ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in

Computer Vision

AUTOMATED PAIN ASSESSMENT FROM FACIAL EXPRESSIONS OF ELDERLY PEOPLE USING MACHINE LEARNING

CANDIDATE	SUPERVISOR
Lorenzo Massa	Prof. Luigi Di Stefano
	CO-SUPERVISOR
	Mr. Dries Lorent

Academic year 2023-2024

Session 2nd

Contents

Abstract

This thesis presents the development of an advanced machine learning model designed to accurately assess pain levels in dementia patients residing in elderly care homes. The project, conducted in collaboration with Sentigrate, a start-up focused on data science company, aims to create a predictive model that assigns pain scores ranging from 0 (no pain) to 6 (maximum pain) based on facial expressions. The research employs computer vision techniques, primarily convolutional neural networks, to extract meaningful features from facial images. A comparative study of various predictive techniques is conducted to determine the most effective approach. This project addresses the critical issue of inadequate pain management in dementia patients due to communication challenges. The objective is to provide an objective pain assessment tool that will significantly improve pain management strategies and enhance the quality of life for dementia patients in elderly care settings. The findings of this research have the potential to transform elderly care practices, offering valuable insights into pain management and contributing to the broader field of healthcare technology.

Chapter 1

Introduction

1.1 Motivation

The assessment of pain in elderly patients, particularly those with dementia, represents a significant challenge within the field of healthcare. These individuals often encounter difficulties in verbally communicating their pain, which may result in inadequate treatment and a reduction in quality of life. The lack of adequate pain management can precipitate a decline in cognitive function, elevate the risk of developing depressive and anxiety disorders, and give rise to behavioral disturbances such as aggression and agitation. Moreover, chronic pain can result in a diminished capacity to perform activities of daily living, which in turn can compromise the patient's autonomy and overall well-being. It is therefore imperative to address pain in patients with dementia, not only for their comfort but also to improve their overall health outcomes. Furthermore, the urgency of addressing this issue is compounded by the growing aging population and the increasing prevalence of dementia worldwide. Conventional pain assessment techniques rely predominantly on self-reporting or observational scales, which are susceptible to subjectivity and may be time-consuming. Moreover, these methods present significant challenges when applied to patients who are non-verbal, which may result in potential under-treatment. The advent of machine learning and computer vision technologies offers a promising opportunity for the development of automated, objective pain assessment tools. By analyzing facial expressions, which have been demonstrated to be reliable indicators of pain, we may be able to develop a non-invasive, efficient method for pain detection and monitoring in people affected by dementia. Such an automated system has the potential to enhance the quality of care and patient outcomes by facilitating more precise and personalized pain management strategies. Furthermore, it could furnish healthcare professionals with valuable insights into pain patterns and intensities, enabling more timely interventions and better-informed treatment decisions.

1.2 Objectives of the Study

The objective of this study is to develop and evaluate an automated pain assessment system that employs machine learning techniques to analyze the facial expressions of elderly individuals, with a particular focus on those with dementia. The primary objectives are as follows:

- To design and implement a robust machine learning model capable of accurately detecting and classifying pain levels from facial expressions in elderly patients.

- To compare the performance of different deep learning architectures, including Convolutional Neural Networks (CNNs) and Vision Transformers, for this specific task.

- To investigate the interpretability of the developed models using attribution methods, providing insights into which facial features are most indicative of pain.

- To evaluate the ethical implications and potential clinical applications of such an automated pain assessment system in elderly care settings.

This research is conducted in collaboration with Sentigrate, a data-centric consulting startup based in Leuven, Belgium. Sentigrate specializes in sensor data and machine learning, and its dynamic team of data scientists is known for creating innovative solutions that extract valuable insights from complex data sets. The company's expertise and collaborative environment make it an ideal partner for this ambitious project.

1.3 Structure of the Thesis

This thesis is structured into multiple chapters, each addressing a specific research objective. Following this introduction, Chapter 2 presents a review of the existing literature on dementia, pain in dementia patients, and current pain assessment methods. Chapter 3 explores the theoretical background of facial expression recognition, relevant machine learning architectures, and ethical considerations. Chapter 4 outlines the methodologies employed, including details on dataset preparation, model design, and evaluation metrics. Chapter 5 presents the results of the experiments, with an analysis of model performance and interpretability. Chapter 6 discusses the findings, addressing limitations and implications for practice and research. Finally, Chapter 7 concludes the thesis, summarizing key findings, discussing challenges, and proposing directions for future research and clinical integration.

Chapter 2

Research Background

2.1 Dementia: An Overview

Dementia is a debilitating neurological disorder that is characterized by a progressive decline in cognitive functions, including memory, reasoning, language, and executive abilities. Dementia is not a single disease entity but an umbrella term encompassing a range of conditions. Alzheimer's disease represents the most common form of dementia, followed by vascular dementia and Lewy body dementia, among others [33]. The pathophysiology of dementia involves complex changes in the brain, including the accumulation of abnormal protein deposits, neuronal death, and the disruption of synaptic connections, leading to a gradual loss of brain function [25]. As dementia progresses, individuals encounter increasing challenges in performing daily activities, necessitating the involvement of caregivers for basic needs. The global burden of dementia is considerable, affecting over 55 million people worldwide, with projections indicating a near doubling of cases every 20 years as populations age [21]. This presents significant challenges for healthcare systems, necessitating a multifaceted approach that includes early diagnosis, comprehensive care, and support for caregivers to effectively manage the disease.

2.2 Pain in Dementia Patients

Pain management in patients with dementia represents a particularly complex and critical issue, as cognitive decline severely impairs the ability to communicate discomfort, leading to widespread underdiagnosis and undertreatment of pain in this population. As dementia progresses, patients may lose the capacity to express pain verbally, relying instead on non-verbal cues such as facial expressions, vocalizations, or changes in behavior, which can be easily overlooked or misinterpreted by caregivers and healthcare professionals. The undertreatment of pain in patients with dementia not only exacerbates their suffering but also contributes to a decline in their overall quality of life. This manifests in increased agitation, depression, and a reduction in functional abilities. Furthermore, the pharmacological management of pain in dementia is further complicated by the potential adverse effects of analgesic medications, which may include increased cognitive impairment or interactions with other medications. Therefore, a careful and sophisticated approach to pain assessment and management is essential, one that integrates non-verbal pain assessment tools, individualized treatment plans, and a multidisciplinary team approach to ensure that the pain in dementia patients is adequately addressed.

2.3 Pain Assessment Metrics

The assessment and management of pain in clinical settings presents a significant challenge, particularly among populations with impaired communication abilities, such as patients with advanced dementia. The most prevalent approach to pain assessment is patient self-report, which entails individuals rating their pain levels using a range of scales. This method is convenient and does not necessitate the use of advanced technology or the acquisition of specialized skills, thereby ensuring its accessibility in the majority of healthcare settings. However, self-report measures have notable limitations. These include variability in the metric properties across different scale dimensions, susceptibility to suggestion, and differences in the conceptualization of pain between clinicians and patients. These challenges are further compounded in populations unable to communicate effectively, such as young children and those with severe cognitive impairments, where self-report becomes an ineffective methodology. Additionally, self-report measures provide only a snapshot of pain at a specific moment, often at the emotional apex of the patient, without offering continuous data on the patient's pain experience or emotional state over time [28]. To address these limitations, alternative pain assessment methods have been developed, including observer-based scales and facial expression analysis.

2.3.1 Visual Analog Scale

The Visual Analog Scale (VAS) is a frequently utilized unidimensional instrument that assesses pain intensity. The VAS can be presented in a variety of formats, including numerical rating scales, graphic rating scales, and box scales. It is a commonly utilized tool in clinical and epidemiological research, facilitating the tracking of pain progression and the comparison of pain severity across patients. The simplicity of the VAS, often represented as a straight horizontal line with endpoints defined as the extreme limits of the symptom being measured (Figure 2.1), makes it a versatile tool in pain assessment. However, as with self-report methods, the applicability of the VAS is limited by the patient's ability to comprehend and interact with the scale [34].

2.3.2 Facial Action Coding System

The Facial Action Coding System (FACS) is a comprehensive method for categorizing human facial movements by their visual appearance, providing a

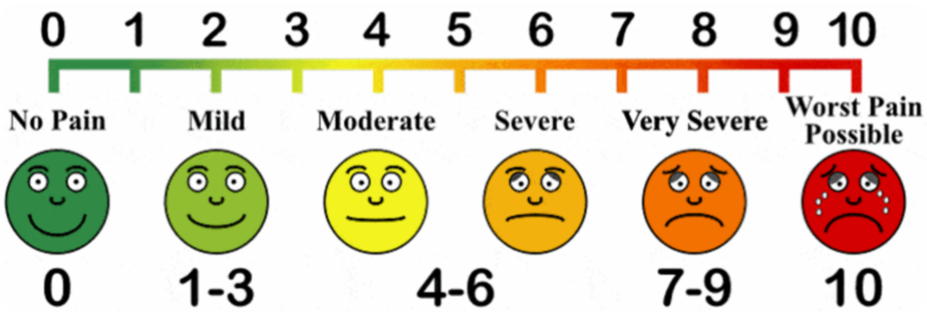
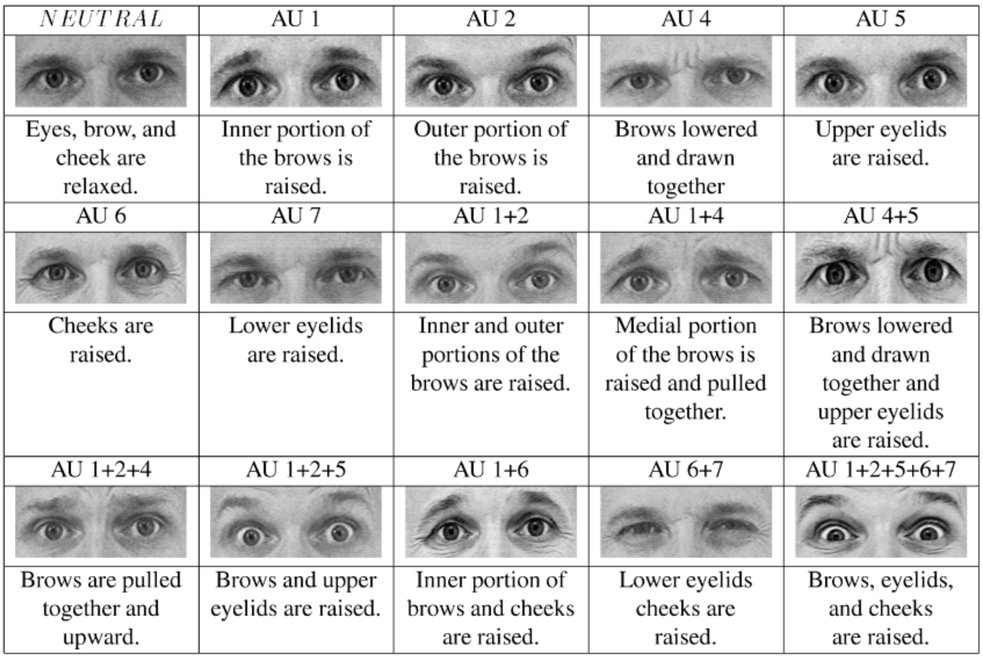


Figure 2.1: Visual Analog Scale

reliable framework to assess pain in individuals who are unable to communicate verbally. FACS identifies facial expressions through Action Units (AUs), which correspond to the contraction or relaxation of specific facial muscles. Out of the 44 defined AUs, 30 are directly linked to muscle contractions: 12 pertain to the upper face and 18 to the lower face. AUs can manifest either independently or in combination, where combinations can be additive or non-additive. Additive combinations do not alter the appearance of the individual AUs, while non-additive combinations result in a distinct appearance that differs from the constituent AUs. Despite the limited number of individual AUs, over 7,000 unique AU combinations have been documented.

Figure 2.2 illustrates some commonly occurring AUs along with examples of both additive and non-additive AU combinations. For instance, AU 4 (brow lowering) looks different when it occurs independently versus when it is paired with AU 1 (inner brow raiser) in the combination AU 1 4. When AU 4 occurs alone, the eyebrows are pulled together and lowered, but in AU 1 4, the eyebrows are drawn together yet raised due to the influence of AU 1. Another example is AU 1 2, in which AU 2 (outer brow raiser) not only lifts the outer brow but often elevates the inner brow as well, creating an appearance similar to AU 1 2. These non-additive effects complicate the recognition of AUs [41]. Although originally developed to analyze emotions, such as depression, FACS has demonstrated considerable potential in pain assessment,



NEUTRAL	AU 1	AU 2	AU 4	AU 5
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 2.2: Upper Face Action Units and Some Combinations

particularly in populations where traditional self-report methods are not feasible. Additionally, there is ongoing research focused on automating FACS coding through computer algorithms; however, progress is hindered by the scarcity of extensive, manually coded datasets that could serve as a benchmark for training these systems [44].

2.3.3 Prkachin and Solomon Pain Intensity

The Prkachin and Solomon Pain Intensity (PSPI) metric is a widely utilized tool for the assessment of pain intensity through the analysis of facial expressions. It provides a quantitative measure of pain by evaluating specific facial movements that are associated with discomfort [17]. The PSPI scale is defined as:

$$PSPI = AU_4 + Max(AU_6, AU_7) + Max(AU_9, AU_{10}) + AU_{43}$$

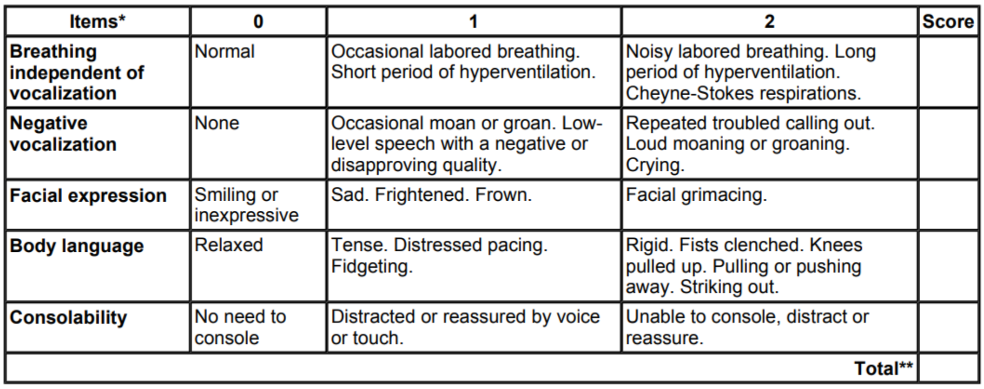
The PSPI has a theoretical range of 0 to 16, with higher scores indicating greater pain intensity. In practice, scores typically range from 0 to 12 [1].

