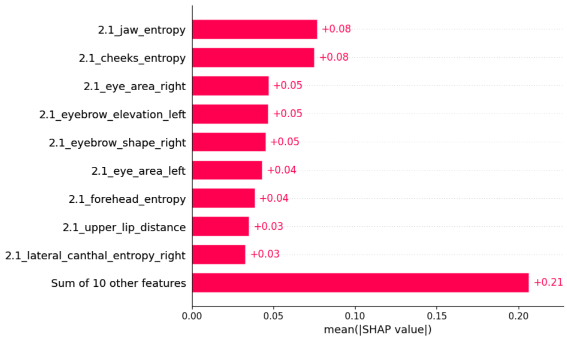
(a) SHAP values of model 1.1



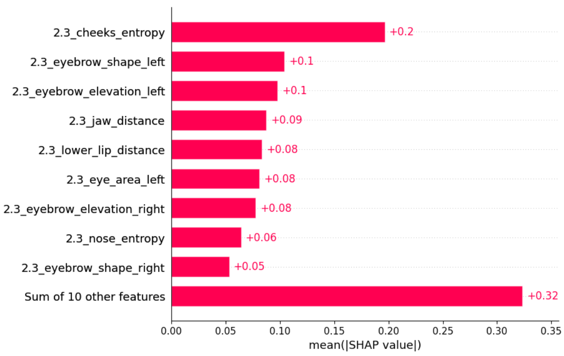
(b) SHAP values of model 1.4



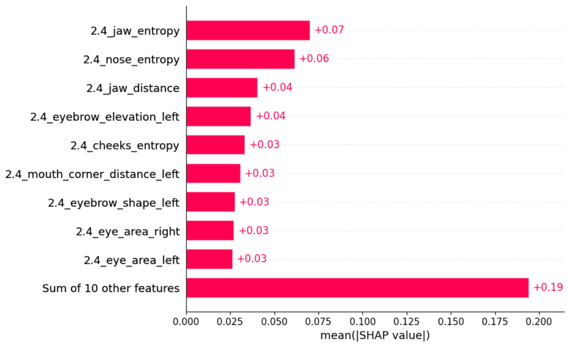
(c) SHAP values of model 1.5



(d) SHAP values of model 2.1

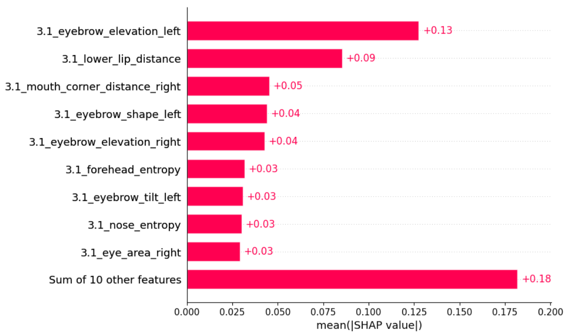


(e) SHAP values of model 2.3

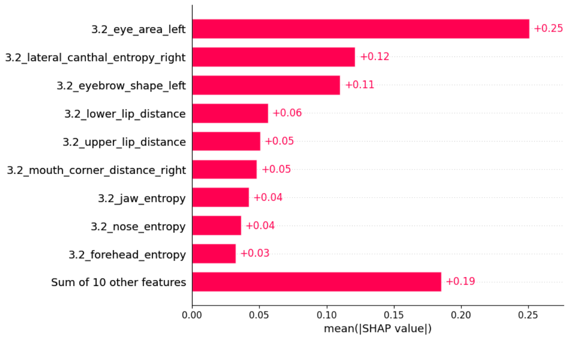


(f) SHAP values of model 2.4

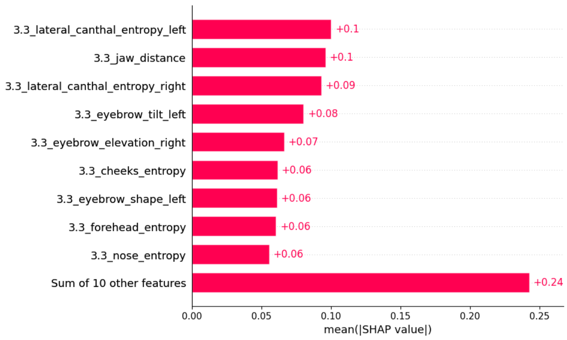
Fig. A.3: Mean absolute SHAP values of features in models with oversampling, for models 1.1, 1.4, 1.5, 2.1, 2.3, 2.4



