Deep learning facial expression recognition for pain assessment

Aaron Chen
St Peter's College

4^{th}-Year Project Report

Supervised by
Professor Mauricio Villarroel

Department of Engineering Science
University of Oxford

Trinity Term, 2024

Acknowledgements

First and foremost, I would like to thank my supervisor, Professor Mauricio Villarroel for his invaluable guidance, patience, and encouragement throughout the past 8 months. I feel extremely grateful and honoured to have had the opportunity to work under his supervision. His wisdom and enthusiasm have not only inspired me to pursue a career in this field but have also made me a better engineer. I am also thankful to everyone in the research group for their selfless support and life advice.

Thanks for all of my marvelous friends, Ebed, Eunsoo, Tina, James, Ploy, for the countless moments we have shared together, and for enduring all the rants I had. A special thank goes to Nicolas from the Berks Group, Department of Biochemistry, for all the social events and emotional support during the most challenging times of the year.

And finally, I must thank my brother, Jonathan, and my parents, Brenda, and Ping, for always being there and believing in me, despite being spread across three different continents.

Abstract

Pain assessment remains a significant challenge in clinical settings, especially for patients who cannot communicate their discomfort. This report develops and evaluates two automatic pain assessment models, using facial expression analysis, aiming to enhance accuracy and reliability in pain detection.

The first model employs handcrafted feature extraction to analyse specific facial micro-movements associated with pain expressions, such as nasolabial folds, lowering eyebrows, and mouth closures. The model showed great predictive performances with 75.0% accuracy in classifying between no pain and the highest level of pain. Explainability AI methods supported the selection of handcrafted features and suggested more emphasis on mouth movements, such as widening and opening of the mouth.

The second model utilises 3D Convolutional Neural Network (3D-CNN) architectures, namely Res3D and R(2+1)D, with novel preprocessing techniques to improve performance by minimizing data redundancy. The models were rigorously tested, demonstrating their potential to outperform traditional pain assessment tools. Guided Grad-CAM was integrated to provide deeper insights into the features that contribute the most to pain assessment. The results agreed with the first model, with emphasis on the nasolabial folds but provided less attention to the mouth itself.

This research advances the field of pain measurement by assessing the viability of current pain scales, namely the Prkachin and Solomon Pain Intensity (PSPI) scale. It provides further insights into how pain is manifested from facial movements and suggests a change in the current pain scales with more attention on the mouth and the inclusion of the duration of facial movements. This study also presents the limitations of making automatic pain assessment tools clinically ready. It paves the way towards a scalable product that can be integrated into both clinical and domestic environments, achieving a patient-oriented future.

Table of contents

List of figures

List of tables

Glossary

AU	Action Unit
AUC	Area Under the Curve
CNN	Convolutional Neural Network
ECG	Electrocardiogram
EEG	Electroencephalography
EMG	Electromyogram
FACS	Facial Action Coding System
Grad-CAM	Gradient-weighted Class Activation Mapping
HOG	Histogram of Oriented Gradients
PAT	Pain Assessment Tool
PCA	Principal Component Analysis
PSPI	Prkachin and Solomon Pain Intensity
ReLU	Rectified Linear Unit
RF	Random Forest Classifier
ROC	Receiver Operating Characteristic
SCL	Skin Conductance Level
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
XAI	Explainable AI
XGBoost	eXtreme Gradient Boosting

Chapter 1

Introduction

1.1 Clinical motivation

Pain is defined by the International Association of the Study of Pain (IASP) [1] as "*an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such tissue damage*". It is a distinctive, personal experience that can greatly vary according to the patient's life experiences or other psychological and social factors. There are two main types of pain, Acute and Chronic Pain, with duration as the primary difference between the two. Acute Pain is often nociceptive pain that originates in the peripheral nervous system and lasts no more than three to six months. It is relatively easy to identify and treat as the pain is often associated with a cause and goes away when the underlying cause is healed. On the other hand, when the pain persists beyond the expected healing time of over three months, it is classified as Chronic Pain [2]. It is often accompanied by more significant burden and complexity due to its uncertain origin.

Chronic pain has been a substantial challenge to the global healthcare systems, with high prevalence among the worldwide population. In 2017, the Health Survey for England reported a nationwide prevalence of over 34% of all ages, with over half of the respondents over the age of 75 suffering from long-term chronic pain [3]. Other studies have reported even more daunting results, with over one-third to one-half of the UK population suffering from chronic pain, that is, over 28 million adults, with the number increasing with an ageing population [4]. The Centers for Disease Control and Prevention (CDC) in the US reported an estimated prevalence of 20.9% in the US [5] and 31.54% in China's 31 provinces in a community study in 2020 [6].

Chronic pain, specifically chronic lower back pain, has been the leading contributor to the global population's years lived with disability measurement in a recent study [7]. It often comes with physical and emotional consequences, including fatigue, irritability, and depression. More significantly, there have been links established between Chronic Pain with dementia [8], higher suicide risk [9], and increasing use of substances such as cannabis and alcohol [10]. Chronic Pain is now not only a clinical problem but also a global economic and social problem. It is now

widely considered a disease instead of a symptom, with the World Health Organization (WHO) including chronic pain as an individual category in the new 11th edition of the International Classification of Diseases (ICD-11) [11].

With such prevalence and impact, providing accurate and effective pain management plans is crucial. However, proper management of pain is often not delivered. A survey suggested that over 50% of chronic pain patients have not received adequate pain management [12, 13, 14], with factors including underestimation of pain, lack of training, as well as social and psychological problems. Lack of pain assessment training and staff shortage were other main factors for inadequate pain management. In Intensive Care Units (ICU), patients' pain should be assessed routinely; however, due to lack of training, more than half of the nurses were reported to not use any assessment tool for patients who were not able to self-report [15]. Therefore, there is a need for an accurate and effective method of assessing pain to aid diagnosis and provide adequate pain management.

1.2 Pain assessment

Most of the pain assessment tools used in clinical practice require good communication skills both from the caretakers and the patients to get an accurate assessment. The gold standard for pain assessment is self-reported pain assessment, which includes verbal description and unidimensional scale ratings. Analogies are commonly used for first screening but lack comparability and contain potential biases from patients' psychological status as well as pain catastrophising. The National Health Service (NHS) suggests the use of the Numerical Pain Rating Scale (NPRS) [16] or the Visual Analogue Scale (VAS) [17] to quantify pain on a scale of 0-10, with instructions and questions such as ”Is your pain mild, moderate, or severe?” and ”Please describe your pain on a scale of zero to ten”. For simpler communications, especially for school-age patients, visual aids were used on top of the analogue scale. For example, the Wong-Baker FACES Pain Rating Scale [18] puts cartoon-like faces or emojis for each rating on the analogue scale and has been shown to be effective across different gender and ethnic groups [19].

For individuals who cannot express their pain directly, such as infants or people with cognitive impairment, the current pain assessment tools are less effective and sometimes misleading, resulting in improper pain treatment[20]. The best method currently to assess pain for these vulnerable patients is observer-based scales where a third party, mostly a nurse or caretaker,

assesses pain through behavioural indicators such as facial grimacing, crying, or specific body language that suggests the presence of pain. There have been numerous studies suggesting a positive correlation of the observer-based scales with the gold standard self-reported scales, however, the ability of observer-based methods to identify and separate pain levels is still unclear, and depends highly on the methods used and the settings.

With the advancement of Machine Learning methods, automatic pain assessment approaches with the use of behavioural and facial cues would play a pivotal role in providing an accurate, objective, and reliable pain assessment. With the rise of the new healthcare model putting more emphasis on home care and Information and Communications Technology (ICT) usage, such new pain assessment tools can be integrated into homes and families to help educate the public about long-term chronic pain and act as a screening process for primary care.

1.3 Objectives

The following are the main objectives of this report:

1. To assess and evaluate the facial expressions and micro-movements that best reflect the experience of pain using handcrafted features.
2. To develop a robust and reliable deep representation learning method for automatic pain assessment.
3. To explore the interpretability of automatic pain assessment methods to provide more context to clinicians.

1.4 Outline of document

Chapter 2 reviews previous literature and the current state-of-the-art methods used for automatic pain assessment. Chapter 3 provides a detailed description of the dataset used for this project. Chapter 4 describes a low-level feature approach utilising handcrafted facial features. Chapter 5 presents a novel deep learning architecture model for automated pain assessment. Chapter 6 provides concluding remarks and limitations of the study and presents potential future research opportunities on automated pain assessment.

Chapter 2

Literature review

2.1 Introduction

The main approaches for automatic pain assessment can be separated into two categories, physiological and behavioural. Physiological indicators include vital signs such as pulse rate, temperature, and respiratory rate as well as bio-potential signals (i.e. Electroencephalography (EEG), Electrocardiogram (ECG), Electromyogram (EMG)). These modalities have been proven to have strong correlations to patients' mental stress levels that can be induced by pain [21]. However, these methods require accurate measurement from the sensors in very controlled settings and increase the difficulty in acquiring large-scale datasets.

Behavioural methods prioritise visible indicators of pain, similar to the decision-making process of the observer-based pain scales in clinical settings. These visible indicators involve facial expressions, body movements, and audio, which allow for easier capture of signals without a need for signal sensors, enabling a vast amount of research with the introduction of multiple datasets. In this report, I focused primarily on a behavioural-based approach with facial expressions.

2.2 Facial pain assessment

Facial expressions are the most common and effective modalities in quantifying pain due to a higher understanding of the mechanisms and their expressiveness. The Facial Action Coding System (FACS) [22] was developed in 1978 aiming to gain further understanding of the movement of facial muscles in relation to different emotions. The FACS system breaks a face down into different Action Unit (AU), and by scoring changes of each AU based on intensity, it allows quantification of emotions by utilising empirically derived equations with combinations of AU. Figure 2.1 illustrates 30 example AUs in the FACS system. The expression of pain, too, was decoded using the FACS AU and formulated with The Prkachin and Solomon Pain Intensity (PSPI). These AU include brow lowering (AU4), cheek raising (AU6), lid tightening (AU7), nose wrinkling (AU9), upper lip raising (AU10), and eye closing (AU43). This subset of AU was each

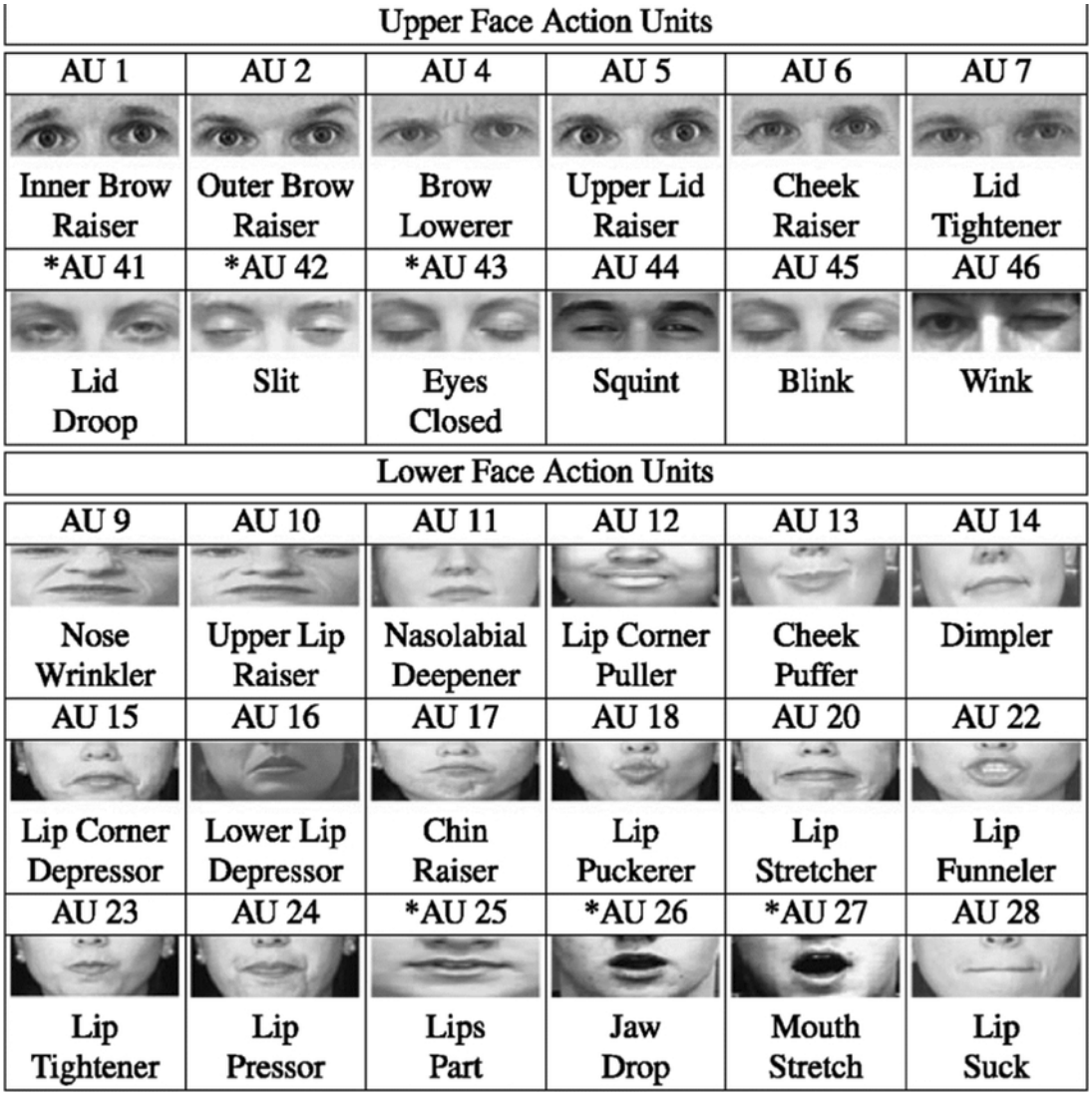


Figure 2.1: FACS Action Units. Reprinted, with permission, from [23].

scored on a scale of 0-5 with 0 as absence and 5 as maximum and was used to quantify pain with the PSPI equation defined as:

$$Pain = AU4 + (AU6 \| AU7) + (AU9 \| AU10) + AU43 \tag{2.1}$$

This pain intensity scale allows an objective method to quantify pain purely from facial cues and was commonly used in most pain datasets. However, some researchers were still reluctant to use PSPI as the ground truth. Firstly, PSPI was derived statistically on a sequence level; hence, a frame-level classification of pain with PSPI has not been justified as suggested by Werner et al. [24] Moreover, PSPI only considers the facial reaction of pain, but lacks the direct link to the experience of pain. Some patients might still experience pain with little to no

expressions while the PSPI intensity would be 0. Therefore, it should not be used as an estimate of the intensity of pain but instead of the intensity of facial reaction to pain. Finally, there is no evidence that the AU considered in the PSPI equation is the only facial feature related to painful expressions. For example, Kunz et al. [25] suggested that more AU that were left out by the PSPI equation, such as the raising of eyebrows (AU1, 2) or mouth opening (AU 25, 26, 27) should all be considered as painful expression components. The advance in deep learning, especially Convolutional Neural Network (CNN), allowed the learning of the entire facial features and was considered by researchers as a solution to the challenges above.

Current vision-based automatic pain assessment methodologies can be separated into two categories: Non-temporal (frame level classification for static images) and Temporal (sequence level classification for facial analysis in videos) methods. Both methods involved one crucial step: face extraction or landmark extraction.

2.3 Face detection

Almost all of the methods for automatic pain assessment involve a crucial preprocessing step, face or landmark extraction. The data from clinical datasets were typically captured to include as much information as possible, with minimal processing, often retaining the participants' surroundings and upper body. This could distract the representation learning process, resulting in sub-optimal model performance. Therefore, one essential step was extracting and aligning the faces in the data before feeding into the model.