2.3.4 Pain Assessment in Advanced Dementia

The Pain Assessment in Advanced Dementia (PAINAD) scale has been developed as a standardized method for assessing pain in patients with advanced dementia where communication barriers are significant. The PAINAD scale has been developed for the purpose of assessing pain in patients with advanced dementia who are unable to communicate verbally. It focuses on observable indicators of pain, such as breathing patterns, vocalization, facial expressions, body language, and comfort. The total score ranges from 0 to 10, with higher scores indicating more severe pain. This scale provides a practical solution for assessing pain in non-verbal patients, ensuring that their pain is identified and managed appropriately [43]. The PAINAD scale consists of five key observational items:

- Breathing: Normal, occasional labored, or noisy labored breathing.
- Negative vocalization: Ranging from none to loud moaning or crying.
- Facial expression: From smiling to sad or frightened expressions.
- Body language: Observing whether the individual appears relaxed or exhibits signs of distress such as fidgeting or rigidity.
- Consolability: Assessing whether the individual needs consolation or can be reassured by voice or touch.

As in Figure 2.3, each item is assigned a score from 0 to 2, and the total score is calculated by summing these individual scores, thereby providing a comprehensive assessment of pain levels in patients who are unable to verbally express their discomfort. However, some limitations have been identified,



Items*	0	1	2	Score
Breathing independent of vocalization	Normal	Occasional labored breathing. Short period of hyperventilation.	Noisy labored breathing. Long period of hyperventilation. Cheyne-Stokes respirations.	
Negative vocalization	None	Occasional moan or groan. Low-level speech with a negative or disapproving quality.	Repeated troubled calling out. Loud moaning or groaning. Crying.	
Facial expression	Smiling or inexpressive	Sad. Frightened. Frown.	Facial grimacing.	
Body language	Relaxed	Tense. Distressed pacing. Fidgeting.	Rigid. Fists clenched. Knees pulled up. Pulling or pushing away. Striking out.	
Consolability	No need to console	Distracted or reassured by voice or touch.	Unable to console, distract or reassure.	
			Total**	

Figure 2.3: Pain Assessment in Advanced Dementia (PAINAD) Scale [4]

including moderate internal consistency and variability in the use of certain items, particularly those related to breathing and consolability. Nurses have reported challenges in applying the scale due to its brevity and the potential for missing subtle pain cues in patients.

2.4 Performance Metrics

To comprehensively evaluate the model's performance, three distinct metrics were employed, each serving to assess different aspects of the model's output. Given the dual approach of the task, in which both regression and classification objectives were considered, the selected metrics provide a balanced evaluation framework that accounts for both continuous and categorical predictions.

2.4.1 Regression Metrics: MSE and MAE

For the regression task, Mean Squared Error (MSE) and Mean Absolute Error (MAE) were employed. These metrics are foundational in regression analysis, offering insights that are complementary to one another and that provide a comprehensive understanding of the model's predictive accuracy. MSE is defined as the average of the squared differences between the predicted values

($\hat{y}_i$) and the actual values (y_i). It is mathematically expressed as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

MSE is particularly sensitive to larger errors due to the squaring of differences, which amplifies the impact of outliers. This sensitivity can be advantageous during model training, as it drives the optimization process to prioritize the reduction of substantial errors. MAE, on the other hand, measures the average magnitude of errors in a set of predictions. MAE is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

In contrast to MSE, MAE treats all errors in a uniform manner by calculating the absolute difference between the predicted and actual values. This metric is more robust to outliers, as it does not disproportionately penalize larger errors. Consequently, MAE provides a more balanced view of the overall prediction error, making it a preferred choice when a more uniform error treatment is desired. The combination of MSE and MAE allows for a more comprehensive evaluation of the model's performance. MSE identifies major discrepancies, prompting the model to refine its predictions in areas where it is most inaccurate. In contrast, MAE provides a straightforward estimation of the average error, reflecting the model's overall prediction precision.

2.4.2 Classification Metrics: Accuracy, Precision, Recall, and F1-score

In order to evaluate the classification aspect of the task, the metric employed was accuracy. Accuracy is a straightforward yet effective metric, representing the proportion of correct predictions out of the total number of predictions made.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

While accuracy is particularly effective when the classes are balanced, providing a clear and intuitive measure of the model's overall effectiveness in categorizing data points correctly, its utility extends to certain multi-class scenarios as well. In a multi-class classification setting, accuracy remains a suitable metric as it reflects the proportion of correctly predicted instances across all classes. This indicates that if the model demonstrates consistent accuracy across multiple classes, accuracy can serve as a reliable measure of performance. However, it is important to note that accuracy alone may not always be an adequate metric for evaluating multi-class classification, particularly in cases where the classes are imbalanced or where some classes are more critical than others. In such cases, the model may still achieve high accuracy by performing well on the majority class while underperforming on the minority classes. This is where additional metrics such as precision, recall, and F1-score become essential for providing a more comprehensive evaluation. Precision measures the proportion of true positive predictions out of all positive predictions made by the model. Precision is especially important in situations where the cost of a false positive is high. A higher precision means fewer false alarms, which is critical in applications where incorrect positive predictions could lead to significant negative consequences. Precision is mathematically defined as:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions out of all actual positive instances. Recall is crucial when the cost of a false negative is high, such as in medical diagnostics where failing to identify a condition could have serious consequences. A high recall indicates that the model successfully captures most of the actual positive cases, minimizing the chances of missing critical positive instances.

Recall is defined as:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

F1-Score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. The F1-score is particularly useful in cases where there is an uneven class distribution or when there is a need to balance precision and recall. It is defined as:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1-score ranges from 0 to 1, with 1 indicating perfect precision and recall. This metric is ideal when both false positives and false negatives carry significant consequences, and neither precision nor recall can be optimized without adversely affecting the other. In practice, these metrics are often calculated for each class in a multi-class classification problem and can be averaged (using micro, macro, or weighted averaging) to provide an overall performance metric.

- Micro-Averaging: Aggregates the contributions of all classes to compute the average metric, providing an overall view that accounts for class imbalance.

- Macro-Averaging: Calculates the metric independently for each class and then averages them, treating all classes equally regardless of their frequency.

- Weighted-Averaging: Similar to macro-averaging but takes into account the support (i.e., the number of true instances) of each class, giving more weight to the performance on more frequent classes.

The choice of averaging method depends on the specific requirements of the application and the importance of each class.

2.5 Current Methods of Pain Assessment

Facial expressions are among the most reliable indicators of pain severity and are often used to convey a patient's level of pain. Given their effectiveness, the development of a system capable of accurately identifying pain intensity by extracting the most relevant facial cues is highly promising. Facial expressions can be derived from dynamic sources, such as the temporal dimensions of video, or from static images. This dissertation focuses specifically on the use of static facial expressions to assess pain levels. In recent years, with the rise of machine learning algorithms in predictive modeling, static facial expression-based methods have increasingly used these techniques to estimate pain intensity. These methods can be classified into three categories: traditional machine learning, deep learning, and hybrid model-based approaches [7].

2.5.1 Accessible Pain Assessment Datasets

In recent years, the availability of various facial image and video datasets has led to significant advances in the field of automated pain assessment. Among these, the UNBC McMaster Shoulder Pain Expression Archive Database [29] is one of the most widely used. This dataset was collected from 25 adult participants with shoulder pain and consists of 48,398 RGB frames from 200 variable-length videos. The images are primarily labeled with 17 levels of the PSPI scale, ranging from 0 to 16, and 11 levels of the VAS, ranging from 0 to 10. However, like many other image datasets, the UNBC McMaster dataset has a significant imbalance problem, with over 80% of the data labeled with a PSPI score of zero, indicating ”no pain”. This imbalance creates challenges for model training and can lead to biased predictions. To mitigate this, many studies have used under-sampling techniques to reduce the prevalence of the ”no pain” class, thereby creating a more balanced dataset for training pain assessment models.

In addition to these technical challenges, a growing concern in recent years has been the increasing unavailability of many facial datasets due to privacy and ethical concerns. As awareness of privacy and the ethical implications of using sensitive biometric data, such as facial expressions, has grown, several datasets have been withdrawn or restricted from public access. This trend has created significant challenges for researchers in the field, as reduced access to high-quality, diverse datasets limits the ability to effectively develop, validate, and benchmark new models.

2.5.2 Machine Learning-Based Methods

Traditional machine learning models, such as SVM and KNN, typically rely on the classification of pre-extracted, hand-crafted features. The effectiveness of these models is highly dependent on the quality and relevance of the extracted features, which often requires significant domain expertise. While machine learning-based methods have been widely used for the past fifteen years due to their promising results, their use in pain intensity estimation has declined with the advent of more advanced models, such as deep learning and hybrid approaches, which offer superior performance. In [37], a relatively flat CNN architecture with three convolutional layers was proposed. This computationally efficient network, with its minimal number of parameters, achieved an accuracy of 93.34% when evaluated on the UNBC McMaster dataset. In [48], a hierarchical network architecture using two per-frame feature modules was introduced to improve pain estimation. The first module extracts low-level features from image patches, which are then assembled using second-order pooling. The second module extracts deep learning features using a deep CNN. The outputs of these two modules are then weighted and combined to form a holistic face representation, which significantly improves the pain estimation process. This combined feature is then classified using a linear L2-regularized L2-loss SVR, resulting in an MSE of 1.45 on the UNBC McMaster dataset.

Transfer learning has also gained traction in image classification tasks, using pre-trained models for specific tasks to acquire transferable knowledge. This approach reduces training time and improves model performance. For example, in [13], a pre-trained DenseNet-161 model was fine-tuned on the UNBC McMaster dataset. Features were extracted from ten middle layers of the fine-tuned network and used as inputs to an SVR classifier. This model achieved an MSE of 0.34 on the UNBC-McMaster dataset. In addition, [6] proposed a KNN-based pain assessment method. In this approach, facial features are extracted from face patches using a pre-trained DarkNet19 model, followed by feature selection using the iterative neighborhood component analysis technique. The selected features are then classified using the KNN algorithm to efficiently predict pain intensities, achieving a pain intensity estimation accuracy of 95.57% on the UNBC McMaster dataset.

2.5.3 Deep Learning-Based Methods

Since 2018, deep learning models have demonstrated remarkable advancement in the field of automatic pain assessment, largely due to their robust performance in data classification and the increasing availability of extensive pain-related datasets. Among these models, CNNs have exhibited particularly success, with more sophisticated variants like ResNet, DenseNet, and InceptionV3 contributing significantly to this progress. In 2021, Semwal et al. [40] presented an enhanced version of their earlier work [37] by introducing SPANET, a compact and shallow CNN model specifically designed for pain severity assessment. The model incorporates a false positive reduction technique, achieving an MSE of 1.1 on the UNBC–McMaster dataset. In order to enhance the focus on pain-related facial regions, another study [47] proposed a nine-layer CNN incorporating an attention mechanism. This approach involved the weighting of different facial regions according to their expressiveness of pain. This innovation was further developed into a multi-task pain assessment architecture [46], designated as the Locality and Identity-Aware Network (LIAN). LIAN employs a dual-branch locality-aware module to prioritize pain-related facial information, followed by an Identity-Aware Module (IAM) that decouples pain assessment from identity recognition. This approach markedly enhanced the precision of pain detection, attaining an accuracy rate of 89.17% on the UNBC–McMaster dataset. Similarly, Cui et al. [9] proposed the Multi-Scale Regional Attention Network (MSRAN), a method that applies adaptive learning to accurately assess pain intensity by capturing detailed information from pain-specific facial regions. This model integrates self-attention and relation attention modules to enhance the understanding of facial pain expressions and their interrelationships, resulting in an accuracy of 91.13% on the UNBC–McMaster dataset. Building on these advances, another approach [23] involved modifying the VGG16 architecture to create a deeper, customized CNN model. This model, designed specifically for pain intensity estimation, applied rigorous pre-processing techniques to the input images, including gray-scaling, histogram equalization, face detection, image cropping, mean filtering, and normalization. The modified VGG16 model achieved an accuracy of 92.5% on the UNBC–McMaster dataset, which highlights the significance of image pre-processing in enhancing model performance. The most recent and notable advancement in this domain [2] employed two concurrent CNNs based on the InceptionV3 architecture, optimized using stochastic gradient descent. In this model, all convolutional blocks were frozen, and the classifier layer was replaced with a shallow CNN. The outputs from these models were then concatenated and passed through a dense layer, followed by a fully connected layer. This resulted in an unprecedented pain intensity estimation accuracy of 99.1% on the UNBC–McMaster dataset.

2.5.4 Hybrid Methods

The success of machine learning and deep learning models in automatic pain assessment has prompted the development of a variety of ensemble learning methods, which aim to combine the strengths of multiple models. Ensemble techniques have demonstrated efficacy in enhancing classification accuracy by leveraging the complementary capabilities of different models. In [38], Semwal and Londhe introduced the ECCNET model, which integrates three distinct CNNs (VGG-16, MobileNet, and GoogleNet) using an average ensemble rule for aggregating their predictions. The results of their experiments demonstrated that the combination of these CNNs resulted in superior classification performance compared to the use of each network individually. The ECCNET model achieved a notable accuracy of 93.87% on the UNBC–McMaster dataset, thereby demonstrating the effectiveness of the ensemble approach. In a subsequent study, the same authors [39] further refined this model by reapplying the CNN fusion technique, this time incorporating advanced data augmentation strategies to address overfitting. This refinement resulted in a 2.13% improvement in pain level detection accuracy, bringing the overall accuracy to 96% on the same dataset. Another noteworthy approach is the EDLM, which was proposed in [5]. The model begins with the use of a fine-tuned VGGFace network to extract facial features. This is followed by the application of PCA, which serves to reduce feature dimensionality while preserving the most informative features. The dimensionality reduction not only helps to minimize the risk of overfitting but also decreases the training time, thus making the model more efficient. Subsequently, the reduced features are classified using three independent CNN-RNN deep learning models, each with different weights. Despite the computational efficiency achieved through PCA, the EDLM model attained a respectable accuracy of 86% on the UNBC–McMaster dataset. Yeten et al. [49] proposed a parallel

CNN framework that incorporates regional attention to focus on the most pain-sensitive regions of the face. The method employs a combination of VGGNet and ResNet architectures to extract detailed facial features, which are then classified using a SoftMax classifier. By focusing on key facial regions that are most indicative of pain, the model achieved a high accuracy of 95.11% on the UNBC–McMaster dataset, underscoring the importance of targeted feature extraction in pain assessment. In a recent contribution to the field, Sait and Dutta [42] proposed a sophisticated ensemble learning approach that leverages a fine-tuned ShuffleNet V2 model for feature extraction. They employed class activation maps and fusion feature techniques to enhance the model's ability to capture relevant facial features. Subsequently, a stacking ensemble strategy was employed, in which XGBoost and CatBoost were utilized as base models, and a Support Vector Machine (SVM) was employed as a meta-learner to predict pain intensities. This ensemble approach resulted in an accuracy of 98.7% on the UNBC–McMaster dataset, indicating the robustness and potential real-world applicability of the model in healthcare settings.