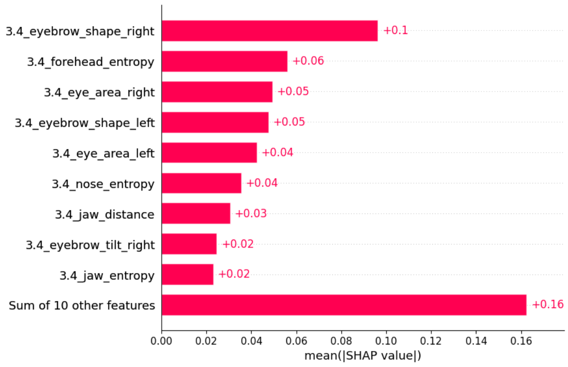
(a) SHAP values of model 3.1



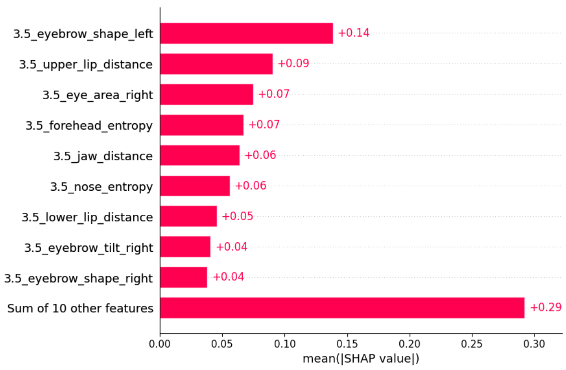
(b) SHAP values of model 3.2



(c) SHAP values of model 3.3

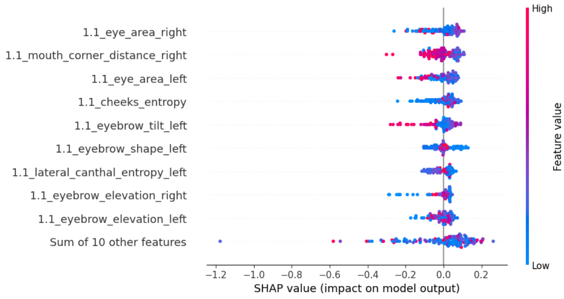


(d) SHAP values of model 3.4

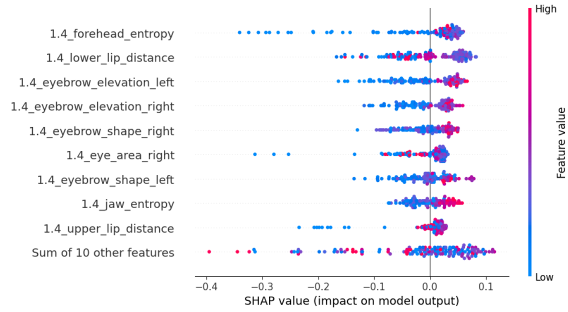


(e) SHAP values of model 3.5

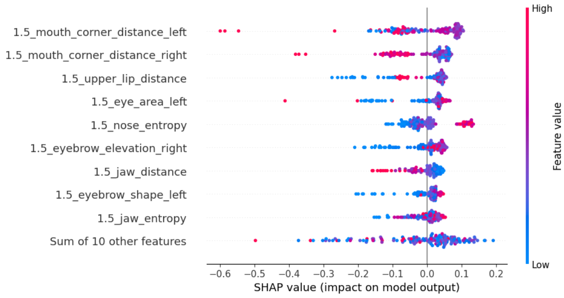
Fig. A.4: Mean absolute SHAP values of features in models with oversampling, for models 3.1, 3.2, 3.3, 3.4, 3.5



