

Faculdade de Engenharia da Universidade do Porto



User Sentiment-analysis during Physical Rehabilitation

Diogo Ricardo Castro Gonçalves

MASTER'S THESIS

MESTRADO INTEGRADO EM BIOENGENHARIA - ENGENHARIA BIOMÉDICA

Supervisor: Hélder Oliveira, PhD, FCUP | INESC TEC

Co-supervisor: Luís Ungaro, MSc, SWORD HEALTH

Co-supervisor: Marta Ferreira, MSc, SWORD HEALTH

June 2020

User Sentiment-analysis during Physical Rehabilitation

Diogo Ricardo Castro Gonçalves

**MESTRADO INTEGRADO EM BIOENGENHARIA - ENGENHARIA
BIOMÉDICA**

June 2020

Resumo

As lesões músculo-esqueléticas são um conjunto de condições com prevalência e incidência significativa em todo o mundo. Estas têm impacto negativo tanto no paciente (sendo, a principal causa de incapacidade para trabalhar) como no sistema de saúde. A reabilitação física é frequentemente empregue para ultrapassar este problema crónico, sendo considerada a melhor solução disponível. Com o objetivo de minimizar custos e de otimizar recursos humanos, a reabilitação digital surge como solução, proporcionando resultados clínicos favoráveis em comparação com os métodos tradicionais.

O seguinte trabalho, desenvolvido na SWORD Health, uma empresa de dispositivos médicos que permite precisamente a pacientes realizarem fisioterapia em casa, descreve o desenvolvimento de um sistema de classificação do estado emocional dos pacientes durante esse processo de reabilitação. A extrema importância do estado emocional do paciente na sua interação com o fisioterapeuta tendo em vista um tratamento bem sucedido justifica o âmbito da dissertação.

A revisão literária realizada sobre técnicas de deteção de estados afetivos demonstrou que diferentes abordagens são atualmente utilizadas, sustentadas por distintas modalidades tal como reconhecimento de expressões faciais, áudio ou texto. No entanto, a maioria destas depende de configurações e sinais específicos não disponíveis durante o programa de reabilitação digital já implementado. Dadas as limitações da tecnologia e os requisitos do produto, uma abordagem baseada em dados fora escolhida suportada pela grande quantidade de dados inexplorados existentes e o imenso potencial de diferentes características comportamentais, introduzindo uma perspetiva diferente na deteção do estado afetivo.

Tendo isso em consideração e uma vez que não há uma medida objetiva do estado emocional, a seguinte dissertação foca-se particularmente na exploração de diferentes variáveis inseridas manualmente no sistema pelo paciente – dor, fadiga e satisfação, denominadas de *variáveis associadas a sentimento* – visando prevê-las separadamente ou combinando-as de forma a criar uma representação do estado emocional validada clinicamente. Independentemente da variável a prever, a metodologia é centrada nos dados recolhidos pelo sistema, a partir dos quais são construídos algoritmos de Aprendizagem Computacional supervisionados. A análise exploratória da base de dados forneceu as bases para as experiências de Aprendizagem Computacional, permitindo a extração de características baseadas no domínio. Embora vários algoritmos de base tenham sido estudados, a aplicação de técnicas de aprendizagem em dados não balanceados superou os resultados dos restantes modelos testados, incluindo a abordagem sequencial projetada que fez uso de uma rede neuronal *Long Short-Term Memory*. Estas técnicas de aprendizagem em dados não balanceados, combinadas com *Support Vector Machines*, atingiram o melhor resultado nos dados de teste com valores de F1 de 0.646 e 0.447 em classificação binária de dor e fadiga, respetivamente. Por outro lado, o melhor desempenho em termos da classificação multi-classe das *variáveis associadas a sentimento* é partilhado com as técnicas implementadas de classificação ordinal.

Embora os resultados demonstrem a exequibilidade de utilizar dados para deteção de emoções, é necessário trabalho futuro de forma a ser possível integrar tal funcionalidade no produto da SWORD Health. De qualquer forma, o objetivo principal de fornecer uma base para um futuro sistema de feedback adaptativo em tempo real durante o processo de reabilitação fora atingido.

Abstract

Musculoskeletal disorders are a set of conditions with significant prevalence and incidence worldwide. They have an impact on both the patient (the leading cause of work disability) and the healthcare system. Physical rehabilitation is frequently employed to overcome this lifetime problem, appearing as the best available solution. In the wake of minimising costs and optimise human resources, digital rehabilitation is arising as a solution, providing superior clinical outcomes.

This work, developed at SWORD Health, a medical devices company allowing exactly patients to carry out their therapy at home, describes the development of a pipeline for the classification of patients affective-state during that rehabilitation procedure. This emotional aspect is of utmost importance in patient-therapist interaction towards a successful treatment, justifying the scope of the dissertation.

A literature review on Affect Detection techniques yields several approaches that are being employed supported by different modalities as facial emotion recognition, audio or text. However, they all depend on specific settings and signals not available during the therapy protocol already implemented. Given the technology limitations and product requirements, a data-based approach is chosen sustained by the large amount of unexplored existing data and the immense potential of the behavioural information, introducing a different perspective concerning affective-state detection.

Bearing that in mind and since there is no objective measure of emotional-state, this thesis focuses particularly on the exploration of different variables inserted manually in the system by the patient - pain, fatigue and satisfaction, named as sentiment-related variables - aiming to predict them separately or combining them to create a clinically-validated emotional-state representation. Regardless of the variable to predict, the methodology is centred around system collected data, on which supervised machine learning (ML) algorithms are built. Exploratory analyses of the dataset provided the foundations for the ML experiments, extracting domain-based features. Although various baseline algorithms have been studied, the application of imbalanced learning techniques outperformed the rest of the tested models, including a designed sequential approach, making use of Long Short-Term Memory Neural Networks. These imbalance learning techniques, combined with Support Vector Machines, achieved the best testing F1-scores of 0.646 and 0.447 in pain and fatigue binary classification, respectively. At the same time, the top performance in terms of sentiment-related variables multi-class classification is shared with the implemented ordinal classification techniques.

Although results proved the feasibility of employing data to affect detection, future work is required in order to reach a production-ready state. In any case, the final goal of providing a basis for a future real-time adaptive feedback system during rehabilitation was accomplished.

Acknowledgments

Dizem que os anos mais ricos da minha vida estão a chegar ao fim. Resta-me a agradecer a todas as pessoas, que por diferentes razões, me fizeram ser quem sou e estar aqui nesta noite a escrever a última página do meu percurso académico. Assim, sem mais demoras e sem qualquer ordem especial:

Ao meu orientador Prof. Hélder Oliveira, por ter aceitado orientar a tese sem saber bem do que se tratava. Um obrigado pelo apoio científico e pelas sempre interessantes conversas. Ao Luís e à Marta, e a toda a SWORD Health. Obrigado pela confiança e apoio durante estes últimos meses.

Aos meus pais, porque sem vocês nada teria sido possível. Obrigado por sempre terem acreditado em mim, por tudo o que me proporcionaram e pela dedicação total. Isto tudo, sem nunca me tirarem a possibilidade de voar e seguir o meu caminho. Ao meu irmão, por saber sempre tudo e pela alegria/mau-humor contagiante, mano há só um. A toda a minha família, como não poderia deixar de ser.

Ao 3ºD, palco de tantas histórias e de tantas memórias. A uma casa que nunca faltou vida durante 5 anos, seria impossível pedir melhor. Ao Zé e ao Miguel por terem partilhado comigo este tempo, por saber que ia chegar a casa e estar sempre em família.

Ao Zé, novamente, porque não me lembro de existir sem tu existires. Já lá vão mais de 20 anos desde que, segundo consta, nos conhecemos. Sempre juntos, chegamos agora a esta fase com uma única certeza: muitos mais anos estão para vir.

À FEUP, por ter proporcionado estes memoráveis 5 anos, nos quais nada mudaria. Uma faculdade que rapidamente se tornou casa e que tanto me fez crescer academicamente e como pessoa. É um orgulho sair desta casa Engenheiro.

Ao NEB, a uma causa na qual tenho um orgulho imenso e que vai deixar muitas saudades, ainda que fique a certeza que estará bem entregue. A todas as pessoas fantásticas, em particular, à Ana e à Joana por terem tornado este último ano tão especial. Prometemos nunca falar mal do NEB e, assim, cumprimos.

A todos os meus amigos que viveram tantas histórias comigo e que tornaram estes 5 anos, os anos mais incríveis da minha vida. Ao Álvaro, à Ana e à Joana, companheiros da aventura que foi Glasgow, companheiros desde o primeiro dia e amigos para sempre. Um obrigado por tantos momentos que não caberiam nestes agradecimentos. À Natália, por ser a minha *slav* favorita. Ao Malafaya, por ser um dos primeiros e o único que conseguiu acabar o curso com menos bigode do que eu. Ao incomparável Meneses, pela amizade durante estes anos, por todos os áudios e por todas as vezes que alegraste o dia de todos nós. Ao Bruno, à Leonor, ao Migas, ao Pedrão, ao Pedrinho, à Raquel, amigos que Biomédica juntou mas sem os quais muitas páginas iriam ficar por escrever. A minha casa estará sempre aberta para vos receber. Agora, que venha Peniche.

À Catarina, porque não deve ser fácil aturar-me e este percurso não ficaria completo sem ti. Um obrigado por seres quem és, desde a pessoa mais chatinha do mundo à melhor namorada que alguém pode sonhar ter. Que estes sejam apenas os primeiros 3 anos da nossa história.

E com uma lágrima no olho, fica a certeza que não poderia ter sido mais feliz.

Diogo Gonçalves

*“I realized that intimidation didn’t really exist
if you’re in the right frame of mind.”*

Kobe Bryant

Contents

List of Figures	xi
List of Tables	xv
List of Abbreviations	xvii
1 Introduction	1
1.1 Motivation	3
1.2 Goals	5
1.3 Contributions	6
1.4 Document Structure	7
2 Background	9
2.1 Musculoskeletal Disorders	9
2.1.1 Risk Factors of Musculoskeletal Disorders	11
2.2 SWORD Health System	11
2.3 Sentiment-Analysis-Related Concepts in Rehabilitation	13
2.3.1 Motivation Concept	13
2.3.2 Pain	14
2.3.3 Fatigue	15
2.3.4 Satisfaction	15
2.4 Machine Learning	16
2.4.1 The Data Mining Process	17
2.5 Summary	18
3 Literature Review	19
3.1 Affective computing	19
3.1.1 Affective States and emotions categorisation	20
3.2 Affect Detection	22
3.2.1 Facial Emotion Recognition	23
3.2.2 Audio Expression Recognition	30
3.2.3 Body Gestures and Posture	33
3.2.4 Emotion Recognition from Text	34
3.2.5 Physiological Data and EEG	36
3.2.6 Multimodal Affect Recognition	38
3.3 Summary	40

4	Dataset and Statistical Analysis	41
4.1	Dataset Description	41
4.2	Target Distribution	44
4.3	Feature Engineering	46
4.4	Feature Analysis	48
4.5	Summary	53
5	Sentiment-Related Variables Classification	55
5.1	Experiments Design	56
5.2	Experiments Workflow	57
5.3	Classification with Tree-based Emsembles	59
5.3.1	Random Forest	60
5.3.2	Gradient Boosted Trees	63
5.4	Classification with Support Vector Machines	66
5.4.1	Classification with SVM and Feature Selection	68
5.4.2	Classification with SVM and PCA	70
5.5	Classification with Neural Network	73
5.5.1	Results & Discussion	76
5.6	Imbalanced Learning Techniques	77
5.6.1	Cost-Sensitive Method	78
5.6.2	Undersampling and Oversampling Methods	81
5.7	Ordinal Classification Techniques	86
5.7.1	Ordinal Classification: a Simple Approach	87
5.7.2	Ordinal Classification: a NN approach	90
5.8	Sequential approach (LSTMs)	93
5.8.1	Results & Discussion	96
5.9	Comparative analysis & Conclusions	98
5.10	Summary	99
6	Emotional-State Inference	101
6.1	Clustering	101
6.1.1	K-Means	102
6.1.2	Non-parametric clustering	110
6.1.3	Comparative Analysis of Clustering Methodologies	114
6.2	Emotional-state Classification Task	115
6.2.1	Results & Discussion	116
6.3	Summary	118
7	Conclusions and Future Work	119
	Bibliography	123
A	Complementary results of models for sentiment-related variables classification	141
B	Complementary results of models for emotional-state classification	171

List of Figures

2.1	Changes within cancellous bone in Osteoporosis.	10
2.2	SWORD Phoenix Main Components.	12
3.1	Wheel of Emotions.	21
3.2	Hourglass of Emotions 3D model.	22
3.3	Procedure used in conventional FER approaches	24
3.4	List of AUs related to the six basic emotions.	24
3.5	Example of Geometric Features.	25
3.6	Feature Extraction process from face region.	26
3.7	Dynamic and hybrid method using visual features.	27
3.8	Overall structure of FER approach using CNN.	28
3.9	Overview of an hybrid DL framework.	29
3.10	C3D network used in feature extraction from videos.	29
3.11	Emotion recognition from audio modality using CNN.	32
3.12	CNNs and LSTMs for audio Emotion Recognition.	32
3.13	Posture analysis system used in Emotion Recognition.	34
3.14	RNN and CNN combined in Emotion Classification using Text.	36
3.15	Examples of physiological signals used in Affect Detection.	37
3.16	AMIGOS dataset distribution of labels.	38
3.17	Sentiment analysis from linguistic information.	39
4.1	Entity-relationship of the dataset used over the dissertation	42
4.2	Dataset Distribution according to patient gender and type of therapy.	43
4.3	Age of the patients in the dataset, stratified by gender and therapy type	44
4.4	Target Distribution in the dataset.	45
4.5	Linear Interpolation applied in the labelling procedure of satisfaction values.	46
4.6	Feature Distribution Analysis in terms of Fatigue and Pain values.	49
4.7	Example of the Transformation of continuous variables into binary data.	50
4.8	Correlation Matrix of feature set.	52
4.9	Distribution scatter of a pair of attributes to evaluate feature correlation.	52
5.1	Example of an ensemble of two simple decision trees.	59
5.2	Performance of RF classifier on sentiment-related variables prediction.	62
5.3	Feature importance evaluation of a RF classifier.	63
5.4	Performance of GBT classifier on sentiment-related variables prediction.	65

5.5	Example of classification of data by Support Vector Machines	67
5.6	Performance of SVM-RFE classifier on sentiment-related variables prediction. . .	69
5.7	Example of data transformation through Principal Component Analysis.	70
5.8	Performance of SVM-PCA classifier on sentiment-related variables prediction. .	72
5.9	A simple MLP neural network for classification.	74
5.10	Performance of MLP classifier on sentiment-related variables prediction.	76
5.11	Cost-sensitive classification results.	80
5.12	Binary classification of Pain using SVM-RFE cost-sensitive	80
5.13	Graphical representation of the Clustering-based undersampling algorithm. . . .	82
5.14	Graphical representation of the SMOTE algorithm.	83
5.15	Data-balancing classification results at testing set.	85
5.16	Binary classification of Pain using SVM-RFE with data balancing: hybrid approach.	86
5.17	Transformation of an ordinal dataset with four values into three binary attributes.	88
5.18	Performance of Ordinal classifier: Simple approach.	89
5.19	Multi-class classification of Pain using ordinal classifier: Simple approach. . . .	90
5.20	Comparison between a standard classification neural network and an ordinal regression neural network.	92
5.21	Performance of Ordinal classifier: a NN approach.	92
5.22	The processing blocks inside an LSTM cell.	94
5.23	Performance of LSTM classification models on sentiment-related variables prediction.	97
6.1	Illustration of K-means algorithm.	103
6.2	The Elbow method applied at the sentiment-related variables.	105
6.3	Emotional-state representation by K-Means.	106
6.4	Emotional-state representation distribution.	108
6.5	Analysis of the initialisation effect in K-means.	109
6.6	Influence of the regularisation parameter on DP-Means results.	111
6.7	Emotional-state representation by DP-Means - λ defined as 15.	112
6.8	Emotional-state representation by DP-Means - number of clusters fixed at 3. . .	113
6.9	Emotional-state classification results.	117
A.2.1	Confusion matrices for sentiment-related variables classification using RF classifier.	144
A.2.2	Confusion matrices for sentiment-related variables classification using GBT classifier.	145
A.2.3	Confusion matrices for sentiment-related variables classification using SVM-RFE classifier.	146
A.2.4	Confusion matrices for sentiment-related variables classification using SVM-PCA classifier.	147
A.2.5	Confusion matrices for sentiment-related variables classification using MLP. . .	148
A.3.1	Sentiment-related variables classification results using cost-sensitive RF.	149
A.3.2	Sentiment-related variables classification results using cost-sensitive GBT.	150
A.3.3	Sentiment-related variables classification results using cost-sensitive SVM-RFE. .	151
A.3.4	Sentiment-related variables classification results using cost-sensitive SVM-PCA. .	152
A.3.5	Sentiment-related variables classification results using cost-sensitive MLP.	153
A.4.1	Data-balancing classification results at validation set: Hybrid Approach.	154

A.4.2	Sentiment-related variables classification results using RF with data balanced at mean of all classes number of instances: Hybrid Approach.	155
A.4.3	Sentiment-related variables classification results using GBT with data balanced at mean of all classes number of instances: Hybrid Approach.	156
A.4.4	Sentiment-related variables classification results using SVM-RFE with data balanced at mean of all classes number of instances: Hybrid Approach.	157
A.4.5	Sentiment-related variables classification results using SVM-PCA with data balanced at mean of all classes number of instances: Hybrid Approach.	158
A.4.6	Sentiment-related variables classification results using MLP with data balanced at mean of all classes number of instances: Hybrid Approach.	159
A.5.1	Data-balancing classification results at validation set: Undersampling Approach.	160
A.5.2	Sentiment-related variables classification results using RF with data balanced at minority class: Undersampling approach.	161
A.5.3	Sentiment-related variables classification results using GBT with data balanced at minority class: Undersampling approach.	162
A.5.4	Sentiment-related variables classification results using SVM-RFE with data balanced at minority class: Undersampling approach.	163
A.5.5	Sentiment-related variables classification results using SVM-PCA with data balanced at minority class: Undersampling approach.	164
A.5.6	Sentiment-related variables classification results using MLP with data balanced at minority class: Undersampling approach.	165
A.6.1	Ordinal classification results: Simple approach with RF binary classifiers.	166
A.6.2	Ordinal classification results: Simple approach with GBT binary classifiers.	166
A.6.3	Confusion matrices of Ordinal Classifier: An NN approach	167
A.7.1	LSTM complementary results: 10 sessions with overlap.	168
A.7.2	LSTM complementary results: 5 sessions without overlap.	169
B.1.1	Emotional representation by K-means with 3 clusters.	172
B.1.2	Emotional representation by K-means with 4 clusters.	172
B.2.1	Emotional representation by DP-means: Regularisation parameter fixed at 15.	173
B.2.2	Emotional representation by DP-means: Number of clusters set at 3.	173
B.2.3	Distribution of data in clusters generated by DP-means.	174
B.2.4	Analysis of input sequences order effect in DP-means.	174
B.3.1	Emotional state classification results using baseline classifiers.	175
B.3.2	Emotional state classification results using cost-sensitive classifiers.	176
B.3.3	Emotional state classification results with data balanced: Hybrid approach.	177
B.3.4	Emotional state classification results with data balanced: Undersampling approach.	178

List of Tables

5.1	Optimised hyperparameters in the RF sentiment-related variables classifier, with the considered values and definition.	61
5.2	Optimised hyperparameters in the GBT sentiment-related variables classifier, with the considered values and definition.	65
5.3	Optimised hyperparameters in the SVM-RFE sentiment-related variables classifier, with the considered values and definition.	69
5.4	Optimised hyperparameters in the SVM-PCA sentiment-related variables classifier, with the considered values and definition.	72
5.5	Optimised hyperparameters in the MLP neural network for sentiment-related variables classification, with the considered values and definition.	75
5.6	Example of a Cost-Matrix in a binary classification problem.	78
5.7	Optimised hyperparameters during LSTM training for sentiment-related variables classification, with the considered values and definition.	96

List of Abbreviations

AAM	Active Appearance Model
AC	Affective Computing
AI	Artificial Intelligence
AU	Action Unit
BOW	Bag Of Words
c3D	3D Convolutional Networks
CNN	Convolutional Neural Networks
DL	Deep Learning
DM	Dirichlet Process
EDA	Electrodermal Activity
EEG	Electroencephalography
EEG	Expectation Maximisation
EMG	Electromyogram
ERD	Entity-Relationship Diagram
FACS	Facial Action Coding System
FER	Facial Emotion Recognition
FS	Feature Selection
GBT	Gradient-Boosted Trees
LBP	Local Binary Pattern
LL	Lower Limbs
LSTM	Long Short-Term Memory
MFCC	Mel Frequency Cepstral Coefficients
ML	Machine Learning
MLP	Multilayer Perceptron
MSD	Musculoskeletal disorder
NLP	Natural Language Processing
NN	Neural Network
NPS	Net Promoter Score
NSAID	Non-steroidal Anti-Anflammatory Drug
PCA	Principal Component Analysis
RA	Rheumatoid Arthritis
RBF	Radial Basis Function
RF	Random Forest
RNN	Recurrent Neural Networks
SD	Standard Deviation
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
SVM-RFE	Support Vector Machine Recursive Feature Elimination
THA	Total Hip Arthroplasty
TKA	Total Knee Arthroplasty
UL	Upper Limbs
US	United States
YLDs	Years Lived With Disability

Chapter 1

Introduction

Musculoskeletal disorders (MSDs) have been defined as injuries or disorders which impact the human body's movement or musculoskeletal system - from muscles to bones and joints. Regarding pathophysiology, these conditions are a diverse group, but they are linked anatomically and always associated with pain and impaired physical function [1]. They include a set of conditions, from chronic disorders as osteoarthritis, rheumatoid arthritis, osteoporosis, and low back pain to short duration and acute symptoms. Although it may affect all parts of the body, the back, neck, shoulders, hip, and knees are the most commonly affected [2].

MSDs are a widespread and increasing occupational health problem, being the most prevalent occupational disease in the European Union [3]. Also across Europe, MSDs are considered the leading cause of work disability, sickness absence from work, presenteeism and loss of productivity. Overall, according to the ad hoc analysis of the European Labour Force Survey examined in [3], MSDs are responsible for 50% off all lost workdays for more than three days and 49% if discussing absences lasting a couple of weeks or more. In the United States (US), it is estimated that work-related musculoskeletal disorders account for 600,000 injuries, summing 34% of all absences from work [4]. This problem does not affect only the workers themselves, it leads to high costs for enterprises and society as whole [3].

According to the Global Burden of Disease Study analysed in [5], this issue is not exclusive to Europe or US, it is one of the significant causes of disability worldwide and affects hundreds of millions of people. Globally, the musculoskeletal disorders account for 21.3% of the total years lived with disability (YLDs), only below mental and behavioural problems. If studying both death and disability, it combined caused 6.7% of the total global disability-adjusted life years, ranked fourth in the greatest burdens on the health of the world's population [5], [6].

MSDs have financial repercussion for both the patient and the healthcare system. In the latest Large Employers' Health Care Strategy and Plan Design Survey [7], musculoskeletal conditions came in at the top spot as the specific disease that drives up patients healthcare costs. According to analysis already mentioned performed by the European Labour Force Survey [3], MSDs are accountable for 40 to 50% of the costs of all work-related health problems.

As consequence of the different factors as the ageing of the population or increasing obesity rates, it is noticed an exponential rise in the burden of musculoskeletal disorders in the wake of the increasing number of people suffering from MSDs throughout the world. According to

the Global Burden of Disease Study from 2010, the prevalence estimated of MSDs was already extremely high - accounting together more than 25% of worldwide population [5] - and it is expected to increase. This way, the treatment and care of them will be discussed in the current section.

Although MSDs comprise different regions of the body, people with musculoskeletal disorders share similar characteristics, prognostic and may respond to identical treatments. Pain acts as the symptom responsible for the diminishing of quality of life (affecting social functionality and mental health) and increased demands on healthcare [1], therefore it is managed by plenty of treatment options. Care is, mostly, delivered in primary care by first contact clinicians such as general practitioners, physiotherapists, chiropractors and osteopaths [8]. Essentially, musculoskeletal pain is being addressed in three different ways: non-pharmacological treatments as exercise therapy, manual therapy or psychosocial interventions; pharmacological interventions as analgesics, non-steroidal anti-inflammatory drugs (NSAIDs) or opioids; and, finally, surgical interventions.

As for exercise therapy, it was found to be beneficial for pain, function and quality of life. A systematic review of effective treatment options that identified different guidelines and clinical pathways showed evidence for the five major musculoskeletal disorders [8]. It noted that optimal content and way of delivering exercise may vary according to the group of musculoskeletal pain, however, Babatunde *et al.* conclude that functional exercises (that allow patients' exercises to be adapted on their daily activities and enable patients to perform them easier and injury-free) may be more effective than exercises with no specific targeted function. Benefits of physical rehabilitation include improved not only musculoskeletal function and mental health but also, improved cardiovascular and pulmonary health [9].

Looking now at pharmacological interventions, NSAIDs and opioids had significant consequences at pain relief but evidence show effect only in the short-term [10]–[14]. Furthermore, the potential for non-favourable effects such as gastrointestinal bleeding and opioids-induced hyperalgesia need careful attention. Opioids have high addiction rates that way their prescription should be avoided unless the instantaneous reduction of pain is mandatory given patients' conditions¹. Thus, doctors should be particularly cautious when prescribing them. The last guidelines issued by the Centre for Disease Control and Prevention [15] also provided considerations on whether chronic pain patients need opioid treatment or not according to their condition. Pharmacological injections consist of a distinct delivery mode that is offered for acute pain relief or to enable patients to tolerate physical therapy. However, there was no evidence that this treatment is more effective in long-term [8].

Surgical procedures work as other treatment option, appointed as a solution when in the presence of severe pathology and substantial pain; either way, early surgery may always be considered on a case by case ground. The work of Peul *et al.* [16] shows that early surgical intervention provides favourable results - benefits for pain and perceived recovery - compared to previous conservative options (including analgesia and physical therapy) in the short term. On the other side, physical therapy and non-pharmacological treatments strongly continue to appear as the best available solution to outcome long time problems since, as demonstrated by

¹SWORD Health Blog post - "Round-up on the Opioid Epidemic: Musculoskeletal Disorders and Opioids": swordhealth.com/blog_post/msds-and-opioids/. Accessed on 17-06-2020

a series of studies, there are no significant differences in results in pain and function between conservatively and surgically treated patients [17]–[20]. A surgery turns out to be an expensive treatment solution that would only be justifiable if these procedures had a nearly perfect rate of success. Regarding back surgery, the likelihood of failure is particularly high, existing a known syndrome - Failed Back Surgery Syndrome - that impacts 20 to 40% of patients submitted to spine intervention [21]. Because of possible lower risks for complications and similar efficiency, conservative treatment as physical rehabilitation tends to be preferred.

1.1 Motivation

Rehabilitation with an emphasis on physiotherapy and exercise is widely promoted after different surgical procedures as total knee arthroplasty (TKA) or total hip arthroplasty (THA). These surgeries may be considered for someone who has severe musculoskeletal disorders (e.g. arthritis or fractures) at knee or hip. Physical rehabilitation is, that way, the standard of care immediately following this kind of procedures as stated in [22]. It is concluded that the results of these surgeries are improved when resorting to physical therapy, as shown in [23], [24] for TKA and in [24], [25] for hip replacement. Rehabilitation studies state that the programs should start as early as possible after intervention and be intensive in order to achieve effective recovery [26], greater satisfaction and adherence, and reduction of complications and expenses [27], [28]. It is not clear neither consensual the duration and frequency of rehabilitation services that should be provided to patients [29], [30]. Also, little information is available describing the content of the practice, although it seems agreed the need to increase strength and stretching [30], [31].

In rehabilitative healthcare, personalised exercise programs are prescribed by the therapist individually to each patient to address his problems and therapeutic goals. The clinician must address and have into account the patient's specific pathology and disabilities, along with morbidity and additional impairments [32]. Traditional physical therapy relies on clinicians experience and training, lacking an objective and normalised analysis to evaluate patient's performance and evolution, representative variables of treatment results. Moreover, not all patients may have access to rehabilitation clinics, for different reasons as: difficulties to get to the centres (patients that live in isolated areas far from clinical centres or do not have means of transport); having difficulty in being transported (might be dependents on wheelchairs); difficulties finding openings at the clinic and timetable compatibility.

Concerning rehabilitation location, it is proven that both conventional and home-based rehabilitation or home-based have produced identical improvements [33]–[37]. Nowadays trends in healthcare delivery suggest an home-based approach, demonstrated to be cost-effective and a preferable alternative to the traditional one. Despite that, this approach at home may be very challenging due to the scarcity and high-cost of human resources as well as other logistic reasons. In order to overcome these problems, to maximise access and minimise costs, digital rehabilitation is rising as a solution. This methodology capable of providing rehabilitation without permanent supervision was defined as rehabilitation services delivered at home from a remote location through a telecommunication system and information technology [38]. Rodriguez *et al.* [39] establishes tele-rehabilitation as "the application of telecommunication, remote sensing and operation technologies, and computing technologies to assist with the provision

of medical rehabilitation services at a distance". Some potential advantages of this system approach include cost saving due to the reduction of specialised human resources, improvement of the patient comfort and lifestyle [40] and may increase adherence of patients to rehabilitation treatment [37], [41]. However, this alternative remains to be dependent on human monitoring, whether throughout sessions performed at home or during additional supervised sessions, limiting widespread availability [42].

That said, technological solutions that empower patients and enable rehabilitation from a home setting to occur without constant requirement of human supervision are seen as the best solution in order to achieve effectiveness and reduce costs. This should incorporate appropriate instruments for the quantitative measurements to capture motion position or specific details of task execution, aiming to evaluate patients performance. Currently, the inclusion of biofeedback systems is seen as a requirement, since an increase in performance and patient adherence has been demonstrated [43], [44]. This is where SWORD Health comes in, a Portugal based-startup that allows patients to perform their therapy at home, maximising engagement and delivering superior clinical outcomes². Through inertial motion sensors, the system controls patient motion and provides real-time feedback on performance using a mobile app. A web-based platform is also included, allowing therapists to prescribe, monitor and adapt the rehabilitation content from a distance. Thus, SWORD Health product enables patients to carry out rehabilitation sessions by themselves not requiring constant clinician monitoring but assuring remote feedback during the rehabilitation program.

Thereby, SWORD Health developed SWORD Phoenix, an Artificial intelligence (AI) therapist that guides patients through the session in real-time under the remote control of a team of clinicians, that check periodically and set the exercises to be performed by the user. The system offers a scalable solution far less demanding in terms of human resources, and that maximises the reach of existing ones. Patient discomfort and the need for travelling back and forth is also minimised. Fit a therapy appointment in the routine schedule is necessary for a high number of patients but it is not always easy, that way, products as the one developed by SWORD Health may help people to mitigate this inconvenience.

SWORD Health has already published different studies following groups of patients recovering from knee injury. In Dias Correia *et al.*'s study [42], the patients were divided in half, where a part was taking conventional therapy and the other part using SWORD Health's solution. The primary outcome of the study was the change in patient performance after the end of rehabilitation period (8 weeks), measured with the Timed Up and Go Test [45]. Still in the same study, it was concluded that, in the experimental group, the progress was twice as much as the traditional therapy group. The patients who were under a home-based treatment also showed higher engagement due to the comfort of use. The other study, published in February 2019 [46], reported similar positive results in favour of the digital intervention, where higher adherence and satisfaction levels were measured. The novel digital intervention for independent home-based rehabilitation demonstrated, also, the capability to maximise clinical outcomes comparing with typical therapy. The outcomes remained favourable both at three and six months assessments, without the need of extra care. Moreover, a distinct clinical study [47] demonstrated that this solution holds promise in rehabilitation not only after knee injury but after hip replacement. It

²SWORD Health Webpage: [/swordhealth.com](https://swordhealth.com). Accessed on 17-06-2020

managed to ensure better clinical outcomes than conventional rehabilitation.

The factors that explain the better results are summarised in [47] as follows: (1) beneficial effect of biofeedback and gamification on patient involvement and performance; (2) greater patient empowerment, coupled with the effect of monitoring patient effort; and (3) program changes based on objective data due to the incorporation of instruments for quantitative measurement.

In digital health rehabilitation, the digital therapist should be able to replace the physiotherapist during session in real-time, while the latter monitors the patient remotely. Given this, merely classifying the movements as correct or incorrect and providing feedback for correction, is just a part of the conventional physiotherapist-patient relation. Another part of this interaction is the emotional connotation behind the interaction, which directly impacts the patient's motivation and thus, his performance in the rehabilitation sessions [48].

Different studies are suggesting that the identification and expression of emotions are a significant factor for good clinical outcomes as [49], [50]. Research has shown that rehabilitation professionals should be aware of the underlying patients' emotions during treatment and that the interaction between physiotherapists and patients is fundamental to improve clinical outcomes. Thus, the consciousness of patients emotional reactions is an essential prerequisite for a successful treatment.

Then, SWORD Health's physiotherapy professionals justified the interest behind understanding patients emotional-state during the recovery process. Low satisfaction levels or high levels of pain may result in patient's demotivation. A part of patients with negative states of mind gives up the program after some sessions/weeks, without understanding reason. There is still a part of patients that do not give up; however, they show signs of demotivation that result in interactions with the therapist - phone calls and complaints during the therapists' on-site visits - or other actions as answering with exaggerated values of pain and fatigue or not finishing the session or exercise.

Taking into consideration the fact that the system does not include any feature able to understand patient emotional-state during the rehabilitation program, it is possible to find here a specific need. Physical therapists should learn to identify emotions in their different components during the treatment. That said, regarding the clinical importance of emotions, their assessment appears as a necessity that should be addressed, either in or between-sessions. There are no algorithms or instruments developed to infer the emotional-state of patients throughout the process, although the critical role of keeping patients motivated is known. The availability of the prediction of variables whose importance is known as fatigue or pain, given patient behaviour during a rehabilitation session could provide much value for both physiotherapy professionals and patients.

1.2 Goals

The main goal of this work is the development of a pipeline for the classification of patients emotional-state during a rehabilitation program. This emotional-state may be characterised by a set of patient's inputs that includes their satisfaction with the system, fatigue and pain levels. It

is intended to identify relevant changes in the patient behaviour associated with certain system events (e.g. exercise timeouts, exercise skips, wrong movements, performance during session) that may be related to his affective state.

Within this scope, the ultimate goal is to provide a basis for the development of an adaptive feedback algorithm that should be used to keep the patient's motivation high during the physical recovery process. The identification of critical situations will provide information for therapists that they may use to take some actions, as change prescribed exercises, make a call to the patient or even a home visit.

While the regular process is already implemented, and that way, the collection of data occurs for all the patients on-board in the company's plan, on the other side, there is no validated measure of patients emotional-state, whether qualitative or quantitative. Therefore, it should be considered the different variables inserted manually in the system by the patient that may be representative of patients sentiment-related state. This state may be characterised by merely one or a combination of them, and, thus, used as a label of the affective state; therefore, both possibilities should be explored. It is planned to predict these patients inputs and emotion-representation at the end of the session using in-session features. This approach can lead to adaptive feedback throughout the program when certain levels of pain or fatigue are being estimated. The availability of this prediction could be compared to the work of physiotherapists in conventional rehabilitation whose indications to the patient vary according to the state of satisfaction or pain noticed.

The goals contemplate the development of Machine Learning (ML) algorithms with relevant performance assessments to predict the mentioned emotional-state during the rehabilitation process. This includes the development of algorithms capable of predicting the independent input measures (pain/fatigue and satisfaction) based on the data obtained during the session. Moreover, the emotional-state may be seen as a quantitative variable to infer through the set of the mentioned quantitative measures inserted by the patient. This representation is to be found and tested in order to assure clinical relevance and validation. After the approval by the domain-experts, the inference and the development of predicting models of such representation complete thesis' objectives.

Along with the exploration of the output label that should be used, it is also fundamental to understand which are the best features from the database that may be relevant to solve the problem. This data may be interpreted in a sequential format; a fact that may be relevant in the design of models that aim to learn this temporal relationship between sessions. The possibility of introducing features that also reflect this temporal-dependency has to be considered to, if necessary, complement the remaining existing features.

1.3 Contributions

The main contributions of this work are:

- Comprehensive analysis of SWORD Health's dataset, providing an overview of patient population and data distribution.

- Definition of the most relevant features to be used in affective state recognition, by the analysis of behavioural and performance features collected during physical therapy.
- A Machine Learning approach to predict sentiment-analysis related variables as pain or fatigue within a rehabilitation program based on the extracted features.
- A quantitative label representation of patients emotional-state through the exploration of user input variables, considering two different methods.
- Clinical validation of the emotional-state representation.
- A ML framework to predict patients emotional-state validated representation, exploring the data gathered during the rehabilitation program.

1.4 Document Structure

This document is organised as follows: Chapter 2 provides a description of Musculoskeletal Disorders and the risk factors related to it. The SWORD Health main technology is presented as well as the components that the product encompasses. Also, the different concepts associated with Sentiment Analysis in Rehabilitation are explored. A brief introduction to Machine Learning finishes the chapter. In Chapter 3, the basic principles of Affective Computing are defined along with the definition of Affective States and categorisation of emotions. The literature and state-of-the-art techniques for affect detection are then reviewed. The employed dataset is described and assessed throughout Chapter 4, that also comprises feature engineering and analysis work. Based on that data, a set of ML experiments for sentiment-related variables classification was designed, with Chapter 5 reporting on those designs and the corresponding results. In addition, Chapter 6 aims to find an emotional-state representation combining the former variables, exploring two different methods. An ML classification task on the created representation concludes the chapter. Finally, Chapter 7 offers some concluding observations and a brief summary of the knowledge acquired with this research.

Chapter 2

Background

Regarding the context of the dissertation, understanding the main pathologies which comprise the set of conditions named Musculoskeletal Disorders is fundamental. To address this group of afflictions, SWORD Health is recreating physical rehabilitation offering a digital solution. Furthermore, it is essential to link the context of physical therapy with fundamental sentimental-analysis concepts as pain, fatigue and motivation. An introduction to Machine Learning and its typical workflow concerning the methodology to be employed also appeared as requisite.

This chapter explores exactly these topics intending to provide a groundwork where the thesis will be constructed.

2.1 Musculoskeletal Disorders

Musculoskeletal disorders may be defined as a musculoskeletal system and soft-tissue injuries or dysfunctions that affect human natural motion - comprising muscles, bones, nerves, ligaments and joints. Generally, these disorders are caused by repetitive motion, force, vibration, and uncomfortable positions [1], [2]. MSDs do not include disorders caused by accidental events as slips, trips, falls or similar incidents. They considerably affect the psychosocial status of people afflicted with it together with their families and caregivers. Even though anatomically linked, these conditions are a wide group from those of intense beginning and short duration to lifetime disorders. It may affect diverse parts of the body as lower limbs, neck, lower back or shoulder.

The set of conditions include osteoarthritis, rheumatoid arthritis, osteoporosis, low back pain and work-related upper limb disorders. If measuring concerning disability-adjusted life years, osteoarthritis is the 4th most prevalent predicted cause of health problems around the world in women and the 8th in men; rheumatoid arthritis reduces people's work capacity by a third just within the first year; the overall prevalence of osteoporosis is estimated at 5% among women aged 50 years and almost 50% at women of 85 years of age; it is estimated that any particular point of time, 4 to 33% of the population are suffering back pain [51]. According to the Global Burden of Disease data, more than 23% of European workers inform that they have experienced muscular pain in their neck, shoulders or upper limbs. That said, these conditions will now be shortly reviewed due to their high prevalence [1], [6].

- **Osteoarthritis:** Osteoarthritis is the most common joint disorder worldwide [52] and may occur in the hip, knee and the joints of the hand. It is characterised by focal damage and regions of articular cartilage loss within synovial joints - may be related to bone hypertrophy and thickening of the capsule. The principal symptoms include joint pain, stiffness, limitation of movement, occasional effusion, and may lead to local inflammation. The progress of the disease differs a lot, but it is often progressive - this evolution is followed by radiographs - and difficult to prevent. Symptoms may be relieved and function improved specifically by resorting to joint replacement procedures.
- **Rheumatoid arthritis:** Rheumatoid arthritis (RA) is a usual autoimmune disorder which is associated with gradual disability, systemic complications and early death [53]. It also has synovial joint involvement and leads to inflammatory complications. Despite being a systemic condition, it affects mostly peripheral joints. The cause of the condition is unknown; however, the advances in understanding the pathogenesis had lead to the development of new treatments that are effective at controlling progress and reducing long-term disability. Remission in people diagnosed with RA is at present around 10 to 30%.
- **Osteoporosis:** Osteoporosis is characterised by a systemic impairment of bone mass and microarchitecture deterioration that leads to bone fragility and susceptibility to fracture [54]. Currently, the diagnostic criteria for osteoporosis are based on the bone mineral density where this value is compared to the mean bone mineral density of young adult women. It is diagnosed as osteoporosis if below the standard value for more than 2.5 standard deviations. Clinically, it is acknowledged by the appearance of characteristic fractures after low-energy trauma - frequently at the hip, vertebrae, or distal forearm. Figure 2.1 represents the difference between a normal bone 2.1a and one debilitated 2.1b as a consequence of mass loss. Individual trabecular plates of bone are lost resulting in bone weakness and higher fracture risk.
- **Low Back:** Back pain is a common, episodic but recurrent and usually self-limiting condition [1]. It is the health disorder with the greatest influence on YLDs in the Global Burden of Disease. Low back pain may be defined as pain located below the line of the twelfth rib and above the lower gluteal folds, including or not leg pain. To be classified as acute, it must last no longer than six weeks, subacute if the duration is between six weeks and three months, and chronic if it lasts more than three months. Often, low back pain is never

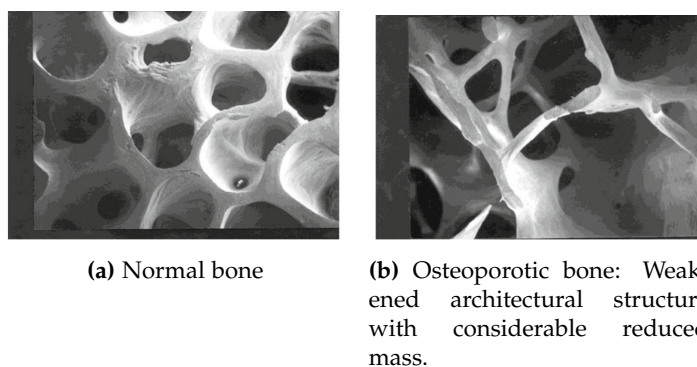


Figure 2.1: Changes within cancellous bone in Osteoporosis as consequences of bone loss, adapted from [55].

completely overcome, and patients experience periodic episodes of this condition.

- **Work-related upper limb disorders:** These are MSDs of the neck and upper limbs - which include the shoulders, any region of the arm, elbows, wrists and hands. It incorporates problems with soft tissues, muscles or ligaments, as well as circulatory and nerve supply to the limb. These conditions are caused or aggravated by work and may be a result of excessive static loading, repetitive low magnitude dynamic loading or poor posture. For this kind of repetitive strain injury, recovery is often unsatisfactory and may end with permanent disability and loss of employment [3], [56].

With all this into account, it seems relevant to take a step back and understand the underlying risk factors of such a syndrome.

2.1.1 Risk Factors of Musculoskeletal Disorders

Looking at the risk factors, the prevalence of these conditions increases significantly with age, and many are affected by lifestyle factors, like obesity and lack of physical activity. The ageing of the population gives rise to an additional concern since age is one of the most relevant risk factors. As a person gets older, musculoskeletal functions tend to decrease due to the advance of age-related degenerative disorders and loss of tissue strength - leading to an increase in the probability or severity of soft tissue damage [57]. As a consequence of the reduced resilience, many studies as [58]–[63], show that ageing has a strong correlation with the increase in the prevalence of MSDs.

Other risk factors associated with MSDs comprise obesity, as mentioned before, increased motorisation and work-related matters [5]. Low back pain is one of the MSDs where different studies found obesity or high body mass index to be a relevant factor associated with an increased occurrence [64], [65]. This factor is present in both developed and developing countries (even though developed countries have a higher rate of obese adults) and it is estimated, by two distinct studies, that obesity continues to increase - globally, until 2030 [66] and in Europe, until 2025 [67]. Additionally, the levels of motorisation are on the rise leading to a higher number of road accidents and, consequently, whiplash-associated disorders [5].

Lastly, MSDs are frequently seen as a work-related disorder - defined by the World Health Organization as a condition where the work environment and repetitive tasks contribute largely to the development or aggravation of the disease. These conditions are, generally, related to multiple factors that include physical, ergonomic and psychosocial factors. Repetitive work, painful or tiring positions, carrying or moving heavy loads are among the most common risk factors. Each profession may have specific contributions to the development of any MSD, such as exposure to vibrations, lifting or moving people, and prolonged standing or walking [3].

2.2 SWORD Health System

SWORD Health is a healthcare startup that is reinventing physical rehabilitation, enabling patients to perform their therapy at home while offering them an effective way to treat MSDs. SWORD Health has developed SWORD Phoenix, a novel digital biofeedback system, to overcome the dependence of current rehabilitation frameworks on specialised human resources,

which are expensive and insufficient. The technology is facilitating the distribution of care through the combination of science-driven therapeutic approaches with powerful technologies. Using both portable motion sensors and robust software, it provides home-based physical rehabilitation, a more efficient and cost-effective approach to physical therapy.

The solution is composed of three main components [68] (Figure 2.2):

- **Web-based Portal** (Fig 2.2a): A web-based platform that allows physiotherapists to carry out a detailed analysis of all patients individually, assessing his progress and adapting rehabilitation program in accordance to objective data, maximising efficacy. The Portal enables the team to control each patient continuously.
To prescribe a session, the clinician has to select the exercises, number of sets, number of repetitions and range of motion for each exercise. After the performance of a session, the results are uploaded to the platform and available for review. Results are available at any time from a global point of view (that gives the general picture, facilitating detection of red alerts), through a prompt analysis of each exercise, to a micro-level review of a repetition within a specific exercise. Based on this, the clinician may edit the prescription remotely. Clinical teams can keep in touch with patients during all the program, using for that a messaging app included in the Portal.
- **Inertial motion Trackers** (Fig 2.2b): Small and lightweight motion trackers placed on strategic and studied parts of the body, that digitise patient's movement with high precision. Each tracker comprises gyroscopes, accelerometers and magnetometers, allowing 3D movement quantification. The communication occurs via Bluetooth LE with a tablet computer. The trackers are placed on body segments using straps, in specific positions [42].
- **Digital therapist** (Fig 2.2c): AI-engine that guides the patient over each rehabilitation session with clinical-grade instant feedback, aiming to assure perfect exercise performance. Different feedbacks are provided to the patient, helping him to know how to perform each movement correctly. If the user violates a constraint, an error message is shown giving advises to achieve better results in the subsequent attempts. Minimal physical input to guide the patient throughout the session is required to him given the designed User Interface. Before each exercise, the system demonstrates to the patient, using for it a video and audio explanation. Also, by the digital therapist, patients are incited to get back any symptoms to the appointed clinician and are able to message the clinical team using the messaging app.

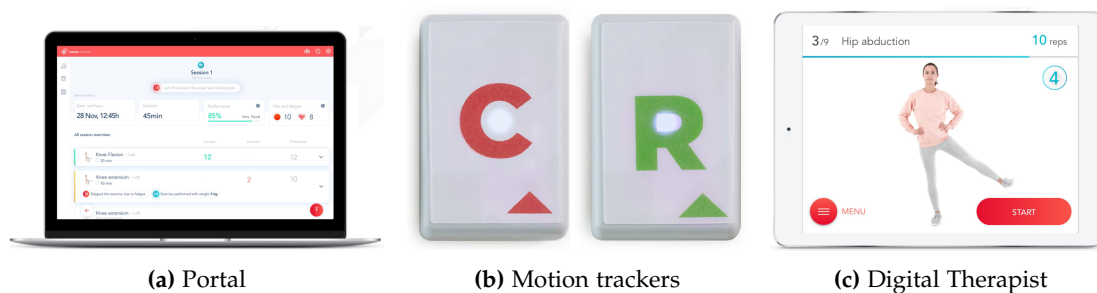


Figure 2.2: SWORD Phoenix Main Components.

By combining these three main components, the system appears as a scalable solution that allows patients to carry out therapy sessions without requiring real-time clinical monitoring but guaranteeing remote feedback. Thereby, patient discomfort due to travelling needs is minimised and, clinical resources optimised.

2.3 Sentiment-Analysis-Related Concepts in Rehabilitation

The classification of any affective-state representation requires understanding of precisely what are the variables that will be used in their characterisation. Regarding the final objective of developing an adaptive algorithm aiming to keep motivation levels high, this concept should also be revised.

That being said, this section explores the different concepts of motivation, pain, fatigue and satisfaction precisely in the context in which the problem takes place: during a rehabilitation program.

2.3.1 Motivation Concept

As stated in Section 1.1, motivation is seen as one of the main factors of influence during rehabilitation programs. Despite that, rehabilitation clinics are not in agreement about the concept of motivation and are usually doubtful about what should be seen as motivated behaviour. Scientifically, both [69] and [48] works state that within the field under study, motivation is generally viewed as an important concept, but also one that is difficult to measure objectively and prone to value judgement. That said, the concept should be explored with further attention.

Concerning the origins of the term, Eccles and Wigfield stated that the word *motivation* came from the Latin word *motus* that means *to move* and, basically, the study of motivation is the study of action [70]. *Motives* are the forces that encourage a person to act in a specific way, in order to accomplish a specific human need in a moment. It is rooted in a fundamental need as to minimise pain or maximise pleasure, or it may include other particular needs. Motivation may be evaluated concerning innate and internal drives or needs, inner stimuli and responses, and according to goals or directions. It includes intrinsic and extrinsic circumstances responsible for changes in the intensity, quality and direction of an underway behaviour.

In the Self-Determination Theory proposed by Ryan and Deci [71], it was distinguished different types of motivation based on the diverse reasons and purposes that cause an activity. The most primary distinction is between Intrinsic and Extrinsic motivation. The former pertains being engaged in an action only for the pleasure and satisfaction derived from doing the activity itself. If a person is intrinsically motivated, she performs the behaviour voluntarily, not requiring any reward or external approval. The action is conducted only for personal satisfaction, interest and enjoyment. Regarding the extrinsic domain, it does not refer to self-oriented behaviours, but behaviours driven by external rewards or interests. They are performed because of a belief to be an instrument for attaining an expected consequence or separable outcome.

There are recent works in the area aiming to review the concept of motivation in physical rehabilitation [48], [69]. Thus, they observed that studies might be divided into three different

groups. The first incorporates clinical research where motivation is studied as a persistent aspect of personality. It mostly depends on the individual inborn qualities, and it is not affected by external factors. A second group focus on the effect of social and environment on motivation. It tends to look at how the rehabilitation process and team, and the patient's social background may influence its motivation. The third group evaluates the importance of both individual characteristics and the therapy settings - also analysing the interactions between them. Although Maclean and Pound - authors of the aforementioned works - explored motivation as a characteristic of each person, they noted that this kind of approach might be problematic since it may be correlated with a moral viewpoint where the patient is blamed. It was also stated that motivation during rehabilitation should be associated with proactivity - although shall not be manifested in rejection of therapy - and reasonable compliance.

In conclusion, we may define motivation as the process that provides the individual with its energy and direction. It deals with an internal process that ultimately influences a person's persistence and strengths in goal-oriented behaviour.

2.3.2 Pain

Pain is an endless topic when discussing musculoskeletal conditions. Even though each patient interprets and feels pain differently, it plays a key role in their daily life. This section explores exactly the interplay between pain in rehabilitation and its effect on patients state of mind in order to support the exploration of pain as a potential representation of the emotional-state.

Pain may be defined as localised or generalised unpleasant bodily sensation. It is the way our body reacts to a threat by prompting protective behaviours as focusing on pain sources and avoidance. These complex sensations cause physical discomfort and emotional distress. That said, pain should be understood, so that, recovery will not look so difficult and unsustainable. A previous study showed that when pain is controlled, patients participation in rehabilitation increases [72]. A lack of pain education hindered pain management and increase the adherence to rehabilitative exercises, consequently, accelerating the recovery of their physical function. Remarkably, Vlaeyen *et al.* [73] noted that pain means more than just the vital signal of the body. The authors stated that it is the fear of movement and the incapacity of performing specific actions because of the concern of them being harmful.

Assessing pain might be notoriously hard due to its highly subjective nature. Pain is manifested by symptoms like anxiety and frustration what may be reflected in demotivation of the patient [74]. A previous study [74] showed that pain influences both function and mental well-being, such as the emotional-state. Pain reduction is of extreme importance during treatment since it was reported to be the primary stress factor of a patient. Cognitive and emotional factors carry out an important influence on pain: a negative emotional-state increases pain, whereas a positive state lowers it [75]. That said, it is advisable to monitor pain continuously during the recovery process.

2.3.3 Fatigue

Besides pain, fatigue is seen as the most disturbing and distressing symptom among patients with MSDs [76], [77]. Patients place high importance on fatigue due to its impact on the quality of life. This way, it is mandatory to define fatigue and understand its importance in patients affective state and in their motivation for the success of rehabilitation therapy.

Fatigue is defined as a progressive disability of the force-generating ability of muscle resulting in a reduced work-capacity and lower efficiency [78]. Generally, feelings of exhaustion, sleepiness and anger are linked to it.

People with musculoskeletal conditions who are experiencing fatigue report that it occurs on most days, albeit with varying intensity and frequency. Additionally, patients clearly differentiate between the normal tiredness felt before the MSD and the one resulting of it [76]. In the same study, physical fatigue has been associated with cognitive fatigue, manifested as an incapability to think clearly, maintain normal concentration and motivation levels. This emotional impact is portrayed by patients as frustration, irritability and bitterness.

The role of fatigue in musculoskeletal physical rehabilitation has already been highlighted [78], so that, it makes sense to follow fatigue levels during therapy progress as it is implemented in SWORD Health's product. Moreover, the lack of motivation and energy may be related to fatigue justifying the interest of assessing and exploring it as a component used to infer the affective state of the patient.

2.3.4 Satisfaction

Nowadays, patient-centred healthcare is the predominant framework of health service delivery. Home services, like the one provided by SWORD Health, have emerged within this paradigm of minimising patients inconveniences and increasing their adherence while maintaining remote control of their performance.

Nevertheless, patient satisfaction has arisen as an important patient-focused indicator of the quality of patient care. The relevance of keeping track of this measure is further emphasised when revealed that satisfied patients are more likely to adhere to treatment, benefit from their health care and have a higher health-related quality of life. [79], [80]. The feedback provided by patients about the services they attend should be used by physiotherapists to optimise quality and outcomes of patient care.

Therefore, this patient satisfaction should be measured. With that purpose, patient experience surveys are frequently presented to the patients where satisfaction is included. In search of a simple and straight-forward way of assessing satisfaction, Net Promoter Score (NPS) is gaining attention as a summary of client (patient) satisfaction. Bearing this in mind, NPS will be further visited in the next section.

2.3.4.1 Net Promoter Score

In 2003, the Net Promoter Score was introduced by Frederick Reichheld in the Harvard Business Review as the only number a company needs to grow [81]. According to Reichheld, you do

not need any "measurements of customer satisfaction or retention. You simply need to know what your customers tell their friends about you". It was claimed that companies require to stay abreast of their individual NPS, and this number establishes customer loyalty and the growth rate of the companies. He claimed that one question based on the probability of the recommendation of the customer to friends or colleagues is all it takes to determine the above-mentioned factors. Also, Reichheld rejected the necessity to take extensive surveys and other detailed statistical analysis as NPS may be linked with the financial performance and the growth of the company better than any other indicator.

From his point of view, the growth of a company is proportional to the number of promoters, and the willingness to promote is a strong indicator of loyalty. It was justified by the fact that customers put their reputation into play when they recommend a product or service to friends and family, so that, they only do it when they are deeply faithful. The customers may give an answer ranging from 0 - "not at all likely" - to 10 - "extremely likely". This measure groups the customers into three partitions: individuals that score a 9 or 10 are called *promoters* and it is assumed that they will give positive word-of-mouth advertising; individuals answering 7 or 8 are named *passive* and are considered indifferent; lastly, persons that answer from 0 to 6 are considered *detractors* and are likely to be displeased clients. Finally, the NPS is computed by subtracting the percentage of detractors from the percentage of promoters.

Despite the global acceptance by company managers, many people - from practitioners have not accepted and adopted the NPS concept and had raised doubts about the claims made by Reichheld [82]–[84]. Throughout these studies, different critiques are made to NPS, stating that it does not appear to be enough evidence that NPS is a lead indicator of business outcomes or that NPS focuses only on the user which is not always the person that makes the purchasing decision.

In the context of the problem, NPS is not looked as a measure of the company's growth, but individually, as a value of each patient satisfaction. The actual NPS of an organisation is computed at an aggregate level as explained; however, the data provided by this question in this scope has relevance on a personal basis. Typical *promoters* of NPS metric are seen as satisfied patients, while *detractors* considered unsatisfied. By assessing the likelihood of the patient recommending the healthcare received to another, it is intended to infer knowledge about his motivation and emotional-state. It is believed that there is a similarity between recommendation answers and satisfaction, since, if patients are not broadly satisfied and motivated, they will not suggest to a friend or family that they undergo the same procedure. Thus, this metric will also be explored, aiming to characterise user affective state.

2.4 Machine Learning

Traditionally, when approaching a specific task on a computer, we need to develop and apply an algorithm that uses specific *inputs* to transform them into several *outputs*. However, for some specific tasks, there is no specific algorithm that can be created, so the problem must be treated differently. In case the algorithm that transforms the inputs into outputs (*labels*) is not known, the computer must be able to generate its own to solve the problem automatically. If it is missing knowledge, we need to leverage data to support the decisions.

With the advent of computer technology, it is possible to store and handle large amounts of data and access it from remote locations through a computer connection. Thus, these facts opened the opportunity to make up for the lack of knowledge with algorithms based on large amounts of data [85].

Bearing this in mind, algorithms are being created to solve some learning tasks without using explicit instructions, relying on patterns and deductions. As follows, machines are able to extract these patterns from existing data to create a useful model for predicting new and unobserved data. Another important aspect to be highlighted is that, when employing this methodology, it is believed that, somehow, there is a process that can explain the results generated. The ability of a computer has to discover this process is named *Machine Learning*. Although it is not possible to fully identify the process, it is assumed that a good and useful approximation may be built by detecting specific patterns in data.

For this reason, Machine Learning is frequently linked with *Pattern Recognition* [86] since what is done is the search and recognition of certain patterns and regularities in data. Moreover, application of machine learning on large datasets is denominated *Data Mining* [85]. In this case, the construction of a simple model with valuable use and high predictive accuracy rests on a fundamental pillar: a large and reliable volume of data. Data mining is seen as a broad area comprising the study of collecting, cleaning, processing, analysing, and gaining useful insights from data.

To sum up, Machine learning is a data-driven approach, the scientific study of algorithms and mathematical models that a computer uses in order to perform a particular task without the need of explicit commands. The next Subsection will be focused on the typical data mining process.

2.4.1 The Data Mining Process

The data mining process includes different phases already enumerated in the previous section. With the emergence of knowledge discovery in databases, it is possible to define the different steps of a data-driven approach that may be defined in accordance with [87], [88]:

- **Problem Definition:** The final goals and the knowledge that one wants to extract should be accurately defined. An appropriate understanding of the application domain is required, as well as a practical application of the generated knowledge.
- **Data Collection:** The collection of the dataset is one of the first steps of every machine learning/data mining problem. Data Collection may require the aid of an expert (he could give suggestions on which attributes or features may be the most informative), and the use of specialised hardware such as sensors or software tools as Web forms. This specific stage is highly application-specific and often outside the domain of the data mining analyst; however, it is critical since any decision will have a significant impact in all data mining process.
- **Data Cleaning and Pre-Processing:** Data should be put in a specific format to make it suitable for processing and friendly to data mining algorithms. Before the extraction of relevant features, it is crucial to clean the data. Missing entries and outliers such as extreme values should be handled, some records may require to be discarded, or missing values

should be estimated. In the case of the data being extracted from multiple sources, they should be integrated into a uniform format to proceed with the processing.

- **Data Transformation and Feature Extraction:** A diversity of methods may be used either to eliminate insignificant features or to transform the current set to a new space that is more compliant for subsequent analysis. It is followed by the extraction of significant features to achieve the final goal. This procedure should result in a nicely structured data set that may be effectively used by the classification algorithms. Extracting the proper features is generally an ability that demands an *a priori* understanding of the particular application domain.
- **Feature Selection:** Analysis regarding the physical meaning of the transformed features obtained from feature extraction techniques. It is an automatic process of selection of attributes in the data that are most relevant to the predictive modelling problem that one is working on. This selection comprises the elimination of attributes that are irrelevant and may decrease the model accuracy and quality.
- **Analytical Processing and Classification Algorithms:** Machine Learning algorithms are employed in order to extract meaningful data patterns and create the knowledge expected. The processed data is used to design effective analytical methods. This step may be sub-categorised since it includes the choice of the data mining algorithm to employ (given the characteristics of the problem and the data) and the hyper-parameters values that should be assigned.
- **Interpretation and Use of Knowledge:** The results obtained must be analysed in order to understand the real meaning of them. They should be evaluated in terms of the application realm to reach the final goal. The integration of the acquired knowledge into the operating system and the use of it in the practical context are the final tasks to be met.

2.5 Summary

Throughout this chapter, musculoskeletal disorders were introduced, as well as the risk factors associated with it. SWORD Phoenix, the system developed by SWORD Health, was introduced as an effective solution to treat MSDs throughout physical therapy. The three main components were introduced, these being the Web-based Portal, the Inertial Motion Trackers and the Digital Therapist.