The development of reliable face detectors is itself a complex task with heavy investments from big companies such as Google and Meta, and is out of the scope of this report. Here, I present some popular and state-of-the-art publicly available face detection and landmark detection algorithms, namely, BlazeFace [26], RetinaFace [27], and YuNet [28].

2.3.1 BlazeFace

BlazeFace, developed by Bazarevsky et al. (2019) at Google [26], is designed for real-time face detection on mobile devices, emphasising speed and efficiency. The model's core is a convolutional neural network constructed with BlazeBlocks. It separated conventional convolutional layers by a depth-wise convolution and a point-wise convolution that significantly reduced the latency without too much performance degradation. A traditional convolutional layer with input

tensor size $s \times s \times c$ with kernel $k \times k$ for d output channels would require s^2ck^2d multiplication operations. The number of computations is significantly reduced when separated into depth and point-wise layers. The combination would only require $s^2ck^2 + s^2cd$. Furthermore, 5×5 convolutional layers are used to increase the receptive field of the model, increasing the detection performance.

2.3.2 RetinaFace

RetinaFace [27] is a deep learning-based face detector, introduced by researchers from InsightFace. This model is considered the state-of-the-art face detector for its robust performance in detecting faces across various challenging scenarios, from face occlusion to different illuminations. The model uses a feature pyramid network (FPN) and a ResNet backbone, which allows for efficient face detection at multiple scales and in different orientations.

2.3.3 YuNet

YuNet [28], developed by OpenCV, is a more recent and lightweight face detection model. Similar to BlazeFace, it also utilises depthwise convolutional blocks to enable real-time face detections on mobile CPUs. However, the model does lack training on different facial orientations, hurting its performance in complex situations.

2.4 Non-temporal pain assessment

Non-temporal approaches focus on image analysis to identify pain on a frame-by-frame level. For the past few years, with the rise of computer vision and CNN, promising results were achieved by researchers. State-of-the-art result was achieved by combining the low-level handcrafted features with the deep CNN features at the output of the last convolutional layer before passing to a deep neural network for classification. Egede et al. [29] took key facial landmarks and the Histogram of Oriented Gradients (HOG) descriptors with the 4096 dimensional flattened feature vectors from pretrained VGG-16 [30] and ResNet-50 [31] model, outperforming methods using purely CNNs. Other researchers built upon this concept of using both CNNs and handcrafted facial information. One popular approach took advantage of the attention framework that prioritises regions or ”pay attention” to the areas of interests that best reflects pain. Xin et al. [32] implemented the attention mechanism with a 9-layer CNN, yielding a 19% increase in overall accuracy for frame-level

classification compared to any individual CNN approaches. Huang et al. [33], on the other hand, proposed a multi-stream CNN consisting of 4 sub-CNNs for each of the regions around the left eye, right eye, mouth, and nose, achieving comparable results with the state-of-the-art.

These methods have shown great accuracy at the frame level comparing with the labelled PSPI scale as ground truths. However, they only considered the spatial information of facial expressions but completely ignored the temporal aspects. For example, the model would not be able to differentiate between a pain-induced eye closure and a simple blink if only given a single frame of information. It has, therefore, been widely agreed that pain assessment should be conducted with temporal data such as facial video recordings.

2.5 Temporal approaches

Temporal methods classify pain on a sequential level, identifying if there is an occurrence of pain in a specific time window. They allow the evolution of facial features to be considered but, at the same time, dramatically increases the complexity and number of parameters in the model. There were two main methods to handle sequential video inputs, 3D CNNs and CNN with sequential models.

Temporal evolutions were implicitly exploited and captured with 3D-CNNs by considering the entire stacked video signal as an input. 3D CNNs allowed simultaneous learning of spatial and temporal features in a single model; however, they suffered from high computational costs with the extra temporal dimension and high training data requirements. Wang and Sun [34] first attempted using a 3D ConvNet introduced by Tran et al. [35] with 8 convolutional layers of 3x3x3 kernels. An outstanding accuracy of 0.94 Root Mean Squared Error (RMSE) was achieved with the UNBC Shoulder Pain dataset; however, the author expressed concerns about the generalisability of such a model with insufficient training and limited datasets. To increase the training process and convergence, Tavakolian and Hadid [36] proposed a Spatiotemporal Convolutional Network (SCN) architecture by separating the 3D convolutional layers. The 3D model was first trained with a 2D pre-trained ResNet and further fine-tuned with the pain assessment dataset to allow the learning of deep features with limited data. The model showed state-of-the-art performance in multiple datasets, outperforming the standard 3D ResNet. However, only 32 frames were selected with a temporal stride of 2, though yielding the highest Area Under the Curve (AUC), the decision was not justified or tested with different datasets and may have lead to loss of

information. The random selection of frames also affected the performance of other datasets with partial or self-occlusions, as suggested by the author.

Sequential models with temporal modules were also a popular approach to exploit the evolution of features by considering a video input as a sequence of frames. They allowed an efficient learning of the temporal dynamics while providing a more flexible representation learning process. Bargshady et al. [37] achieved the best performance on the UNBC McMaster dataset by applying Principal Component Analysis (PCA) on VGG-Face [38] deep features before passing into the bi-LSTM (Bidirectional Long-Short Term Memory) model [39].

2.6 Pain datasets

The basis of any Machine Learning problem is the availability of datasets, even more so now for deep learning models. The performance of these methods depends heavily on the quality and quantity of training data. For automatic pain assessment, the recording of both visual and biopotential signals of patients experiencing pain is necessary. This was normally done by either recruiting patients suffering from chronic pain during clinical visits or inducing artificial sources of pain in healthy patients from heat or electrical stimuli. The first method provides a more accurate estimation of the real-world setting in which automatic pain assessment tools would be utilised. However, it is harder to control the level of pain experienced, and the risk of biases is high. As the pain persisted, the pain tolerance would increase, resulting in a smaller reaction. At the same time, chronic pain patients are vulnerable to pain catastrophising (the exaggerated feeling of pain due to anxiety) as pain persists, exceeding their expectations. These datasets are also limited to a smaller scale due to specific recruiting processes and criteria, creating an unbalanced situation classification problem. Induced pain experiments, on the other hand, allow large-scale collection of data from the recruitment of healthy volunteers under highly controlled settings and precise stimuli. However, compared to patients suffering from chronic pain who were continuously experiencing pain and gaining tolerance, external stimuli could result in an over-intensified reaction due to the lack of tolerance and unexpectedness.

Table 2.1 provides an overview of the popular publicly available datasets for pain assessment with the labels and metrics used to measure pain. In this report, the BioVid Heat Pain dataset was used; details of the dataset are further discussed in Chapter 3.

Table 2.1: Summary of publicly available pain assessment datasets.

Dataset	Population	Annotation Labels
BioVid Heat Pain [40]	90 healthy adults	Calibrated heat stimuli
UNBS-McMaster [41]	25 patients with shoulder pain	Self-reported, VAS[1], OPI[2], PSPI
MIntPAIN [42]	20 healthy adults	Self-reported, and calibrated stimuli
iCOPE [43]	26 healthy neonates	Pain, cry, rest, friction, air puff
iCOPEvid[44]	49 neonates	Pain, no pain binary
NPAD-I[45]	36 healthy neonates	NIPS (Neonatal Infant Pain Scale)
EmoPAIN [46]	22 adults with chronic pain & 28 healthy adults	Self-reported, and OPI[2]

[1] VAS: Visual Analogue Scale; [2] OPI: Observer Pain Intensity

2.7 Conclusion

Automatic pain assessment has made significant progress in recent years, which has been encouraged by the development of publicly available datasets and the advancement of deep learning networks. This chapter presented and outlined some of the prior works done for automatic pain assessment and their limitations. The literature suggested that frame-level analysis is not sufficient, and temporal evolution has to be considered when performing pain assessment to fully reflect expressions of pain. However, this gave rise to a significant computational burden and lack of generalisability with smaller datasets. Moreover, none of the methods presented interpret how pain was implied from facial expressions.

Chapter 3

Dataset

3.1 Introduction

The dataset used in this study is the Biopotential and Video(BioVid) Heat Pain Database [40]. It is a multimodal dataset containing both biopotential (i.e. ECG, EMG, and EEG) and video signals for facial expressions collected in a highly controlled setting.

3.2 Experiment and design

The BioVid Heat Pain Database utilised a unique pain intensity metric based on the subjective experience of pain. Pain was stimulated through inducing heat above the participants' pain threshold. A Medoc PATHWAY (Pain & Sensory Evaluation System, https://www.medoc-web.com/pathway) was used on participants' right wrist when they were in a comfortable position with their right arm resting on a desk in front of them. This ensured a uniform contact between the heating thermode and participants' skin while inducing and accurately quantifying the heat in a highly controlled setting. A maximum temperature of 50.5°C was strictly avoided to prevent skin damage throughout the experiment. Two experiments were performed for pain per participant:

3.2.1 Pain calibration

The pain intensity metric in this dataset was classified on a discrete scale of 0 to 4, with 0 being no experience of pain and 4 being the maximum amount of pain participants could tolerate. The experiment aimed to estimate the pain intensity from induced heat on participants' wrists; therefore, a calibration or mapping needs to be established between the amount of heat and the pain metric. Levels 1 and 4 were two key metrics measured in the experiment, and they are known as the pain threshold and pain tolerance, respectively.

The pain threshold was defined as the minimum intensity at which the participant begins to sense pain, and it was collected by slowly increasing the heat until the participants just starting to experience a feeling of burn or sting, at which point the participants stopped the heating by pressing the stop button and the temperature was recorded. Pain tolerance recorded the pain

when participants cannot accept any more, and it was similarly collected by slowly increasing the pain until the stop button is pressed by the participant. Both pain metrics were highly subjective and dependent on individuals' life experiences and even the anatomy of the skin. Therefore, the heat required to achieve each level of pain was calibrated individually. Pain level 0 was maintained at 32°C for all participants. Pain levels 1 and 4 were measured from the experiment above, and the intermediate pain levels (levels 2 and 3) were derived by linearly interpolating as shown in Figure 3.1.

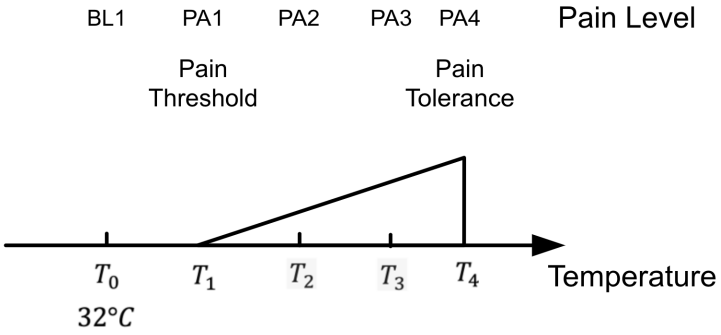


Figure 3.1: Pain level calibration from tolerance and threshold. Baseline temperature was set at $32°C$. Pain level 1 and 4 were collected as pain threshold and tolerance, respectively. The intermediate pain levels were linearly interpolated.

3.2.2 Main stimulation

The main stimulation process for each participant was around 25 minutes, with 20 separate stimulations for each pain level (levels 1-4 determined from the previous pain calibration experiment). Each stimulation includes two phases: a heating phase where the randomly selected pain level heat was held for 4 seconds and a resting phase with a randomized period of 8 to 12 seconds at a baseline heat level of 32°C). Figure 3.2 shows the heating profile for two example stimulus cycles, each with one heating phase and one resting phase. This process resulted in a total of 80 pain stimulations and 80 resting phases with no pain. In the 80 resting phases, 20 were randomly selected as the pain level 0 data to ensure data balance for each label. For part A of the dataset, the 5.5-second footage that starts 1 second after heating start (see figure 3.2 shaded region) was selected as a single data video, resulting in 100 5.5-second videos with 20 videos for each pain level, including baseline pain level 0.

The same experiment was repeated for the 87 participants with facial EMG sensors for part B of the dataset at the same individual heat levels collected from the pain calibration experiment.

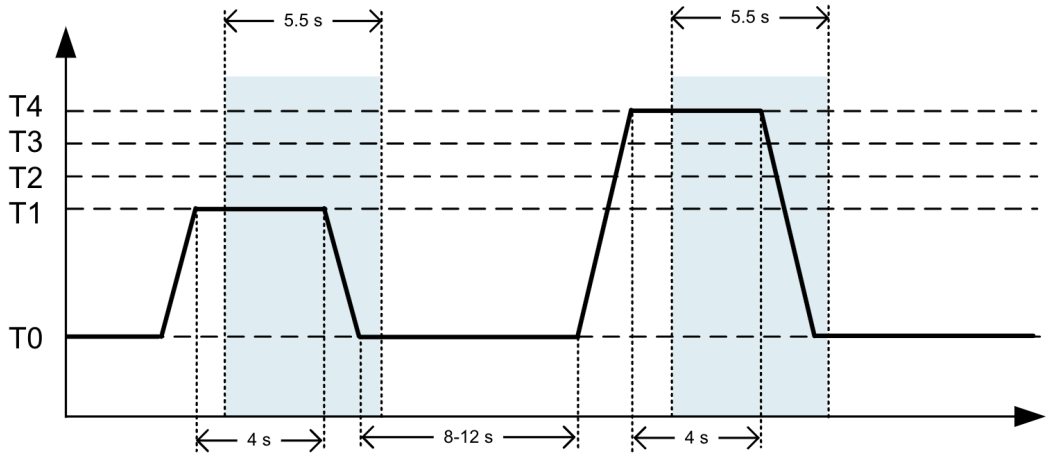


Figure 3.2: Heating profile for two example stimuli cycles. The shaded 5.5 seconds region was extracted for Part A of the Dataset.

3.3 Collection protocol

3.3.1 Biopotential signals

Pain assessment using biopotential signals is one of the aims of this dataset, with multiple biopsychological data recorded, including the Skin conductance level (SCL), ECG, EMG, and EEG, with the Nexus-32 amplifier (http://www.mindmedia.nl). However, pain assessment with biopotential signals is out of the scope of this report.

3.3.2 Video signals

The facial video signal was captured with three AVT Pike F-145C cameras at a synchronized frame rate of $25Hz$. The main camera was directly in front of the participants, while the other two were positioned on each side of the main camera, pointing to the face of the participant with a 45° angle. The triple camera setup allowed a face capture even with the participants moving and turning their heads. The cameras recorded at a resolution of 1388×1038 coloured pixels. In addition to the three Pike cameras, a Kinect Sensor was placed alongside the main camera in front of the participant to record the depth map with a resolution of 640×480 pixels at $30Hz$, and its frontal image of 1280×1024 coloured pixels at $10Hz$. The researcher synchronised the two video streams manually by adding and locating a clapperboard action before and after the experiments. Table 3.1 describes the specification for the AVT Pike F-145 C camera used in the experiment.

Table 3.1: General features and specifications for the AVT Pike F-145 C camera. (Source: Allied Vision Technologies)

Item	Description
Imaging Sensor	Type 2/3 (diag. 11.2 mm) progressive scan, SONY IT CCD
Image size (pixels)	1388 (H) x 1038 (V)
Pixel Size	6.45 µm x 6.45 µm
Frame rate	25 Hz
Video Data Output	8, 12, 14 and 16-bit digital data
Gain & Exposure	Manual 0 - 32dB, auto gain
Lens Mount	C-mount
Interface	IEEE1394b
Physical dimensions	96.8 mm x 44 mm x 44 mm (L x W x H)

3.4 Dataset overview

The dataset consists of five parts: parts A, B, and C focus on pain stimulation, part D contains posed pain and emotions, and part E includes emotion elicitation. The complete demographics for each part are described in Table 3.2. The database was collected from 90 participants and distributed equally between the three age groups. It is worth noting that all participants are white in ethnicity. This is further discussed in Section 3.5.