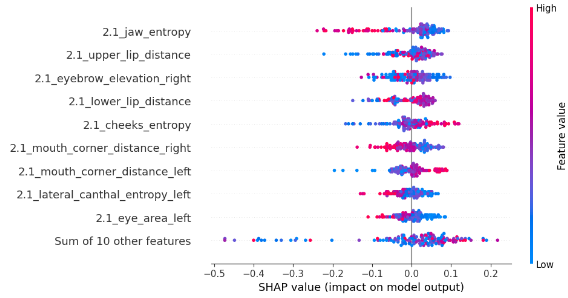
(a) SHAP values of model 1.1



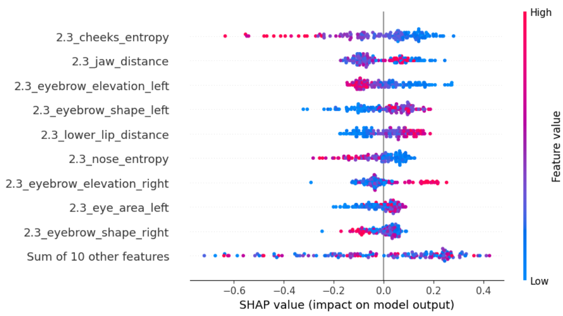
(b) SHAP values of model 1.4



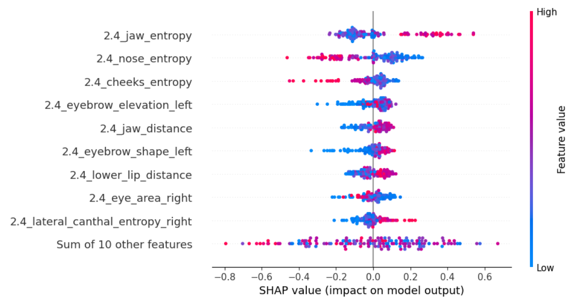
(c) SHAP values of model 1.5



(d) SHAP values of model 2.1

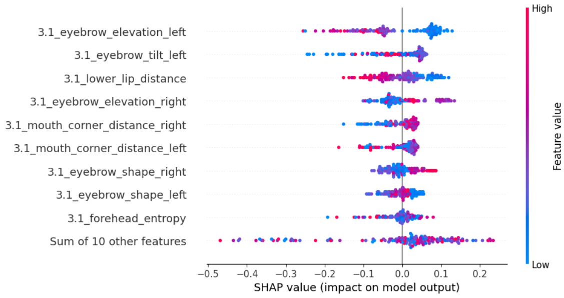


(e) SHAP values of model 2.3

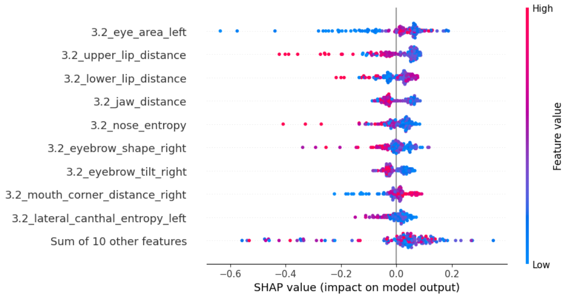


(f) SHAP values of model 2.4

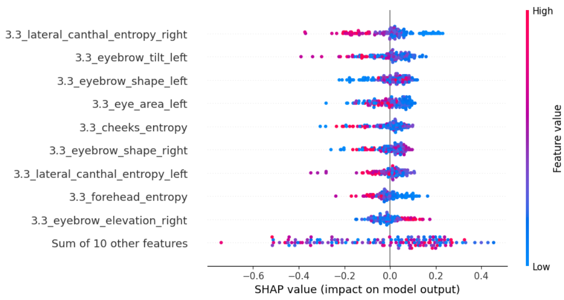
Fig. A.5: Beeswarm plots of SHAP values of features in models without oversampling – impact on model output, for models 1.1, 1.4, 1.5, 2.1, 2.3, 2.4