Afterwards, concepts related to sentiment-analysis in rehabilitation were defined, such as motivation, pain, fatigue and satisfaction. The future application of these variables in emotional-state recognition has been justified across the section, given the importance of patient compliance and quality of treatment.

To address the affect-detection task, Machine Learning may be an option; therefore, a brief introduction to the topic finishes the current chapter. The typical process that constitutes the phases of this data-oriented approach has been highlighted, having in mind the application in the context of the problem.

Chapter 3

Literature Review

Revised the problem fundamentals and the concepts related to the topic of the thesis, this chapter focuses on the up-to-date solutions and prevailing techniques present in the literature. Thus, it includes the research currently adopted to overcome the problem and the methods considered state-of-the-art. The focus of the chapter is precisely this, obtaining knowledge concerning Affective Computing fundamentals and the standard modalities that are being employed in the identification of affective states.

3.1 Affective computing

Affective computing (AC) has been defined by Rosalind Picard as "computing that relates to, arises from, or influence emotions" [89]. Nowadays, computers are beginning to get the ability to show and identify affect, and even the ability to have their own emotions. AC is responsible to assign computers human-like aptitudes as observation, interpretation and production of affect features. The machine should understand the emotional-state of humans and adapt its behaviour to them, giving an appropriate response to those emotions.

AC is one of the most emerging topics in research, having a broad potential. It is getting attention from researchers of different areas due to its wide spectrum of applications and multidisciplinary knowledge background (from psychology, cognitive, physiology and computer sciences) [90]. AC may have application in several scenarios and companies which may include the analysis of emotion and sentiments in their system. It has great potential as a subcomponent technology for other systems like recommender systems as it allows to expose which features customers enjoy or give negative feedback.

Although sentimental analysis and emotion recognition are two distinct research topics, they are connected under the domain of Affective Computing Research. Sentiment analysis is, generally, seen as a binary classification task - positive versus negative - performing a coarse-grained affect recognition. On the other hand, emotion recognition acts as a methodology under the topic of affect detection which involves fine detail, focusing on the classification of input data in accordance with a specific set of emotion labels [91].

This area of computer science may be divided into two main veins: detection and recognition of emotions and simulation of emotion in computational devices. The first one should hold the capability of perception and interpretation of human's feelings whether the last should give intelligent, sensitive and appropriate responses. Together they build an "affect model".

Affective computing technologies sense the affected state of a user through different components like sensors, microphone, cameras, among others. The computer may gather clues to user emotion from these diverse sources. Identifying emotional information demands the extraction of meaningful patterns from the extracted data. This is done using Machine Learning or Deep Learning (DL) methods that process different modalities, such as facial expression detection, natural language processing, physiological data, body gestures and speech recognition. These different modalities of affect detection will be reviewed in Section 3.2.

3.1.1 Affective States and emotions categorisation

Affective states may be defined as psychophysiological constructs that influence behaviour. It refers to "any psychological state that is felt and in same way is evaluative and valenced (positive or negative)" [92]. The affective phenomena include emotions, feelings, moods, attitudes, temperament and interpersonal stances [93]. Emotions and feelings assume an essential role in people's day-to-day life: they are crucial at decision-making, education, communication and social interactions. Focusing on the emotion, it is possible to define it, citing Scherer [94], as "an episode of interrelated, synchronised changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism".

Although it might look simplistic to describe emotions in discrete lists (a discretisation fails to characterise the full range of emotional experiences intrinsic to any of us) -, different studies divide them into categorical sets. Philosophical studies on emotions were initiated at ancient times with the Greeks and Romans, however, studies on the evolutionary theory of emotions date back to the late 19th century by Darwin. Darwin's theory reflected his evolutive view and stated that also emotions advanced via natural selection and, consequently, have universal counterparts across cultures. At the beginning of the 1970 decade, Ekman found proof suggesting that humans share six basic emotions. A basic emotion is defined by a group or neural correlates in the brain, a set of bodily expressions and a motivational part for action tendencies. Still one of the most popular divisions, he separated the basic emotions in the following categories: happiness, sadness, fear, anger, disgust, and surprise [95]. Having in mind that any description will cover a restrict portion of daily emotional demonstrations, the selection of affect categories should be made in a realistic and context-dependent mean. It is possible to find work in the bibliography of different attempts do distinguish other affective states as satisfaction, confusion, frustration, anxiety and fatigue [92].

The dimensional description acts as an alternative to categorical representation where an affective state is represented as coordinates in a set of dimensions defined by psychological concepts. This methodology is attractive because they allow describing emotional-states in a way more flexible than using just words. Furthermore, they allow a non-discretization of emotions and variations over time.

Not only for theoretical but also for practical reasons, several researchers in the area prefer to define emotions in a multi-dimensional space [96], [97]. An early study by Russell was the circumplex model that used a couple of dimensions - pleasure/displeasure and degree of arousal - to plot 150 affective labels [96]. In a model of this type, proximal affective states are arranged to lie adjacently on a circle, while dissimilar ones are disposed of in opposition to each other. Different bi-dimensional approaches were proposed, as it is the case of the work of Whissell [98] whose dimensions considered are evaluation (how a person feels from negative to positive) and activation (likelihood of a human to take some action under the emotional-state from passive to active). The same author developed the *Dictionary of Affect in Language* which attributes a pair of values to more than 4,000 affective-related words [98]. In the following of Darwin's thought, it emerged the Plutchik's *Wheel of emotions* which had offered a theory based on evolutionary thoughts [99]. Today's emotions are considered adaptations that had evolved for the primitive will of keeping the subject safe. This *Wheel of emotions* consists of 8 basic emotions and 8 advanced (result of the combination of two basics). In this model, shown in Figure 3.1, the vertical dimensions assign intensity, and the radial one depicts the level of similarity among emotions.

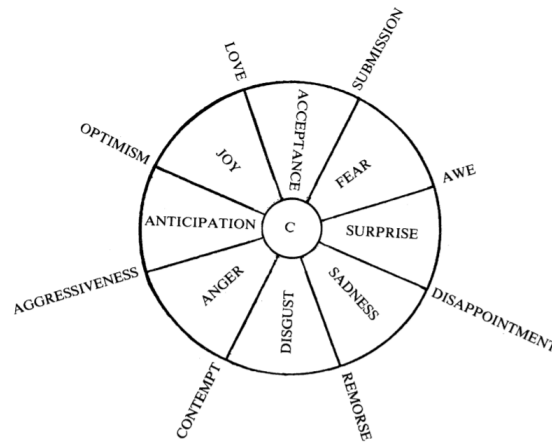


Figure 3.1: Wheel of Emotions proposed by Plutchik, adapted from [99].

In addition to 2D spaces, a commonly used structure for emotion representation is the <arousal, valence, dominance> set which may be found in the literature with different names as <evaluation, activation, power> or <pleasure, arousal, dominance> [100]. A typical model applied in affective recognition is the PAD-model which uses the dimensions just mentioned. *Arousal* represents the activation, level of mental alertness and physical activity and *dominance* represents the feeling of control over one's situation and surroundings versus feeling under control or influenced by external conditions. Finally, *pleasure* corresponds to cognitive judgements of evaluation - higher evaluations of stimulus are associated with greater pleasure induced by that stimulus [100]. Recent works indicate a set of 4 dimensions, specifically <valence, potency, arousal, unpredictability> [101].

Despite all the advantages of dimensional approaches, they present some limitations. Although it enables the comparison between affect words -in terms of their difference -, it normally does not allow operation between them neither have in account the fact that two or more emotions may be felt at the same time. Also, these approaches work at word level what may be a

problem in some specific languages. As consequence of these limitations, it appeared the Hourglass of Emotions [102], a new affective representation based on discussed Plutchik's method. The following model, shown in Figure 3.2, illustrates affective states through labels, and four independent and dissimilar dimensions. The model assumes an hourglass shape given that the affective states are represented relative to their strength (from strongly positive to null to strongly negative).

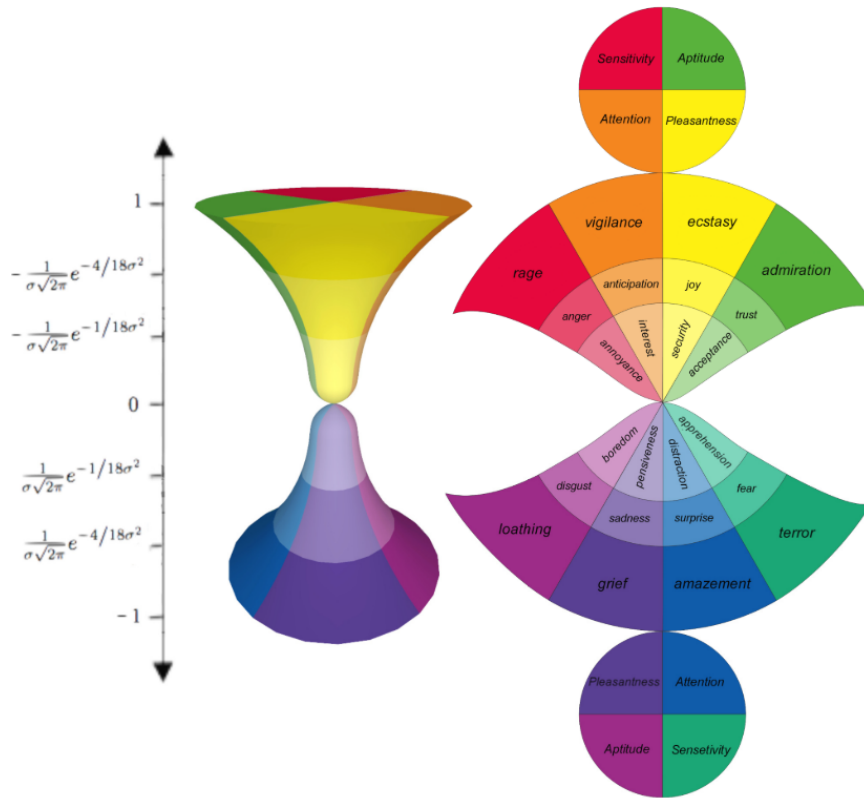


Figure 3.2: The 3D model and the net of the Hourglass of Emotions, adapted from [102].

It has potential to characterise the entire scope of emotions occurring in a natural communication setting. These multiple activation levels permit to the Hourglass of Emotions covering situations where up till four emotions are expressed. Besides that, it is possible to address emotions in an algebraic manner and to compute polarity based on emotions - what was seen as the first tentative to bridge sentiment analysis and emotion recognition.

3.2 Affect Detection

Any approach to an Affective Computing system rests on a fundamental pillar: the recognition of the affective states and detection of emotions. In order to develop any interface emotion-sensitive that responds automatically during the interaction of the user with the computer, it is essential to sense user's affective states primarily. The focus of this section is on the current and upcoming techniques for affect detection and data that support the *Affect Detection*.

Affective detection should be as accurate as possible, though, it is a very challenging problem due to the nature of emotions. Emotions are not a measurable quantity, they are *constructs* with unclear boundaries. Each person has their own way of manifesting an emotion, resulting in tremendous individual variations in expression and experience.

The review of the current affect detection methods is organised with regard to each modality or channel. The review includes the inference of emotional through facial emotion, audio expression, body gestures, text and physiological data. All individual modalities have their advantages and limitations towards its use as an applicable emotion recognition method. The validity and reliability of the signals in real-world settings, the time resolution of the signal, and the cost and intrusiveness for the user of the modality are factors that affect the value of each technique.

3.2.1 Facial Emotion Recognition

The detection of the basic emotions from facial expression has been the greatest focus of research since AC science appeared. The research of facial emotion has gained even more attention over the past decades as results of its application not only in AC but also in perceptual and cognitive sciences and computer animations. Facial expression is seen as a primary signal for understanding people emotions and sentiments. In a natural communication environment, the face is fundamental to form an impression of any subject state of mind and to understand the intention of others. It regulates face-to-face interactions, indicates reciprocity and interpersonal attraction or repulsion, and enables the communication of subject feelings even between members of different cultures [103]. Ekman and Friesen claimed that it is possible to detect the 6 basic emotions (suggested by Ekman in [95]) from cues of facial expression.

It is possible to divide researches on Facial Emotion Recognition (FER) into two groups according to whether the features are handmade or automatically produced by the output of a deep neural network. Regardless of the approach used, there are numerous challenges that should be considered. These include: non-frontal pose and significant head motion; facial actions are by its nature subtle, making them hard to model; the temporal dimension is not always constant; inherent differences between subjects' appearance and face shape make generalisation difficult.

3.2.1.1 Conventional Facial Emotion Recognition Approaches

In conventional FER approaches, the procedure may be divided into three main steps, as shown in Figure 3.3: (1) face and facial components detection, (2) feature extraction, and (3) expression classification. In the first step, the face is identified from the input image, and its parts or landmarks are emphasised from the face region. Secondly, distinct features may be extracted from the components and they include spatial and temporal features. Last, the classifiers (that must be trained previously) use the extracted features to produce the classification results [105].

As consequence of this increased traction gained by facial expression, a significant number of measurement systems for facial recognition was developed [104], [106]. These systems encode the movement of specific facial muscles and may be used for further feature extraction.

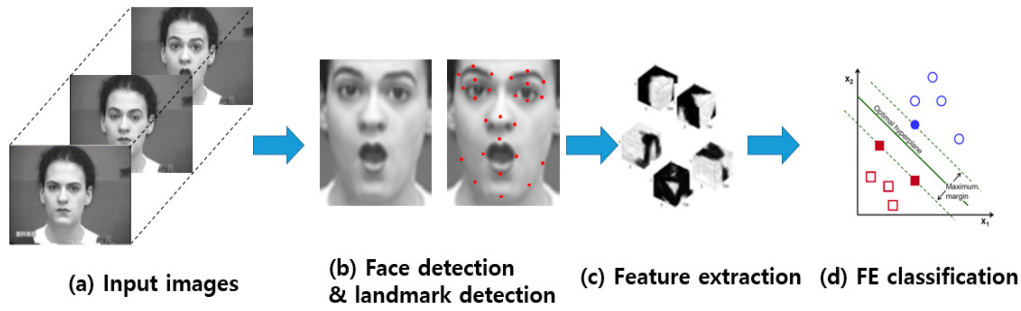


Figure 3.3: Procedure used in conventional FER approaches, adapted from [105]. In the input images (a), the face region and main salient points are detected. (b) From face components and landmarks, spacial and temporal features are extracted, (c) and the facial expression recognized using for that purpose pre-trained classifiers.

A significant number of facial expression recognition systems are based on a detailed annotation scheme named Facial Action Coding System (FACS) developed by Ekman and Friesen [104]. It should be highlighted since it turned to be the leading method for objective classification of expressions. This system applies objective component facial motions or *facial actions* to measure frequency and intensity of facial activity and, consequently, deduces what is mentioned as an Action Unit (AU). A set of AUs - the smallest discriminable movements detectable in facial expression - were identified and then used by trained human coders to reconstruct expressions in terms of them. The facial AUs belonging to the six main emotions are illustrated in Figure 3.4, which is adapted from [107].

AU1  Inner Brow Raiser	AU2  Outer Brow Raiser	AU4  Brow Lowerer	AU5  Upper Lid Raiser	AU6  Cheek Raiser	AU7  Lid Tightener
AU9  Nose Wrinkler	AU10  Upper Lip Raiser	AU12  Lip Corner Puller	AU15  Lip Corner Depressor	AU16  Lower Lip Depressor	AU17  Chin Raiser
AU20  Lip Stretcher	AU23  Lip Tightener	AU24  Lip Pressor	AU25  Lips part	AU26  Jaw Drop	AU27  Mouth Stretch

Figure 3.4: A List of AUs that are pertained to the six basic emotions, adapted from [107].

Although FACS allows the distinction between facials actions, it gives no direct inference on emotions. The Investigators' Guide, work of Ekman and Friesen, may be used therefore to infer emotions and, this way, works as the last step of a conventional FER normal pipeline. This guide was updated more recently, including at this day details not only on head movements but also eye positions. A seventh emotion - *contempt* - was also added where each AU description and AU combinations were aim of further study [108].

The work of Ekman and Friesen had inspired many researchers in the area to employ image and video processing methods - most of them already tested with different applications - to ex-

amine facial expressions. Various types of conventional approaches have been studied, but they all follow the typical pipeline, varying only the type of features extracted (geometric, appearance or hybrid) and the way of detecting the face region.

Thus, any algorithm begins with face detection. The most widely used face detector, with proven excellent results in the case of a relatively frontal pose, is the Viola and Jones [109]. However, there are several approaches as it may be found in the review performed by Zhang and Zhang [110]. Then, it is important to keep track of the face - whether using a landmark or a group of facial features. The latter one may be advantageous since it affords information that may be used to infer a 3D pose. To track a set of facial features, Active Appearance Models (AAMs) are frequently employed [111], [112]. AAMs dissociate the shape and appearance of a face image. They align the shape model to an unseen image containing the face and facial expression of interest when a pre-defined linear shape model with linear appearance variation is given. It is important to refer that AAMs are dependent on the person and should be trained individually. To overcome this problem, Saragih *et al.* [113] developed a technique known as constrained local model. These allow the generalisation regarding unseen appearance variation and offer invariance to global illumination variation and occlusion. Although they are fast enough to maintain real-time tracking, their detection of shape is less accurate.

About the type of features extracted, geometric features refer to facial landmarks - as the eyes or brows - that may be used to discriminate between the mentioned action units. For instance, geometric features with a multi-scale and a multi-orientation Gabor Wavelet-based representation were applied to the identified purpose [114]. Also for the geometric features, both Suk and Prabhakaran [115], and Ghimire *et al.* [116] used the relationship between facial landmarks to construct feature vectors for training. Figure 3.5 represents a few examples of landmarks that are used in feature extraction as well as some geometric features create from distances between these landmarks.

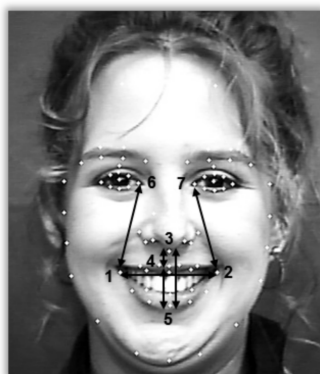


Figure 3.5: Example of landmarks and some points and distances that may be used as geometric features, adapted from [115].

For appearance features, they represent changes in skin texture such as wrinkling and deepening of facial grooves and pouching of the skin [103]. Techniques that allow the description of local image texture as local binary pattern (LBP) have been proposed. They may be extracted from the global face region [117] or from independent face regions that contain different type of information [118]. LBP histogram may be applied to extract global features and then used

as feature vectors or may be applied into categories based on the components of the face that contribute more toward an expression (Stepwise linear Linear Discriminant Analysis was employed in [118]). For the former example, LBP histogram of different block sizes was extracted and used as features that are classified in diverse facial expressions using a principal component analysis (PCA). A global approach may have some limitations since it cannot reflect local variations of the facial components to the feature factor. Although it may be carried out in real-time, it does always give the same importance to the different regions whether or not they carry relevant information. For instance, the forehead or cheek does not contain any pertinent information if comparing with eyes or mouth. Ghimire *et al.* [119] divided the total area of the face into domain-specific local partitions in order to extract appearance features specific to the region. In this study, the most important local regions were found through an incremental search approach, resulting in feature dimension reduction and improvements in recognition accuracy. The face is divided into 29 regions based on expert knowledge regarding AUs given the result shown in Figure 3.6. The same figure, adapted from [119], also presents the feature extraction process where a uniform LBP pattern from each selected local region is computed.

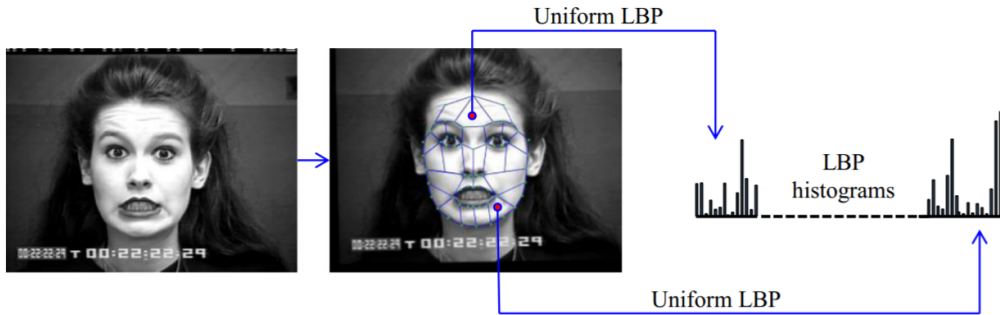


Figure 3.6: Local region face representation and uniform LBP feature extraction process from each region, adapted from [119].

In the work developed by Haro *et al.* [120], PCA techniques are also used to enhance facial features. Furthermore, in Gunawan *et al.*'s study [121], Kinect motion sensor is used to identify the face region on depth information and AAM track the detected face during a video recording. The function of AAM is to adjust shape and texture model in a new face when it is noted a variation of these characteristics in comparison with the training dataset. The knowledge obtained from Ekman is essential since, in recognition of the facial emotion, the basic logic behind FACS is used.

Concerning hybrid features, different works in the area have merged geometric and appearance features to overcome the limitations of each one of the approaches that may lead to better results in certain cases [119], [122]. A typical example of geometric and appearance feature-based methods is the one of Zhang and Ji [107], who combined the use of facial points (26 points including a few around the eyes, eyebrows, and mouth) and temporary features like crow-feet wrinkles and nasal-labial furrows. In Figure 3.7, it is represented the method suggested by Zhang and Ji.

As conclusion, in conventional approaches is essential to determine the features and classifiers. For the first case, many state-of-the-art techniques may be applied to obtain handcrafted

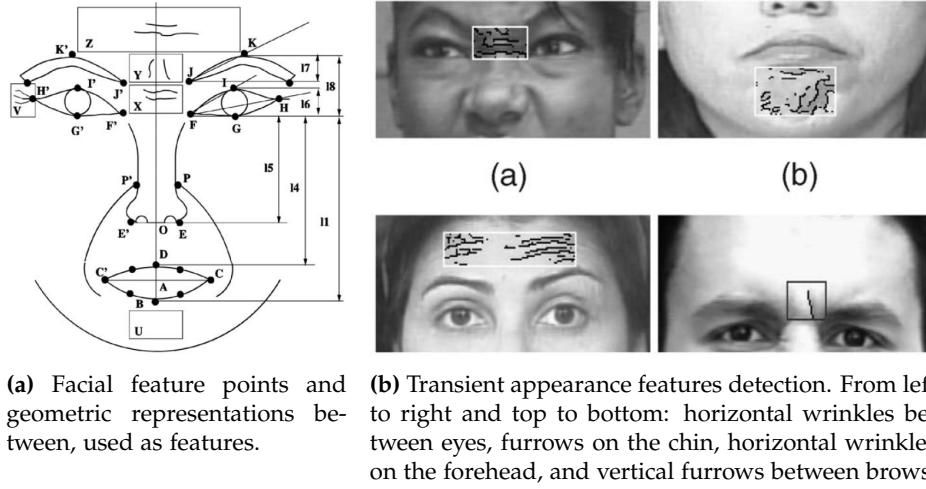


Figure 3.7: Dynamic method proposed by [107], combining geometrical and appearance features.

features (as Histogram of Gradients, LBP or simply the distance and angle between face landmarks). In terms of the pre-trained classifiers, Support Vector Machine (SVM), AdaBoost and Random Forest are the state-of-the-art algorithms applied in facial expression recognition. This kind of approaches are computationally less demanding in terms of power and memory than Deep Learning techniques that way they are still being studied for application in real-time systems. The programmer should optimize feature extraction and classifiers in order to achieve the expected performance.

Although it was addressed the key principles when dealing with the recognition of facial expressions from images, most of these methods are not appropriate to model temporal information. It is important to discuss some methodologies that can be used to affect recognition from videos. Deformable models are the standard techniques that have been used to track facial features in videos [123]. Maintaining accurate tracking across the video sequence is of critical interest, and these deformable methods have demonstrated a significant advance in terms of facial tracking and facial expression analysis accuracy [123]. Following this topic, a few automatic methods (both image and video-based) for detection of facial features and, consequently, facial expressions were proposed and they may be revisited in [124]. It is also possible to extract some specific motion features as optical flow and dynamic textures. Applying these techniques, Cohen *et al.* indirectly restored motion representations by developing features in a manner that each feature motion corresponds to a simple deformation on the face [125].

The last step is the classification of the extracted features, and for that, an extensive number of classifiers may be used according to the feature extraction approach. Multi-class AdaBoost with dynamic time warping and SVMs are among the most popular algorithms when dealing with video both in supervisor or unsupervised learning.

3.2.1.2 Deep Learning based Facial Emotion Recognition Approaches

After describing the use of handcrafted feature extraction from a visual modality, it is time to review Deep Learning approaches. With the breakthrough in deep learning algorithms - namely Conventional Neural Networks (CNN) and recurrent neural networks (RNN) -, it is possible now

to extract features automatically without previous intervention. The Deep Learning framework has achieved state-of-the-art results in diverse fields (e.g. computer vision, signal processing and Natural Language Processing (NLP)) and based in this recent success, emotion and sentiment analysis have initiated the adoption of these algorithms.

Thus, CNN visualization techniques were employed by Breuer and Kimmel [126] to understand images. It demonstrated (by application on different datasets and FER tasks) the ability of networks trained on affect detection. With the same purpose, it was designed a novel visual sentiment prediction network inspired on transfer learning techniques from a CNN pre-trained on large scale data for object recognition [127]. The main objective was to show that there is not any requirement of domain knowledge to perform sentiment prediction. Jung *et al.* [128] combined (using an integration method) two different types of CNN to improve the performance of the algorithm of expression recognition. The first is used to extract temporal appearance features from the image sequences, whereas the second is responsible for the extraction of geometric features from the temporal facial landmark. The approach under discussing is represented in Figure 3.8 where the purple box comprises the joint fine-tuning method responsible for integrating the outputs of both networks.

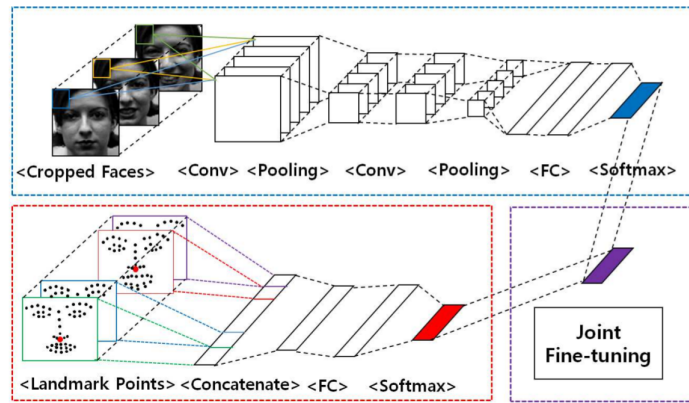


Figure 3.8: Overall structure of the approach extracted from [128]. Blue and red boxes represent the two architectures of CNN implemented. The integration method is represented in the purple box. Conv and FC relate to Convolutional and Fully Connected layers, respectively.

However, CNN-based approaches do not take into account temporal variations in the facial components. It was of paramount importance the development of methods that could deal with the temporal sequence of images - videos. Therefore, a different approach that combines CNN (for spatial features and individual frames) and long short-term memory (LSTM) (for the temporal features of consecutive frames) was employed. LSTM networks are a specific type of RNN, which are designed to learn long-term dependencies and are capable of overcoming the vanishing gradient problem that affects traditional RNNs. The problem is solved by using short-term memory. As all recurrent neural networks, LSTM has a chain-like structure, even though the repeating units present a different design.

This application of LSTM (or other RNN models) for analysing sequential images have different advantages when compared with the existing procedures. Among them, we should highlight the fact that it supports both fixed and variable-length inputs and outputs. There are several works that use this combination of CNN and LSTM's stacked and trained together [129], [130].

All the works rest on a similar structure (combination of an LSTM with a deep hierarchical visual feature extractor) that enables the recognition and the synthesising of temporal dynamics for tasks involving consecutive images. The overview of the structure proposed by Chu *et al.* [130] is shown in Figure 3.9. The network includes both CNNs and LSTMs and makes use of both cues to perform an automated AU detection.

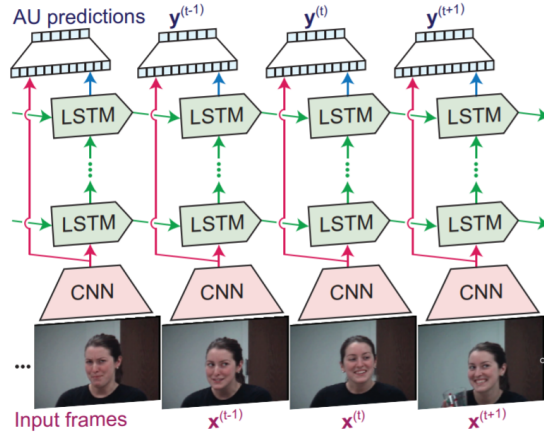


Figure 3.9: Overview of the suggest hybrid Deep Learning framework, extracted from [130].

Additionally, recently deep 3D convolutional networks (C3D) have been proposed for spatio-temporal learning [131]. This network, represented in Figure 3.10 consists of 8 convolution layers, 5 pooling layers, 2 fully connected layers, and a softmax output layer. Comparing with the results obtained with traditional 2D CNNs, the network has proved to be more flexible for spatio-temporal feature learning. Only using a simple classifier with the learned features, it outperformed the referred state-of-the-art methods by creating a more accurate architecture with 3D convolutional kernels in all layers.

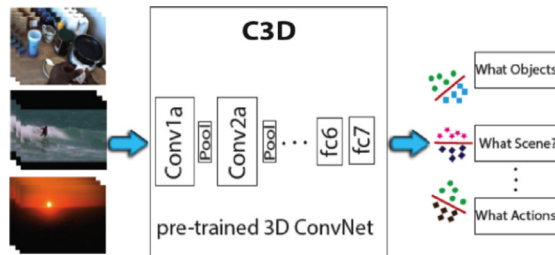


Figure 3.10: C3D network used in the extraction of spatio-temporal video features, adapted from [91]

As final remarks, the main advantage of Deep Learning based FER approaches is to highly reduce the dependence on pre-processing techniques or features extraction algorithms by performing an end-to-end learning in the pipeline immediately from the input images (or video frames). On the other hand, these approaches have a few limitations namely: the need for extensive datasets, huge computing power and amount of memory, and the consume of a lot of time both during training and testing stages.

3.2.2 Audio Expression Recognition

Similarly to the visual modality, research on audio expression recognition is also influenced by the basic emotion theory already explained. That said, most of the works present in the literature that will be reviewed chose this direction, aiming to distinguish the basic emotions or a subset of them. Speech conveys affective information not only through the explicit linguistic message - what is said - but also through the implicit linguistic features of the expression - the way it is said. The latter will be the focus of this section, whereas the explicit part (text) will be discussed throughout Section 3.2.4.

Different types of features can be extracted from everyday sounds that may lead to emotion and sentimental analysis. The main used features that are teaching machines how to identify emotions are prosodic and acoustic. The earliest research in the area aiming to extract audio features has focused on the phonetic and acoustic properties of spoken language. Based on psychological studies linked with emotion, it was revealed the influence of vocal parameters (as intensity, pitch, voice quality and speaking rate) in recognition of emotion and sentiment analysis [132]. Bearing this in mind, researchers started targeting affective reactions in daily sounds, and it led to a set of applications able to decode the basic emotions using prosody or other non-linguistic vocalisations as laughs or cries.

Given that emotional characteristics are more notable in prosodic features, these are widely applied to this purpose and are present in the majority of applications and research systems. Prosodic features that are features based on the pitch and energy contours of speech are known to give information about the emotional-state of the speaker [133]. In the work of Koolagudi, Kumar, and Rao, energy and pitch parameters are extracted from utterances, words and syllables individually in order to model emotions.

Acoustic parameters work as a different set of features, however, it was showed that these are dependent on personality traits and the setting of the oral presentation. Furthermore, there is some ambiguity regarding the performance of acoustic features with basic emotions. These drawbacks lead to a series of works intended to identify features that allow a better analysis [134].

Besides pitch (quality of the sound controlled by the rate of vibrations) and energy-related features - already noted their key role in affect recognition -, other features as spectral ones have been used by some researchers for feature extraction. All of the mentioned features and others may be reviewed briefly in accordance with [91]:

- **Mel Frequency Cepstral Coefficients (MFCC)** are coefficients that together form a short-term power spectrum of a sound or an audio clip, known as mel-frequency cepstrum. The MFCC performs the best possible approximation between a linearly spaced frequency band distribution and the human auditory system. The linear cosine transform of the log power spectrum is used to compute the discussed feature.
- **Spectral centroid** identifies the centre of mass of the magnitude spectrum of the sound. It provides an indication of the brightness of the sound that may be used in emotion analysis.
- **Spectral flux** measures how quickly the power spectrum of the audio signal changes over its entire duration. The Euclidean distance between two normalised spectra should be calculated.

- **Beat Histogram** is a histogram that demonstrates the relevance of several rhythmic periodicities in a sound. It is normally computed by taking the Root Mean Square of 256 windows and then the Fast Fourier Transform is applied at the input.
- **Beat Sum** is employed to highlight the basic beats of a signal. It is estimated as the sum of all entries in the beat histogram.
- **Strongest beat** is the strongest beat in a signal and, the beat histogram is again used for identifying it.
- **Pause duration** measures the time a speaker is silent during the whole audio record.
- **Pitch** shows the quality of the sound by controlling the rate of vibrations producing it.

It is important to mention that the features are employed to classify both discrete feeling states (the known universal emotions) and states based on the dimension as valence, arousal or dominance. Regarding the speaker-dependency of the approach, it was shown that speaker-dependent one outperforms the speaker-independent as shown in the work of Navas *et al.* [135]. 98% of accuracy was obtained with a Gaussian mixture model as a classifier and several audio features, namely prosodic, voice quality along with speech features (for example, MFCCs). Although the state-of-the-art results of this approach are relatively better, this technique may not be applied in practical applications that address a considerable number of users. One of the best classification accuracies found for speaker-independent applications was 81% where a two-stage classifier and a singular set of features (from voice to prosodic and spectral features) was implemented [136]. This result was obtained on a benchmark database [137] of emotional speech where ten actors simulated emotions, producing ten statements that may be used in daily basis communication and may be interpreted in all tested emotions.

Additionally, audio affect classification may follow to distinctive logics: extracting local features or global features [91]. The former approach is seen as the most common where each audio or utterance is divided into several segments and examined separately. The signal is considered static and the features removed from these parts named local features. In contrast, global features measure global statistics of the whole signal. It includes average, deviation and other metrics of the local features. The main advantage of this approach rests on the enhancement of speed computation - fewer numbers compared to local features lead to fast computation. Like any approach in the literature, there are limitations associated as they cannot be used to distinguish between affective states of similar arousal (anger and disgust are two examples of high arousal not easy to discriminate). Moreover, the absence of temporal information and relations between segments in a statement may be seen as drawbacks.

Concerning the classifiers used for the research of emotion recognition, there is no clear restriction of what can or can not be employed. The classifier selected should be dependent on the set of features used and what kind of emotions we hope to identify. For instance, using prosody features, Dellaert *et al.* [138] compared the efficiency of three distinctive classifiers: maximum likelihood Bayes classification, kernel regression, and k-nearest neighbour in recognition of 4 basic emotions - sadness, anger, happiness and fear. Other examples of classifiers comprise SVM or Linear discriminant classification [90].

As it was described in Section 3.2.1.2 for visual modality, Deep Learning is gaining popularity in affect recognition research. In the context of audio emotion classification, Deep Learning is also being applied with promising results. In the work of Huang *et al.* [139], it was proposed

the use of an autoencoder followed by a CNN in order to learn affect-salient features for Speech Emotion Recognition. The CCN is responsible for deducing a hierarchical representation of input data that promotes categorisation. The overview of all system is presented in Figure 3.11. It was used a spectrogram of speech signal as input and an SVM classifier to identify emotion class. The network had to be trained on the features extracted from all time frames since, as already discussed, this type of model is not capable of modelling temporal information. However, the experiment of Huang *et al.* showed, testing on benchmark datasets, that it is possible to achieve a stable and robust recognition performance in complex scenes. Again to overcome this problem, LSTMs are frequently employed on handcrafted audio features.

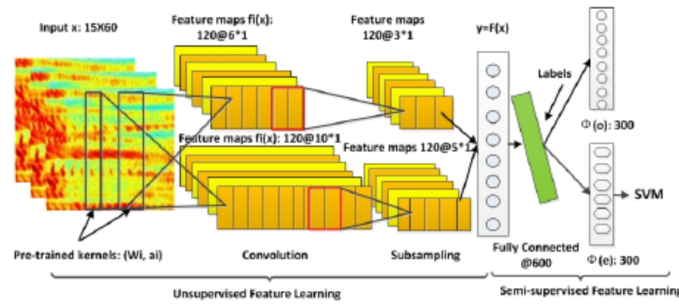


Figure 3.11: Overall system structure, extracted from [139]. The system comprises an input layer, one convolutional layer, one fully connected layer and a SVM classifier.

Contrarily to what is normally done in the visual modality, deep networks are not being used to extract features, but only for classification. An expectable future approach is to construct neural networks capable of automatically extract features from raw audio data. It has already been demonstrated through a pilot study [140] that convolutional networks have the potential to produce their own features from aural data that may afterwards be trained for the final task of emotional classification. Through this work, it was demonstrated the aptitude of CNNs and LSTMs individually for feature extraction and emotion classification, respectively. The full approach is shown in Figure 3.12.

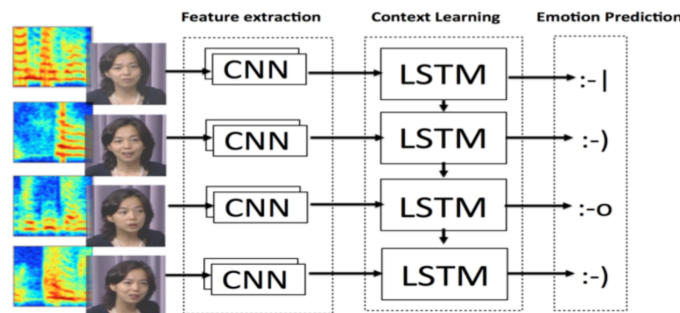


Figure 3.12: Overall implemented approach, extracted from [140]. Features are extracted from aural data using CNNs whether LSTMs are used in emotion prediction.

Last, it was found in the literature review another significant method that does not require manual feature removal and demonstrated improvements comparing to other state-of-the-art techniques [141]. In this paper, the authors proposed the utilisation of deep neural networks to

both extract high-level characteristics from non-processed data and to generate an emotion state probability distribution for each speech utterance. The segment-level probability distributions are used to build features of a higher level (of the entire speech recorded). These features are then fed into an extreme learning machine - a simple and efficient single-hidden-layer neural network - to identify emotions.

3.2.3 Body Gestures and Posture

Although the majority of the research performed in the area of Affective Computing had overlooked body gestures and posture as a serious contender to facial expressions and acoustic features, it is known that these may be related to emotions. Research in psychology had suggested that body gestures can provide an important source of features for emotion and sentiment recognition [91]. Kapur *et al.* [142] shown how simple statistical measures of motion's dynamics caused by body movements may be employed to distinguish between the basic emotions.

Moreover, there are quite a few benefits of using body movements and position to diagnose the affective states of a user [93]. Comparing only to the face, the whole body is relatively larger and, thereby, have multiple degrees of freedom and countless singular configurations. A potential affective communicative channel may arise from the innumerable combination of static positions or gestures aligned with temporal dependency. Besides, posture and gestures may offer information that may be unavailable from the conventional features such as the face or speech when registered over longer distances. Recognition of facial features when recording at a distant range may be impaired due to a lower resolution of the image [143]. In the same conditions, it is also challenging to record speech and, this way, use the audio modality to infer any psychological state.

However, the most significant benefit of gestures or posture-based affect detection is that natural body movements are generally unconscious, unintentional, and consequently not susceptible to social manipulation. Facial expressions or speech intonation and rhythm may be more prone to conscious editing, and it leads to increased difficulty in distinguishing emotions through conventional channels.

Inspired by this motivation, an automatic recognition framework was proposed using a set of postural, kinematic and geometrical features from sequences 3D skeletons [144]. These are acquired through two different systems: a professional grade optical motion capture system and Microsoft Kinect, and fed to a multiclass SVM classifier for emotion recognition. The method was assessed on the six basic emotions and reached an overall accuracy of 61.3%.

Mota and Picard [145] presented a system for recognising naturally occurring postures to infer the affective states related to the user's interest in a learning environment. The position of the user was gathered using a body pressure measurement system that consisted of two matrices of pressure sensors. This system can be mounted on a variety of surfaces as the seat and back of a chair as shown in Figure 3.13. The example pressure patterns of both back and seat bottom are also present in the underlying figure. Posture features are extracted using a mixture of four Gaussians and input to a neural network that classifies the nine postures. Then, Hidden Markov Models are used to analyse temporal transitions among these posture sequences with the purpose of determining a child's level of interest. It was obtained an overall accuracy

of 76.5% when tested with posture sequences coming from new users. Different works are extending the system just mentioned and were also employed during a learning task. A system based in Body Pressure Measurement was suggested by D'Mello and Graesser [146] and was developed to identify other affective states than interest as engagement, confusion, frustration or delight.

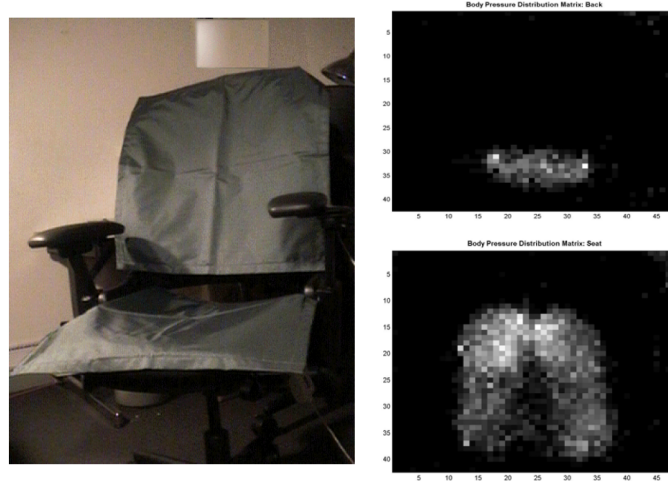


Figure 3.13: Figure representative of system idealized in [145]. Left: body pressure measurement system. Top right: Example pressure patterns for back. Bottom right: Example pressure patterns at seat bottom.

Comparing with affect recognition using facial expression and speech as inputs, we may say that body gestures and posture approaches require, generally, more expensive, not practical and sophisticated equipment. On the other hand, posture tracking does not interfere with normal user experience.

3.2.4 Emotion Recognition from Text

In the following section, it will be reviewed the state of art of emotion recognition and sentiment analysis using text to extract relevant features. The automatic identification of emotion and sentiments from text with the objective of distinguishing between simple emotions as anger, joy or sadness, among others, has been addressed by several researchers. We roughly may divide the existing studies into two main groups: Emotion recognition based on traditional methods or based on Deep Learning approaches.

3.2.4.1 Traditional Methods based Emotion Recognition from Text

The traditional methods found in the state-of-the-art may be now sub-categorised into two partitions: 1) methods based in lexicon; 2) Machine Learning approach.

Lexicon-based methods utilise pre-defined lists of words that are categorised in accordance with diverse emotions. Since these lexicons are normally compiled manually, they may be exploited for keyword matching. As an example, the NRC Word-Emotion Association lexicon [147] - that include the annotation of around 14,000 words in English - was derived manually with

the help of crowdsourcing. Another popular emotion lexicon that was not only generated by experts from the area of psychology is the WordNet-Affect dictionary [148]. WordNet-Affect dictionary was created starting with a seed of words labelled as affect by experts and then, it was expanded by assigning scores to a larger number of words in terms of the proximity to the seed words. Other attempts to generate emotional lexicons that may be found in the literature include DepecheMood or LIWC lexicon [149].

Machine Learning may be exploited with the same purpose of detecting emotions. The classifiers are frequently dependent on the subset of affect emotions and the set of features extracted from the documents used. The more common approaches Machine-Learning-based rely upon general linguistic features. Bearing this in mind, text representation acts as a fundamental task for any text classification algorithm. Bag of Words (BoW) is a simple model, easy to understand, and implement. It is used in feature extraction from text to be next employed in Machine Learning algorithms. This representation of the text of a document results in a vector of the word counts that appear in it. The occurrence of words is perhaps the most common set of features used in classification tasks where each word count as a feature [150].

Another relevant feature set that is frequently considered is the *N-grams*. It is defined as sequences of words of length n . *N-grams* are used for finding syntactic patterns in text, and it includes important text features as negations. Negation is of paramount importance since it changes the meaning of the sentence completely. This way, it is claimed to be an improvement compared to the BoW.

Vector Space Model is also often used in the information retrieval task where each document is converted in a vector, and each dimension corresponds to a separate term [151]. A term has a non-zero value in a document if it occurs in it. The lexicon is represented as a group of terms whether a document is presented as a vector of each term weight. Since the final goal is to identify emotions, each emotion class has its vector representation. Then, the similarity between vectors may be used to identify the emotions present in a document.

Example of classifiers that were already explored with in the context of emotion recognition from text include Naïve Bayes [152], Support Vector Machines [152], Logistic Regression [153] and Random Forest [153].

3.2.4.2 Deep Learning based Emotion Recognition from Text

During the following section, it will be discussed a representative set of attempts at applying deep learning to the identification of affective states. CNN with a sliding window and subsequent max-pooling was employed to predict aggression through NLP [154]. Despite the acceptable resources in the specific task, the network was projected aiming to only deal with a single dimension (aggression), what does not work as a demonstration of the applicability of it in a multi-class environment.

It was found in the literature several approaches that are introducing RNN for emotion recognition tasks. In order to train these networks in large datasets and since there are not many emotion-labelled datasets, most of these algorithms are making use of data gathered from social networks like Twitter. Hashtags or emoticons are used as the emotion label for data as it is the case of the dataset built by Abdul-Mageed and Ungar [155]. They used a Gated Recurrent

Unit network to identify eight primary emotion dimensions. A different approach used LSTMs to attribute emotions to full conversations in social media means [156]. Due to the specific characteristics and emotions found on this type of data, this approach was unable to generalise their experiments to emotion recognition. Regardless of the chosen deep learning approach, they mostly use distributed word vectors as GloVe [157] (which stands for global vectors) for word representation. These global vectors enable the models to learn word representation by providing trillions of word with low computational costs.

Recently, there are a few comparative studies [158], which analyse the results of Machine Learning and Deep-Learning-based algorithms on extensive Twitter data. It was compared the performance of several CNN and RNN (with GloVe word embeddings) with the performance on bag-of-words indexing models. In general, the neural networks approaches outperformed the Machine Learning ones in the identification of emotions. Figure 3.14 presents the model suggested in [158] where it was employed a combination of CNN and RNN having as final goal emotion classification.

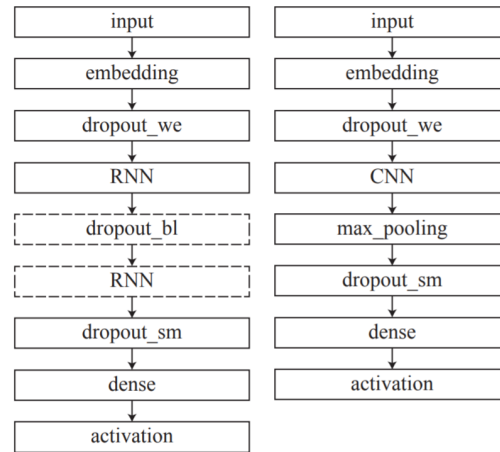


Figure 3.14: Overview of the architecture of RNN (left) and CNN (right) models proposed in [158]

3.2.5 Physiological Data and EEG

Several Affective Computing application focuses on the detection of affects by applying Machine Learning methods to identify patterns in the physiological activity which may be related to the expression of different emotions. The strong traditions of psychophysiology [159] inspired this type of AC research. Psychophysiology is the study of the interrelationships between mind and body, and it relates behaviours and cognitive neuroscience. The main goal is to understand the physiology of behaviour. It notes that one may understand psychological states and processes through observing or measuring physiological changes [159]. The emotions and sentiments are, this way, included in the concept of psychophysiology.

The majority of the measures used to monitor physiological states are non-invasive as they are based on recordings of physiological signals induced by brain, heart, muscles or skin. The data usually gathered includes Electromyogram (EMG), Electrodermal Activity (EDA), Electrocardiogram, Electroencephalography (EEG) or other neuroimaging techniques. These set of

signals offer singular advantages to other modalities already discussed. First, there is an increasing availability of wearable devices that may be used to extract some of these measures. Furthermore, they are not variant against external auditive or visual noises (as illumination), making them quite robust and versatile. These measures are, also, very hard to replicate or mask in order to hide specific emotions. Finally, it allows efficient processing due to a relatively low dimensional structure comparing, for instance, with an image.

Physiological data is being used for automatic emotion estimation even before the definition of Affective Computing by Picard in 1995 [89]. Fridlund and Izard [160] are responsible for one of the earliest works where they applied Linear Discriminant Analysis on the facial EMG activity. This study was a landmark in the area since it showed there is a correlation between personal biological signal and affective states. A multimodal approach based only in physiological data (heart rate, respiration, skin conductance and facial electromyogram) was introduced in [161]. One of the main objectives was to demonstrate that using more extensive data, it is possible to achieve relatively better results. An example of 100 seconds of the response of the diverse physiological signals is demonstrated in Figure 3.15. The two segments correspond to two distinct pretended emotions (anger and grief) where it is visible differences between the signals in general.

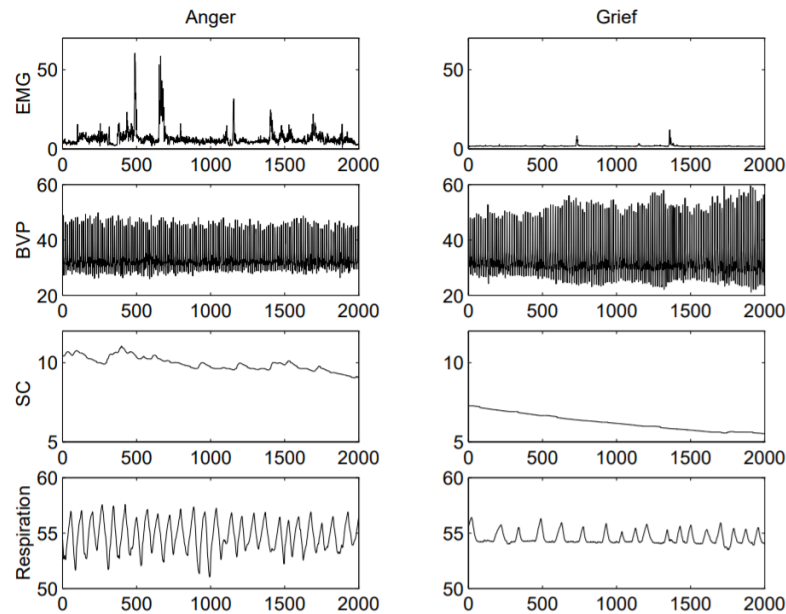


Figure 3.15: Examples of physiological signals measured from a user when expressing anger (left) and grief (right). From top to bottom: electromyogram (microvolts), blood volume pressure (percent reflectance), skin conductivity (microSiemens), and respiration (percent maximum expansion). Figure extracted from [161].

Moreover, other recent works of automatic systems for emotion classification found in the literature comprise the use of SVM classifiers based on EDA signals with wavelet-based features and time-frequency analysis [162], and the introduction of features from EEG, pupillary response, and gaze distance [163].

Recently, the rise of a relevant dataset named AMIGOS [164] - a dataset for multimodal

investigation of affect, personality characteristics and mood on individuals and groups through physiological signals and facial expressions - has unlocked the opportunity of applying Deep Learning to this specific field of AC. The dataset uses the dimensional variables of valence and arousal, as noted in Figure 3.16. The same figure allows the visualisation of the distribution of labels in the dataset under discussion. Bearing this in mind, Comas, Aspandi, and Binefa have suggested a multimodal model based on deep neural networks using the dataset AMIGOS [165]. Electrocardiogram and galvanic skin response of each participant was used in order to estimate valence and arousal and, consequently, emotional-state.

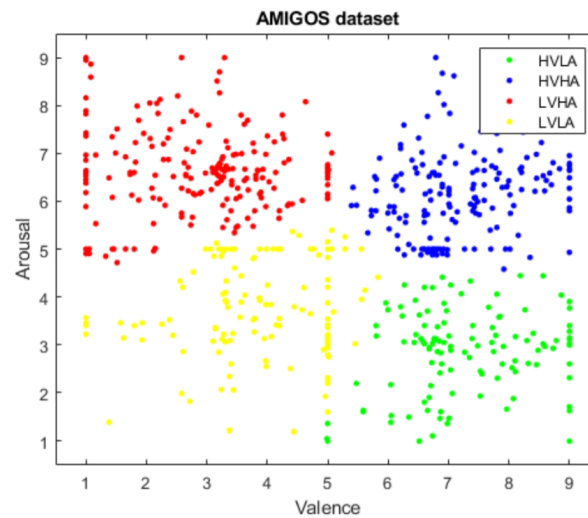


Figure 3.16: Distribution of valence and arousal label from self- assessment labels of AMIGOS dataset (scale from 1 to 9), extracted from [165]. The image is divided into 4 quadrants (HVLA: high-valence / low- arousal; HVHA: high-valence / high-arousal; LVHA: low- valence / high-arousal; LVLA: low-valence / high-arousal).

Regarding EEG, some problematic issues should be taken into account. Setting up an experimental protocol that resembles the real-world may be complex, time-consuming and involve high costs. These factors may hinder the development of practical applications. However, researchers are looking for ways to develop emotion-dependent systems that could be applied in the context of Brain-Computer-Interface.

3.2.6 Multimodal Affect Recognition

In the field of affective computing, multimodal affect analysis is already achieving promising results. After the discussion of the unimodal methods in the previous sections, it is appropriate to discuss the approaches to overcome the multimodal problem. The different modalities that may be used simultaneously must be fused in a proper way. As defined in [91], multimodal fusion is the procedure responsible for combing the data gathered from various modalities for analysis tasks - in this case, affect recognition tasks. It is a task with application not only in the area of Affective Computing as it is also being tested with diverse final goals. The fusion of data from different sources may lead to extra information and consequent increase in accuracy [166]. To date, there are mainly two types of fusion studied, being feature-level fusion or early fusion,

and decision-level fusion or late fusion. These have also been combined in several works as part of a hybrid fusion solution. They may be briefly reviewed:

- **Feature-Level Fusion:** This fusion method joins the features extracted from the diverse modalities in one only general feature vector that is fed into a classifier. The greatest advantage is that it correlates the various features generated at an early stage and the classifier may be optimised having that into account. It may accomplish better results, as the classifier will be able to understand which are the most important features in the all set and give them extra weight during classification. Features from the various modalities can differ broadly in many aspects so, before the fusion process, they should be brought, for instance, into the same temporal format. The limitation of this fusion process may arise from difficulties in time synchronisation. Busso *et al.*'s study [167] is a classic example of feature-level fusion, which concatenated prosodic and facial features to build a feature vector that is then used to train an affect recogniser system.
- **Decision-Level Fusion:** In this fusion process, the features are assessed and classified separately. The fusion occurs at the decision-level, where the output labels of all classifiers are combined into a vector to generate a final decision. The advantage of this method is that the decisions of all classifiers will be easier to compare since they are expected to have the same form. Each modality may use its best adequate classifier or model to learn its features, which is seen as another advantage. Though, since different classifiers are used for the examination task, the learning process may take longer and require higher computing power.
- **Hybrid Multimodal Fusion:** This fusion approach acts as a combination of the feature-level and decision-level methods discussed earlier. This methodology aims to explore the advantages of each fusion methods and overcome their limitations. Wöllmer *et al.* [168] have proposed a hybrid fusion strategy as shown in Figure 3.17. In their method, the visual and aural features are merged and then classified using a Bidirectional LSTM - feature-level fusion. Then, the prediction obtained will be fused with the result of the textual classifier that employs an SVM for the classification task - decision-level fusion.

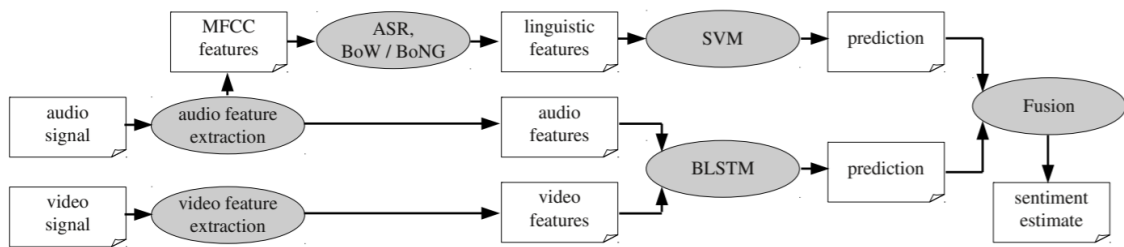


Figure 3.17: System architecture for hybrid fusion of audio-visual and linguistic information, which produces a sentiment analysis in YouTube videos, adapted from [168]

As conclusion, it is still an open issue how to build appropriate fused feature vectors composed of features from diverse modalities with different time scales, metric levels or dynamic structures. Most researchers tend to prefer decision-level fusion due to the simplification of the fusion problem [169]. Hybrid-fusion that combines the benefits of both methodologies may be

the best choice.

3.3 Summary

The conducted review on affective computing was fundamental to demonstrate the various scenarios where the analysis of emotions may be beneficial. The categorisation of the affective states has been approached in diverse perspectives; however, the Hourglass of Emotions, due to the capability of describing the full range of emotional experiences, is growing in popularity and has been applied in several affect recognition studies.

Any AC workflow has to begin with the identification of user's emotional-state and the capture of data. This question has been mostly solved, making use of passive sensors and is strongly dependent on specific settings or signals. Most of the works analysed are targeted at the identification of specific emotions, thereby built around the functionalities of the sensors in those systems, and having real-time requirements. The detection of basic emotions based on image data (facial recognition) proved to be the greatest focus of investigation identified in AC literature review, demanding for that purpose the inclusion of digital cameras in the implemented systems. The literature reviewed lacks generalisation, thus not enabling an application over a rehabilitation program.

Limited by the existing technology, the work aims to open different perspectives regarding affective-state detection. Remarkably, there is a consensus on the potential of data as a tool for a broad spectrum of purposes. Given the large amount of available data, and the fact that the systems reviewed are not flexible to employ this kind of information, a data-driven approach appears as a valuable option to embrace the problem on hands.

Chapter 4

Dataset and Statistical Analysis

As it was concluded in the previous chapter, a data-driven approach is to be followed. Consequently, before any kind of ML development can happen, and to guide the classification tasks, a comprehensive understanding of the data is required. This chapter introduces the dataset, by detailing the patient population and the distributions of the targets, and includes typical (but mandatory) steps of any ML workflow: feature engineering and analysis.

4.1 Dataset Description

Assessing the data distribution and preparing the dataset for the modelling steps are of the utmost importance in any ML pipeline, as the models to build should have in consideration that data under study.

First of all, it could be defined a set of exclusion and inclusion criteria as follows:

- Dataset includes all data generated as long as SWORD Health's product is in use; thus, some of the available data recorded during product testing or from the earliest ages of the system when some essential features had not yet been implemented will be discarded.
- SWORD Health's product is being used not only in Europe but also in Australia and in the US. Given the potentially different psychological boundaries that may exist in the population using the system, possible different perspectives on the scales used in patient's input values, and different distribution of rehabilitation therapies, it will only be used data gathered in Europe. The greater amount of historical information and the existence of more standardised methods justify the use of the data collected in Europe rather than the use of data obtained in the others marketplaces.
- Patients that did not complete more than five rehabilitation sessions will also be discarded. It was assumed that patients who did not meet this imposed condition did not complete the rehabilitation process - it may be the case of users testing the product or early withdrawals; thus, not representative of the class of patients to address in the present work.

Since SWORD Health's database addresses different purposes, and it is used by diverse teams, it is composed of various entities or classes with no usefulness in the problem that aims

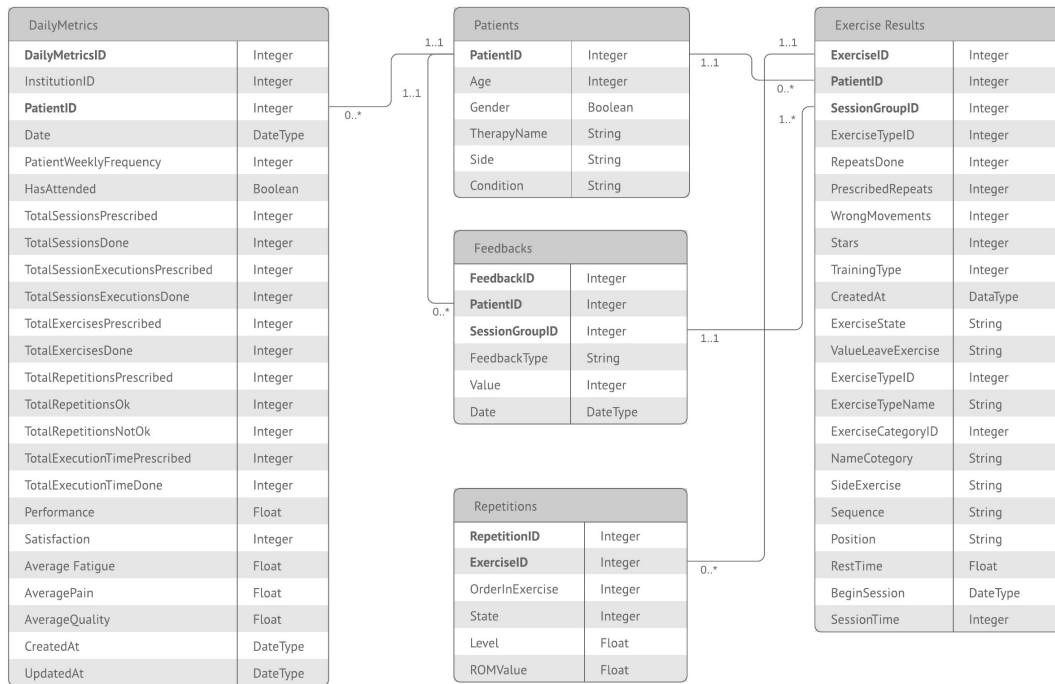


Figure 4.1: Entity-relationship of the dataset used over the dissertation

to be solved. Hence, a specific organisation will be employed throughout this thesis. In this respect, the following Figure 4.1 presents the entity-relationship diagram (ERD) of the dataset utilised. The ERD shows the relationships of the entity sets stored in the database, illustrating the dataset logical structure.