This report focused primarily on the facial expression-based unimodal approach to pain assessment, and therefore, Part A was chosen, with a total of 8,700 videos (100 videos for each of the 87 healthy participants), for the absence of a facial sensor connected to the corrugator and zygomaticus muscles, allowing for a clearer and more accurate extraction of face and facial landmarks.

Table 3.2: Summary of dataset participant demographics

Parts	Number of subjects	Number of recordings	Gender		Age group			Ethnicity[1]	
			Male	Female	18-35	36-50	51-65	W	O
A	87	100	44	43	30	29	28	87	0
B	86	100	42	44	30	29	27	86	0
C	87	1	44	43	30	29	28	87	0
D	90	7	40	40	30	30	30	90	0
E	86	5	42	44	29	29	28	86	0

[1] W = White, O = Other.

3.5 Limitations

Despite its good design, this database has several limitations. Firstly, as the demographic suggested, the participants recruited were purely ethnically white and from Germany due to local constraints. However, the reaction to emotions, especially facial expressions, varies greatly with cultural and ethical diversity. They can greatly influence the expressiveness of pain and different emotions. These correlations have not been thoroughly established yet but still raise questions about the dataset's generalisability to be used on a different population.

Moreover, heat was selected as the method to induce pain as it is highly controllable and easy to quantify. But at the same time, heat is highly accessible in everyday life and resistance can easily be developed through repeated exposure. For example, participants with more experience cooking would have higher heat tolerance and threshold with less reaction to the heat at these thresholds. This could potentially result in biased data from the individual resistance and amount of exposure in their life experience. The overall experiment time of 25 minutes also raised concern about participants building up resistance during the experiment and reacting inconsistently to the same heat level. The resting phase of 8 to 12 seconds may not be enough to completely remove the impact of the previous stimuli, especially if it was at pain tolerance (highest level). The body temperature changes may also lead to inconsistent reactions to the same heat level. This would lead to incorrect and misleading data for the training and evaluation process which affected the overall performance of the model.

The author and researcher of the dataset have also suggested some limitations to the quality of the dataset, notably the less expressive participants during the experiment. During both the main stimulation and emotion collection experiments, several participants continued to show little to no facial expression or reaction. Sellner et al. [47] performed a graphical analysis of the facial AU intensities between pain and neutral emotions, and the results suggested high variations in the responses between different individuals in the dataset. There were subjects with clear discrimination quality, while some responses had no differences, or exhibited misleading differences in intensity. This significantly hurt the performance of any classifier and challenge the classifier's ability to work with noisy data. This remained a major concern for the quality of this dataset. To accommodate this, the author suggested filtering out the 20 less expressive individuals and focusing all future work on the more expressive group of 67 participants in part A of the dataset [48].

Chapter 4

Pain assessment using handcrafted features

4.1 Introduction

This chapter focuses on assessing pain by utilising handcrafted features inspired by the FACS AU system to further investigate the mechanisms behind pain expressions and discover the dominant facial features related to pain.

4.2 Model overview

The proposed model pipeline, shown in Figure 4.1 can be separated into three main steps: Frame-level feature extraction, Sequence-level descriptor construction, and pain classification. The input video data was first analysed frame-by-frame through a face detector, a facial landmark detector, and a feature extractor. A total of 13 facial features were selected, producing 13×1 signal data point in each frame. To track the change over the video data, the signal data points were then combined to construct the feature signals (13×138) for the entire 5.5-second video data (138 frames in total) and passed into a feature extractor. 48 descriptors were extracted per signal, resulting in a final descriptor vector of 624×1 (48×13 = 624). This was passed into the classifier for both binary (pain vs. no pain) and multi-class (pain level 0 to 4) pain assessment.

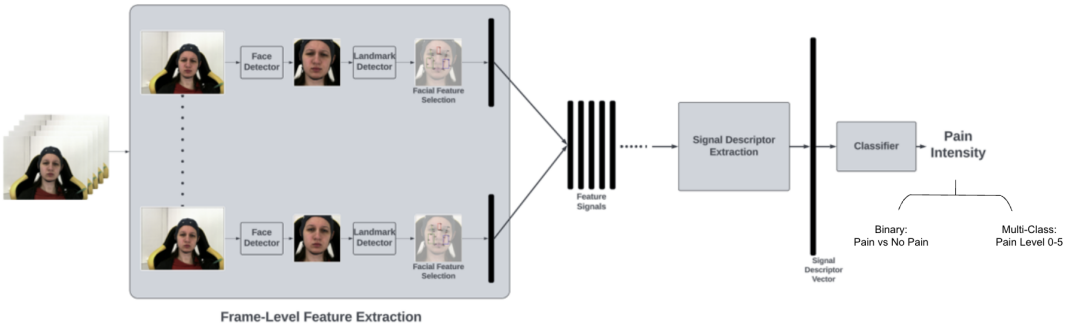


Figure 4.1: Proposed model pipeline.

4.3 Frame-level feature selection

4.3.1 Face detection

The face of the participants was extracted with the BlazeFace full-range Face Detector in the mediapipe library [49]. The model is a variant of the BlazeFace model described in chapter 2, which used a technique similar to CenterNet [50] by representing objects such as faces with a centre key point and a pair of corner key points. The face was extracted from the input frame with size (1038×1388), aligned, and cropped to (256×256) with 20% margin on all sides before feeding into the landmark detector.

The model was chosen due to its state-of-the-art latency, which allows for real-time analysis, and its simplicity of use, as it is integrated into the mediapipe Python library. The model also outperforms most face detectors commonly used for pain assessment, such as the dlib CNN or MTCNN [51] models.

4.3.2 Facial landmark detection

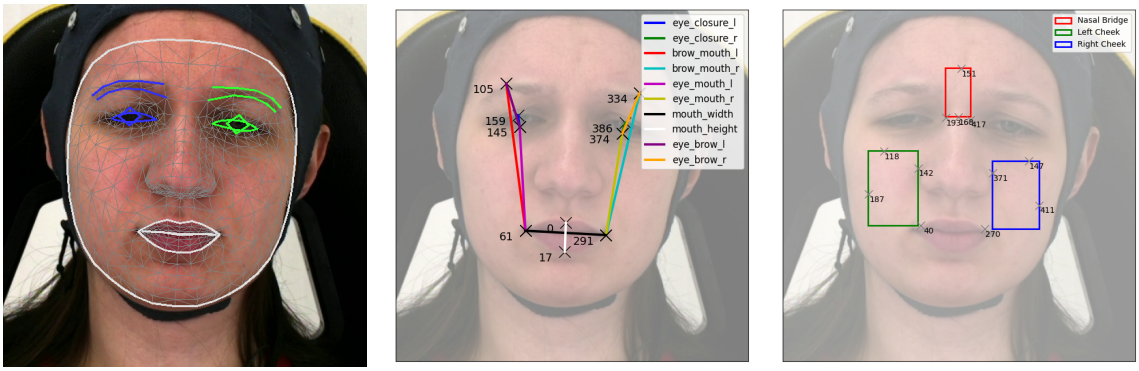
I extracted the facial landmarks by using the mediapipe FaceMesh-V2 [52]. The face mesh model is a CNN with a similar architecture to MobileNetV2 that takes 256×256 cropped face images and can detect 478 3-dimensional facial landmarks with depth information. Figure 4.2(a) shows an example output face mesh with 478 facial landmarks. [40]. The comprehensive 478 landmark labelling allowed for a significantly more accurate feature selection process, and the number of facial landmarks from FaceMesh-V2 is unrivalled by any other landmark detector.

4.3.3 Feature selection

Out of the 478 landmarks returned, I selected 10 landmarks and constructed 10 feature vectors by connecting the landmarks that can best represent the changes and actions induced by pain, according to previous studies of pain, including Werner et al. [40], and Prkachin [53]. Actions such as raising cheeks (AU 6), lowering eyebrows (AU 4), mouth opening (AU 25), closing of eyes, and stretching of the mouth can be easily captured by taking the vector distances of key landmarks. I selected the top of the eyebrows (highest point), corner of the lips, and corner of the eye on both sides as key points and measured the distance between the eyebrows to eyes, the eyebrows to mouth, and the eyes to mouth. 3D Euclidean distance described in Equation

4.1 was used as 3-dimensional information is available from the landmark detector and helped reduce the error when there is a change in head orientation. The width and height of the mouth were also recorded, as well as the distance between the upper and lower eyelids to capture the closing of the eyes. Figure 4.2(b) lists and presents the feature vectors constructed from landmarks.

$$d = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2} \tag{4.1}$$



(a) (b) (c)

Figure 4.2: Frame-level features construction. (a) 478 landmark face mesh from the Mediapipe landmark detector. (b) 3D feature vectors constructed from a subset of landmarks. (c) Texture features and their anchor points.

Some other facial expressions, such as nasal wrinkling or squeeze of cheeks, cannot be easily represented by vector distances. Facial textures were analysed instead. I selected 3 key regions, including the nasal bridge and root, lower left cheek, and lower right cheek. I expressed the changes in texture by analysing the gradient of the key regions. Figure 4.2(c) shows the key regions and the anchor landmarks to construct the key regions. For each region, the mean gradient magnitude was measured by taking the mean of all pixel gradient magnitudes. For each pixel, the gradient in the x- and y-direction was calculated by applying the Sobel filter in the x- and y-direction, respectively, and the magnitude was then obtained by taking the vector magnitude of the gradients with the following equation:

$$G = \sqrt{G_x^2 + G_y^2} \tag{4.2}$$

where G_x and G_y are the gradients in the x- and y-direction, respectively.

4.4 Sequence-level signal descriptor

4.4.1 Signal construction

From frame-level feature extraction, I utilised a 10D feature distance vector and a 3D texture descriptor per frame. I combined the 13D for each frame in the 5.5-second video input (138 frames) to yield the overall feature signal with size (13×138).

To capture more temporal information and correlations within the signal, The smoothed signal was computed with a first-order Butterworth filter at a cutoff frequency of 1Hz. With the smoothed signal, I took the first and second derivatives for each of the 13 features, marked as *v (velocity)* and *a (acceleration)*. Figure 4.3 presents the processing of one example signal captured from the same subject in two different experiments. The signals on the left were collected from the subject when experiencing pain level 1 (lowest level of pain), while the signals on the right were from the subject experiencing pain level 4. The signal with lower pain level showed a lower variability indicating less facial response to pain, while signals for high pain level experiment presented a large increase in signal intensity at 2.5 seconds. This change was correctly captured with the first and second derivative signal as the peak and zero-crossing, respectively.

4.4.2 Handling missing data

Due to occlusions and participants' head movements, several video data or frames resulted in failed face or landmark extraction. As no landmark coordinates were extracted, it resulted in missing data in the final signals. I employed a mean replacement method to accommodate these missing data points by replacing the missing data with the mean of the 30 data around them (15 before and 15 after, if possible). The missing first or last data of the signal was handled differently. Instead, it was replaced by the value of the adjacent data point, as using a mean window for this missing information may be misleading. This best preserved the property of the signal while reducing the risk of corrupting the first and second derivatives of the smoothed signal. If there were more than 10 missing data consecutively or more than 30 missing data overall, the whole data would be excluded from the experiment. This resulted in 213 video data excluded from the overall 8,700 datasets, or around 2.45%

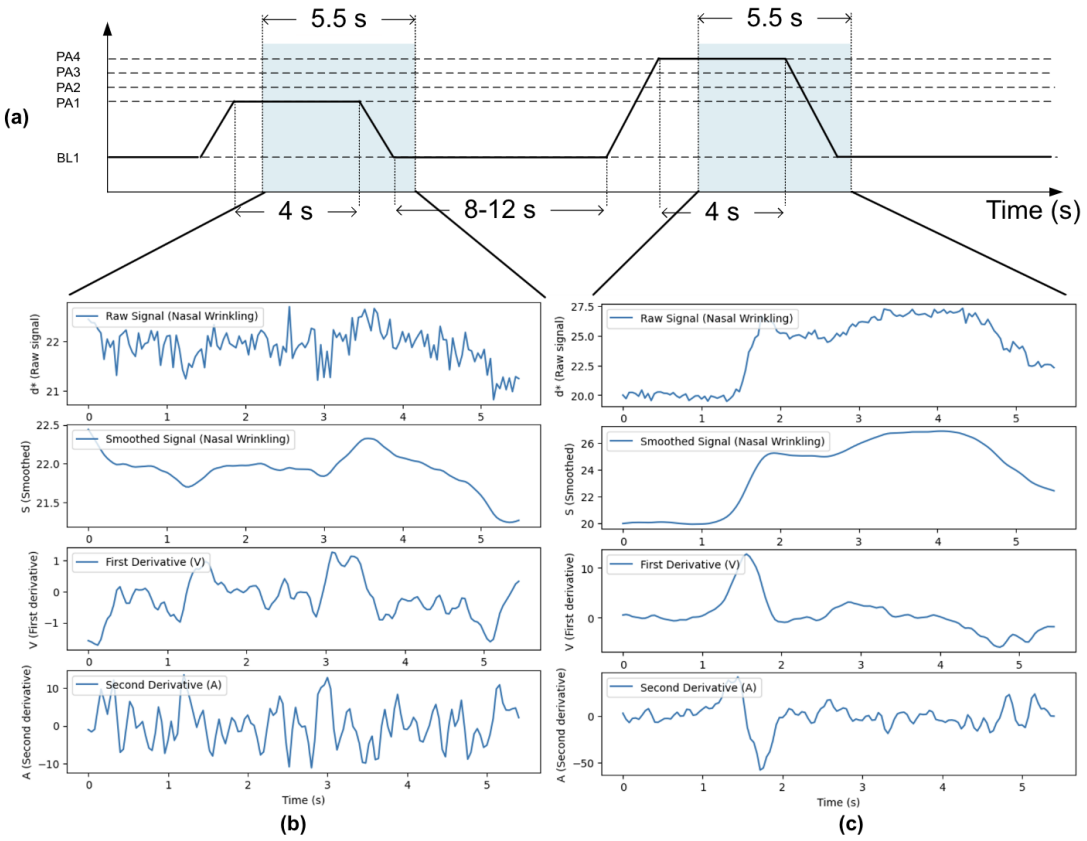


Figure 4.3: Example of the mean gradient magnitude signal of the nasal bridge region collected and processed. Top figure (a) shows the heat stimulation intensity for Pain level 1 (PA1) and Pain level 4 (PA4) and a resting period of 8-12 seconds. The bottom left (b) signal was collected from the PA1 experiment with the raw signal on top, followed by the smoothed signal with its first and second derivatives. The bottom right (c) signal was collected from the PA4 experiment with a more noticeable facial reaction. As expected, the facial expression of lower pain level showed less variability, while the high pain level experiment showed a huge change in facial movement after 2.5 seconds.

4.4.3 Signal descriptor

Lists of temporal signals cannot be passed into a linear classifier; signal descriptors were used to reduce the extracted feature signals' dimensions. For each signal, 16 descriptors were measured to capture different aspects of the signal, including the variability, duration, and intensity, inspired by Werner et al. [40] Table 4.1 listed all the 16 signal descriptors and the properties each measure. The signal descriptor components were inspired by the FACS system and the prior work by Werner et al. [40]; the statistical descriptors, mean, median, min, and max were designed to account for the intensity of the facial movements, whereas the variability can be captured by both the range and area based descriptors such as the ratio of the area under the curve to the maximum possible rectangular area. Duration-based descriptors were also employed to express the temporal evolution of the facial features, such as tmax, which measures the time when the

intensity is at its apex. Positive Mean Crossing count (PMC) and Positive Average Crossing count (PAC) also reflect key insights on the signal behaviour, including the signal's volatility. High PMC may imply that the experiment data is of low quality and involves excessive facial movement. Overall, with 16 descriptors for each processed signal, I have 48 descriptors per feature signal (3 processed signals per feature signal). Hence, with 13 features, my final feature vector has 624 dimensions. Before passing into the classifier, I standardised the feature vector as:

$$z = \frac{x_i - \mu}{\sigma} \tag{4.3}$$

where μ is the descriptor mean, and σ is the descriptor standard deviation by individual participants to make it more comparable and remove noisy data where no actions were present.