Chapter 3

Theoretical Background

3.1 Introduction

In recent years, the deep learning (DL) paradigm has emerged as the gold standard within the machine learning (ML) community. It has rapidly become the most widely adopted computational approach in the field, delivering remarkable performance on a range of complex cognitive tasks, often matching or even surpassing human capabilities [3]. A key factor in the effectiveness of any ML algorithm is the quality of the representation of the input data. Research has consistently shown that well-represented data significantly improves algorithm performance compared to poorly represented data. As a result, feature engineering - which focuses on constructing meaningful features from raw data - has been a major research trend in ML for many years. This approach is typically domain-specific and often requires considerable human expertise and effort. For example, in computer vision, various feature extraction techniques such as Histogram of Oriented Gradients (HOG) [11], Scale-Invariant Feature Transform (SIFT) [27], and Bag of Words (BoW) [45] have been developed and extensively researched. Whenever a new feature extraction method proves to be effective, it often opens a new research direction that can last for decades. Deep learning algorithms, particularly convolutional neural networks (CNNs), have revolutionized feature extraction by automating the process and reducing the need for human intervention and specialized domain knowledge. These algorithms utilize a multi-layer architecture where the initial layers capture low-level features and the deeper layers extract high-level, more abstract features. This hierarchical approach mirrors the way the human brain processes information and underscores the primary advantage of DL. Among the various DL models, CNNs have become particularly popular and widely used. Their ability to automatically identify and prioritize significant features without human supervision distinguishes them from previous models, making CNNs a key component of modern DL applications.

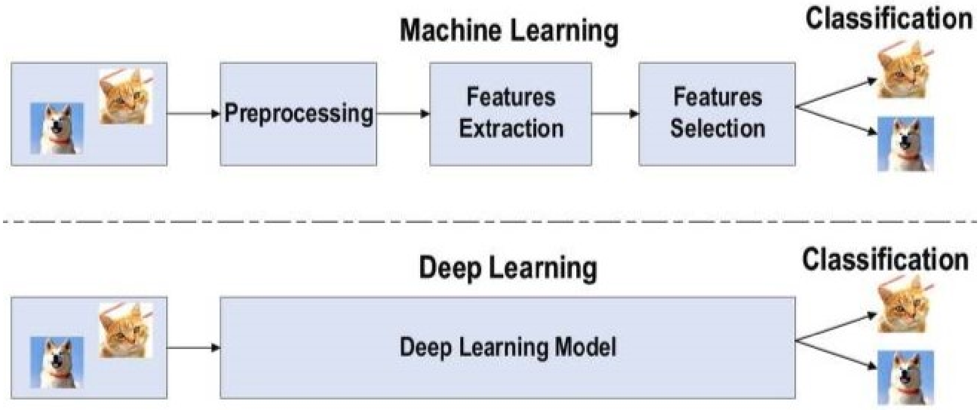


Figure 3.1: Machine Learning vs Deep Learning

3.2 Deep Learning

Deep learning, a specialized subset of machine learning, draws inspiration from the information processing patterns observed in the human brain. In contrast to conventional ML methodologies, DL does not necessitate the formulation of human-designed rules or the handcrafting of features. Instead, it employs vast quantities of data to facilitate the automatic mapping of inputs to designated outputs or labels (Figure 3.1). The fundamental structure of DL comprises multiple layers of artificial neural networks, wherein each layer progressively extracts increasingly abstract and complex features from the input data. In the context of conventional ML, the accomplishment of tasks such as classification necessitates the completion of a series of sequential steps, including pre-processing, feature extraction, feature selection, learning, and classification. The efficacy of these conventional techniques is contingent upon the quality of the selected features. Inadequate or biased feature selection can result in erroneous classification outcomes, as it may fail to capture the true discriminative characteristics between classes. In contrast, DL is particularly effective in automating the process of feature learning, thereby reducing the necessity for manual feature engineering. The capacity to learn feature representations directly from raw data in a single integrated process has contributed to the popularity of DL, particularly in the context of big data. Among the various deep learning models, recurrent neural networks (RNNs) and CNNs are the most well-known and widely used. The following section provides an overview of RNNs, while CNNs are discussed in more detail due to their critical importance to the research project at hand and their broad applicability across multiple domains.

3.2.1 Recurrent Neural Networks

RNNs are predominantly applied in the fields of speech processing and natural language processing (NLP). Unlike conventional feedforward neural networks, RNNs are designed to handle sequential data by maintaining a memory of previous inputs within the network. This sequential data processing is crucial in applications where the context of information is of particular importance. For instance, in language modeling, understanding the context of a sentence is essential to determine the meaning of a specific word. The structure of an RNN can be compared to a short-term memory unit, where x represents the input layer, y is the output layer, and h denotes the hidden state layer. Figure 3.2 depicts a typical unfolded RNN diagram for a given input sequence.

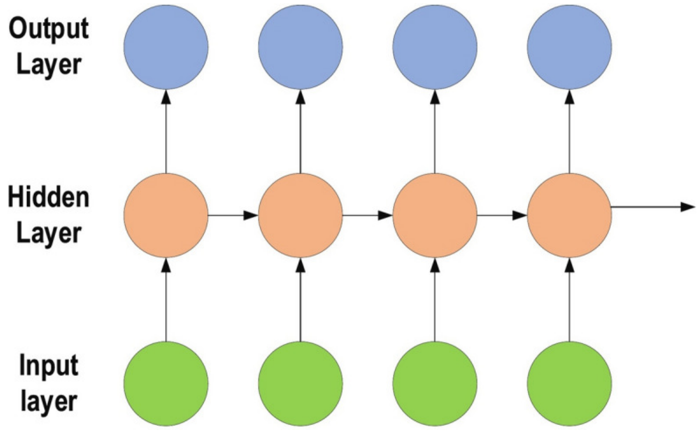


Figure 3.2: Recurrent Neural Network

However, RNNs have their challenges. One of the primary issues associated with RNNs is their sensitivity to the exploding and vanishing gradient problems [16]. During the training process, repeated multiplication of large or small gradients can cause the gradients to either explode or vanish, respectively, leading to instability or inefficiency in learning. As new inputs are fed into the network, earlier inputs can be ”forgotten,” causing the network's performance to degrade over time. This issue can be mitigated through the utilization of long short-term memory (LSTM) networks [15], a variant of RNNs. LSTMs incorporate memory blocks that maintain information over longer periods, addressing the vanishing gradient problem by allowing the network to retain important information for extended durations. Each memory block in an LSTM contains memory cells capable of storing temporal states, along with gated units that regulate the flow of information in and out of the cells. In networks of considerable depth, residual connections can also assist in mitigating the impact of vanishing gradients, enabling the network to maintain more stable and effective learning [19].

3.2.2 Convolutional Neural Networks

CNNs are among the most prominent and widely utilized architecture in the field of deep learning [24]. One of the key advantages of CNNs over their predecessors is their ability to automatically identify and learn relevant features from input data without the need for human intervention. CNNs have been successfully applied across various domains, including computer vision, speech processing, and face recognition [26], making them an indispensable tool in modern artificial intelligence. In contrast to conventional fully connected networks, CNNs leverage shared weights and local connections to efficiently process structured 2D input data, such as images. This approach considerably reduces the number of parameters required, thus simplifying the training process and enhancing the network's computational efficiency. A typical CNN architecture includes multiple convolutional layers followed by sub-sampling (pooling) layers, with FC layers at the end of the network. In a CNN model, the input x to each layer is organized in three dimensions: height, width, and depth. This is often denoted as $m \times m \times d$, where m represents the height, which is equal to the width, and d corresponds to the depth. For instance, in an RGB image, the depth d is three, corresponding to the red, green, and blue channels. Each convolutional layer contains several filters, or kernels, denoted by k, which are also three-dimensional ($n \times n \times q$), where n is the kernel size and q is the depth. It is necessary to note that, n is smaller than m, and q is equal to or less than d. These kernels perform local connections by convolving with the input data to produce feature maps h_k of size $(m - n + 1)$. The shared parameters, including the bias b_k and the weight W_k, are used across the input data, enabling the generation of these feature

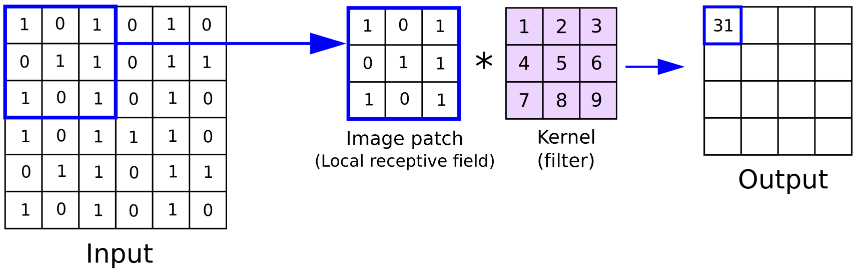


Figure 3.3: Convolutional Layer

maps. The output of the convolutional layer is then passed through a non-linear activation function, typically a rectified linear unit (ReLU), with the objective of introducing non-linearity into the model. This can be expressed as follows:

$$h_k = f(W_k * x + b_k)$$

Here, $*$ denotes the convolution operation, and f is the activation function. Following the convolutional layers, sub-sampling (pooling) layers are applied to each feature map. Pooling reduces the dimensionality of the feature maps, which has the dual benefit of accelerating the training process and mitigating the risk of overfitting by reducing the network's complexity. The pooling function, such as max-pooling or average-pooling, is typically applied over a region of size $p \times p$, where p is the kernel size used in pooling. This down-sampling process retains the most salient features while discarding less important information. Finally, the FC layers process the output from the convolutional and pooling layers to create high-level abstractions. These layers serve as the final decision-making component of the network, combining the mid- and low-level features learned earlier in the network to produce the final output.

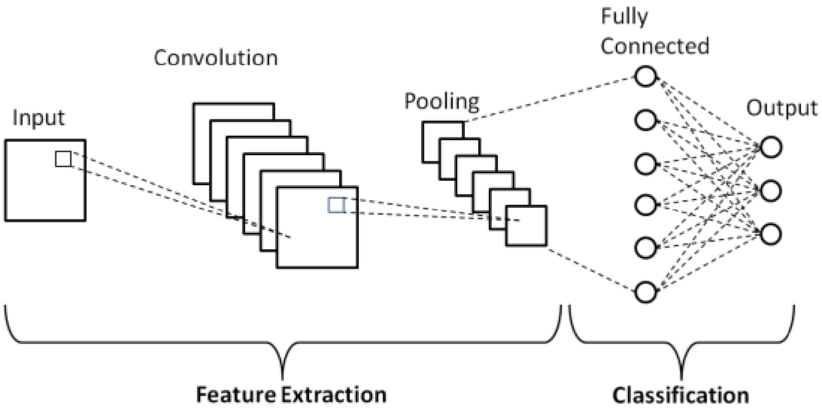


Figure 3.4: Convolutional Neural Network

The advantages of using CNNs over traditional neural networks, particularly in the domain of computer vision, include:

- Weight Sharing: CNNs employ weight sharing in their convolutional layers, significantly reducing the number of trainable parameters. This reduction simplifies the model and enhances its ability to generalize, reducing the likelihood of overfitting.

- Integrated Feature Learning and Classification: CNNs concurrently learn feature extraction and classification within a unified framework. This integration ensures that the learned features are highly relevant to the task at hand, leading to more organized and accurate model outputs.

- Scalability: CNNs are highly scalable and can be effectively implemented in large-scale applications, making them more practical and efficient compared to other neural network architectures.

3.2.3 Vision Transformers

Vision Transformers (ViTs) represent a significant advancement in computer vision. First introduced in 2020 as an extension of the transformer model initially designed for NLP in 2017, ViTs have become a prominent approach in the field of computer vision [22]. By 2021, ViTs had begun to outperform CNNs in terms of both performance and efficiency, particularly in image classification tasks [12]. The key innovation of ViTs lies in their attention-based mechanism, which enables them to capture complex patterns in images more effectively than traditional convolution-based architectures. This ability has positioned ViTs as a powerful alternative in the field of deep learning applied to vision, offering new perspectives and approaches in image classification. The Vision Transformer operates by dividing an input image into fixed-size patches, which are then linearly embedded. In order to retain spatial information, positional embeddings are added to these patches. The resulting sequence of vectors is then fed into a standard transformer encoder, as illustrated in Figure 3.5. The main Vision Transformer steps for image classification are

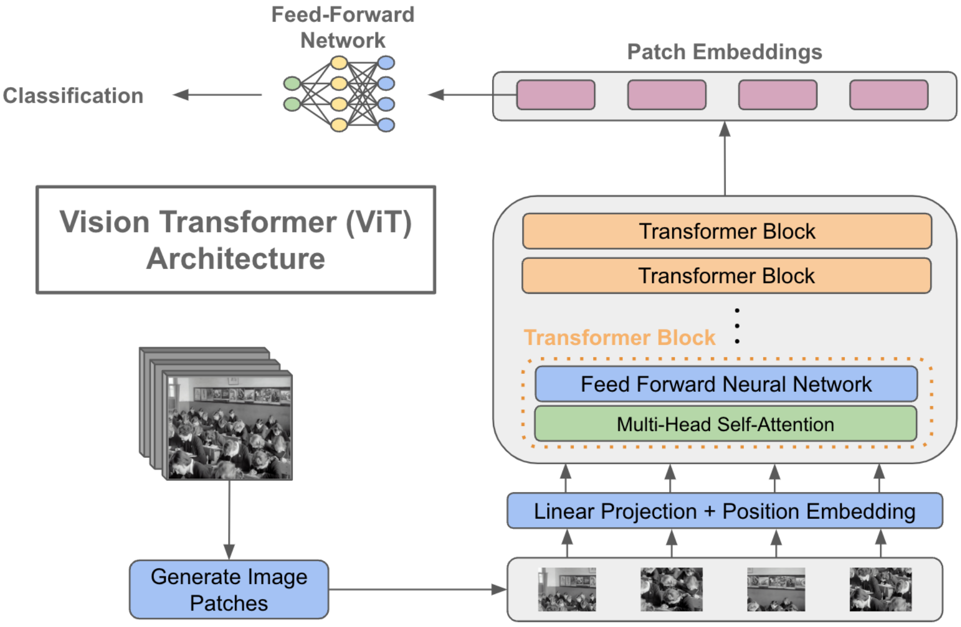

Feed-Forward Network
Patch Embeddings
Classification
Vision Transformer (ViT) Architecture
Transformer Block
Transformer Block
Transformer Block
Feed Forward Neural Network
Multi-Head Self-Attention
Linear Projection + Position Embedding
Generate Image Patches

Figure 3.5: Vision Transformer

:

- Image Patching: The image of size $(H \times W \times C)$ is divided into N patches of size $(P \times P \times C)$, where H is the height, W is the width, and C is the number of channels of the image. P is the resolution of the image patch.