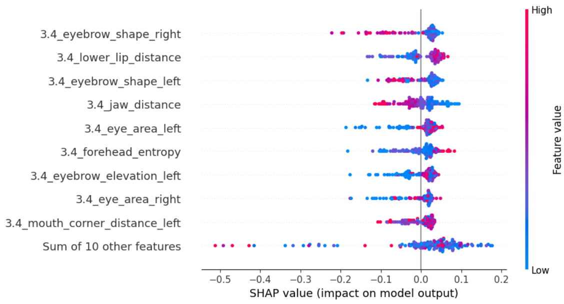
(a) SHAP values of model 3.1



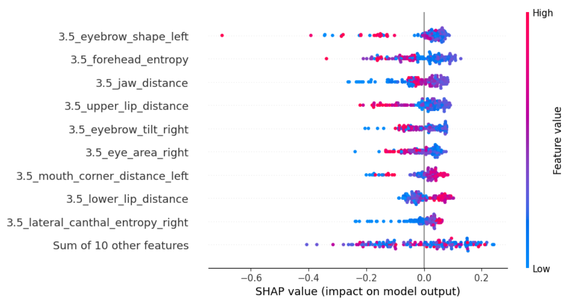
(b) SHAP values of model 3.2



(c) SHAP values of model 3.3

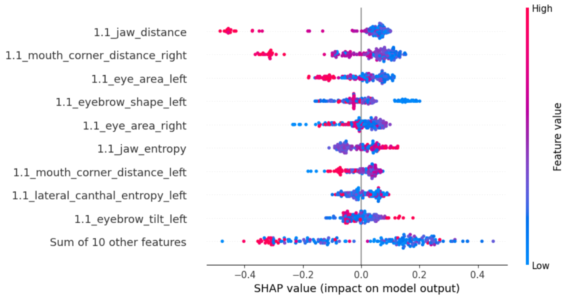


(d) SHAP values of model 3.4



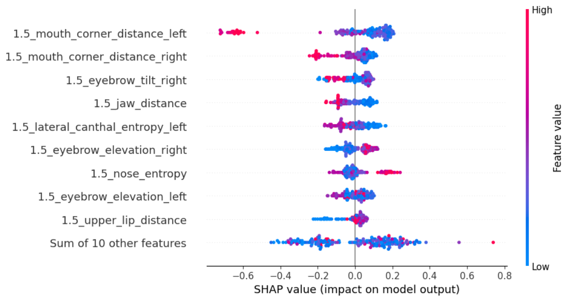
(e) SHAP values of model 3.5

Fig. A.6: Beeswarm plots of SHAP values of features in models without oversampling – impact on model output, for models 3.1, 3.2, 3.3, 3.4, 3.5

