Federated learning for automated pain detection

Artificial Intelligence Master Thesis

Author
Jimena Mejía de Miguel
8180067
j.mejiademiguel@students.uu.nl

May 2024

1st supervisor: Assist. Prof. dr. I. Onal Ertugrul
2nd supervisor: Prof. dr. A.A. Salah

Department of Information and Computing Sciences
Faculty of Science

Universiteit Utrecht

Contents

Abstract **4**

1 Introduction **5**

2 Research questions **6**

3 Literature Review **8**

3.1 Automatic Pain Detection 8

3.2 Federated Learning 12

3.3 Pain detection using federated learning 16

3.4 Differential Privacy 17

3.4.1 Differential Privacy in a Federated Learning system . . 18

4 Data **19**

4.1 UNBC-McMaster shoulder pain expression 20

4.2 The BioVid Heat Pain 22

5 Methodology **24**

5.1 Preprocessing 24

5.1.1 Data Selection 24

5.1.2 Feature Extraction 25

5.1.3 Data Relabeling 26

5.1.4 Data Modifications 27

5.2 Experimental setup 27

5.2.1 Centralized model 28

5.2.2 Local model 29

5.2.3 Federated learning model using a single database . . . 30

5.2.4 Federated learning model with two databases 32

5.2.5 Federated learning model with differential privacy . . . 34

6 Results **37**

6.1 Centralized model 37

6.2 Local model 41

6.3 Federated learning model using a single database 42

6.4 Federated learning model with two databases 45

6.5 Federated learning model with differential privacy 46

6.6 Summary of the results 48

7 Discussion and Conclusion **49**

8 Limitations and Future work **54**

Abstract

This thesis focuses on exploring the performance of automated pain detection systems through the application of federated learning. The study involves the use of two different databases (UNBC-McMaster Shoulder Pain Expression Archive database and the BioVid Heat Pain database) and investigates various aspects of this approach. More specifically, it evaluates performance disparities between individual database training and federated learning methods, with the goal of determining the feasibility and benefits of federated learning. This research also explores the use of federated learning within a single database, treating each patient as an individual "client" to evaluate the potential benefits in terms of data privacy, while maintaining or even improving the accuracy of pain detection. Furthermore, this project evaluates the result of enhancing the privacy using differential privacy. This study aims to provide valuable insights into the application of federated learning in the context of pain detection systems.

1 Introduction

Pain is described by the International Association for the Study of Pain (IASP), as “An unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage”[1]. Pain is an inherent aspect of the human condition. It acts as an alarm system, allowing people to be alert to possible damage or injury.

Pain, therefore, plays a crucial role in protecting the integrity of the human body. Acute pain usually goes away on its own with healing. However, if the pain lasts longer than 3 months, it is called chronic or persistent pain. Pain is a serious problem for many people. In fact, according to P. Mäntyselkä et al. it is the main reason that leads people to seek medical attention [2]. According to Cordell et al., 52.2% of emergency department visits are due to pain, while only 34.1% are unrelated to pain [3].

According to Gregory et al. acute pain is one of the main significant symptoms in hospitalised patients. Up to 35% of patients report severe pain and approximately 50% of patients report pain [4]. In a study by Zoëga et al. the prevalence of pain in hospitals has been even 83% [5]. However, the subjective nature of pain perception makes its assessment and treatment a complex challenge.

For differential diagnosis, appropriate therapy selection, progress monitoring, and determining whether or not a treatment should be continued or modified, a valid and reliable assessment of pain is required. Since uncontrolled pain not only causes suffering and lowers quality of life but also jeopardizes the nervous system [6], endocrine system [7], and immune function [8], pain assessment and management are crucial not only to provide comfort but also to prevent both immediate and long-term consequences that are harmful to the person’s overall health [6]. Chronic pain syndrome, which is frequently accompanied by anorexia, poor immunity, low focus, and sleep difficulties, can develop as a result of untreated pain. Furthermore, patients may experience issues and hazards as a result of improper therapy [9].

Health care and human-computer interaction are two areas where being able to precisely detect and quantify pain has major significance. In the medical field, a precise evaluation of pain is essential for diagnosis, planning treatments, and keeping track of patients’ wellbeing. Effective pain manage-

ment enhances patient quality of life while also ensuring prompt administration of necessary therapies, minimizing suffering and improving medical outcomes[6].
Automatic pain detection offers fascinating prospects to improve user experiences and broaden accessibility of technology in the context of human-computer interaction. With these developments, smartphones and computers might be able to detect discomfort during telemedicine consultations and modify their user interfaces to accommodate patients' pain-related constraints. These future-facing prospective uses hold the promise of more flexible and sympathetic user interfaces for technology [10].