- Linear Transformation of Patches to Vectors: Each patch is flattened into a vector of size $(1 \times P\check{s} * C)$. This linear transformation converts the patches into a format suitable for processing by the model.

- Adding Position Tokens: To maintain the positional information of patches, positional embeddings are added to the patch vectors. Additionally, a special classification token (CLS) is prepended to the sequence. The combined positional embeddings and patch vectors are then fed into the Transformer Encoder.

- Encoder Layer: The transformer encoder consists of alternating layers of multi-head self-attention (MSA) and multilayer perceptron (MLP) blocks. Before each self-attention and MLP block, layer normalization is applied, and residual connections are used post-block to facilitate effective learning.

- Classification Layer: A classification head is constructed using an MLP with a hidden layer for feature extraction and a final linear layer for classification. This layer processes the final representation output by the encoder to generate class scores for image classification.

3.2.4 Hybrid Models

Hybrid architectures have recently emerged as a promising approach to address the limitations inherent in both CNNs and ViTs. CNNs, while effective at capturing local features and inductive biases, struggle with capturing long-range dependencies and require a fixed input size. Conversely, ViTs excel at modeling global context but may suffer from weak local feature extraction, sensitivity to noise, and high memory consumption [10]. The integration of the strengths of CNNs and ViTs in hybrid models has been demonstrated to enhance performance in image classification tasks. These architectures are designed to combine the local feature extraction capabilities of CNNs with the global attention mechanisms of ViTs, creating a more balanced and robust model. Researchers have explored various methods to combine these architectures, including parallel and sequential integration approaches. Parallel approaches may involve processing the same input through both a CNN and a

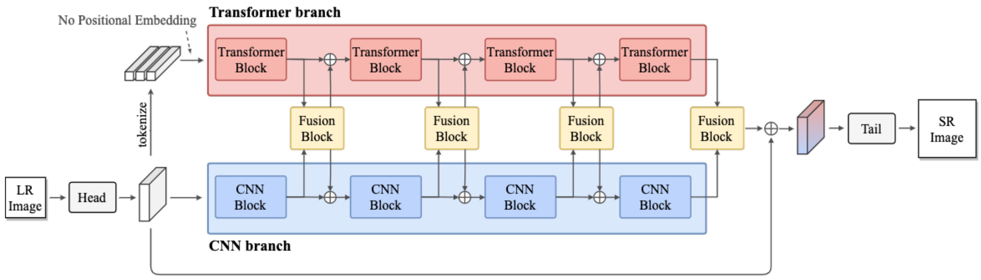


Figure 3.6: Parallel integration of CNNs and ViTs

ViT in tandem, then combining the outputs, while sequential approaches might use a CNN to extract initial features, followed by a ViT to model long-range dependencies. By addressing the limitations of each architecture through hybridization, these models can achieve superior performance and generalization in complex computer vision tasks. Ongoing research continues to refine these hybrid strategies, optimizing the synergy between CNNs and ViTs for various applications.

3.3 Model Interpretability

The term ”model interpretability” is used to describe the capacity to comprehend and explain the internal operations and decision-making processes of a machine learning model. It is a pivotal element in the implementation of AI systems, particularly in domains where trust and accountability are imperative, such as healthcare, finance, and law. High interpretability enables stakeholders to trace decisions back to specific input features, facilitating the identification of potential biases, errors, or unintended consequences in the model’s logic. Furthermore, it smooths the way for debugging, model refinement, and ensures that the model’s behavior aligns with human values and legal standards. However, there is often a trade-off between interpretability and model complexity, as more complex models like deep neural networks tend to be less interpretable despite their higher predictive performance. Various methods have been developed to enhance interpretability, including model-specific approaches, which are tailored to particular algorithms, and model-agnostic techniques, such as Occlusion, which will be covered in the next section.

3.3.1 Occlusion

Occlusion is a technique used to interpret and understand machine learning models, particularly deep learning models like convolutional CNNs. The method involves systematically masking or occluding parts of the input data (such as sections of an image) to observe the resulting alterations in the model's predictions. By analyzing these changes, one can determine which parts of the input are most important for the model's decision-making process [50]. The key steps are:

1. Input Data Selection: The initial stage of the process is the selection of the data set to be analyzed. This is typically an image, although the occlusion technique can also be applied to other data types, such as text or tabular data.

2. Masking Parts of the Input: The input is systematically occluded by covering specific regions with a neutral value, such as blacking out sections of an image or replacing words in a text with a placeholder. The size and shape of the occlusion mask may vary depending on the task at hand. For instance, in image classification, a square patch may be employed.

3. Prediction with Occluded Input: The occluded input is then fed into the model to generate a prediction. This process is repeated multiple times, with different parts of the input occluded each time.

4. Analyzing Prediction Changes: The model's output is compared across different occlusions to identify which areas of the input, when occluded, cause the most significant change in the model's prediction. If the model's confidence in its prediction drops significantly when a particular part of the input is occluded, that part is likely crucial for the model's decision.

Occlusion is a valuable technique for model interpretability, as it helps to reveal the inner workings of ”black-box” models by showing which features are most influential. This technique can be particularly useful in fields like healthcare, where understanding which parts of a medical image are driving a diagnosis is crucial for trust and decision-making. Furthermore, it aids in identifying potential biases or flaws in models by revealing unexpected areas of focus.

3.4 Ethical Considerations

The incorporation of ML into the field of healthcare has the potential to transform medical diagnostics, treatment planning, and patient care. Nevertheless, the implementation of these technologies also gives rise to considerable ethical concerns that must be carefully addressed in order to guarantee that the benefits of ML are fully realized without compromising patient rights, safety, or trust. One of the most critical ethical issues is the potential for bias in algorithms. If the training data used to develop machine learning models is unrepresentative or reflects existing societal biases, the models may produce biased outcomes. This can result in disparities where certain groups, such as racial minorities or disadvantaged individuals, may receive less accurate diagnoses or suboptimal treatment recommendations. To ensure fairness, rigorous testing of models across diverse populations and the implementation of strategies to mitigate bias, such as the use of balanced datasets and the incorporation of fairness constraints in the model development process, are essential. Accountability and transparency are also important. Clinicians and patients alike must be able to understand the reasoning behind a model's predictions or recommendations to trust and effectively use them. Black-box models, which offer high accuracy but low interpretability, pose significant ethical challenges. Efforts must be made to either enhance the interpretability of these models or develop methods, such as model-agnostic interpretability techniques, that provide understandable explanations for their decisions. A further significant issue is that of patient privacy. The use of ML frequently entails the utilization of extensive datasets comprising sensitive patient information. It is therefore of primary importance to guarantee the confidentiality and security of this data. There is a risk that patient data could be revealed through data breaches or inappropriate use, which could result in potential harm, discrimination, or a loss of trust in healthcare providers. Ethical ML practices in healthcare must encompass the implementation of robust data protection measures and techniques such as data anonymization and differential privacy to safeguard patient information. Moreover, patients must be informed of the manner in which their data is being utilized, including the potential risks and benefits associated with this process. In many cases, data may be repurposed for research or model development without the explicit consent of the data subjects, which can raise ethical concerns. It is of the utmost importance that communication is transparent and that informed consent is obtained. This ensures that patients are able to opt in or out of having their data used in ML applications. Furthermore, in instances where an ML model issues an erroneous diagnosis or treatment recommendation that ultimately causes harm to a patient, determining the responsible actor can prove to be a challenging task. It is essential to establish clear guidelines and frameworks for accountability in order to address these issues, including defining the role of clinicians in overseeing and validating model outputs before they are acted upon. Finally, the implementation of these applications in healthcare may result in a transformation of the traditional patient-provider relationship. While machine learning has the potential to enhance decision-making by providing data-driven insights, there is a risk that it may depersonalize care if clinicians place excessive reliance on algorithms. It is necessary to consider ethical implications that ensure the maintenance of the human element in healthcare, ensuring that technology serves as a tool to support, rather than replace, the clinician-patient interaction.

Chapter 4

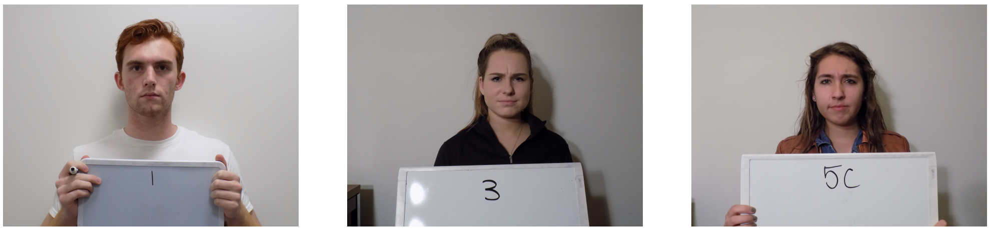
Methodologies

4.1 Dataset

The dataset utilized for training a machine learning model is of significant importance with regard to the model's overall success and generalizability. A well-balanced and diverse dataset ensures that the model learns to recognize patterns that are representative of real-world scenarios, reducing the risk of overfitting. Overfitting occurs when a model demonstrates excellent performance on the training data but exhibits poor generalization to new, unseen data. This is often due to the limited diversity or skewed distribution of outcomes in the dataset. By incorporating a diverse and balanced dataset, the model is exposed to a wide range of examples, allowing it to develop robust predictive capabilities and perform more accurately across different situations. This highlights the crucial role of the dataset in the model development process. The *UNBC-McMaster Shoulder Pain Expression Archive* [29] has served as the primary data set in the field of pain recognition and expression analysis, due to its comprehensive FACS encodings and VAS scores as labels, which make it a valuable resource for researchers. These features have made the dataset highly versatile and explainable, thereby facilitating the development of numerous DL models. However, the recent **removal of this dataset from public availability** due to privacy concerns has posed a significant challenge for researchers. The unavailability of this critical resource underscores a considerable challenge in the field: the absence of alternative datasets with sufficient size and diversity to train deep learning models from scratch. The limited access to large, high-quality datasets like UNBC-McMaster impedes the development of robust and accurate models, thus constraining progress and innovation in pain recognition research.

4.1.1 Delaware Pain Database

The dataset used for training the model is the *Delaware Pain Database* [31], a well-characterized and diverse collection of facial expression images, specifically designed to capture both painful and neutral expressions. The dataset comprises photographs of 127 female and 113 male subjects, ensuring balanced gender representation. The *Delaware Pain Database* comprises seven discrete pain levels, ranging from 0 (no pain) to 6 (maximum pain), derived from 240 individual subjects. The dataset is publicly accessible for research purposes and it is important to note that all expressions are simulated, meaning that the subjects were not actually experiencing pain when the images were captured. This presents a significant challenge in model development, as the artificial nature of these expressions may introduce ambiguity. The lack of genuine pain responses could limit the model's ability to accurately generalize to real-world scenarios, where the subtle nuances of true pain might differ from those depicted in simulations. Therefore it is essential to address this challenge in order to ensure the model's effectiveness in practical applications. Figure 4.1 depicts three subjects exhibiting varying degrees of pain, exemplifying a pivotal challenge inherent to this research: the intrinsic subjectivity of pain perception. Upon examination of the expressions, it becomes evident that



(a) Male, Pain Level 1 (b) Female, Pain Level 3 (c) Female, Pain Level 5

Figure 4.1: Figures form Delaware Pain Database

distinguishing between the various pain levels is challenging, even for human observers. For example, despite the assertion that Figure 4.1a and Figure 4.1c are separated by four levels of pain intensity, the visual distinction between these expressions is limited. This ambiguity, evident even to the human eye, represents a significant issue for the DL model. If humans, with their nuanced perception of facial expressions, struggle to differentiate these levels of pain, it poses an even greater challenge for a DL model to accurately interpret and classify such minor differences. This underscores the importance of addressing these ambiguities in the model's design and data preprocessing, which will be the focus of the next section.

4.2 Data Preprocessing

Data preprocessing represents a crucial phase in machine learning, as it enables the input data to be prepared in a manner that enhances the performance and reliability of the model. In this case, the initial step is to resize the input images to a uniform dimension (256x256 pixels) via bicubic interpolation. This guarantees that all images possess identical dimensions, which is essential for feeding them into a neural network, which requires consistent input sizes. A center crop of 224x224 pixels is then applied, focusing the model's attention on the most salient image regions. This approach is particularly effective for datasets like Delaware, where patient faces are typically centered. To enhance the model's generalization capabilities, data augmentation techniques are employed. Random horizontal flipping with a 50% probability is implemented, which helps the model learn invariance to horizontal orientation. This augmentation strategy can significantly improve the model's ability to recognize features regardless of their left-right orientation in the image. Pixel value normalization follows, using specific mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225) values. This normalization step ensures that input data is on a comparable scale, allowing for faster model convergence and potentially improving overall performance. To address the common issue of class imbalance in medical datasets, Synthetic Minority Oversampling Technique (SMOTE) is applied. SMOTE is an oversampling method that creates synthetic examples of the minority class. The method works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line. Specifically, SMOTE generates new instances of the minority class by interpolating between existing minority instances that are close together. This technique effectively balances the dataset, preventing the model from being biased towards the majority class and improving its ability to learn from underrepresented classes. By combining these preprocessing steps (image resizing, center cropping, data augmentation through horizontal flipping, pixel normalization, and SMOTE for class balancing) a robust foundation for training the machine learning model is created. This comprehensive approach standardizes the input data and enriches it, ultimately leading to a more effective and generalizable model.

4.3 Model Architecture and Design

The model architecture is composed of two primary components: a feature extraction stage and a subsequent neural network for classification and regression tasks.

4.3.1 Feature extraction

The feature extraction stage leverages a pre-trained model to obtain rich facial features from input images. The following models have been used.

MediaPipe Face Landmarker

MediaPipe Face Landmarker is an advanced real-time face landmark detection solution capable of identifying 468 3D landmarks on human faces. This model utilizes a multi-layer architecture consisting of three key components: a face detection model, a face mesh model, and a blend shape prediction model.