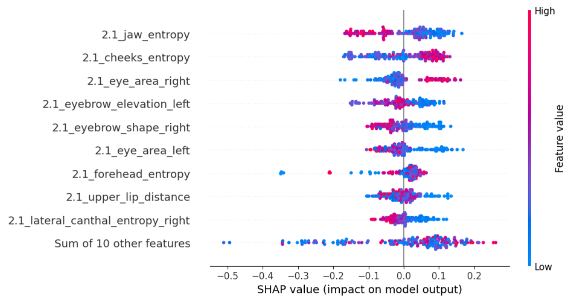


(a) SHAP values of model 1.1

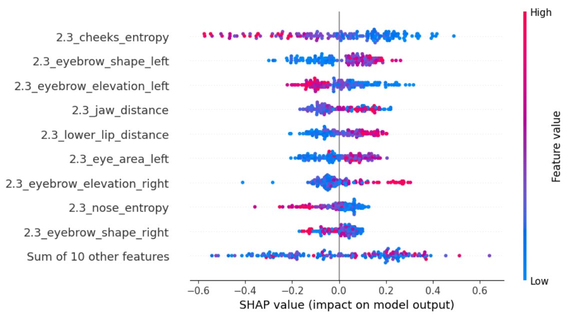
(b) SHAP values of model 1.4



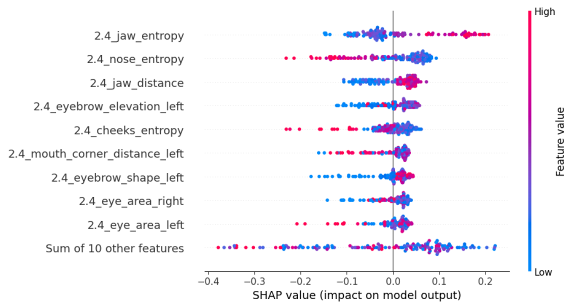
(c) SHAP values of model 1.5



(d) SHAP values of model 2.1

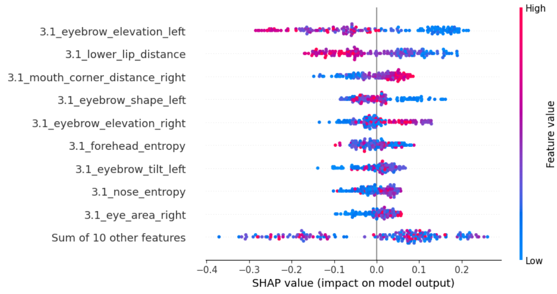


(e) SHAP values of model 2.3

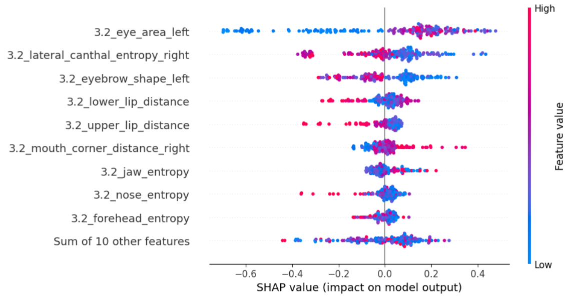


(f) SHAP values of model 2.4

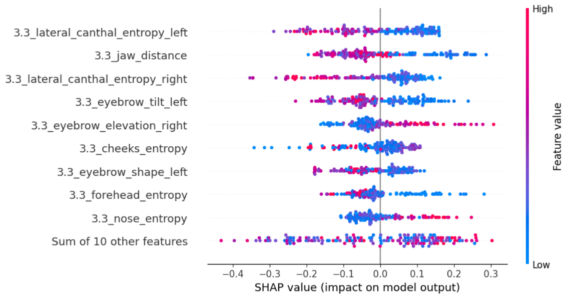
Fig. A.7: Beeswarm plots of SHAP values of features in models with oversampling – impact on model output, for models 1.1, 1.4, 1.5, 2.1, 2.3, 2.4