The dataset provided includes five entities, namely: *Patients*, *Exercise Results*, *Feedbacks*, *Daily Metrics* and *Repetitions*. A session does not appear as an individual class, but all information of a session may be easily obtained by grouping the exercises. Providing a brief description of the diverse classes of the database:

- **Patients:** refers to all the patients enrolled in SWORD Health's system.
- **Exercise Results:** comprises data about an exercise, referring the patient who performed the session in which the exercise was part.
- **Repetitions:** An *Exercise Result* combines a set of repetitions; thereby, the entity *Repetitions* combine all the information associated with a specific repeat.
- **Feedbacks:** An object of this class is always created when an input variable is inserted into the system by a patient (including the input type, session associated and the value inserted by the patient);
- **Daily Metrics:** generated every day during the period the patient is attending SWORD Health's rehabilitation program.

In order to assess a possible generalisation of the work to be done, patient population will be detailed. Of the existing patients in the database, and after removing the patients who did not meet the necessary criteria mentioned, it is possible to analyse some aspects related to demography as age and gender. The distribution of the patient population regarding the gender and type of treatment performed with SWORD Health's product is shown in Figure 4.2.

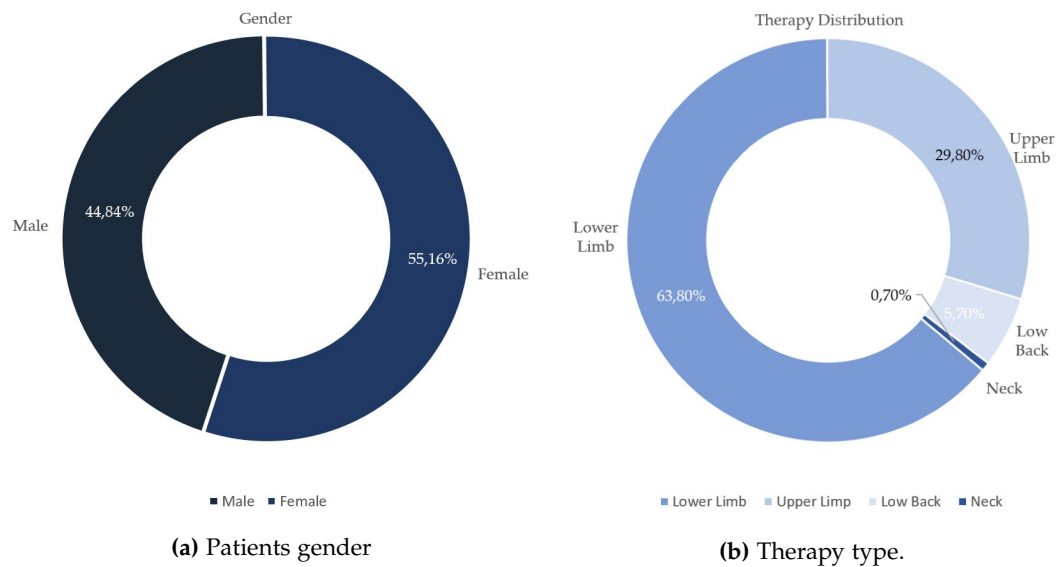


Figure 4.2: Dataset Distribution according to patient gender and type of therapy.

Figure 4.2a shows an equilibrium in the dataset regarding gender, avoiding bias in this criteria when building the models. On the other hand, the physical therapy type is not equally distributed across the dataset as demonstrated in Figure 4.2b. It shows that the number of patients treating low back and neck disorders is substantially lower when compared to therapies of the upper limbs (also comprises shoulder specific therapies) or the lower limbs (including knee and hip-focused). Addressing patients with these injuries is not the main focus of SWORD Health's product in Europe as these procedures are not fully implemented in that market, what explains the later statement. That said, given the short amount of data resulting from patients undergoing neck and low back treatment, data from these patients will be not be employed. It is intended to remove any unfavorable influence that could arise from this latest implementation in the product, although it is known that it may affect the generalisation of the solution for all possible treatments.

Thus, it is fundamental to understand the amount of data to be used henceforth, for the remaining of the analysis and models design. The total values - concerning data registered in Europe and after the exclusions discussed - obtained were the follows:

- **Total number of patients included:** 562 (383 on lower limbs and 179 on upper limbs therapy).
- **Total number of sessions registered:** 21,564.
- **Total number of exercises registered:** 500,299.

Figure 4.3 analyses the population from a chronological perspective: the age of the patients is reported with stratifications by gender and therapy type.

Although a broad range of patient ages is not covered by the data (most patient ages range from 50 to 60), Figure 4.3 shows the balance between the two remaining therapy types - Upper Limbs (UL) and Lower Limbs (LL) - and similar age distribution in both genders. These conclusions are important to assure the diversity in patients profile, important towards a general solution in which any application of the developed system hold for a variety of populations and

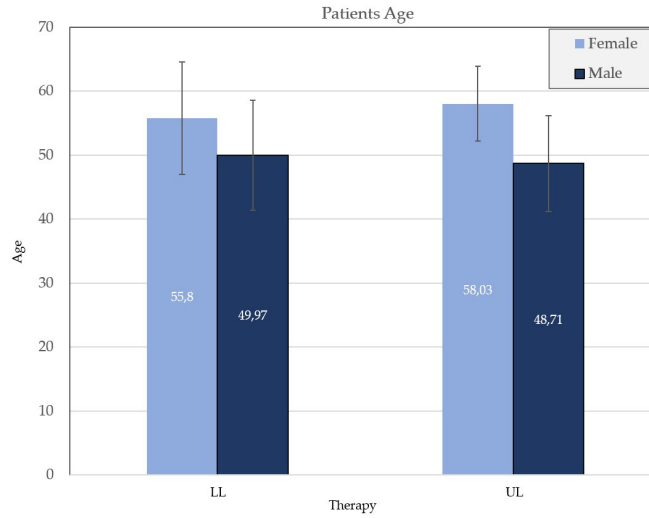


Figure 4.3: Age (mean $\pm$ Standard Deviation) of the patients in the dataset, stratified by gender and therapy type

settings.

4.2 Target Distribution

Several ML algorithms assume that the number of samples in the considered target classes considered is roughly similar. Therefore, assessing the underlying data distribution before the application of any ML or statistical modelling approach is a fundamental step. Bearing in mind that the current work deals with a real-life database, in many situations, the distribution of the examples is skewed - some classes appear much more frequently. In this particular problem, the distribution of patient input variables must be evaluated. Figure 4.4 presents the distribution of all three characteristics (pain, fatigue and satisfaction) in the dataset. The target values are stored in the database as follows: Pain and Fatigue range from 0 to 5 - five as the maximum level of pain/fatigue felt by the patient-, while Satisfaction has values from 0 to 10 - 0 meaning that patient is the least possible satisfied.

Furthermore, it is noted that data is not equally distributed around the possible values. By the evaluation of the histograms, it is observed the most significant disproportionality in the satisfaction values - users tend to be highly satisfied with the service provided. Overall, pain and fatigue values show also an uneven distribution - more frequent at low levels, with level 2 emerging as the most habitual response. Although the dataset includes pain and fatigue outputs at the end of each session (except if the system fails to collect it for any reason), patients satisfaction is not asked to them in all sessions, but periodically - sensibly every ten days. Thus, to enable the inference of all sentiment-related variables on every session and to avoid discarding a considerable number of sessions, the remaining satisfaction values may be interpolated. Thereby, a linear interpolation provides the satisfaction values for the previously unlabeled sessions. This approach was seen as the most appropriate choice in a conceptual point of view (in comparison with, e.g. considering the same value of the next or the previous session) as it is assumed a

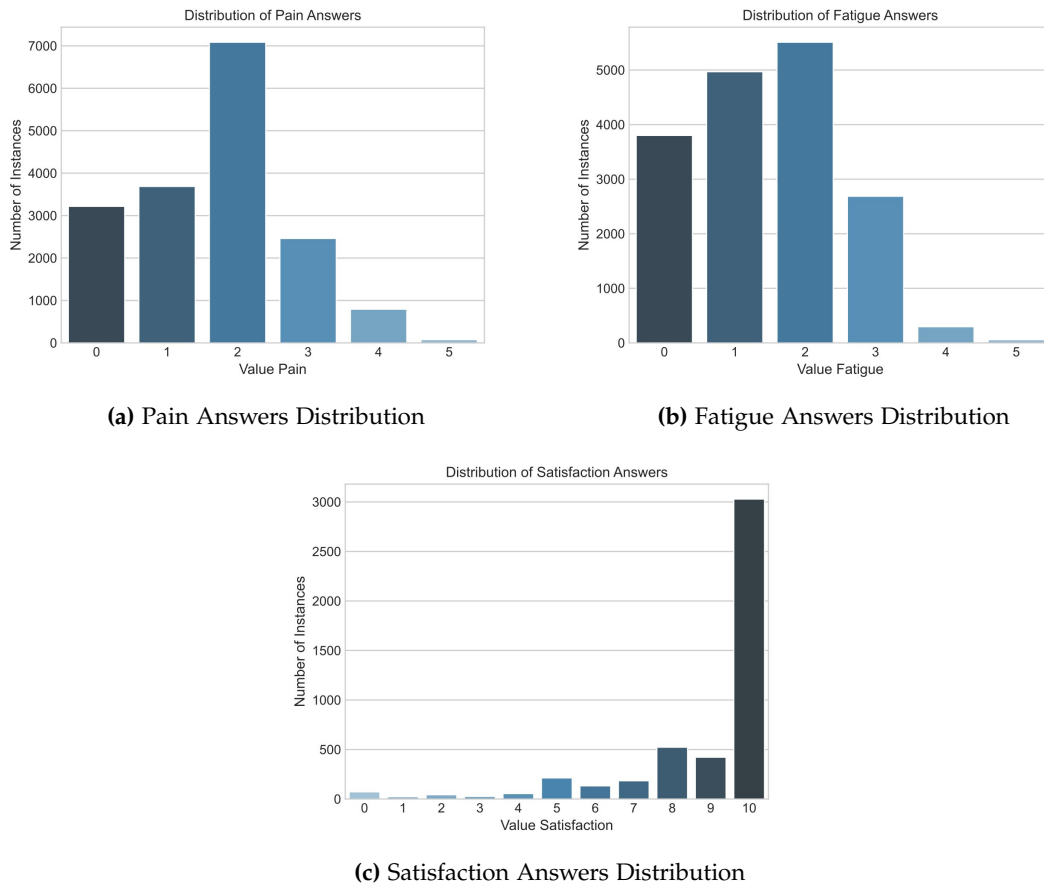


Figure 4.4: Target Distribution concerning SWORD Health's system use in Europe.

linear variation between sessions. The input values are sequential along with the sessions, and a steady increase/decrease in levels of satisfaction are assumed. This interpolation is made on patient and temporal-basis and only intends to label sessions when patients input of previous and later sessions are present; otherwise, those sessions are discarded. Since the learning task is to be assessed as a classification problem, and it aims to predict a finite number of labels, the interpolated values are rounded to the nearest integer. Figure 4.5 shows a representation of the labelling procedure that was applied.

In summary, this representation revealed that the existing data is not equally balanced across the classes. Imbalanced classes are a common problem in ML classification, that said, this analysis shows that the methods must be built around this finding. It raises a difficulty for learning algorithms, as they will be biased towards the majority group. One must develop a sensible system able to overcome such a tendency and take into account that distribution. In these cases, the minority classes are closely related with the most critical issues - low levels of satisfaction or higher levels of pain/fatigue -, hence we require methods to improve its recognition rates.

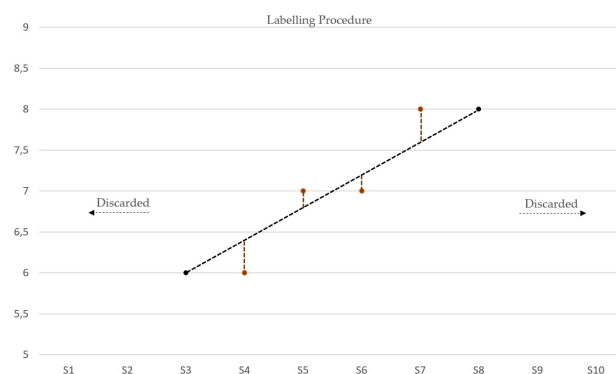


Figure 4.5: Linear Interpolation applied in the labelling procedure of satisfaction values. The example represents an exemplificative situation in which only two satisfaction values were gathered (black points), then, all sessions before and after that period are discarded (Session 1 to 2 and, Session 9 to 10). Red points represent the interpolated values, hereafter, satisfaction values associated with Sessions 4 to 7.

4.3 Feature Engineering

As mentioned in Section 2.4.1, Feature Engineering plays a major role in any classification task. Feature engineering implies the extraction of a subset of informative features and/or the combination of different features into new ones, to be passed afterwards to the ML algorithm. The process must be context-based and requires prior domain knowledge, used to guide this task. In order to address the described problem, for each session will be generated a representative cluster of features. This feature extraction may be done using only the information of the specific session individually or by comparison with the previous sessions of the patient under study.

As we are dealing with a real-life application, a few attributes of some records are expected to be missing in the data set. It may be caused by different reasons, such as a failure in the data collection device or as a consequence of a particular user action. Since the models are not designed to handle a dataset with missing values, they have to be processed in advance. During this feature engineering task and to avoid sample removal, missing data will be explicitly handled for each attribute based on domain knowledge (whether filling in with zeros or making estimates supported by other values of the same attribute in the other samples of the corresponding patient).

Thus, to achieve the objectives proposed, the data was organised by the indicative of a session and information structured accordingly. The whole set of features extracted may be divided into subgroups as follows:

- **Patient Data:** This first subset of features comprises a groups of features that reflect information about the patient. Appearing as categorical data, information about the type of therapy in which the patient is enrolled (whether focused on the upper or lower limbs) and the condition that affects him (the lesion that justified the need for rehabilitation) are included in the set of features. Moreover, the demographic data - age and gender - complete this subset of features.
- **Therapist Input:** Although not directly related to patients performance and actual physical

and emotional state, the prescription made by the therapist has consequences on patients behaviour throughout the session. The total number of exercises, repetitions and total training time prescribed for the session and the rest time given to the user between exercises may be directly extracted from the dataset and will be used as features. In addition, information on the days since the previous session, if the session was performed on a day with more than a session, the number of different exercises and different positions (whether exercises are performed sitting, standing or lying) were also computed and included in this subgroup of features, as they only depend on physiotherapist input.

These features, although extracted as continuous variables, were also collapsed into binary ones by splitting the sample at a designated threshold (based on domain knowledge or feature distribution) [170]. The transformation of these variables was done to make the analysis and the interpretation simpler. Given the uneven distribution of data in certain features - as the total number of sessions in a day (where most samples have value one) -, this dichotomisation is made to distinguish the regular values of the abnormal ones easily. This simplification may lead to loss of information and power of the features when employed in linear models so that both continuous and binary variables are kept and will be further analysed in the following tasks of the ML pipeline.

Concluding, the adaptation of the rehabilitation prescription by the therapist is a vital part of the interaction between the patient and the system; therefore, therapist inputs are considered in the feature extraction process.

- **Patient Performance:** This subset of features comprises the majority of features related to patient performance, that, as already stated, is directly affected by the patient's motivation and physical state. Information on the number of wrong movements, percentage of repeats well performed and the number of times the patient dropped the session without finishing it was extracted and included in this group of attributes. The average time it takes a patient to complete an exercise may also be representative of his physical state, so it was also introduced.

Besides, the average stars (representations of patient performance in each repetition) per repetition achieved in the exercise set is included along with the minimum stars per repeat in a specific exercise during the session. The standard deviation of the patient's performance between all the exercises within the session was computed in order to analyse the consistency of the patient during the program. Since the variation of the physical/emotional state may be gradual and be reflected not only in the current session but also in the previous ones, the mean value of these just-mentioned features in the five previous sessions is included - if there is not any previous session, it will be set that session values; otherwise, the mean value from previous existing sessions is used. On the other hand, a significant decrease in performance when compared to last sessions may be associated with a drastic increment in pain/fatigue levels - a feature comparing both values was included based on this assumption (set to 0 in the first session).

Based on the previous sessions, it was also designed a couple of features that account for the exercises that show lower performance. An exercise will be considered to have a low performance when presents less than 2.5 stars on average (stars values range from 0 to 5) in previous sessions. This way, the total exercises and repeats prescribed and performed

of the mentioned exercise are extracted as features, since the inclusion of this type of exercises may lead to higher levels of pain/fatigue, assuming the relationship between poor performance and feeble physical state. Although these features resort to the information introduced by the therapist, they are present in this section as they are a consequence of the low performance of the patient in specific exercises (hereafter referred as difficult exercises).

Following the logic described in the previous subset of features behind the conversion of continuous variables to binary ones, this transformation has also been made in some of the recent-mentioned features, such as the number of difficult exercises or the number of exercises dropped within a section. Again, the inclusion of the features as continuous or categorical will be further analysed in the following sections.

- **Exercise-level comparison:** This set of features comprises features not only directly dependent on the physical therapist prescriptions but also on patient performance. The performance of the patient is not analysed only based on information of the current session nor by comparing the information of all the exercises with the previous sessions; however, therapists' prescription is considered in this comparison. For each session, only information resulting from the set of exercises that were already present in the last prescription is gathered. The goal is to avoid bias introduced by the removal or addition of different exercises and to perform a direct comparison using the same set of exercises. Thus, similar features are extracted, but always subtracting information of the previous session into the current one. The total number of repeats/series of those exercises prescribed, wrong movements noticed, performance and time spent per repetition, minimum performance value in an exercise are examples of features extracted with this comparative logic.
- **Daily Metrics:** An object of the Daily Metrics class is generated all days during the period in which a patient is enrolled in the system, with total values resulting from the sessions performed by the patient on the corresponding day. Since it may contain information about more than one session, to avoid bias that could be created from days with double sessions, only relative or percentage features are employed: e.g. percentage of exercises/repetitions and execution time performed in terms of the prescribed. Thus, only adherence metrics are extracted in this set of features that is also dependent on patients behaviour. The attendance of the patient to the program prescribed may act as relevant information, as low engagement rates may be related to higher levels of pain/fatigue or dissatisfaction with the therapy plan.

All the just discussed subsets of features had been joined to be further analysed and selected according to the designed models. A total number of 56 features has been extracted. A complete list may be found in Appendix [A.1](#).

4.4 Feature Analysis

After the discussed data preparation and feature extraction, it is possible to explore the features and draw some conclusions, with a view to future use in classification tasks. The possible application of these features in the prediction of pain/fatigue and satisfaction levels at the end

of the session may be evaluated. In this phase, the levels of the variables to predict are treated as discrete values from 0 to 5, if pain and fatigue and from 0 to 10, if predicting satisfaction values. This exploration of the features extracted is a critical step in a typical data science project. This investigation may help to get some insights on the data available and build a broad picture of significant trends. The process of excluding extracted attributes that may be insignificant should also begin.

To get a general picture of the potential of features related with patient performance, Figure 4.6 shows the possible application of two different features in the prediction of fatigue and pain levels, respectively, at the end of the session. Figure 4.6a takes advantage of the boxplot representation to show the distribution of the total wrong movements (number of incorrectly performed repetitions) during a session in terms of fatigue values. Thereby, it is seen an increase in *Wrong Movements* variable when higher values of fatigue are noticed. Then, Figure 4.6b denotes the correlation between the average stars per repetition accomplished during the session and pain levels. The relation may also be intuitively explained as higher values of pain may lead to lower performance values. This representative analysis intends to confirm the ability of this particular subset of features that are expectable to have remarkable importance when training an ML model.

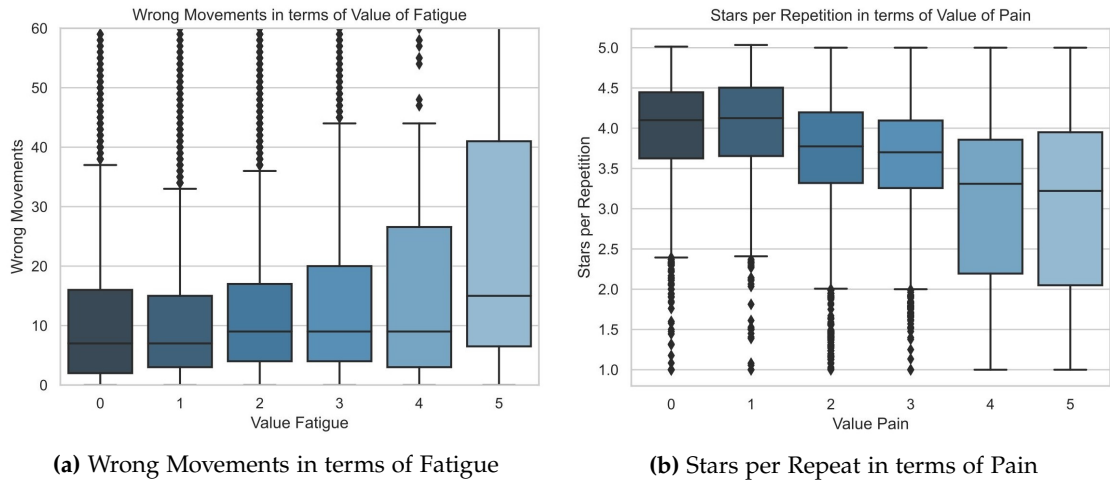


Figure 4.6: Feature Distribution Analysis in terms of Fatigue and Pain values, using boxplot representation.

To support the transformation made on the features that have been converted to binary, Figure 4.7 takes as an example the feature that represents the total number of repetitions of exercises which presented performance lower than 2.5 stars on average - named *difficult exercises*.

Figure 4.7a shows the uneven distribution of the *Repetitions of difficult exercises* feature, where a considerable share of values is equal to zero. Consequently, making the mentioned modification allows distinguishing the most common value from the unusual ones simply. Although some information about the absolute total number is lost, the simple fact of having (or not) repetitions of difficult exercises carry significant information, as demonstrated in Figure 4.7b and Figure 4.7c where for higher values of pain and fatigue the ratio of sessions with repetitions of difficult exercises increase.

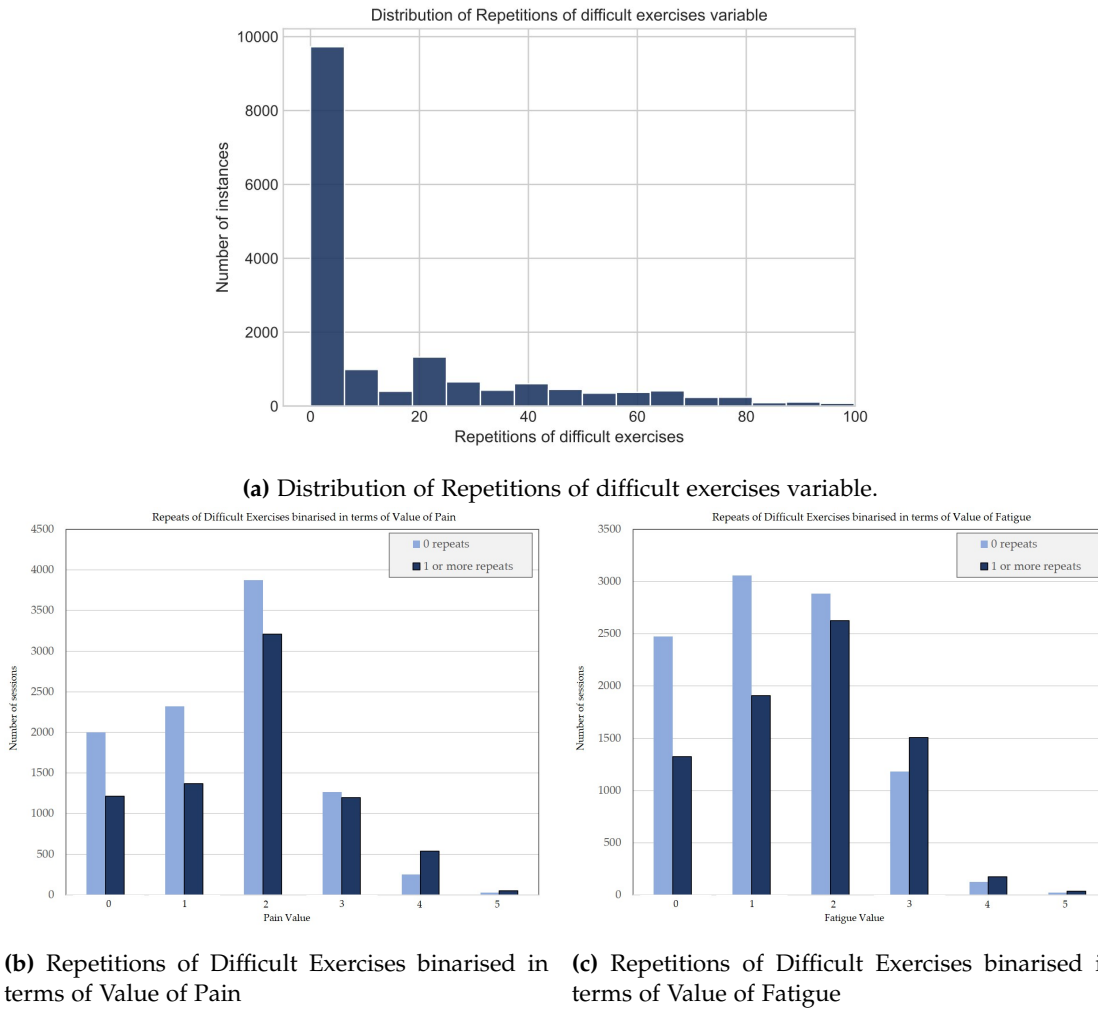


Figure 4.7: Example of the Transformation of continuous variables into binary data.

The relation between pairs of features should also be analysed. Data and feature correlation are deemed a fundamental step in this initial feature selection stage, included in the Features Exploration. It is possible to get some insights, such as whether one or more attributes depend on another attribute or are the cause for another. Analysing the correlation may be useful to indicate the presence of this causal relationship. Features with high correlation are more linearly dependent and therefore have a similar effect on the target variable. That way, it is important to minimise the redundancy between the features when employing feature selection methods and to reduce data dimensionality. Whenever a couple of features contain resembling information, removing one of the two or combining them act as a strong option. Performing this selection before modelling the data can also be useful to reduce overfitting problems - decreasing possible harmful bias, as less redundant data means less opportunity to make decisions based on noise -, and to reduce training time - fewer data points may reduce algorithm complexity.

Pearson correlation coefficient was used to indicate the presence of a causal relationship between continuous features. *Pearson's R* appears as the most usually employed correlation coefficient, calculated by a parametric test which requires normally distributed data [171]. Re-

garding the strength of the relationship, the coefficient value ranges from +1 (perfect positive association) to -1 (strong negative association), then values close to zero imply weak or any linear relationship. Equation 4.1 provides the formula used to compute the referred coefficient for a data sample (r_{xy}), where n is the sample size, x_i, y_i the individual samples indexed with i and $\bar{x}$ the sample mean (analogously for $\bar{y}$).

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.1)$$

However, as it might be noticed from the equation above, *Pearson's R* is not defined when the data is categorical. Encoding the feature could be an option to handle it, but it can result in a higher-dimensional space and make the analysis more complex. Thus, when both sequences of data are nominal-valued, Cramer's V-statistic was used as the alternative measure of correlation based on a matching in a bipartite graph [172], giving a value between 0 and 1 (maximum correlation). Proposed by Cramer in 1946 [173], it measures the correlation between 2 data sequences (x and y) where n is the total of observations, p the minimum number of columns/rows and χ derived from Pearson's chi-squared test:

$$corr_{xy} = \sqrt{\frac{\chi_{xy}^2}{n(p-1)}} \quad (4.2)$$

Lastly, in order to measure the association between a continuous and a categorical feature, it was used the Correlation Ratio, the weighted variance of the mean of each category divided by the variance of all samples. As the Cramer's V-statistic just presented, the output ranges from 0 (no dispersion among the means of the different categories) to 1 (no dispersion within the respective categories). Suppose each observation is y_{xi} where x points out the value of the categorical variable and i is the label of the particular observation, and n_x the number of observations in category x . Setting $\bar{y}_x$ as the mean of the continuous variable in category x and $\bar{y}$ the mean of the whole population, correlation ratio (η) is computed using the following equation:

$$\eta^2 = \frac{\sum_x n_x (\bar{y}_x - \bar{y})^2}{\sum_{x,i} (y_{xi} - \bar{y})^2} \quad (4.3)$$

Therefore, Figure 4.8 displays the correlation matrix of the all set of features where the measure is used accordingly to the variables' type.

It is possible to denote high correlation between different pairs of features, mainly in features of the same subgroup that form blocks near the diagonal. This way, in order to avoid keeping features with a similar effect in the prediction of the dependent variables, one of the two highly correlated features can be dropped. In the next tasks of the problem, it may be assumed that correlations higher than a specific threshold are not giving new information and, therefore, only keep one of the involved features. At this stage all the available features are maintained, referencing this analysis when intended in the chapters that follow.

Furthermore, Figure 4.9 provides an analysis of the correlation of two different pairs of variables as an illustration. Figure 4.9a shows an example of two correlated features (*Total Prescribed Series* and *Total Prescribed Repeats*) that carry similar information and may result into

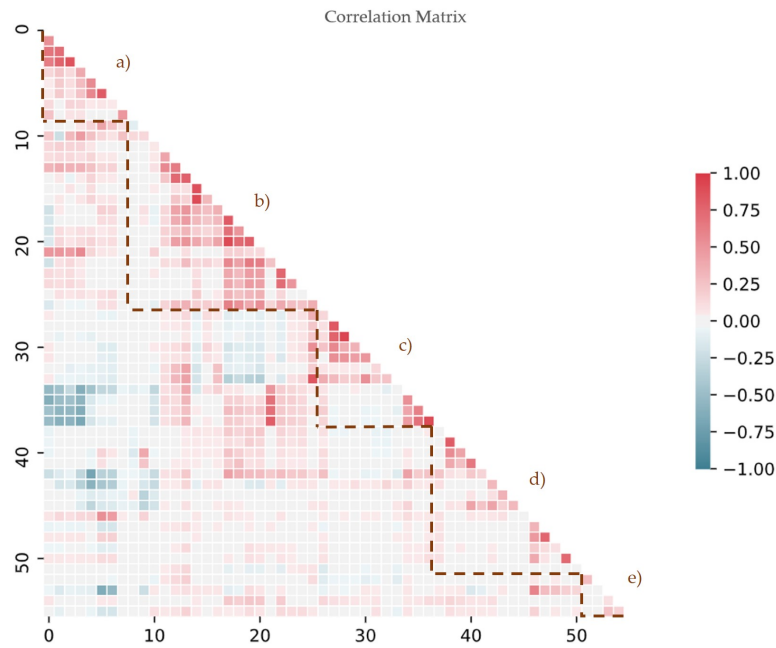
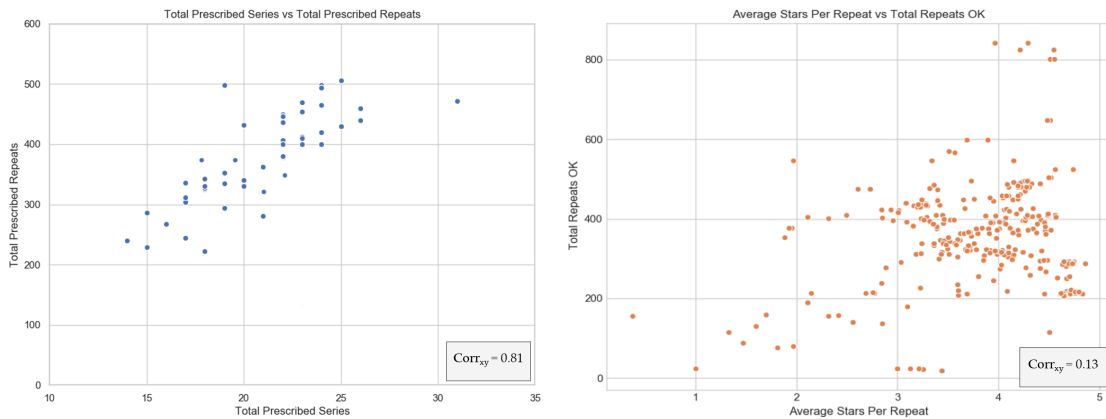


Figure 4.8: Correlation Matrix of feature set. Dashed lines highlight the following different subgroups in the feature set: a) *Daily Metrics*; b) *Patient Performance*; c) *Exercise-level comparison*; d) *Therapist Input*; e) *Patient Data*.

the removal of one of them. Although the number of repetitions per exercise is not constant, the correlation of these features was already expected. In opposition, Figure 4.9b shows an example of a uncorrelated pair of features: *Average Stars per Repeat* and *Total Repetitions Completed*. The figure ensures that this pair of features hold dissimilar information that may be relevant later.



(a) Example of a correlated features pair.

(b) Example of a non-correlated features pair.

Figure 4.9: Distribution scatter of a pair of attributes to evaluate feature correlation.

Although this analysis is model-independent, it may be used to guarantee that similar features are not passing redundant information to the model.

4.5 Summary

Through the dataset exploration and characterisation conducted in this chapter, the richness of the dataset was assured. The chapter starts with the description of the employed dataset, presenting the complete scheme of the data used throughout the dissertation. Patient population is detailed, allowing to conclude about the generalisation of the work and, defining different exclusion criteria what results in a total number of 562 patients data to use hereafter.

The analysis of the target distribution showed that the various sentiment-related variables are not equally distributed around the possible values, asking for the application of imbalance learning techniques in the following chapters. Moreover, the essential step of feature engineering allowed the author to extract the relevant features that are to be applied in the ML pipeline. The division in subgroups aims to analyse the importance of the different features types in the classification task. Furthermore and finishing the chapter, some of the extracted features are analysed in order to conclude about its future use in the supervised learning task. Correlation matrix of feature set denoted high correlation in features of the same subgroup - a finding that should be considered in the subsequent chapters.

To sum up, valuable insights from this dataset exploration were acquired, providing a groundwork where the models will be constructed. The feature set power should be further examined in the development of suitable classification models with proper regularisation, overfitting countermeasures and fair assessments of generalisation power. In the following chapter, the development of such models is addressed.

Chapter 5

Sentiment-Related Variables Classification

The main objective of this experiment is to use the behavioural features extracted to predict, at the end of the session, the different variables of patient input (whether pain, fatigue or satisfaction). The behavioural features comprise all information based on patients' performance during a session, or on the therapist's prescription. A detailed description of the features generated was provided in Section 4.3.

In this task, data is evaluated as a cluster of features representing only a particular session. The output label to be used corresponds to the value inserted by the patient at the end of the session in the case of pain and fatigue variables. The satisfaction values of the sessions whose information was not gathered were labelled through the procedure described in Section 4.2. Thus, the learning task will be assessed as a classification problem, where it is intended to predict a finite number of values. Given the imbalance of the target variables (discussed in Section 4.2) and the small number of samples of the minority classes, these former categories will be combined throughout this chapter without harming the clinical relevance. In terms of pain and fatigue predictions, the highest values (3 to 5) - considered by SWORD Health's clinical professionals as the group of patients that is of relevance to identify - are merged and will be treated as one class hereafter. Moreover, the problem will be assessed in two distinct ways:

- Binary classification task: majority classes are merged. The purpose will be differentiating the highest values of pain/fatigue (patients that should be followed by the health professionals) from the merged low levels (patients who do not require clinical intervention).
- Multi-class classification task: it is also aimed to distinguish the lowest values of pain/fatigue between them, intending to classify four different classes.

Regarding the satisfaction values, it is not of interest to distinguish between all the existing values. As highlighted in Section 2.3.4.1, satisfaction is measured resorting to the NPS as a simple survey presented to the client - in this case, a rehabilitation patient. This measure clusters patients into three partitions: *promoters* - patients that score 9 or 10; patients answering 7 or 8 are named *passive*; and, patients who respond from 0 to 6 are considered *detractors*. It is of clinical interest to keep these divisions; thus, it will be the basis of the multi-class classification

henceforth when discussing satisfaction predictions. For the binary classification, *promoters* and *passive* are joined and evaluated against *detractors*.

Furthermore, sessions may not only be evaluated separately but also studied in a sequential matter; therefore, algorithms that take that into account may also be designed to solve the presented problem. Although the temporal-relation is already present in some extracted features, sequential patterns learning are to be exploited throughout this chapter.

5.1 Experiments Design

Several traits are common to all the experiments in the following chapter. The entire set of trials used a fixed training and testing set, built from the 562 patients in the database described along Chapter 4. Dataset split was random but ensured the subsequent conditions:

- An 80/20 training/testing division, stratified by the percentage of samples for each class. A similar target distribution in both generated groups of patients is this way assured.
- Since rehabilitation sessions are to be addressed individually, a similar average of sessions per patient should be placed in the training and testing set. This had resulted in an average of 36 samples per patient in both sets.

To provide an unbiased evaluation of the model when optimising and tuning the hyperparameters, all experiments will use validation sets through K-Fold cross-validation using for that purpose the training data. To accurately support the obtained results, validation sets are also generated using a Group and stratified K-Fold. This procedure enables whole data from a patient to be present on the development set or on the validation one, but never on both. The distribution of classes across all folds is also preserved. The number of folds employed during all the experiments was 8. The results obtained regarding the application of the K-Fold cross-validation on training data will be referred as validation results for the remainder of the chapter.

In order to assess the performance of the models, the main metric used was the F1-score (harmonic average of precision and recall - values between 0 and 1 where 1 is the maximum value). However, precision (a measure of accuracy for each class) and recall (a measure of sensitivity) may also be mentioned if justified. Equation 5.1 presents the formulas used to compute the mentioned metrics:

$$\begin{aligned}
 Precision &= \frac{True\ Positive}{True\ Positive + False\ Positive} \\
 Recall &= \frac{True\ Positive}{True\ Positive + False\ Negative} \\
 F1Score &= 2 \times \frac{Precision \times Recall}{Precision + Recall}
 \end{aligned} \tag{5.1}$$

For multi-class examples, the average F1 score of all classes is pointed out in an unweighted form - arithmetic mean of all classes. The values of the positive class are displayed when the problem is transformed into a binary classification - in order to allow a fair comparison. This metric was also selected as the optimisation goal given the unbalanced distribution of the classes as it disfavours models inclined towards the most prevalent classes.

All the experiments were conducted in Python, using packages such as scikit-learn ³(ML models and other useful functions), Keras ⁴with Tensorflow ⁵backend (Neural Networks training) and HyperOpt ⁶(for Bayesian Hyperparameter Optimizattion).

5.2 Experiments Workflow

In order to produce the best performing models, the features extracted (and described along Section 4.3) are not in the ideal format. There are a few transformational steps, such as encoding categorical variables or feature scaling that need to be carried out.

Regarding the categorical variables (in this particular problem, only the therapy type and physical condition), ML algorithms accept only numerical inputs; hence, it is necessary to encode these categorical variables. There is a wide variety of encoding techniques that may be used in classification tasks as *One-Hot Encoding*, *Ordinal*, *Helmert*, *Frequency* and *Target*, among others [174].

In *One-Hot Encoding*, each category is mapped to a vector that contains 1 or 0, denoting the presence or absence of the dichotomous binary variable. This way, it transforms a single variable with n observations and d distinct values into d binary features (or $d-1$ if first column dropped) with n observations each [175]. This methodology is the one applied to the *Therapy type* feature, as it only results in one feature, considering the only two possible values.

Even though *One-Hot Encoding* has, generally, favourable performance, it can lead to a huge increase in the dimensionality of the feature representations for high-cardinality variables. This would induce sparsity into the dataset, which is undesirable, particularly if working with tree-based ensembles - continuous variables will be given more importance. This feature expansion may also create serious memory issues in a dataset with various high cardinality features.

As it is intended to explore tree-based algorithms (like Random Forest or Gradient Boosting), the previous encoder will not be applied to the other mentioned categorical feature - *Condition*. On the other hand, *Target Encoders* do not affect the volume of data and appears as a solution. This encoder, as suggested by the name, uses information from the dependent variable in the encoding process. The mean value of the target variable on the training data is used for each category in the feature under analysis. This encoding approach reveals the relation between identical categories, but the bonds are restricted within the categories and the target itself. In order to avoid any possible overfitting, different regularisation approaches are applied in most circumstances. In the *Leave One Out* - the sub-type of *Target Encoders* applied in the present feature - the sample under consideration is left out when computing the average value of each category.

Besides, the feature set contains attributes that highly vary in magnitudes, units and range. Feature scaling is, then, essential for the application of certain ML algorithms that compute distances between data or assume normality in data. Standardisation is a specific method of scaling where the values are centred around the mean with a unit standard deviation. The

³Scikit-Learn: scikit-learn.org. Accessed on 23-07-2020

⁴Keras: keras.io. Accessed on 23-07-2020

⁵Tensorflow: tensorflow.org. Accessed on 23-07-2020

⁶HyperOpt: hyperopt.github.io/hyperopt/. Accessed on 23-07-2020

formula of the standardisation is the follows where μ is the mean of the feature values and σ its standard deviation:

$$X' = \frac{X - \mu}{\sigma} \quad (5.2)$$

Thus, the standardisation will be applied to the extracted features in order to avoid any bias that could be created by their scales. This process is applied regardless of the classifier to employ to assure that all methods receive the same features values. Although it is intended to apply algorithms that do not require normalisation/scaling - as tree-based algorithms that rely on rules -, that does not constitute a problem as they are not affected by any monotonic transformations of variables as the standardisation under discussion (also, with favourable impact on computational costs).

The referred transformations (categorical variables encoding and feature scaling) should be applied during the workflow when training the model and again on any new data to predict on. This way, during this work, these two procedures will be implemented in a pipeline, responsible for applying the transformers and fitting the final estimator used. In order to avoid data leakage (information from outside the training dataset used to create the model) during the development stage and to create overly optimistic performance, these transformations will always be performed within the cross-validation K-Folds. The data re-scaling process and the target encoding will not be aware of the full distribution of the training data, but will only calculate its parameters in each iteration and use those to prepare the data on the held-out test fold.

In the following sections, it will be tested a series of models (employing the features extracted in Section 4.3 and with no application of data-balancing approaches) in order to obtain results used as baseline and for comparison with future techniques (imbalanced, ordinal or sequential learning). The author decided to present the diverse results also to allow to compare the models among themselves as well as study different strategies concerning feature selection (FS).

Tree-based algorithms are intuitive and straightforward methods, used worldwide in classification tasks. They are capable of handling a large amount of data (requiring less data cleaning compared to some other modelling techniques) with high dimensionality, and map non-linear relationships quite well, justifying their applicability in the current work. They may also be used as dimensionality reduction methods as tree-based models output feature importance, an advantage to explore. A pair of ensemble methods are the topic of Section 5.3 - Random Forest (RF) and Gradient-Boosted Trees (GBT). Support Vector Machines are also an extremely popular algorithm, able to capture complex relationships between data points and efficiently perform a non-linear classification, mapping the inputs into high-dimensional feature spaces. During this work, it is combined with a pair of feature reduction techniques (selection and transformation) as it is intended to compare their performances. Finally, the last baseline model aims to evaluate the efficiency of a simple Deep Learning approach and the effect of not performing any previous feature transformation or selection. The multilayer perceptron results will be later used with comparative purposes - when combined with the aforementioned learning techniques.

5.3 Classification with Tree-based Ensembles

The sentiment-related variables identification in this section is based on classification through the efficient techniques of ensemble methods. Ensemble methods combine several predictive models according to a given simpler algorithm, named the base classifier, in order to produce one more powerful learning algorithm. The prediction of a new observation is then obtained by voting the decisions of the set of models that may (or not) have different weights. The most popular techniques when constructing these ensembles are bootstrap aggregation, *bagging*, and the Adaboost family of algorithms - *boosting*. Both of them invoke multiple times the same base model with different training sets [176].

In the current work, the base learner corresponds to a decision tree. Decision trees are a type of supervised classifiers that use tree structured graphs or models to support the decisions. It presents a flowchart-like framework that divides the feature space using sequential threshold-based splits on single attributes - this way, not demanding feature scaling/standardisation as discussed in Section 5.2. The objective of any split is to reduce the impurity of each node - a tree node is considered *pure* if only a single class is represented by all training samples it contains. The decision of evolving the tree structure by consecutive splits is based on a greedy search across the feature space, selecting the value that generates the most pure nodes. With regard to the impurity measures, Gini index and entropy arise as the most common [177]. An example of a decision-tree ensemble is shown in Figure 5.1.

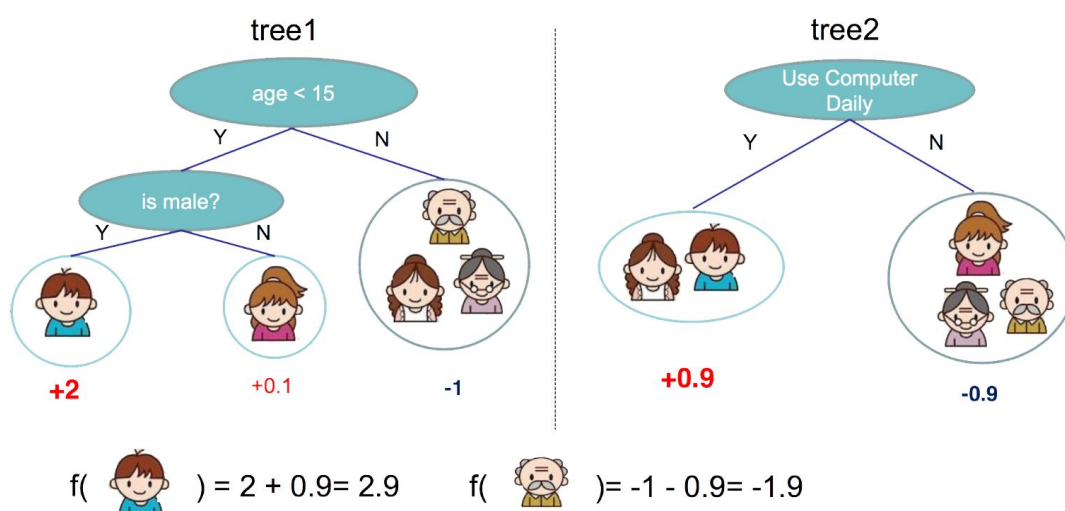


Figure 5.1: Example of an ensemble of two simple decision trees depth of 1 with both categorical and numerical features. The final prediction for each sample is the combination of the predictions from both trees which had evaluated a different set of features. Adapted from [178]

Bearing this in mind, two particular ensemble decision trees are further discussed and implemented in the designed ML pipeline.

5.3.1 Random Forest

Aiming to reduce the high variance of a decision tree, Bagging (Bootstrap Aggregation) was developed by Breiman in 1996 [179] as an ensemble approach that combines a series of independent models in which each learner is trained using a data subset taken from the whole training set as replacement. An equal number of instances as in the original dataset is generally included in each data subset in order to guarantee a sufficient amount of data per inducer. All models are trained independently, enabling the reduction of model's variance and the sensitivity to the data on which an individual tree is trained. The process of selecting rows randomly from the training set with replacement was defined as bootstrap.

Random Forest emerged as an extension over bagging and was also introduced by Breiman in 2001 [180]. In addition to the bootstrap method where a random subset of data is taken, RF injects a second randomisation process: rather than using all features, the inducer also takes a random selection of features in the learning process [181]. At each split when growing a tree, instead of selecting the best attribute every time, only a random subset of the features are inspected, avoiding high correlation between the bagged trees. Finally, a large number of trees may be generated, and the unweighted vote is used to provide the ultimate decision of the random forest algorithm.

RF may be considered an easy to tune method when compared to other methods as Gradient Boosted Trees (to discuss next section) that require a more careful training process. Concerning the number of trees, several theoretical studies and empirical results have shown that beyond a certain number of trees, adding more to the RF will not improve the results [182]. On the other hand, RF do not present substantial risk of overfitting and it may be used a large number of unpruned decision trees.

Moreover, another benefit of using RF for predicting models is the ability to handle features set with a large number of predictor attributes; even though, a lower number of features may improve efficiency. Although not all set of features is seen by each tree due to the random sampling of the subset of attributes, Random Forest allows deriving the importance of each feature in the model. During the construction of trees, it is possible to calculate the reduction in the impurity caused by a feature. This reduction is directly associated with the feature importance and may be averaged across trees to get to know the final importance of the variable in the RF. Intuitively, the top splits are associated with more important features since they lead to larger information gains [183].

Random Forest may act as a straightforward method for FS under the category of *Embedded* methods, requiring no previous attribute choice methodology. There are, however, some limitations to the use of RF with no previous selection, as correlated features will be given similar importance but reduced in absolute value compared to an equivalent tree without correlated counterparts. In the standpoint of the model, if the dataset has several correlated features, when selecting the best split, any of them can be used with no real preference of one over the others. Although it may have no influence in models results, it could lead to incorrect interpretations of features importances - lower or different importance may be found in features with high correlation. The previous correlation matrix shown in Section 4.4 was, this way, used to overcome this RF limitation. By dropping one the correlated features (threshold set at 0.75), it is assured that

redundant data is not given to the classifier. Unlike the methods of Section 5.4, in this section, it is not intended to apply any other feature reduction method (either selection or transformation), so it was decided to remove the correlated features *a priori* (which does not happen in the other methods for the opposite reasons). The goals include reducing feature dimensionality (and, consequently reduce computational costs) and allowing proper evaluation of feature importance.

For this classifier, a Bayesian hyperparameter optimisation was done through a Stratified Group 8-Fold, as noted in Section 5.1. Bayesian optimisation provides a principled technique based on Bayes Theorem usually applied in hyperparameters tuning of a given model on a validation dataset [184]. It had appeared as a smarter alternative to grid or random search by building a probability model of an objective function. It considers the past evaluations, avoiding wasting time evaluating inadequate parameters. In the different classifiers training, the improvement of F1-score over the hyperparameter setting is used as acquisition function. The F1-score with macro average (unweighted average form) was used for evaluating the best model for multi-class classification, and F1 of the positive class for binary classification. Optimised hyperparameters and their tested values are detailed in the following table, Table 5.1. Since Random Forest may deal with unpruned trees without significant risk of overfitting (since it is mitigated by the intrinsic aspects of the algorithms already discussed as bootstraps and column sample), both *max depth* and *minimum samples to split* parameters range include the default values in which trees are not pruned (*None* and 2, respectively). The number of estimators was set to 250, a number considered to be sufficiently high to benefit from RF power keeping computation costs at an acceptable level. In the case of *max features* and *min samples split*, tested values are defined as discrete uniform distributions, this is, each value in that range is equally likely.

Table 5.1: Optimised hyperparameters in the RF sentiment-related variables classifier, with the considered values and definition.

Hyperparameter	Definition	Values	Distribution
Criterion	The function to measure the quality of a split.	{Gini Index, Entropy}	-
Max Depth	The maximum depth of each tree comprised in the model.	{2, 6, 10, None}	-
Min Samples Split	The minimum number of samples required to split an internal node.	[2, 10]	Discrete uniform Step: 1
Max Features	Number of features to consider in each split.	[0.3, 1]	Uniform

Thus, the best hyperparameter-optimised model was re-trained on the entire training set and assessed on the test set, for all the tree sentiment-related variables: pain, fatigue and satisfaction. In the following sub-section, the results are presented while the comparison between models will be topic of Section 5.9.

5.3.1.1 Results & Discussion

Both the results obtained for multi-class classification and for binary classification are included in Figure 5.2. The performance comprises the test results on data from the patients that had been left out from model design and features analysis present in Chapter 4 - 20% of the total number of samples. The validation results of the 8-fold Cross-Validation are also reported in Figure 5.2 as it allows discussing the generalisation power of the model and the existence (or not) of overfitting.

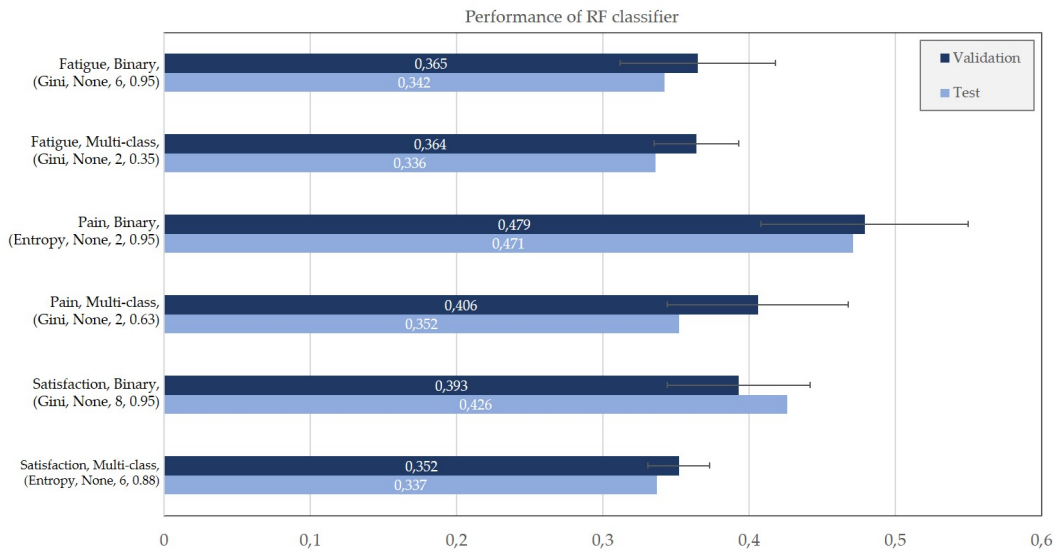


Figure 5.2: Performance of RF classifier on sentiment-related variables prediction. X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (criterion, max_depth, min_samples_split, max_features)*.

The first main takeaway from this model results is the unsatisfactory performance for all variables that are being predicted. Generally, performance decays for multi-class classification, with the binary classification of pain values taking the top performance spot.

From the confusion matrices analysis (see Appendix A.2.1), it may be concluded that for multi-class classification, the model is not accurately differentiating neither between the low levels of the variables nor the principal class of interest, as the predictions are biased towards the majority class. For the binary classification problem, as a severe class imbalance is noted, the threshold used to map these probabilities to class labels is adjusted by locating the value that maximises the F1-score. Thus, the best results using this classifier appear for the binary classification of pain and fatigue. In opposition, satisfaction prediction maintains the poor results, maybe also due the strong class imbalance.

Moreover, other two interesting observations arise from these results: firstly, testing score is similar to the validation score (even higher in satisfaction binary classification) and always within the range defined by the standard deviation (SD). It may allow concluding that models are generalising relatively well to unseen data (even if with reasonably low performance) and

overfitting seems to have been avoided. The second one is that models tend to elect practically unpruned trees. Nodes are allowed to expand until all leaves are pure - no restriction of depth - or contain less than *min samples split* samples, another parameter that seems to have been minimised for most models.

Concerning the remainder parameters, criteria and fraction of the number of features to be considered appear to have no influence on the selection of the the optimised models - both criteria were selected, and practically the full range of *max features* is represented.

Of the features employed in the model (after the removal of the redundant ones), it is possible to analyse directly their importance. Figure 5.3 presents the mentioned evaluation taking the binary pain classification model as an example.

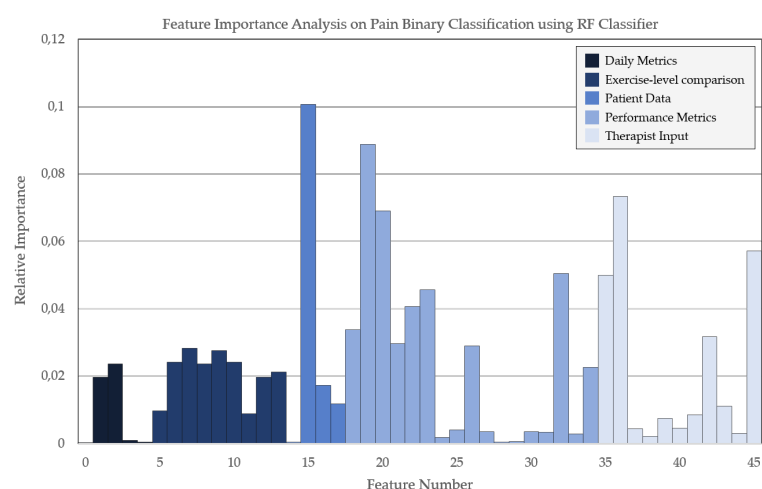


Figure 5.3: Feature importance evaluation of a RF classifier, making use of the pain binary model. Legend unveils the group of each feature.

Not surprisingly, the *Performance Metric's* features comprise the majority of the features with higher importance. As previously explored in Section 1.1 or 2.3.1, patients' performance in the rehabilitation program may be considerably impacted by their emotional and physical state; therefore, the significance of this subset only confirms this expectation. All sets of features had contributed to the model, justifying their inclusion into the classification algorithms. Another evidence of this analysis is the notable importance of patient's age - feature with the highest value included in the *Patient Data* subset. This analysis was performed with the other tree-based algorithm (next subsection) in which the results obtained were similar. Some features demonstrated to have no importance; thus, providing a groundwork for the application of features selection or transformation methods (topics of Section 5.4).

5.3.2 Gradient Boosted Trees

Based on the definition of boosting where a number of diverse weak learners are blended to achieve a better performance than their individual predictions, it had emerged boosting ensembles. Boosting conjectures the idea of filtering observations, putting more weight on the misclassified instances and sequentially adding new learners to the model that focus their training on

the formerly unrecognised patterns [176].

The first great success application of boosting was AdaBoost where each weak learner were the discussed Decision Trees with a single split, known as decision stumps. As practice in ensemble methods, predictions are made by majority vote, in this case, weighted by learners' particular accuracy. A generalisation of AdaBoost was made by Friedman [185] through a statistical framework that cast boosting as a numerical optimisation problem whose goal is the minimisation of a loss function by a gradient-descent-like technique given raise to the Gradient Boosted Trees algorithm. By opposition to the so-far developed boosting classifiers, Gradient Boosted Trees do not fit to the previous errors, but to the residuals between the ground truth label and the value predicted. This enabled the definition of a differentiable objective function where the gradient descent is applied during training to optimise the learner.

The defined methodology carries significant risk of high variance and subsequent overfitting since the consecutive trees focus on misclassified samples. As a greedy algorithm, it may benefit from regularisation methods that penalise various parts of the model. Besides the different constraints that can be imposed on the construction of the decision trees (as the tree depth or minimum accepted impurity in a tree node), the complexity penalty (fundamental to keep the base models as weak learners) and the control over the number of estimators (that is equivalent to the number of times classification residuals are improved), there are other specific GBT regularisation techniques as shrinkage or row sub-sampling (the last one not used in this work). Similar to a learning rate, shrinkage reduces the influence of each tree, assuring a trade-off between generalisation and number of iterations. The application of a stochastic gradient boosting method - in which trees are sequentially trained with small subsets - may also be employed [186]. Furthermore, feature sampling (column sub-sampling) may work in the reduction of overfitting in the greedy process of growing more trees.

The mentioned Bayesian optimisation was, once again, employed making use of the Stratified Group 8-Fold described in Section 5.1. F1-score (*macro* average or of the positive class, depending on the type of classification) is used to optimise the parameters specified in Table 5.2. These main parameters already highlighted during this section are tuned both to combat overfitting and to guarantee the best performance of the model. As regards to *learning_rate* and *complex_penalty*, a log uniform distribution which distributes values evenly on a logarithmic scale is chosen in order to avoid the algorithm to waste time trying to distinguish between values of the same scale where the influence in the performance should not be felt. Concerning the number of estimators, only discrete values multiples of five will be evaluated.

Therefore, the best model is one more time trained on the whole training set and evaluated at the test set.

5.3.2.1 Results & Discussion

Following the example of Section 5.3.1.1, results are shown in Figure 5.4, incorporating validation and test results of the three variables under analysis.

In general terms, the first observation to be done is the similarity with the results of the other tree-based model tested, Random Forest. Again, performance declines for multi-class classification comparing with binary approaches, holding the binary distinction of pain values

Table 5.2: Optimised hyperparameters in the GBT sentiment-related variables classifier, with the considered values and definition.

Hyperparameter	Definition	Values	Distribution
Number of estimators	Number of gradient boosted trees (equivalent to number of boosting rounds).	[10, 200]	Discrete uniform Step: 5
Learning rate	Step size shrinkage used in update to control the influence of the most recent trees.	[0.001, 0.3]	Log uniform
Max Depth	The maximum depth of each tree comprised in the model.	[2, 12]	Discrete uniform Step: 1
Min Child weight	The minimum accepted impurity in a tree's node.	[2, 12]	Discrete uniform Step: 5
Column Subsampling by tree	The subsample ratio of columns when constructing each tree.	[0.5, 1]	Uniform
Complexity penalty	L2 regularisation parameter penalising excessive complexity in tree structure.	[1 ⁻⁶ , 10]	Log uniform

the best performance.