1. Face Detection Model: The initial stage employs the BlazeFace model, which is capable of detecting the presence of faces in images and provides basic facial landmarks, thus establishing a region of interest for subsequent processing.

2. Face Mesh Model: Once a face is detected, the face mesh model takes over, estimating a comprehensive mapping of the face's geometry as in Figure 4.2. This model outputs 478 3D landmarks, capturing intricate facial details and contours. The system operates efficiently under a variety of conditions, ensuring accurate landmark detection even in real-time scenarios.

3. Blendshape Prediction Model: The final component analyzes the output from the face mesh model to predict 52 blendshape scores, which represent different facial expressions. These scores allow for the interpretation of emotional states and facilitate the application of facial filters and effects in augmented reality environments.

The entire pipeline is optimized for performance, utilizing GPU acceleration to maintain high processing speeds, essential for applications such as virtual

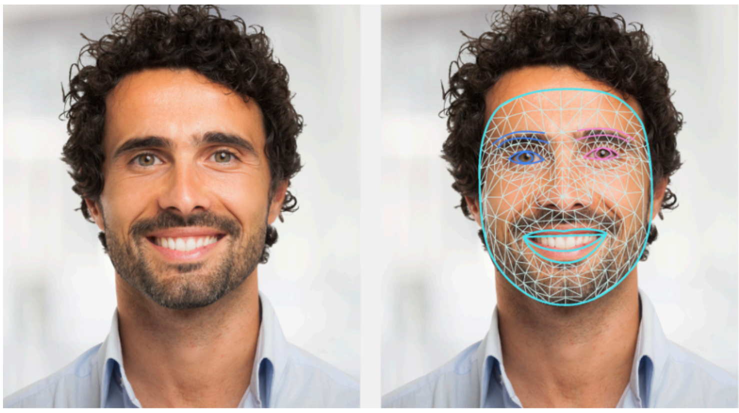


Figure 4.2: Face Mesh

avatars and interactive experiences. Additionally, the architectural design allows for the input of diverse data formats, including single images, video frames, and live streams, which makes it versatile for different use cases. It is worth noting that this approach is particularly valuable for its ability to generate interpretable face blend shapes, which makes it highly suitable for applications in healthcare where model transparency is crucial.

MobileNetV3

MobileNet V3 [20] is an advanced DL model that has been specifically designed for efficient deployment in mobile and embedded vision applications. The architectural design is based on the inverted residual structure introduced in MobileNet V2, where each block incorporates a bottleneck layer to reduce the dimensionality of the data, followed by depthwise separable convolutions that significantly reduce the number of parameters and the computational cost (Figure 4.3). MobileNet V3 incorporates the Squeeze-and-Excite (SE) module, which adaptively recalibrates channel-wise feature responses, enhancing the network's ability to capture salient features. Moreover, MobileNet V3 utilizes a Neural Architecture Search (NAS) technique to achieve an optimal balance between latency and accuracy. This results in two main variants: MobileNet V3 Small, optimized for lower latency and smaller models, and MobileNet V3 Large, which provides higher accuracy for more demanding tasks.

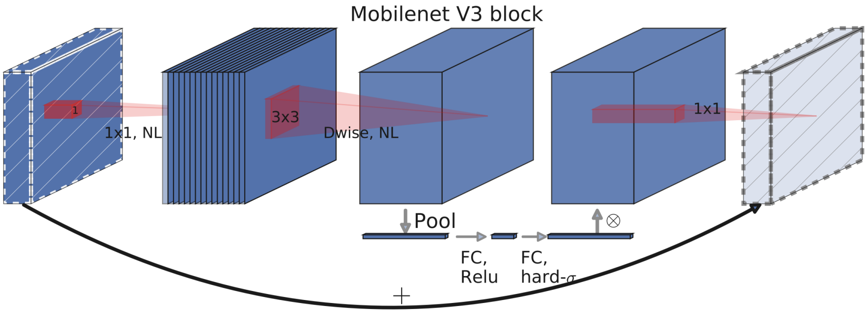


Figure 4.3: MobileNet V3 Architecture

The final layer includes a global average pooling, a fully connected layer, and a softmax output, completing the model's architecture. This is both compact and powerful, making it ideal for deployment in resource-constrained environments.

FaceNet

FaceNet [36] is a pioneering facial recognition system developed by researchers at Google and first introduced in 2015. A deep convolutional neural network is employed to map facial images into a compact 128-dimensional Euclidean space, wherein the distance between points is directly correlated with facial similarity. This innovative approach enables the efficient execution of face recognition, verification, and clustering tasks through the utilization of a distinctive triplet loss function during the training phase (Figure 4.4). The triplet loss function encourages the model to minimize the distance between the embeddings of the same identity, while simultaneously maximizing the distance between those of different identities.

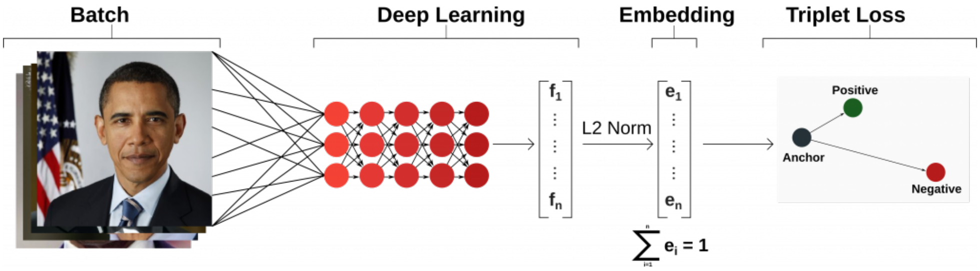


Figure 4.4: FaceNet High Level Architecture

The architecture of FaceNet is noteworthy for its capacity to generate high-quality face embeddings, which can be utilized in a multitude of applications, including security systems and social media platforms. The model attained an exceptional degree of accuracy, establishing a record of 99.63% on the Labeled Faces in the Wild dataset, markedly exceeding the performance of previous methods and establishing a new benchmark in the domain of face recognition.

DINO V2

DIstillation of knowledge with No labels and vIsion transformers (DINO) V2 [32] is a state-of-the-art self-supervised learning model designed for vision tasks, building upon the success of its predecessor, DINO. Meta AI developed DINO V2, which demonstrates excellence in visual data comprehension without the necessity of labeled datasets, making it highly versatile across diverse applications. A significant advancement in DINO V2 is its capacity to generate high-quality, semantically rich features that are useful not only for classification but also for a range of other tasks, including segmentation, object detection, and even fine-grained image retrieval. This version represents a notable improvement in the scalability and generalization capabilities of self-supervised vision models, enabling DINO V2 to perform well across diverse tasks without requiring additional fine-tuning. One of the standout features

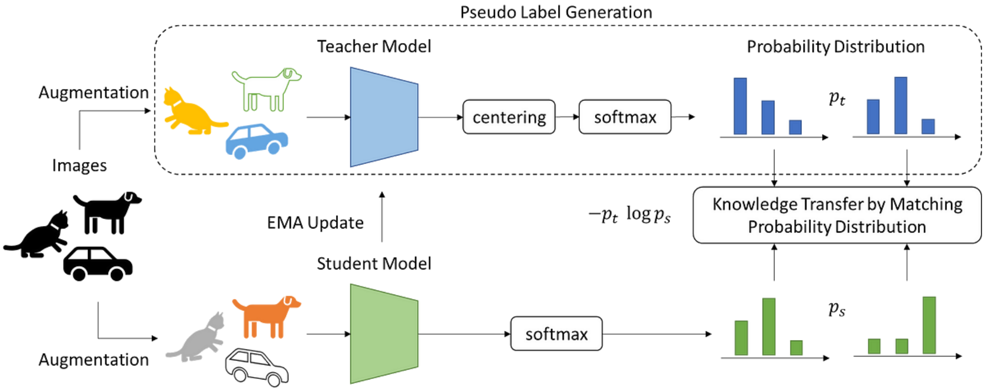


Figure 4.5: DINO V2 Training Process

of DINO V2 is its use of self-distillation during the training phase. This approach enables the model to learn to predict its own outputs as it processes data. Specifically, DINO V2 employs a teacher-student setup, where two versions of the model (a teacher and a student) are maintained during training (Figure 4.5). The teacher model is updated at a slower rate (through an exponential moving average of the student weights), and it generates ”soft” targets or predictions for the student to learn from. This method enables the model to develop high-quality, stable representations without requiring labeled data. The training process of DINO V2 focuses on optimizing these self-distillation objectives while using techniques like multi-crop augmentation, where multiple views of the same image at different scales are processed simultaneously, encouraging the model to learn invariant features across these views. This, combined with the inherent strength of ViTs in capturing contextual and semantic information, allows DINO V2 to achieve robust and generalized visual representations. The result is a model that demonstrates excellence across a diverse range of vision tasks, exhibiting robust performance even on datasets and tasks that were not the explicit focus of the training process.

4.3.2 Classification and Regression

After the features have been extracted, they are presented as input into a neural network that functions as classification or regression. Both methods have been explored as they can have different levels of explainability and different loss functions can be applied. The flexibility to experiment with both methods allows for a comprehensive comparison of performance, enabling the selection of the most appropriate approach based on the specific requirements of the application.

Adaptation of Backbone Models

Each backbone model (DINO V2, FaceNet, MobileNet V3, and MediaPipe Face Landmarker) was carefully adapted to suit the target dataset and task. For the DINO V2, two linear layers were added after the class token output of the Vision Transformer. During fine-tuning, only these additional layers were trained, keeping the pre-trained DINO V2 weights frozen. For classification tasks, the linear layers output the number of classes in the target dataset, while for regression tasks the linear layers output a single value. In the case of FaceNet, a small Multi-Layer Perceptron (MLP) was added on top of the embedding. For classification, this MLP consists of two hidden layers (64 and 32 neurons) and a final output layer matching the number of target classes. For regression, a similar MLP structure is used but with a single output neuron. During fine-tuning, the FaceNet weights were frozen, and only the MLP was trained. For MobileNet V3, the final classification layer of the pre-trained MobileNet V3 was removed. A Global Average Pooling layer was added after the convolutional features. For classification, two new dense layers were added with output neurons corresponding to the number of target classes. Fine-tuning involved training the newly added layers keeping earlier layers frozen. The MediaPipe Face Landmarker adaptation began by flattening the 468 3D landmarks into a 1404-dimensional vector (468×3). A feature reduction layer (100 neurons) was added to compress the high-dimensional input. For classification, another dense layer was added with neurons equal to the number of target classes. For regression, a linear layer with appropriate output neurons for the regression task was added. Only these additional layers were trained, keeping the weights of the face landmarker model fixed. In all models, the appropriate activation function was utilized in the final layer. For multi-class classification, a softmax activation was employed, whereas a sigmoid activation was utilized for binary classification. In the case of regression tasks, no activation was applied.

4.4 Training and Validation

In the context of model training, the selection of appropriate optimization strategies, loss functions, and regularization techniques is of considerable importance. These choices directly influence the model's learning, its ability to generalize, and its overall performance on both seen and unseen data. This section provides an examination of the methodology employed, encompassing the optimizer, loss functions, and key regularization techniques. The Adam (Adaptive Moment Estimation) optimizer with weight decay was employed during the training process. This choice combines the benefits of Adam's adaptive learning rates for each parameter with the regularization effect of weight decay, which helps to prevent overfitting and improve generalization. The addition of weight decay to Adam introduces a form of L2 regularization, which lead to simpler models with improved performance on unseen data. To prevent overfitting and optimize computational resources, an early stopping mechanism was implemented. This technique monitors the model's performance on a validation set during training and halts the process when the performance ceases to improve. A maximum of 100 epochs was set as an upper bound for the training duration, striking a balance between sufficient time for learning and excessive computation. The following subsections present an in-depth analysis of the various loss functions evaluated for both classification and regression tasks. The analysis explores how each function shaped the model's learning process and influenced its performance.

4.4.1 Classification Loss Functions

In order to evaluate the efficacy of different loss functions in addressing the inherent challenges associated with various types of classification problems, a series of loss functions were tested. These functions were evaluated to determine their effectiveness in improving the model's predictive accuracy and robustness.

- Cross-Entropy Loss: This is the standard loss function for multi-class classification problems. It measures the divergence between the predicted probability distribution and the actual distribution. Cross-entropy loss is effective in driving the model to predict probabilities that are close to 1 for the correct class and close to 0 for others.

- Binary Cross-Entropy Loss: Specifically used for binary classification tasks, this loss function is a special case of cross-entropy loss where only two classes are considered. It is highly effective in scenarios where the outcome is binary, pushing the model to make confident predictions between the two classes.

- Ordinal Categorical Loss: For ordinal classification problems, where the classes have a natural order. This loss function penalizes the model based on the distance between the predicted and actual classes, which is crucial in maintaining the ordinal relationship between classes.

- Weighted Cross-Entropy Loss: To address class imbalance, weighted cross-entropy loss was employed. By assigning higher weights to underrepresented classes, this loss function ensures that the model does not disproportionately favor the majority class, improving the performance on minority classes.

- Focal Loss: This loss function reduces the relative loss for well-classified examples, allowing the model to focus more on difficult cases, which is particularly useful in highly imbalanced datasets.

- KL Divergence Loss: For tasks requiring the model to match or approximate a probability distribution. This loss function measures the difference between the predicted probability distribution and a target distribution.

- Estimated Error Loss: This experimental loss function was tested to directly minimize the estimated error in classification. It attempts to reduce the expected classification error by incorporating a measure of prediction uncertainty into the loss calculation.

4.4.2 Regression Loss Functions

In the regression tasks, a variety of loss functions were evaluated to identify the optimal one for minimizing the error between the predicted and actual continuous values. These loss functions correspond to different aspects of regression, including absolute errors and distributional predictions.

- Mean Squared Error Loss: The most commonly used loss function for regression tasks, MSE calculates the average of the squared differences between predicted and actual values. MSE is sensitive to large errors, making it suitable for applications where large deviations are particularly undesirable.

- Mean Absolute Error Loss: MAE was tested to complement MSE by measuring the average absolute difference between predicted and actual values. Unlike MSE, MAE treats all errors equally, making it more robust to outliers and providing a more balanced error metric.

- Weighted MSE Loss: Similar to the weighted cross-entropy in classification, weighted MSE loss was used to address situations where certain predictions are more critical than others. By assigning different weights to different errors, the model can prioritize minimizing errors in more important predictions.

- Distributional Regression Loss: This loss function was tested for tasks requiring the model to predict a probability distribution rather than a single point estimate. Distributional regression loss helps in scenarios where understanding the uncertainty or variability of predictions is crucial.