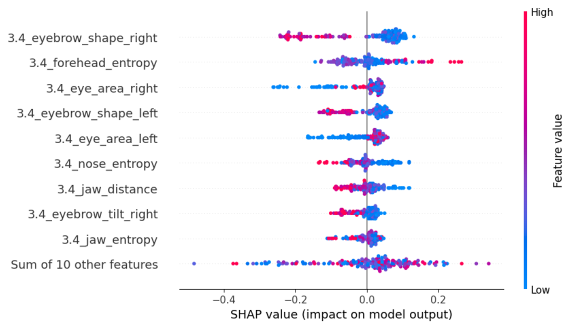
(a) SHAP values of model 3.1



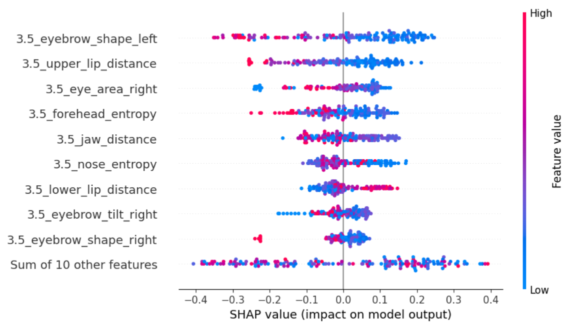
(b) SHAP values of model 3.2



(c) SHAP values of model 3.3

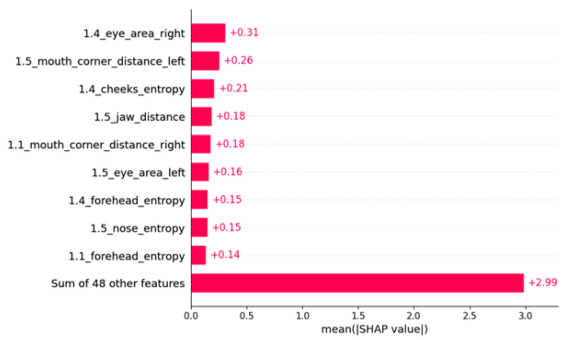


(d) SHAP values of model 3.4

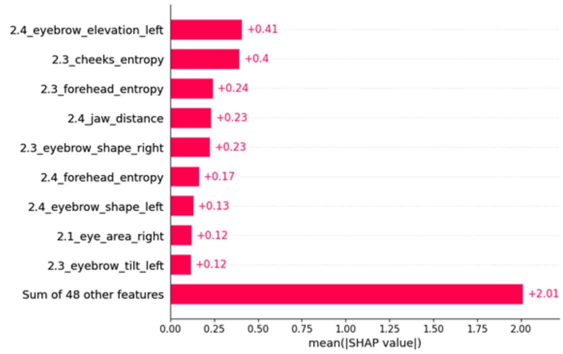


(e) SHAP values of model 3.5

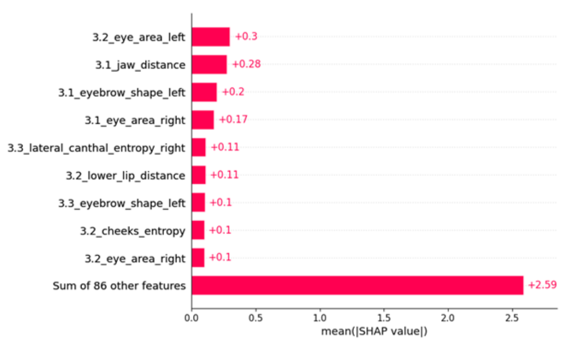
Fig. A.8: Beeswarm plots of SHAP values of features in models with oversampling – impact on model output, for models 3.1, 3.2, 3.3, 3.4, 3.5