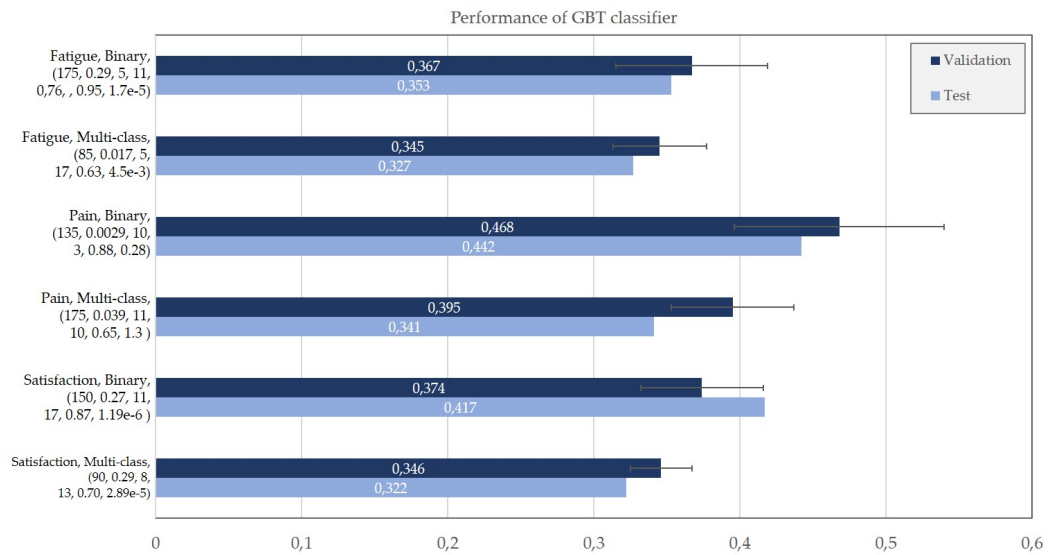


Figure 5.4: Performance of GBT classifier on sentiment-related variables prediction. X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme: *Variable to predict, classification type, (n_estimators, learning_rate, max_depth, min_child_weight, column_subsampling, complex_penalty)*.

Looking at the confusion matrices (Appendix A.2.2), particularly of the multi-class classification, the majority classes tend to monopolise all the predictions (fatigue model trying to counteract this known bias). The need to find appropriate methods to overcome these data imbalance start to arise as the minority class fails to be recognised. Moreover, GBT confirms the tendency found in RF that justifies the combination of low values of pain/fatigue, this is, the distinction between those values is not properly achieved, and the binary results retain the top performances. Concerning the threshold adjustment in the binary classification, quite low thresholds turn to be chosen (decreasing precision rates), allowing a significant part of the positive classes to be distinguished (even satisfaction that had been misclassified for the majority of instances of the positive class in RF model).

These results also show testing metrics similar to validation metrics. The same phenomena occurred in the previous experiment, suggesting that the built models are adapting properly to unseen data (in this case, through the Group 8-Fold training scheme and the limitations in hyperparameter ranges). Regarding the hyperparameters selected, results show that GBTs prefer shallow trees with low variance in opposition to the deep trees of the bagging ensemble. As concerns to the remaining regularisation parameters, it seems that they have been employed in a mixed way; regardless of the combination of the attributes found, the main goal was accomplished: avoid overfitting.

5.4 Classification with Support Vector Machines

Support Vector Machines [187] are perhaps one of the most commonly employed ML algorithms and the go-to method for a different type of problems. In summary, the SVM works by finding a hyperplane decision boundary in an N-dimensional space that best distinctly classifies the training examples. This division of the feature space into two classes of data points may be performed in different ways. However, the final goal should be finding the plane that maximises the margin; this is, the maximum distance between the nearest data point (regardless of the class) and the hyperplane.

While the hyperplanes are the decision boundaries that separate the data points of each class, the support vectors are the closest samples on either side and influence the orientation of the hyperplane that intends to maximise its perpendicular distance to those samples. The solution of the optimal hyperplane can be written as a combination of these few input points. Figure 5.5 shows an example of classification supported by SVMs.

Without any knowledge of the mapping, SVM finds the ideal hyperplane through the dot product functions in the feature set that are named kernels [189]. Kernel functions are responsible for mapping the feature space into higher dimensions and use linear approaches to define similarity between new data and the support vectors [190]. The most common kernel types are the linear - use simply a linear combination of the inputs -, Polynomial Kernel and Radial Basis Function (RBF) Kernel. The last one, in contrast to the linear one, adopts these transformations to higher-dimensional spaces, aiming to find the best separation between the two classes.

In practice, real-life data is not easily isolated by a hyperplane, and for that reason, the constraint of maximising model-margin must be soft, allowing a number of data points to violate the

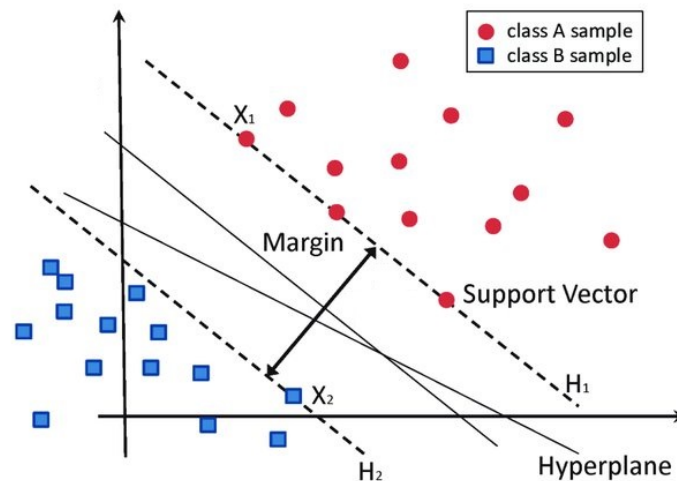


Figure 5.5: Example of classification of data by Support Vector Machines. The distance between hyperplanes (namely, the margin), two possible hyperplanes and the sides defining the margin - support vectors - are shown in the figure. Adapted from [188]

partitioning data. A regularisation parameter is introduced, named C , defining the magnitude of the violation allowed across all dimensions. The parameter trades off correct classification of the training samples against margin maximisation. C must be tuned during the learning phase as it is directly related to overfitting - using smaller values encourage a larger margin, a simpler decision at the cost of training accuracy. In opposition, for larger C values, the decision function will be better at classifying all training examples since a smaller margin value is accepted, losing the generalisation properties of the model.

SVM is capable of handling particular aspects as non-linearity and high-dimensionality of the data, which are recognised characteristics of current work data. This quality is especially improved if combining SVM with the RBF kernel function. As aforementioned, SVM is a binary classifier, and thus, in the case of the multiclass problem contemplated in this thesis, it is mandatory to convert the multiclass classification task into several binary classification problems, using a One-Versus-One approach [190].

The reduction of feature space is of utmost importance in classification with SVM. In typical pattern recognition tasks, the reason for this reduction is twofold: reduce computational complexity and cost, and reduce risk of overfitting, improving classifier's performance. It is advisable to discard features with no significant contribution as the integration of redundant information can jeopardise algorithm generalisation ability [191]. Dimensionality reduction could be done by both feature selection methods as well as feature transformation methods. These two different approaches will be further detailed afterwards. Correlated features are not removed in advance (as was done with the tree-based classifiers making use of the correlation matrix - Figure 4.8) since both features reduction methods assure that redundant information is not kept - either discarding features not relevant to the model or transforming the features.

5.4.1 Classification with SVM and Feature Selection

Feature Selection is the procedure of identifying and selecting the most relevant features from an entire set. This process was already mentioned in Section 5.3, where redundant features were removed based on the correlation between pairs of features, independently of the tree-based classifier used afterwards. On the contrary, during this section, a wrapper method will be implemented, which means a subset of features is selected in a model-dependent manner.

The Support Vector Machine Recursive Feature Elimination (SVM-RFE) is a wrapper technique relying on a recursive process of feature elimination [191]. The algorithm is implemented to create a feature ranking, sorted by order of significance in terms of a classification metric, intended to find the most important features. The process starts with all the available features and removes the feature with the least impact on classification performance at each step. The process ends with a stopping criterion, denoting that the desired features subset has been found.

As any regular ML algorithm, SVM-RFE parameters kernel type and C should be tuned for each specific situation. In accordance with the original SVM-RFE implementation, it was considered a fixed kernel type, the linear kernel, for the SVM-RFE method [192]. To avoid excessive computational cost, a formerly limited array of SVM-RFE C values was assessed. Since it was already shown that low C values improve the performance of SVM-RFE [193], the following possible values were evaluated: $C = 0.1, 1$. These were the only two values assessed for computation reasons and time concern, although, it is known that a broad range should be scope of future work. Of this process, different features ranking for each one of the regularisation parameter values were generated. In order to take into consideration the existing class imbalance, the most appropriate rankings were selected considering the F1-Score (from the positive class or unweighted average). The subset of features that maximises the defined metric - only setting a minimum number of 5 features - was determined. These features ranking are obtained relying on a cross-validation (group 8-fold) scheme using the training data.

Having obtained the subset of features for each provided regularisation value, the following step is to find the classification model and relevant parameters that guarantee the best possible performance. Unlike the FS phase, the kernel function employed was RBF, since, although it denotes a higher complexity, performs better than the linear kernel with nonlinear data [194], as it is the case of the data under study. Therefore, the non-linearity of data does not justify tuning that parameter and an SVM-RBF was employed for classification. A crucial regularisation parameter when using the RBF kernel is *Gamma* that accounts for how far the influence of a single training sample reaches. If using higher values, it considers only points close to the plausible hyperplane, and low *Gamma* will consider points at greater distances; therefore, more points are considered when finding the most accurate hyperplane. Consequently, the best combination of C and *gamma* should be tuned in the performance assessment.

Following the Cross-validation strategy described in Section 5.1, a Bayesian hyperparameter optimisation was performed using a Stratified Group 8-Fold. The following range for C and *gamma* were screened for SVM using the Cross-validation methodology already described, and ranges are present in the following table (Table 5.3).

A log uniform distribution was chosen in C and *gamma* parameters since, one more time, one did not pretend the optimiser to explore the range uniformly but make distinctions considering

Table 5.3: Optimised hyperparameters in the SVM-RFE sentiment-related variables classifier, with the considered values and definition.

Hyperparameter	Definition	Values	Distribution
C	Regularisation parameter to control the trade-off between margin-maximisation and the number of misclassified instances on training data.	[0.01, 20]	Log uniform
Gamma	Kernel coefficient for non-linear SVMs. Defines the curvature of the decision boundary by setting the influence of a single instance.	[0.01, 100]	Log uniform

different scales. For each feature subset obtained by the SVM-RFE (two for each sentiment-related variable and type of classification), the best combination of C and gamma regarding SVM-RBF has been obtained; then, the model is trained making use of all training instances and tested on the held-out patients' data.

5.4.1.1 Results & Discussion

The results obtained using the just described methods are summarised in Figure 5.6. Although the optimisation has happened for both C values used in the linear-SVM to perform feature selection, only the best combination results are present in the figure. This way, the model description follows the pattern: *Variable to predict, classification type, (linear-SVM C used in FS, SVM-RBF C, SVM-RFE gamma)*.

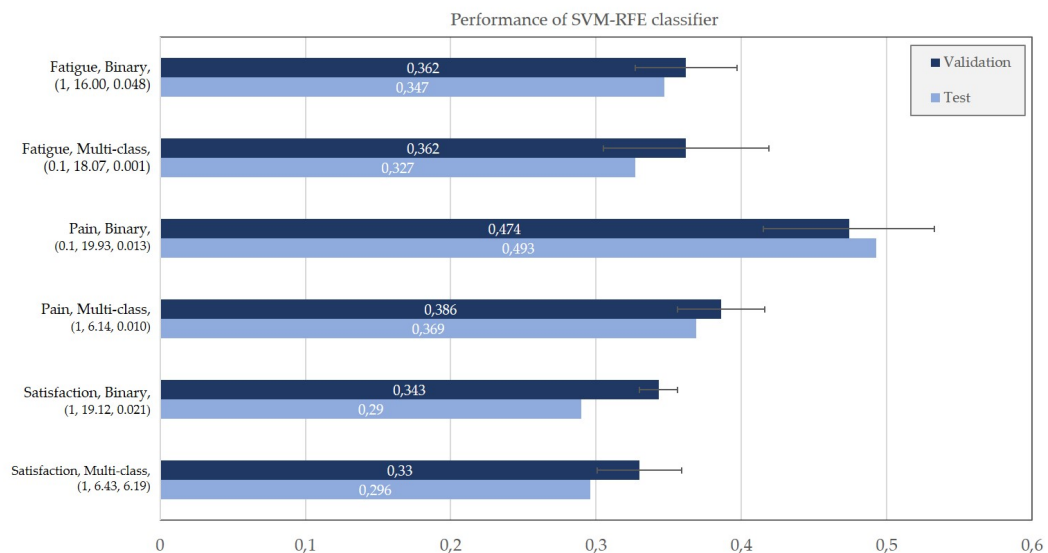


Figure 5.6: Performance of SVM-RFE classifier on sentiment-related variables prediction. X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (linear-SVM C, SVM-RBF C, SVM-RFE gamma)*.

The first observation emerging from results' analysis is the inconsistent performance of the models depending on the target. The results clearly show an unsatisfactory performance at satisfaction and fatigue classification both at binary and multi-class classification. Merging the classes of low values did not have favourable consequences given the close scores of each model. In opposition, binary pain classification outperforms the mentioned models, showing encouraging results.

An in-depth analysis of the confusion matrices (Appendix A.2.3) shows two distinct properties: in all multi-class models, the predictions are absorbed by the majority class (label 2) with the model clearly influenced by target distribution, as the instances of the most relevant class are roughly always misclassified. The other interesting pattern (this time, from binary classification analysis) is that admitting an increase in the false-positive rate, the model is capable of identifying a significant part of the examples of the positive class. The improvements are not as drastic in fatigue/satisfaction as the ones reported by the identification of sessions with high levels of pain. Additional ways to minimise the low precision measured should be considered.

Moreover, test and cross-validation scores are substantially equal, indicating that generalisation was guaranteed even though the feature selection was performed only resorting to the training data. In terms of the hyperparameter optimisation, some models seem to maximise the range of the regularisation parameter C , calling for experiments with prolonged ranges. Besides the computational overload of increasing C ranges, care is required: large increases in this regularisation parameter may very rapidly lead to overfitting since fewer misclassifications on the training set are accepted.

5.4.2 Classification with SVM and PCA

Principal Component Analysis is a popularly employed method for dimensionality reduction in continuous data with a wide range of applications in science and engineering [195]. It works by rotating and projecting data along the direction of increasing variance, creating features with maximum variance, the principal components. PCA is based on a simple idea of reducing a feature set, while preserving as much variability/statistical information as possible. It does so by generating these new uncorrelated variables that successively maximise variance. Figure 5.7 provides an example of data transformation using PCA.

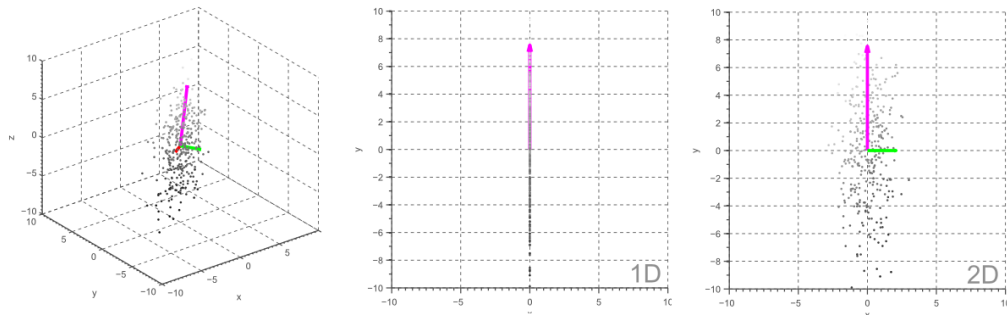


Figure 5.7: Example of data transformation through Principal Component Analysis. 3D data projected onto a 2D or 1D linear subspace by means of Principal Component Analysis.

Contrarily to FS method used in the last subsection (Subsection 5.4.1), it generates new features from the existing ones, keeping the most meaningful information. Moreover, this technique is not model-dependent, presenting this way some downsides. PCA may eliminate components with minor contributions to the general variance, which could nonetheless have a meaningful contribution to the classification task. Another possible limitation of PCA is that the principal components originated are less interpretable since they are built as linear combinations of the original variables. Prior to the computation of the principal components, it is critical to perform standardisation of the data since the latter is considered sensitive regarding the variances of the initial variables.

Given that it intends to find a set of linearly uncorrelated variables, the procedure starts by computing the correlation matrix of the formerly variables summarising the correlations between all the pairs of attributes. Eigenvectors and eigenvalues are computed from the covariance matrix with the purpose of finding the principal components of data. Eigenvectors are the directions of the axes in which there is the most information, and the eigenvalues, the coefficients attached to them. These new variables appear as linear combinations of the initial ones, where the most information within the original features are compressed into the first components - the largest possible variance is accounted by the first principal component. This organisation allows reducing dimensionality without losing information by discarding the components with low information, and thus, the number of components to use must be set. During this work, this number will be chosen based on the amount of variance that is supposed to be explained by the generated components. This explained variance was set at 95%.

The combination of PCA dimensionality reduction method and SVM classifier is proved to be efficient in different applications [196]–[198]; thereby, this combination was also tested in the following work. The cutoff of redundant information and a feasible reduction in computational cost are the basis for the development of this SVM-PCA model that could also lead to a performance gain.

The experiment designed followed during all the experiments already performed throughout this chapter and detailed along with Sections 5.1 and 5.2, was once again repeated in order to avoid any bias when comparing the results of the different trials. The hyperparameters to optimise, their brief description and values tested are present in table 5.4.

Contrarily to SVM-RFE, the kernel was also focus of the optimisation - the linear kernel can lead to promising results given the linear-nature of the PCA transformation. Choosing the appropriate kernel is crucial - an incorrect transformation can lead to poor results - and, that said, the two alternatives were tested. Gamma is a kernel coefficient only employed in non-linear hyperplanes; thus, it is ignored when the linear kernel-function is applied.

Hence, the optimised model is re-trained on the entire training set and the performance evaluated on training set.

5.4.2.1 Results & Discussion

The sentiment-related variables prediction performance using the SVM-PCA model are displayed in the following figure (Figure 5.8). Both validation and test results are included.

Table 5.4: Optimised hyperparameters in the SVM-PCA sentiment-related variables classifier, with the considered values and definition.

Hyperparameter	Definition	Values	Distribution
Kernel	Kernel function to map the data, and based on these transformations an optimal boundary will be used to separated the feature space.	{rbf, linear}	-
C	Regularisation parameter to control the trade-off between margin-maximisation and the number of misclassified instances on training data.	[0.01, 20]	Log uniform
Gamma	Kernel coefficient for non-linear SVMs. Defines the curvature of the decision boundary by setting the influence of a single instance.	[0.01, 100]	Log uniform

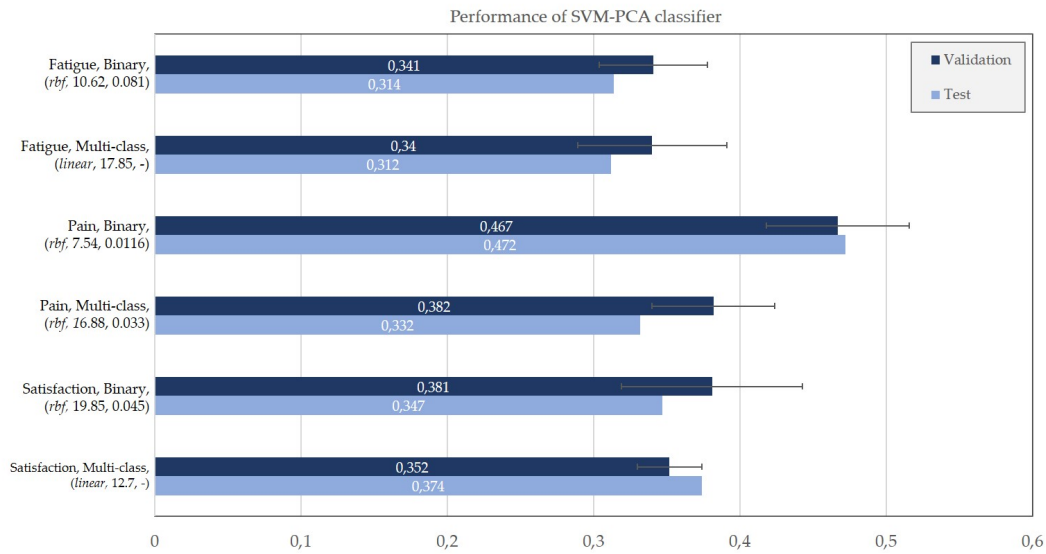


Figure 5.8: Performance of SVM-PCA classifier on sentiment-related variables prediction. X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (kernel, C, gamma)*.

From Figure 5.8, the chosen approach does not succeed in distinguishing the different values of the sentiment-related variables in a consistent way. However, the performance slightly increases when the problem is transformed into a binary classification challenge. The top performance is occupied, once again, by the distinction between patients experiencing (or not) pain - about 47% of F1-score of the positive class.

By the analysis of the confusion matrices present in Appendix A.2.4, it is possible to take further conclusions on variable-basis. Regarding pain prediction, the first observation that stands

out is the monopolisation of predictions by the majority class (label 2) in the multi-class approach. For the binary classification, by adjusting the threshold and penalising the precision, it is possible to get promising identification rates of the positive class (around 77%). The cost of this adjust is the elevated number of false-positives that could result in practice on specific attention by the therapist in a patient that did not require that special concern.

An interesting observation arises from the fatigue prediction confusion matrix given the fact that predictions are not that biased towards the majority class. In the multi-class approach, it may be noticed an increased distinction between the classes of low levels of fatigue what was not achieved in pain classification. Either way, the identification of the positive class increases almost 30% when classified against only one class.

Finally, satisfaction shows one more time the least acceptable classification - algorithm predicts nearly always the most frequent class - label 2 in the confusion matrix (Appendix A.2.4). The excessive imbalance (even if compared with the distribution of pain and fatigue variables) results in the lowest performance and null identification of the classes of interest. For the binary approach, in order to obtain a relevant percentage of true positives, the precision of the positive class has to be overly sacrificed.

As regards to the parameters selected, in a similar way to the SVM-RFE (Section 5.4.1), the maximum values of the regularisation parameter C tend to be preferred. Back again, feature work should contemplate extra tests with higher values of C , bearing in mind that special care should be taken in order to guarantee that overfitting is avoided. It may also be noted that the best classification performance in some cases was achieved with the linear kernel, justifying its inclusion in the tuning parameters. Gamma appeared not to have so much influence seeing that distinct values were picked.

The transformation of features resorting to PCA method did not have promising outcomes even if compared to values obtained in the previous section. PCA could have resulted in some loss of spatial information or damaged the practical meaning of some extracted features. As it was applied on meaningful features extracted from the dataset and not on raw-data, their combination in other linear representations may have compromised the employed feature space. Different features reduction techniques might be further explored in order to overcome the mentioned limitations.

5.5 Classification with Neural Network

A Neural Network (NN) is a common name for a group of classifiers whose the most notorious implementation is the multilayer perceptron (MLP). These brain-inspired systems that intend to replicate the way a human learns are part of the subfield of ML named Deep Learning, frequently employed in the field of Affective Computing as was denoted in the Literature Review (Sections 3.2.1.2, 3.2.2 and 3.2.4.2). MLP is one of the classes of networks, providing much flexibility and have demonstrated to be beneficial and reliable in an extensive range of problems [199]. MLPs are appropriate for classification problems as the one understudy.

Hence, an MLP is a class of feedforward NNs comprising a set of perceptron classifiers arranged in a multi-layer architecture. A perceptron or neuron, the building block of a NN inspired

on the biological neural, are simple computational units that evaluate a weighted combination of the input signals and produce a numeric output using a non-linear activation function [86]. Thereby, a multi-layer NN is comprised by a considerable number of neurons together combined in a pattern of connections [200]. These units comprise three different types where the features are propagated: an input layer of neurons (receive the information to be analysed), the output layer (where are found the results and its number usually is equal to the number of classes) and several hidden units/layers that form all the blocks between the previous ones. All the successive layers are connected to each other by a group of weights (in addition to an independent bias factor per neuron). A representation of a simple NN may be found in Figure 5.9.

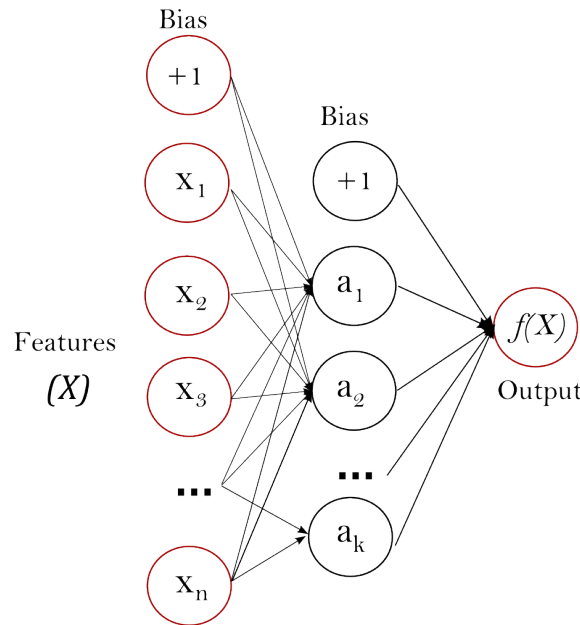


Figure 5.9: A simple MLP neural network for classification, adapted from ⁷. The leftmost layer comprises a set of neurons representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation, followed by a non-linear activation function. The output layer receives the values from the last hidden layer and transforms them into output values.

Throughout the learning phase, the weight value of each connection is adapted to minimise a cost function. The final goal of this optimisation process is to assure that, given a training instance of a specific class, the respective output neuron presents the highest value. Given the discrete format of our classes, categorical cross-entropy appears as the most-used cost function. Equation 5.3 represents the mathematical formulation for the categorical cross-entropy cost function $J(\omega)$ where y is the binary indicator (0 or 1) if class label c is the correct classification for observation o and, p is the predicted probability of observation o belonging to class c . M represents the number of classes.

$$J(\omega) = - \sum_{c=1}^M y_{o,c} \log(p_{o,c}) \quad (5.3)$$

⁷Scikit-Learn Theoretical Modules: scikit-learn.org/stable/modules/neural_networks_supervised. Accessed on 16-06-2020

In a multiclass problem, a separate loss for each class per observation is computed and the result summed as noticed above. In binary classification, cross-entropy may be simplified to the following equation (Equation 5.4).

$$J(\omega) = -(y \log(p) + (1 - y) \log(1 - p)) \quad (5.4)$$

Backpropagation is the technique that allows the minimisation of the cost function by adjusting the weights and biases of the connections. The gradients of the cost function determine this level of adjustment with respect to the learning rate, a time-varying correction factor that usually decreases as training advances (to avoid $J(\omega)$ overshooting) [86].

The performance of an MLP relies on different characteristics as the number of layers (and the number of hidden neurons of each), the units activation functions, among other aspects. The determination of the most suitable network architecture (between a single layer or a pair of sequential layers, with a varying number of output units); the dropout rates and the learning rate will be the target of the Bayesian Optimisation. Dropout regularisation is an overfitting counter-measure where randomly selected neurons are ignored during training - connection between network cells is temporally removed, and the neuron weight is not updated on the backward pass. The consequence is that the network becomes less sensitive to specific weights, consequently, less likely to overfit the training data.

The hyperparameters (and a brief description of them) and the considered ranges/values are demonstrated in the following table (Table 5.5).

Table 5.5: Optimised hyperparameters in the MLP neural network for sentiment-related variables classification, with the considered values and definition.

Hyperparameter	Definition	Values	Distribution
Learning Rate	The rate at which the model weights are updated during backpropagation through time in response to estimated error.	[0.001, 0.1]	Log uniform
Num. units dense layer 1	The dimensionality of the output of the first dense layer, to be passed to the second layer.	[10, 100]	Discrete uniform Step: 10
Num. units dense layer 2 and 3	The dimensionality of the second and third layer output. If 0 in layer 2, both cells are non-existent and outputs flow from the first cell to the final dense layer. If 0 in layer 3, outputs flows from the second cell to the output layer.	[0, 50]	Discrete uniform Step: 10
Layer 1, Layer 2 and Layer 3 dropout rates	The fraction of Dense cells' output connections to be randomly dropped.	[0. 0.5]	Uniform - Step: 0.1

The training model of each model was performed using the Adam optimiser with step decay (reducing learning rate as the training progresses may help to avoid gradient explosion); and ei-

ther *categorical cross-entropy* or *binary cross-entropy* depending on classification type, respectively, multi-class and binary classification. Adam optimiser is an extension to stochastic gradient descent that has recently seen broader deep learning applications. It is one of the most popular choices and computationally efficient (low memory requirement), justifying, this way, its use.

Models were trained with early stopping triggered by no changes greater than 0.5% in validation F1-score throughout 10 epochs. It was avoided the definition of a constant number of epochs (an arbitrary large number of training epochs was specified) and overfitting of the training set due to too many epochs or underfitting by the opposite reason prevented. The batch size (hyperparameter that controls the number of training instanced to work through before the model's internal weights are updated) was defined at 64. The goal was to find a compromise between memory-requirements and low generalisation error, and an accurate estimation of the error gradient. The all set of extracted features is used - without the removal of correlated features - as it is also intended to evaluate the power of DL methodologies as an embedded FS method.

5.5.1 Results & Discussion

After the optimisation through cross-validation using the Stratified and Group 8-fold (further detailed in Section 5.1), the best model was re-trained in the whole training set and evaluated on the test set as usual. The performance of the MLP classifier, including the average scores obtained with CV (with associated standard deviation) and test scores are displayed in Figure 5.10, with corresponding confusion matrices in Appendix A.2.5.

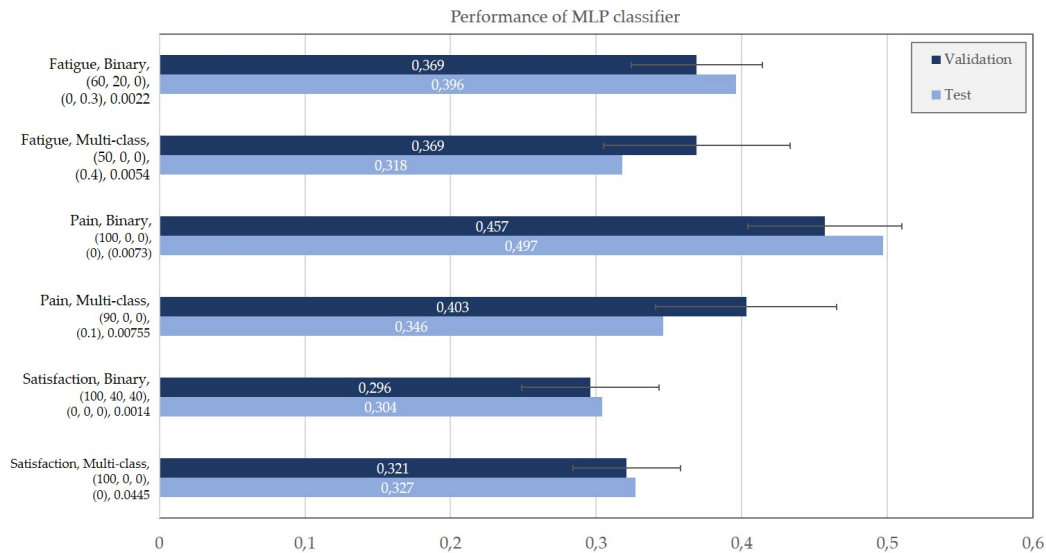


Figure 5.10: Performance of MLP classifier on sentiment-related variables prediction. X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable, classification type, (Num. units layer 1, Num. units layer 2, Num units layer 3), (Layer 1 dropout, Layer 2 dropout, Layer 3 dropout), learning rate*.

The models did not show any improvement when compared to the previously implemented models. Except for pain binary classification, models show sign of underfitting (particularly satisfaction models in which the most represented class is constantly predicted), not being able to differentiate between the diverse stats of the patient at the end of a therapy session. Another relevant observation is the expansion of the standard deviation values, showing no consistency over the different data splits. Following the same logic, it is noted testing scores higher than the average validation scores, suggesting data is not being modelled adequately by the diverse NNs architectures designed.

Looking at the confusion matrices, the most reasonable performance is perceived at pain binary classification in which almost 70% of the instances of interest are identified. However, the high percentage of false positives harms the global F1-score, asking for the implementation of balancing techniques.

Regarding the MLP architecture resultant from the Bayesian optimisation, it is possible to conclude that the model has not improved with the number of sequential layers - three layers preferred only in Satisfaction Binary classification whose results are not auspicious. Generally, dropout rates are kept close to 0 what may indicate that the designed model is not demanding additional regularisation methods. In its turn, the present results demonstrate that low learning rates are favoured for convergence, assuring neither any local minima are missed, nor gradient descent overshoots/diverges.

In conclusion, the implemented Deep Learning approach did not provide promising results, meaning that either the architecture does not fit to the data employed or the feature extraction methodology was not appropriated to the model under study. More complex NN structures should be the focus of future work (as CNNs or RNNs).

5.6 Imbalanced Learning Techniques

The issue of learning from class-imbalanced data has been receiving increasing attention in several different areas from academia to industry or government agencies [201]. The rationale behind the imbalanced learning problem is the capability of imbalanced data to considerably endanger the performance of the most ordinary learning algorithms. A typical approach assume or await balanced class distributions or similar misclassification costs. Thus, when facing not balanced data sets, as it is the case of the problem under study, these algorithms tend to fail to represent the distributive characteristics of data and, consequently, leading to non-satisfactory results across all classes of the data.

An extra difficulty appears for learning algorithms, as they tend to be biased towards the majority group, neglecting the minority class - commonly the class of interest in the real-life situations (again, the case of the present sentiment-related classification). The induction rules that explain the minority concepts are generally fewer and weaker than those of majority objects as the minority class is both outnumbered and under-represented. Therefore, this conceptual bias should be overcome and proper systems designed.

The typical imbalanced learning approaches may be distinguished into three types according to [202]: Data-level methods (modification of the set of instances to balance distributions),

algorithm-level approaches (directly modify learning algorithm to mitigate the inclination towards majority class) and hybrid methods that combine the former strategies.

During this section, the task defined in the beginning of the chapter will be addressed, making use of the combination of the different classifiers already tested during this chapter, plus a couple of imbalanced learning techniques. The main motivation is to counteracting the unequal distribution of data in the three variables that have been predicted - pain, fatigue and satisfaction. The section comprises both an algorithmic and a data-level approach.

5.6.1 Cost-Sensitive Method

Unlike sampling methods that try to balance distributions by amending the proportions of class samples in target distribution, algorithm-level approaches concentrate on modifying the existing learners to alleviate their bias. The most popular branch is the cost-sensitive approaches [203]. These algorithms try to overcome the non-balance by setting high costs to the misclassification of the less represented class instances, resulting in varying penalties for each of the considered groups of examples. Through the assignment of higher costs to minority classes, its importance is boosted during the learning process whose objective is to minimise the global cost associated with classification errors. In specific problems, the actual values of the cost-matrix may be defined based on domain-expertise. If that is not the case, and the misclassification errors can not be pre-set, a typical approach is to estimate the sample weights with regard to classes distribution.

An essential foundation directly associated with cost-sensitive learning is the concept of cost-matrix. It may be defined as a numerical representation of the penalty of classifying examples from one class to another. In a binary classification problem, the cost of misclassifying a majority class instance as a minority class example is defined as $C(M_{in}, M_{aj})$ and $C(M_{aj}, M_{in})$ represents the cost of the opposite case. Then, a representation of a cost-matrix in a binary classification problem (Table 5.6):

Table 5.6: Example of a Cost-Matrix in a binary classification problem.

Cost Matrix		Predicted Class	
		M_{aj}	M_{in}
True Classes	M_{aj}	0	$C(M_{in}, M_{aj})$
	M_{in}	$C(M_{aj}, M_{in})$	0

The cost of correct classification is set to zero and, in this approach under discussion, the cost of misclassifying minority instances will be higher than the contrary case ($C(M_{aj}, M_{in}) > C(M_{in}, M_{aj})$). During this work, the algorithm will automatically adjust samples weights inversely proportional to class frequencies in the input data, as no *a priori* quantitative information is available based on context. Data points from the majority class maintain their weight as one, while the minority class instances' weights are computed based on the proportion between the number of samples in each class. Then the algorithms will take these cost-proportionate weights into account, training to minimise the cumulative error. This methodology assumes that the class

distribution of the training data is representative of the broader problem, an assumption that can be made in the actual problem as training/testing split was conducted on that condition.

The aforementioned sample-weighting method may be applied based on all models that have been tested throughout this chapter: both tree-ensemble methods, SVM with the two feature reduction methodologies and the designed MLP. Thus, keeping the same experimental design and workflow, it was possible to implement the imbalanced learning technique in order to obtain comparative results. The algorithms were tune based on the same optimisation technique to avoid any possible bias. That way, Bayesian approach is kept as well as the tuned parameters and their respective range of values. Regarding the features employed, there was no apparent reason to change the set of features that are being applied, maintaining also the correspondent pipeline of each described model (for instances, removing correlated features before tree-based ensembles). A small difference may be found in the SVM-RFE pipeline when compared with Section 5.4.1: the regularisation parameter C in the linear-SVM used in the FS stage was set as 1 in all experiments, only creating a features ranking per variable, instead of the assessment of two possible values (0.1 and 1).

Therefore, the subsequent section presents the results & discussion of the application of cost-sensitive learning in the sentiment-related variables classification.

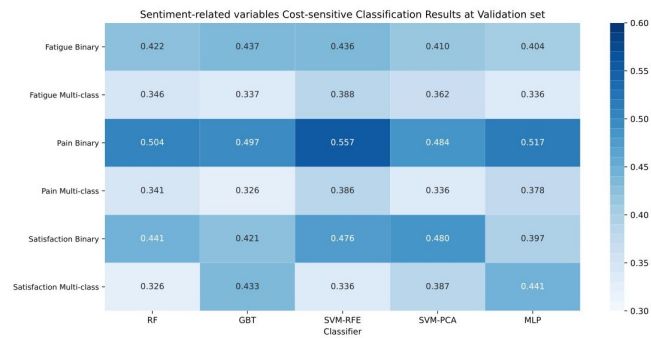
5.6.1.1 Results & Discussion

The models' performance is shown in the following figure (Figure 5.11). The subsequent matrices enable the comparison between the performance of the different algorithms in the several variables predictions. Both test and validation results are displayed. Additional results (from confusion matrices of each variable-model combination to global results of each model with parameters found in the optimisation process) are present in Appendix A.3.

The opening comment on the results above is the significantly higher performance of binary classification compared to the multi-class approach, making it the key factor in the variation between the results. The direct differentiation of the class of interest from all the others merged has provided higher results, not affecting the clinical relevance. Addressing the problem in this way allows a simplified identification of patients who require special attention (due to high levels of pain/fatigue or low satisfaction) such as onsite interventions or prescription adjustment.

Moreover, the binary classification of pain levels granted the best overall results with the highest F1-score of around 56%. The confusion matrix is now highlighted for additional comments (Figure 5.12).

The analysis of the confusion matrix shows very consistent detection of the positive class, with misclassifications only in 25% of the cases. In fact, the main drawback is the low precision-rate, responsible for lowering the f1-score of this class. The existence of a considerable share of false-positives (keeping in mind that 30% of the most represented class correspond to a large absolute number) suggests that other methods to decrease this false detection should be studied. In practical terms, this could lead to needless resources allocation or special concerns for patients who are not experiencing worrying levels of pain. This trade-off between false detections and correct identifications should be improved in order to achieve the desired clinical pertinence.



(a) Validation set results



(b) Testing set results

Figure 5.11: Results of classification of sentiment-related variables using cost-sensitive classifiers. Y axis refer to the variable and type of classification, while x-axis to the different classifiers employed. Values presented correspond to the F1-Score.

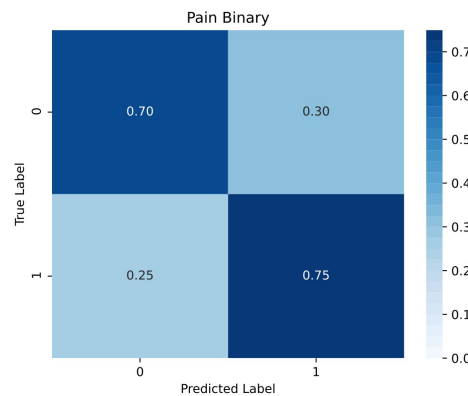


Figure 5.12: Binary classification of Pain using SVM-RFE cost-sensitive

Furthermore, comparing the diverse algorithms between themselves, both tree-based ensembles (RF and GBT) provided similar results, as was also the case in baseline approaches. On the other hand, the results obtained employing SVMs demonstrated divergent performances depending on the dimensionality-reduction technique applied before the learning task, as the recursive feature selection had substantially higher results than the transformation resorting to

PCA. As already verified in Section 5.4.2, PCA transformation is not leading to favourable results - perhaps due to loss of information or features impairment. Regarding the classification with the MLP neural network, the results have improved compared to the baseline method (in Section 5.5), confirming as well the benefits of this imbalanced learning technique in the tested deep-learning approach.

In summary, the technique had a positive influence on the overall results, increasing the detection of the least represented classes. In the baseline approaches, it had been noticed an extreme bias toward higher classes, and even in multi-class classification, it is seen an increased distribution in models prediction. For this last case, even though the global metric did not improve considerably, a higher absolute number of identifications of the most relevant samples is noted.

5.6.2 Undersampling and Oversampling Methods

The data-level approach concentrates on modifying the distribution of the data itself by attempting to provide a balanced distribution between the minority and majority classes. Two procedures arise within this data balance approach: the removal of some examples of the majority class (undersampling) and the generation of the number of minority class instances (oversampling). The data-balancing approach is the most predominant in the literature in terms of imbalance learning methods due to its simplicity and generality - it may be applied regardless of the base classifier to employ and had shown in several studies an improved overall performance [201].

5.6.2.1 Undersampling Technique

As stated, undersampling consists of the removal of instances of the majority class, either by random choice or applying some heuristic models. Random undersampling is seen as a risky strategy as potential significant information may be lost [204]. Although simple and effective, a drawback of this technique is the removal of samples without any consideration of their importance in the decision boundary definition between classes. Several techniques have been proposed in order to make the selected samples more representative. Various methods based on distance use distinct modes as highlighted in [204] to select the representative instances of the majority class. In order to attach the same issue of potential information loss, NearMiss methods [205] have been proposed by Mani and Zhang as different variations of distance-based algorithms.

As further discussed in Chapter 6, clustering is an unsupervised technique which assigns similar examples into the same cluster in terms of their similarity or proximity in the feature space. Therefore, it can be applied to find the mentioned representative instances of the majority class. With the use of clustering, the majority class could maintain the formerly general distribution as similar instances will be grouped and keep their representation in the new set of data. The clustering-based undersampling strategy that will be employed during this work was proposed by Lin *et al.* and the complete theoretical consideration may be found in [206]. The main rationale behind this technique was based on the assumption that each majority class has different clusters, and each of them has distinct characteristics. This assumption perfectly

suits to the problem at hand, as dissimilar patients can be found within each class - there are different reasons, for instance, that lead a patient to feel satisfied; or patients with different behaviours (and consequently in different regions in the feature space) may denote the same levels of pain/fatigue. The objective is to ensure that information concerning different data samples in the majority class is kept, and all sub-concepts are adequately represented. When performing random undersampling, if some sub-clusters were represented by a few instances, the likelihood of these instances being retained was relatively low. Thus, the cluster-based undersampling arose.

The cluster-based undersampling technique considers K-means as the base clustering algorithm since it is regarded as the baseline clustering method and the most widely used. The k-means algorithm is employed to produce k clusters given a defined number of centres (k), and then, these centroids are used to replace the whole majority class - keeping as much data samples as the number of clusters. A graphical representation of technique is displayed in Figure 5.13.

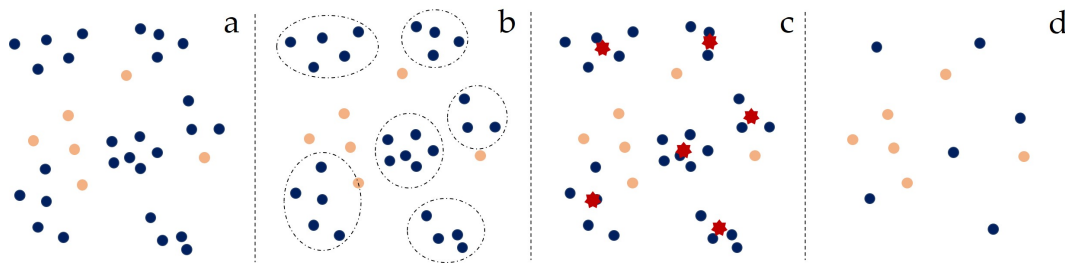


Figure 5.13: Graphical representation of the Clustering-based undersampling algorithm. a) The algorithm starts from a set of positive (orange points) and negative (blue points) examples; b) K-Means algorithm applied on the majority class data with k set as the number of minority samples - dashed lines represent clusters formed; c) clusters centres identified (red stars) - will be used to replace whole data of majority class; d) Final dataset undersampled.

The work of Lin *et al.* suggested to set k as the number of data samples of the minority class; one of the strategies to follow during this master thesis and the one highlighted in Figure 5.13. That being said, the down-size of the majority class henceforth will be performed by the described cluster-based undersampling algorithm, as it is shown to avoid loss of information, keeping the distinct characteristics of the data points in the most represented class, and improving the efficiency of the classification task.

5.6.2.2 Oversampling Technique

On the other hand, the other procedure covered by the data-level approach is oversampling. It comprises the replication of minority classes examples or the creation of new minority class patterns. Oversampling has the advantage of compensating the shortage of data with the generation of extra data, guaranteeing that there is no relevant information being discarded. However, the main challenge is to assure that the distribution and patterns of the minority class data are not neglected nor correctly emulated [207].

Several studies have assayed oversampling by data points duplication and notice that it does not considerably improve minority classes recognition [208]. Fundamentally, oversampling

minority class by instances replication will have a negative effect on classifiers as more specific regions in the feature space will be identified. In, for example, tree-based algorithms, it will cause new splits in particular regions, forcing the algorithm to learn more and more and, eventually overfitting [209]. This way, simple oversampling with data duplication was discarded as a viable solution to implement in this work.

A successful method for increasing minority class examples is the so-called *Synthetic Minority Oversampling Technique* (SMOTE) proposed by Chawla *et al.* [209] in 2002. SMOTE works by sampling data from the minor class, selecting examples that are close in the feature space and creating instances on the line segment that connects that point and one of their nearest neighbours. The first step is the random selection of the data point, then the defined k closest points are found, and a synthetic example is generated as a convex combination of the original data point and an again random selected point of the previously k identified [210]. Figure 5.14 provides a graphical representation of the described algorithm.

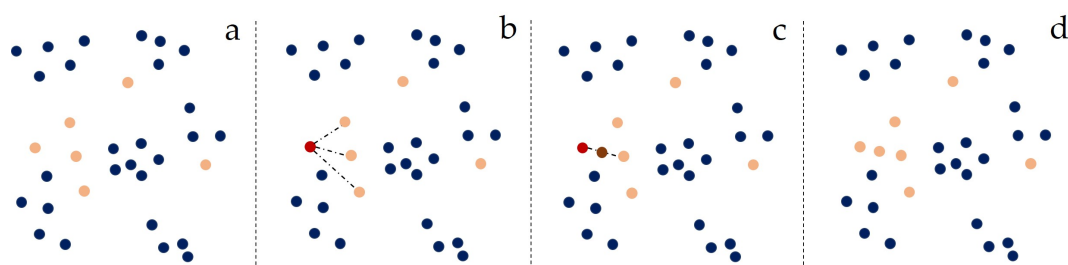


Figure 5.14: Graphical representation of the SMOTE algorithm. a) SMOTE starts from a set of positive (orange points) and negative (blue points) examples; b) selects a point from the minority class (red point) and its k nearest neighbours (defined as 3 in the example); c) one of the k nearest neighbours is randomly selected and a new synthetic example generated (brown point) along the straight line that connects the previous points; d) a new point of the minority class is added to the dataset.

Selecting the appropriate number of k -neighbours is also of paramount importance. If the algorithm only makes use of the nearest point, all generated instances will be formed on the line connecting the point in question and its closest neighbour. It would result in a high correlation between the generated samples and make any new point less valuable - similar information would be given to the model. On the contrary, large k may lead to higher errors given the possible influence of distant points in the feature space. As general conduct, selecting k in the range of 5 to 10 appears to be an appropriate choice [207]. Thus, 5 nearest neighbours will be used along all tests performed with the application of the SMOTE algorithm in this dissertation.

Thereby, SMOTE will be used to synthesise as many instances for the minority classes as required effectively. The generation of relatively close samples in feature space to original examples cause the classifier to establish larger decision regions that include nearby minority class points [209]. A general limitation of the SMOTE approach is that it blindly creates synthetic samples without considering majority class samples - possibly resulting in ambiguous instances if there is overlap for the classes - and may cause overgeneralisation.

5.6.2.3 Tested Methods

Several works suggest that the combination of over and undersampling performs better than plain undersampling or the application of SMOTE in isolation. A modest amount of oversampling can increase the bias towards the minority class, whilst the undersampling of more represented classes can help to reduce the bias on that class. An hybrid approach will, therefore, be designed and tested during this section. Following the technique described in [211], the mean of the number of instances of all the classes will be used as the threshold to define whether a class is oversampled or undersampled. Computing that number will be the first stage of the current algorithm fixing the desired number of instances per class. Moreover, it will be defined that any class cannot have its total number of samples more than duplicated. This intends to avoid having more samples generated than real ones, which could harm the data points real meaning and hinder the model generalisation (with adverse consequences on test results). The hybrid approach enables to create synthetic instances, without the addition of too much noise, and to reduce the redundancy of the majority class. It is known that, in extreme cases, it will not result in equally-balanced data; however, the intention was to avoid creating excessive differences in training and testing set distributions.

Therefore, the mixed over and undersampling (mentioned as hybrid approach hereafter) may be summarised in the following steps:

1. Calculate the average samples present in each class.
2. For each class, verify if the number of samples is below or above the average number of samples computed in the previous step for the specific class.
3. If the number of samples is above average, cut the exceeding samples resorting to cluster-based undersampler to obtain the average number of samples.
4. If the number of instances is below average, check if the average number is bigger than two times the original number of samples within that class. If true, only duplicate the number of samples of the original class (not equalising the number of samples in each class). If not true, using SMOTE technique, create the number of samples required to obtain the average number.

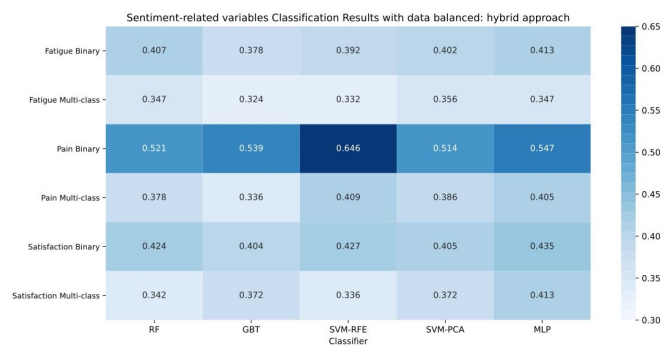
To investigate the effect of undersampling methods as standalone balancing technique, an additional approach was implemented and tested. In this methodology, it will not be applied any oversample technique (keeping the original number of instances of the minority class) and the cluster-based undersampling technique is employed to balance the dataset. As suggested in [206], this technique aims to demonstrate the applicability of using a small number of clusters (equal to the minority class) to represent all the information of the majority class. The oversampling technique is not applied to equalise the dataset independently (without combining with undersampling) since it is not intended to generate more news sample than those that already existed, as has already been stated.

Given the generality of data-level approaches, any classifier may be applied after the distribution adjustment. That way and following the pipeline of the last section (application of cost-sensitive learning, Section 5.6.1), all the classifiers will be tested (C parameter in feature selector

linear-SVM is again fixed as 1). Once again, parameters are tuned and the same set of features employed (maintaining the same preparatory work and workflow detailed in Section 5.2). Then, next section comprises the results obtained with both designed data-level approaches.

5.6.2.4 Results & Discussion

The comparison between the two data-level approaches is enabled by the following figure (Figure 5.15). All supplementary results are presented in Appendices A.4 and A.5) comprising: general results at validation set, results by classifier with associated best-tuned parameters per predicted variable and, confusion matrices (normalised at true classes) from all the possible combinations.



(a) Hybrid approach.



(b) Undersampling approach.

Figure 5.15: Results of classification of sentiment-related variables using data-balance approaches at testing set. Y axis refer to the variable and type of classification, while x-axis to the different classifiers employed. Values presented correspond to the F1-Score.

The results above encompass the results obtained from all classifiers at the different variables using both data-balance approaches. It is possible to note that, in general terms, the hybrid approach outperforms the undersampling approach in which data is equalised at the minority class. It could mean that a full equalisation in extreme circumstances may not be beneficial when evaluating the models on validation/testing set. As the data used to evaluate the models is not balanced (otherwise, the real-life situation would be distorted, and performance metrics jeopardised), equalising classes in extreme-unbalanced situations originates a prejudicial difference

in the distribution of training and testing set, and it may prevent models from achieving higher performances. Since, in the hybrid approach, a maximum value is set for the oversampling task, classes are regularly (especially in binary tasks) not equalised although balanced, leading to an improvement of the baseline results.

Moreover, there is clearly a combination of sentiment-related variable and classification type whose results are highlighted: binary classification of pain. It was achieved the best overall results with an F1-score of 64% in the testing set through the combination of SVM with previous FS and the hybrid approach. The confusion matrix may, again, be emphasised (Figure 5.16):

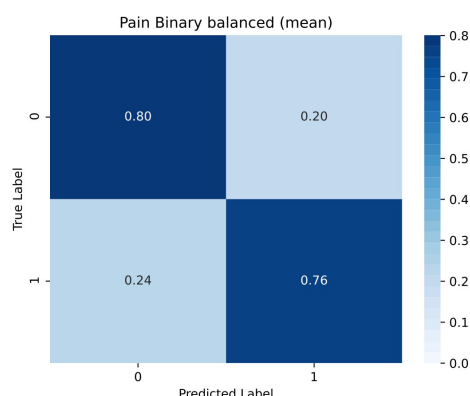


Figure 5.16: Binary classification of Pain using SVM-RFE with data balancing: hybrid approach.

Compared to the previous section (cost-sensitive approach: Section 5.6.1) whose best results were achieved with the same classifier, the percentage of false positives is lowered, resulting in increased precision and consequent F1-score of the positive class. Additionally, it may be reaffirmed that similarly to what is being remarked, binary classification surpasses multi-label classification. From the confusion matrices, it is noted that satisfaction distinction is not holding favourable results due to its high unbalancing - even with threshold adjustment, classes separation was not accomplished.

Regarding the different classifiers, the best performance was achieved through the SVM with recursive feature selection. Nevertheless, regardless of the classifier and compared to the baseline models, it was verified an increment in classes of interest predictions. The main drawback is the almost proportional rise of false-positive instances what prevents the global score from a significant improvement (as it is unweighted averaged in multi-class classification).

5.7 Ordinal Classification Techniques

In classification algorithms, unknown labels are estimated from a set of finite and unordered categories by mapping a group of attribute values. In regression, these numeric outputs assume continuous values. Practical ML applications regularly involve circumstances that exhibit an order among the different categories represented by the class attribute. Ordinal classification appears as a supervised learning problem, which has properties of both classification and regression. The learning task of ordinal classification (also called ordinal regression) is to assign

data points into a set of finite ordered categories [212]. It differs from classification due to the order of categories; moreover, unlike regression, the output variables (categories) are discrete and finite.

Decision systems which incorporate the classification of examples into classes with natural ordering are ubiquitous in everyday life. Frequent applications examples include recommender systems (where preference order of items are analysed), stock market analysis or breast cancer diagnosis [213], for example. Ordinal classification shows up frequently in rating situations, modelling human levels of preference where categories typically correspond to quotations or linguist terms [214]—varying from "very bad" to "excellent," for example - or in cases in which users are asked to rate a product/service based on a scale.

That being said, a clear relation to the problem at hand during this work may be found. Regarding the multi-class classification, there is an inherent order between the classes, although without meaningful numerical difference between them. As already explained, patients are asked by SWORD Health's system to rate their level of pain/fatigue and also satisfaction based on a relative scale converted into the numeric labels employed throughout this dissertation. These numeric outputs are used as class labels, without any absolute significance. Thus, even though addressed as a typical classification problem of K classes (numbered from 0 to $K-1$) throughout all chapter, this sentiment-related identification task may be interpreted as an ordinal classification problem. The natural order from 0 to $K-1$ is considered, and a couple of techniques designed specifically for ordered classes explored, aiming to accomplish performance improvements while maintaining classifiers generalisation.

The first works on Ordinal Classification date back to 1980 from McCullagh [215], in which a regression model was prepared considering ordinal information on the data and eliminating the need for labels assignment. Nowadays, different variations of traditional classifiers have been developed intending to take advantage of this order among labels, namely: systems based on DTs [216], [217], DTs ensembles [218], SVMs [219] or NNs [220], just mentioning a few works. More extensive reviews on the topic are offered elsewhere [221].

A couple of methodologies will now be explored and applied to the current problem: a application of $K-1$ binary classifiers to address a K -class ordinal problem, proposed by Frank and Hall [222] and, an adaptation of a classical NN to learn ordinal categories, work of Cheng *et al.* [223].

5.7.1 Ordinal Classification: a Simple Approach

Frank and Hall [222] proposed a simple trick that allows any conventional classification algorithm to make use of ordering information in class attributes. The methodology consists of the conversion of the original ordinal class problem into a series of binary tasks that encode the order. This method proved to be effective in different applications and showed it outperforms the naive approach (class values treated as an unordered set) on several multi-class dataset problems, also demonstrating that the improvements were not simply a side effect of transforming the problem into a set of binary-class ones.

The rationale behind the algorithm comprises using $K-1$ standard binary classifiers to handle the K class data problem. For that purpose, the training of the i -th learning algorithm is carried out by converting the ordinal dataset with classes $C_0, \dots, C_{K-1}$ into a binary dataset, opposing

$C_0, \dots, C_i$ against $C_{i+1}, \dots, C_{K-1}$. The following figure (Figure 5.17) shows the process of converting an ordinal dataset with four ordered values into $k-1$ binary datasets in which the classifiers will be trained.

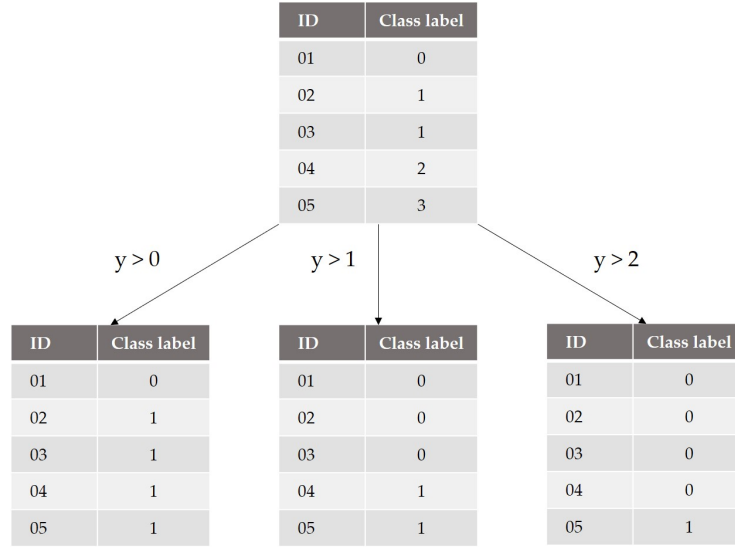


Figure 5.17: Transformation of an ordinal dataset with four values into three binary attributes.

Trained all the $K-1$ classifiers in an independent manner, their probabilities may be combined to produce a single estimation and, consequently, predict the ordinal value. The estimation of the probability of the first and last class depends only on a single classifier. The employment of a pair of classifiers is required for class values in the middle of the range. In general terms, class probabilities may be computed as follows (y as target):

$$\begin{aligned}
 Pr(C_0) &= 1 - Pr(y > C_0) \\
 Pr(C_i) &= Pr(y > C_{i-1}) - Pr(y > C_i), \quad 0 < i < K - 1 \\
 Pr(C_{K-1}) &= Pr(y > C_{K-2})
 \end{aligned} \tag{5.5}$$

Upon prediction time, a data sample of an unknown class is predicted by each of the classifiers; the equations above are used to compute the probabilities for each of the K ordinal classes, and the class with the maximum probability is allocated to the example.

One of the main advantages of the method compared to custom-made algorithms for ordinal classification is the generalisation ability, as it is not required any modification to the standard binary classifiers. Moreover, any classifier may be employed as the basic learner of this strategy. Concerning the original work of Frank and Hall, decision trees were used as the building block and, given the favourable results, this study will focus on the application of tree-based algorithms: RF and GBT.

Therefore, both tree-based ensembles will be employed as binary classifiers, and the results will be presented in the next section. Both learning models were optimised following the scheme used throughout the current chapter (Bayesian optimisation through 8-Fold cross-validation) and the considered values of Tables 5.1 and 5.2, respectively. All the binary classifiers are optimised independently; thus, the resulting parameters will not be displayed. Concerning the feature set,

the rationale of the tests performed in Section 5.3 is kept, maintaining the removal of the correlated features (although feature importance is not to be assessed) and the same pre-processing detailed in Section 5.2 - encoding of the categorical features and standardisation of the feature set (even if not mandatory in tree-based algorithms).

Finally, fatigue and pain values ordinal dataset will consider only the typical four classes of the so-far presented multi-label classification - in order to enable direct comparison and reduce the difficulties created by the small number of instances in the highest levels (even though Frank and Hall [222] showed that the performance gap increases with the number of classes). Thus, three classifiers are trained in a single fit of the ordinal classifier. Regarding satisfaction and since the multi-class problem has been addressed with only three classes, only two binary classifiers would be required. Given that the binary classification of satisfaction in the tree-based baseline models did not provide promising results and the extreme imbalance between classes (and since it is not intended to apply imbalance learning techniques), satisfaction will be withdrawn from the current exploration.