Although precise pain detection is obviously important, the subjectivity of pain perception, the impact of cultural and societal influences on how pain is expressed, and the possibility of erroneous or deficient self-reports, particularly in those with cognitive impairments or communication challenges, are some of these limitations [11]. These difficulties highlight the urgent need for automated, unbiased pain assessment techniques.

The development of autonomous pain detection models has advanced significantly in the fields of artificial intelligence (AI) and machine learning in response to these difficulties. These models use a variety of data sources to infer the presence and severity of pain, such as physiological signs, facial expressions, vocal characteristics, and even neuronal activity [12]. However, the availability of vast and varied training datasets is essential for the performance of such models, which frequently prompts worries about data security and privacy, particularly in healthcare settings [13].

This is where federated learning, comes into play. Federated learning solves the privacy and security issues associated with conventional centralized machine learning techniques by enabling several institutions or edge devices to cooperatively train a global pain detection model while keeping sensitive data localized. It presents a system that upholds privacy laws and protects the privacy of patient data while improving the precision and effectiveness of pain detection [13].

2 Research questions

In this thesis, the following research topic and related sub-questions were formulated in the research proposal. The importance of these questions will be demonstrated in the literature review part. The approach, results, and discussion will be subsequently covered.

Research Question: To what extent is federated learning a suitable approach to improve the performance of automated pain detection systems?

Federated learning is a machine learning technique that allows multiple parties to collaboratively train a model without sharing their raw data. In the context of pain detection systems, this approach is expected to be advantageous because it can improve accuracy through collaborative training while respecting the privacy of different sources. By comparing the results of conventional centralized and local models with federated learning approaches, the performance of federated learning will be identified.

Sub-question 1: When a single database is used, how does the model trained using federated learning compare to centralized and local models?

Within a single database we simulate three scenarios: (1) centralized model which consists of aggregating all data in a center and performing conventional training (2) local model which assumes that individual models are trained only with the small datasets of local users and (3) federated learning model in which user's data are not shared directly but the local neural network weights are shared with the central server that trains a network to detect pain in a decentralized manner.

Sub-question 2: How do the models trained using multiple databases (UNBC and BioVid) in a federated learning setup compare to the centralized models trained on these models separately?

Through a comparative analysis of individual database training and federated training, this thesis aims to provide information on the feasibility of federated learning as an effective solution to improve automatic pain detection. This experiment help to clarify the practical applicability of federated learning, and whether it really offers an improvement over individual training.

Sub-question 3: How does performance vary when using differential privacy in the models compared to a less private model?

Through a comparative analysis of a federated model in which differential privacy is applied and one that does not, this study aims to establish how adding noise affects the final performance of the model. It is expected that the performance will be lower but at the same time it is aimed to establish whether it is worth the performance lost compared to the privacy gained.

3 Literature Review

3.1 Automatic Pain Detection

Pain is a complex phenomenon that affects the senses and the mind. It has important effects on healthcare. The efficacy of diagnosis, treatment and patient well-being depend on the rapid and accurate detection of pain[12]. For this reason, in recent years, great efforts have been devoted to this research topic, achieving significant advances. Early studies use more traditional methods such as manual assessments and conventional physiological markers.[14]. However, over the years, more complex[15][16] and multi-modal methods[17] [18] have been developed. These new approaches include the use of neural networks, among others[19]. This literature review provides an overview of the evolution of technologies for autonomous pain detection.

In pain detection, traditional approaches used human observations and self-reports from individuals. Although this way of assessing pain has significant value, it also has certain drawbacks[20]. With these methods, patients either verbally reported their pain or used measures, such as the visual analog scale or the numerical rating scale, to rate the intensity of their pain [21][22][23]. Although these techniques are very informative, they also have a high degree of subjectivity and can be affected by cognitive variables, cultural variations and personal pain tolerances[24]. All of this makes these methods less reliable since it is not possible to measure pain or pain level objectively for a large number of patients.

On the other hand, in the clinical setting, inter-observer variability is a major source of concern. How various medical experts interpret the pain of the same patient can vary, which can impact treatment choice and patient care by leading to differences in pain assessment[24]. These methods may also overlook the complex and continuous nature of painful sensations. In addition, recall bias and emotional states may influence self-reported pain levels, further reducing their reliability. Pain can be expressed non-verbally through body language, vocal intonation, and facial expressions. [25].