- Quantile Regression Loss: For regression tasks where different quantiles of the target distribution are of interest. This loss function is particularly useful in predicting the median, lower, or upper quantiles of the distribution, providing a more comprehensive understanding of the possible outcomes.

4.4.3 Loss Function Selection

Based on extensive experimentation and analysis of various loss functions, this study ultimately employed two distinct loss functions for the different aspects of the task at hand. For the regression component, MAE loss was selected as the optimal choice. This decision was grounded in empirical evidence from numerous experiments, where MAE consistently outperformed other regression loss functions in terms of model performance and prediction accuracy. The robustness of MAE to outliers and its ability to provide stable gradients during training likely contributed to its superior performance in this context. With regard to the classification task, the Ordinal Categorical loss function was adopted. This choice was primarily motivated by the inherent ordinal nature of the classification problem at hand. The incorporation of ordinal categorical loss enabled the explicit incorporation of the ordinal relationship between classes into the learning process, effectively forcing the model to understand and respect the sequential order of the categories. This approach proved to be highly effective, as evidenced by the superior results achieved in subsequent evaluations.

4.4.4 Validation

To robustly validate the model's performance and ensure its generalizability across different subsets of data, k-fold cross-validation [14] was employed as the primary evaluation technique. This method is particularly advantageous in providing a reliable estimate of the model's performance by mitigating the risks associated with overfitting or the influence of data partitioning. In k-fold cross-validation, the dataset is partitioned into k equal-sized subsets, or ”folds.” The model is trained and evaluated k times, with each iteration utilizing a distinct fold as the validation set and the remaining k-1 folds as the training set. As illustrated in Figure 4.6, this process is repeated until each fold has served as the validation set exactly once. The overall performance metric is then calculated as the mean of the metrics obtained from each of the k iterations.

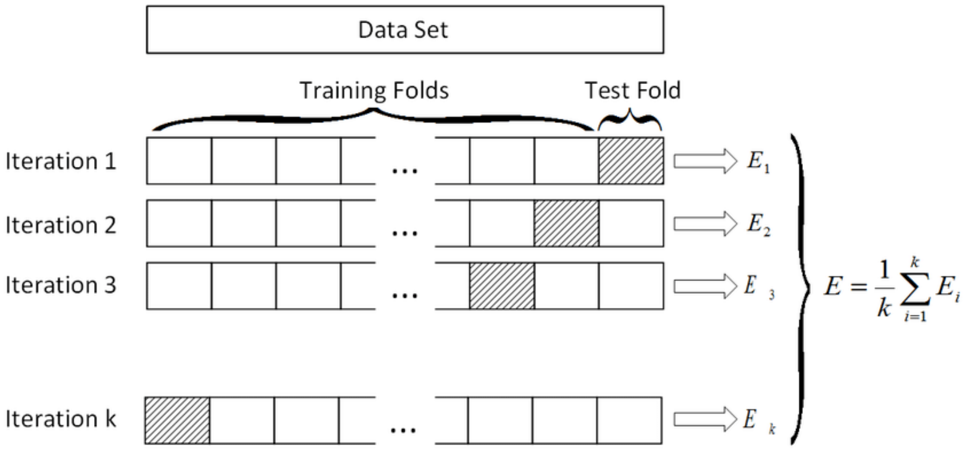


Figure 4.6: K-fold cross-validation

4.5 Hyperparameter Optimization

Hyperparameter optimization was conducted to fine-tune the model and enhance its performance by selecting the most effective combination of hyperparameters. To this end, the tool Weights and Biases (W&B) was employed, which is designed to facilitate the tracking of experiments and the automation of hyperparameter tuning. W&B enabled the management of numerous experiments by logging the results of different hyperparameter configurations and visualizing their impact on model performance (Figure 4.7). Through the integration of W&B, the optimization process was streamlined, enabling efficient exploration of the hyperparameter space and leading to the identification of optimal settings that significantly improved the model's accuracy and generalization capabilities.

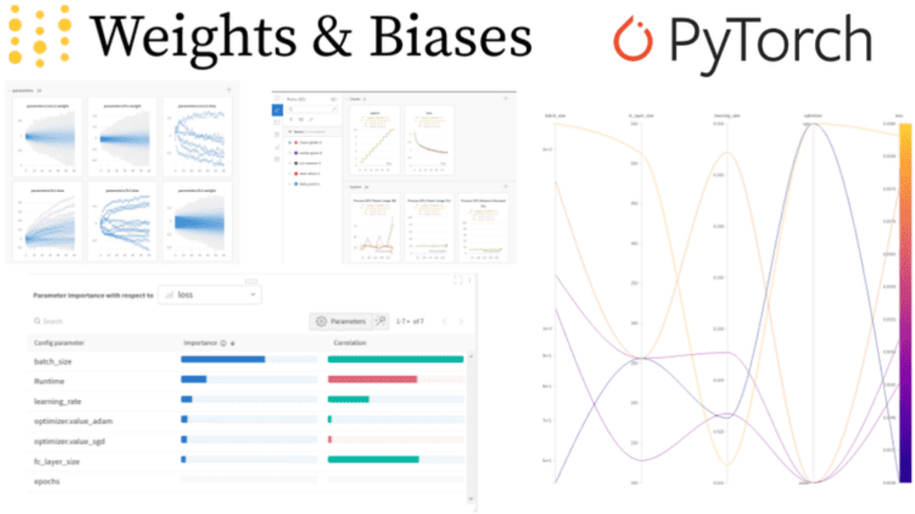


Figure 4.7: Weight and Biases

Chapter 5

Results

5.1 Experimental Setup

5.1.1 Hardware Configuration

All experiments were performed on an Amazon SageMaker notebook instance with the following hardware configuration:

- Instance type: ml.g4dn.2xlarge
- CPU: 8 vCPU Intel Xeon Platinum 8259CL CPU @ 2.50GHz
- GPU: 1 x NVIDIA T4 Tensor Core GPU
- Memory: 32 GB RAM
- Storage: 50 GB of attached EBS storage

5.1.2 Software Environment

The notebook instance was running on Amazon Linux 2, utilizing a Conda environment with Python 3.10.10. The environment was configured with essential packages:

- NumPy 1.26.3
- Torch 2.2.2
- Torchvision 0.17.2
- OpenCV 4.7.0.72

5.2 Experimental Results

The results of our study are presented in Tables 5.1 and 5.2, which show the performance in classification and regression tasks, respectively. In the classification task, DINO V2 (L) demonstrated the best performance with a MAE of 1.02 and an accuracy of 48%. This was closely followed by DINO V2 (S) with an MAE of 1.13 and 47% accuracy. The Face Landmarker (FL) model showed competitive results with an MAE of 1.28 and 46% accuracy. MobileNet 3 (MN 3) and FaceNet (FN) had comparatively lower performance. For the regression task, DINO V2 (L) again outperformed other methods with an MAE of 1.04 and an accuracy of 41%. Overall, the DINO V2 variants, particularly the larger model, consistently demonstrated superior performance across both classification and regression tasks. It is noteworthy that the accuracy in the regression task was calculated by approximating the result to the nearest integer, providing a discrete measure of performance for this continuous prediction task.

Method	**FL**	**MN 3**	**FN**	**DINO V2 (S)**	**DINO V2 (L)**
MAE	1.28	1.34	1.44	1.13	**1.02**
Accuracy (%)	46	39	41	47	**48**

Table 5.1: Classification Results

To provide a more detailed view of the DINO V2 (L) model's performance in the classification task, Figure 5.1 presents its confusion matrix. This visualization reveals the model's strengths and weaknesses across different classes.

Method	FL	MN 3	FN	DINO V2 (S)	DINO V2 (L)
MAE	1.41	1.39	1.47	1.29	**1.04**
Accuracy (%)	26	26	23	38	**41**

Table 5.2: Regression Results

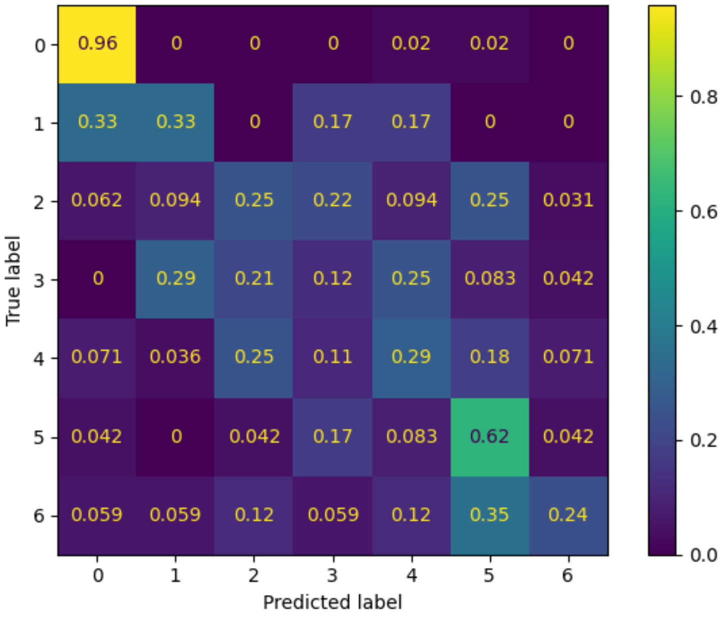


Figure 5.1: Confusion Matrix (DINO V2 (L))

Notably, the model shows strong performance in identifying class 0, with 96% correct predictions. It also performs well for class 5, correctly identifying 62% of instances. However, the matrix also highlights areas for improvement, particularly for classes 1 and 6, where the model shows more significant confusion with other classes. For instance, 33% of true class 1 instances were misclassified as either class 0 or class 1. The confusion matrix also reveals some interesting patterns of misclassification, such as a tendency to confuse classes 2, 3, and 4, which could indicate some shared features among these classes that the model is struggling to differentiate. This detailed breakdown complements the overall accuracy metric and provides valuable insights for potential model refinement and understanding of class-specific challenges in the dataset. The classification report of the model is presented in Table 5.3. The model achieved an overall accuracy of 0.48 across all classes. Class 0 demonstrated the best performance with a precision of 0.85, a recall of 0.96, and an F1-score of 0.90. In contrast, Class 1 showed the lowest performance with a precision of 0.14, a recall of 0.33, and an F1-score of 0.20. The macro-average scores, which afford equal weight to each class, yielded the following results: 0.39 for precision, 0.40 for recall, and 0.38 for F1-score. The weighted averages, which account for class imbalance, exhibited slight improvement at 0.48, 0.48, and 0.47 for precision, recall, and F1-score, respectively. These findings indicate that while the model demonstrates efficacy for certain classes, there is potential for enhancement, particularly for classes with lower F1 scores.

Class	Precision	Recall	F1-Score	Support
0	0.85	0.96	0.90	49
1	0.14	0.33	0.20	6
2	0.35	0.25	0.29	32
3	0.16	0.12	0.14	24
4	0.35	0.29	0.31	28
5	0.41	0.62	0.49	24
6	0.44	0.24	0.31	17
Accuracy: 0.48				
Macro Avg: 0.39 / 0.40 / 0.38 / 180				
Weighted Avg: 0.48 / 0.48 / 0.47 / 180				

Table 5.3: Classification Performance Metrics

To further assess the model's efficacy, a binary classification analysis was conducted, as illustrated in Figure 5.2. This simplified task serves to illustrate the model's considerable discriminative capacity when the problem is reduced to two classes. The confusion matrix for binary classification demonstrates the model's efficacy, with 96% accuracy for class 0 and 94% accuracy for class 1. This high performance in the binary setting indicates that the model is particularly adept at distinguishing between two broad categories, although it may face more challenges in the multi-class scenario. The misclassification rates are notably low, with only 4.1% of class 0 instances being incorrectly labeled

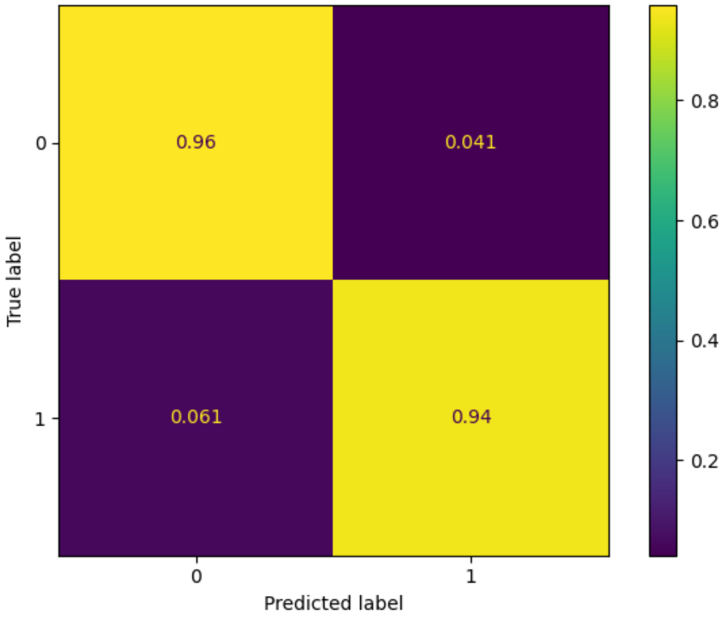


Figure 5.2: Binary Confusion Matrix (DINO V2 (L))

as class 1, and 6.1% of class 1 instances mislabeled as class 0. This binary classification performance provides a valuable benchmark for comparing DINO V2 (L) with other models and highlights its robustness in a simplified classification task. The significant improvement in accuracy from the multi-class to binary classification demonstrates the impact of increasing task complexity as the number of classes grows. The binary classification report of the model are presented in Table 5.4. For the purposes of this analysis, class 0 is considered the negative class, while all other classes (1-6) are combined into the positive class. The model demonstrated an overall accuracy of 0.95, indicating a robust capacity to differentiate between the positive and negative classes. The model exhibited optimal performance in class 1, with a precision of 0.99, a recall of 0.94, and an F1-score of 0.96. This indicates that the model is highly effective at identifying faces exhibiting pain, with a minimal false negative rate. Even the class 0 demonstrated exceptional performance, with a precision of 0.85, a recall of 0.96, and an F1-score of 0.90. The macro-average scores, which give equal weight to both classes, were 0.92 for precision, 0.95 for recall, and 0.93 for F1-score. The weighted averages, which account for class imbalance, were similarly high at 0.95, 0.95, and 0.95 for precision, recall, and F1-score, respectively. These results suggest that the model performs exceptionally well in binary classification, effectively distinguishing between class 0 and all other classes combined. The high accuracy and balanced performance across both classes indicate a robust and reliable classification model for this binary task.