5.7.1.1 Results & Discussion

The following Figure 5.18 presents the results obtained resorting to the ordinal classification by application of multiple binary classifiers. The confusion matrices obtained at the testing set with both tree-based algorithms may be found in Appendix A.6.1.

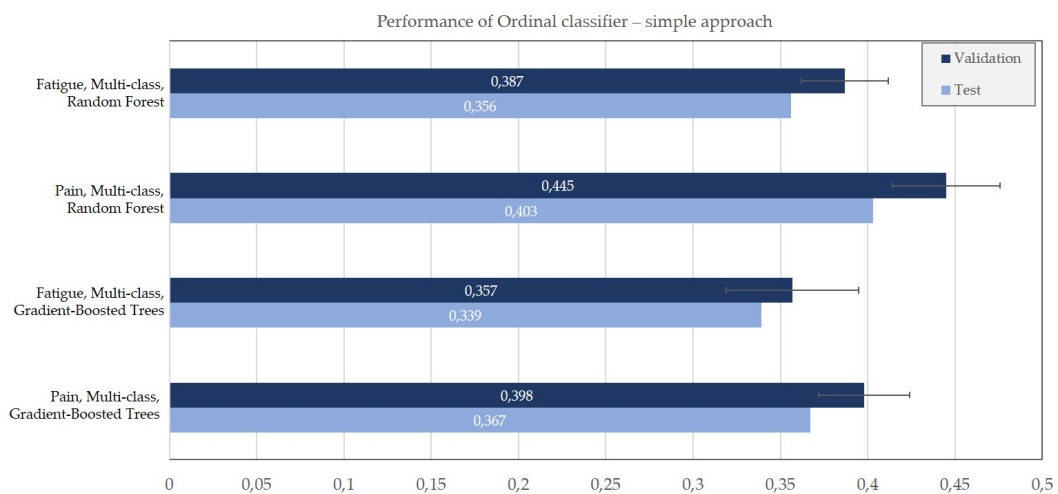


Figure 5.18: Performance of Ordinal classifiers: Simple approach. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable, classification type, Base classifier*.

From the figure above, it is possible to highlight the higher performance of the application of RF as base classifier when compared to GBT. Since the ordinal classifier uses prediction probabilities to estimate the class of each example and given that GBT makes use of shallower trees than the bagging ensemble, it may originate less accentuated (larger number of samples in final nodes) probabilities in the binary classifiers and, consequently, less promising results.

If compared to the standard classifiers (previous sections), it is noted a higher difference between validation and test results. The classifiers were optimised independently without fixing the cross-validation folds what could result in dissimilar divisions, hence, harming model generalisation. As future work, this observation should be taken into account and optimisation framework altered when tuning the individual classifiers hyperparameters.

Nevertheless, the results turned out to hold the top performance in terms of multi-class classification, only comparable with the results obtained employing data-level balancing techniques, specifically, the hybrid approach of Section 5.6.2. Having achieved the highest performance for pain classification, it is possible to analyse carefully the confusion matrix concerning the testing set evaluation (Figure 5.19).

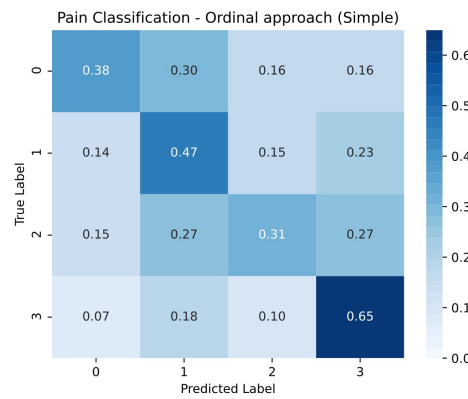


Figure 5.19: Multi-class classification of Pain using ordinal classifier: Simple approach by the application of binary RFs.

Comparing with the typical confusion matrices obtained by the application of the baseline approaches, it is possible to conclude that the ordinal classifier has dealt considerably better with the imbalance of the dataset. It is observed a clear decrease in the bias towards the most represented class (label 2) noted in the standard RF model (Section 5.3.1.1). The class of interest - patients with the most apprehensive levels of pain - are identified sensibly 65% of the cases which constitutes a promising result.

In conclusion, this ordinal classifier technique outperformed the baseline approaches in what multi-class classification concerns, with results only comparable with imbalance learning methodologies.

5.7.2 Ordinal Classification: a NN approach

An effective approach to adapt the traditional neural network to ordinal classification was suggested by Cheng *et al.* [223] and appeared as a generalisation of the perceptron learning into multilayer perceptrons (similarly to the method of Section 5.5). Given the promising performance displayed in the original paper and the diverse advantages noticed (as learning in batch mode and allow training on large and non-linear datasets with fast predictions), the method will be tested on the sentiment-related variables classification task.

For a classical classification network - as already implement throughout this thesis-, the goal is to predict the likelihood of a data point belonging to one class ($y = k$), and with that purpose, the target is encoded as the following vector $v = (0, \dots, 1, \dots, 0)$ where only the element v_k is set to 1 and all others to 0. The NN objective is to map the input vector x to a probability vector $p = (p_1, \dots, p_k, \dots, p_{K-1})$ where p_k should be close to 1 and other elements close to zero (constrained by the fact that sum of probabilities must be 1). By opposition, and as it is intended to take into account categories ordering, the target vector will be encoded following a different principle. When a particular data point is assigned to category k , it will also be assigned into lower-order categories, this is, from 0 to $k-1$. The target vector will be encoded as follows $v = (1, \dots, 1, \dots, 0)$ where $v_i (0 \leq i \leq k)$ is set to 1 and the other elements, set to 0. The goal is altered as the NN should map all $p_i (i \leq k)$ to values close to 1. The constraint is, consequently, also amended as the sum of probabilities must correspond to the number of categories to which the point belongs.

The learning of the method proceeds similarly as traditional neural networks, and a multi-layer NN may be built. Equally to the classical MLP, the number of dimensions of the input should correspond to the number of features; a varying number of hidden layers may be added without loss of generality; all the NN nodes may be fully-connected with previous and subsequent layers' nodes up to output layer, and any transfer function may be used in these connections (whether *sigmoid*, *tanh*, among others). The main difference lies in the output layer whose activation function cannot be softmax - generally employed to satisfy the constraint aforementioned and used in Section 5.5. By contrast, each output node (number equal to the number of classes) will use a standard sigmoid function to estimate the probability of a data point belonging to a category regardless of the other output nodes. The NN is not restricted to any normalisation; thus, it aims to adjust weights to generate probability outputs as similar as possible to the target vector.

During the learning phase, standard back-propagation processes are employed and *categorical cross-entropy* used as loss function, such as suggested by the original work [223]. The typical batch configuration may also be used for weights updates. The following Figure 5.20 presents a summary of the comparison between a typical NN and the developed ordinal classification NN.

In regard to the prediction step, the method scans output nodes in increasing order (from C_0 to C_{K-1}), stopping when the value of a node is smaller than a pre-set threshold (0.5 in the original work, so also the value of current experiment) or when no nodes are remaining. The predicted category will be the index k of the last node whose output is bigger than the threshold. The main drawback of such strategy and the use of independent sigmoid functions is that the monotonic relation ($p_0 \geq p_1 \geq \dots \geq p_{K-1}$) is not guaranteed and inconsistencies can be found.

This multi-threshold approach will, then, be applied to the classification at hand. Contrarily to the work of Cheng *et al.* the number of hidden layers will not be fixed to one. However, the addition of an extra one or two will also be studied (as the method enables the application of different architectures) making use of the typical Bayesian Optimisation (considered number of units and dropout rates outlined in Table 5.5). The overall configuration of Section 5.5 classification task is kept - Adam optimiser, batch size of 64 with training while early stopping is not triggered - in order to enable direct comparison. Moreover, no extra processing is performed at the feature set. Similarly to the previous sections, pain and fatigue are evaluated in accordance with the four classes configuration; satisfaction prediction is removed from the tests for

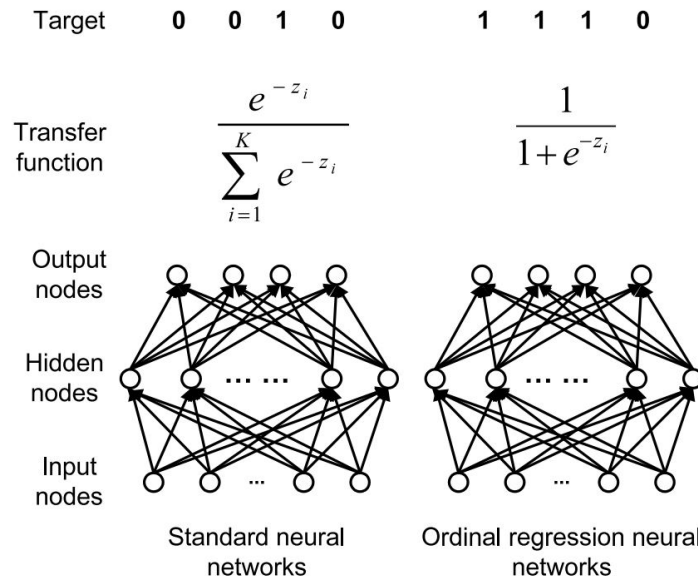


Figure 5.20: Comparison between a standard classification neural network and an ordinal regression neural network representing a 4 class ordinal dataset. Transfer output function corresponds to the softmax function in standard NN and sigmoid function in ordinal classification NN. Figure extracted from [223].

the same reasons - the extreme imbalance between classes and not favourable results obtained in the baseline method (Section 5.5).

5.7.2.1 Results & Discussion

The performance of the NN approach that considers the inherent order between classes is presented in the following figure - Figure 5.21. The models were re-trained on the entire training set in order to evaluate their performance on the test set. The respective confusion matrices may be found in Appendix A.6.2.

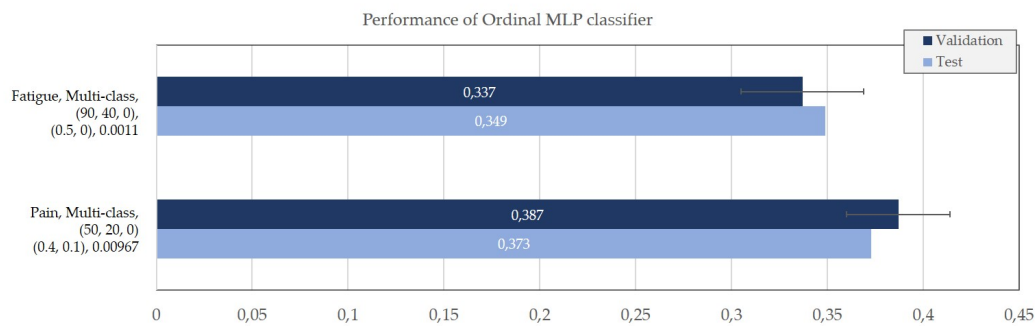


Figure 5.21: Performance of Ordinal classifier: a NN approach. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable, classification type, (Num. units layer 1, Num. units layer 2, Num units layer 3), (Layer 1 dropout, Layer 2 dropout, Layer 3 dropout), learning rate*.

The analysis of the figure above demonstrates the highest performance on pain prediction, in agreement with the results obtained throughout the entire chapter and, particularly with the ordinal classification approach of the previous section. Comparing directly with the standard NN performance (Section 5.5), it may be noticed an improvement in the evaluation on the testing set, with only equivalent performance to the application of data-balancing techniques.

Once more, although the results are not as remarkable as the ones obtained applying the simple approach of the previous section, it is denoted a substantial increase in the detection of the class of interest (from the confusion matrices), enabling to identify the maximum levels of pain and fatigue more than 50% of the cases.

Notwithstanding the fact that more tests are needed, from the results it may be affirmed that the model has benefited from the employment of multiple hidden layers - the optimisation procedure in both methods had caused the addition of an extra hidden layer in comparison to the original work of Cheng *et al.*.

To sum up, the current ordinal classification has shown favourable results comparing to the baseline results, calling for future work including threshold adjustments - thresholds used in the predictive stage could be dependent on the target distribution (not being fixed at 0.5) what could benefit the identification of the most relevant class.

5.8 Sequential approach (LSTMs)

Bearing in mind that data is sequential along with the session, different algorithms may also be tested. LSTMs are a class of Neural Networks (fundamental overview of NNs present in Section 5.5), more specifically, a Recurrent Neural Network, that intends to recognise patterns in sequences of data. It may apply to data from different domains as numerical time series emanating from sensors, stock markets but also text or audio and visual features as was reviewed in Section 3.2.2 and 3.2.1.2, respectively. The contemplation of the temporal dimension is the key differentiating factor of RNNs (and LSTMs) from the other typical neural networks.

The key aspect of these networks is that an RNN cell takes both information of the current input, and the output of the last step, enabling the processing of series of different lengths. LSTM was designed by Hochreiter and Schmidhuber [224] and tackled the problem of long term-dependencies of RNNs in which only the information of the most recent timesteps was considered. When a sequence was long enough, RNN suffered from short-term memory not being able to carry information from earlier steps to later ones. This appears as a consequence of the vanishing gradient problem where the gradient (used to update a NN weight) shrank as it backpropagated through time. The gradient turned to be extremely small, not contributing much in the learning phase.

LSTMs can by default retain the information for long periods of time, despite having a similar control flow as an RNN. The main difference relies on the processing inside the memory cell used to allow LSTM to keep or forget information. The following figure shows the processing block inside an LSTM cell (Figure 5.22).

The critical component is the vector C, the long-term cell state, that act as a transport highway that transfers relevant information all the way down the sequence chain. The main com-

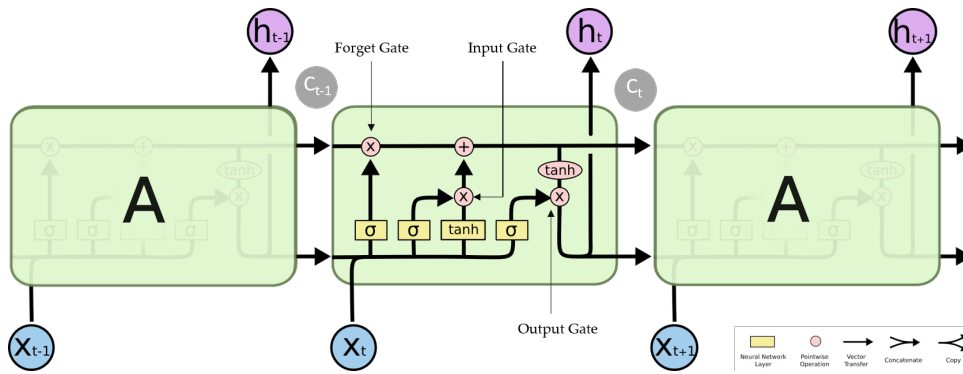


Figure 5.22: The processing blocks inside an LSTM cell, adapted from ⁸. σ and $\tanh$ represent fully connected layers with logistic and hyperbolic tangent activation functions, respectively. Inputs h_t and h_{t-1} are filtered and their influence on h_t and C_t controlled by the three logical gates.

putational task of A is maintaining C, deciding what information to keep, what to add and to be removed, in regard to the analysis and screening of the present input and of the previous output. The memory manipulations are done by four fully connected layers (represented by σ and $\tanh$) and three logical gates with specific tasks:

- **Forget gate:** the data no longer useful in the cell state is erased with the forget gate.
- **Input gate:** filter the data to be remembered, adding useful information to the cell state.
- **Output gate:** decide what should be written as current step output, extracting valuable information from the current cell state.

As it was the case of the MLP developed in Section 5.5, LSTM may combine in higher-level architectures with multiple hidden layers. Similarly to a hidden layer of a feedforward NN with a number of units, cells may be vertically stacked whose activations are sent forward to the next timestep. Regardless of the architecture to employ, the classification of each temporal step is conducted by placing a fully connected layer (with, during this work, softmax activation function) after the last LSTM layer.

LSTMs recurrent neural networks are able to seamlessly model problems with multiple input variables what may be considered a great benefit in time series forecasting. A time series (sequence of data points in the time-domain) may be univariate - sequence of measurements of a unique variable - or multivariate, where a sequence of measurements from multiple variables are gathered. Multivariate time series forecasting has significant interest nowadays, and LSTMs appear as the most commonly employed algorithms [225].

In the current work, data may also be employed in a sequential manner - the emotional and physical state of a patient is not only dependent on the current session but influenced by past sessions. In the same way, the actual state of a patient may be better predicted if taking into account the behaviour of the patient not only of the just-finished session. Features extracted may be seen as temporal sequences and be used to predict pain or fatigue levels of a patient at the last temporal point. The problem is again structured as a supervised learning task whose objective is to predict a specific sentiment-related variable given the patient behaviour of prior

⁸Understanding LSTM Networks: colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed on 17-06-2020

sessions. Thus, the application of the LSTM network.

Comparing to the last sections, data should be prepared differently. A sample corresponds not only to the data collected from the current session but to all the features from the most sessions. This way, it arises the possibility of a session being part of more than one sample. This option will be studied through the implementation of two data preparation approaches, whether including or not information from a single session in more than an instance. This strategy will also allow contrasting the number of previous sessions included in a data sample, either 5 or 10 last sessions (always including the just-finished session). When testing the methodology in which only five sessions are included, data will be introduced without overlap; this is, data from a session is only present in a sample in training/testing set. To note that the different number of sessions introduced in each approach aims to keep the total number of examples used in the model design similar (if used ten sessions with no overlap, number of samples would be considerably reduced). The input features are shaped into a 3D dimension (as usually expected by an LSTM), namely [number of samples, number of sessions, number of features]. In its turn, output samples shape is defined as [number of samples, 1] as it is only intended to predict the patient state at the end of the introduced sessions.

Furthermore, another difference to remark is the different set of features to employ. All the models already studied are not aimed explicitly at sequential data so the time notions were encoded in some features as: *Patient Performance* features that compare the current session with previous ones or *Exercise-level comparisons* that analyse specific exercises and subtracted information of the previous session into the present one (Section 4.3 for full details). As LSTMs are targeted at sequential data, these features that injected sequentially were removed in order to prevent any data duplication. Given the exploratory nature of this work, no additional feature selection was carried out (maintaining only feature normalisation and procedures discussed in Section 5.2).

Moreover, only pain and fatigue values are to be evaluated throughout this sequential approach. Satisfaction values are not collected in every session; thus, to avoid employing an interpolated value as a representation of a series of previous sessions, this variable was withdrawn from the sequential analysis. The reason behind this decision also relies on the distribution of classes in the satisfaction data (the imbalance is maximum); none imbalanced learning technique will be employed, that way, the variable will not be studied in order to avoid compromising the analysis.

Even for simple architectures, the amount of learnable parameters of LSTMs makes the computational cost of training in a Group K-Fold framework excessive. To keep computation times at an acceptable level and to avoid sampling the dataset, only four folds generated by the Stratified-Group 8-fold will be fixed as validation sets and used across LSTM parameter tuning. Learning rate, network architecture (between a single LSTM cell or stacked cells, with changing the number of units, and a possible dense layer after LSTM cells) and dropout rates (overfitting counter-measure discussed in Section 5.5) were the targets of the typical Bayesian hyperparameter optimisation procedure. Considered parameters and respective ranges/values are displayed in Table 5.7.

Models were kept simple in order to avoid overfitting - low dimensionality of LSTM cells. After the final layer (either an LSTM or dense), a fully connected dense layer with a softmax

Table 5.7: Optimised hyperparameters during LSTM training for sentiment-related variables classification, with the considered values and definition.

Hyperparameter	Definition	Values	Distribution
Learning rate	The rate at which the model weights are updated during backpropagation through time in response to estimated error.	[0.0001, 0.1]	Log uniform
Num. units LSTM layer 1	The dimensionality of the output of the first LSTM layer, to be passed to the second layer.	[10, 100]	Discrete uniform Step: 10
Num. units LSTM layer 2	The dimensionality of the second LSTM cell's output. If 0, cell is non-existent and outputs flow from the first cell to the dense layer (final or not).	[0, 100]	Discrete uniform Step: 10
Num. units dense layer 1	The dimensionality of the first dense cell. If 0, cell is non-existent and outputs flow from the last LSTM cell to the final dense layer.	[0, 50]	Discrete uniform Step: 10
Input dropout rate	Fraction of input connections to be randomly dropped.	[0, 0.3]	Uniform - Step: 0.1
LSTM Layers 1 and 2 dropout rates	Fraction of LSTM cells' output connections to be randomly dropped.	[0, 0.5]	Uniform - Step: 0.1

activation is placed. This cell with varying number of nodes according to classification type (if binary or multi-class) aims to make predictions for the last value of the sequence (fatigue/pain level) based on the output sequence of the last network layer. For all experiments, Adam optimiser was employed (again with gradient clipping) and models trained with early stopping activated if no changes greater than 0.5 % in validation F1-score are noticed across ten epochs, following the methodology of Section 5.5. Batch size was defined at 64 and *categorical* or *binary cross-entropy* loss used in accordance with classification type. Results are presented in the next section after re-training of the best model in the whole training set and evaluation in the testing one.

5.8.1 Results & Discussion

Following the Bayesian hyperparameter optimisation and evaluation of the best model on the testing set, a set of LSTM models associated with a sentiment-related variable and a number of sessions was generated. The performance of these relevant models, including validation (this time, only using four folds) and testing results, is displayed in Figure 5.23, with corresponding confusion matrices in Appendix A.7.

The first observation emerging from results' analysis is the similarity with the baseline models that only take advantage of data sequentially through the extracted features. From the

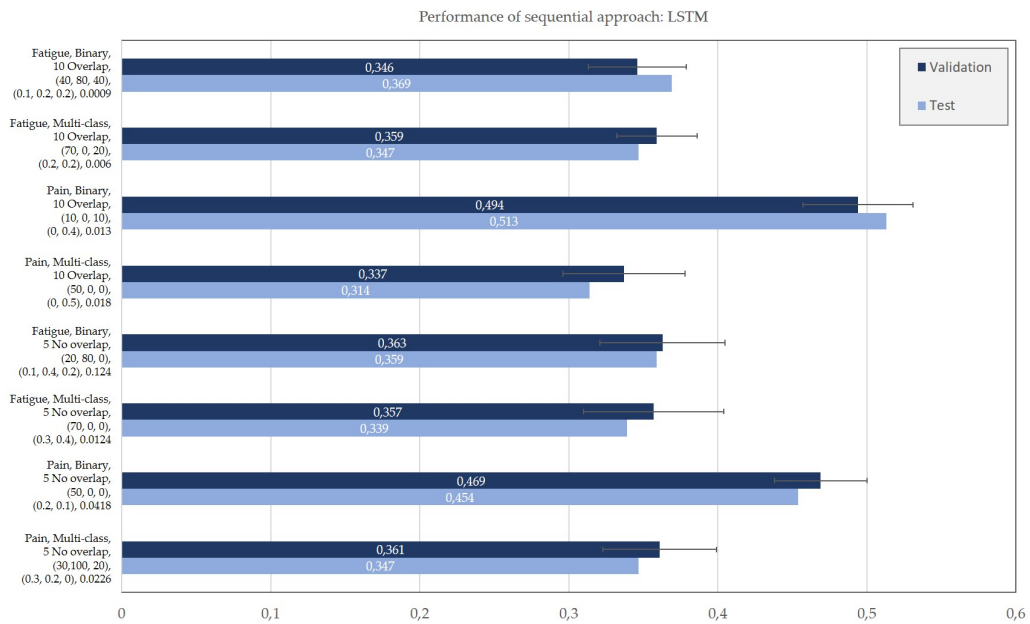


Figure 5.23: Performance of LSTM classification models on sentiment-related variables prediction. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, strategy used, (Num. units layer 1, Num. units layer 2, Num. units dense layer 1), (Input dropout, Layer 1 dropout, Layer 2 dropout), learning rate*.

confusion matrices, it is noted that these neural networks may be affected by the unequal distribution of classes in the data used. In multi-class classification, the most prevalent class tends to over influence the prediction, while in binary classification the separation between the two classes was not clearly found. The greatest exception is, once again, binary classification of pain presenting the highest performance with this sequential approach; however, the results obtained are clearly inferior to the ones achieved with the imbalance learning techniques of Section 5.6 and on par with the other deep learning approach tested (MLP classifier in Section 5.5).

Concerning the two alternatives studied regarding data preparation, the use of the previous 10 sessions with overlap outperforms the other tested approach - only 5 sessions without overlap. Even though the difference is minor, this slight improvement may indicate that the model benefits from the increase in past sessions.

While the results clearly show that lower learning rates are required for convergence, they are not as clear regarding the ideal architecture. Both stacked (2 LSTM cells) and unstacked (1 LSTM cell) architectures show up in the top models, with similar performances. The inclusion of an extra dense layer was also beneficial in some cases, not allowing to take definitive conclusions about its usefulness. In order to fully understand the impact on the results of deeper architectures, additional experiments with successively higher levels of stacking are necessary.

In conclusion, the results may be considered promising given the limited hyperparameter optimisation, architecture engineering, not specific feature engineering (though the removal of some temporal-dependent features did not prevent the models from achieving the results of the

other baseline models), and also, the non-application of balancing methodologies.

5.9 Comparative analysis & Conclusions

The main conclusions and considerations concerning the different methodologies tested may be outlined based on the following topics:

- **Sentiment-related Variables:** As the problem encompassed the prediction of three different variables, the performance between the prediction of those may be compared. Pain may be seen as the sentiment-related variable whose results are the most promising: in the majority of the conducted experiments pain prediction outperformed the other two variables occupying the top performance spots (64% of F1-score using the hybrid oversampling/undersampling approach). Satisfaction, maybe due to its extreme class imbalance, presents, in general, the least encouraging outcomes, even with the application of balancing techniques. The fact that all variables are dependent on patients' state-of-mind with different perspectives on the scale used as labels may be restraining the models from achieving better results. This is an important factor given that it had been tried multiple models and all of them revealed not-excellent accuracies. The difficulty of the problem may be related to the nature of the targets (extremely subjective) and its association with the data collected for their prediction.
- **Classification Type:** Regarding the two classification types embraced during this chapter, it may be stated that binary classification achieved the highest performances. The merge between the least-relevant classes (low levels of pain/fatigue and high levels of satisfaction) proved to be favourable as simply separating the rehabilitation sessions in two classes was better accomplished. Overall, the top performances of all classifiers correspond to binary classification tasks (generally, pain assessment). Clinical relevance is not compromised with the binary approach as the aim of identifying patients requiring clinical intervention (whether a call to the patient, prescription adjustments or even a home visit) is kept.
- **Baseline algorithms:** For all the classifiers, remarkable coherence among the results is observed. While this could be the expected behaviour when comparing the tree-based classifiers, the application of additional features reduction techniques or a deep-learning approach could lead to significant performance differences that have not been verified. The application of roughly the same feature set and the targets subjectivity could be behind these findings, calling for additional feature engineering work or the inclusion of different data sources (such as gesture-based-features extracted from IMU's or image data). Regardless of the classifier, it may be noted that the application of diverse overfitting counter-measures was effective and guaranteed the generalisation of the implemented pipeline.
- **Features Reduction Techniques:** In order to avoid misinterpretations of features importance in tree-based models, correlated features were dropped before the application of these models and showed no prejudicial effect. For the pair of reduction techniques employed with the SVM classifier, FS by means of recursive feature elimination outperformed the results with PCA transformation. PCA may have led to loss of important spatial information or adulteration of the realistic meaning of the extracted features - compromising

the results. It is important to note that the overall best result was obtained employing the SVM-RFE combination, thus fully supporting its applicability.

- **Imbalanced Learning Techniques:** Given the high-class imbalance that was verified in all sentiment-related variables, the improvement in the results with the application of imbalanced learning techniques was somehow expected. The data-level strategy, particularly the hybrid approach, accomplished the best results. A complete equalisation in extreme conditions showed not to be the most effective strategy, perhaps due to generation of training/testing sets with highly dissimilar distributions. Moreover, cost-sensitive learning provided a beneficial effect on the overall results, although increasing the number of false-positives with the identification of the classes of interest.
- **Ordinal Classification Techniques:** The main idea of these techniques was to retain the ordinality of the classes in multi-classification problems. The so-called simple approach achieved higher results when compared to the baseline results (mainly using binary RF classifiers), performing at identical levels of the imbalance learning methodologies. The NN approach, although not achieving the results of the other ordinal classification method, also proved to be beneficial for counteracting the observed biased towards the most represented class in the classic MLP designed.
- **Sequential approach:** The sequential approach showed promising results as a simple LSTM architecture was implemented - the number of bayesian optimisation iterations and the number of folds has also been reduced throughout this search for the most suitable architecture. Further specific feature engineering and application of balancing techniques should also be the focus of future work. Nevertheless, the results are comparable with baseline algorithms what may be considered encouraging also due to the removal of the extracted temporal-dependent features.

5.10 Summary

After defining the task to be addressed, the design and the typical workflow of the experiments to conduct initialise the chapter. Then, a set of baseline models was studied, presenting some theoretical considerations and providing benchmark results for future comparisons. The various methods also allowed to compare the models between themselves and evaluate the diverse feature selection methodologies followed. Given the observed target imbalance, a pair of imbalanced learning techniques was studied, providing the top results regarding binary classification. A designed hybrid approach (mixed over and undersampling) achieved the best overall results with an F1-score of 64% on pain classification through the combination with SVM and the recursive feature elimination method. The subsequent section proved the applicability of ordinal classification techniques by a simple approach that combines binary classifiers and a variant of an MLP-NN. It potentially opened up new research lines, as these methods are suited to the natural order existing between the sentiment-related variables, focus of the current chapter. A sequential approach based on LSTMs was also tested intending to recognise patterns from sequences of rehabilitation sessions. Given the experimental nature of this particular technical work, the

results may be considered encouraging and should be further explored. A comparative analysis finished the chapter, reflecting on the developed work and comparing the sentiment-related variables employed, the different classifiers tested, and the remainder approaches.

Chapter 6

Emotional-State Inference

The main goal of this task is to explore the user input variables to infer a quantitative label representative of emotional-state and design a proper algorithm to predict this same variable. The system has not incorporated any direct measurement of patients motivation level nor any emotion categorisation as was the case of the sentiment-related variables (fatigue, pain and satisfaction). This way, there is no implicit label to validate the algorithms when implemented. A combination of user input variables will be considered in order to find this implicit label and associated with the emotional-state.

Similarly to the task presented in the beginning of Chapter 5, this task also encompasses a typical ML problem. An identical set of the behavioural features extracted from patients performance and clinical prescription (described in Section 4.3) may again be used, as it is intended to perform the classification task with all the information acquired during a session.

Regarding the labelling procedure, the exploration of the input variables aims to find a possible representation of the emotional-state. Since it is not a supervised task, and there are no evaluation metrics that could be used to support the new partitions, context-based knowledge will be required in this validation procedure. With this purpose, SWORD Health's clinical and development team had been asked to, in an initial stage, hand their viewpoint on different subjects and details under discussion in order to find the aforementioned categorisation, and subsequently, to validate all process and the formed groups. After the emotional representation being approved and guaranteed the clinical interest in inferring patients' state based on the partitions created, the classification task will be employed.

A different non-parametric method that does not require domain knowledge or clinical validation will be studied in order to carry out a comparative analysis.

6.1 Clustering

Grouping objects is necessary for several purposes in different areas of engineering, health sciences, technology and other fields of daily life. Taking, as example, the medical area, patients with common symptoms should be placed in a group marked with an appropriate label - such as the name of a disease. Another set of people not holding those symptoms should not be

placed in the same group and be handled differently. This situation may be considered trivial as the labels are known *a priori*. However, in various circumstances, no such label is provided in advance, and the objects must be grouped based on their resemblance. So, it is the case of the task presented in the last section.

In generic terms, these cases are under the scope of unsupervised classification - classification task in which the class label is not previously tagged on the object, thus, not provided to the learning algorithm. The main goal behind this task is to develop a tool that may be used to predict the unknown object class, finding meaningful patterns in data and drawing references from the dataset, representative of each class. The task of dividing data points into several groups, creating a collection of objects in subsets based on the similarity and dissimilarity between them, is known as Clustering [226]. By convention, the clustering structure can be represented as a set S of subsets $S_1, S_2, \dots, S_k$, such that $S_1 \cap S_2 \dots \cap S_k = \emptyset$, meaning each instance belongs to exactly one subset and does not to any other.

As there is no precise definition of clustering, there are multiple clustering approaches with inherent techniques. Fraley and Raftery [227] suggested separating the clustering methods into two different categories: hierarchical (clusters formed by a tree-type structure based on hierarchy) and partitioning (data assigned into k -clusters, optimizing a particular objective function). Moreover, the similarity of objects within a cluster plays a crucial role in the clustering process. Typically, numerical values represent the features of the objects and, that said, the only way to extract any information pertaining to the relationship among patterns is using a measure of distance between the data points (e.g. Euclidean distance). Therefore, combining the different techniques and the multiple measures of similarity, there are countless alternatives to perform unsupervised learning, only resorting to clustering approaches.

Given the details of the current task, clustering appeared as a simple alternative to be applied in the mentioned labelling procedure. Taking the known sentiment-related variables, pain and fatigue introduced by the patient at the end of all sessions and satisfaction obtained by the labelling procedure (Figure 4.5 in Section 4.2), clustering will be applied to group the different combinations of these variables. Theretofore used as classification target (in Chapter 5), during the following experiments, the three patient-input variables are now employed as the attributes of a data points whose association is intended to provide an emotional-state quantitative label.

The subsequent sections explore two different approaches aiming to find this representation of the emotional-state. Similar to the design implemented in the last chapter, a training/testing split was carried out at the start, following the conditions detailed in Section 5.1.

6.1.1 K-Means

K-Means is the most well-known, benchmark and simplest partitional clustering algorithm, applied in different areas to solve clustering problems. In this procedure, data instances are classified through a user-specified number of clusters, k . The algorithm aims to find partitions such that the squared error between the empirical mean of a cluster (cluster centroid) and the points in that cluster is minimized. Let $x_1, \dots, x_n$ be the input data and k the number of clusters. The goal of k-Means is, thus, to find the set of k clusters, $L = \{l_1, \dots, l_k\}$, that minimise the following

objective function (Equation 6.1):

$$J(L) = \sum_{c=1}^k \sum_{x \in l_c} \|x - \mu_c\|^2 \quad (6.1)$$

K-Means is a greedy algorithm, converging to a local minimum, even there is a large probability of converging to a global optimum if clusters are well-separated [228]. As the squared error always decreases with an increase in the number of clusters, minimisation of the algorithm objective may only occur for a fixed number of clusters. The critical steps of the algorithm are as follows [228]:

In-

put: $x_1, \dots, x_n$: input data, k : Number of clusters.

Output: Clustering set $L = l_1, \dots, l_k$.

1. Initialise initial clusters taking k distinct points randomly, $p_1, \dots, p_k$. These points represent initial group centroids: $\mu_1 = p_1, \dots, \mu_k = p_k$.
2. Repeat until convergence (no changes in any data point assignment):
 - Assign each point to clusters $l_1, \dots, l_k$ based on the minimum distance between each point and cluster centroids, this is, $z_1, \dots, z_k : l_j = x_i | z_i = j$, generating the new partitions.
 - For each cluster l_j , compute the new centre, $\mu_j = \frac{1}{|l_j|} \sum_{x \in l_j} x$.

Figure 6.1 shows a representation of the K-Means algorithm on a two-dimensional dataset with three clusters.

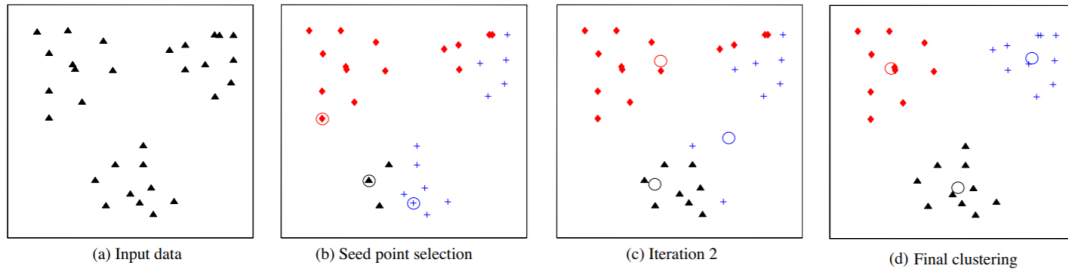


Figure 6.1: Illustration of K-Means algorithm, adapted from [228]. (a) Two-dimensional input data with three clusters; (b) three seed points selected as cluster centres and initial assignment of the data points to clusters; (c) intermediate iteration updating cluster labels and their centres; (d) final clustering obtained by K-Means algorithm at convergence.

The K-Means algorithm demands three user-specified parameters: distance metric, number of clusters K , and cluster initialisation. All parameters will be further analysed and discussed since they influence the results obtained. Concerning the distance metric, the Euclidean distance acts as the most commonly employed metric to compute distances between points and clusters centres, evaluating the mentioned similarity among data. This metric tends to find spherical data or ball-shaped clusters in data. It was selected to be used during this work, as it is the most intuitive and computational-efficient, and implicitly on the foundation of K-Means given the objective function already presented (Equation 6.1).

As a consequence of distance-based measurements applied to differentiate data points, the first decision that implies clinical assessment arises: the scale of variables. In the hope of achieving objectivity and giving all variables an equal weight, data standardisation or scaling is mandatory as different units and variations will have consequences for the groups formed. The standardisation is the practice to follow when no prior knowledge is possessed.

As it is known, satisfaction values are distributed between 0 and 10, while pain and fatigue variables have only six possible values (0 to 5). Variables should be adequately scaled in order to ensure that none of them has increased weight in the clustering process; however, the assignment of different weights may be justified based on context-matter knowledge [229]. Hence, the possibility of maintaining the original scales was presented to the clinical team, calling attention that this may result in increased importance for the satisfaction values. This option turned out to be the preferable alternative of clinical professionals. Two reasons were presented: since that it is the scale presented to patients, it would be desirable to work on these scales, and that way avoid transforming original values. The second reason is the risk associated with amplifying record errors or intermediate values. Thus, patients will be mapped based on the original scale, accepting that satisfaction may have a substantial effect on the resulting structure. This may also be beneficial given the distribution of the variable (extremely skewed towards high levels), allowing to distinguish the classes of most interest (low satisfaction values) from the regular ones.

Following with K-means parameters, the most critical choice is K. While no perfect mathematical criterion exists, a number of heuristics are available for choosing K. For that, it may be evaluated how well the models are performing based on different K clusters since they are used in the downstream modelling. As the objective of the algorithm during the learning phase is to minimise the squared error, a typical (but naive) approach known as "Elbow method" makes use of this minimisation goal [230]. The technique exists upon the idea that one should choose a number of clusters so that adding another cluster does not improve the data modelling significantly. The intra-cluster variation (or total Within-Cluster Sum of Squares) is plotted against k, and the optimal number is chosen at a point where the gain of increasing number of clusters is marginal (usually after relevant information gains). An angle in the graph is formed; hence, the elbow method. We obtain the following plot for the Within-Cluster Sum of Squares versus the number of clusters for the data under study (Figure 6.2).

The graph presented above is not particularly conclusive because the curve is monotonically decreasing and does not show a distinct point where it starts flattening out. Nevertheless, the inertia declines the most when a third cluster is added. Thus, if there was no domain-expertise or context, more tests should be performed to find the optimal number of clusters. Naming an alternative that could be tested to find the appropriate number of partitions, the Silhouette score is a measure of the average similarity of the objects within a cluster and their distance to the other objects in the other clusters. This analysis allows to determine the degree of separation between clusters and, therefore, be used to select k. Among others, Gap statistic developed by Tibshirani *et al.* [231] is one of the most accepted methodologies, comparing the within intra-cluster variation for different values of k with their expected values under the null reference distribution of the data.

Therefore, the suggested tests could also be the basis of an analysis in which the algorithm

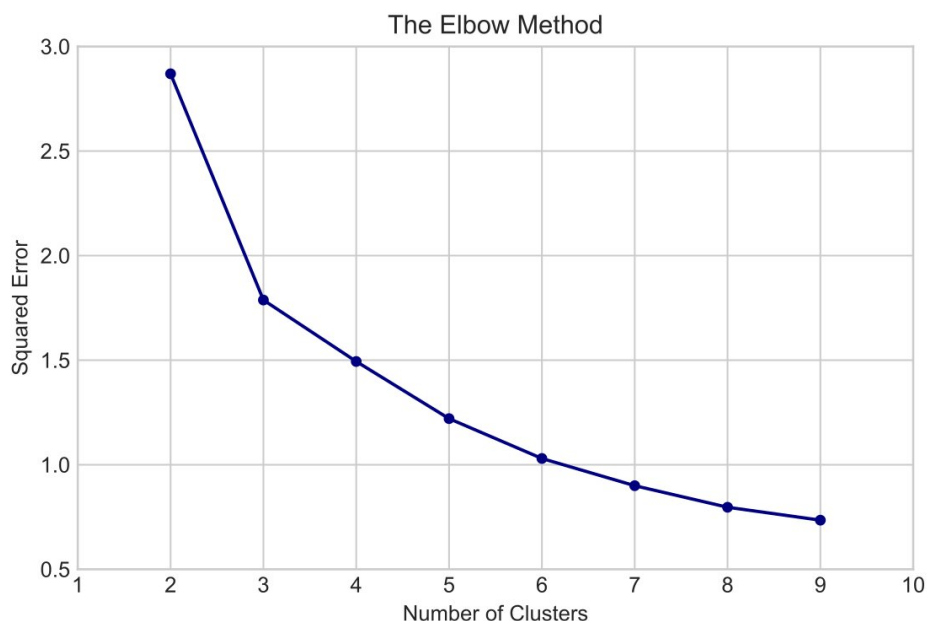


Figure 6.2: The Elbow method applied at the sentiment-related variable showing the variance of Squared-error in terms of the number of clusters (k).

runs independently for different k values as if no domain-knowledge was available. Since that is not the case, and the objective of this emotional representation is to have clinical relevance - as real data from the rehabilitation procedures is being employed -, again, domain expertise will be required to select the most meaningful number of partitions.

As above referred, the last parameter that requires user definition is the cluster initialisation. It is known that different initialisation can result in dissimilar clustering as K-Means only converges to local minima [228]. To overcome this limitation, K-Means will be run multiple times, resulting in the same number of partitions of which it is selected the one with the smallest squared error. An analysis of the arrangement of the centroids (after the selection of k) according to the initialisation will be present in the next section. Also, to reduce the sensitivity of K-Means to the first centroids definition, a smarter initialisation is ensured by k-Means++. By initialising centroids that are far from each other and not completely random, it speeds up the convergence and improves the quality of the clustering.

6.1.1.1 Results & Discussion

Since we are not familiar with how many clusters of patients are present in data, the choice of representation must be guided by the domain knowledge. As it is of interest to obtain interpretable partitions, a constraint has been provided by clinical and medical experts, suggesting that the number of clusters should never exceed four. Any division should segment the type of patients by the sensations felt during sessions and by the satisfaction with the product and, for their expertise, any higher number could form too specific groups that do not generalise typical patients behaviour. Bearing it in mind, only two partitions will be presented in this section. The

domain expertise will be used to validate the significance of the created groups, as no ground truth could be used to compute other performance metrics. These experiments were performed only on the training data; thereby, the sessions of the group of patients that comprise test dataset will be later transformed according to the found centroids.

The following figure (Figure 6.3) present the results obtained by the application of the algorithm described during 30 times - all of them run until convergence is reached (no changes in data points assignments). The partitions that had resulted in a minimum within-cluster squares sum are selected. These are, also, the results that were presented to the medical professionals of SWORD Health. Complementary figures may be found in Appendix B.1.

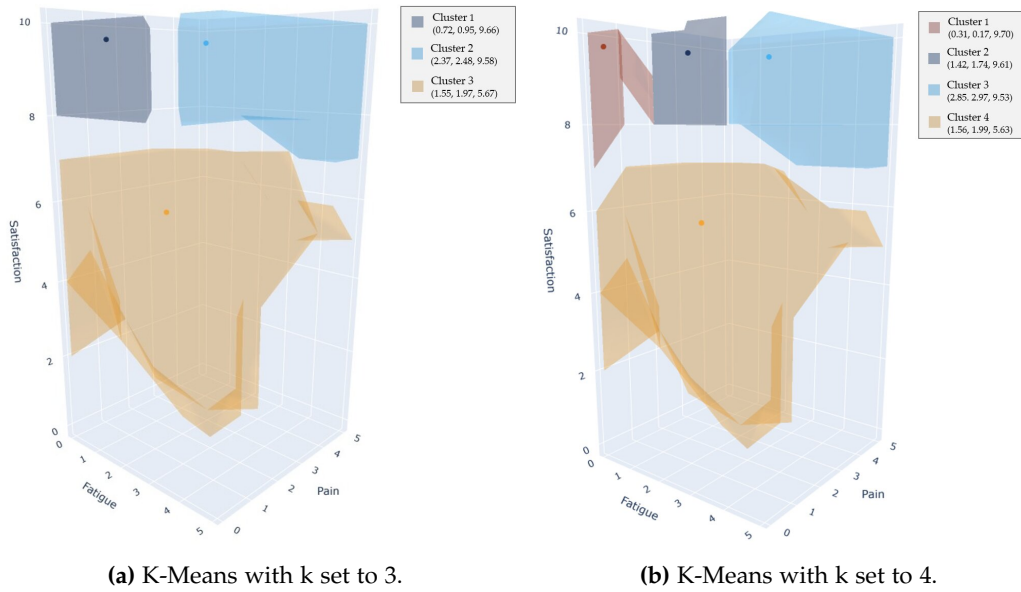


Figure 6.3: Results obtained by the application of K-Means on the sentiment-related variables, aiming to find a emotional-state representation. Legend follows the scheme: Cluster *label*, and cluster centroid as (*fatigue*, *pain*, *satisfaction*)

The first main takeaway from this figure is the resemblance between 2 pairs of clusters: cluster 2 of Figure 6.3a and Cluster 3 of 6.3b and, cluster 3 of Figure 6.3a and cluster 4 of Figure 6.3a. The difference found when comparing the two representations lies in a specific region of the 3D-representation that corresponds to a set of data points of low values of pain/fatigue and high values of satisfaction. The addition of a new cluster results in a subpartition of that region, generating an extra cluster with high satisfaction values. The mathematical rationale for this division lies in the strong share of sessions that are present in that area (as may be confirmed in Section 4.2). Then, the maximum decrease of inertia (the minimisation of the objective function) that may be reached when creating an additional cluster occurs, as expected, in that section. Even though cluster 3 (of K-Means with K=3) covers the broader region, the amount of points - as will be verified shortly - is lower when comparing to the clusters comprehending high satisfactory levels; thus, the effect of sub-partitioning that area in two, based on the other two variables, would be more residual.

Looking at the centroids, others relevant observations may be done: Cluster 3 of Figure 6.3a and Cluster 4 of Figure 6.3b have practically the same values; at the same time, Cluster 3 of the

representation with 4 clusters suffers a slight shift towards higher values of pain/fatigue when compared with K-Means with 3 clusters, consequence of the addition of a new cluster with a nearer centroid - Cluster 2 of Figure 6.3b.

At this moment, the point of view of the domain-experts was asked in order to understand which of the representations was better suited to all the types of patients by them identified. In their perspective, the subdivision of Cluster 1 is not justified, as the two partitions created have no practical differentiation when approaching patients with those impressions during rehabilitation. Although mathematically explainable by the cost function, both segment patients with almost no sensations of pain/fatigue and satisfied with the undergoing treatment, not requiring special attention by clinical professionals nor intervention. That said, the clusters created by K-Means with k equal to three were the one selected by the clinical team and that presents more clinical relevance.

Following the technical know-how of the domain experts, the clusters may be interpreted as follows:

- **Cluster 1:** Group created by sessions of patients who are experiencing irrelevant levels of pain and fatigue at the end of the sessions and demonstrate to be satisfied with the rehabilitation program powered by SWORD Health. Patients show comfort with the therapist prescriptions, able to perform the majority of inquired tasks. According to the professionals, such range of responses are associated with patients noting progress in their physical condition and, this way, keeping motivation at high levels. This partition comprises the patients who are not dependent on home-interventions or any specialised remote assistance.
- **Cluster 2:** Unlike the previous cluster, this group comprises sessions of patients who experienced some or a lot of pain/fatigue during the session. This sentiment-related levels may be caused by an intrinsic condition of the patient that, although motivated to progress, continues impressing some pain and/or fatigue. These sensations may be associated with an overload of inquired exercises, so, an action should be taken in order to understand if any prescription change would be appropriate. On the other hand, satisfaction proceeds with high values which may be explained by intrinsic motivation and will to recover from the actual condition and, to achieve that, controlled levels of pain and fatigue are accepted. Nevertheless, as the most extreme values of pain and fatigue are contemplated in this region, this group of patients may need particular attention if these levels persist or if no progress is observed.
- **Cluster 3:** Cluster created based only on satisfaction values. Particularly by the Figure B.1.1 present in Appendix B.1, it may be noticed that there are reported values in both extremes of pain and fatigue and, for that reason, no conclusions may be made regarding the mentioned affections. Therefore, this representation will be associated with not pleased patients what *per se* comprises considerable interest. Satisfaction is seen as an indicator of the quality of patient care and directly associated with patient's motivation. Keeping track of the patients present in this group is of extreme relevance and further emphasised when unveiled that high satisfied patients tend to show better outcomes and higher-adherence rates. The lack of motivation shown by the patients may have different reasons as intrinsic characteristics (not feeling stimulated by the desire of improving his physical state); pa-

tients may be sceptical to the rehabilitation program they are engaged; or, simply, having a different expectation about the treatment and his progress that is not being accomplished. It should be noted that the class combines both satisfaction values of individuals considered *detractors* but also a few *passive* (Section 2.3.4.1 to remind the terminology) as answers with value 7 may be observed.

After the clarification about the nature of the sessions, patients associated to each cluster and assured the relevance of distinguishing the three classes, these results will be, from now on, referred as the emotional-state representation. The intervention of the clinical team was fundamental in the validation of the unsupervised learning technique applied upon the sentiment-related variables.

Similarly to the analysis carried out in section 4.2 and for the reasons already outlined, it is important to analyse the distribution of the data among the three clusters. So Figure 6.4:

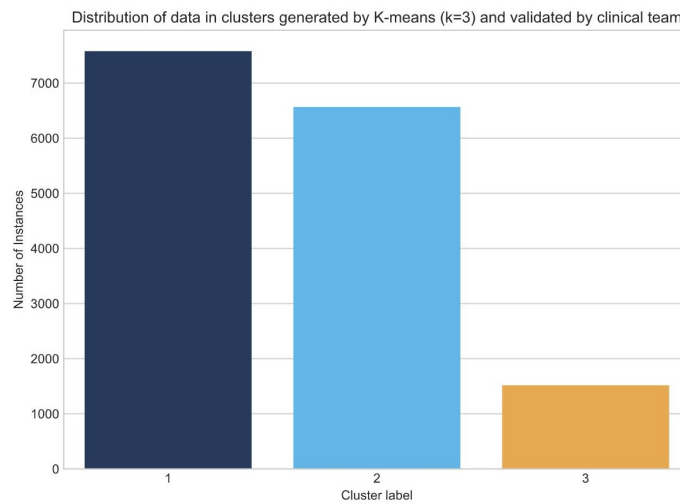


Figure 6.4: Distribution of data in clusters generated by K-Means ($k=3$) and validated by the clinical team, from now on, the emotional-state representation.

As first remark, the analysis of the emotional-state representation allows confirming the observation made during this section that the maximum share of values should be part of Cluster 1. It is now clear that even though cluster 3 covers a broader region, the algorithm benefits from adding a new cluster on the top part of the 3D-representation given the total amount of data distributed amongst cluster 1 and cluster 2. Cluster 3 presents not only the smallest number of samples but also the wider region what may be associated with an enormous variation in data points, and consequently, low coherence in features values, not representative of the concerned class.

Moreover, we are once again facing an unbalanced dataset. The distribution of the satisfaction values created an unequal distribution of data, as the cluster whose patients do not show to be satisfied with the product includes significantly fewer instances. Again, imbalanced learning techniques should be considered in order to avoid failing the recognition, principally, of the minority class.

Combining the analysis of the distribution of the emotional-state label created and the 3D

representation of Figure 6.3a, there is another observation to be made regarding the scale of the employed variables. Given the short amount of data with low satisfaction values, it is believed that if the same scale for all variables had been used during these tests, a partition that took into account the satisfaction data would not have been obtained. The option to maintain the original scale proved to be favourable, counteracting the skewed distribution of data points and empowering the distinction of a set of patients not satisfied with SWORD Health's product; otherwise, only the two variables collected every session would have been used in the clustering process. Even though there is no distinction among those patients regarding the pain/fatigue felt, the remarkable importance of having a class comprising non-satisfied (or at least, non-promoters) patients was confirmed by the health professionals. The decision beforehand made, therefore, arose as the most opportune and with recognised clinical pertinence.

Furthermore, the influence of the initialisation may be analysed. As aforementioned, the algorithm was run for 30 iterations and the arrangement that minimises the cost function (Equation 6.1) chosen. Figure 6.5 shows the positioning of the cluster centroids over all iterations.

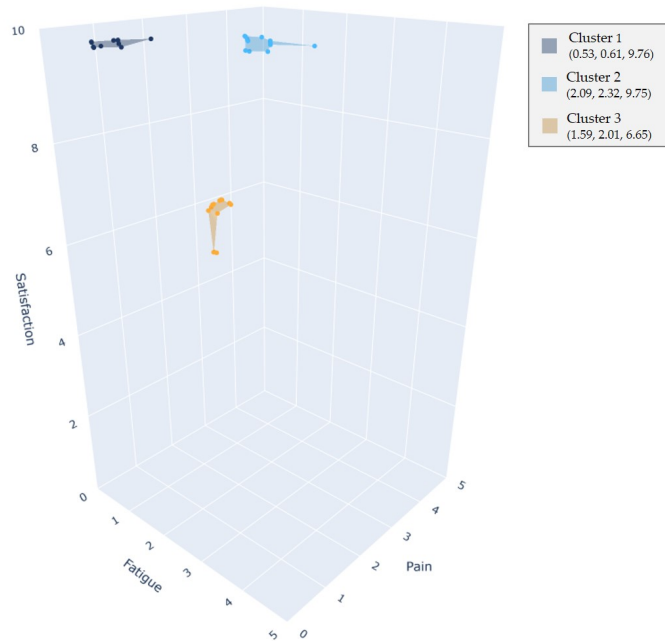


Figure 6.5: Analysis of the initialisation effect in K-means: representation of cluster centroids over the 30 iterations. Legend follows the scheme: Cluster *label*, and median of the 30 found centroids, as (*fatigue*, *pain*, *satisfaction*).

Evaluating the difference between the median of the cluster centres over the 30 repetitions (presented in Figure 6.5 legend) and the clusters centres of the emotional-state representation to be used, an average of 8.70% of relative deviation is noticed. Moreover, by the analysis of the figure above, it is possible to affirm that all the iterations result in similar cluster centres - for the 3 clusters, their centroids are placed at nearly the same region with tiny displacements. It may be a consequence of the K-Means++ initialisation that tries to minimise the effect of the first initialisation, aiming to guarantee the best possible algorithm performance in all iterations. Nevertheless, a short deviation may influence several data points assignments (mostly at the upper volume of 3D-representation given the greatest amount of points); therefore, running the

algorithm for consecutive iterations and output the best result in terms of inertial seems to be a proper and justified approach.

6.1.2 Non-parametric clustering

In opposition to traditional clustering methods, there have been several approaches from a statistical point of view in which clusters are modelled according to specific distributions. Soft-clustering via expectation maximisation (EM) acts as one of the most commonly employed and was proposed in [232], modelling clusters as Wishart distributions. However, this algorithm requires specifying the number of clusters and, therefore, presenting the same drawback as the classic clustering algorithms. Thus, an approach of this unsupervised problem from a non-parametric Bayesian perspective was studied, with the aim of employing a solution in which the number of components does not need to be fixed upfront and can be dynamically updated.

Bayesian Non-parametrics models are a class of models with a potentially infinite number of parameters, allowing flexible models in various configurations. These models, as it is the case of the Dirichlet process mixture [233], result in infinite mixture models which do not fix the number of clusters in data *a priori*. However, this class of models requires sampling or variational inference methods, generally hard to implement and not easily scalable, compared to more straightforward methods as K-Means.

A combination of the advantages of each approach (K-Means hard-clustering and Dirichlet processes) was designed by Kulis and Jordan in [234]. The authors had proposed a "scalable hard clustering algorithm from a Bayesian non-parametric viewpoint". This way, motivated by the link existing between K-Means and mixtures of Gaussians (k-Means may be seen as the limit of the expectation-maximisation algorithm [234]), the new algorithm makes use of the Dirichlet process mixture model, performing a similar limiting argument as part of a simple sampler - Gibbs sampler. This had resulted in an algorithm with hard cluster assignments, identical to the classical k-Means, excluding the fact that a new cluster is only created when a point is enough far away from all current cluster centres. To avoid overfitting, a penalty for the number of clusters is included in the k-Means-like objective that the algorithm aims to minimise.

The so-called DP-Means (originally, Dirichlet Process- means) algorithm is now detailed, based on [234]:

Input: $x_1, \dots, x_n$: input data, λ : cluster penalty parameter.

Output: Clustering set $L = l_1, \dots, l_k$ and number of clusters k .

1. Initialise the first cluster: $K = 1$; $l_1 = x_1, \dots, x_n$ and μ_1 the global mean.
2. Initialize clusters indicators $z_i = 1$ for all points ($i = 1, \dots, n$).
3. Repeat until convergence (no changes in any data point assignment):
 - For each point x_i :
 - Compute squared euclidean distance: $d_{ic} = \|x_i - \mu_c\|^2$ for all centres $c = 1, \dots, k$.
 - If $\min_c d_{ic} > \lambda$, create new cluster: $k = k + 1$, $z_i = k$, and $\mu_k = x_i$.

- Else, set $z_i = \operatorname{argmin}_c d_{ic}$.
- Update clusters $l_1, \dots, l_k$ based on the minimum distance between each point and cluster centroids, this is, $z_1, \dots, z_k : l_j = x_i | z_i = j$.
- For each cluster l_j , compute mean, $\mu_j = \frac{1}{|l_j|} \sum_{x \in l_j} x$.

To note that although the first cluster is initialised with the mean of all points, DP-Means depends on the order in which data points are processed, in opposition to K-Means. The parameter λ acts as a regularisation parameter, controlling the rate at which new clusters are created. Thus, the objective function (Equation 6.2) of the algorithm DP-Means may be defined (where μ_c is the centroid of each cluster):

$$J(L) = \min_{\{l_c\}_{c=1}^k} \sum_{c=1}^k \sum_{x \in l_c} \|x - \mu_c\|^2 + \lambda K \quad (6.2)$$

This function is simply the k-Means objective function with the additional penalty based on the number of clusters, λ . It is now possible to analyse the performance of DP-Means to be compared with classic clustering approach of Section 6.1.1.

6.1.2.1 Results & Discussion

The following figure (Figure 6.6) presents both the influence of the regularisation parameter on the number of clusters and the behaviour of the objective function in terms of the independent variable.

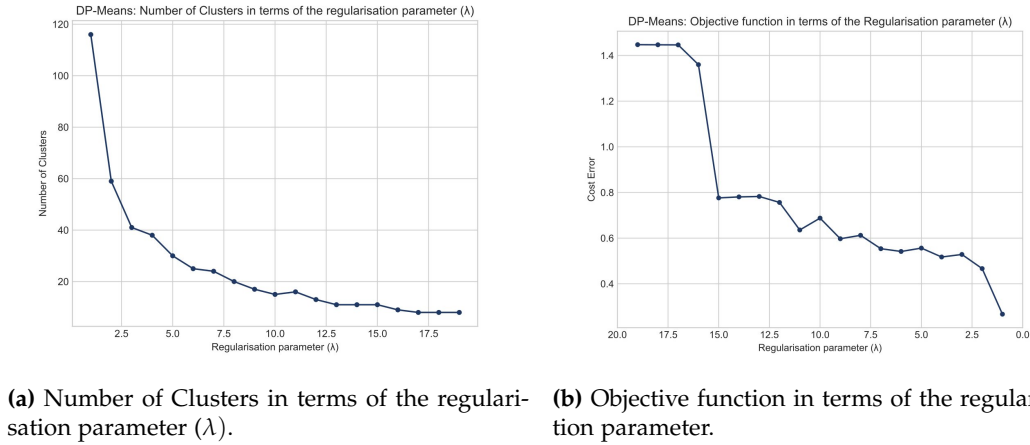


Figure 6.6: Influence of the regularisation parameter on DP-Means results.

In the figure above, it is possible to see the non-parametric clustering applied to the sentiment-related variables in order to form groups representative of the emotional-state. As may be confirmed, the strength of regularization parameter lambda controls the number of clusters created and as expected, the number of clusters decrease when the maximum accepted distance between a point and a cluster centre increases. Figure 6.6a shows a monotonically decreasing function where an elbow may be identified (between 2 and 3) - the number of clusters decreased markedly

for an increase in λ , point that corresponds to the emergence of clusters according to [235]. It is essential to highlight that the minimum possible λ was set at one since it corresponds to the minimum distance between points (as we are dealing with discrete variables). Any decrease of the regularisation parameter, regardless of how diminutive, would set the number of clusters to the number of unique data points.