Table 4.1: Signal Descriptor Components

Variable	Description	Domain
mean	mean value of signal	value
median	median value of signal	
min	minimum value of signal	
max	maximum value of signal	
range	range of signal	value variability
STD	standard deviation of signal	
IQR	inter-quartile range of signal	
IDR	inter-decile range of signal	
MAD	median absolute deviation of signal	
tmax	time when the signal is at its maximum	time
TGM	duration of the signal is greater than mean	duration
TGA	duration of the signal is greater than the average of mean and min	
PMC	number of times where the signal crosses over the mean	count
PAC	number of times where the signal crosses over the average of mean and min	
area	area under the curve between signal and its minimum	value $\times$ duration
areaR	quotient of the area and the rectangular area between max and min	

4.5 Pain classifier

In this experiment, three classifiers were tested, namely, eXtreme Gradient Boosting (XGBoost) [54], Random Forest Classifier (RF) [55], and Support Vector Machine (SVM). RF and SVM have shown great results in past experiments for pain assessment. This section briefly introduces the classifier and the method to assess the importance of the feature selected.

4.5.1 Random Forest Classifier

RF is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes. It uses a ”bagging” ensemble training method that splits the dataset into multiple subsets by random sampling with replacement, uses each subset to train a weak learner in parallel and predicts with majority voting. An additional layer of randomness is employed by only selecting a random subset of features at each node of the trees. The independent training of multiple models reduces the variance or noise of the final classifier even when each individual model has a high variance or bias.

4.5.2 SVM

SVM is the most widely used classifier for all purposes. SVM classifies binary classes by finding the optimal hyperplane in the feature space that maximises the distance between classes given a margin. It employs hinge loss, which penalises wrong predictions and even correct predictions but falls within the predefined margin of the hyperplane. By performing SVM on a high dimensional space, it can perform non-linear classification tasks using kernel functions. This experiment tested SVM with both Linear and Radial Basis Function (RBF) kernels. For multi-class classification, I used SVM with a one-vs-one strategy where a classifier is trained for each pair of classes.

4.5.3 XGBoost

XGBoost is an advanced ensemble model utilising the ”boosting” mechanism that subsequently adds predictors to the ensemble, with each new predictor correcting its predecessor. It improves upon the gradient boosting algorithms by using both the gradient and the Hessian (second-order derivative of the loss), achieving a faster convergence for each predictor.

XGBoost also has its built-in feature importance measurement, more specifically, it provides three metrics on how each feature improves the model performance: weight, gain, and cover. The weight importance of a feature is similar to the feature's frequency in the ensemble model. It counts for the number of times the feature is used to split the data in all trees, with the features used more often considered more important. Gain, on the other hand, evaluates the performance or accuracy gained by using a feature in the model. It sums up the gain in accuracy of each feature when it is used in a decision tree. Finally, the cover measures the sum of the Hessian of the loss function for all data points through the branch created by splitting the feature. It provides insights into the amount of data affected by the feature.

4.5.4 SHapley Additive exPlanations (SHAP)

SHAP [56] is a widely used approach for machine learning interpretability. The method utilised cooperative game theory to explain predictions of any machine learning model by decomposing the output of the model into sums of the effects of each feature with conditional probability. It provides crucial insights into the model behaviour and comprehensive breakdowns of how each feature impacts the model prediction. SHAP is also model-agnostic, meaning it can be used to interpret any model after training (post-hoc) without any modification of the original model. In this experiment, I used SHAP to understand each feature's effects on top of the built-in variable importance metrics.

4.6 Evaluation

Prediction outputs can be viewed with a confusion matrix, which splits the output into four categories: TP (True Positives), TN (True Negatives), FP (False Positives), and FN (False Negatives). The matrix contrasts the actual target label with the model's prediction, providing a detailed breakdown of the model's performance and its correct and incorrect predictions. Accuracy was the metric primarily used because it is simple and highly intuitive, defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

For each of the three models, a combination of grid search technique is used to find the optimal hyperparameters. From prior works by Werner et al. [57], the author discovered the number of decision trees in the model was the most important feature when tuning the model

for pain assessment. Therefore, I employed a coarse grid search with cross-validation with a higher number of trees with other hyperparameters before a fine-grained grid search around that result for both the RF and XGBoost model. SVM was fine-tuned similarly with a coarse and fine-grained grid search for parameter C and γ. Finally, the model performance was evaluated on a predetermined validation set suggested by Stefanos Gkikas for best generalisability [58].

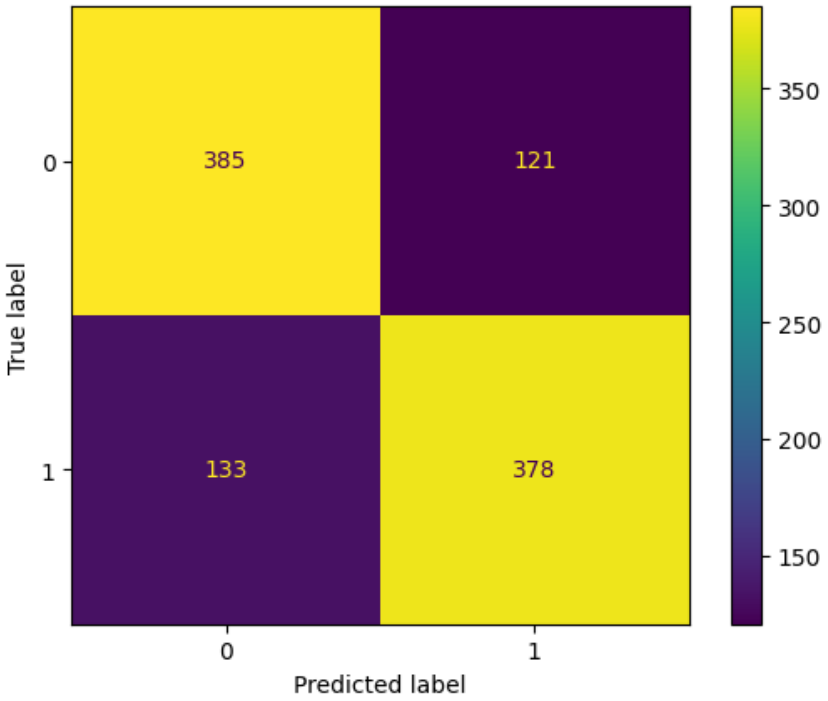
4.7 Results

Table 4.2 lists the classifier accuracy with the best hyperparameters in comparison with previous work done with the BioVid Heat Pain dataset, with the best performing results in bold. Binary classification was tested by classifying different levels of pain with no pain, while multi-class classification quantify pain into 5 levels (0-4). XGBoost has shown the best overall performance for both binary and multiclass pain assessment with an accuracy of, respectively. The discriminative ability of the classifier was measured with the Receiver Operating Characteristic (ROC) curve and the AUC value, and the binary (pain vs no pain) classifier achieved a high AUC ROC of 0.841. The confusion matrix of the best model for both binary and multi-class classification was shown in figure 4.4. The confusion matrix has shown a good balance between False Negatives and False Positives, with similar precision and recall metrics.

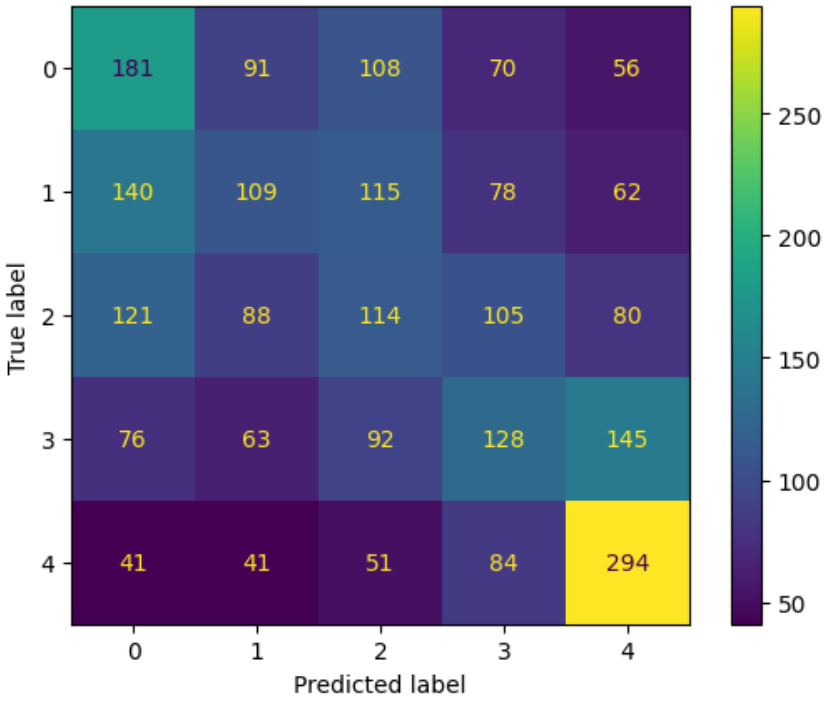
To evaluate the feature importance, both the built-in XGBoost importance metrics and the SHAP analysis were plotted in Figure 4.5 and Figure 4.6.

Table 4.2: Predictive performance of the three classifiers for different classification tasks.

Classifier	Binary Classification				Multi-Class Classification
	[0-1]	[0-2]	[0-3]	[0-4]	[all-all]
Facial Activity Descriptor (Werner et al. [40])	**0.533**	0.560	0.640	0.724	0.308
RF (1000 trees)	0.501	**0.561**	**0.666**	0.742	0.288
SVM	0.527	0.519	0.660	0.684	0.3034
XGBoost (200 trees)	0.521	0.542	0.645	**0.750**	**0.326**
XGBoost (1000 trees)	0.529	0.542	0.643	0.749	0.313

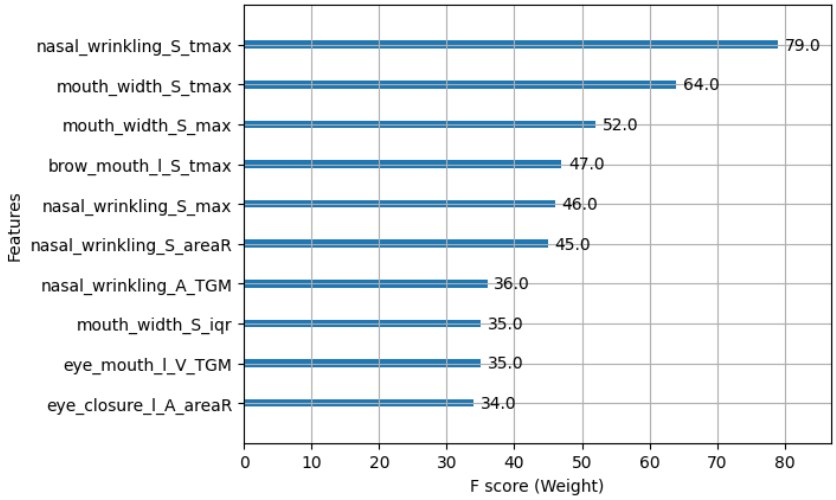


(a)

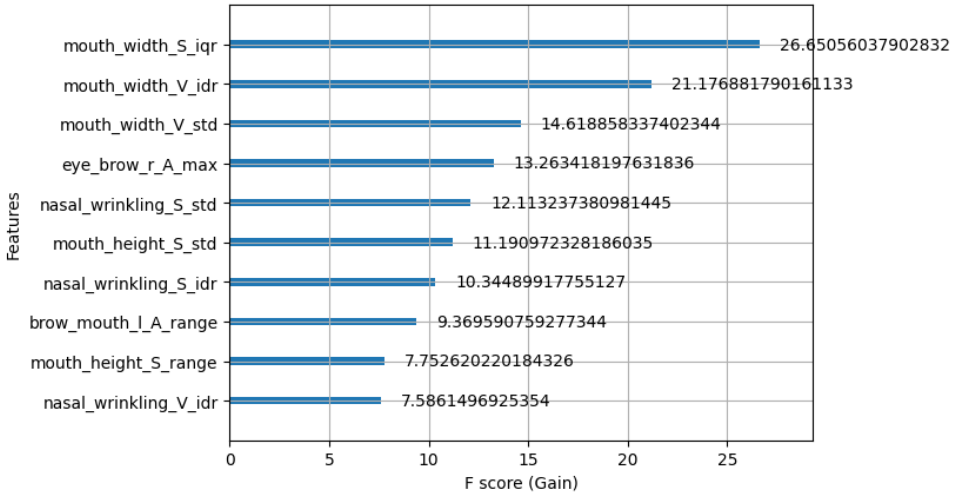


(b)

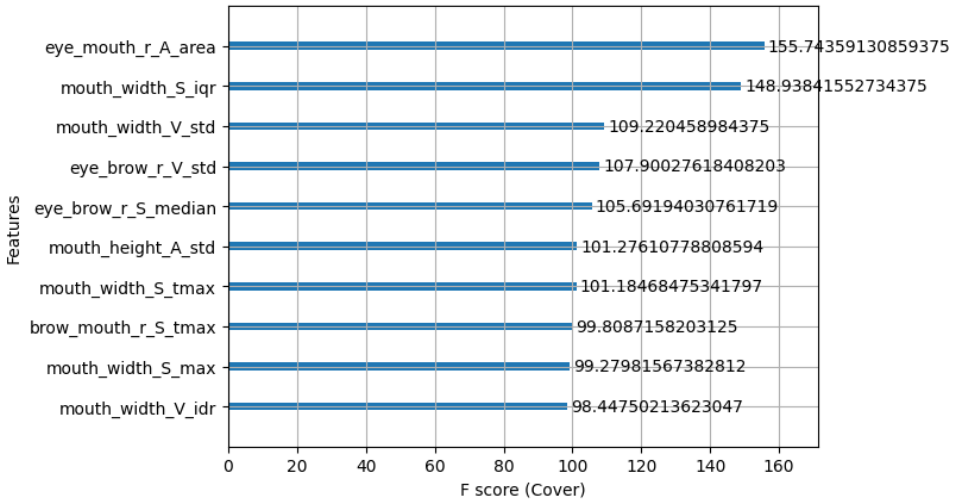
Figure 4.4: The confusion matrix of the XGBoost classifier. (a) Binary classification between pain intensity 0 and 4. (b) 5-class pain classification from 0 to 4.



(a)



(b)



(c)

Figure 4.5: XGBoost variable importance metrics. (a) 10 features with the highest weight importance. (b) 10 features with highest gain importance. (c) 10 features with highest cover importance.