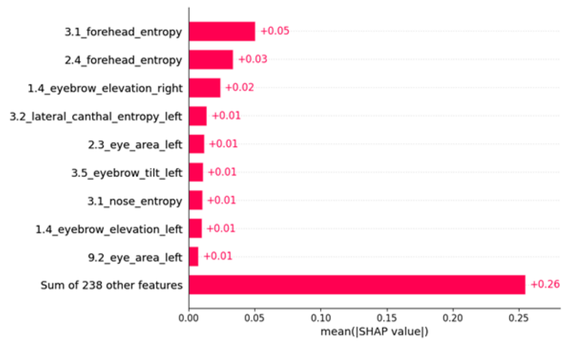
(a) Group 1: Lips



(b) Group 2: Chin

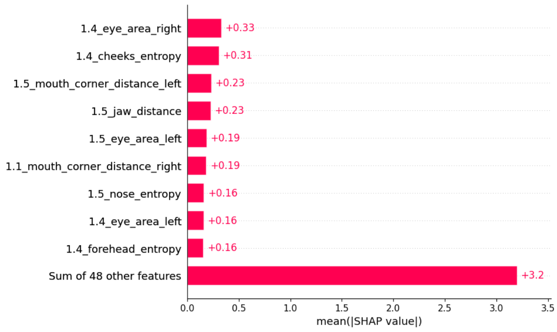


(c) Group 3: Tongue

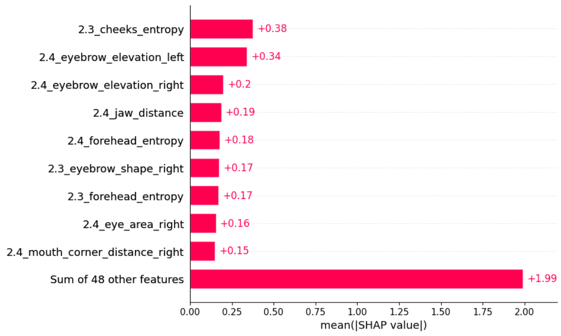


(d) F1

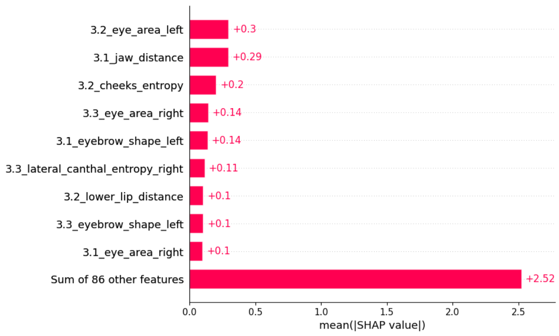
Fig. A.9: SHAP values for models without oversampling



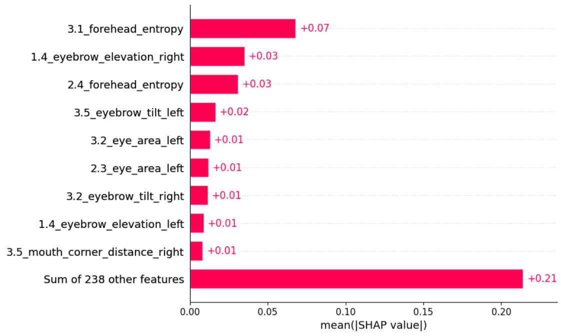
(a) Group 1: Lips



(b) Group 2: Chin

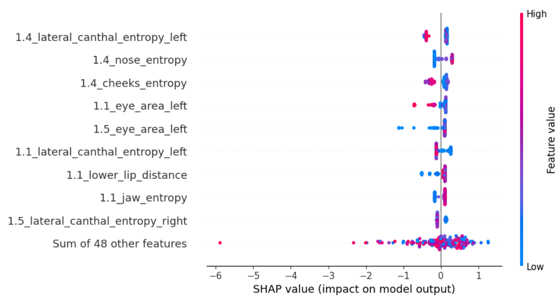


(c) Group 3: Tongue

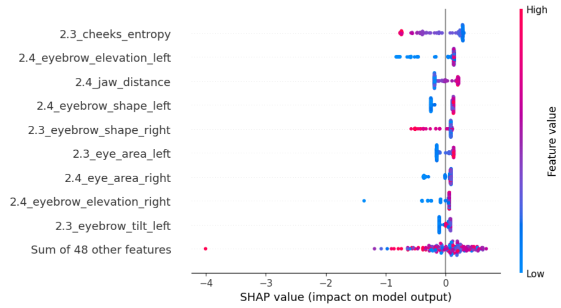


(d) F1

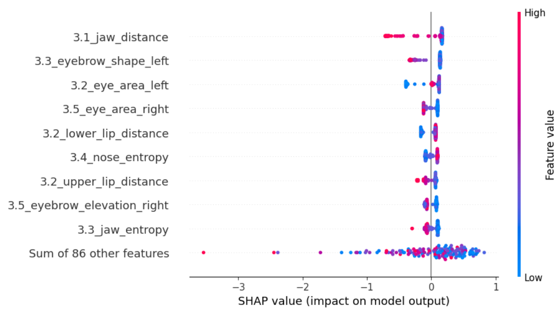
Fig. A.10: SHAP values for models with oversampling