Moreover, Figure 6.2 enables the author to examine the effect of the regularisation parameter (and consequently, the number of clusters) on the objective function. Although the objective function comprises an additional penalty for the number of clusters, the behaviour of the curve is quite similar to the one found in K-Means (Figure 6.2). The existence of the penalty term produces small increases in the cost error, notwithstanding the increase in the number of clusters, phenomena not possible in K-Means. Once again, the Elbow method already described in Section 6.1.1 could be employed, this time, to define the regularisation parameter, that is, the maximum distance a data point can be to the centre of the cluster is assigned. It may be noted that the cost error decreases the most when the regularisation parameter reaches the value of 15 and an elbow is formed in the graph. The produced results with the non-parametric clustering method with this optimum value of λ (15) are displayed in the following figure - Figure 6.7.

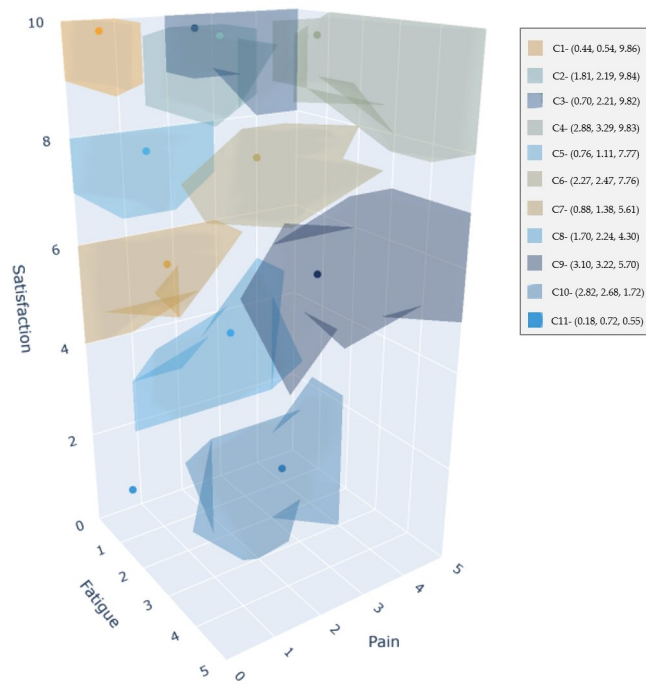


Figure 6.7: Results obtained by the application of DP-Means on the sentiment-related variables, aiming to find an emotional-state representation with regularisation parameter set at 15. Legend follows the scheme: Cluster *label*, and cluster centroid as (*fatigue*, *pain*, *satisfaction*).

From Figure 6.7, it is observed that the application of the algorithm with lambda set as 15 gave rise to the total number of 11 clusters. Despite being the ideal number resulting from the combination of the non-parametric clustering methodology and the Elbow method - which assures the maximum decrease of the cost function - such approach does not result in interpretable groups, not being possible to associate the clusters formed to the same number of patients seg-

mentation. Very specific groups of patients that do not generalise regular patients behaviour and whose clinical relevance cannot be guaranteed have been obtained. Moreover, the existence of a cluster created only by a unique value - cluster 11 - does not act as a desirable behaviour in order to find an appropriate representation of the emotional-state.

Notwithstanding, the need to specify the number of clusters K ahead of time was eliminated and all the linked methods used to find the optimal number of clusters. Several studies affirm that ensuring a minimum separation between cluster centroids leads to higher quality clusters [235]; however, the underlying distribution and context knowledge may once again be taken into consideration in order to set λ adequately and search for an adequate number of groups given the known patients' classes.

Thus, in order to enable a direct comparison between the clusters generated by both algorithms (K-Means and DP-Means), it was found the value of the regularisation parameter that originates three partitions in sentiment-related data. This value corresponds to approximately 30 (dependent on the order of the data), and the obtained results are displayed in the following figure (Figure 6.8, complementary figures in Appendix B.2). Although fixing the number of clusters was not mandatory, in this following analysis, it was fixed for comparative purposes.

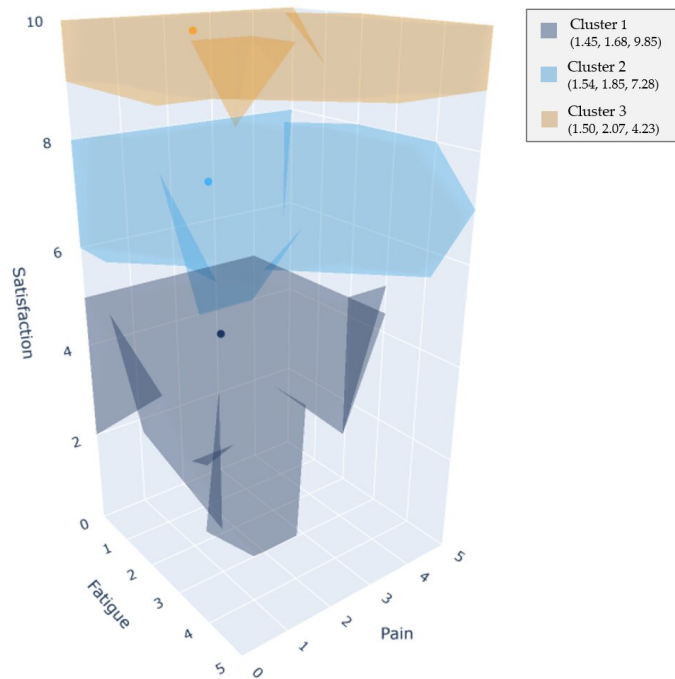


Figure 6.8: Results obtained by the application of DP-Means on the sentiment-related variables, aiming to find an emotional-state representation with number of clusters fixed at 3. Legend follows the scheme: Cluster *label*, and cluster centroid as (*fatigue*, *pain*, *satisfaction*).

Looking at the 3D-representation above, the first and more important observation is that the algorithm practically has only considered the satisfaction values. As DP-Means does not reflect the distribution of the data - this is, only a point is sufficient to generate a new cluster - the acknowledged great share of points at high satisfaction levels do not have any effect. The fact the data was employed with variables on different scales proved to be non-beneficial for this

non-parametric approach as the groups were organised accordingly to one and only variable - satisfaction. The application of the variables using the original scale was one of the suggestions provided by the clinical team; despite that, the partitions created showed no clinical relevance as they are essentially defining the typical NPS groups (*detractors*, *passive* and *promoters*).

A pair of complementary analysis may be found in appendix B.2: distribution of data in clusters and the influence of the order of the data points. The first one confirms what was already expected given the known distribution of satisfaction levels - pretty skewed towards high-levels. This way, this grouping approach had resulted in an extremely unbalanced dataset.

Moreover, as already referred, unlike standard K-Means which depends on the initial clustering of the data, DP-Means depends on the order in which data points are processed. Figure B.2.4 represents the different placement of the centroids resultant of a series of applications of the algorithm on the data with shuffled order. It is important to note that the algorithm was run for 30 iterations and similarly to the strategy designed in K-Means clustering, the partition that minimised the objective function was selected. Perhaps surprisingly, the influence of the order was not that significant as the clusters centres are placed realistically always in the same region, and the algorithm is not relying on the order of the input sequences. Jiang *et al.* [236] mention that DP-Means is considerably robust and not sensitive to the order of the input data because the model adjusts the centre of each cluster after the assignment step and repeat that until convergence - achieving a more accurate estimation of the cluster centre.

As a conclusion, during this section, it was demonstrated the possible application of an algorithm that combines Dirichlet Process mixture models and hard clustering algorithms. The classical k-Means algorithm is revised from a Bayesian Non-parametric perspective and applied to the underlying data, resulting in a different representation of the emotional-state. This experience objective did not include any clinical validation, having been carried out only for comparative purposes.

6.1.3 Comparative Analysis of Clustering Methodologies

The main differences and considerations regarding the results of two clustering methods employed may be summarised according to some specific topics as follow:

- **Data Features Scaling:** In both approaches, the preference of the clinical team of maintaining the original scale of the variables was followed. However, the consequences of such a decision proved to be distinct when comparing the results obtained by the two algorithms. For K-Means clustering, the scale was beneficial to counteracting the unbalanced data distribution and, thus, originate, a partition in which patients not satisfied were distinguished. On the contrary, regarding DP-Means (with number of clusters defined), the results show that only the variable with a broader scale (satisfaction) was considered in the clustering process, allowing to conclude that scaling the input sequences in advance would have been advantageous.
- **User-Specified Parameters:** The primary motivation behind the application of non-parametric clustering is to avoid pre-defining the number of clusters. Although this has been achieved through DP-means algorithm, a regularisation parameter (maximum distance of a point to its cluster centroid) has, anyhow, to be set ahead. Consequently, domain knowledge

should be considered regardless of the selected approach either to define the number of clusters or the regularisation parameter λ . During this work, SWORD Health clinical team was asked its position regarding this matter, more specifically, the ideal number of clusters that better define the known type of patients. If this were not the case, empirical methods should have been required to set the optimal values of these user-specified parameters. Moreover, both methods required the definition of a distance-based similarity metric; during this work, the Euclidean distance was picked.

- **Data distribution in the clusters formed:** The emotional-state representation carried out for both algorithms had created, once more, an uneven distribution of data. Nevertheless, at K-Means, the majority of data is distributed along with the top clusters, resulting in a minority class. In contrast, in the non-parametric approach, this imbalance is even more pronounced given that most of the data is present only at the class comprising high-satisfied patients as it follows satisfaction distribution (review Figure 4.4c).
- **Initialisation effect:** K-Means depends on the initialisation of the clusters in the first stage of its algorithm. During this section was analysed its influence on the final results, having been assured that its effect is reduced - running the algorithm for 30 iterations and selecting the one that minimises the cost function. Regarding K-Means, the iterative process depends on the order of the input sequence. This effect was also analysed and counteracting following the same logic. In any case, it was demonstrated that in both approaches, all the iterations had resulted in similar cluster centroids.
- **Clinical Relevance:** Since the clinical team provide their suggestions during the clustering process and had interpreted the groups formed, it may be affirmed that K-Means resulted in an emotional-state representation with clinical relevance. All the process was verified and explained to the SWORD Health's clinical team and, consequently, the unsupervised learning technique validated. The clinical interest in the inference of the emotional-state based on the created partitions was this way, assured. The non-parametric method was performed with comparative intentions only, with no need for domain-experts viewpoints on the matter.

6.2 Emotional-state Classification Task

After the combination of the user input variables and found the emotional-state representation, it is possible to design proper algorithms to predict this latest target. We are now facing a typical supervised learning problem, similar to the one embraced in Chapter 5. As the emotion representation validated by the health-professional team comprises three classes, the problem rests on a multi-class classification task. The set of behavioural features extracted and detailed in Chapter 4, may once again be employed in order to make use of all the information acquired during a rehabilitation session.

A similar experiments workflow was followed, also keeping the same training/testing division. It is important to refer that the identification of clustering centroids was carried out resorting only to the train set and the assignments of data points in test set performed afterwards.

Following the same rationale of previous experiments, F1-score with unweighted average was selected as the preferred metric to employ.

A set of pertinent models already detailed throughout Chapter 5, from baseline models to models with the application of imbalance learning techniques, were tested during this classification task. The features were prepared according to each model following the principles of the previous Chapter. For the tree-based ensembles, correlated features are removed *a priori* in order to avoid inaccurate interpretations of features importance and reduce data dimensionality (consequently, avoiding pointless computational-costs). Regarding the application of the SVM classifier, either features are transformed using PCA or previously selected by the Recursive Feature Elimination technique). Finally, for the Deep Learning approach, no transformations or feature selection stage is included given the inherent characteristics of the model employed.

Given the inequality demonstrated among classes instances (Figure 6.4) and the existence of a minority class that appear to be (as typical) the class of most interest, the imbalanced learning techniques discussed in Section 5.6 will be again employed in the current task. Both cost-sensitive learners and data-balance approaches (equalising at minority class or at mean) are tested, and the results will be presented next section. Bayesian Optimisation relying on the Stratified Group 8-fold was used to optimise each single classifier based on the parameters pointed out on the respective session of Chapter 5 (Random Forest Classifier Table 5.1, Gradient Boosting parameters on Table 5.2, SVM-RFE considered values at Table 5.3, SVM-PCA Table 5.4, and finally MLP classifier attributes in 5.5).

6.2.1 Results & Discussion

After the hyperparameter tuning and best model training on the whole training set, the performance of the different models in the same classification task may be compared. The following figure comprises not only the results of model evaluation on testing set but also the outcomes of the Cross-validation 8-fold (average value displayed) used in parameter optimisation (Figure 6.9). Again, it is possible to find in the appendix (Appendix B.3) complementary results: direct association between testing and validation results with reference to the selected parameters and, confusion matrices of each classifier organised by the imbalanced learning technique employed (or the baseline approach in which none is used).

By the analysis of both figures, the first observation that emerges is that results are not as satisfactory as necessary to affirm that the emotional-state is being inferred through the pipeline implemented. In order to enable clinical application and distinguish patients based on the representation found in last section, higher predictions rates are mandatory.

From both matrices, it is noticed the low variation between the performance of all implemented classifiers - as was already the case in sentiment-related variables classification. F1-scores unweighted averaged values consistently fall around 36-40% in test set, regardless of the model, showing that the models are not extracting relevant information from data. Data proved not to be effective on this emotional-state prediction, calling for additional inputs to achieve the final goal and provide relevant results to the clinical team.

The scores obtained are coherent with those obtained in multi-class classification of Chapter 5 as best scores tend to assume values in the same range (around 40%). The justification for

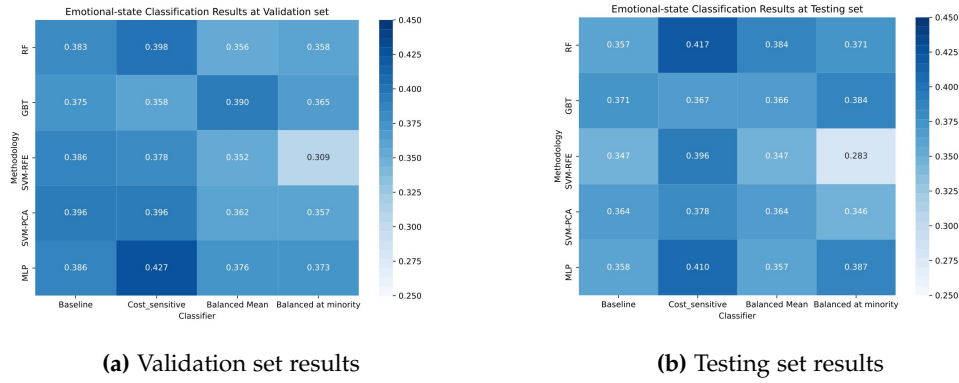


Figure 6.9: Results of the multi-class classification emotional-state representation. Y axis refer to the different classifiers employed, while x-axis to the methodology. Validation results are reported with the average value.

this observation lies on the fact that this problem target is conceived based on a combination of the former variables (pain, fatigue and satisfaction) and the dataset and features employed are also the same. Given the inaccurate distinction between the different levels of satisfaction in the previous chapter and its crucial role in the definition of the class of most interest in the representation created (cluster 2, formed only on the basis of satisfaction values), an identical result was already expected.

As for the effect of training with imbalanced-learning techniques, global scores do not present significant improvements; nevertheless, the models designed that obtained the highest results correspond to cost-sensitive RF and MLP. The unbalancing of data has again proved not to be the critical obstacle that prevents models from achieving higher performances and, reveals that other base problems may exist; either associated with the data employed, the feature extraction stage or the relativeness of emotional-based targets.

Despite that, by the interpretation of the confusion matrices present in Appendix B.3, there are significant differences between the number of minority class predictions applying imbalanced learning techniques or not. In the baseline classifiers, it is noted that class 2 (minority class) is rarely predicted, showing that classifiers are not learning to recognise this class, due to the low share of instances. When employing the mentioned techniques, the number of true positives increases substantially (mainly in cost-sensitive learning and balance at minority class). Along with this fact and as expected, the false-negative rate of class 2 also increases. It may be argued that from a clinical point of view, although the global score is similar, the performance resulting from the utilisation of these techniques is preferable, given the higher identification of non-satisfied patients - which can lead to clinical-specialised interventions. The benefits of these identifications are superior to an eventual cost of approaching patients not actually requiring (in false-positive situations).

Finally, overfitting issues were not revealed from Figure 6.9 as validation and test scores are not dissimilar. Despite the unsatisfactory results, it may be affirmed that model generalisation was guaranteed.

6.3 Summary

The experiments conducted over the course of this chapter allowed the creation of a quantitative label representative of the patient's emotional-state. The sentiment-related variables (user input variables) have been combined through the application of a pair of clustering algorithms: the classic K-Means and the non-parametric method based on the Dirichlet Process, DP-Means.

Regarding the parametric method, SWORD Health's clinical team was essential throughout the process, providing valuable insights on different issues as the scale of variables to use or the preferable number of clusters. Domain-expertise was again required for the validation of the formed clusters and fundamental to guarantee the clinical relevance of identifying the various partitions generated. Thus, the emotional-state representation emerged. Besides, the distribution of data in clusters and the effect of the initialisation step were analysed, enabling to justify some of the decisions previously made.

For comparative purposes, DP-Means method was applied upon the sentiment-related data without any special domain-knowledge and avoiding the pre-definition of the number of clusters. The results obtained with both the optimal regularisation parameter set and the number of clusters defined are displayed, allowing a different examination. The influence of the order in which data instances are processed, as well as the distribution of classes, were also object of analysis.

Finishing the chapter, proper algorithms are designed aiming to classify the established emotional-state representation. The imbalance of the data turned out to be again the most significant obstacle preventing the models from achieving more promising results.

Chapter 7

Conclusions and Future Work

Millions of people worldwide are affected every day for Musculoskeletal Disorders. The high prevalence and incidence are even emphasised if measured in the impact on patients, as it limits their workability, causing mental and behavioural issues and having financial repercussion. Physical therapy is widely promoted and acts as the standard of care addressing this kind of injuries. Given the limitations of conventional physical rehabilitation, namely high dependence on human resources and patient discomfort due to travelling inconveniences and schedule incompatibilities, home-based therapy is emerging as a preferable solution.

Despite the better clinical outcomes achieved with digital rehabilitation, this kind of approach lacks an emotional component. Rehabilitation therapists should be conscious of patients' emotional-state during treatment, keeping track of pain and fatigue levels throughout an exercise session. This awareness of patients reactions was seen both in several works and for SWORD Health's clinicians as an essential prerequisite for a successful treatment. This Master's Thesis provided the basis for the development of an adaptive feedback algorithm dependent on patient's emotional-state during the recovery process.

Working towards making this adjustable algorithm readily available, the current thesis focused on the development of an emotional-state classification system based on a data-driven approach. While enabling patients to perform their rehabilitation therapy from home, offering an efficient way to treat Musculoskeletal Disorders, SWORD Health's system allows the collection of data for all the patients on-board in the rehabilitation program, comprehending the basis of the implemented framework. Its development involved statistical analysis of the dataset, pre-processing and cleaning routines, the definition of the patients' emotional-state representation and use of those to implement ML techniques for classification. The interpretation and comparative analysis of results in order to understand the real potential of them are also part of the work contributions.

The developed review work accomplished important insights into the techniques typically employed to deal with affect detection. The literature revealed the reliability of the application of several input channels; yet deficiencies were identified: most of the techniques used input information not available in SWORD Phoenix (audio, text and even visual) or the systems developed were dependent on particular settings. Such approaches lack on generalisation and could not be applied during a rehabilitation program. In anytime the thesis' pathway could be disso-

ciated from the product requirements of the SWORD Health system and disregard technology limitations. It is the author's firm conviction that those product requirements led, in the final analysis, to significant scientific work, opening different perspectives regarding affective-state detection through a data-conducted methodology.

Various difficulties arose from the lack of previous exploration of the dataset and the use of data in AC. The primary one encountered was the impossibility to make performance comparisons with prior art as no benchmark dataset was employed, but real-data from the rehabilitation procedures. The next obstacle was the nonexistence of a validated qualitative or quantitative measure of patients emotional-state. Therefore, the different variables inserted by the patient were explored as the representative target, having been followed two distinct paths: assessment of these variables (pain, fatigue and satisfaction, named sentiment-related variables) on individual basis, or combining them intending to create a tag of the emotional-state.

Regardless of the subsequent application, the extraction of relevant features to the predicting model was a fundamental step. The extraction of proper features demands domain knowledge and *a priori* understanding of the dataset, an essential fact considered in this stage. Careful database analysis and comprehension was fundamental to gain acquaintance with the data, confirming the potential of the utilisation of behavioural features.

As for the sentiment-related variables classification part, a set of classifiers and feature selection approaches are tested, aiming to obtain baseline performances. Given the imbalance of data and the inherent order between classes, there was a clear opportunity for the application of imbalanced learning and ordinal classification techniques, methods that enabled the accomplishment of the best results. Although requiring further refinement, the performance of LSTM models was very encouraging due to the relatively little investment in specific feature engineering work, and the absence of combination with balancing techniques.

Concerning the emotional-state inference, a pair of clustering methods was applied to the sentiment-related variable with the purpose of finding a possible representation of patient's affective state. Context-based knowledge and SWORD Health's clinical team were essential throughout the process and enabled the validation of the novel representation, assuring clinical relevance and the interest of distinguishing patients according to those labels. With that in mind, ML models are tested, making use of the baseline models designed and the balancing techniques.

Notwithstanding the classification target, there is still a considerable gap between its current state and a production-ready state. The fact that the targets are dependent on patient's emotional-state with different viewpoints on the scale used for obtaining the values of the sentiment-related variables may be preventing the models from reaching higher results. Given that a considerable number of models were tested and most of them did not provide favourable results, it may be suggested that the main difficulty may rely on the extremely subjectivity of the targets and the nature of the problem that aimed to be solved.

Future work comprises the possibility of transforming target variables, making them only relative to each patient or, perhaps, dependent on the previously recorded labels. Moreover, other suggestions include the implementation of a system that could work on real-time, instead of using all information acquired during a session for end-sessions prediction; or explore anomaly-detection-specific methods. The refinement of particular models should be considered as occasionally referred throughout the dissertation. Furthermore, based on the literature re-

view, questions can be raised as to whether the addition of audio or visual data sources could improve global results.

Bibliography

- [1] A. Woolf and B. Pfleger, “Burden of major musculoskeletal conditions”, Tech. Rep.
- [2] D. Van Eerd, C. Munhall, E. Irvin, D. Rempel, S. Brewer, A. J. Van Der Beek, J. T. Dennerlein, J. Tullar, K. Skivington, C. Pinion, and B. Amick, “Effectiveness of workplace interventions in the prevention of upper extremity musculoskeletal disorders and symptoms: an update of the evidence”, 2015. DOI: [10.1136/oemed-2015-102992](https://doi.org/10.1136/oemed-2015-102992).
- [3] E. A. f. S. at Work and Health, *Work-related musculoskeletal disorders in the EU - Facts and figures*, 11. 2010, vol. 60, pp. 366–373, ISBN: 9789291912612. DOI: [10.1007/BF03344315](https://doi.org/10.1007/BF03344315).
- [4] O. S. Administration and Health, *Prevention of Work-related musculoskeletal disorders*. [Online]. Available: https://www.osha.gov/pls/oshaweb/owadisp.show_document?p_id=4481&p_table=UNIFIED_AGENDA (visited on 11/29/2019).
- [5] L. March, E. U. Smith, D. G. Hoy, M. J. Cross, L. Sanchez-Riera, F. Blyth, R. Buchbinder, T. Vos, and A. D. Woolf, *Burden of disability due to musculoskeletal (MSK) disorders*, 2014. DOI: [10.1016/j.berh.2014.08.002](https://doi.org/10.1016/j.berh.2014.08.002).
- [6] A. D. Woolf, “Global burden of osteoarthritis and musculoskeletal diseases”, *Musculoskeletal Health in the 21st Century*, 2015.
- [7] *Large Employers’ 2020 Health Care Strategy and Plan Design Survey - National Business Group on Health*. [Online]. Available: <https://www.businessgrouphealth.org/benchmarking/survey-reports/2020-report/> (visited on 11/29/2019).
- [8] O. Babatunde, J. Jordan, D. Van Der Windt, J. Hill, N. Foster, and J. Protheroe, “Effective treatment options for musculoskeletal pain in primary care: A systematic overview of current evidence”, 2017. DOI: [10.1371/journal.pone.0178621](https://doi.org/10.1371/journal.pone.0178621). [Online]. Available: <https://doi.org/10.1371/journal.pone.0178621>.
- [9] S. M. Mcphail, M. Schippers, A. L. Marshall, M. Waite, and P. Kuipers, “Perceived barriers and facilitators to increasing physical activity among people with musculoskeletal disorders: a qualitative investigation to inform intervention development”, *Clinical Interventions in Aging*, 2014. DOI: [10.2147/CIA.S72731](https://doi.org/10.2147/CIA.S72731). [Online]. Available: <http://dx.doi.org/10.2147/CIA.S72731>.
- [10] T. Kuijpers, M. Van Middelkoop, S. M. Rubinstein, R. Ostelo, A. Verhagen, B. W. Koes, and M. W. Van Tulder, *A systematic review on the effectiveness of pharmacological interventions for chronic non-specific low-back pain*, Jan. 2011. DOI: [10.1007/s00586-010-1541-4](https://doi.org/10.1007/s00586-010-1541-4).

- [11] E. M. Heintjes, M. Berger, S. M. Bierma-Zeinstra, R. M. Bernsen, J. A. Verhaar, and B. W. Koes, "Pharmacotherapy for patellofemoral pain syndrome", *Cochrane Database of Systematic Reviews*, Jul. 2004, ISSN: 1469-493X. DOI: [10.1002/14651858.cd003470.pub2](https://doi.org/10.1002/14651858.cd003470.pub2).
- [12] The British Pain Society, *Guidelines for Pain Management Programmes for adults*. 2013, ISBN: 9780956138644. [Online]. Available: www.britishpainsociety.org.
- [13] C. Mitchell, A. Adebajo, E. Hay, and A. Carr, *Shoulder pain: Diagnosis and management in primary care*, Nov. 2005. DOI: [10.1136/bmj.331.7525.1124](https://doi.org/10.1136/bmj.331.7525.1124).
- [14] R. Van Der Sande, W. D. Rinkel, L. Gebremariam, E. M. Hay, B. W. Koes, and B. M. Huisstede, *Subacromial impingement syndrome: Effectiveness of pharmaceutical interventions—nonsteroidal anti-inflammatory drugs, corticosteroid, or other injections: A systematic review*, May 2013. DOI: [10.1016/j.apmr.2012.11.041](https://doi.org/10.1016/j.apmr.2012.11.041).
- [15] C. for Disease Control, U. D. o. H. Prevention Public Health Service, and H. Services, "Guideline for Prescribing Opioids for Chronic Pain", *Journal of Pain and Palliative Care Pharmacotherapy*, vol. 30, no. 2, pp. 138–140, Apr. 2016, ISSN: 15360539. DOI: [10.3109/15360288.2016.1173761](https://doi.org/10.3109/15360288.2016.1173761).
- [16] W. C. Peul, H. C. van Houwelingen, W. B. van den Hout, R. Brand, J. A. Eekhof, J. T. Tans, R. T. Thomeer, and B. W. Koes, "Surgery versus Prolonged Conservative Treatment for Sciatica", *New England Journal of Medicine*, vol. 356, no. 22, pp. 2245–2256, May 2007, ISSN: 0028-4793. DOI: [10.1056/NEJMoa064039](https://doi.org/10.1056/NEJMoa064039). [Online]. Available: <http://www.nejm.org/doi/abs/10.1056/NEJMoa064039>.
- [17] L. Gebremariam, E. M. Hay, B. W. Koes, and B. M. Huisstede, *Effectiveness of surgical and postsurgical interventions for the subacromial impingement syndrome: A systematic review*, Nov. 2011. DOI: [10.1016/j.apmr.2011.06.006](https://doi.org/10.1016/j.apmr.2011.06.006).
- [18] M. J. S. Tamaoki, J. C. Belloti, M. Lenza, M. H. Matsumoto, J. B. Gomes dos Santos, and F. Faloppa, "Surgical versus conservative interventions for treating acromioclavicular dislocation of the shoulder in adults", *Cochrane Database of Systematic Reviews*, Aug. 2010, ISSN: 1469-493X. DOI: [10.1002/14651858.cd007429.pub2](https://doi.org/10.1002/14651858.cd007429.pub2).
- [19] J. A. Coghlan, R. Buchbinder, S. Green, R. V. Johnston, and S. N. Bell, "Surgery for rotator cuff disease", *Cochrane Database of Systematic Reviews*, Jan. 2008, ISSN: 14651858. DOI: [10.1002/14651858.CD005619.pub2](https://doi.org/10.1002/14651858.CD005619.pub2). [Online]. Available: <http://doi.wiley.com/10.1002/14651858.CD005619.pub2>.
- [20] O. Dorrestijn, M. Stevens, J. C. Winters, K. van der Meer, and R. L. Diercks, "Conservative or surgical treatment for subacromial impingement syndrome? a systematic review", *Journal of shoulder and elbow surgery*, vol. 18, no. 4, pp. 652–660, 2009.
- [21] Z. Baber and M. A. Erdek, "Failed back surgery syndrome: Current perspectives", *Journal of pain research*, vol. 9, p. 979, 2016.
- [22] M. G. Joice, S. Bhowmick, and D. F. Amanatullah, "Perioperative physiotherapy in total knee arthroplasty", *Orthopedics*, vol. 40, no. 5, e765–e773, 2017.

- [23] N. Artz, K. T. Elvers, C. M. Lowe, C. Sackley, P. Jepson, and A. D. Beswick, "Effectiveness of physiotherapy exercise following total knee replacement: systematic review and meta-analysis", *BMC Musculoskeletal Disorders*, vol. 16, no. 1, p. 15, Dec. 2015, ISSN: 1471-2474. DOI: [10.1186/s12891-015-0469-6](https://doi.org/10.1186/s12891-015-0469-6). [Online]. Available: <https://bmcmusculoskeletaldisord.biomedcentral.com/articles/10.1186/s12891-015-0469-6>.
- [24] T. Okoro, A. B. Lemmey, P. Maddison, and J. G. Andrew, "An appraisal of rehabilitation regimes used for improving functional outcome after total hip replacement surgery", *Sports Medicine, Arthroscopy, Rehabilitation, Therapy & Technology*, vol. 4, no. 1, p. 5, Dec. 2012, ISSN: 1758-2555. DOI: [10.1186/1758-2555-4-5](https://doi.org/10.1186/1758-2555-4-5). [Online]. Available: <http://bmcsportsscimedrehabil.biomedcentral.com/articles/10.1186/1758-2555-4-5>.
- [25] C. L. Coulter, J. M. Scarvell, T. M. Neeman, and P. N. Smith, "Physiotherapist-directed rehabilitation exercises in the outpatient or home setting improve strength, gait speed and cadence after elective total hip replacement: A systematic review", *Journal of Physiotherapy*, vol. 59, no. 4, pp. 219–226, 2013, ISSN: 18369561. DOI: [10.1016/S1836-9553\(13\)70198-X](https://doi.org/10.1016/S1836-9553(13)70198-X).
- [26] M. Zampolini, E. Todeschini, M. B. Guitart, H. Hermens, S. Ilsbrouckx, V. Macellari, R. Magni, M. Rogante, S. S. Marchese, M. Vollenbroek, and C. Giacomozzi, *Tele-rehabilitation: Present and future*, 2008.
- [27] L. R. Mikkelsen, I. Mechlenburg, K. Søballe, L. B. Jørgensen, S. Mikkelsen, T. Bandholm, and A. K. Petersen, "Effect of early supervised progressive resistance training compared to unsupervised home-based exercise after fast-track total hip replacement applied to patients with preoperative functional limitations. A single-blinded randomised controlled trial", *Osteoarthritis and Cartilage*, vol. 22, no. 12, pp. 2051–2058, 2014, ISSN: 15229653. DOI: [10.1016/j.joca.2014.09.025](https://doi.org/10.1016/j.joca.2014.09.025).
- [28] R. Haas, M. Sarkies, K.-A. Bowles, L. O'Brien, and T. Haines, "Early commencement of physical therapy in the acute phase following elective lower limb arthroplasty produces favorable outcomes: a systematic review and meta-analysis examining allied health service models", *Osteoarthritis and Cartilage*, vol. 24, pp. 1667–1681, 2016. DOI: [10.1016/j.joca.2016.05.005](https://doi.org/10.1016/j.joca.2016.05.005). [Online]. Available: <http://dx.doi.org/10.1016/j.joca.2016.05.005>.
- [29] C. A. Oatis, W. Li, J. M. Dirusso, M. J. Hoover, K. K. Johnston, M. K. Butz, A. L. Phillips, K. M. Nanovic, E. C. Cummings, M. C. Rosal, D. C. Ayers, and P. D. Franklin, "Variations in Delivery and Exercise Content of Physical Therapy Rehabilitation Following Total Knee Replacement Surgery: A Cross-Sectional Observation Study HHS Public Access", *Int J Phys Med Rehabil*, vol. 5, 2014. DOI: [10.4172/2329-9096.S5-002](https://doi.org/10.4172/2329-9096.S5-002).
- [30] N. Artz, S. Dixon, V. Wylde, A. Beswick, A. Blom, and R. Gooberman-Hill, "Physiotherapy Provision Following Discharge after Total Hip and Total Knee Replacement: A Survey of Current Practice at High-Volume NHS Hospitals in England and Wales", *Musculoskeletal Care*, vol. 11, no. 1, pp. 31–38, Mar. 2013, ISSN: 14782189. DOI: [10.1002/msc.1027](https://doi.org/10.1002/msc.1027). [Online]. Available: <http://doi.wiley.com/10.1002/msc.1027>.
- [31] M. D. Westby, A. Brittain, and C. L. Backman, "Expert Consensus on Best Practices for Post-Acute Rehabilitation After Total Hip and Knee Arthroplasty: A Canada and United States Delphi Study", *Arthritis Care & Research*, vol. 66, no. 3, pp. 411–423, Mar. 2014, ISSN:

- 2151464X. doi: [10.1002/acr.22164](https://doi.org/10.1002/acr.22164). [Online]. Available: <http://doi.wiley.com/10.1002/acr.22164>.
- [32] W. Zhao, D. D. Espy, M. Ann Reinthal, and H. Feng, "A Feasibility Study of Using a Single Kinect Sensor for Rehabilitation Exercises Monitoring: A Rule Based Approach", Tech. Rep.
- [33] M. P. Galea, P. Levinger, N. Lythgo, C. Cimoli, R. Weller, E. Tully, J. McMeeken, and R. Westh, "A targeted home- and center-based exercise program for people after total hip replacement: A randomized clinical trial", *Archives of Physical Medicine and Rehabilitation*, vol. 89, no. 8, pp. 1442–1447, 2008, issn: 00039993. doi: [10.1016/j.apmr.2007.11.058](https://doi.org/10.1016/j.apmr.2007.11.058).
- [34] D. Li, Z. Yang, P. Kang, and X. Xie, "Home-Based Compared with Hospital-Based Rehabilitation Program for Patients Undergoing Total Knee Arthroplasty for Osteoarthritis", *American Journal of Physical Medicine & Rehabilitation*, vol. 96, no. 6, pp. 440–447, Jun. 2017, issn: 0894-9115. doi: [10.1097/PHM.0000000000000621](https://doi.org/10.1097/PHM.0000000000000621). [Online]. Available: <http://insights.ovid.com/crossref?an=00002060-201706000-00012>.
- [35] R. López-Liria, D. Padilla-Góngora, D. Catalan-Matamoros, P. Rocamora-Pérez, S. Pérez-De La Cruz, and M. Fernández-Sánchez, "Home-based versus hospital-based rehabilitation program after total knee replacement", *BioMed Research International*, vol. 2015, 2015, issn: 23146141. doi: [10.1155/2015/450421](https://doi.org/10.1155/2015/450421).
- [36] J. F. Kramer, M. Speechley, R. Bourne, C. Rorabeck, and M. Vaz, "Comparison of clinic- and home-based rehabilitation programs after total knee arthroplasty", *Clinical Orthopaedics and Related Research*, no. 410, pp. 225–234, May 2003, issn: 15281132. doi: [10.1097/01.blo.0000063600.67412.11](https://doi.org/10.1097/01.blo.0000063600.67412.11).
- [37] K. J. Picha and D. M. Howell, "A model to increase rehabilitation adherence to home exercise programmes in patients with varying levels of self-efficacy", *Musculoskeletal Care*, vol. 16, no. 1, pp. 233–237, Mar. 2018, issn: 14782189. doi: [10.1002/msc.1194](https://doi.org/10.1002/msc.1194). [Online]. Available: <http://doi.wiley.com/10.1002/msc.1194>.
- [38] R. Bashshur, G. Shannon, E. Krupinski, and J. Grigsby, "The taxonomy of telemedicine", *Telemedicine and e-Health*, vol. 17, no. 6, pp. 484–494, Jul. 2011, issn: 15305627. doi: [10.1089/tmj.2011.0103](https://doi.org/10.1089/tmj.2011.0103).
- [39] A. C. Rodriguez, C. Roda, P. González, and E. Navarro, "Contextualizing tasks in tele-rehabilitation systems for older people", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9455, Springer Verlag, 2015, pp. 29–41, isbn: 9783319264097. doi: [10.1007/978-3-319-26410-3_4](https://doi.org/10.1007/978-3-319-26410-3_4).
- [40] M. Mutingi and C. Mbohwa, "Developing Multi-agent Systems for mHealth Drug Delivery", in 2015, pp. 671–683. doi: [10.1007/978-3-319-12817-7_29](https://doi.org/10.1007/978-3-319-12817-7_29).
- [41] P. A. Morreale, "Wireless sensor network applications in urban telehealth", in *Proceedings - 21st International Conference on Advanced Information Networking and Applications Workshops/Symposia, AINAW'07*, vol. 1, 2007, pp. 810–814, isbn: 0769528473. doi: [10.1109/AINAW.2007.375](https://doi.org/10.1109/AINAW.2007.375).

- [42] F. Dias Correia, A. Nogueira, I. Magalhães, J. Guimarães, M. Moreira, I. Barradas, L. Teixeira, J. Tulha, R. Seabra, J. Lains, and V. Bento, "Home-based Rehabilitation With A Novel Digital Biofeedback System versus Conventional In-person Rehabilitation after Total Knee Replacement: a feasibility study OPEN", *SCIENTIFIC RepORTS* 1, vol. 8, p. 11299, 2018. doi: [10.1038/s41598-018-29668-0](https://doi.org/10.1038/s41598-018-29668-0). [Online]. Available: www.nature.com/scientificreports/.
- [43] M. Raaben, H. C. Vogely, and T. J. Blokhuis, "Real-time visual biofeedback to improve therapy compliance after total hip arthroplasty: A pilot randomized controlled trial", *Gait and Posture*, vol. 61, pp. 306–310, Mar. 2018, issn: 18792219. doi: [10.1016/j.gaitpost.2018.01.038](https://doi.org/10.1016/j.gaitpost.2018.01.038).
- [44] Z. Pataky, D. De León Rodriguez, A. Golay, M. Assal, J. P. Assal, and C. A. Hauert, "Biofeedback Training for Partial Weight Bearing in Patients After Total Hip Arthroplasty", *Archives of Physical Medicine and Rehabilitation*, vol. 90, no. 8, pp. 1435–1438, Aug. 2009, issn: 00039993. doi: [10.1016/j.apmr.2009.02.011](https://doi.org/10.1016/j.apmr.2009.02.011).
- [45] S. Mathias, U. S. Nayak, and B. Isaacs, "Balance in elderly patients: The 'get-up and go' test", *Archives of Physical Medicine and Rehabilitation*, vol. 67, no. 6, pp. 387–389, 1986, issn: 00039993.
- [46] F. D. Correia, A. Nogueira, I. Magalhães, J. Guimarães, M. Moreira, I. Barradas, M. Molinos, L. Teixeira, J. Tulha, R. Seabra, J. Lains, and V. Bento, "Medium-Term Outcomes of Digital Versus Conventional Home-Based Rehabilitation After Total Knee Arthroplasty: Prospective, Parallel-Group Feasibility Study.", *JMIR rehabilitation and assistive technologies*, vol. 6, no. 1, e13111, Feb. 2019, issn: 2369-2529. doi: [10.2196/13111](https://doi.org/10.2196/13111). [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/30816849%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6416534>.
- [47] F. D. Correia, A. Nogueira, I. Magalhães, J. Guimarães, M. Moreira, I. Barradas, M. Molinos, L. Teixeira, J. Pires, R. Seabra, J. Lains, and V. Bento, "Digital versus conventional rehabilitation after total hip arthroplasty: A single-center, parallel-group pilot study", *Journal of Medical Internet Research*, vol. 21, no. 6, 2019, issn: 14388871. doi: [10.2196/14523](https://doi.org/10.2196/14523).
- [48] N. Maclean, P. Pound, C. Wolfe, and A. Rudd, "Qualitative analysis of stroke patients' motivation for rehabilitation", doi: [10.1136/bmj.321.7268.1051](https://doi.org/10.1136/bmj.321.7268.1051). [Online]. Available: <http://www.bmj.com/>.
- [49] G. Gard and A. L. Gyllensten, "The Importance of Emotions in Physiotherapeutic Practice", *Physical Therapy Reviews*, vol. 5, no. 3, pp. 155–160, Sep. 2000, issn: 1743288X. doi: [10.1179/ptr.2000.5.3.155](https://doi.org/10.1179/ptr.2000.5.3.155).
- [50] J. Tracey, "The Emotional Response to the Injury and Rehabilitation Process", *Journal of Applied Sport Psychology*, vol. 15, no. 4, pp. 279–293, Oct. 2003, issn: 15331571. doi: [10.1080/714044197](https://doi.org/10.1080/714044197).
- [51] A. D. Woolf and K. Åkesson, *Understanding the burden of musculoskeletal conditions*, May 2001. doi: [10.1136/bmj.322.7294.1079](https://doi.org/10.1136/bmj.322.7294.1079). [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1120225/>.
- [52] N. Arden and M. C. Nevitt, *Osteoarthritis: Epidemiology*, Feb. 2006. doi: [10.1016/j.berh.2005.09.007](https://doi.org/10.1016/j.berh.2005.09.007).

- [53] I. B. McInnes and G. Schett, "The Pathogenesis of Rheumatoid Arthritis", *New England Journal of Medicine*, vol. 365, no. 23, pp. 2205–2219, Dec. 2011, issn: 0028-4793. doi: [10.1056/NEJMra1004965](https://doi.org/10.1056/NEJMra1004965). [Online]. Available: <http://www.nejm.org/doi/abs/10.1056/NEJMra1004965>.
- [54] T. D. Rachner, S. Khosla, and L. C. Hofbauer, *Osteoporosis: Now and the future*, 2011. doi: [10.1016/S0140-6736\(10\)62349-5](https://doi.org/10.1016/S0140-6736(10)62349-5).
- [55] F. Cosman, S. J. de Beur, M. LeBoff, E. Lewiecki, B. Tanner, S. Randall, and R. Lindsay, "Clinician's guide to prevention and treatment of osteoporosis", *Osteoporosis international*, vol. 25, no. 10, pp. 2359–2381, 2014.
- [56] V. C. Hoe, D. M. Urquhart, H. L. Kelsall, and M. R. Sim, "Ergonomic design and training for preventing work-related musculoskeletal disorders of the upper limb and neck in adults", *Cochrane Database of Systematic Reviews*, Aug. 2012, issn: 14651858. doi: [10.1002/14651858.CD008570.pub2](https://doi.org/10.1002/14651858.CD008570.pub2). [Online]. Available: <http://doi.wiley.com/10.1002/14651858.CD008570.pub2>.
- [57] B. P. Bernard, V. Putz-Anderson, M. E. Susan Burt Libby L Cole, C. J. Fairfield-Estill Lawrence Fine, D. A. Katharyn Grant, C. J. Gjessing Lynn Jenkins Joseph Hurrell Jr, N. Nelson, D. Pfirman Robert Roberts Diana Stetson, M. Haring-Sweeney, and S. Tanaka, "Musculoskeletal Disorders and Workplace Factors A Critical Review of Epidemiologic Evidence for Work-Related Musculoskeletal Disorders of the Neck, Upper Extremity, and Low Back", Tech. Rep., 1997. [Online]. Available: <http://www.cdc.gov/niosh>.
- [58] J. J. F. Soares, Ö. Sundin, and G. Grossi, "Age and Musculoskeletal Pain", Tech. Rep., 2003.
- [59] S. Parsons, A. Breen, N. Foster, L. Letley, T. Pincus, S. Vogel, and M. Underwood, "Prevalence and comparative troublesomeness by age of musculoskeletal pain in different body locations", *Family Practice*, vol. 24, no. 4, pp. 308–316, Aug. 2007, issn: 0263-2136. doi: [10.1093/fampra/cmm027](https://doi.org/10.1093/fampra/cmm027). [Online]. Available: <https://academic.oup.com/fampra/article-lookup/doi/10.1093/fampra/cmm027>.
- [60] E. Holmström and G. Engholm, "Musculoskeletal disorders in relation to age and occupation in Swedish construction workers", *American Journal of Industrial Medicine*, vol. 44, no. 4, pp. 377–384, Oct. 2003, issn: 02713586. doi: [10.1002/ajim.10281](https://doi.org/10.1002/ajim.10281). [Online]. Available: <http://doi.wiley.com/10.1002/ajim.10281>.
- [61] J. Collins and L. O'Sullivan, "Psychosocial risk exposures and musculoskeletal disorders across working-age males and females", *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 20, no. 4, pp. 272–286, Jul. 2010, issn: 10908471. doi: [10.1002/hfm.20220](https://doi.org/10.1002/hfm.20220). [Online]. Available: <http://doi.wiley.com/10.1002/hfm.20220>.
- [62] B. Heiden, M. Weigl, P. Angerer, and A. Müller, "Association of age and physical job demands with musculoskeletal disorders in nurses", *Applied Ergonomics*, vol. 44, no. 4, pp. 652–658, 2013, issn: 18729126. doi: [10.1016/j.apergo.2013.01.001](https://doi.org/10.1016/j.apergo.2013.01.001).
- [63] J. D. Collins and L. W. O'Sullivan, "Musculoskeletal disorder prevalence and psychosocial risk exposures by age and gender in a cohort of office based employees in two academic institutions", *International Journal of Industrial Ergonomics*, vol. 46, pp. 85–97, Mar. 2015, issn: 18728219. doi: [10.1016/j.ergon.2014.12.013](https://doi.org/10.1016/j.ergon.2014.12.013).

- [64] R. Webb, T. Brammah, M. Lunt, M. Urwin, T. Allison, and D. Symmons, "Prevalence and Predictors of Intense, Chronic, and Disabling Neck and Back Pain in the UK General Population", *Spine*, vol. 28, no. 11, pp. 1195–1202, Jun. 2003, issn: 0362-2436. doi: [10.1097/01.brs.0000067430.49169.01](https://doi.org/10.1097/01.brs.0000067430.49169.01).
- [65] D. Hoy, P. Brooks, F. Blyth, and R. Buchbinder, *The Epidemiology of low back pain*, Dec. 2010. doi: [10.1016/j.berh.2010.10.002](https://doi.org/10.1016/j.berh.2010.10.002).
- [66] T. Kelly, W. Yang, C.-S. Chen, K. Reynolds, and J. He, "Global burden of obesity in 2005 and projections to 2030", *International Journal of Obesity*, vol. 32, pp. 1431–1437, 2008. doi: [10.1038/ijo.2008.102](https://doi.org/10.1038/ijo.2008.102). [Online]. Available: www.nature.com/ijo.
- [67] E. Pineda, L. M. Sanchez-Romero, M. Brown, A. Jaccard, J. Jewell, G. Galea, L. Webber, and J. Breda, "Forecasting Future Trends in Obesity across Europe: The Value of Improving Surveillance", *Obes Facts*, vol. 11, pp. 360–371, 2018. doi: [10.1159/000492115](https://doi.org/10.1159/000492115). [Online]. Available: www.karger.com/ofa.
- [68] B. Marcus, "Product Desk", pp. 75–83, 2012. doi: [10.1007/978-1-4419-6126-6_10](https://doi.org/10.1007/978-1-4419-6126-6_10).
- [69] N. Maclean and P. Pound, "A critical review of the concept of patient motivation in the literature on physical rehabilitation", doi: [10.1016/S0277-9536\(99\)00334-2](https://doi.org/10.1016/S0277-9536(99)00334-2). [Online]. Available: www.elsevier.com/locate/socscimed.
- [70] J. S. Eccles and A. Wigfield, "Motivational beliefs, values, and goals", *Annual review of psychology*, vol. 53, no. 1, pp. 109–132, 2002.
- [71] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.", *American psychologist*, vol. 55, no. 1, p. 68, 2000.
- [72] S. R. Chen, C. S. Chen, and P. C. Lin, "The effect of educational intervention on the pain and rehabilitation performance of patients who undergo a total knee replacement", *Journal of Clinical Nursing*, vol. 23, no. 1-2, pp. 279–287, 2014, issn: 09621067. doi: [10.1111/jocn.12466](https://doi.org/10.1111/jocn.12466).
- [73] J. W. Vlaeyen, G. Crombez, and S. J. Linton, "The fear-avoidance model of pain", *Pain*, vol. 157, no. 8, pp. 1588–1589, Aug. 2016, issn: 18726623. doi: [10.1097/j.pain.0000000000000574](https://doi.org/10.1097/j.pain.0000000000000574). [Online]. Available: http://content.wkhealth.com/linkback/openurl?sid=WKP_TLP:landingpage%7B%5C%7D&Dan=00006396-201608000-00005.
- [74] L. C. Persson and A. Lilja, "Pain, coping, emotional state and physical function in patients with chronic radicular neck pain. A comparison between patients treated with surgery, physiotherapy or neck collar - A blinded, prospective randomized study", *Disability and Rehabilitation*, vol. 23, no. 8, pp. 325–335, May 2001, issn: 09638288. doi: [10.1080/09638280010005567](https://doi.org/10.1080/09638280010005567).
- [75] M. C. Bushnell, M. Čeko, and L. A. Low, "Cognitive and emotional control of pain and its disruption in chronic pain", *Nature Reviews Neuroscience*, vol. 14, no. 7, pp. 502–511, 2013.
- [76] S. Hewlett, J. Nicklin, and G. Treharne, "Fatigue in musculoskeletal conditions", *The Journal Of Rheumatology Supplement*, vol. 67, no. Series 6, pp. 42–44, 2003. [Online]. Available: <http://eprints.uwe.ac.uk/10682/>.

- [77] H. Repping-Wuts, R. Uitterhoeve, P. van Riel, and T. van Achterberg, "Fatigue as experienced by patients with rheumatoid arthritis (RA): A qualitative study", *International Journal of Nursing Studies*, vol. 45, no. 7, pp. 995–1002, Jul. 2008, ISSN: 00207489. DOI: [10.1016/j.ijnurstu.2007.06.007](https://doi.org/10.1016/j.ijnurstu.2007.06.007).
- [78] Y. Yildirim and G. Ergin, "A validity and reliability study of the Turkish Multidimensional Assessment of Fatigue (MAF) scale in chronic musculoskeletal physical therapy patients", *Journal of Back and Musculoskeletal Rehabilitation*, vol. 26, pp. 307–316, 2013. DOI: [10.3233/BMR-130385](https://doi.org/10.3233/BMR-130385).
- [79] J. M. Hush, K. Cameron, and M. Mackey, "Patient Satisfaction With Musculoskeletal Physical Therapy Care: A Systematic Review", Tech. Rep., 2011. [Online]. Available: <https://academic.oup.com/ptj/article-abstract/91/1/25/2735069>.
- [80] J. M. Hush, V. Yung, M. Mackey, R. Adams, B. M. Wand, R. Nelson, and P. Beattie, "Patient satisfaction with musculoskeletal physiotherapy care in Australia: an international comparison", ISSN: 2042-6186. DOI: [10.1179/2042618612Y.0000000009](https://doi.org/10.1179/2042618612Y.0000000009). [Online]. Available: <https://www.tandfonline.com/action/journalInformation?journalCode=yjmt20>.
- [81] F. F. Reichheld, "The One Number You Need to Grow", Tech. Rep., 2003. [Online]. Available: www.hbr.org.
- [82] T. L. Keiningham, B. Cooil, W. Andreassen, and L. Aksoy, "A Longitudinal Examination of Net Promoter and Firm Revenue Growth", *Journal of Marketing*, vol. 71, pp. 39–51, 2007, ISSN: 1547-7185. [Online]. Available: <http://www.marketingpower.com/jmblog..>
- [83] N. I. Fisher and R. E. Kordupleski, "What's Wrong with Net-Promoter Score?", Tech. Rep.
- [84] P. C. Mandal, "Net promoter score: a conceptual analysis", *International Journal of Management Concepts and Philosophy*, vol. 8, no. 4, p. 209, 2014, ISSN: 1478-1484. DOI: [10.1504/ijmcp.2014.066899](https://doi.org/10.1504/ijmcp.2014.066899).
- [85] Y. Baştanlar and M. Ozuysal, *Introduction to Machine Learning Second Edition*. 2014, vol. 1107, pp. 105–28, ISBN: 9780262012430. DOI: [10.1007/978-1-62703-748-8_7](https://doi.org/10.1007/978-1-62703-748-8_7). [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/24272434>.
- [86] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.
- [87] J. M. Moreira, A. C. P. L. F. de Carvalho, and T. Horváth, *A General Introduction to Data Analytics*. 2018, ISBN: 9781119296256. DOI: [10.1002/9781119296294](https://doi.org/10.1002/9781119296294).
- [88] D. Mining and T. Textbook, *No Title*, ISBN: 9783319141411.
- [89] R. W. Picard, "Affective Computing", Tech. Rep., 1995. [Online]. Available: <http://www.media.mit.edu/%CB%9Cpicard/>.
- [90] J. Tao and T. Tan, "Affective computing: A review", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3784 LNCS, 2005, pp. 981–995, ISBN: 3540296212. DOI: [10.1007/11573548_125](https://doi.org/10.1007/11573548_125).
- [91] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion", *Information Fusion*, vol. 37, pp. 98–125, 2017, ISSN: 15662535. DOI: [10.1016/j.inffus.2017.02.003](https://doi.org/10.1016/j.inffus.2017.02.003). [Online]. Available: <https://doi.org/10.1016/j.inffus.2017.02.003>.

- [92] W. G. Parrott, *Emotions in social psychology: Essential readings*. Psychology Press, 2001.
- [93] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications", *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010, ISSN: 19493045. DOI: [10.1109/T-AFFC.2010.1](https://doi.org/10.1109/T-AFFC.2010.1).
- [94] K. R. Scherer, "What are emotions? and how can they be measured?", *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.
- [95] P. Ekman, "Universal and cultural differences in facial expression of emotion", in *Nebraska symposium on motivation*, vol. 19, 1972, pp. 207–284.
- [96] J. A. Russell, "Affective space is bipolar.", *Journal of personality and social psychology*, vol. 37, no. 3, p. 345, 1979.
- [97] C. E. Osgood, "The nature and measurement of meaning", *Psychological Bulletin*, vol. 49, no. 3, pp. 197–237, May 1952, ISSN: 00332909. DOI: [10.1037/h0055737](https://doi.org/10.1037/h0055737).
- [98] C. M. Whissell, "The dictionary of affect in language", in *The measurement of emotions*, Elsevier, 1989, pp. 113–131.
- [99] R. Plutchik, *A psychoevolutionary theory of emotions*, 1982.
- [100] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament", *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.
- [101] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions Is Not Two-Dimensional", Tech. Rep., 2007.
- [102] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions", in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7403 LNCS, 2012, pp. 144–157, ISBN: 9783642345838. DOI: [10.1007/978-3-642-34584-5_11](https://doi.org/10.1007/978-3-642-34584-5_11).
- [103] J. F. Cohn and F. De la Torre, "Automated face analysis for affective computing.", 2015.
- [104] P. Ekman and W. V. Friesen, *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.
- [105] B. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information", *Sensors*, vol. 18, no. 2, p. 401, Jan. 2018, ISSN: 1424-8220. DOI: [10.3390/s18020401](https://doi.org/10.3390/s18020401). [Online]. Available: <http://www.mdpi.com/1424-8220/18/2/401>.
- [106] A. M. Kring and D. Sloan, "The Facial Expression Coding System (FACES): A Users Guide", Tech. Rep.
- [107] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences", *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 5, pp. 699–714, 2005.
- [108] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
- [109] P. Viola and M. J. Jones, "Robust real-time face detection", *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [110] C. Zhang and Z. Zhang, "A survey of recent advances in face detection", 2010.

- [111] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Automatic face identification system using flexible appearance models", *Image and vision computing*, vol. 13, no. 5, pp. 393–401, 1995.
- [112] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face—pain expression recognition using active appearance models", *Image and vision computing*, vol. 27, no. 12, pp. 1788–1796, 2009.
- [113] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift", *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.
- [114] Z. Zhang, "Feature-based facial expression recognition: Sensitivity analysis and experiments with a multilayer perceptron", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, no. 06, pp. 893–911, 1999.
- [115] M. Suk and B. Prabhakaran, "Real-time mobile facial expression recognition system—a case study", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 132–137.
- [116] D. Ghimire and J. Lee, "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines", *Sensors*, vol. 13, no. 6, pp. 7714–7734, 2013.
- [117] S. Happy, A. George, and A. Routray, "A real time facial expression classification system using local binary patterns", in *2012 4th International conference on intelligent human computer interaction (IHCI)*, IEEE, 2012, pp. 1–5.
- [118] M. H. Siddiqi, R. Ali, A. M. Khan, Y.-T. Park, and S. Lee, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields", *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1386–1398, 2015.
- [119] D. Ghimire, S. Jeong, J. Lee, and S. H. Park, "Facial expression recognition based on local region specific features and support vector machines", *Multimedia Tools and Applications*, vol. 76, no. 6, pp. 7803–7821, 2017.
- [120] A. Haro, M. Flickner, and I. Essa, "Detecting and tracking eyes by using their physiological properties, dynamics, and appearance", in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, IEEE, vol. 1, 2000, pp. 163–168.
- [121] A. A. Gunawan *et al.*, "Face expression detection on kinect using active appearance model and fuzzy logic", *Procedia Computer Science*, vol. 59, pp. 268–274, 2015.
- [122] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570.
- [123] Z. Wen *et al.*, "Capturing subtle facial motions in 3d face tracking", in *Proceedings Ninth IEEE International Conference on Computer Vision*, IEEE, 2003, pp. 1343–1350.
- [124] B. Fasel and J. Luetttin, "Automatic facial expression analysis: A survey", *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

- [125] I. Cohen, N. Sebe, A. Garg, M. S. Lew, and T. S. Huang, "Facial expression recognition from video sequences", in *Proceedings. IEEE International Conference on Multimedia and Expo*, IEEE, vol. 2, 2002, pp. 121–124.
- [126] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions", *arXiv preprint arXiv:1705.01842*, 2017.
- [127] C. Xu, S. Cetintas, K. Lee, and L. Li, "Visual sentiment prediction with deep convolutional neural networks (2014)", *arXiv preprint arXiv:1411.5731*,
- [128] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2983–2991.
- [129] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video", in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 467–474.
- [130] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection", in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 2017, pp. 25–32.
- [131] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks", in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
- [132] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.
- [133] S. G. Koolagudi, N. Kumar, and K. S. Rao, "Speech emotion recognition using segmental level prosodic analysis", in *2011 International Conference on Devices and Communications (ICDeCom)*, Feb. 2011, pp. 1–5. DOI: [10.1109/ICDECOM.2011.5738536](https://doi.org/10.1109/ICDECOM.2011.5738536).
- [134] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection", *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005, ISSN: 08936080. DOI: [10.1016/j.neunet.2005.03.007](https://doi.org/10.1016/j.neunet.2005.03.007).
- [135] E. Navas, I. Hernáez, and I. Luengo, "An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional tts", *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1117–1127, 2006.
- [136] H. Atassi and A. Esposito, "A speaker independent approach to the classification of emotional vocal expressions", in *2008 20th IEEE International Conference on Tools with Artificial Intelligence*, IEEE, vol. 2, 2008, pp. 147–152.
- [137] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech", in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [138] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech", in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, IEEE, vol. 3, 1996, pp. 1970–1973.

- [139] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn", in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.
- [140] N. Anand and P. Verma, "Convoluting feelings convolutional and recurrent nets for detecting emotion from audio data", in *Technical Report*, Stanford University, 2015.
- [141] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine", in *Fifteenth annual conference of the international speech communication association*, 2014.
- [142] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen, "Gesture-based affective computing on motion capture data", in *International conference on affective computing and intelligent interaction*, Springer, 2005, pp. 1–7.
- [143] R. Walk and K. Walters, "Perception of the smile and other emotions of the body and face at different distances", 1988.
- [144] S. Piana, A. Stagliano, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures", *arXiv preprint arXiv:1402.5047*, 2014.
- [145] S. Mota and R. W. Picard, "Automated posture analysis for detecting learner's interest level", in *2003 Conference on Computer Vision and Pattern Recognition Workshop*, IEEE, vol. 5, 2003, pp. 49–49.
- [146] S. D'Mello and A. Graesser, "Automatic detection of learner's affect from gross body language", *Applied Artificial Intelligence*, vol. 23, no. 2, pp. 123–150, 2009.
- [147] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon", *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.
- [148] C. Strapparava, A. Valitutti, *et al.*, "Wordnet affect: An affective extension of wordnet.", in *Lrec*, Citeseer, vol. 4, 2004, p. 40.
- [149] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-Emotion Neural Network for Emotion Recognition From Text", *IEEE Access*, vol. 7, pp. 111 866–111 878, Aug. 2019. doi: [10.1109/access.2019.2934529](https://doi.org/10.1109/access.2019.2934529).
- [150] G. Mishne *et al.*, "Experiments with mood classification in blog posts", in *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, vol. 19, 2005, pp. 321–327.
- [151] T. Danisman and A. Alpkocak, "Feeler: Emotion classification of text using vector space model", in *AISB 2008 Convention Communication, Interaction and Social Intelligence*, vol. 1, 2008, p. 53.
- [152] M. Hasan, E. Rundensteiner, and E. Agu, "Emotex: Detecting emotions in twitter messages", 2014.
- [153] A. Seyeditabari, S. Levens, C. D. Maestas, S. Shaikh, J. I. Walsh, W. Zadrozny, C. Danis, and O. P. Thompson, "Cross corpus emotion classification using survey data", *This paper was presented at AISB*, 2017.
- [154] D. Gordeev, "Detecting state of aggression in sentences using cnn", in *International Conference on Speech and Computer*, Springer, 2016, pp. 240–245.