However, as these indicators can differ greatly from person to person, understanding them accurately can be difficult[26]. In addition, it has also been shown that depending on expectations or external cues, patients may change the way they respond, which could bias clinical judgments and lead to incorrect assessments of pain[24].

To overcome these problems, researchers began studying objective physiological measures such as blood pressure[27], electrodermal activity[28], and heart rate[29]. In addition to these markers, nonverbal cues such as self-reports, vocal intonation and facial expressions have been examined. Initially, when these measures were first used, studies tended to be based on a single marker. However, over the years, researchers began to incorporate more than one indicator and take a multi-modal approach. To examine multi-modal data for pain recognition, researchers used various machine learning algorithms, such as random forests, deep neural networks, and support vector machines (SVM)[30].

The following years have seen the integration of multi modal data for pain detection, recognizing pain as a complex experience with both physiological and emotional components. Besides traditional physiological markers, researchers have broadened their approach to include other data modalities, such as medical imaging or functional magnetic resonance imaging (fMRI), among others[31]. Combining these imaging modalities with machine learning algorithms allows pain to be assessed more accurately and objectively[32].

The integration of natural language processing (NLP) is an essential modality in pain detection systems. NLP techniques extract sentiment, context, and linguistic features from patients' descriptions of pain through textual descriptions and self-reports. NLP connects the subjective and objective components of pain assessment. Physiological and imaging data are often integrated with these textual data to provide a comprehensive understanding of pain[31].

The most recent studies use deep learning methods. Deep learning is a subset of machine learning that involves the use of artificial neural networks, in particular deep neural networks with multiple layers[33]. These techniques perform well for tasks involving complex patterns and large data sets because they automatically learn intricate features from raw data[12].

These research projects has been greatly influenced by databases such as UNBC-McMaster Shoulder Pain, or EmoPain which has provided a fundamental platform for method development. Egede et al.[34], presented the EMOPAIN 2020 challenge incorporating a dataset that included both handcrafted features and deep-learned models, such as facial landmarks, HOG, and deep vectors from pre-trained models like VGG-16 and ResNet-50. Pedersen's[30] study used a deep learning approach using a 4-layer contractile autoencoder. Through this, they achieved high accuracy at the frame level.

Convolutional neural networks (CNNs) is one of the most widely used method within the deep learning approach. In fact, in the study by Gkikas and Tsiknakis[12] in which they reviewed many of the studies on automated pain assessment carried out in the past ten years, more than 75% of all studies have used this method. Whether using 1D, 2D or 3D filters, this indicates that the convolution operation is now the fundamental element of deep learning. It is important to mention that studies using deep learning do not start until 2014. Even so, the authors report it, as the most widely used method[12].

Many studies have used deep learning models with handcrafted features to optimize pain assessment[34][35]. These hybrid approaches, using features such as facial landmarks, histograms of oriented gradients (HOG) and deep features extracted from pre-trained models, have shown promise[36]. In addition, lines of research have also been pursued where the importance of specific facial regions in the transmission of pain expressions is studied[37][38][39]. This has led to the development of models that focus on localized features. For instance Huang et al.[37] initially detected facial regions, including the left eye, right eye, nose, and mouth. A multi-stream CNN has been responsible for feature extraction, with a separate sub-CNN for each region. The features have been assigned learned weights to provide attention, considering that each region contributes differently to pain expression.

Many experiments are conducted in controlled lab settings, however some researchers like Semwal et al. [40] focused on pain assessment in uncontrolled environments. They developed a shallow CNN with 3 convolutional layers, achieving high multi-class classification performance comparable to deeper pre-trained models. Later, they conducted a second study with[41] a more complex deep framework, yielding results comparable to other models, such as GoogleNet and VGG. Lee and Wang[42] explored the intensive care unit (ICU) setting, where it is common to have partially occluded faces. They developed a 4-layer CNN combined with an extreme learning machine (ELM) network for the final estimation.

Recurrent neural networks (RNNs) is the second most popular technique[12]. This method is perfect for sequential data analysis, which uses temporal data like video sequences to analyze data in a way that greatly improves pain assessment. RNNs are able to obtain temporal patterns in body movements, vocal intonation, and facial expressions by sequentially processing frames of video data. This allows to better determine how pain changes over time[43].