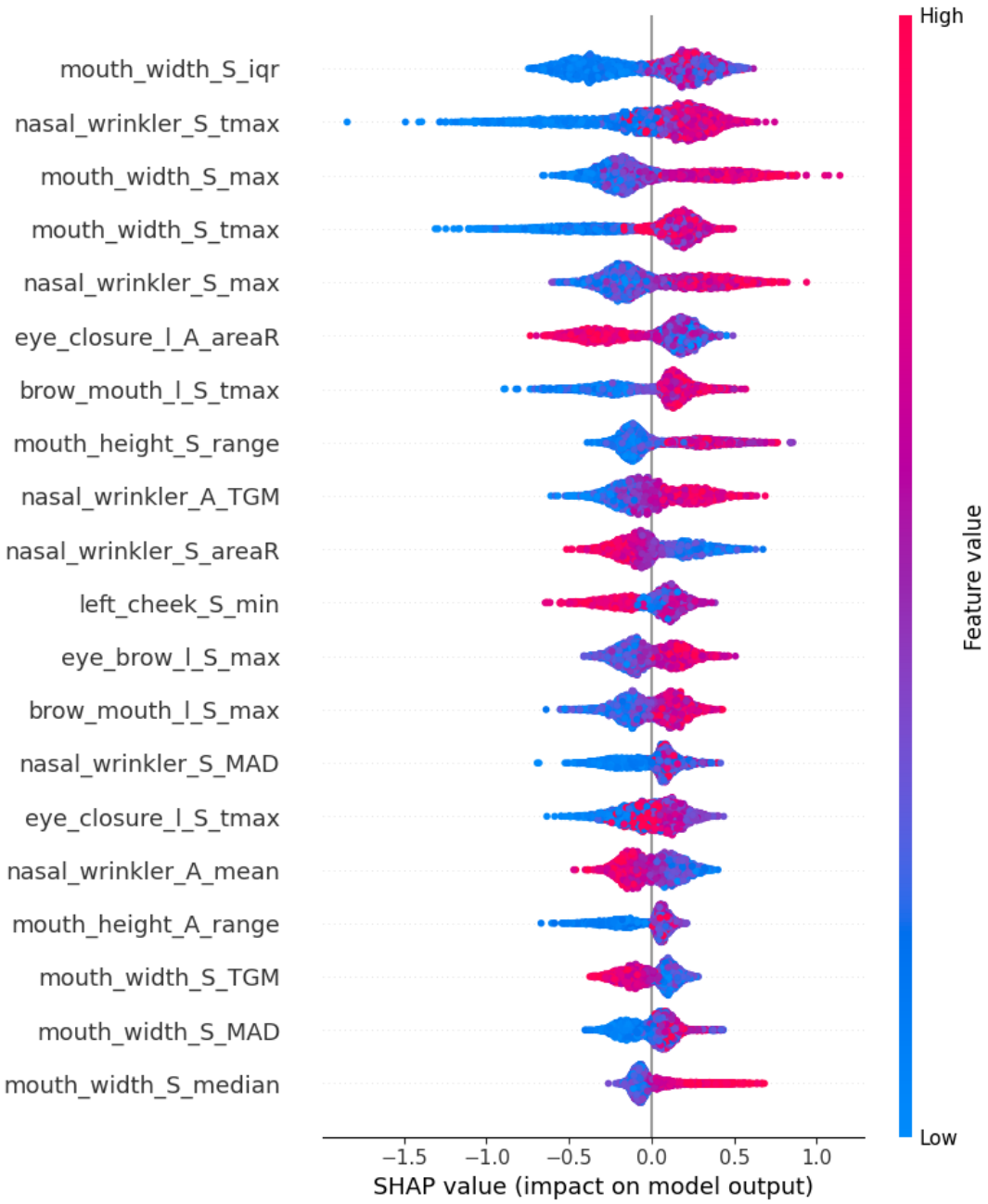


Figure 4.6: SHAP variable importance

4.8 Discussion

XGBoost, with the proposed handcrafted feature extraction pipeline, has shown great predictive performance in binary classification, outperforming the baseline extraction method proposed by the author of the dataset. However, the model struggled with multi-class tasks, especially when classifying between intermediate classes with pain intensity, as shown in the confusion matrix. This was expected as when the pain was low, some participants showed barely any reaction,

and the classifier had to work with these noisy data, which could be misleading. Similarly, good predictive performance was demonstrated for high-intensity tasks for RF and SVM classifiers, while predictions for low-intensity intensity were hardly above flipping a coin. Overall, XGBoost has also shown superior performance to other classifiers and outperformed previous handcrafted feature extraction methods on this dataset. Compared with other methods, head pose information was not considered in this experiment. I observed a large amount of head pose movements in the dataset where participants were not experiencing pain or at low pain intensity. This could add unwanted information to the model, limiting the performance. Furthermore, head pose movements were already considered within the feature distance extraction process by utilising 3D landmark detectors. As suggested by Werner et al. [40], head pose signals were useful indicators alone for pain assessment, but when used together with facial features, the predictive performance only improved slightly compared to using facial features only (0.3%).

The importance metrics have given us an understanding of the features that contributed the most to the classification. Both the height and width change of the mouth were the two of the most important features for all three measures of the XGBoost importance metrics. The high weight and gain importance of the mouth features indicated that they were the most frequently used features in the ensemble model and provided the biggest improvement in the predictive performance. The SHAP analysis agreed with the evaluation, with the mouth width feature occupying the 3 places in the 10 most important signals. These features have also shown great discriminative ability, with higher feature values having a higher positive impact on the model prediction. This is reasonable as an increase in mouth width or a decrease in height were identified as pain-related actions in the FACS system with lips tightening (AU 23), lips stretching (AU 20), and the opening of the mouth (AU 25,26,27). However, none of these AU were considered in the PSPI pain intensity scale. It is, therefore, worth considering mouth movement as part of the pain-induced facial expressions and putting more attention on this area. The distance between eye and mouth was another important feature in all three importance metrics. This is expected as it captures both AU 6 (raising cheeks) and AU 10 (raising upper lips), both were used in the PSPI intensity scale. Eye closure, on the other hand, did not show to have a high importance. This agreed with my observation that participants in the dataset tend to close their eyes even without the presence of pain. The SHAP analysis has also supported the claim as some signals, such as $tmax$ (time of the amplitude) of eye closure, are hardly distinguishable.

However, the area ratio of the eye closure signal has shown great discriminative ability. The area ratio captures both the variability and the duration of eye closure and allows the model to distinguish between a pain-induced eye closure and a blink. This suggests that the duration of eye closure has to be considered in relation to pain and that temporal information should not be ignored when assessing pain.

For texture features, the mean gradient magnitude of the nasal bridge region has shown dominant importance. This was expected as the feature was designed to capture the increase in skin folds in the nasal region, matching AU 4 (lowering the eyebrows) and AU 9 (wrinkling the nose).

The signal descriptor importance should also be analysed as it increases our understanding of which properties of the feature signal contribute more. In general, signal variability is the most dominant signal descriptor property, with inter-quartile range, range, maximum and standard deviation having a significant contribution. This agreed with my assumption that facial movements would be more prominent when pain is experienced. And for signal importance, the smoothed signal played a larger role. This might suggest that the derivatives of the signal were too complex or redundant in this task.

Despite the accuracy the model has shown in this dataset for binary classification, there were some limitations to the proposed model. Firstly, standardising the signal vectors was a pivotal step to allow the model to predict and learn from participants with vastly different facial anatomy. However, in a real-world setting under primary care, for example, where new patients need to be assessed quickly without prior individual data, the signal vector extracted would have nothing to compare to or standardise with, hence hurting the predictive performance significantly. Moreover, as explained above, head pose movements were not considered in this model. This may not hurt the validation accuracy much, but it reduces the ability of the model to handle high head movements, especially when it starts to cause facial occlusion or failure for face and landmark detection. There is still a need to increase the model's capability to handle partial occlusions and even with facial sensors to be used in medical settings.

It is also worth noting that the model still struggled to classify low pain intensities. Having more features to capture small micro-expressions can help us better understand the facial responses to lower pain intensities, helping to solve the challenge. Another direction is to use biopotential information together with facial expressions to compensate for the lack of facial

responses.

4.9 Conclusion

In this chapter, I introduced a handcrafted facial feature extraction pipeline that tracks the facial movements over the 5.5-second time window utilising a face detector and a landmark detector. The model has shown promising predictive performance, outperforming the baseline accuracy for sequence-level prediction. The results also supported the need for temporal exploitation to best capture the pain response.

I also investigated the importance of the feature selected by interpreting the classifier. The results suggested that the PSPI intensity scale cannot fully capture the pain response and that more emphasis needs to be placed on mouth movements (AU 25,26,27) and temporal properties such as duration.

Chapter 5

Pain assessment using CNN

5.1 Introduction

The method in chapter 4 has shown the viability of using handcrafted features for pain assessment, achieving predictive performance above the baseline accuracy. However, these methods were limited to the facial features selected, losing valuable information about facial micro-expressions. CNN-based methods, on the other hand, can capture and learn the most relevant features, providing more information about the mechanism and theoretically increasing the performance.

In this chapter, I tested two 3D CNN-based methods for sequence-level pain assessment with temporal exploitation and investigated the performance of the representation learning with explainable AI methods.

5.2 Model overview

The proposed model pipeline has two major processes: data preprocessing and CNN model training. Preprocessing simplified the input data to capture the most critical frames (frames that capture the evolution of pain facial expression) and reduced the memory requirement for the training process. 3D CNN was used to learn both the spatial and temporal information for both binary (pain vs no pain) and multi-class (pain levels 0-4) pain classification.

5.3 Preprocessing

Due to the complexity and size of the input video data, I employed a preprocessing pipeline to simplify the classification task and accommodate the hardware limitations. The preprocessing pipeline includes two stages: face detection and time window selection, which selects the one-second window where pain expression occurs.

5.3.1 Face detection

Similar to the face detection pipeline used in chapter 2, the face in each frame of the input video data was extracted and resized to 256×256 pixels using BlazeFace [26]. This process removed

a large amount of redundant information, such as the surroundings or the upper limbs of the body, which also reduced the memory requirement for training deep networks. For each video data, the input size was reduced to $138 \times 256 \times 256$, 138 frames from the 5.5 seconds input data with $25Hz$ recording.

5.3.2 Time window selection

Spatial information was cropped in the face detector. However, a large number of redundancies still remained on the temporal scale. With the video captured at 25 frames per second, there was no need to retain every single frame, as the changes between consecutive frames were minimal. I implemented a time window selection process to select a 25-frame, corresponding to 1-second, window where the most significant changes in facial expression occurred.

Taking the first frame, the keyframe where the participant was in an idle position without experiencing pain, I compared the pixel-wise intensity difference between each frame and the keyframe. The amount of facial changes was then measured by counting the pixels with intensity deviation over the threshold of 10% of maximum intensity. Figure 5.1 presents the pixel intensity deviation with the amount of facial pixel deviations with respect to time in blue. We can clearly see from the figure the moment that the pain expression started at the leading edge of the rising signal, as well as the pain expressions in the pixel difference plot.

To extract the target time window, I first implemented a moving average filter with a size of 20 frames to remove the noise before taking the signal derivative. The time window was ultimately determined by taking 25 frames around the peak of the signal derivative, labelled in red in Figure 5.1.

The proposed time window selection method aims to retain temporal evolution information of facial expressions while filtering out small facial movements. For example, extended eye closures can be correctly captured as they result in a low-frequency offset in the pixel deviation signal. On the other hand, eye blinks would result in high-frequency perturbations in the signal. The moving average filter can remove these effects so that these small facial movements are not considered in the selection process.

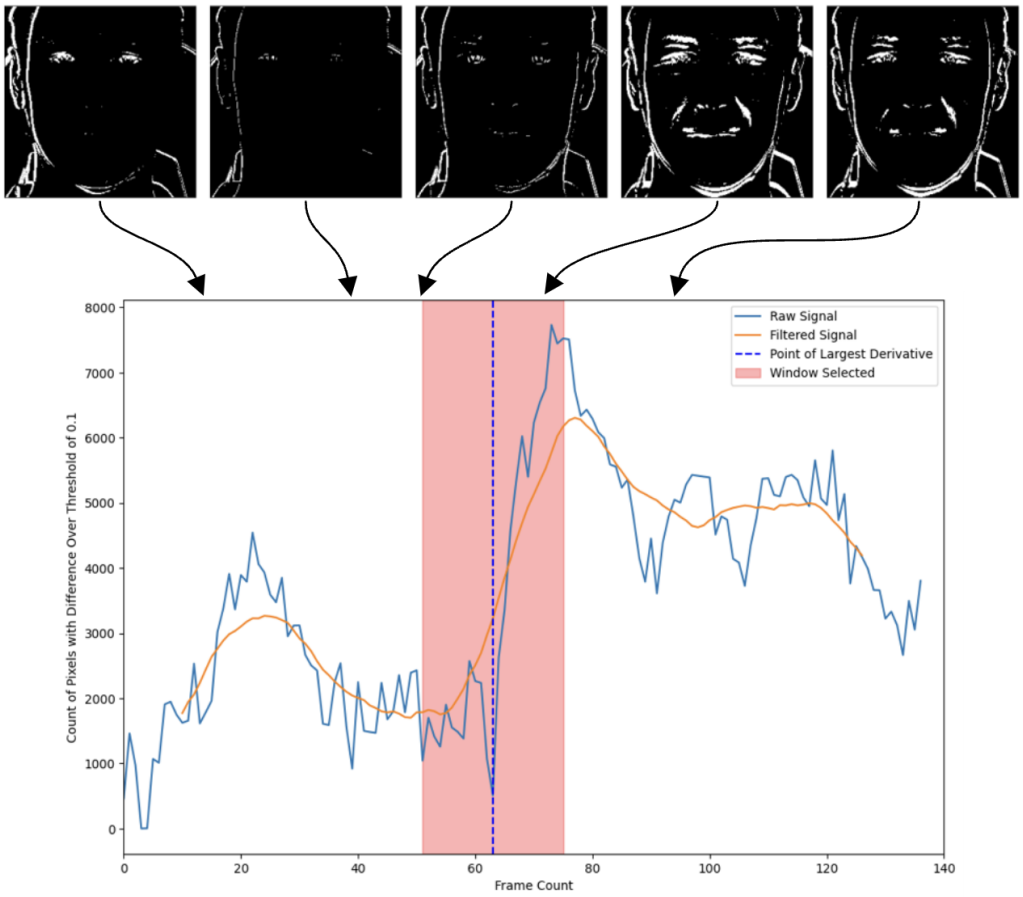


Figure 5.1: Count of pixel changes with respect to frame count and the smoothed signal. The pixel deviations at five different instances are shown. The point of the maximum first derivative is marked in a blue dashed line at the 63rd frame. The 25-frame window extracted was labelled in red.

5.4 Model architecture

Due to the additional dimension of time in the input data, 3D CNN layers were used. Two architectures were tested: Res3D, the same architecture as traditional ResNet [31] with $(3 \times 3 \times 3)$ convolutional layers instead; and R(2+1)D, an architecture proposed by Tran et al. [59] Both architectures build upon the residual network structure with Residual Blocks to allow CNNs to have more depths. Figure 5.2 presents the residual block diagram for both Res3D and R(2+1)D architecture.