Class	Precision	Recall	F1-Score	Support
0	0.85	0.96	0.90	49
1 (1-6 combined)	0.99	0.94	0.96	131
Accuracy: 0.95				
Macro Avg: 0.92 / 0.95 / 0.93 / 180				
Weighted Avg: 0.95 / 0.95 / 0.95 / 180				

Table 5.4: Binary Classification Performance Metrics

5.3 Model Interpretability

Having established that the model utilizing DINO V2 demonstrated superior performance in both classification and regression tasks, the focus will now shift to an examination of interpretability of this model. The interpretability of DL models enables the understanding of their decision-making processes, the validation of their reasoning, and the identification of potential biases or limitations. This section presents an analysis of the model's inner workings, employing three distinct methods to gain insights into the model's behavior: PCA-extracted features, attention map of the class token, and occlusion. Each of these techniques offers a unique perspective on how the model processes information and makes predictions, collectively enhancing our understanding of its strengths and potential limitations.

5.3.1 PCA Extracted Features

The interpretability analysis begins with an examination of the Principal Component Analysis (PCA) of the features extracted by the DINO V2 model. The

DINO V2 paper introduces a novel way to visualize the features learned by the model using PCA:

1. DINO V2 extracts patch features from input images using its self-supervised network.

2. The PCA is conducted on these patch features in two stages. The initial step involves the application of PCA to the complete set of patch features, with the objective of identifying the principal components with the highest variance. Subsequently, the projection of each patch onto the principal components is calculated. This yields a low-dimensional representation of each patch.

3. The low-dimensional patch representations are mapped to RGB colors, with each principal component corresponding to one color channel. This creates a segmented, colorful visualization of the image where each color represents a distinct feature learned by the model.

4. The PCA visualization demonstrates the model's capacity to discern and process disparate image features and its ability to handle complex pixel-level information without supervision from text or captions.

By providing a colorful, segmented view of the image based on the learned features, PCA visualization greatly improves the interpretability of DINO V2. Figure 5.3 presents visualizations of the feature representations extracted by our DINO V2 model for five different input images.

5.3.2 Attention Map of the Class Token

In addition to PCA-extracted features, another powerful tool for interpreting the DINO V2 model's behavior is the analysis of attention maps, particularly focusing on the class token in the last layer. The class token, introduced in Vision Transformers, serves as a global representation of the entire image and

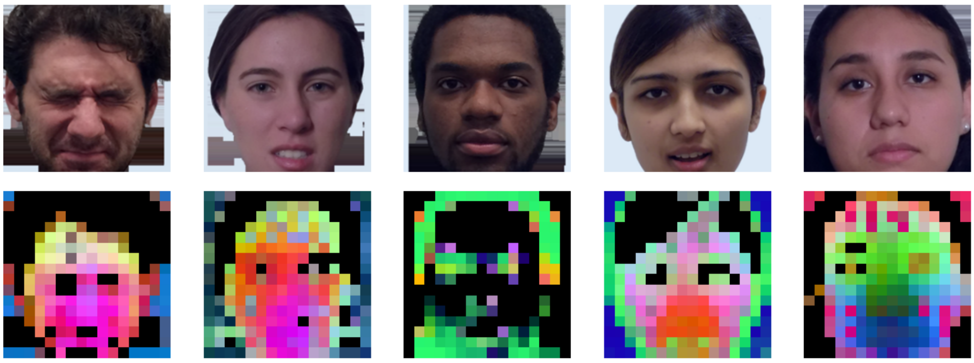


Figure 5.3: PCA extracted features of the penultimate layer of DINO V2

plays a key role in the model's final prediction. The attention mechanism inherent to transformer-based models, such as DINO V2, enables the model to concentrate on different regions of the input when formulating predictions. By visualizing the attention weights associated with the class token, one can gain insights into which regions of the image the model considers most important. In order to generate and interpret the attention map of the class token, the following steps must be taken. The initial step involves a forward pass of an input image through the DINO V2 model is performed. Next, the attention map of the class token is extracted from the final layer. These weights are then reshaped and normalized to match the input dimensions. Ultimately, the resulting heatmap is overlaid on the original image. This process enables the visualization of the model's focus areas, offering insights into its decision-making process. Figure 5.4 shows examples of attention maps for two input images.

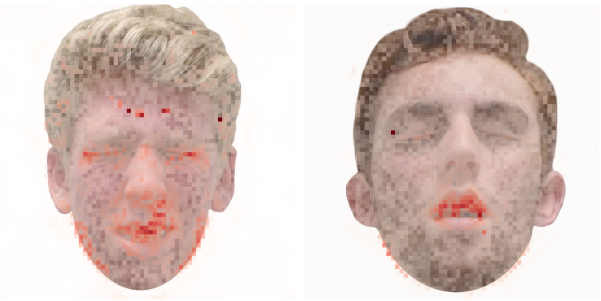


Figure 5.4: Attention maps of the class token for the last layer of DINO V2

5.3.3 Occlusion

The final interpretability technique is the occlusion method, which assists in comprehending the significance of various regions within an input image for the model's classification decision. This method entails the systematic occlusion of different portions of the image, which allows for the observation of changes in the model's prediction. This process enables the generation of importance heatmaps. A sliding window approach was employed, whereby 16x16 pixel patches of each input image were systematically occluded with a black square. For each occluded version, the change in the model's prediction probability for the correct class was recorded. Figure 5.5 presents a series of original images alongside their corresponding occlusion heatmaps. In these heatmaps, blue areas indicate regions where occlusion caused a significant drop in the correct class probability, suggesting these areas are crucial for classification.

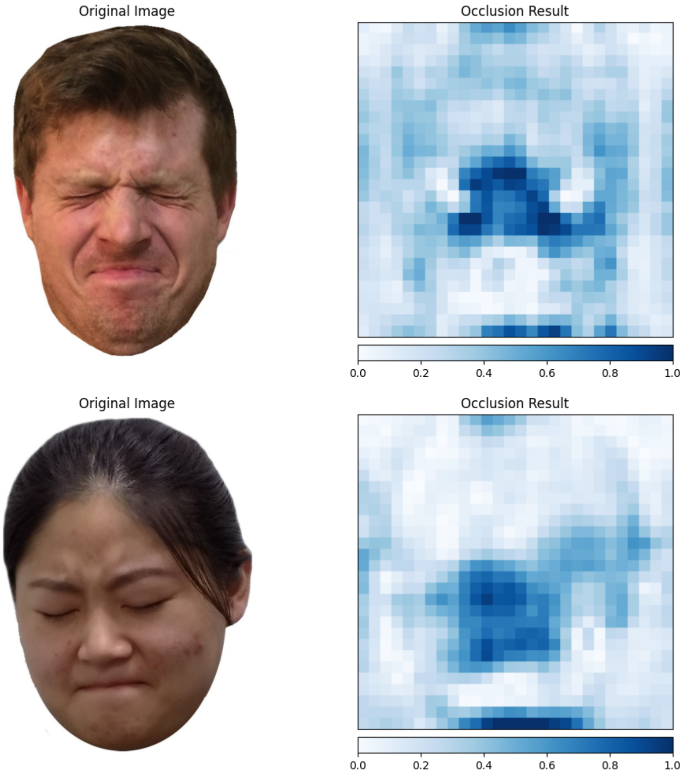


Figure 5.5: Occlusion

Chapter 6

Discussion

6.1 Model Performance Analysis

The results of our study, presented in Tables 5.1 and 5.2, offer valuable insights into model performance across both classification and regression tasks. A key finding is the consistently superior performance of the DINO V2 models, particularly the larger variant (L), across all tasks. DINO V2 (L) consistently achieved the lowest MAE and the highest accuracy, underscoring its ability to capture intricate features and relationships within the data. This observation highlights the strength of the self-supervised learning approach employed by DINO V2, especially when combined with increased model capacity. The strong performance of DINO V2 (L) aligns with the growing research in computer vision and machine learning, where self-supervised learning methods have demonstrated an exceptional ability to learn robust representations without reliance on labeled data. This is particularly advantageous in fields such as healthcare, where labeled data can be scarce, expensive, or challenging to obtain. An interesting trend observed across all models is the generally decreased performance in regression tasks compared to classification. The comparison between regression and classification models is based on both quantitative and qualitative observations, despite the different nature of their evaluation metrics. In order to facilitate a more direct comparison, both accuracy and MAE were calculated for both types of models. In the case of regression models, MAE is calculated traditionally, while accuracy is approximated by rounding the predicted continuous value to the nearest class. For classification models, accuracy is calculated in the standard way, while MAE is adapted by treating class labels as numeric values. The fact that classification models tend to outperform regression models is supported by higher accuracy scores, lower MAE values, and consistency across both metrics. However, it is essential to note that this comparison allows for a relative performance assessment rather than an absolute one, given the adapted use of metrics across different types of tasks. This is evident in the lower accuracy scores and slightly higher MAE values for regression, reflecting the inherent challenges of regression tasks. Unlike classification, which involves assigning discrete categories, regression requires precise numerical predictions, making it a more demanding task. This performance gap highlights the importance of tailoring model architectures and training strategies to the specific nature of the problem, particularly when dealing with continuous versus categorical outputs. The consistency in the ranking of methods across both tasks (DINO V2 (L) > DINO V2 (S) > FL > MN 3 FN) indicates that the relative strengths of these methods are not confined to a specific task. Instead, they reflect the fundamental capabilities of these models in feature extraction and representation learning. This consistency is noteworthy, as it suggests that the underlying methodologies employed by DINO V2 models, particularly the larger variant, are robust and versatile. The performance gap between DINO V2 (L) and DINO V2 (S) further underscores the impact of model size on performance. While the larger model consistently outperforms its smaller counterpart, the difference, particularly in classification tasks, is not overwhelmingly large. This observation raises important considerations regarding the trade-offs between model size, computational resources, and performance gains. In contexts where resources are limited, the smaller DINO V2 (S) may offer an optimal balance between performance and efficiency, making it a viable option without a significant loss of accuracy. In contrast, the relatively lower performance of traditional methods such as FL, MN 3, and FN, particularly in regression tasks, highlights their potential limitations in comparison to more advanced models like DINO V2. As the field of machine learning evolves, there may be a need to reassess and update these traditional approaches to maintain competitiveness with newer, more sophisticated models.

6.1.1 Confusion Matrix

The confusion matrix for DINO V2 (L) in Figure 5.1 reveals a good diagonal tendency, a positive indicator of the model's performance, particularly in ordinal classification tasks. This diagonal reflects the model's ability to achieve accurate classifications and also to grasp the ordinal relationships between classes. When misclassifications occur, they tend to involve adjacent classes, a behavior that mirrors the challenges even humans face when distinguishing between closely related categories, such as pain levels 3 and 4. These performances are particularly valuable in real-world applications where the proximity of predictions to the true class is often more important than exact matches. The model's ability to discern subtle differences and its alignment with human judgment in similar scenarios highlight its sophistication in handling complex ordinal classification tasks.

6.1.2 Interpretability

The interpretability of DINO V2's reasoning process is significantly enhanced by the visualizations presented in Figures 5.3, 5.4, and 5.5. Figure 5.3, showcasing the PCA-extracted features from the penultimate layer of DINO V2, provides a glimpse into the model's high-level representations. The contrast between the original facial images and their corresponding abstract features illustrates how DINO V2 distills complex visual information into meaningful patterns, which, while abstract, likely encode critical information about facial expressions and pain levels. Further insights are provided by Figure 5.3, which highlights the regions most critical to the model's predictions. The concentrated importance around the eyes, nose, and mouth supports the activation patterns observed in the other methods, reinforcing the idea that DINO V2 focuses on physiologically relevant facial features for pain evaluation. The attention maps presented in Figure 5.4 provide additional evidence into the focus areas of the DINO V2 when assessing pain levels. These visualizations, derived from the class token's attention weights in the last layer, reveal the model's areas of emphasis for each input image. In both examples, attention is concentrated on key facial features, particularly the eyes, eyebrows, and mouth regions. The first image illustrates a more distributed attention pattern across the face, with a notable focus on the forehead, eyes, and mouth. This suggests that the model is considering a broader range of facial cues, potentially indicative of a more complex or ambiguous pain expression. Conversely, the second image displays a more localized attention pattern, with an intense focus on the mouth. This concentrated attention may indicate a more pronounced or specific pain expression that the model has identified as particularly salient. These visualizations collectively offer valuable insights into DINO V2's internal decision-making processes, enhancing trust in the model by demonstrating its reliance on human-understandable concepts. Interpretability is particularly important in clinical settings, where transparency and explainability are crucial for adoption. By bridging the gap between complex neural computations and intuitive human concepts, these analyses make a strong case for DINO V2's potential applicability in real-world healthcare scenarios.

6.2 Comparison with Existing Methods

The model presented in this research project, which achieved an accuracy of 95%, demonstrated superior performance when compared to existing methods in the field of automatic pain assessment. This comparison revealed valuable insights into the efficacy of different approaches and the impact of dataset size on model performance. In [8] a binary classifier model is deployed with the objective of discriminating between the absence and presence of pain. The authors employed the OpenFace toolkit to extract AUs from facial expressions, which were then fed into a neural network classifier with two dense layers. After 400 training epochs, the model achieved an accuracy of approximately 94%. It is noteworthy that the training data combined both the *Delaware Pain Database* and the *UNBC McMaster Shoulder Pain dataset*, thereby providing a broader range of pain expressions and contexts. In their study, Sabater-Gárriz, Álvaro, et al. [35] highlighted the necessity of a large and diverse dataset for the training of deep learning models in pain recognition. The researchers merged three extensive databases: the *UNBC McMaster Shoulder Pain Expression Archive Database* [29], the *Multimodal Intensity Pain dataset (MInt PAIN)* [18], and the *Delaware Pain Database* [31]. This approach highlights a common challenge in healthcare AI applications, particularly in facial expression analysis. It is noteworthy that the presented model achieved 95% accuracy, given that only the *Delaware Pain Database* was utilized for training. This is in clear contrast to the approaches in both referenced papers, which relied on combined datasets to ensure sufficient training data. It is crucial to acknowledge that deep learning models often require extensive training data, particularly in the context of healthcare applications where data diversity is of immense importance. The challenge of acquiring comprehensive, high-quality datasets for pain assessment, especially those involving facial expressions, represents a significant obstacle in this field. The necessity for Sabater-Gárriz, Álvaro, et al. [35] to combine multiple datasets highlights the difficulties associated with this process. The performance of the model that is the subject of this research, therefore, not only demonstrates its effectiveness but also highlights a potential pathway for developing robust pain assessment tools with more limited datasets. This approach could be particularly valuable in scenarios where large, diverse datasets are not available or are challenging to compile due to privacy concerns or the specific nature of the pain assessment task. However, it is essential to consider that while the model achieved promising results, further validation on diverse datasets and in various clinical settings would be necessary to ensure its generalizability and robustness across different patient populations and pain conditions.