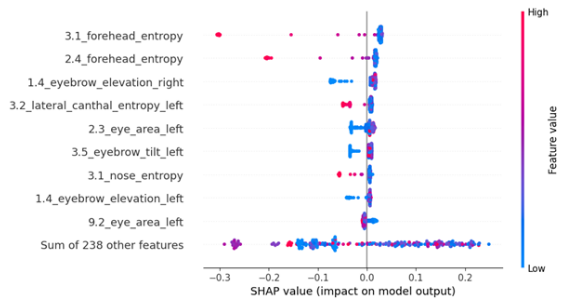
(a) Group 1: Lips



(b) Group 2: Chin

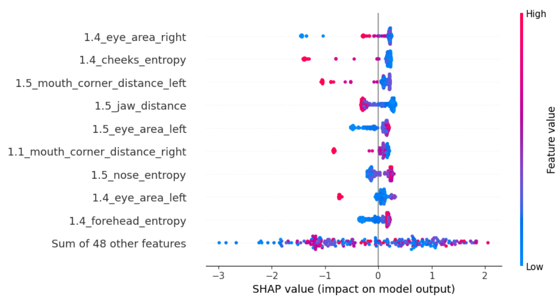


(c) Group 3: Tongue

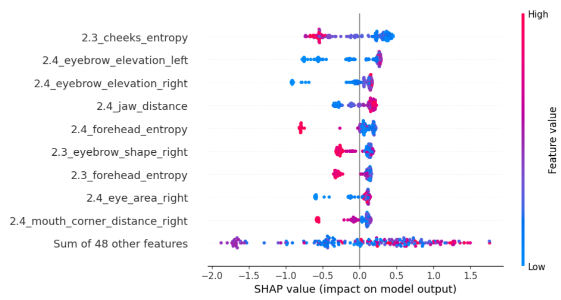


(d) F1

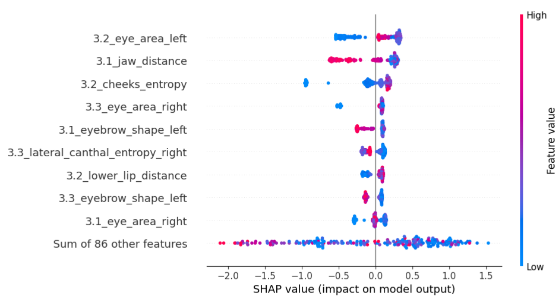
Fig. A.11: Beeswarm plots of SHAP values for models without oversampling



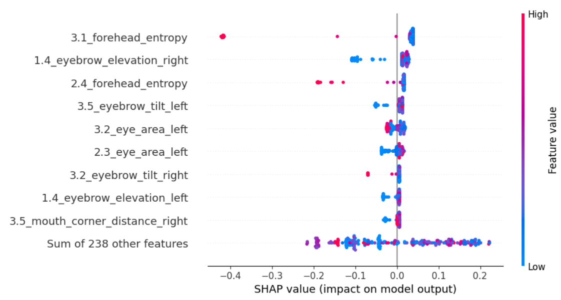
(a) Group 1: Lips



(b) Group 2: Chin



(c) Group 3: Tongue



(d) F1

Fig. A.12: Beeswarm plots of SHAP values for models with oversampling

B Content of the electronic attachment - Used code

The attached appendix contains the code written for this thesis from the beginning of preprocessing, e.g. the facial landmark detection from prepared videos of task executions. It also contains dataset examination and outlier handling, the correlation analysis, and pipelines for training the XGBoost models and CNN models. All of the code was written in Python.

Parts of this project were developed with the assistance of OpenAI's ChatGPT-4o model in accordance with guidelines for generative AI tools issued by the Brno University of Technology. The model was used to generate, debug, and refactor Python code. All generated code was reviewed, adapted, and integrated to meet the specific requirements of this thesis.

```
/ ........................................ root of the attached folder
├── readme.txt ............ important informations for running the codes
├── requirements.txt ........................ used versions of libraries
├── cnn.ipynb
├── corr_analysis.ipynb
├── data_examination+outlier_dropping.ipynb
├── mediapipe_detection+feature_extraction.ipynb
└── xgboost.ipynb
```