5.4.1 Res3D

Res3D used the same philosophy as traditional ResNet of exploiting skip connections in Residual Blocks to allow for a deeper convolutional network. There are several variations of Res3D, with

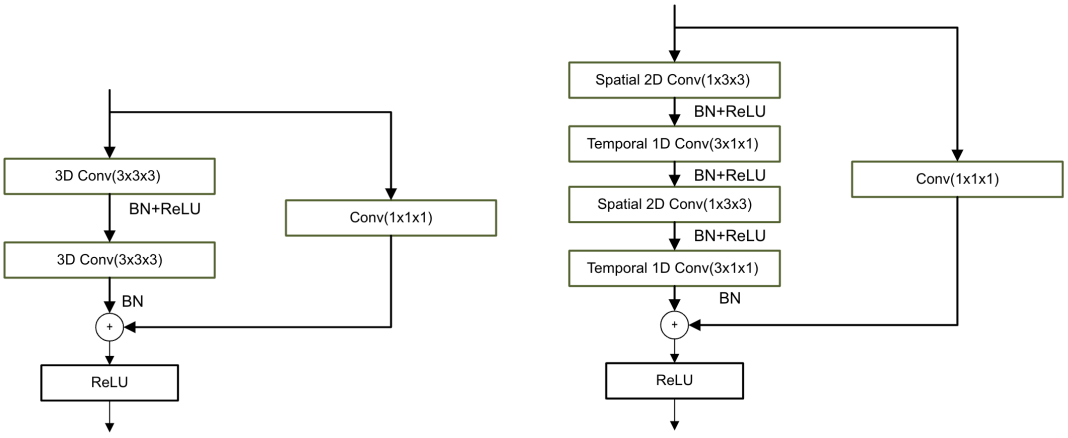


Figure 5.2: Res3D and R(2+1)D Residual block. (a) Res3D Residual Block. (b) R(2+1)D Residual Block.

depths of 10, 18, and 34 (see Figure 5.3). All three were tested in this experiment.

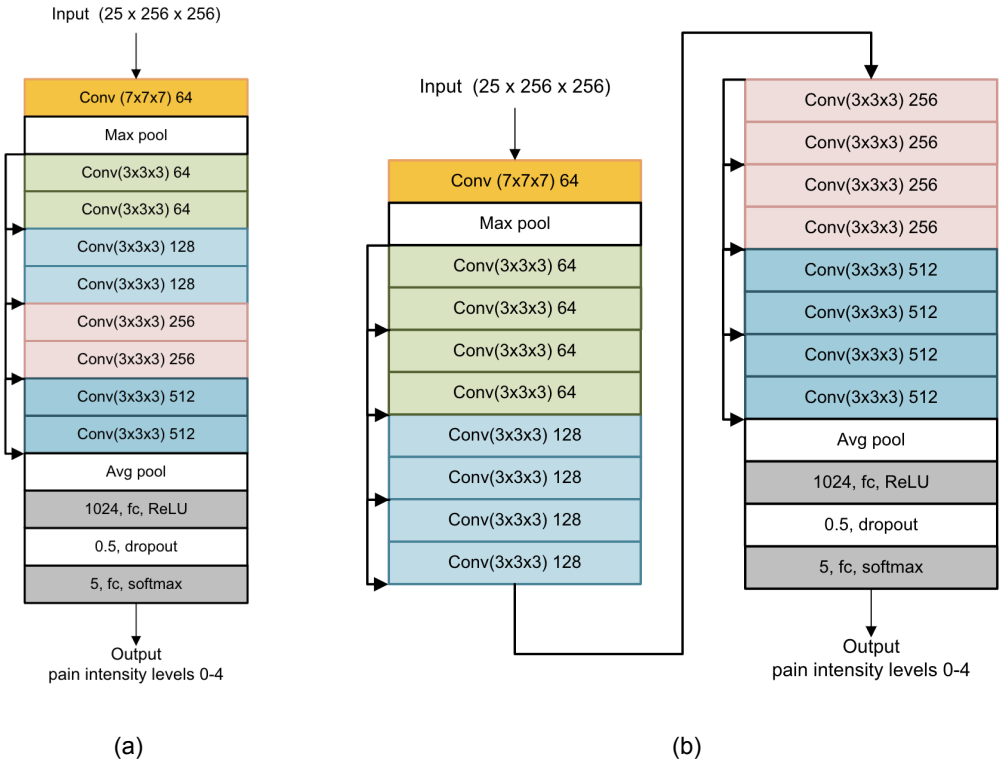


Figure 5.3: Res3D depth variations. (a) Res3D-10 Architecture. (b) Res3D-18 Architecture.

5.4.2 R(2+1)D

R(2+1)D [59] was an architecture developed for human action recognition and achieved state-of-the-art performance on both the Sports-1M [60] and the Kinetics [61] dataset. The core idea behind R(2+1)D is to split the traditional 3D convolution (e.g. a $3 \times 3 \times 3$ kernel into a spatial 2D convolution $1 \times 3 \times 3$ and a temporal convolution 3×1×1). For a 3D convolutional layer of N_i filters with size $(N_{i-1} \times t \times d \times d)$, we can retain the number of parameters by setting the number of intermediate filters M_i as $\left\lfloor \frac{td^2 N_{i-1} N_i}{d^2 N_{i-1} + t N_i} \right\rfloor$. By splitting the 3D convolutional layers, an extra layer of Rectified Linear Unit (ReLU) can be introduced to improve the models' ability to learn more complex problems.

The model has shown better convergence compared to the traditional Res3D with lower training error, especially when the model depth increases. Figure 5.4 presents the model architecture used in this experiment. Similarly, R(2+1)D at depths 10, 18, and 34 were tested.

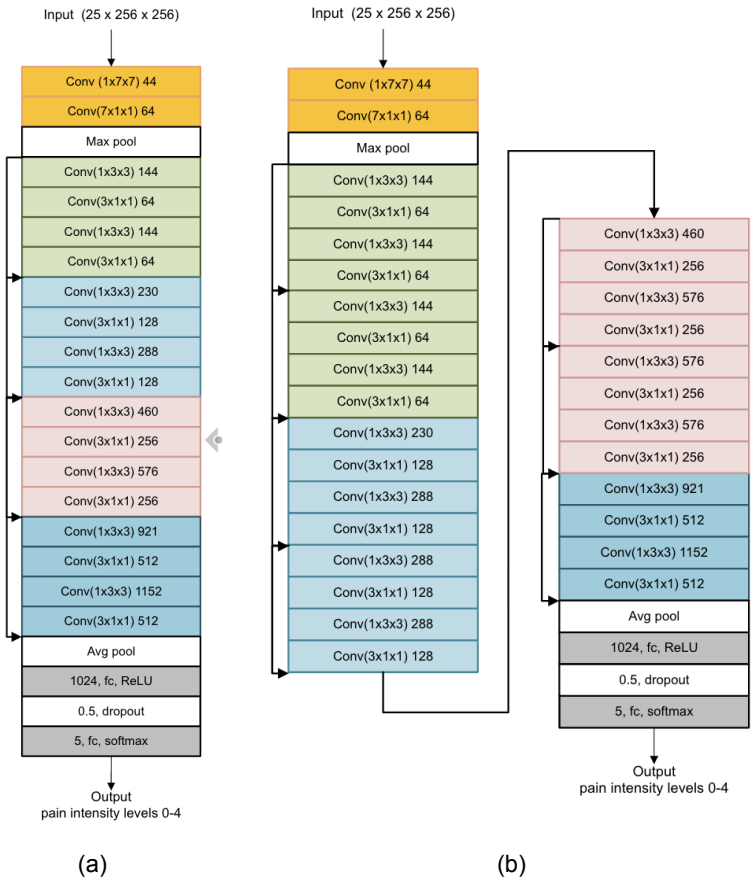


Figure 5.4: R(2+1)D model architecture at various depths. (a) R(2+1)D-10. (b) R(2+1)D-18.

5.5 Model explainability

In this experiment, two popular Explainable AI (XAI) methods were used, Gradient-weighted Class Activation Mapping (Grad-CAM) and guided backpropagation. Due to the additional temporal dimension of the model, the mean of the Grad-CAM and guided backpropagation was taken and superimposed together with the middle frame of the input video data. From XAI methods, the aim is to assess the model's performance and improve upon it by identifying the problem and potential failure mode. At the same time, the transparency allowed me to interpret the features behind the model's decision-making. XAI facilitates a deeper understanding of complex features, such as facial expressions indicative of pain, enriching our knowledge base regarding facial response mechanisms.

5.5.1 Grad-CAM

Grad-CAM [62], or Gradient-weighted Class Activation Mapping, is a popular XAI method that can be implemented on any CNNs. It provides insights into which part of the area contributed the most to the model's prediction by producing a localisation heatmap highlighting the important regions. This is done by first obtaining the average gradient of the target class with respect to the last convolutional layer for each feature map (or the global-average-pooled gradient) as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{5.1}$$

where we can calculate the the neuron importance weight α_k^c of class c for each feature map k, with y^c denoting the output label of class c and A_{ij}^k as the activation for pixel at (i, j) of feature map k.

The overall Grad-CAM localisation map ,$L_{Grad-CAM}^c$, can then be calculated by taking the weighted combination of the importance weight with its corresponding feature map activations, followed by a ReLU layer, defined as:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A_{ij}^k) \tag{5.2}$$

Grad-CAM provides a class-discriminative interpretation method for understanding a complex deep CNNs feature selection process. However, the localisation maps created are constrained by the size of the feature maps of the last layer, and upscaling is required, which signifi-

cantly degrades the detail. Therefore, combining Grad-CAM with the guided backpropagation method is recommended to achieve both high-resolution and class-discriminative visualisations.

5.5.2 Guided backpropagation

Guided backpropagation [63] creates a high-resolution feature map by controlling the gradient backpropagation to the CNN model's top layer (input layer). When the gradients backpropagate through a ReLU layer, it ensures that only non-negative gradients can pass through, applying a ReLU to the backward gradients. This allows the first layer gradients to highlight regions and features that positively correlate to the prediction.

5.6 Evaluation

The same evaluation metrics as section 4.6 were used in this experiment. To evaluate model performance during training, I used Cross-Entropy Loss for multi-class classification (pain level 0-5) defined as:

$$\begin{aligned} J_{CE}(x,y) &= \frac{1}{N}\sum_{n=1}^{N} l_n \\ l_n &= -\sum_{c=1}^{C} log\frac{exp(x_{n,c})}{\sum_{c=1}^{C} exp(x_{n,c})} y_{n,c} \end{aligned} \tag{5.3}$$

and Binary Cross-Entropy Loss (see Equation (5.4)) for binary pain vs no pain classification, defined as:

$$J_{BCE}(x,y) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot log(\sigma(x_i)) + (1-y_i)\cdot log(1-\sigma(x_i))] \tag{5.4}$$

where $\sigma(\cdot)$ denotes a softmax function with (x_i, y_i) as input data and its ground truth label.

The same training-validation split in the previous chapter was used to ensure comparability and generalisability. Due to the input data size and hardware limitations, a coarse grid search with the Stochastic Gradient Descent (SGD) optimiser parameters was implemented and evaluated on the training data for each model. The model with the lowest evaluation loss was selected. The model was then trained with the entire training set and evaluated on the predefined validation set with participants that it had not seen.

Table 5.1: The accuracy of the two 3D CNN models at different depths.

Model	Number of Parameters (Million)	Binary Classification	5-Class Classification
		[0-4]	[all-all]
Res3D-10	14.885	0.6025	0.2446
Res3D-18	33.690	**0.6044**	0.2457
Res3D-34	64.000	0.6034	**0.2469**
R(2+1)D-10	14.885	0.5938	0.2427
R(2+1)D-18	33.690	0.5996	0.2434
R(2+1)D-34	64.000	0.5957	0.2430

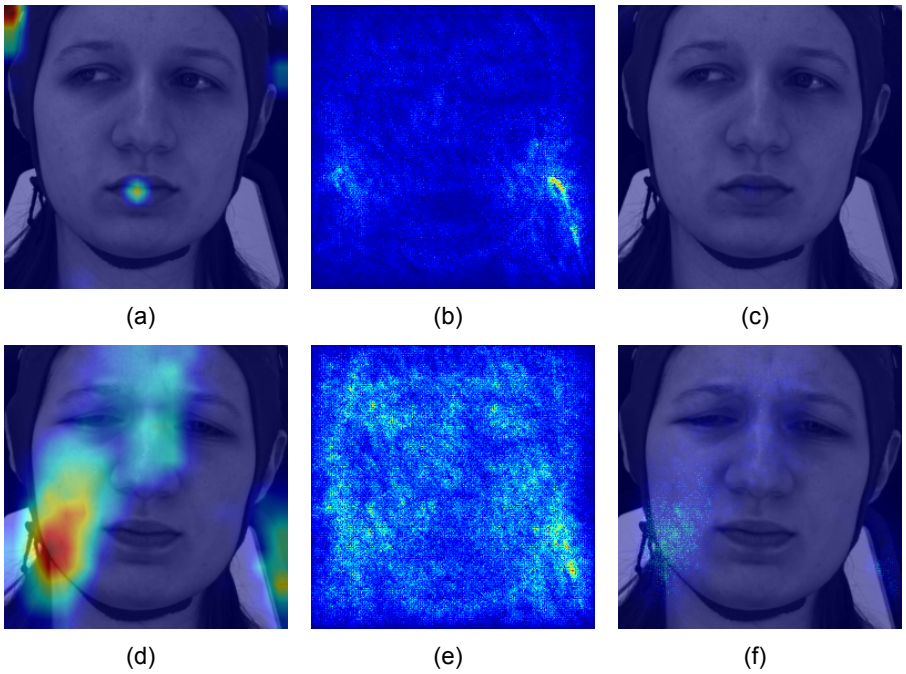


Figure 5.5: Model explainability heatmap on an example participant with no pain (a-c) and pain level 4, the highest level of pain (d-f). (a), (d) Grad-CAM heatmap superimposed with the centre frame of the input. (b), (e) Guided backpropagation heatmap. (c), (f) Guided Grad-CAM superimposed with the centre frame.

5.7 Results

Table 5.1 shows the experiment results for each model and its number of trainable parameters. The Grad-CAM and guided backpropagation result of the same participant experiencing pain level 0 (no pain) and pain level 4 (the highest level of pain) was shown in Figure 5.5. Due to the separated temporal convolutional layers in the R(2+1)D model, I was able to analyse the weights applied to each frame and gain more insights into how the model analysed the temporal evolution (see Figure 5.6).

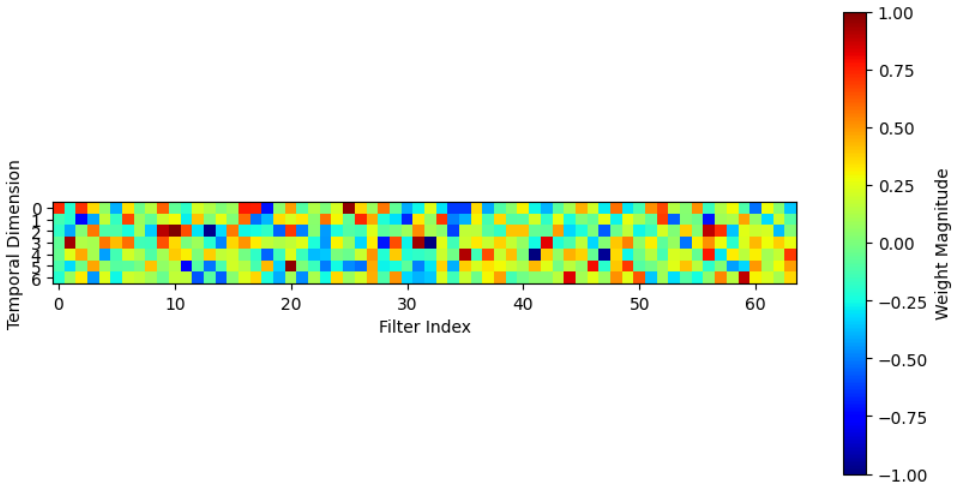


Figure 5.6: The 64 R(2+1)D temporal filters learned at the first layer (Averaged and Normalized). Each filter has size 7×1×1, and the mean was taken for each of the 44 input channels. The weights were then normalised to [-1, 1] for better presentation. For a well-generalised model, the temporal filters should have similar patterns that match the expected temporal evolution of a leading edge. However, there is a high disparity in the temporal filters here with no clear weight patterns.