- [155] M. Abdul-Mageed and L. Ungar, "Emonet: Fine-grained emotion detection with gated recurrent neural networks", in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 718–728.
- [156] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm", *arXiv preprint arXiv:1708.00524*, 2017.
- [157] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [158] N. Colneriç and J. Demsar, "Emotion recognition on twitter: Comparative study and training a unison model", *IEEE transactions on affective computing*, 2018.
- [159] A. Schell and M. Dawson, "Psychophysiology", in *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, 2001, pp. 12 448–12 452. DOI: [10.1016/B0-08-043076-7/03424-0](https://doi.org/10.1016/B0-08-043076-7/03424-0). [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/B0080430767034240>.
- [160] A. Fridlund and C. E. Izard, "Electromyographic studies of facial expressions of emotions and patterns of emotions", *Social psychophysiology: A sourcebook*, pp. 243–286, 1983.
- [161] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state", *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.
- [162] H. Feng, H. M. Golshan, and M. H. Mahoor, "A wavelet-based approach to emotion classification using eda signals", *Expert Systems with Applications*, vol. 112, pp. 77–86, 2018.
- [163] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos", *IEEE transactions on affective computing*, vol. 3, no. 2, pp. 211–223, 2011.
- [164] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups", *IEEE Transactions on Affective Computing*, 2018.
- [165] J. Comas, D. Aspandi, and X. Binefa, "End-to-end facial and physiological model for Affective Computing and applications", Tech. Rep. arXiv: [1912.04711v2](https://arxiv.org/abs/1912.04711v2). [Online]. Available: <http://pytorch.org/docs/stable/torchvision/models.html>.
- [166] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems", *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.
- [167] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information", in *Proceedings of the 6th international conference on Multimodal interfaces*, 2004, pp. 205–211.
- [168] M. Wöllmer, F. Wenginger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context", *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

- [169] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009, issn: 01628828. doi: [10.1109/TPAMI.2008.52](https://doi.org/10.1109/TPAMI.2008.52).
- [170] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. D. Rucker, "On the practice of dichotomization of quantitative variables.", *Psychological methods*, vol. 7, no. 1, p. 19, 2002.
- [171] H. Akoglu, "User's guide to correlation coefficients", *Turkish journal of emergency medicine*, vol. 18, no. 3, pp. 91–93, 2018.
- [172] V. J. Rayward-Smith, "Statistics to measure correlation for data mining applications", *Computational Statistics and Data Analysis*, vol. 51, no. 8, pp. 3968–3982, May 2007, issn: 01679473. doi: [10.1016/j.csda.2006.05.025](https://doi.org/10.1016/j.csda.2006.05.025).
- [173] H. Cramer, "Mathematical methods of statistics", *Princeton U. Press, Princeton*, p. 500, 1946.
- [174] C. Pai and K. Potdar, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers", *Article in International Journal of Computer Applications*, vol. 175, no. 4, pp. 975–8887, 2017. doi: [10.5120/ijca2017915495](https://doi.org/10.5120/ijca2017915495). [Online]. Available: <https://www.researchgate.net/publication/320465713>.
- [175] B. Lantz, *Machine learning with R*. Packt Publishing Ltd, 2013.
- [176] O. Sagi and L. Rokach, "Ensemble learning: A survey", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, e1249, 2018.
- [177] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
- [178] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [179] L. Breiman, "Bagging predictors", *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [180] —, "Random forests", *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [181] M. Pal, "Random forest classifier for remote sensing classification", *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.
- [182] S. Bernard, L. Heutte, and S. Adam, "On the Selection of Decision Trees in Random Forests", *Proceedings of International Joint Conference on Neural Networks*, Jun. 2009. [Online]. Available: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5178693>.
- [183] B. H. Menze, M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data", 2009. doi: [10.1186/1471-2105-10-213](https://doi.org/10.1186/1471-2105-10-213). [Online]. Available: <http://www.biomedcentral.com/1471-2105/10/213>.
- [184] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms", in *Advances in neural information processing systems*, 2012, pp. 2951–2959.

- [185] J. H. Friedman, "Greedy function approximation: A gradient boosting machine", *Annals of statistics*, pp. 1189–1232, 2001.
- [186] —, "Stochastic gradient boosting", *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.
- [187] C. Cortes and V. Vapnik, "Support-vector networks", *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [188] E. Garcí'a-Gonzalo, Z. Fernández-Muñiz, P. J. Garcí'a Nieto, A. Bernardo Sánchez, and M. Menéndez Fernández, "Hard-rock stability analysis for span design in entry-type excavations with learning classifiers", *Materials*, vol. 9, no. 7, p. 531, 2016.
- [189] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines", *IEEE transactions on neural networks*, vol. 13, no. 2, pp. 464–471, 2002.
- [190] J. M. De Sa, *Pattern recognition: concepts, methods and applications*. Springer Science & Business Media, 2012.
- [191] I. Guyon, J. Weston, and S. Barnhill, "Gene Selection for Cancer Classification using Support Vector Machines", Tech. Rep., 2002, pp. 389–422.
- [192] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple svm-rfe for gene selection in cancer classification with expression data", *IEEE transactions on nanobioscience*, vol. 4, no. 3, pp. 228–234, 2005.
- [193] J. Bedo, C. Sanderson, and A. Kowalczyk, "An efficient alternative to svm based recursive feature elimination with applications in natural language processing and bioinformatics", in *Australasian Joint Conference on Artificial Intelligence*, Springer, 2006, pp. 170–180.
- [194] T. Frontzek, T. N. Lal, and R. Eckmiller, "Predicting the nonlinear dynamics of biological neurons using support vector machines with different kernels", in *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, IEEE, vol. 2, 2001, pp. 1492–1497.
- [195] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments", doi: [10.1098/rsta.2015.0202](https://doi.org/10.1098/rsta.2015.0202). [Online]. Available: <http://dx.doi.org/10.1098/rsta.2015.0202>.
- [196] H. Han, Z. Cao, B. Gu, and N. Ren, "Pca-svm-based automated fault detection and diagnosis (afdd) for vapor-compression refrigeration systems", *HVAC&R Research*, vol. 16, no. 3, pp. 295–313, 2010.
- [197] T.-K. Lin, "Pca/svm-based method for pattern detection in a multisensor system", *Mathematical Problems in Engineering*, vol. 2018, 2018.
- [198] R. Noori, A. Karbassi, A. Moghaddamnia, D. Han, M. Zokaei-Ashtiani, A. Farokhnia, and M. G. Gousheh, "Assessment of input variables determination on the svm model performance using pca, gamma test, and forward selection techniques for monthly stream flow prediction", *Journal of Hydrology*, vol. 401, no. 3–4, pp. 177–189, 2011.
- [199] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.
- [200] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques", *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

- [201] H. He and E. A. Garcia, "Learning from imbalanced data", *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [202] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions", *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.
- [203] Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning", *Computational Intelligence*, vol. 26, no. 3, pp. 232–257, 2010.
- [204] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions", *Expert Systems with Applications*, vol. 36, no. 3, pp. 5718–5727, 2009.
- [205] I. Mani and I. Zhang, "Knn approach to unbalanced data distributions: A case study involving information extraction", in *Proceedings of workshop on learning from imbalanced datasets*, vol. 126, 2003.
- [206] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data", *Information Sciences*, vol. 409, pp. 17–26, 2017.
- [207] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance", *Information Sciences*, vol. 505, pp. 32–64, 2019, ISSN: 00200255. DOI: [10.1016/j.ins.2019.07.070](https://doi.org/10.1016/j.ins.2019.07.070). [Online]. Available: <https://doi.org/10.1016/j.ins.2019.07.070>.
- [208] N. Japkowicz, "The class imbalance problem: Significance and strategies", in *Proc. of the Int'l Conf. on Artificial Intelligence*, Citeseer, 2000.
- [209] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique", *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [210] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.
- [211] A. Agrawal, H. L. Viktor, and E. Paquet, "Scut: Multi-class imbalanced data classification using smote and cluster-based undersampling", in *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, IEEE, vol. 1, 2015, pp. 226–234.
- [212] J. S. Cardoso and R. Sousa, "Classification models with global constraints for ordinal data", in *2010 Ninth International Conference on Machine Learning and Applications*, IEEE, 2010, pp. 71–77.
- [213] A. Tagliafico, G. Tagliafico, S. Tosto, F. Chiesa, C. Martinoli, L. E. Derchi, and M. Calabrese, "Mammographic density estimation: Comparison among bi-rads categories, a semi-automated software and a fully automated one", *The Breast*, vol. 18, no. 1, pp. 35–40, 2009.
- [214] J. S. Cardoso and R. Sousa, "Measuring the performance of ordinal classification", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 08, pp. 1173–1195, 2011.
- [215] P. McCullagh, "Regression models for ordinal data", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.
- [216] R. Potharst and J. C. Bioch, "A decision tree algorithm for ordinal classification", in *International Symposium on Intelligent Data Analysis*, Springer, 1999, pp. 187–198.

- [217] F. Xia, W. Zhang, and J. Wang, "An effective tree-based algorithm for ordinal regression.", *IEEE Intell. Informatics Bull.*, vol. 7, no. 1, pp. 22–26, 2006.
- [218] R. Sousa and J. S. Cardoso, "Ensemble of decision trees with global constraints for ordinal classification", in *2011 11th International Conference on Intelligent Systems Design and Applications*, IEEE, 2011, pp. 1164–1169.
- [219] J. F. P. da Costa, R. Sousa, and J. S. Cardoso, "An all-at-once unimodal svm approach for ordinal classification", in *2010 Ninth International Conference on Machine Learning and Applications*, IEEE, 2010, pp. 59–64.
- [220] J. P. Da Costa and J. S. Cardoso, "Classification of ordinal data using neural networks", in *European Conference on Machine Learning*, Springer, 2005, pp. 690–697.
- [221] P. A. Gutiérrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study", *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2015.
- [222] E. Frank and M. Hall, "A simple approach to ordinal classification", in *European Conference on Machine Learning*, Springer, 2001, pp. 145–156.
- [223] J. Cheng, Z. Wang, and G. Pollastri, "A neural network approach to ordinal regression", in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, 2008, pp. 1279–1284.
- [224] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [225] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate lstm-fcns for time series classification", *Neural Networks*, vol. 116, pp. 237–245, 2019.
- [226] L. Rokach and O. Maimon, "Clustering methods", in *Data mining and knowledge discovery handbook*, Springer, 2005, pp. 321–352.
- [227] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis", *The computer journal*, vol. 41, no. 8, pp. 578–588, 1998.
- [228] A. K. Jain, "Data clustering: 50 years beyond k-means", *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.
- [229] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.
- [230] P. Bholowalia and A. Kumar, "Ebk-means: A clustering technique based on elbow method and k-means in wsn", *International Journal of Computer Applications*, vol. 105, no. 9, 2014.
- [231] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.
- [232] S. Hidot and C. Saint-Jean, "An expectation–maximization algorithm for the wishart mixture model: Application to movement clustering", *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2318–2324, 2010.
- [233] N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, *Bayesian nonparametrics*. Cambridge University Press, 2010, vol. 28.

- [234] B. Kulis and M. I. Jordan, "Revisiting k-means: New algorithms via bayesian nonparametrics", *arXiv preprint arXiv:1111.0352*, 2011.
- [235] M. Comiter, M. Cha, H. Kung, and S. Teerapittayanon, "Lambda means clustering: Automatic parameter search and distributed computing implementation", in *2016 23rd international conference on pattern recognition (ICPR)*, IEEE, 2016, pp. 2331–2337.
- [236] L. Jiang, Y. Dong, N. Chen, and T. Chen, "Dace: A scalable dp-means algorithm for clustering extremely large sequence data", *Bioinformatics*, vol. 33, no. 6, pp. 834–842, 2017.

Appendix A

Complementary results of models for sentiment-related variables classification

In this Appendix, confusion matrices of the developed ML models, assessed on test set, are presented. All the matrices are normalised over the true labels (rows); the predicted classes are along the columns. For the imbalanced learning techniques results, comparative graphs that comprise test and validation results are also presented. In the beginning of the Chapter, a list of the behavioural features used organised by subgroups is presented.

A.1 List of Extracted Features

Patient Data:

- Age;
- Condition;
- Age;
- Gender;

Therapist Input:

- Prescribed Series of exercises;
- Prescribed Repeats;
- Prescribed Training time;
- Rest time;
- Days since;
- Days since (2 or more);
- Sessions in day;
- Sessions in day (2 or more);
- Drastic Increment;
- Number of different positions;
- Number of different positions (1 or more);
- Number of different exercises;
- Number of different exercises (3 or more);
- Rest time;
- Prescribed Training time;

Performance Metric:

- Stars per repeat;
- Stars per repeat (min);
- Stars per repeat (STD);
- Wrong Movements;
- Performance rate;
- Repeats Completed (%);
- Left tired (sum);
- Left tired (sum, 1 or more);
- Left pain (sum);
- Left pain (sum, 1 or more);
- Left Exercise (sum);
- Left Exercise (sum, 1 or more);
- Time per repeat;
- Total session time;
- Value Repeat;
- Repeats of difficult exercises Prescribed;
- Repeats of difficult exercises Done;
- Repeats of difficult exercises Done (3 or more);
- Number Difficult exercises Prescribed;

- Number Difficult exercises Done;
- Number Difficult exercises Done (1 or more);
- Mean Stars Sessions Before;
- Difference to last sessions score;

Exercise-level Comparison:

- New Exercises;
- Prescribed Repeats (EL);
- Prescribed Series (EL);
- Repeats Done (% EL);
- Stars per Repeat (EL);
- Stars per Repeat (min, EL);
- Time Per repeat (EL);
- Training Time (EL);
- Util Session Time (EL);
- Wrong Movements (EL);

Daily metrics:

- Executions done (%);
- Sessions done (%);
- Total time done (%);
- Repeats Done (%);
- Repeats OK (%);
- Exercises done (%);

A.2 Confusion matrices for sentiment-related variables classification in baseline models.

A.2.1 Random Forest

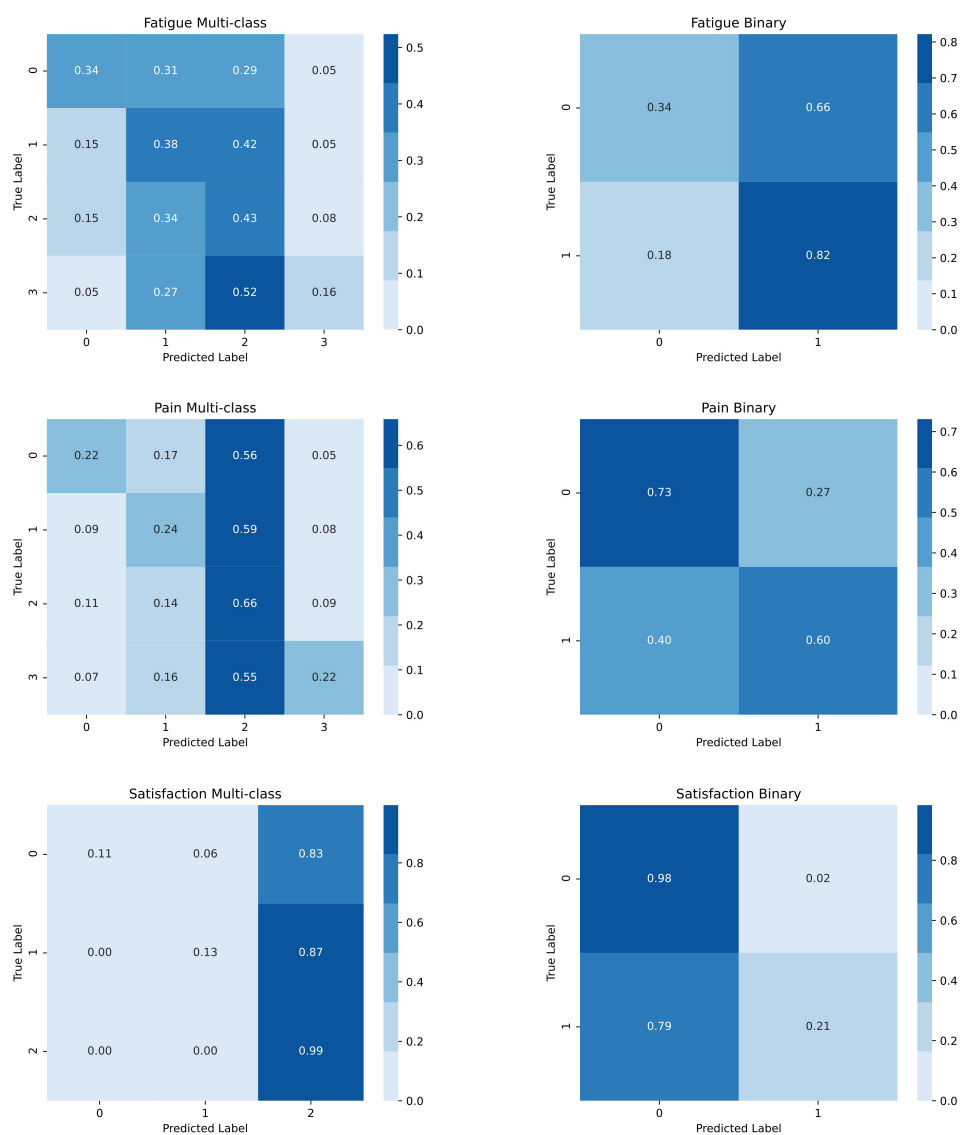


Figure A.2.1: Confusion matrices for sentiment-related variables classification using RF classifier.

A.2.2 Gradient Boosted Trees

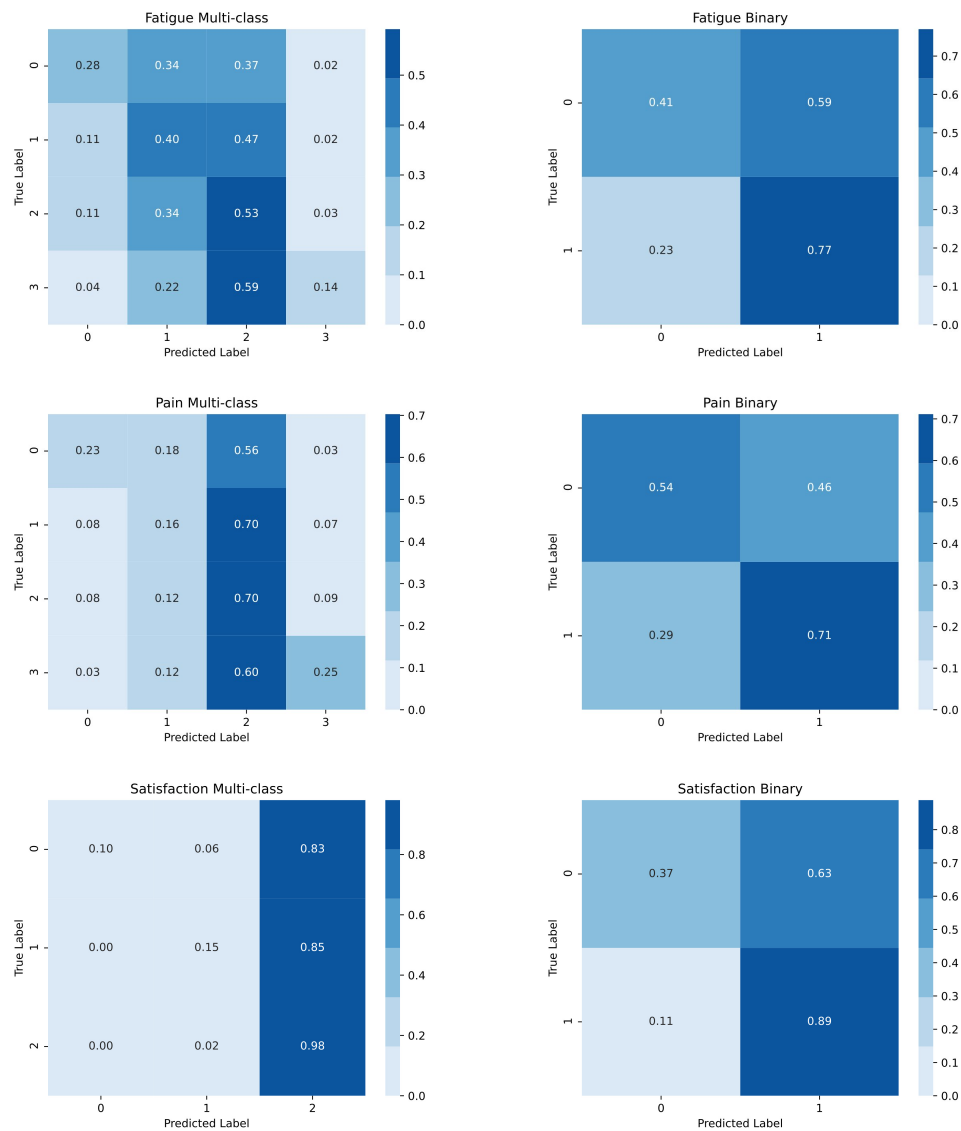


Figure A.2.2: Confusion matrices for sentiment-related variables classification using GBT classifier.

A.2.3 SVM-RFE

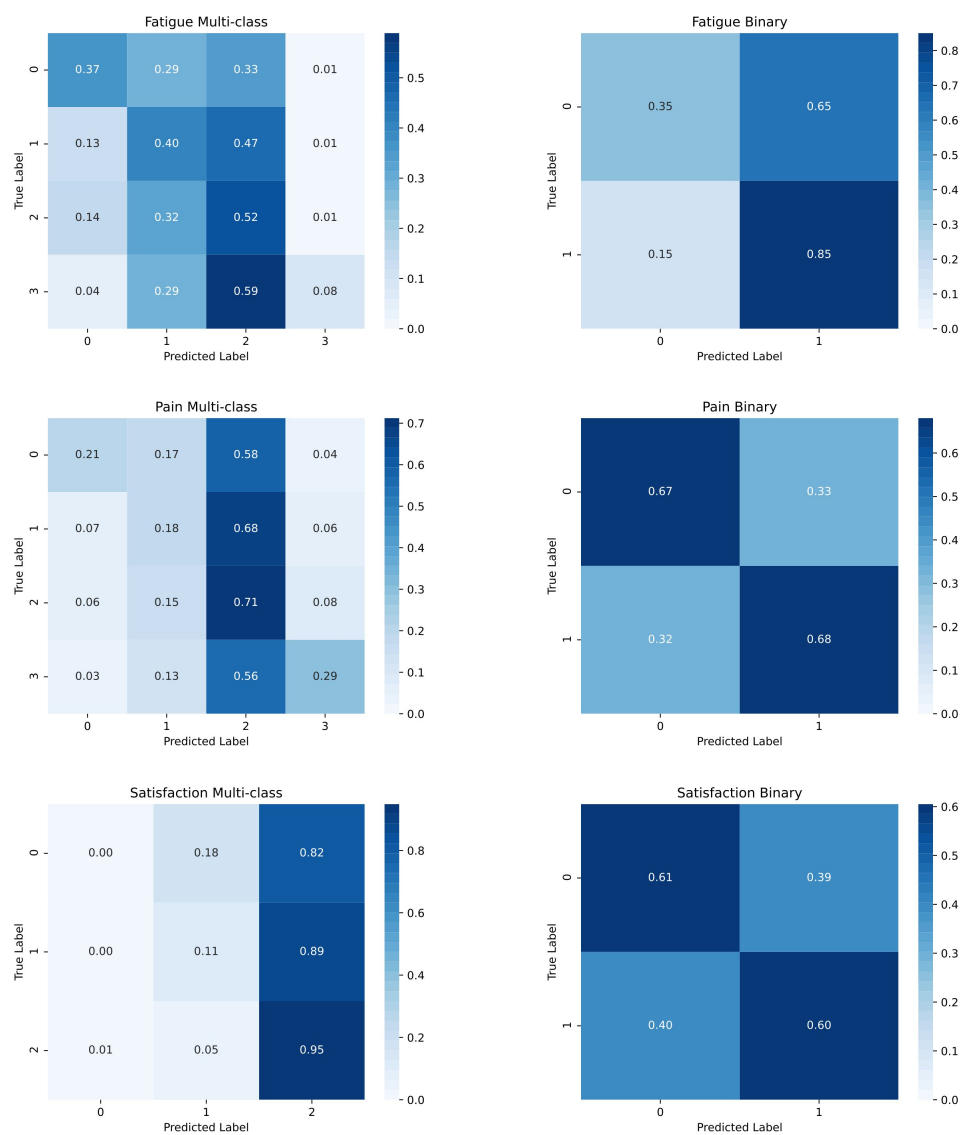


Figure A.2.3: Confusion matrices for sentiment-related variables classification using SVM-RFE classifier.

A.2.4 SVM-PCA

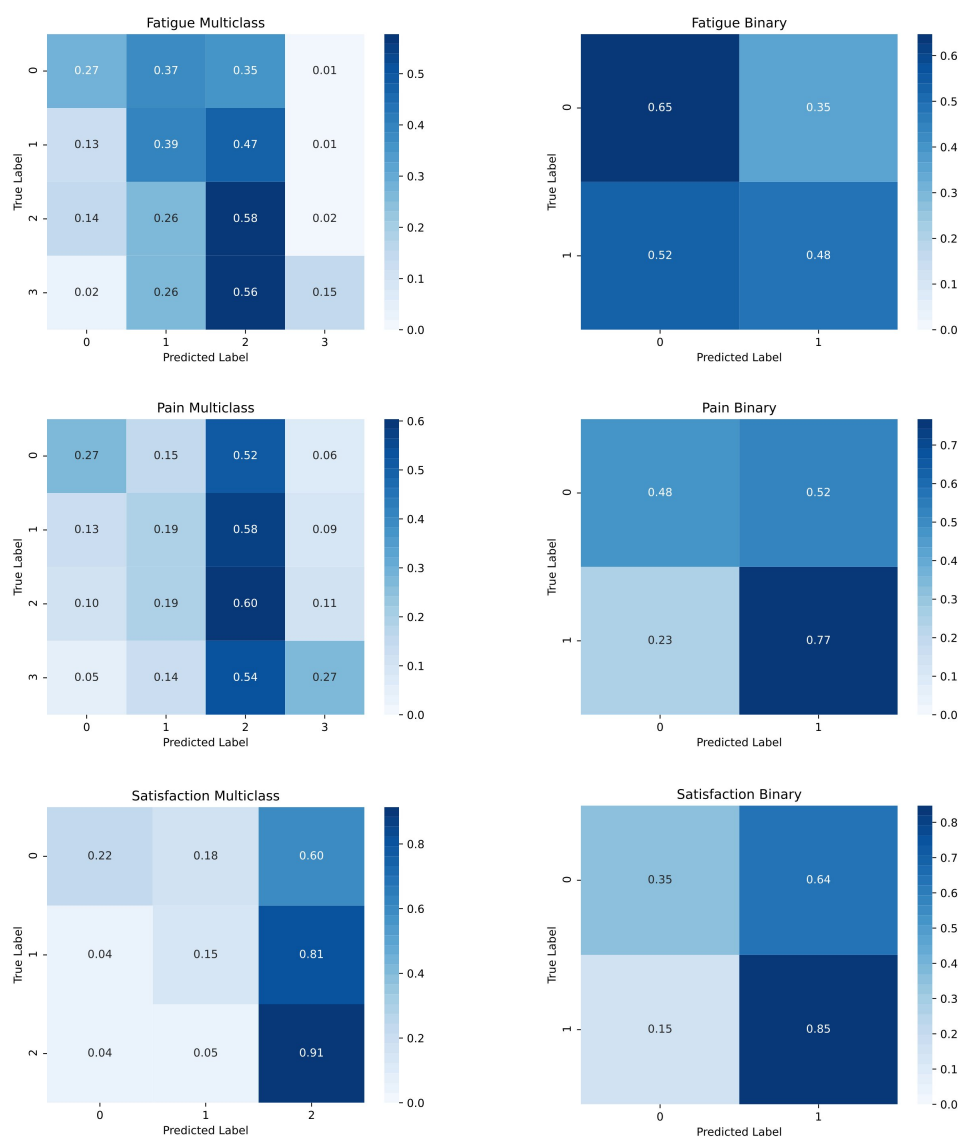


Figure A.2.4: Confusion matrices for sentiment-related variables classification using SVM-PCA classifier.

A.2.5 MLP

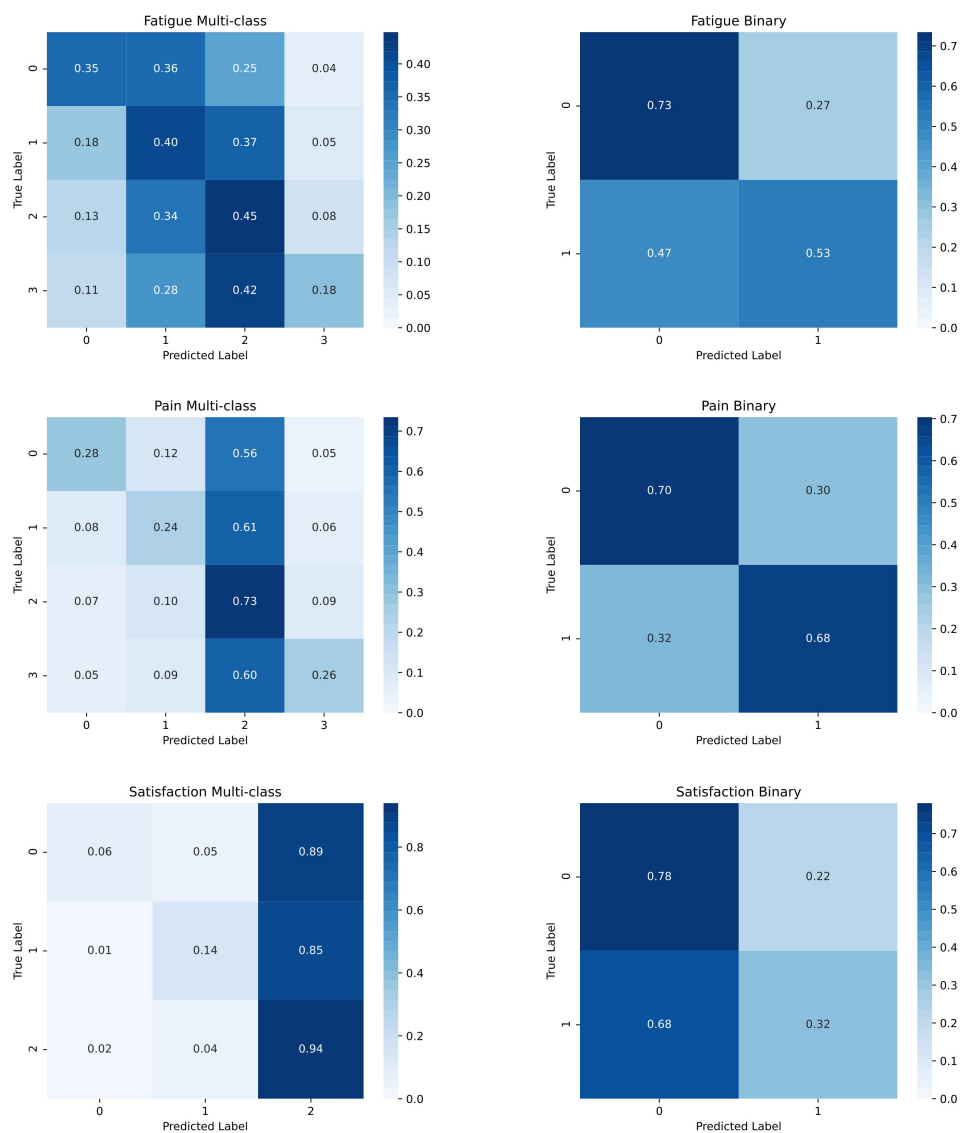
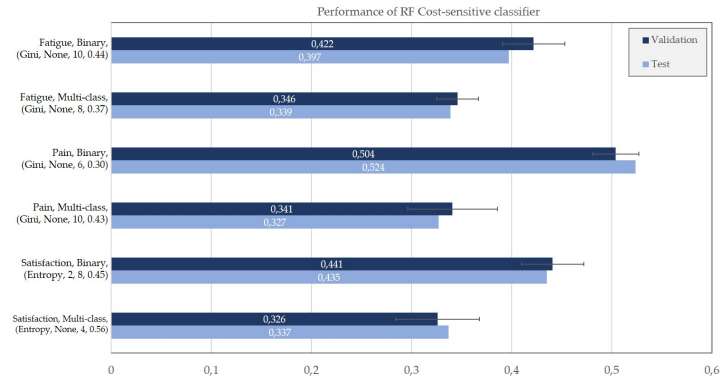


Figure A.2.5: Confusion matrices for sentiment-related variables classification using MLP.

A.3 Complementary results for sentiment-related variables classification using cost-sensitive classifiers

A.3.1 Random Forest Cost-sensitive



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (criterion, max_depth, min_samples_split, max_features)*

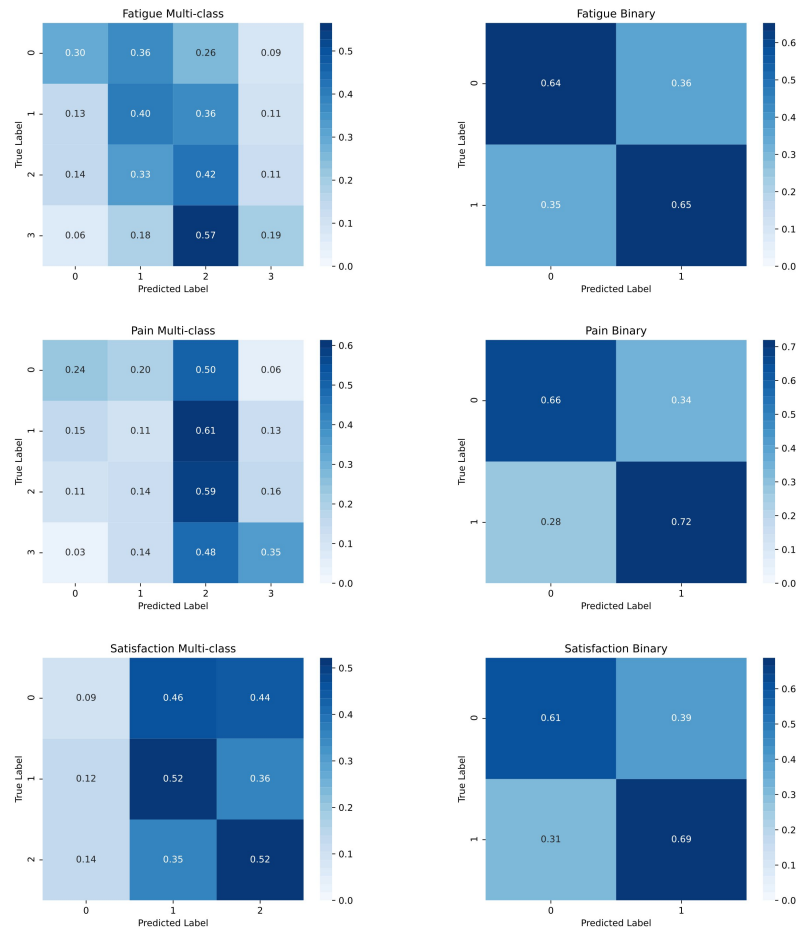
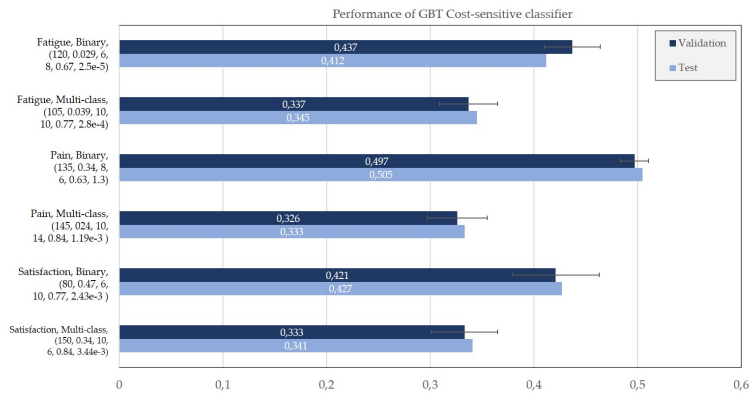


Figure A.3.1: Sentiment-related variables classification results using cost-sensitive RF.

A.3.2 Gradient-Boosted Trees Cost-sensitive



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (n_estimators, learning_rate, max_depth, min_child_weight, column_subsampling, complex_penalty)*.

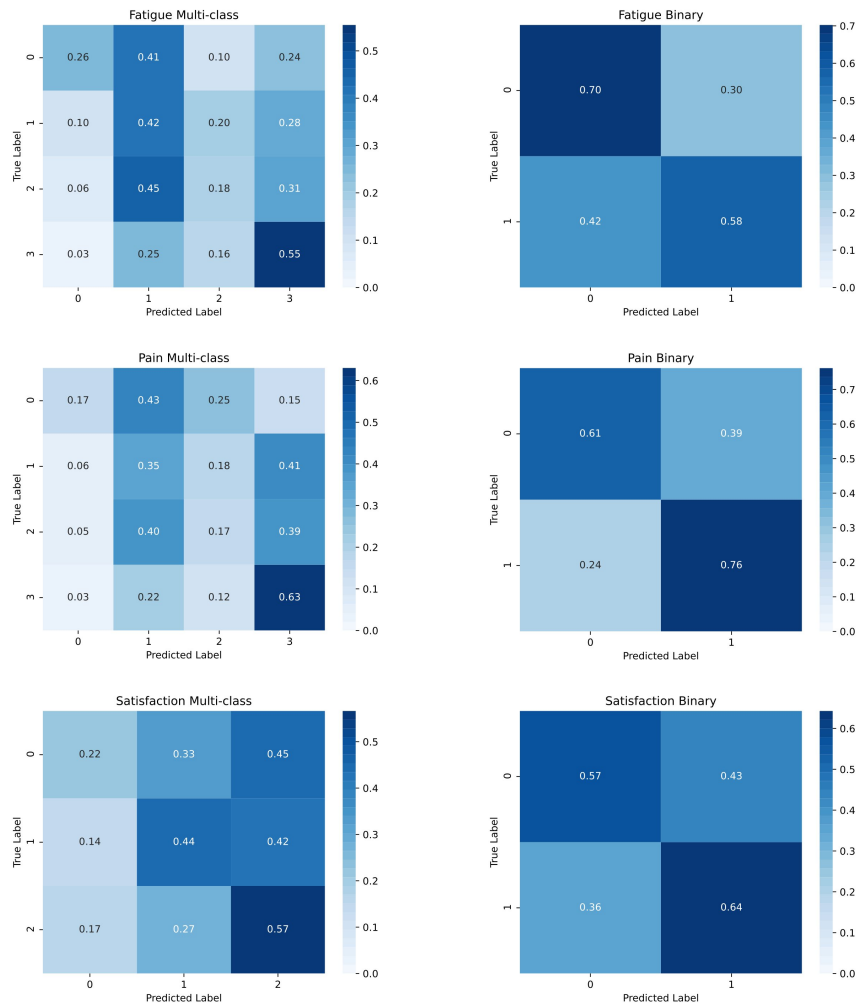
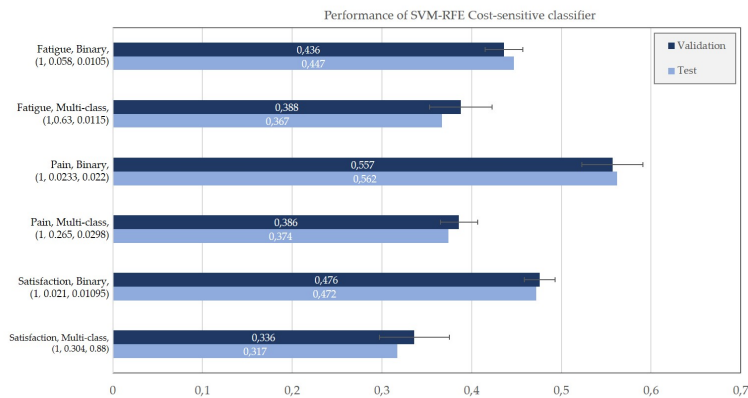


Figure A.3.2: Sentiment-related variables classification results using cost-sensitive GBT.

A.3.3 SVM-RFE Cost-sensitive



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (linear-SVM C, SVM-RBF C, SVM-RBF gamma)*.

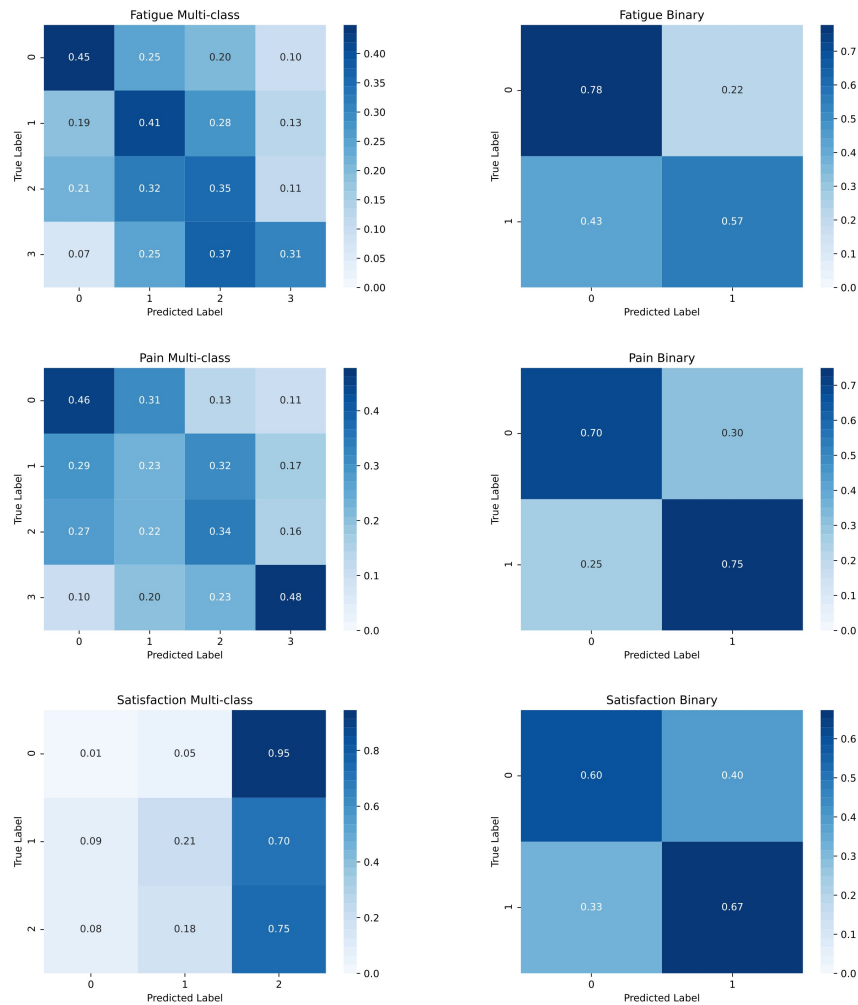
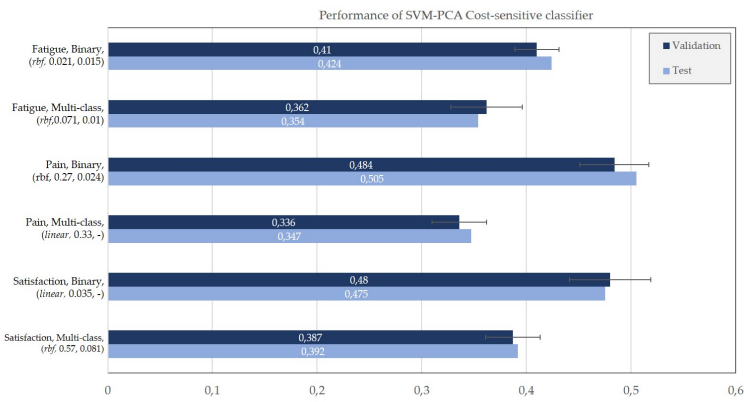


Figure A.3.3: Sentiment-related variables classification results using cost-sensitive SVM-RFE.

A.3.4 SVM-PCA Cost-sensitive



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (kernel, C, gamma)*.

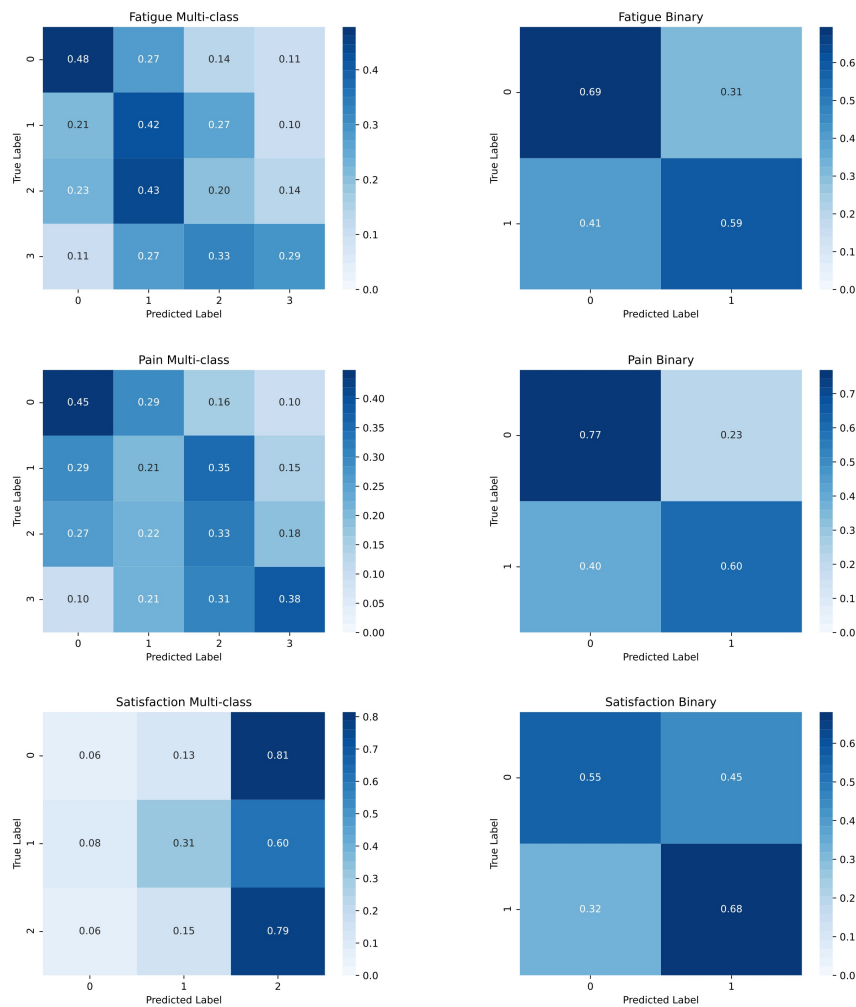
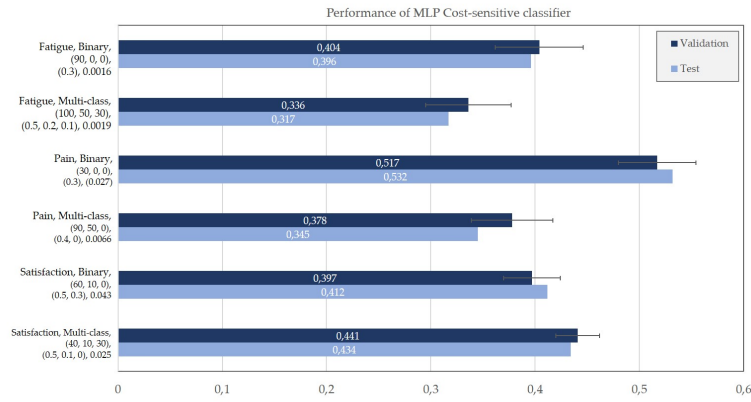


Figure A.3.4: Sentiment-related variables classification results using cost-sensitive SVM-PCA.

A.3.5 MLP Cost-sensitive



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable, classification type, (Num. units layer 1, Num. units layer 2, Num units layer 3), (Layer 1 dropout, Layer 2 dropout, Layer 3 dropout), learning rate*.

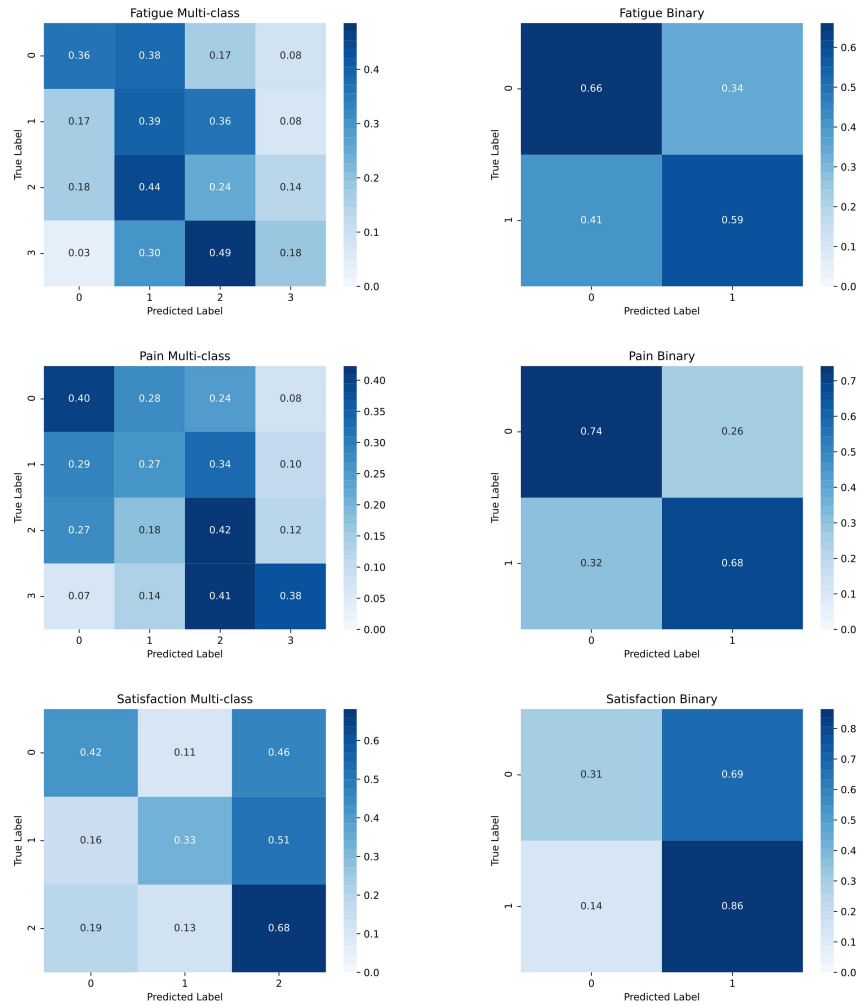


Figure A.3.5: Sentiment-related variables classification results using cost-sensitive MLP.

A.4 Complementary results for sentiment-related variables classification with data balancing: Hybrid approach

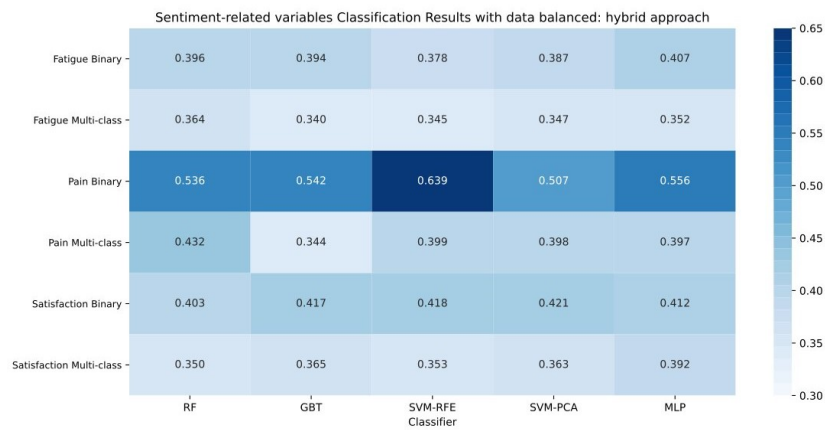
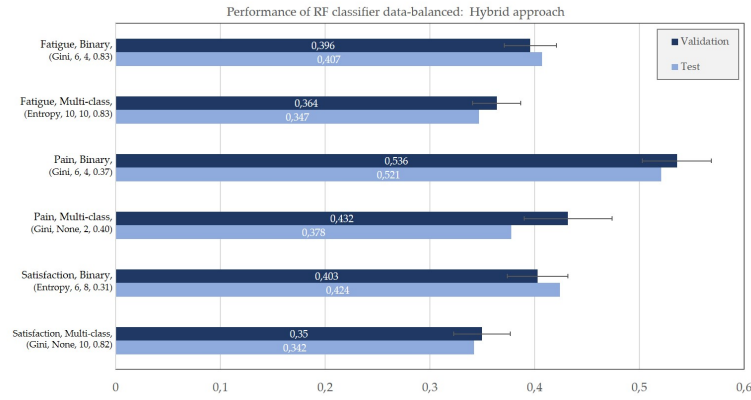


Figure A.4.1: Results of classification of sentiment-related variables using data-balance approaches at validation set: Hybrid Approach. Y axis refer to the variable and type of classification, while x-axis to the different classifiers employed. Values presented correspond to the F1-score

A.4.1 RF with data balancing: hybrid approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (linear-SVM C, SVM-RBF C, SVM-RBF gamma)*.

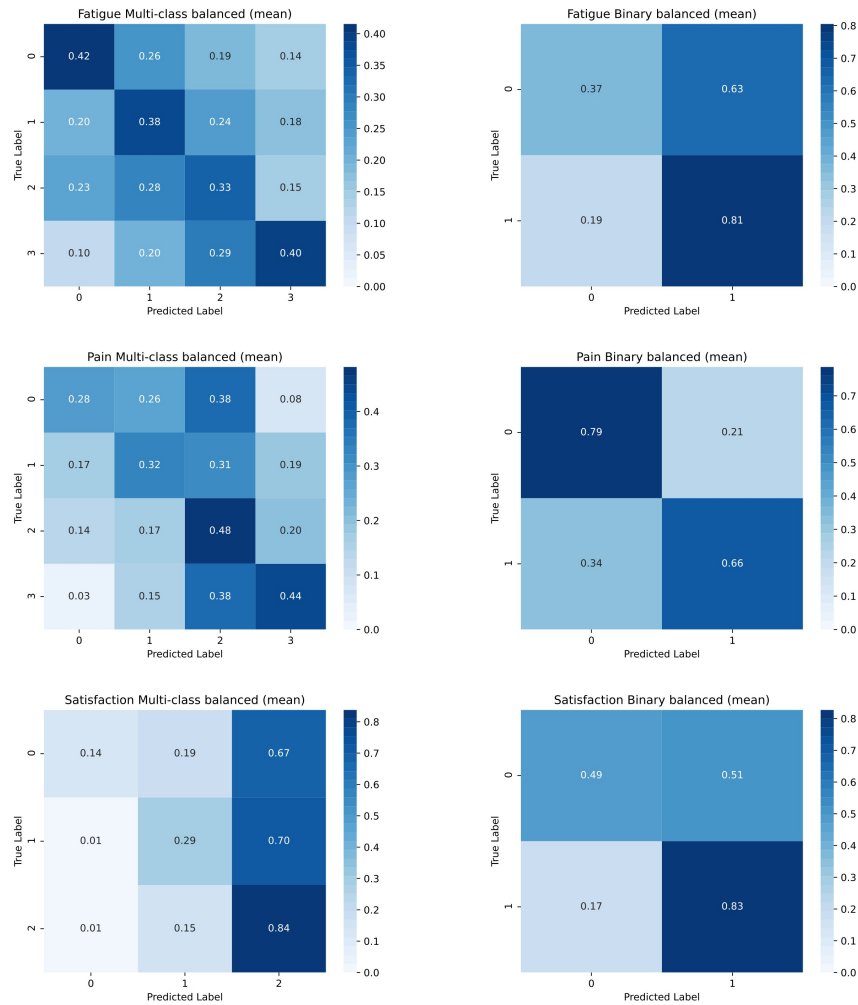
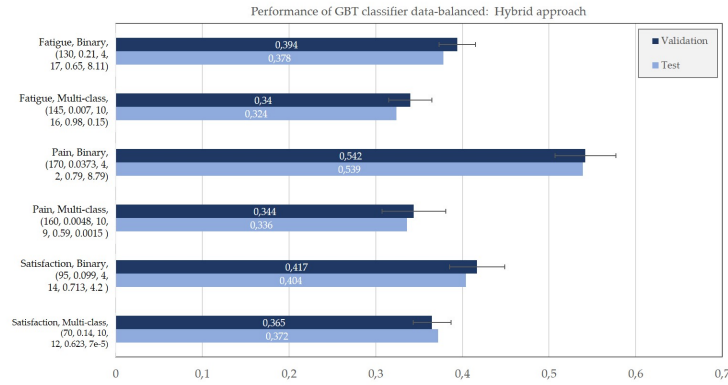


Figure A.4.2: Sentiment-related variables classification results using RF with data balanced at mean of all classes number of instances: Hybrid Approach.

A.4.2 GBT with data balancing: hybrid approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (n_estimators, learning_rate, max_depth, min_child_weight, column_subsampling, complex_penalty)*.

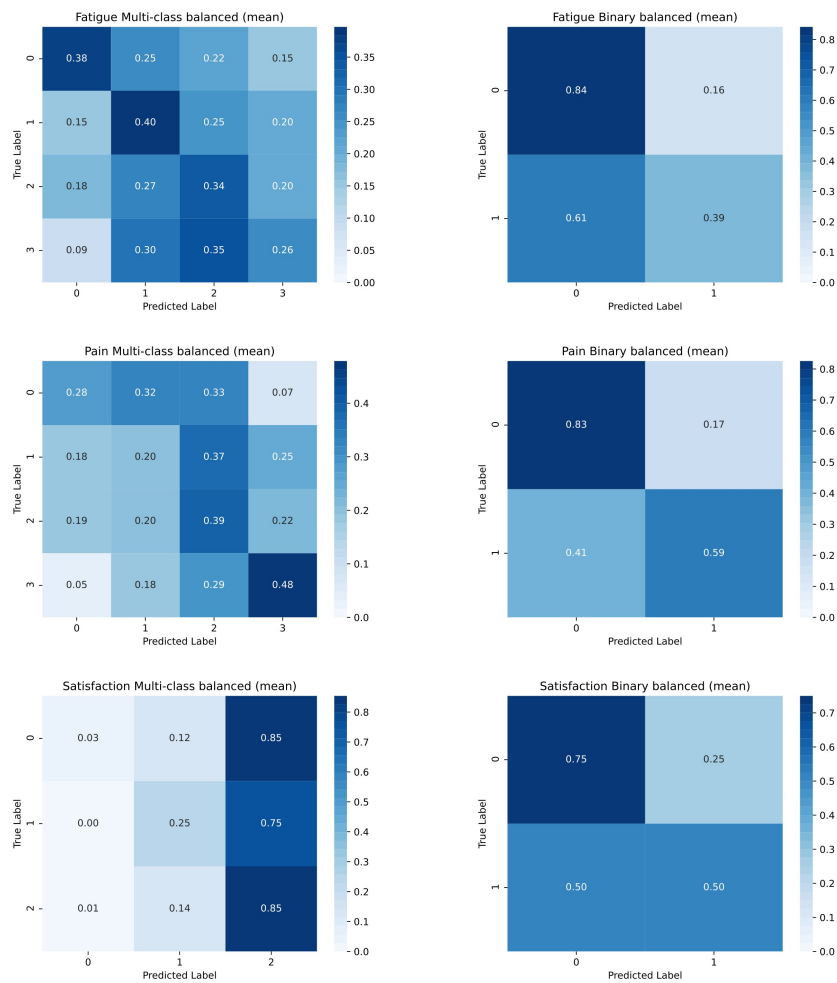
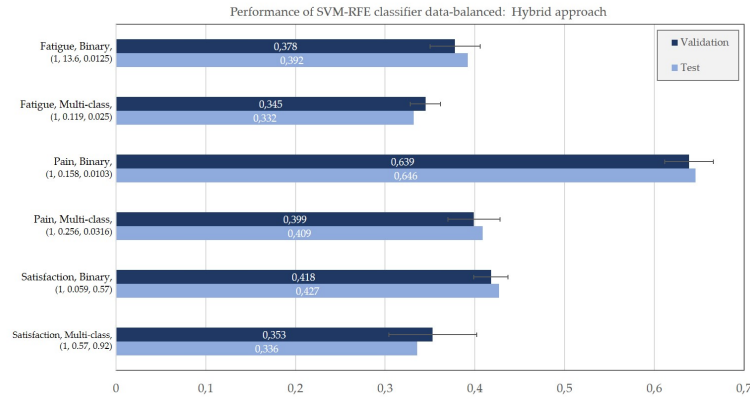


Figure A.4.3: Sentiment-related variables classification results using GBT with data balanced at mean of all classes number of instances: Hybrid Approach.

A.4.3 SVM-RFE with data balancing: hybrid approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (linear-SVM C, SVM-RBF C, SVM-RBF gamma)*.