Apart from these deep learning techniques, other neural network architectures have also been used in pain detection, thus expanding the range of approaches. Within the family of recurrent neural networks (RNNs), it is possible to find gated recurrent unit networks (GRU)[44][45] and long-term memory networks (LSTM)[16][46], which have become best known for their effective modeling of sequential data. These networks work well when analyzing time series data, such as video sequences used in pain assessment, where it is important to capture temporal dependencies. When making predictions, bidirectional LSTMs (biLSTMs)[47][48] incorporate an additional level of complexity by considering both past and future context. Over time, this reciprocal approach has proven useful in identifying subtle patterns in pain expression. These alternative neural network architectures provide researchers with flexible tools in their search for reliable and accurate pain detection models, even if they are not always considered leading deep learning techniques.

Deep learning techniques have shown promising results, providing increased accuracy and robustness in pain assessment[12]. These models have demonstrated the ability to capture complex patterns from diverse data sources, improving the accuracy of pain detection. In addition, they may be able to detect pain in real time[49][50], which could have a significant positive impact on clinical judgment and patient care.

There are still problems to be solved, such as the need for large amounts of labeled data to train deep learning models[51], which may be restrictive in some medical applications. In addition, there are still problems with the interpretability and transparency of the models, especially when dealing with situations where it is crucial to understand the reasoning behind pain detection judgments.

3.2 Federated Learning

Federated learning has become an important development in machine learning, especially when decentralization and data privacy are essential. This literature review highlights the evolution and importance of federated learning by summarizing its main advances and uses.

The origin of federated learning systems come from the concern for preserving data privacy and security in machine learning. It has been developed in response to the difficulties presented by the traditional model of centralized data aggregation. Organizations were hesitant to share their data sets due to legal restrictions, privacy concerns and data ownership issues, especially when it came to sensitive data such as medical data. Early developments in this field focused on decentralized methods that allowed multiple people to work together to jointly create machine learning models without sharing raw data[13].

The concept of federated learning has been first introduced by researchers at Google[52], whose groundbreaking study laid the foundation for the field. The paper offered a novel solution to the basic problem of data privacy in machine learning. It proposed a decentralized training model that eliminated the need to exchange raw data and allowed multiple parties to work together to collaboratively build machine learning models. The key has been to build a global model by aggregating local model updates from each individual data source. Since then, this revolutionary concept has stimulated plenty of research and real-world applications, enabling collaborative, privacy-preserving machine learning across multiple domains while keeping confidential data secure and under the control of data owners[13].

Federated learning has become an influential trend in the medical field[53][54][55] due to its innovative approach to patient data privacy and the dispersed nature of healthcare data. With this novel method, multiple healthcare organizations, from hospitals to research centers, can collaborate on machine learning initiatives without exchanging raw patient data. Federated learning enables the development of predictive models, such as those for automatic pain recognition, in an industry where data security and privacy are paramount, while ensuring that private patient information remains secure within each institution[13]. There are different types of federated learning[13], each adapted to the distribution of data available. Some of the most important ones are reviewed below.

Horizontal federated learning occurs when several entities or clients share distinct data instances, but with comparable characteristics[56]. In the context of automated pain recognition, each institution shares similar characteristics related to pain assessment (facial expressions), but the patients are different. This way, the actual patient data remains secure within each participating healthcare entity but they can collaborate to improve the accuracy of automated pain recognition models for their patients[56][57]. In the context of this project, due to the availability of several databases with similar characteristics but different patients this is the type of federated learning that will be used. Figure 1 shows a diagram of the the method explained.