5.8 Discussion

In general, Res3D architecture has shown a slightly higher classification accuracy than the R(2+1)D models. The increase in models' depth did not significantly improve the predictive performance. This suggests that the complexity at a depth of 10 is sufficient for this task. However, all the models' predictive performances were significantly inferior to the performance of the method proposed in chapter 4. I expected a higher performance due to CNNs' ability to train and select features. Conversely, the inferior performance suggested that the preprocessed data fed into the model contained less meaningful information and had low generalisability. The Guided Grad-CAM supported this argument. Guided backpropagation could not show specific facial features, indicating that the model has not fully learned any facial features with temporal evolution.

The loss of information in this experiment mainly came from participants' head movements captured by the time window selection process. Each clip was part of an entire experiment recording lasting over 20 minutes, and they were extracted by simply matching the recorded heating time. This did not ensure that the subjects were idle at the start of the clip, causing a large amount of pixel changes when the subject returned to the idle head position, hence resulting in the wrong time window captured before training. Figure 5.7 presents one example of

a subject with a non-idle keyframe and the resulting pixel changes signal.

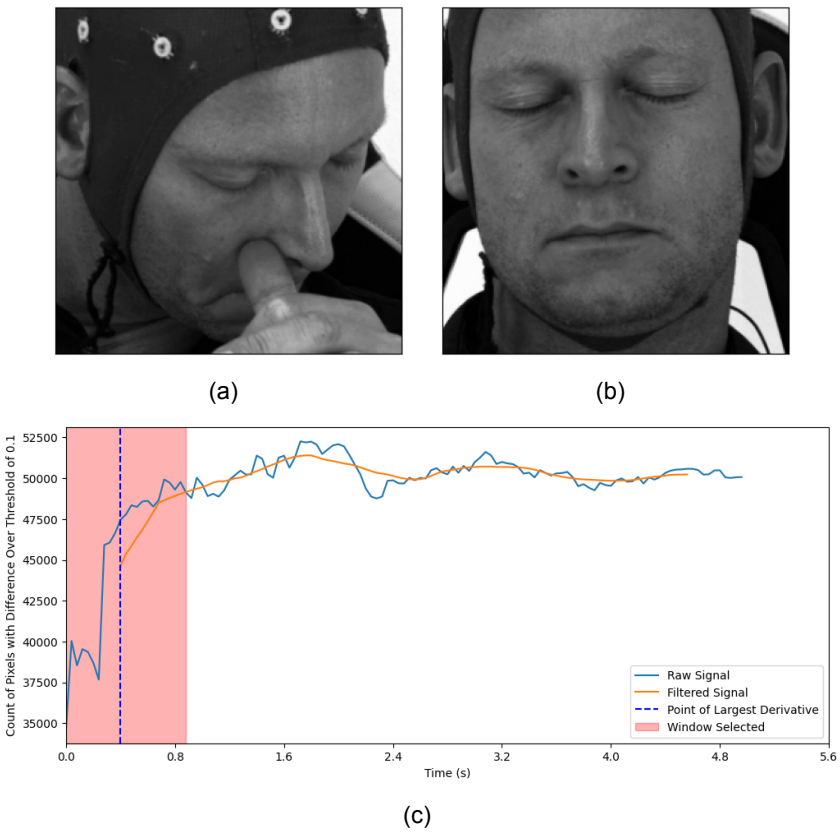


Figure 5.7: The first frame (keyframe) of this data example shows the subject scratching his nose but quickly going back to the normal idle position. (a) Keyframe with unwanted behaviour. (b) Frame 40 when the subject returned back to an idle position. (c) Pixel changes signal with unwanted behaviour, and the red labelled region was selected as the input data.

Head rotations would also result in the wrong time window being selected. Basic face alignment was implemented during the face extraction process. However, the alignment process is limited to translation and cannot handle head rotations when part of the facial information is lost. This would result in the time window selector choosing the head rotation movement as the training data with a large number of pixel changes irrelevant to pain-induced facial expression. The first temporal convolutional layer filter weights supported this claim. There was a high disparity in the weights of the 64 filters, indicating that the model was starting to learn from noise and irrelevant movements.

On the other hand, the spatial information was correctly captured by the tested models. For the subject experiencing high pain, the Grad-CAM heatmap put high attention on the subject's right cheek and nasal bridge, which was expected and supported by the results of Chapter 4. However, different from the behaviours of the handcrafted approach, less weight had been

suggested on the mouth. Grad-CAM showed no attention on the face for data with no pain, while guided backpropagation also highlighted the backgrounds around the cheeks. This suggested that the model uses low activation as the basis for classifying no-pain. The model also showed interesting behaviour by focusing on only one side of the face. This can be a potential indicator of the model overfitting or training with biased data. Further investigations are required to draw the conclusion.

5.9 Conclusion

This chapter presented a 3D CNN-based method of representation learning. A preprocessing pipeline was implemented to simplify the input data and reduce redundant spatial and temporal information before training and testing with two different CNN architectures, Res3D and R(2+1)D. Res3D has shown superior performance, with a depth of 10 being sufficient to extract frame-level facial expressions. Grad-CAM and guided backpropagation were used to provide more insights into the model performance. The spatial facial features highlighted by Grad-CAM agreed with prior studies, such as the FACS and the results from Chapter 4; however, they contradicted the importance of mouth movements.

The model faced limitations in generalisability, evidenced by its poor validation accuracy. The primary contributor to this was the misleading information derived from subjects' head movements. While initial attempts to mitigate these effects through time window preprocessing strategies were implemented, they proved insufficient. Addressing this challenge will remain a focal point for future studies.

Chapter 6

Conclusion and future work

6.1 Introduction

Accurate and efficient pain assessment is vital in providing the appropriate diagnosis and treatment, especially for vulnerable patients who cannot communicate directly. Automatic pain assessment using facial expressions offers a promising solution to this pervasive challenge. This chapter identifies the contributions made by this study, summarises the findings of the report, and presents future research directions.

6.2 Contributions

The main contributions of this report include:

1. The development of two advanced pain assessment models. The first utilising hand-crafted features with an innovative approach to feature extraction. The second with a 3D-CNN architecture that incorporates novel preprocessing pipelines that significantly reduce data redundancy.
2. Critical evaluation revealed the unique limitations of each method and paved the way for future research into developing an automatic pain assessment tool that can be implemented in clinical practice.
3. An investigation to interpret the models' decision-making with explainable AI techniques to pinpoint the most important facial feature that are related to pain.
4. Validation against current pain scales confirms the viability of facial pain assessment and suggests potential improvements over existing pain scales.

6.3 Summary of findings

One of the key challenges of making an automatic pain assessment tool clinically ready is to present the rationale behind such a tool instead of a black box. This requires two steps: identifying facial muscle movements that relate to pain expressions and interpreting models' decision-making process.

Chapter 4 presents an automatic pain assessment algorithm with handcrafted features inspired by prior works and the FACS system. The model tracked 13 facial features constructed by facial landmarks over the entire 5.5-second time window to form feature signals. Each signal was then processed and represented by 16 signal descriptors that capture signal variability and duration before passing into the classifier. This algorithm has shown competitive predictive performance in binary classification between pain and no pain and, more importantly, provided valuable insights into the feature selected. Using importance metrics and SHAP [56], the model suggests that mouth changes (opening or stretching) have shown a dominant discriminative ability to classify pain, contrary to the PSPI pain scale. The nasal bridge, as well, has shown great importance, being one of the most frequently seen features in the ensemble tree models. Due to the temporal consideration of the algorithm, eye closures (commonly interpreted as the key feature) were not of high importance. The experiment suggests a change in the PSPI pain scale to include mouth movements as well as some measures for the duration of the AU changes.

In chapter 5, I proposed a different approach to finding the key facial features by using 3D-CNNs to learn their importance. I tested two 3D-CNN architectures, namely Res3D and R(2+1)D, at various depths to check the capability with a novel frame selection process to extract frames where the expression of pain manifests. The models have shown poor predictive performance due to subjects' excessive head movements, hurting the models' ability to learn the temporal evolution. However, the 3D-CNN models have shown a good understanding of the spatial features, agreeing with the prior experiment, suggesting the left cheek and nasal bridge are the most important regions. Conversely, less attention was paid to the mouth in this experiment, which can be a direction for future work.

6.4 Future work

6.4.1 Multi-Modal pain assessment

One potential solution for pain assessment with poor facial information is a multimodal approach incorporating facial expressions and biophysiological signals. Facial expression approaches provide a quick, objective, and anatomically based analysis of pain reactions but are limited to human observation capabilities. When an individual is experiencing a low level of pain, the subtle facial expressions often cannot be captured or assessed. This is the main reason for the

current models' poor low pain level classification performance. By using more biosignals such as facial muscle EMG and EEG, more subtle reactions can be captured, aiding Pain assessment tool (PAT)s' ability to quantify low pain levels. However, the limitations of facial EMG and the potential trade-off with the quality of facial expression recorded are worth noting. Facial EMG requires sensors over specific facial muscles related to pain, occluding these crucial areas in the facial recording and hurting the predictive performance of the facial expression-based model.

Future work may involve justifying the trade-off when using facial EMG as another modality and determining the optimal position of the EMG sensors to best capture the subtle muscle reactions.

6.4.2 Establishing ground truth

Another challenge for automatic pain assessment is the need for a standardised, globally accepted ground truth. As mentioned in Section 2.2, PSPI scale only quantifies the intensity of facial expression instead of the experience of pain. On top of that, the FACS AU manual labelling is extremely time-consuming and laborious to scale up. Other datasets utilised both the self-reported scale and observer-based pain scale that suffered from the same issue and lacked generalisability if used for cognitively impaired patients. The dataset used in this study, BioVid Heat Pain, used the stimulus heat level as the ground truth that significantly simplified the labelling process. However, it lacks comparability with expert baseline results. To claim that a PAT tool is clinically ready, it has to perform better than human experts and, therefore, requires baseline results by human experts on the same data. Consequently, it is crucial to establish a standardised ground truth that is scalable, comparable with expert results, and can capture the experience of pain instead of facial movements.

6.5 Concluding remarks

Facial expression-based automatic pain assessment is a capable solution to meet the clinical demand for accurate pain assessment of vulnerable patients without communication capabilities. It also opens up whole new opportunities to encourage a patient-oriented healthcare system by providing in-home assessments. Through two experiments, this study discovers the mechanisms of facial reactions to pain and provides a step forward towards a clinical-ready automatic pain assessment tool.

Bibliography

[1] H. Merskey, D. Albe-Fessard, J. Bonica, A. Carmen, R. Dubner, F. Kerr, and C. Pagni, "Editorial: The need of a taxonomy," *Pain*, vol. 6, no. 3, pp. 247–252, 1979.

[2] International Association for the Study of Pain, Subcommittee on Taxonomy, "Classification of chronic pain. descriptions of chronic pain syndromes and definitions of pain terms," *Pain Suppl*, vol. 3, pp. S1–S226, 1986.

[3] "Chronic pain in adults." `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/940858/Chronic_Pain_Report.pdf`, 2017. Accessed 08 Dec 2022.

[4] A. Fayaz, P. Croft, R. M. Langford, L. J. Donaldson, and G. T. Jones, "Prevalence of chronic pain in the uk: a systematic review and meta-analysis of population studies," *BMJ Open*, vol. 6, no. 6, 2016.

[5] S. M. Rikard, A. E. Strahan, K. M. Schmit, and G. P. J. Guy, "Chronic pain among adults - united states, 2019-2021," *MMWR Morb Mortal Wkly Rep*, vol. 72, no. 15, pp. 379–385, 2023.

[6] Z. Yongjun, Z. Tingjie, Y. Xiaoqiu, F. Zhiying, Q. Feng, X. Guangke, L. Jinfeng, N. Fachuan, J. Xiaohong, and L. Yanqing, "A survey of chronic pain in china," *Libyan J Med*, vol. 15, no. 1, p. 1730550, 2020.

[7] T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans, *et al.*, "Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the global burden of disease study 2010," *Lancet*, vol. 380, no. 9859, pp. 2163–2196, 2012. Erratum in: Lancet. 2013 Feb 23;381(9867):628. AlMazroa, Mohammad A [added]; Memish, Ziad A [added].

[8] S. Khalid, U. Sambamoorthi, A. Umer, C. Lilly, D. Gross, and K. Innes, "Increased odds of incident alzheimer's disease and related dementias in presence of common non-cancer chronic pain conditions in appalachian older adults," *Journal of Aging and Health*, vol. 34, pp. 158–172, 2022.

[9] Interagency Pain Research Coordinating Committee, "National pain strategy: A comprehensive population health-level strategy for pain." Washington, DC: US Department of Health and Human Services, National Institutes of Health, 2016.

[10] J. Ditre, E. Zale, and L. LaRowe, "A reciprocal model of pain and substance use: transdiagnostic considerations, clinical implications, and future directions," *Annual Review of Clinical Psychology*, vol. 15, pp. 503–528, 2019.

[11] J. Harrison, S. Weber, R. Jakob, *et al.*, "ICD-11: an international classification of diseases for the twenty-first century," *BMC Medical Informatics and Decision Making*, vol. 21, no. Suppl 6, p. 206, 2021.

[12] S. Deandrea, M. Montanari, L. Moja, and G. Apolone, "Prevalence of undertreatment in cancer pain. a review of published literature," *Annals of Oncology*, vol. 19, pp. 1985–1991, Dec 2008. Epub 2008 Jul 15.

[13] K. Torvik, S. Kaasa, Ø. Kirkevold, and T. Rustøen, "Pain in patients living in norwegian nursing homes," *Palliative Medicine*, vol. 23, pp. 8–16, Jan 2009. Epub 2008 Oct 24.

[14] P. Rahimzadeh, S. Safari, and F. Imani, “Pediatric chronic pain management: Steps toward a neglected area,” *Journal of Comprehensive Pediatrics*, vol. 4, no. 1, pp. 47–8, 2013.

[15] M. Alnajar, R. Shudifat, S. Mosleh, S. Ismaile, M. N’Erat, and K. Amro, “Pain assessment and management in intensive care unit: Nurses’ practices, perceived influencing factors, and educational needs,” *The Open Nursing Journal*, vol. 15, no. 1, pp. 170–178, 2021.

[16] C. von Baeyer, L. Spagrud, J. McCormick, *et al.*, “Three new datasets supporting use of the numerical rating scale (nrs-11) for children’s self-reports of pain intensity,” *Pain*, vol. 143, pp. 223–227, 2009.

[17] W. Camann, “Visual analog scale scores for labor pain,” *Anesthesia Analgesia*, vol. 88, pp. 1421–1429, 1999.

[18] D. L. Wong and C. Baker, “Pain in children: Comparison of assessment scales,” *Pediatric Nursing*, 1998.

[19] R. Luffy, “Examining the validity, reliability, and preference of three pediatric pain measurement tools in african-american children,” *Pediatric Nursing*, vol. 29, no. 1, pp. 54–59, 2003.

[20] K. Herr, P. Coyne, M. McCaffery, R. Manworren, and S. Merkel, “Pain assessment in the patient unable to self-report: Position statement with clinical practice recommendations,” *Pain Management Nursing*, vol. 12, pp. 230–250, Dec 2011.

[21] J. Kim and E. Andre, “Emotion recognition based on physiological changes in music listening,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[22] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[23] P. Ekman, W. V. Friesen, and J. C. Hager, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco, CA: Consulting Psychologists Press, 2002.

[24] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, “Automatic pain assessment with facial activity descriptors,” *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 286–299, 2017.

[25] M. Kunz and S. Lautenbacher, “The faces of pain: A cluster analysis of individual differences in facial activity patterns of pain,” *European Journal of Pain*, vol. 18, no. 6, pp. 813–823, 2014.

[26] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, “Blazeface: Sub-millisecond neural face detection on mobile gpus,” 2019.

[27] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-stage dense face localisation in the wild,” 2019.