The results of this study have led to a significant contribution to the field of computer vision applications in healthcare. The research demonstrated that a vision foundation model, trained in a self-supervised manner on billions of natural images, can effectively address the lack of data often encountered in healthcare-specific computer vision tasks. The good performance of the DINO V2 model underscores the potential of transfer learning from large-scale, general-purpose models to specialized medical applications. This approach presents a promising solution to one of the most persistent challenges in healthcare: the scarcity of large, diverse, and high-quality labeled datasets. As the field progresses, the use of pre-trained vision foundation models may become a key strategy in overcoming data limitations in healthcare AI, potentially accelerating the development and deployment of accurate, reliable, and interpretable computer vision systems across various medical domains. This could facilitate the development and deployment of accurate, reliable, and interpretable computer vision systems across various medical domains.

Chapter 7

Conclusion and Future Work

7.1 Summary of Findings

The study on automatic pain assessment using the DINO V2 model has yielded several significant findings. The DINO V2 model exhibited superior performance in both classification and regression tasks, achieving an accuracy of 48% in classification. These results are particularly notable given the inherent complexity of pain assessment. The confusion matrix indicated a neat diagonal tendency, which suggests that the model effectively captured the ordinal nature of pain levels. This characteristic is of particular value in pain assessment, where near-misses are often clinically acceptable and reflect the complex subjective nature of pain perception. Visualization techniques, including principal component analysis, attention map of the class token, and occlusion analysis, provided substantial insights into the model's decision-making process. These visualizations confirmed that the model focuses on relevant facial areas for pain evaluation, aligning with the clinical understanding of pain expression. It is noteworthy that the model achieved an accuracy of 95% using only the Delaware Pain Database, thereby outperforming existing methods that relied on larger, combined datasets. This achievement underscores the potential of the proposed approach in scenarios where large, diverse datasets are not readily available. The study demonstrate the potential of self-supervised learning approaches in healthcare applications, particularly in scenarios with limited labeled data. This finding is especially relevant in the field of pain assessment, where obtaining large, annotated datasets can be challenging due to privacy concerns and the subjective nature of pain experiences.

7.2 Limitations and Challenges

Despite the promising results, the research project faced several limitations and challenges that require further examination. The relatively small size of the Delaware Pain Database, compared to combined datasets used in other studies, may limit the model's generalizability to diverse populations. This limitation highlights the persistent challenge in AI healthcare applications of balancing model performance with dataset diversity and size. A significant limitation of the current approach is its dependence on static images for pain assessment. Pain is not an instantaneous feeling but a dynamic experience that may fluctuate over time. By examining isolated frames, the model may fail to capture crucial temporal data that could offer a more precise representation of an individual's pain state. This limitation may result in an oversimplification of the intricate and often evolving nature of pain experiences. The inherent subjectivity of pain assessment presents another challenge. Pain is a deeply personal experience, and the ground truth labels in the dataset may not always accurately reflect an individual's pain experience. This subjectivity introduces a level of uncertainty that must be carefully considered when interpreting the model's outputs. Although the model's principal reliance on facial expressions is effective, it may not fully capture other crucial pain indicators, such as body language, physiological measures, or vocal cues. This limitation points to the need for more comprehensive, multimodal approaches to pain assessment that can integrate various sources of information. Furthermore, the deployment of facial recognition technology in healthcare settings gives rise to significant privacy concerns and ethical questions regarding consent and data usage. For this reason, the introduction of the EU AI Act has significant implications for applications like this one. The Act aims to regulate AI technologies based on their risk levels, emphasizing transparency, accountability, and ethical considerations in AI deployment. Given that this application seeks to improve pain assessment, a critical area in healthcare, the AI Act would classify this pain assessment model as a high-risk AI system due to its use in healthcare settings. In accordance with the existing regulations of the AI Act, such an application would be permitted, but subject to strict requirements [30].

- Establish a continuous risk assessment and mitigation system throughout the system lifecycle.
- Conduct data governance, ensuring that training, validation and testing datasets are relevant and sufficiently representative.
- Draw up technical documentation to demonstrate compliance and provide clear information about the AI system's capabilities and limitations.
- Design the AI system for record-keeping to enable it to automatically record events to ensure traceability.
- Allow deployers to implement human oversight on the system.
- Achieve appropriate levels of robustness and cybersecurity.

7.3 Future Research

The findings and limitations provide a foundation for several promising avenues for future research. A principal area of focus is the development of multimodal pain assessment models that integrate additional data sources, including body posture, voice analysis, and physiological signals. This approach could result in the creation of a more comprehensive pain assessment tool, thus enabling a fuller picture of the patient's pain experience to be captured. A key area for future research is the incorporation of temporal information in pain assessment. The creation of models capable of analyzing video sequences instead of static images would facilitate the capture of pain dynamics over time. This approach could provide a more detailed and complex understanding of pain experiences, including the fluctuation and duration of pain episodes. Such temporal models may prove capable of distinguishing between acute and chronic pain with greater efficacy and of capturing subtle changes in pain levels that might otherwise be missed in single-frame analyses. In conjunction with video analysis, the incorporation of audio data represents another promising direction for research. Vocal cues, including tone, pitch, and verbal expressions of discomfort, can provide valuable supplementary data for pain assessment. The creation of models that can integrate visual, temporal, and auditory information has the potential to significantly enhance the accuracy and comprehensiveness of automatic pain assessment systems. Developing larger, more diverse datasets through collaboration with healthcare institutions is essential for advancing the field. These datasets should include video and audio recordings in addition to static images, representing various ethnicities, age groups, and pain conditions in order to ensure the model's applicability across diverse populations. The ethical implications of AI in pain assessment require collaboration with ethicists and policymakers to develop robust ethical guidelines for the deployment of AI-based pain assessment tools. This is crucial for ensuring the responsible and beneficial use of this technology, especially when dealing with more invasive data collection methods like continuous video and audio monitoring. Future research should also explore the potential for developing personalized pain assessment models that can be fine-tuned to individual patients' pain expression patterns over time. This personalized approach could markedly enhance the accuracy and reliability of pain assessments, particularly for patients with atypical pain expressions or those with chronic conditions. Finally, conducting cross-cultural studies on the model's performance is essential to account for potential variations in pain expression and interpretation across different cultures. This research could lead to more culturally sensitive and globally applicable pain assessment tools. By pursuing these future research directions, it is possible to work towards the creation of more accurate, reliable, and ethically sound automatic pain assessment systems that capture the complex, dynamic nature of pain experiences. These advancements have the potential to significantly improve patient care, enhance pain management strategies, and contribute to a more nuanced understanding of pain across diverse populations and clinical contexts.

Bibliography

[1] T. Alghamdi and G. Alaghband. Facial expressions based automatic pain assessment system. *Applied Sciences*, 12(13), 2022. ISSN: 2076-3417. DOI: 10.3390/app12136423. URL: https://www.mdpi.com/2076-3417/12/13/6423.

[2] T. Alghamdi and G. Alaghband. Facial expressions based automatic pain assessment system. *Applied Sciences*, 12(13):6423, 2022.

[3] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[4] A. M. D. Association. Pain assessment in advanced dementia (painad) scale, 2005. URL: %5Curl%7Bhttps://geriatrictoolkit.missouri.edu/cog/painad.pdf%7D. [Online; accessed 11-August-2024].

[5] G. Bargshady, X. Zhou, R. C. Deo, J. Soar, F. Whittaker, and H. Wang. Ensemble neural network approach detecting pain intensity from facial expressions. *Artificial Intelligence in Medicine*, 109:101954, 2020.

[6] P. D. Barua, N. Baygin, S. Dogan, M. Baygin, N. Arunkumar, H. Fujita, T. Tuncer, R.-S. Tan, E. Palmer, M. M. B. Azizan, et al. Automated detection of pain levels using deep feature extraction from shutter blinds-based dynamic-sized horizontal patches with facial images. *Scientific Reports*, 12(1):17297, 2022.

[7] N. Ben Aoun. A review of automatic pain assessment from facial information using machine learning. *Technologies*, 12(6), 2024. ISSN: 2227-7080. DOI: `10.3390/technologies12060092`. URL: `https://www.mdpi.com/2227-7080/12/6/92`.

[8] M. Cascella, V. N. Vitale, F. Mariani, M. Iuorio, and F. Cutugno. Development of a binary classifier model from extended facial codes toward video-based pain recognition in cancer patients. *Scandinavian Journal of Pain*, 23(4):638–645, 2023.

[9] S. Cui, D. Huang, Y. Ni, and X. Feng. Multi-scale regional attention networks for pain estimation. In *Proceedings of the 2021 13th International Conference on Bioinformatics and Biomedical Technology*, pages 1–8, 2021.

[10] Z. Dai, H. Liu, Q. V. Le, and M. Tan. Coatnet: marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] S. El Morabit, A. Rivenq, M.-E.-n. Zighem, A. Hadid, A. Ouahabi, and A. Taleb-Ahmed. Automatic pain estimation from facial expressions: a comparative analysis using off-the-shelf cnn architectures. *Electronics*, 10(16):1926, 2021.

[14] T. Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21:137–146, 2011.

[15] C. Gao, J. Yan, S. Zhou, P. K. Varshney, and H. Liu. Long short-term memory-based deep recurrent neural networks for target tracking. *Information Sciences*, 502:279–296, 2019.

[16] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[17] Z. Hammal and J. F. Cohn. Automatic detection of pain intensity. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 47–52, 2012.

[18] M. A. Haque, R. B. Bautista, F. Noroozi, K. Kulkarni, C. B. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, et al. Deep multimodal pain recognition: a database and comparison of spatio-temporal visual modalities. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 250–257. IEEE, 2018.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[21] A. D. International. Dementia statistics, 2024. URL: `%5Curl%7Bhttps://www.alzint.org/about/dementia-facts-figures/dementia-statistics/%7D`. [Online; accessed 11-August-2024].

[22] S. Jamil, M. Jalil Piran, and O.-J. Kwon. A comprehensive survey of transformers for computer vision. *Drones*, 7(5):287, 2023.

[23] I. Karamitsos, I. Seladji, and S. Modak. A modified cnn network for automatic pain identification using facial expressions. *Journal of Software Engineering and Applications*, 14(8):400–417, 2021.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[25] L. F. e. a. Kumar A Sidhu J. Alzheimer disease, 2024. URL: `%5Curl%7Bhttps://www.ncbi.nlm.nih.gov/books/NBK499922/%7D`. [Online; accessed 11-August-2024].

[26] H.-C. Li, Z.-Y. Deng, and H.-H. Chiang. Lightweight and resource-constrained learning network for face recognition with performance optimization. *Sensors*, 20(21):6114, 2020.

[27] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[28] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(3):664–674, 2010.

[29] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: the unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 57–64. IEEE, 2011.

[30] T. Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

[31] P. Mende-Siedlecki, J. W. Qu-Lee, J. Lin, A. Goharzad, and A. Drain. The delaware pain database: a set of painful expressions and corresponding norming data, March 2019. DOI: `10.31234/osf.io/kjez5`. URL: `osf.io/preprints/psyarxiv/kjez5`.

[32] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[33] W. H. Organization. Dementia, 2023. URL: %5Curl%7Bhttps://www.who.int/news-room/fact-sheets/detail/dementia#:~:text=Alzheimer%20disease%20is%20the%20most%20common%20form%20and%20may%20contribute,frontal%20lobe%20of%20the%20brain).%7D. [Online; accessed 11-August-2024].

[34] Physiopedia. Visual analogue scale — physiopedia, 2024. URL: %5Curl%7Bhttps://www.physio-pedia.com/index.php?title=Visual_Analogue_Scale&oldid=356879%7D. [Online; accessed 7-August-2024].

[35] Á. Sabater-Gárriz, F. X. Gaya-Morey, J. M. Buades-Rubio, C. Manresa-Yee, P. Montoya, and I. Riquelme. Automated facial recognition system using deep learning for pain assessment in adults with cerebral palsy. *Digital health*, 10:20552076241259664, 2024.

[36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: a unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[37] A. Semwal and N. D. Londhe. Automated pain severity detection using convolutional neural network. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pages 66–70. IEEE, 2018.

[38] A. Semwal and N. D. Londhe. Eccnet: an ensemble of compact convolution neural network for pain severity assessment from face images. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 761–766. IEEE, 2021.

[39] A. Semwal and N. D. Londhe. Mvfnet: a multi-view fusion network for pain intensity assessment in unconstrained environment. *Biomedical Signal Processing and Control*, 67:102537, 2021.

[40] A. Semwal and N. D. Londhe. S-panet: a shallow convolutional neural network for pain severity assessment in uncontrolled environment. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0800–0806. IEEE, 2021.

[41] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[42] A. R. Wahab Sait and A. K. Dutta. Ensemble learning-based pain intensity identification model using facial expressions. *Journal of Disability Research*, 3(3):20240029, 2024.

[43] V. Warden, A. C. Hurley, and L. Volicer. Development and psychometric evaluation of the pain assessment in advanced dementia (painad) scale. *Journal of the American Medical Directors Association*, 4(1):9–15, 2003. ISSN: 1525-8610. DOI: `https://doi.org/10.1097/01.JAM.0000043422.31640.F7`. URL: `https://www.sciencedirect.com/science/article/pii/S1525861004702583`.

[44] Wikipedia contributors. Facial action coding system — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Facial_Action_Coding_System&oldid=1236719769`, 2024. [Online; accessed 7-August-2024].

[45] L. Wu, S. C. Hoi, and N. Yu. Semantics-preserving bag-of-words models and applications. *IEEE Transactions on Image Processing*, 19(7):1908–1920, 2010.

[46] X. Xin, X. Li, S. Yang, X. Lin, and X. Zheng. Pain expression assessment based on a locality and identity aware network. *IET Image Processing*, 15(12):2948–2958, 2021.

[47] X. Xin, X. Lin, S. Yang, and X. Zheng. Pain intensity estimation based on a spatial transformation and attention cnn. *Plos one*, 15(8):e0232412, 2020.

[48] R. Yang, X. Hong, J. Peng, X. Feng, and G. Zhao. Incorporating high-level and low-level cues for pain intensity estimation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3495–3500. IEEE, 2018.

[49] X. Ye, X. Liang, J. Hu, and Y. Xie. Image-based pain intensity estimation using parallel cnns with regional attention. *Bioengineering*, 9(12):804, 2022.

[50] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks, 2013. arXiv: `1311.2901 [cs.CV]`. URL: `https://arxiv.org/abs/1311.2901`.