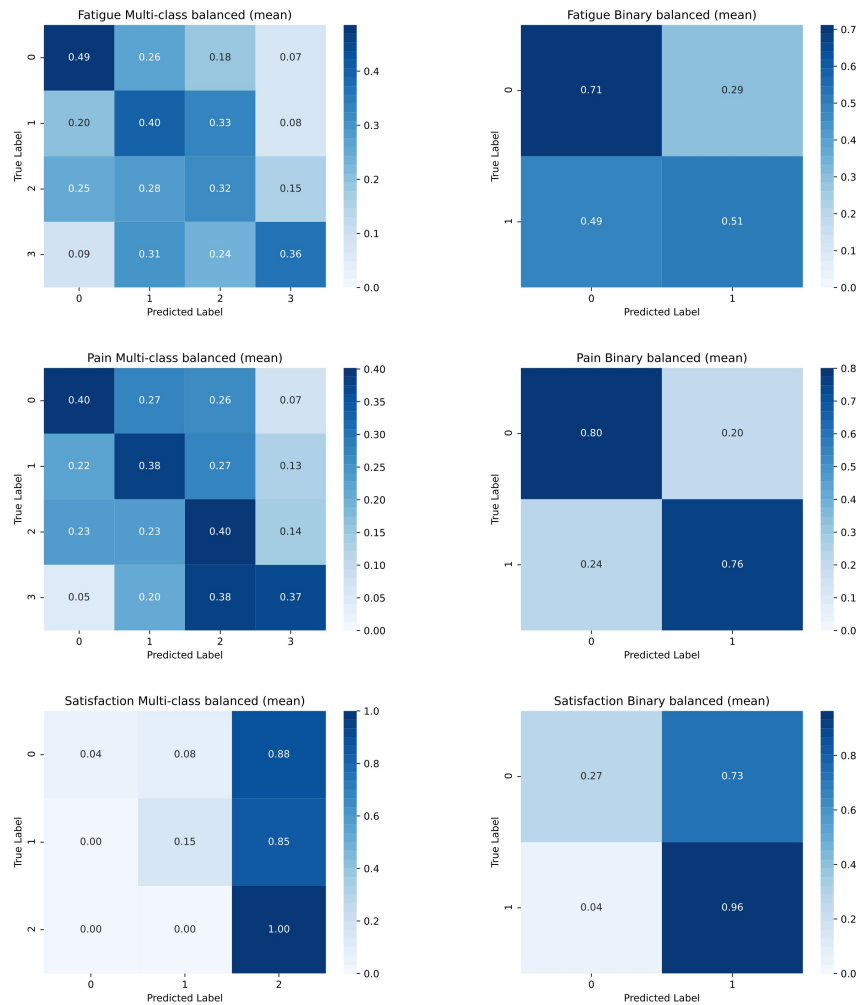
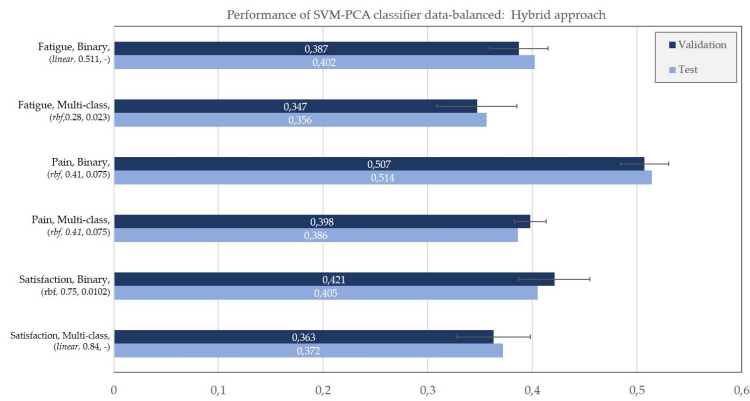


Figure A.4.4: Sentiment-related variables classification results using SVM-RFE with data balanced at mean of all classes number of instances: Hybrid Approach.

A.4.4 SVM-PCA with data balancing: hybrid approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (kernel, C, gamma)*.

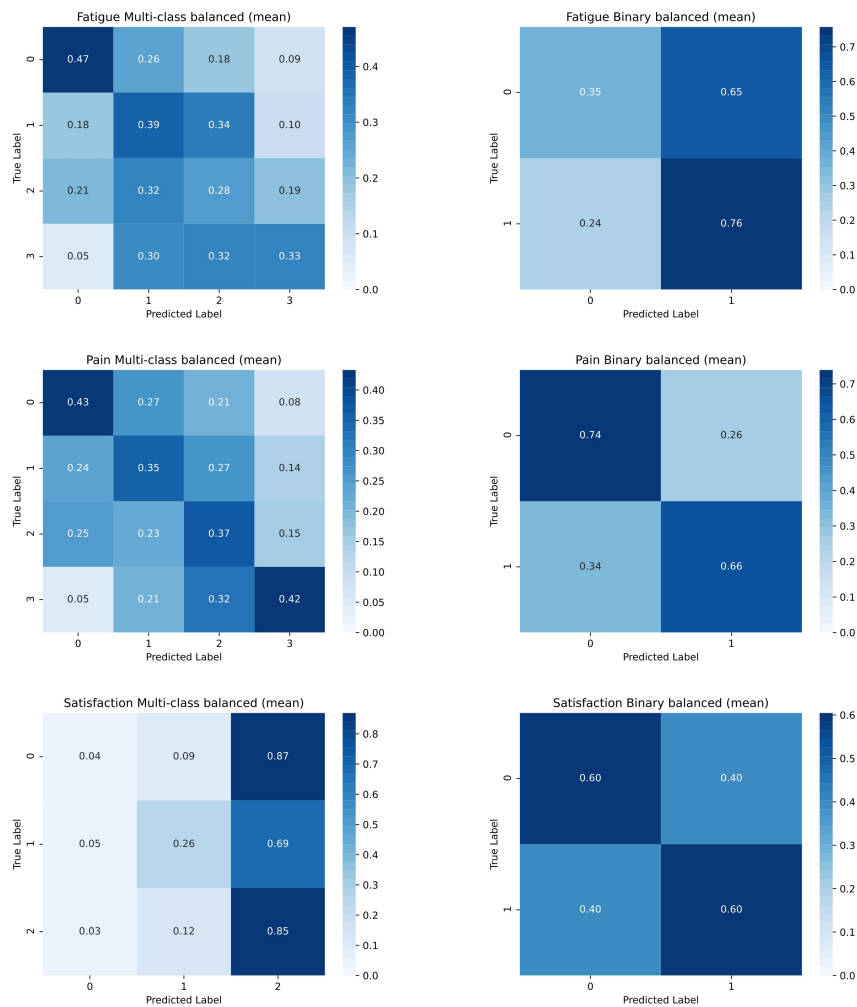
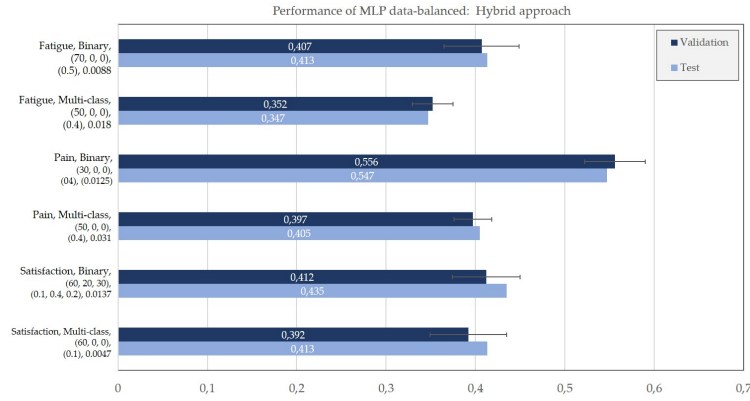


Figure A.4.5: Sentiment-related variables classification results using SVM-PCA with data balanced at mean of all classes number of instances: Hybrid Approach.

A.4.5 MLP with data balancing: hybrid approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable, classification type, (Num. units layer 1, Num. units layer 2, Num units layer 3), (Layer 1 dropout, Layer 2 dropout, Layer 3 dropout), learning rate*.

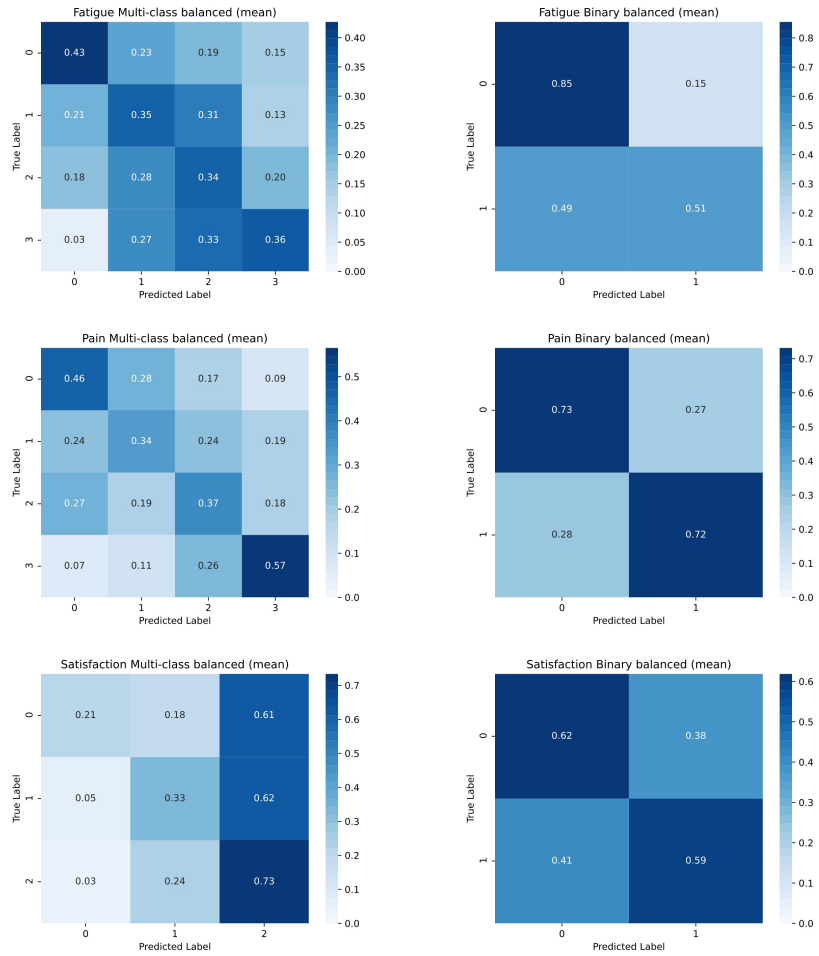


Figure A.4.6: Sentiment-related variables classification results using MLP with data balanced at mean of all classes number of instances: Hybrid Approach.

A.5 Complementary results for sentiment-related variables classification with data balancing: Undersampling approach

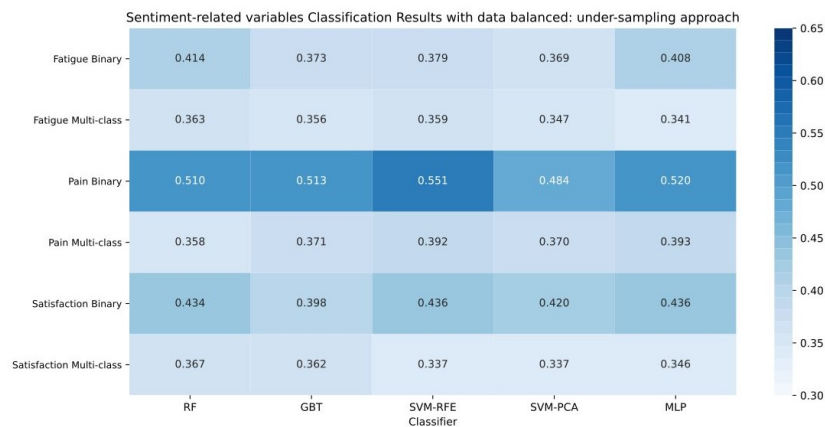
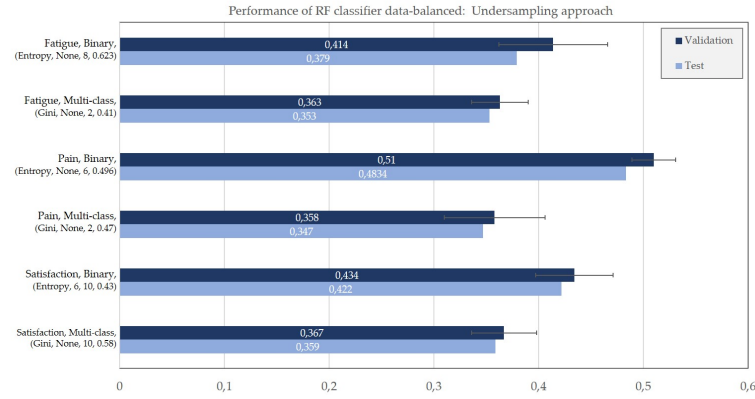


Figure A.5.1: Results of classification of sentiment-related variables using data-balance approaches at validation set. Y axis refer to the variable and type of classification, while x-axis to the different classifiers employed. Values presented correspond to the F1-score.

A.5.1 RF with data balancing: undersampling approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (criterion, max_depth, min_samples_split, max_features)*

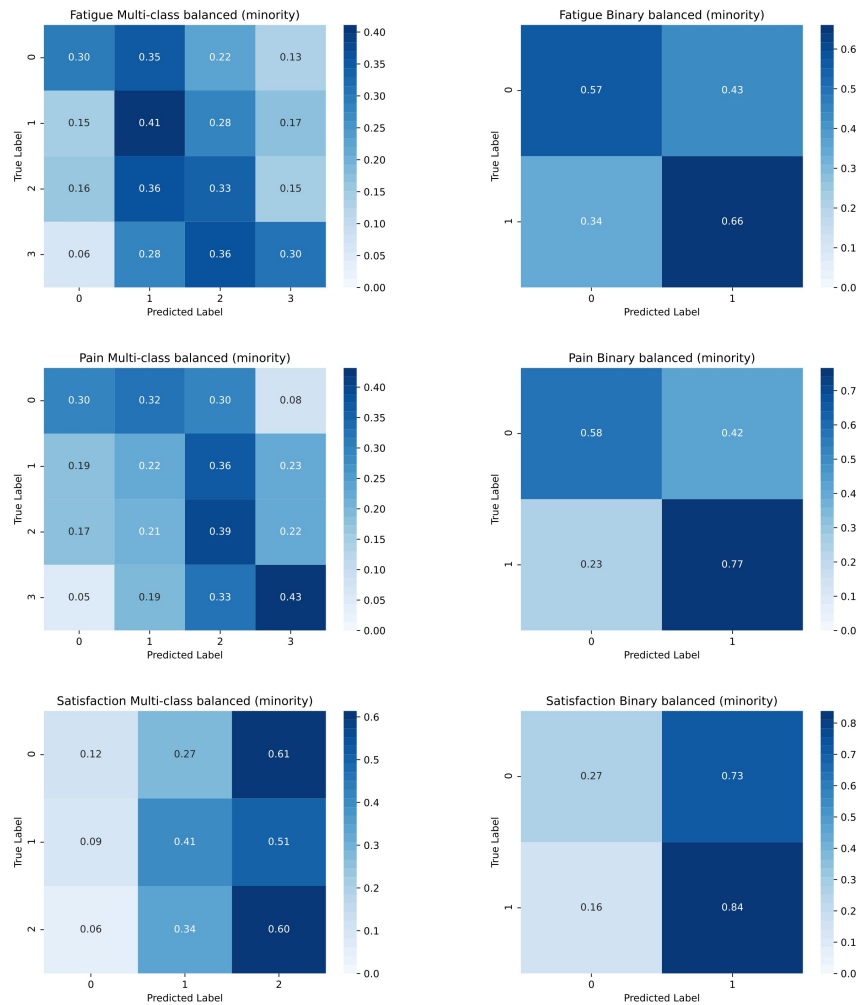
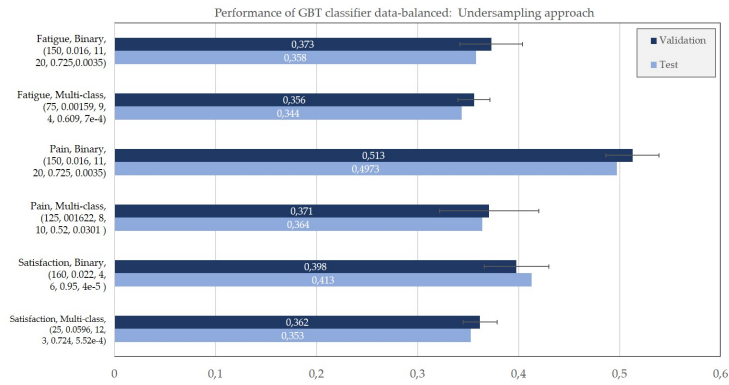


Figure A.5.2: Sentiment-related variables classification results using RF with data balanced at minority class: Undersampling approach.

A.5.2 GBT with data balancing: undersampling approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (n_estimators, learning_rate, max_depth, min_child_weight, column_subsampling, complex_penalty)*.

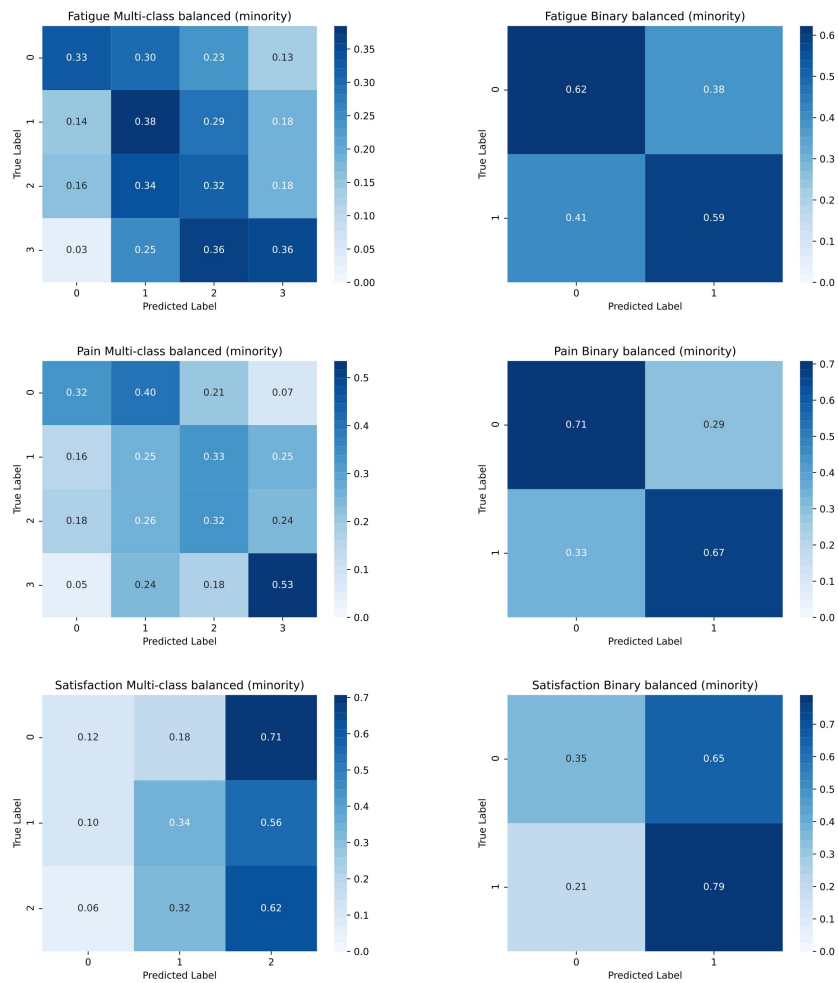
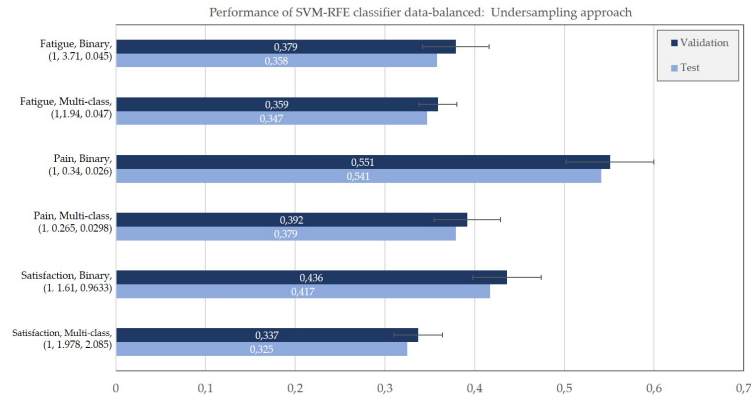


Figure A.5.3: Sentiment-related variables classification results using GBT with data balanced at minority class: Undersampling approach.

A.5.3 SVM-RFE with data balancing: undersampling approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (linear-SVM C, SVM-RBF C, SVM-RBF gamma)*.

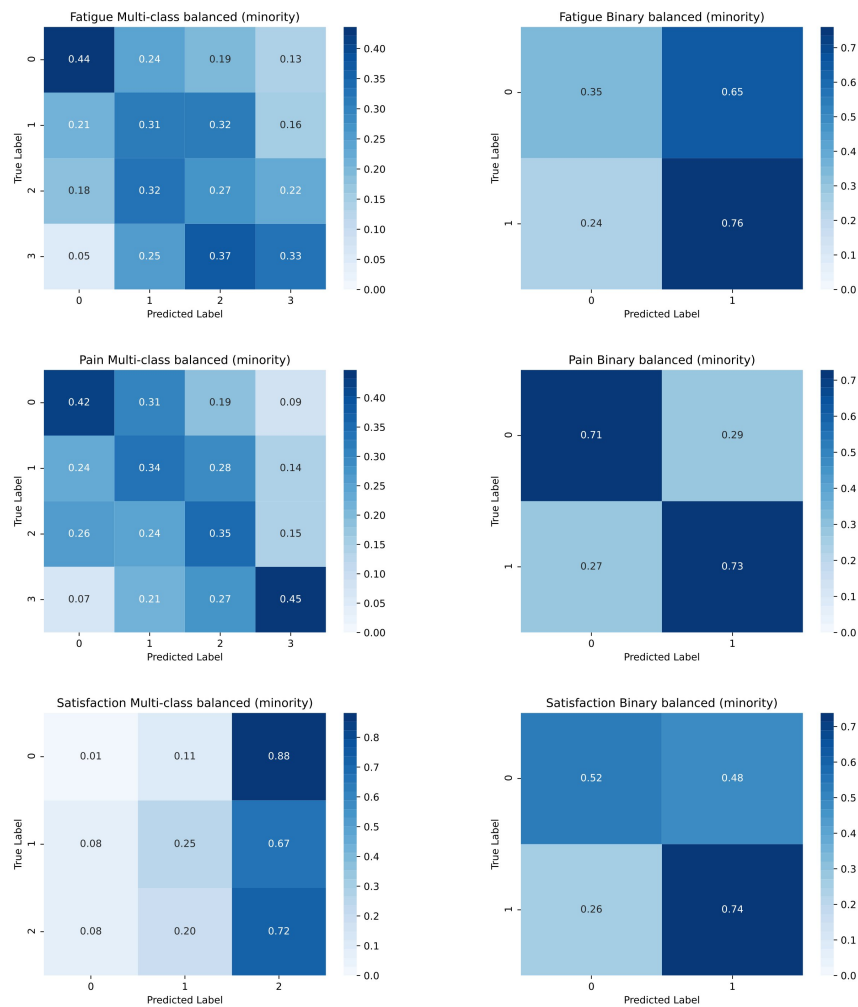
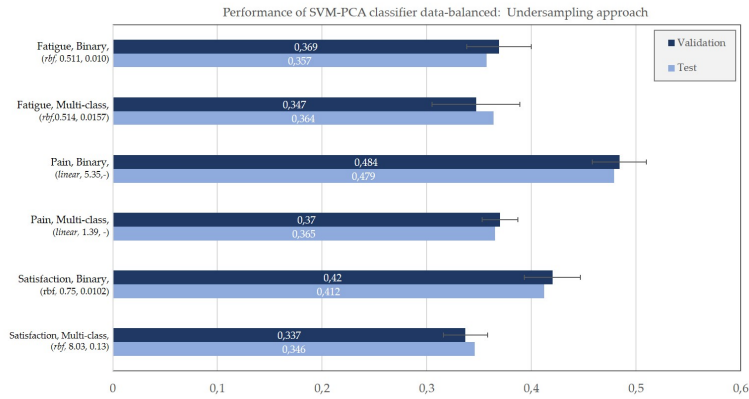


Figure A.5.4: Sentiment-related variables classification results using SVM-RFE with data balanced at minority class: Undersampling approach.

A.5.4 SVM-PCA with data balancing: undersampling approach

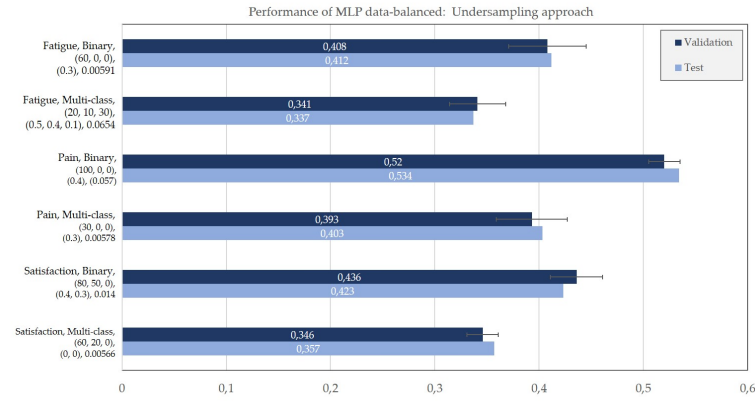


(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable to predict, classification type, (kernel, C, gamma)*.



Figure A.5.5: Sentiment-related variables classification results using SVM-PCA with data balanced at minority class: Undersampling approach.

A.5.5 MLP with data balancing: undersampling approach



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Model description follows the scheme *Variable, classification type, (Num. units layer 1, Num. units layer 2, Num units layer 3), (Layer 1 dropout, Layer 2 dropout, Layer 3 dropout), learning rate*.

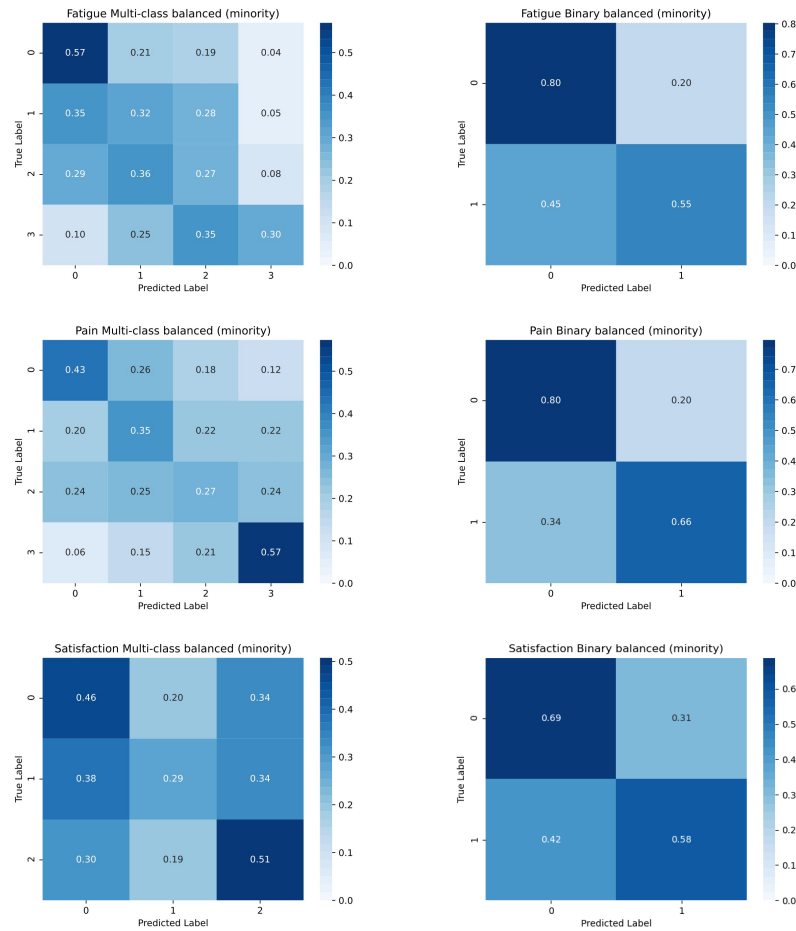


Figure A.5.6: Sentiment-related variables classification results using MLP with data balanced at minority class: Undersampling approach.

A.6 Complementary results for sentiment-related variables classification using Ordinal classification techniques

A.6.1 Ordinal Classification: Simple Approach by the application of binary classifiers

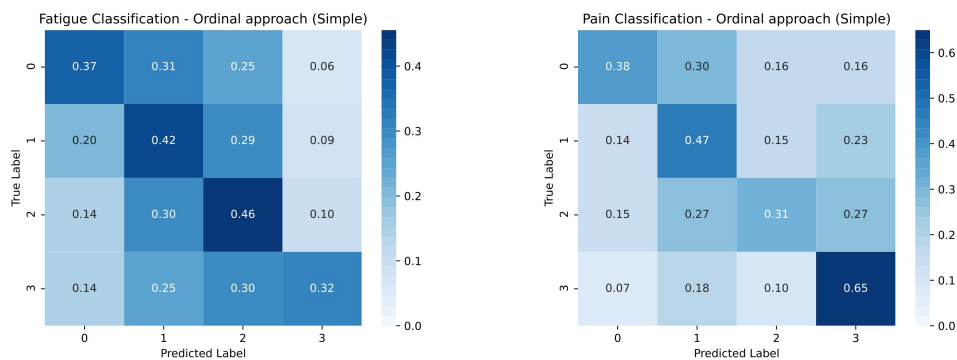


Figure A.6.1: Simple approach: application of RF binary classifiers

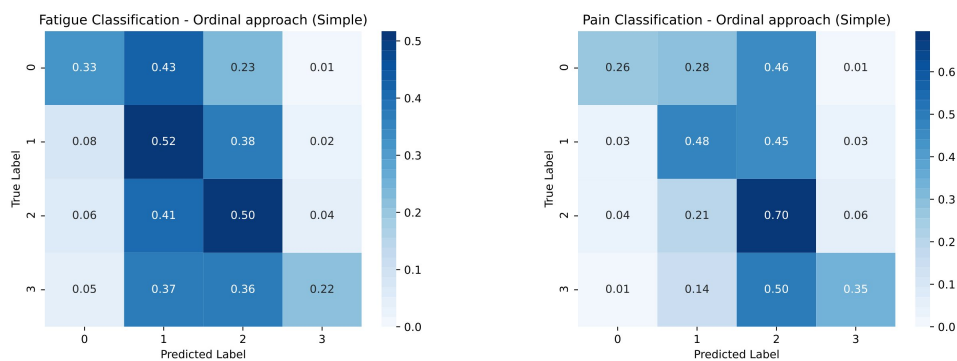


Figure A.6.2: Simple approach: application of GBT binary classifiers

A.6.2 Ordinal Classification: a Neural Network Approach

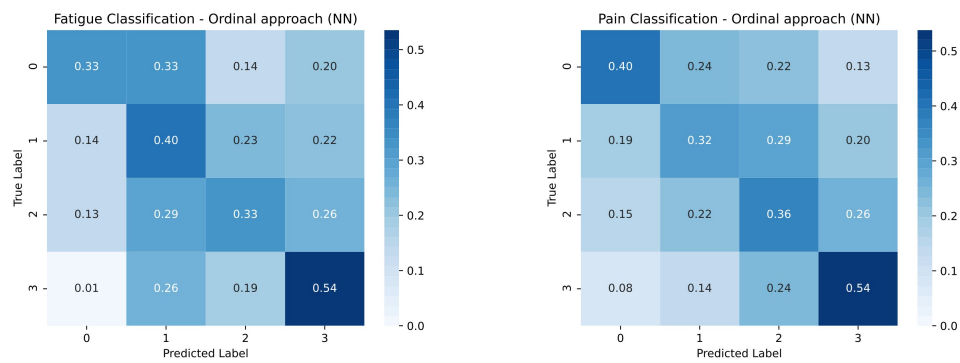


Figure A.6.3: Confusion matrices of Ordinal Classifier: An NN approach

A.7 Complementary results for sentiment-related variables classification using a sequential approach (LSTM)

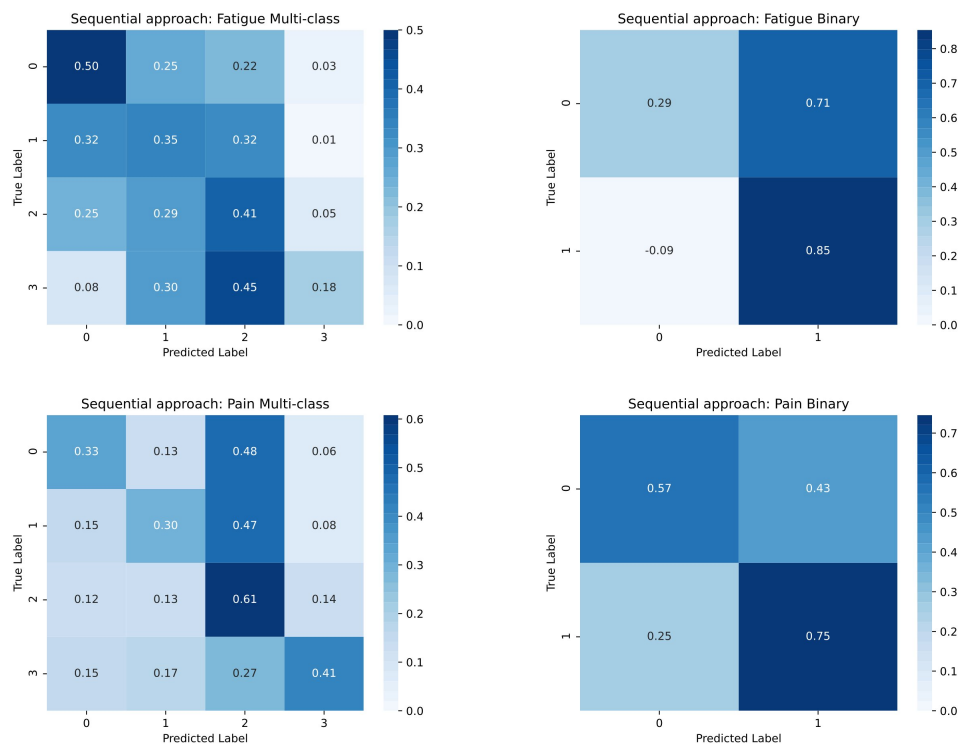


Figure A.7.1: LSTM complementary results: 10 sessions with overlap.

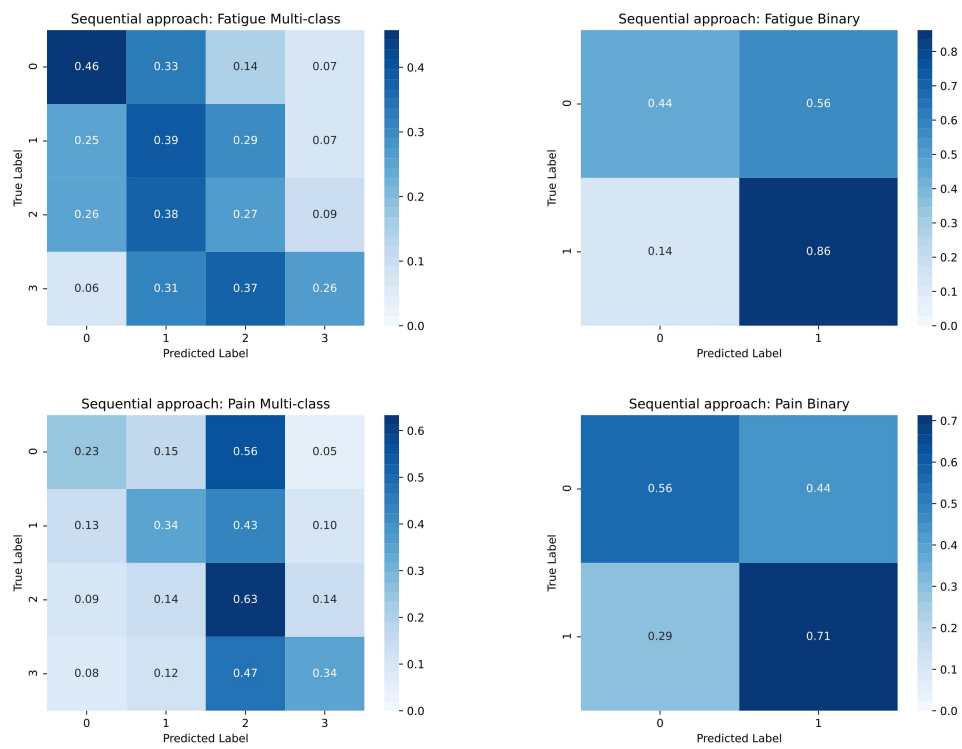


Figure A.7.2: LSTM complementary results: 5 sessions without overlap.

Appendix B

Complementary results of models for emotional-state classification

In this Appendix, complementary results of the emotional-state inference chapter are presented. To note that the confusion matrices of the classification task are normalised over the true labels (rows); the predicted classes are along the columns. Comparative graphs that comprise test and validation results are also included. Regarding the tuning parameters, they follow the scheme of the main chapter.

B.1 Emotional-state representation by K-means

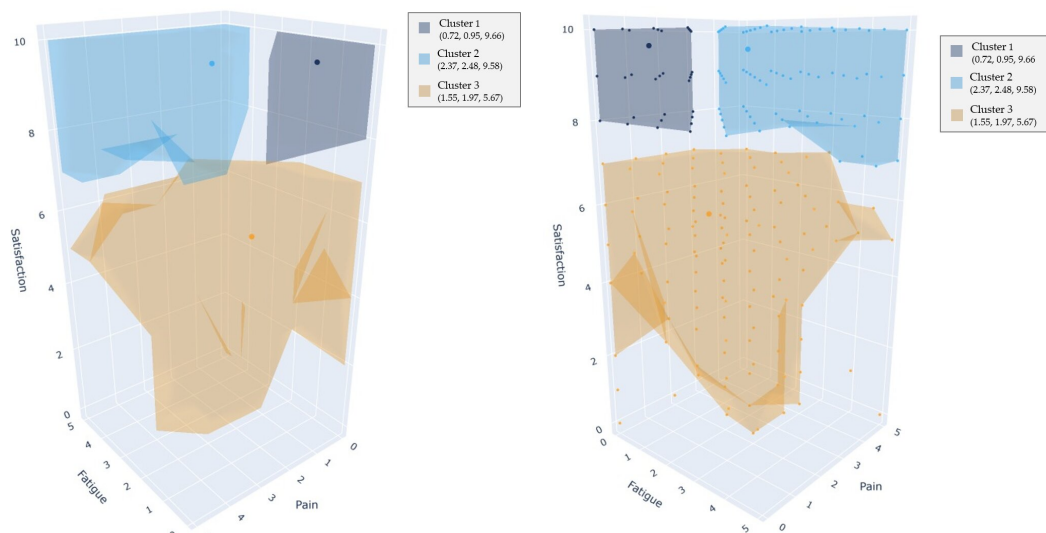


Figure B.1.1: Emotional representation by K-means with 3 clusters.

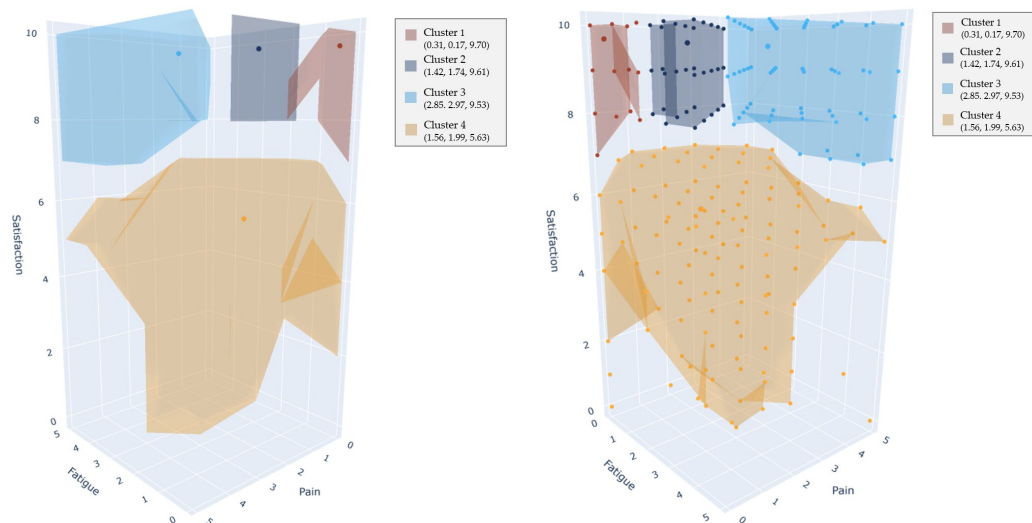


Figure B.1.2: Emotional representation by K-means with 4 clusters.

B.2 Emotional-state representation by DP-means

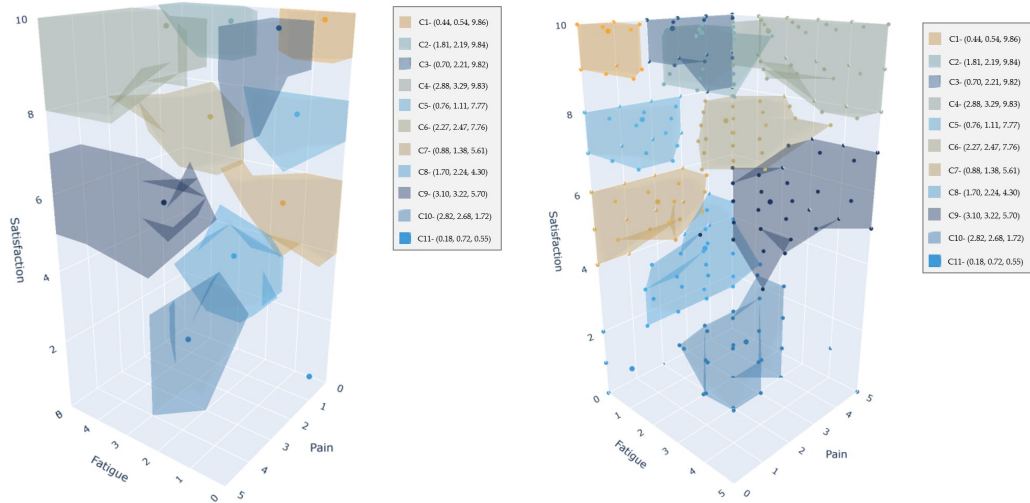


Figure B.2.1: Emotional representation by DP-means: Regularisation parameter fixed at 15.

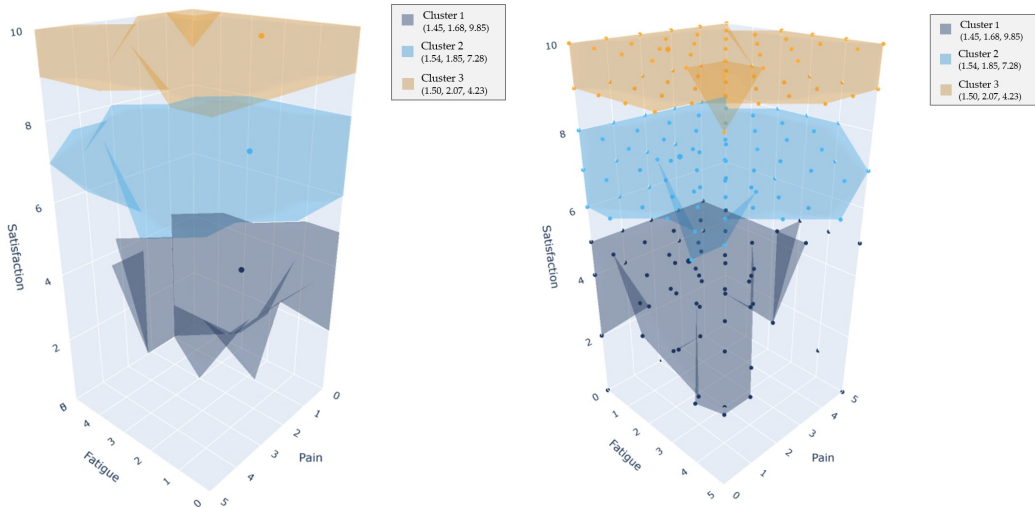


Figure B.2.2: Emotional representation by DP-means: Number of clusters set at 3.

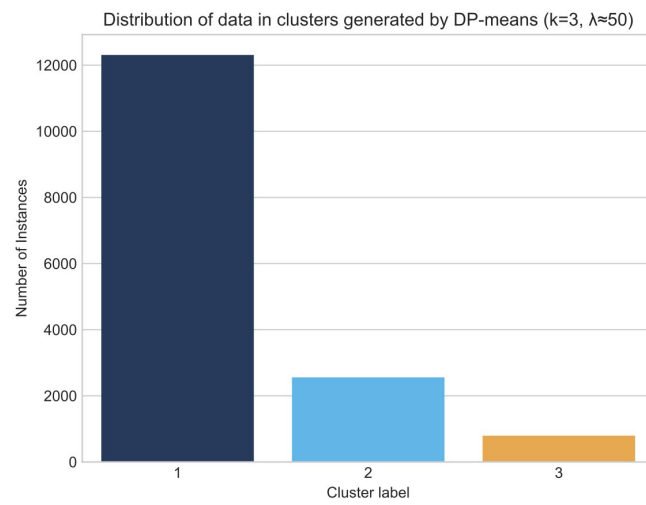


Figure B.2.3: Distribution of data in clusters generated by DP-means (number of clusters set at 3), aiming to find a emotional-state representation.

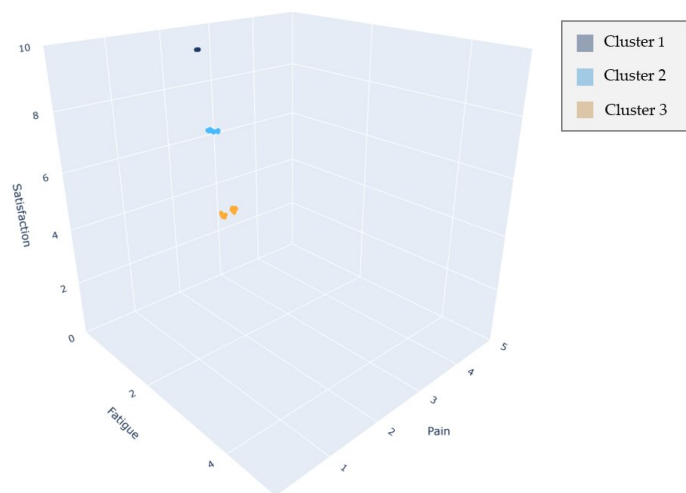
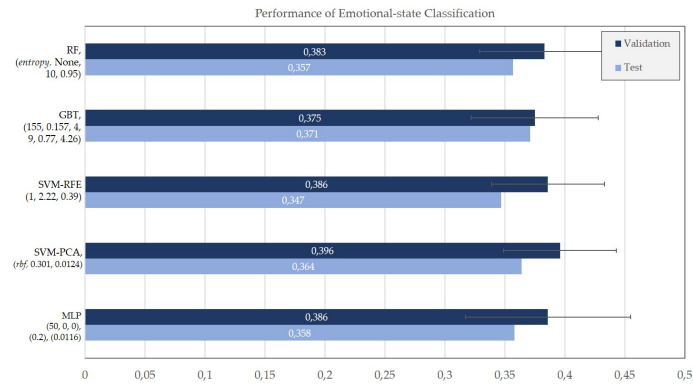


Figure B.2.4: Analysis of input sequences order effect in DP-means (number of clusters set at 3): representation of cluster centroids over 30 iterations.

B.3 Emotional-state Multi-class Classification

B.3.1 Emotional-state Multi-class Classification using baseline classifiers



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Models parameters follow the scheme of previous sections.

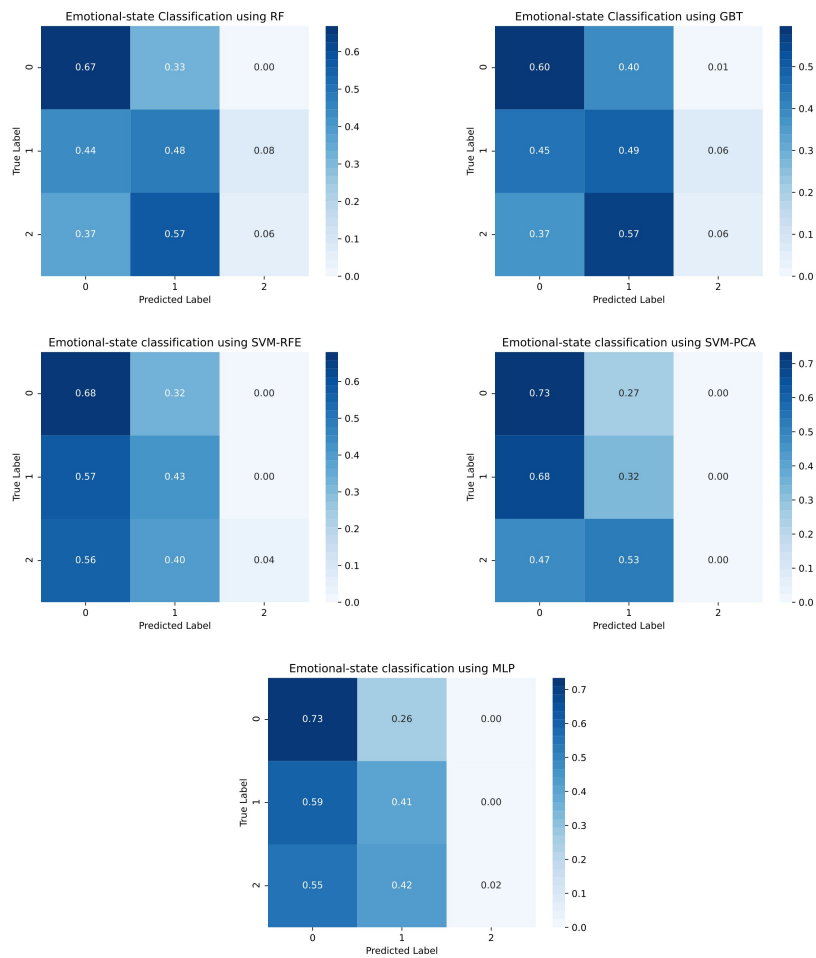
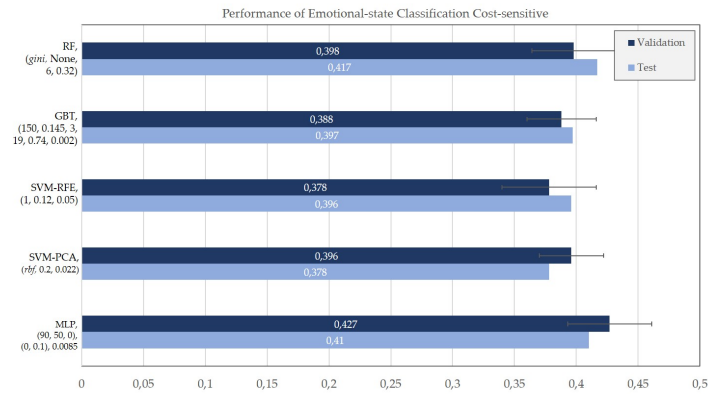


Figure B.3.1: Emotional state classification results using baseline classifiers.

B.3.2 Emotional-state Multi-class Classification using cost-sensitive classifiers



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Models parameters follow the scheme of previous sections.

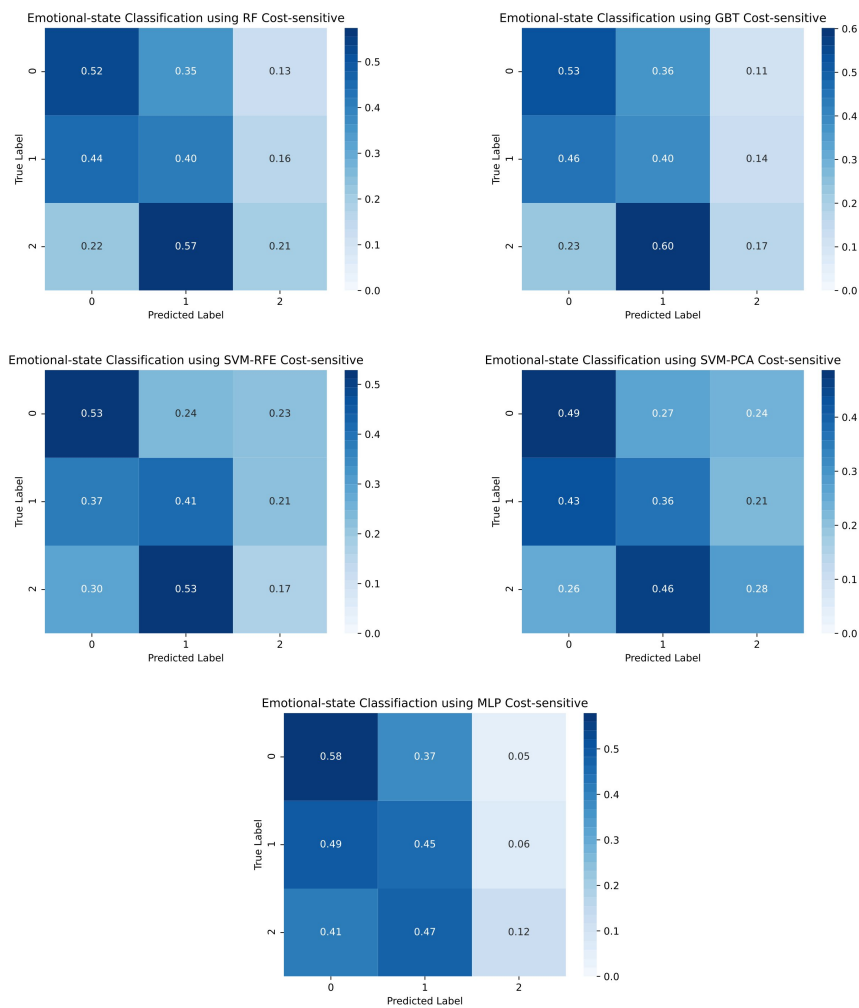
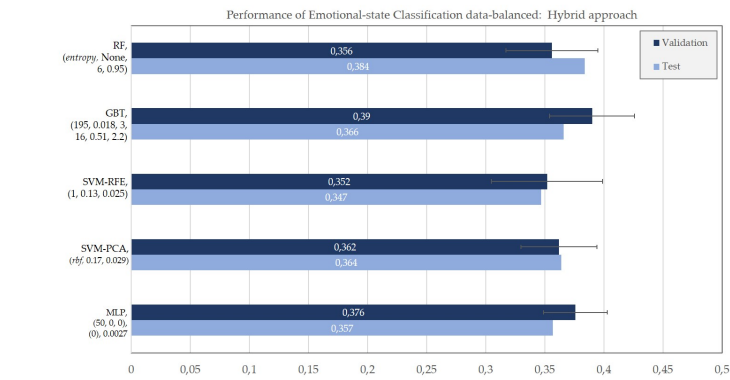


Figure B.3.2: Emotional state classification results using cost-sensitive classifiers.

B.3.3 Emotional-state Multi-class Classification with data balanced at the mean number of instances: Hybrid Approach.



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Models parameters follow the scheme of last sections.

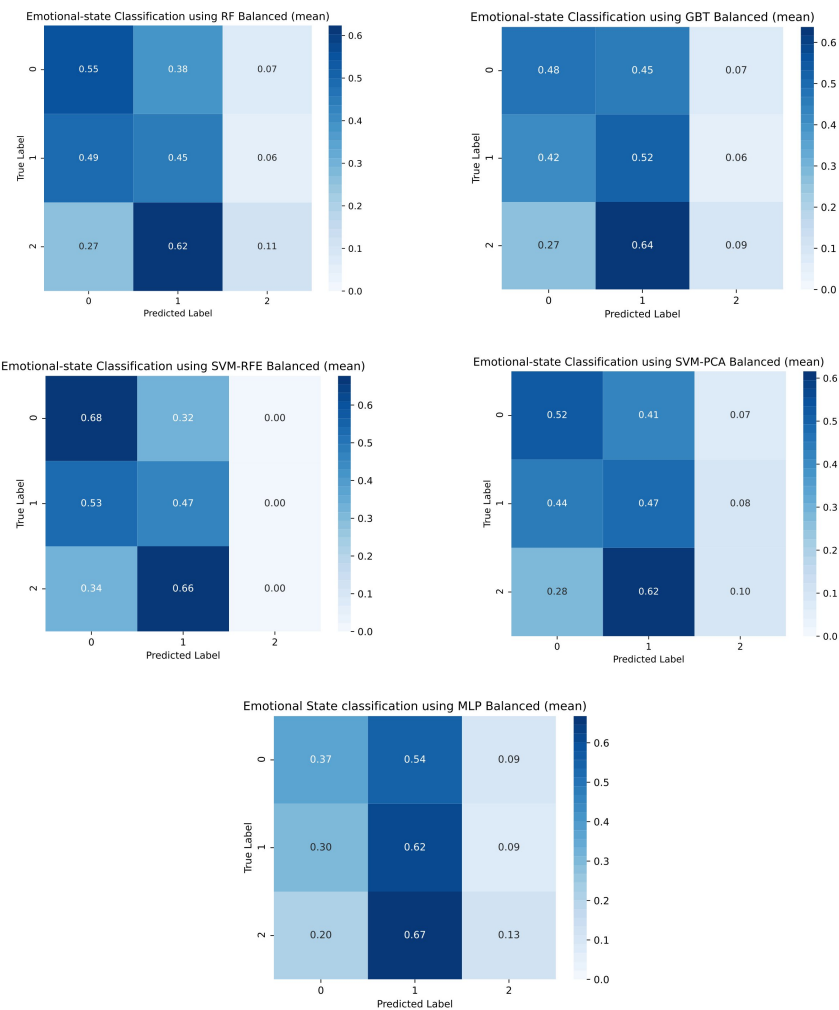
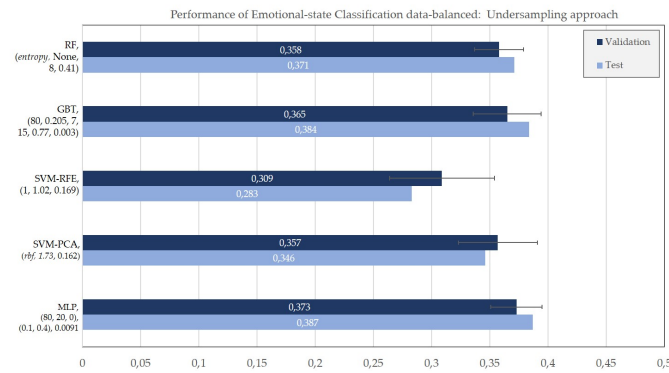


Figure B.3.3: Emotional state classification results with data balanced: Hybrid approach.

B.3.4 Emotional-state Multi-class Classification with data balanced at minority class: Undersampling Approach.



(a) X-axis presents the F1-scores. Validation results are reported as mean $\pm$ SD. Models parameters follow the scheme of last sections.

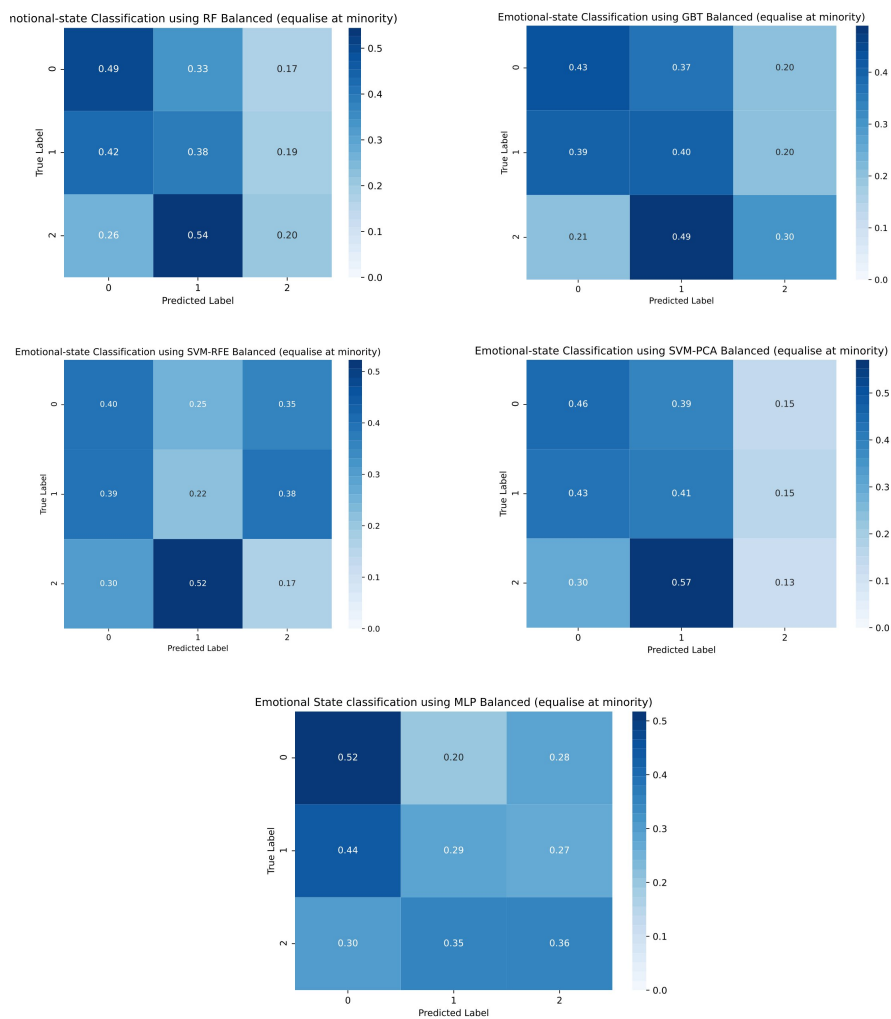


Figure B.3.4: Emotional state classification results with data balanced: Undersampling approach.