[28] W. Wu, H. Peng, and S. Yu, “Yunet: A tiny millisecond-level face detector,” *Machine Intelligence Research*, vol. 20, pp. 656–665, 2023.

[29] J. Egede, S. Song, T. Olugbade, C. Wang, A. Williams, H. Meng, M. Aung, N. Lane, M. Valstar, and N. Bianchi-Berthouze, “Emopain challenge 2020: multimodal pain evaluation from facial and bodily expressions,” in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG2020)*, pp. 849–856, 2020.

[30] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[32] X. Xin, X. Lin, S. Yang, and X. Zheng, “Pain intensity estimation based on a spatial transformation and attention cnn,” *PLoS ONE*, vol. 15, no. 8, p. e0232412, 2020.

[33] D. Huang, Z. Xia, L. Li, K. Wang, and X. Feng, “Pain-awareness multistream convolutional neural network for pain estimation,” *Journal of Electronic Imaging*, vol. 28, p. 043008, Jul 2019. Published 11 July 2019.

[34] J. Wang and H. Sun, “Pain intensity estimation using deep spatiotemporal and hand-crafted features,” *IEICE Transactions on Information and Systems*, vol. E101D, no. 6, pp. 1572–1580, 2018.

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, 2015.

[36] M. Tavakolian and A. Hadid, “Deep spatiotemporal representation of the face for automatic pain intensity estimation,” in *2018 24th International Conference on Pattern Recognition (ICPR)*, vol. 2018-Augus, pp. 350–354, Institute of Electrical and Electronics Engineers Inc., 2018.

[37] G. Bargshady, X. Zhou, R. Deo, J. Soar, F. Whittaker, and H. Wang, “Enhanced deep learning algorithm development to detect pain intensity from facial expression images,” *Expert Systems with Applications*, vol. 149, 2020.

[38] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proceedings of the British Machine Vision Conference 2015*, pp. 1–12, 2015.

[39] Z. Huang, W. Xu, and K. Yu, “Bidirectional lstm-crf models for sequence tagging,” 2015.

[40] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. Moreira da Silva, “The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system,” in *2013 IEEE International Conference on Cybernetics (CYBCO)*, pp. 128–131, 2013.

[41] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, “Painful data: the unbc-mcmaster shoulder pain expression archive database,” in *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011)*, pp. 57–64, IEEE, 2011.

[42] M. Haque, R. Bautista, F. Noroozi, K. Kulkarni, C. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, O. Andersen, E. Spaich, and T. Moeslund, “Deep multimodal pain recognition: a database and comparison of spatio-temporal visual modalities,” in *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pp. 250–257, IEEE, 2018.

[43] S. Brahnam, C.-F. Chuang, F. Shih, and M. Slack, “Svm classification of neonatal facial images of pain,” in *Fuzzy Logic and Applications* (I. Bloch, A. Petrosino, and A. Tettamanzi, eds.), pp. 121–128, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[44] S. Brahnam, L. Nanni, S. McMurtrey, A. Lumini, R. Brattin, M. Slack, and T. Barrier, "Neonatal pain detection in videos using the icopevid dataset and an ensemble of descriptors extracted from gaussian of local descriptors," *Applied Computing and Informatics*, 2019.

[45] G. Zamzmi, P. Chih-Yun, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "A comprehensive and context-sensitive neonatal pain assessment using computer vision," *IEEE Transactions on Affective Computing*, 2019.

[46] M. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh, A. Elkins, N. Kanakam, A. de Rothschild, N. Tyler, P. Watson, A. de C Williams, M. Pantic, and N. Bianchi-Berthouze, "The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 435–451, 2016.

[47] J. Sellner, P. Thiam, and F. Schwenker, "Visualizing facial expression features of pain and emotion data," in *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction* (F. Schwenker and S. Scherer, eds.), vol. 11377 of *Lecture Notes in Computer Science*, (Cham), Springer, 2019.

[48] P. Werner, A. Al-Hamadi, and S. Walter, "Analysis of facial expressiveness during experimentally induced heat pain," in *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017.

[49] Google, "Face detection." `https://developers.google.com/mediapipe/solutions/vision/face_detection`, 2024. Accessed: 20-Apr-2024.

[50] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," 2019.

[51] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, p. 1499–1503, Oct. 2016.

[52] Google, "Facemesh." `https://google.github.io/mediapipe/solutions/face_mesh.html`, 2024. Accessed: 2024-04-20.

[53] K. M. Prkachin, "The consistency of facial expressions of pain: A comparison across modalities," *Pain*, vol. 51, no. 3, pp. 297–306, 1992.

[54] 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*, KDD '16, ACM, Aug. 2016.

[55] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[56] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[57] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain recognition from video and biomedical signals," in *Proceedings of the International Conference on Pattern Recognition*, pp. 4582–4587, 2014.

[58] S. Gkikas and M. Tsiknakis, “Automatic assessment of pain based on deep learning methods: A systematic review,” *Computer Methods and Programs in Biomedicine*, vol. 231, p. 107365, Apr. 2023. Epub 2023 Feb 8.

[59] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” 2018.

[60] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *CVPR*, 2014.

[61] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” 2017.

[62] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” *International Journal of Computer Vision*, vol. 128, p. 336–359, Oct. 2019.

[63] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” 2015.

4YP Risk Assessment 2022

Description of 4YP task or aspect being risk assessed here: **Office work** **Deep Learning based facial expression recognition for pain assessment.**		4YP Project Number: **13092**
Site, Building & Room Number: **Institute of Biomedical Engineering Building**	Other Relevant risk Assessments:	
Assessment undertaken by: Aaron Chen	Signed: Aaron.	Date: 29/10/2023
Assessment Supervisor:	Signed:	Date:

Assessing the Risk*

You can do this for each hazard as follows:

- Consequences: Decide how severe the outcome for each hazard would be if something went wrong (i.e. what are the Consequences?) Death would be "Severe", a minor cut to a finger could be regarded as "Insignificant".
- Likelihood: How likely are these Consequences to actually happen? Highly likely? Remotely likely, or somewhere in between?
- Risk Rating: Start at the left of the coloured Matrix. On your chosen Consequences row, read across until you are in the correct Likelihood column for the hazard in question. For example, an outcome with Severe consequences but with a Low probability of actually happening equates to a Medium risk overall. In this case "Medium" is what should be written in the Risk.

RISK MATRIX

CONSEQUENCES	LIKELIHOOD (or probability)			
	High	**Medium**	**Low**	**Remote**
Severe	**High**	**High**	**Medium**	**Low**
Moderate	**High**	**Medium**	**Medium/Low**	**Effectively Zero**
Insignificant	**Medium/Low**	**Low**	**Low**	**Effectively Zero**
Negligible	**Effectively Zero**	**Effectively Zero**	**Effectively Zero**	**Effectively Zero**

Overall statement of risk

- Carefully consider the risks associated with your project, the nature of the activity with which you will be engaged, and its location.
- Check the information from Health and Safety pages in the intranet including those specifically for the 4YP.

Students must discuss these risks with their supervisor.

☑ **Office work only**. My project involves only basic office work (paper and computers). It does not involve hands-on laboratory or field work of any kind. I am aware of the associated risks, including the health risks associated with the extended use of computers and display screens. No further assessment is required.

☐ **Low Risk**. I consider the health and safety risks associated with my project to be low, working in alignment with existing risk assessments, I have referenced relevant risk assessments above and have agreed with my supervisor that no further assessment is required. For example, collecting data from existing systems within a lab.

☐ **Medium Risk**. I consider there to be additional risks associated with my project as it requires risk assessment authorisation below:

Risk Assessments for Hazardous Substances & Biological Materials. The Biological & Chemical Safety Officer's (BSCO) signature is required for the final sign-off on Engineering Science COSHH Assessments. If the BCSO is unavailable the DSO can provide this signature. For IBME, the IBME Safety Officer can provide this signature. Reference E refers. The BCSO's signature is also required for risk assessments involving the use of biological materials.

Genetically Modified Organisms. Risk assessments involving genetically modified organisms require the BSCO's signature as well as approval from the Genetic Modification Safety Committee for the work to proceed. The department's Safety Policy refers.

Laser Risk Assessments: In addition to the supervisor of the laser equipment/experiment concerned, the Department Laser Safety Officer (DLSO) must also sign risk assessments involving lasers.

Where Specialist Safety Officers Originate Risk Assessments. Where the DSO or Specialist Safety Officers write, co-write or otherwise originate risk assessments they will be required to sign and authorize such risk assessments.

Requirements for review by specialists should be identfied within Safety Requirements section on https://fouryp.eng.ox.ac.uk/resourcetimepreview2.php

☐ **High Risk**. This is a high risk activity as identified by Specialist Safety Officers.

Please review with Specialist Safety Officers where projects are Medium Risk sign below, ask your supervisor to countersign and then submit to Sharepoint site.

Signature of student:	**Signature of supervisor:**
Date: Aaron. 29/10/2023.	**Date:**

4YP Risk Assessment 2022

Description of 4YP task or aspect being risk assessed here: **Office work** **Deep Learning based facial expression recognition for pain assessment.**		4YP Project Number: **13092**
Site, Building & Room Number: **Institute of Biomedical Engineering Building**	Other Relevant risk Assessments:	
Assessment undertaken by: Aaron Chen	Signed: Aaron.	Date: 29/10/2023
Assessment Supervisor:	Signed:	Date:

Assessing the Risk*

You can do this for each hazard as follows:

- Consequences: Decide how severe the outcome for each hazard would be if something went wrong (i.e. what are the Consequences?) Death would be "Severe", a minor cut to a finger could be regarded as "Insignificant".
- Likelihood: How likely are these Consequences to actually happen? Highly likely? Remotely likely, or somewhere in between?
- Risk Rating: Start at the left of the coloured Matrix. On your chosen Consequences row, read across until you are in the correct Likelihood column for the hazard in question. For example, an outcome with Severe consequences but with a Low probability of actually happening equates to a Medium risk overall. In this case "Medium" is what should be written in the Risk.

RISK MATRIX

CONSEQUENCES	LIKELIHOOD (or probability)			
	High	**Medium**	**Low**	**Remote**
Severe	**High**	**High**	**Medium**	**Low**
Moderate	**High**	**Medium**	**Medium/Low**	**Effectively Zero**
Insignificant	**Medium/Low**	**Low**	**Low**	**Effectively Zero**
Negligible	**Effectively Zero**	**Effectively Zero**	**Effectively Zero**	**Effectively Zero**

Overall statement of risk

- Carefully consider the risks associated with your project, the nature of the activity with which you will be engaged, and its location.
- Check the information from Health and Safety pages in the intranet including those specifically for the 4YP.

Students must discuss these risks with their supervisor.

☑ **Office work only**. My project involves only basic office work (paper and computers). It does not involve hands-on laboratory or field work of any kind. I am aware of the associated risks, including the health risks associated with the extended use of computers and display screens. No further assessment is required.

☐ **Low Risk**. I consider the health and safety risks associated with my project to be low, working in alignment with existing risk assessments, I have referenced relevant risk assessments above and have agreed with my supervisor that no further assessment is required. For example, collecting data from existing systems within a lab.

☐ **Medium Risk**. I consider there to be additional risks associated with my project as it requires risk assessment authorisation below:

Risk Assessments for Hazardous Substances & Biological Materials. The Biological & Chemical Safety Officer's (BSCO) signature is required for the final sign-off on Engineering Science COSHH Assessments. If the BCSO is unavailable the DSO can provide this signature. For IBME, the IBME Safety Officer can provide this signature. Reference E refers. The BCSO's signature is also required for risk assessments involving the use of biological materials.

Genetically Modified Organisms. Risk assessments involving genetically modified organisms require the BSCO's signature as well as approval from the Genetic Modification Safety Committee for the work to proceed. The department's Safety Policy refers.

Laser Risk Assessments: In addition to the supervisor of the laser equipment/experiment concerned, the Department Laser Safety Officer (DLSO) must also sign risk assessments involving lasers.

Where Specialist Safety Officers Originate Risk Assessments. Where the DSO or Specialist Safety Officers write, co-write or otherwise originate risk assessments they will be required to sign and authorize such risk assessments.

Requirements for review by specialists should be identfied within Safety Requirements section on https://fouryp.eng.ox.ac.uk/resourcetimepreview2.php

☐ **High Risk**. This is a high risk activity as identified by Specialist Safety Officers.

Please review with Specialist Safety Officers where projects are Medium Risk sign below, ask your supervisor to countersign and then submit to Sharepoint site.

Signature of student:	**Signature of supervisor:**
Date: Aaron. 29/10/2023.	**Date:**

Hazard *(potential for harm)*	**Persons at Risk**	**Risk Controls In Place** *(existing safety precautions)*	**Risk***	**Future Actions identified to Reduce Risks** *(but not in place yet)*
Slips, trips and fall	**Student, Staff**	**General House Keeping.** **No trailing leads and cables.** **Offices cleaned every evening.**	**Low**	
Working at height Filings on top shelves, putting up decorations, etc.	**Student, Staff**	**Stable chairs and tables.** **Staff putting up decorations.**	**Low**	**Buy appropriate stepladder.**
Electrical appliances Electrical shocks, untidy cords.	**Student, Staff**	**Defective equipment taken out of use safely and promptly replaced.** **Regular inspection.** **Adequate grounding of appliances.** **Cords organised in places not posing the risk of tripping.** **Power limits of electrical appliances not exceeded.**	**Low**	**Use labelled or coloured cords to help organise.**
Stress	**Student, Staff**	**Students understand what their duties and projects are.** **Students are free to talk with supervisors if they are feeling unwell or at ease about things in office.**	**Low**	**Office socials to create a more friendly environment**
General Laptop Work Posture problems and pain in arms/ hands, Headaches or sore eyes, Improper use of the workstation, etc.	**Student, Staff**	**Work planned to include regular breaks or change of activities.** **Workstation and equipment set to ensure good posture.** **Noise level controlled.** **Adjustable blinds at window to control natural light on screen.** **Lighting and temperature suitably controlled.**	**Low**	
Fire	**Student, Staff**	**Fire risk assessment done.**	**Low**	**Ensure the actions identified as necessary by the fire risk assessment are done.**

Hazard *(potential for harm)*	**Persons at Risk**	**Risk Controls In Place** *(existing safety precautions)*	**Risk***	**Future Actions identified to Reduce Risks** *(but not in place yet)*
Slips, trips and fall	**Student, Staff**	**General House Keeping.** **No trailing leads and cables.** **Offices cleaned every evening.**	**Low**	
Working at height Filings on top shelves, putting up decorations, etc.	**Student, Staff**	**Stable chairs and tables.** **Staff putting up decorations.**	**Low**	**Buy appropriate stepladder.**
Electrical appliances Electrical shocks, untidy cords.	**Student, Staff**	**Defective equipment taken out of use safely and promptly replaced.** **Regular inspection.** **Adequate grounding of appliances.** **Cords organised in places not posing the risk of tripping.** **Power limits of electrical appliances not exceeded.**	**Low**	**Use labelled or coloured cords to help organise.**
Stress	**Student, Staff**	**Students understand what their duties and projects are.** **Students are free to talk with supervisors if they are feeling unwell or at ease about things in office.**	**Low**	**Office socials to create a more friendly environment**
General Laptop Work Posture problems and pain in arms/ hands, Headaches or sore eyes, Improper use of the workstation, etc.	**Student, Staff**	**Work planned to include regular breaks or change of activities.** **Workstation and equipment set to ensure good posture.** **Noise level controlled.** **Adjustable blinds at window to control natural light on screen.** **Lighting and temperature suitably controlled.**	**Low**	
Fire	**Student, Staff**	**Fire risk assessment done.**	**Low**	**Ensure the actions identified as necessary by the fire risk assessment are done.**