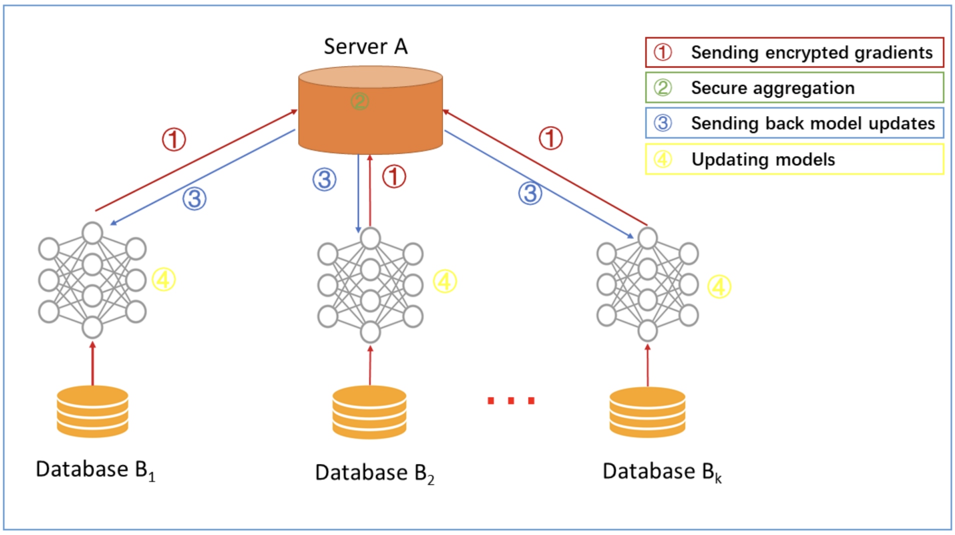


Figure 1. *Architecture for a horizontal federated learning system.[56]*

On the other hand, when entities have different characteristics, but share common data instances, vertical federated learning is used. This type of federated learning works well in situations where several organizations, for example, different healthcare providers want to create a complete patient model. In this case, the patients would be the same but the characteristics being measured would be different. This enables collaborative medical research, where multiple organizations can jointly analyze patient data without compromising the privacy of specific feature data[56]. This type of federated learning would fit this project if, for example, pain detection were performed for the same group of patients, but across related features, such as physiological indicators or textual patient records.

These various forms of federated learning prioritize data security and privacy while providing adaptive responses to a wide range of decentralized machine learning problems. Once the type of federated learning has been decided based on the data, there are several methods or algorithms that can be used to perform the actual learning. One of the most important ones is Federated Averaging (FedAvg)[52][58].

Fundamentally, FedAvg uses a simple but highly effective method to guarantee both model cooperation and data privacy. FedAvg's fundamental principle is to create a global model without exchanging sensitive raw data by averaging model updates from participating clients. Each client, who is frequently represented by organizations, healthcare facilities, or individual devices, starts the process by initializing the global model. This model is used as the foundation for federated training and is frequently selected according to the type of machine learning task, like language modeling or image classification. The training occurs locally on each client's data without any direct data exchange. Every client's local model is updated during the training process using data from that particular dataset.

These updates are specific to the distinct data distribution and characteristics of each client and represent the knowledge obtained from the local data. The clients send updates to a central server about their model parameters after completing local training. Only the updates are being sent in this exchange; raw data is not being sent. This measure guarantees that confidential data stays on the client's end and is not shared. The model updates that are received from the different clients are combined at the central server.

The updates are averaged in this aggregation, which can be represented as a conventional arithmetic mean. The end product is a new global model that combines the insights from every client that took part. The updated model for the subsequent federated training cycle is the recently aggregated global model. The updated global model protects the security and privacy of individual data sources while reflecting the collective knowledge of all clients. By performing these steps repeatedly over several rounds, the global model is able to improve in accuracy and refinement with the combined efforts of all clients. It is possible to established the number of rounds or base it on convergence criteria[52].

This is the approach that will be used in this project because it is the one most commonly used in other research projects. This implies that there are more data and results available that will allow to develop the project in a more efficient way and to compare the results obtained. Figure 2 shows a diagram of how the above explained process works.

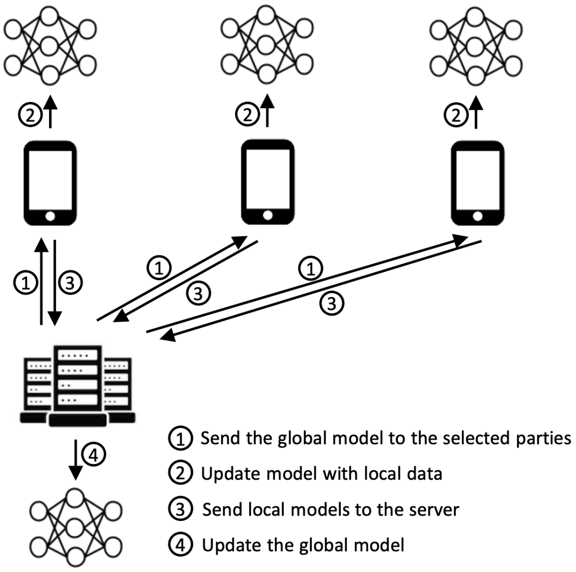

②
②
②
① ③
①
③
①
③
④
① Send the global model to the selected parties
② Update model with local data
③ Send local models to the server
④ Update the global model

Figure 2. *Federated Averaging (FedAvg).[13]*

Other federated learning approaches include Federated transfer learning which uses pre-trained models to accelerate learning across dispersed data sources. Federated Transfer Learning uses the fundamentals of federated learning to securely modify pre-trained models to fit particular local data sources[59]. It is a method for securely and privately sharing knowledge from a centrally trained model to decentralized entities for adaptation and fine-tuning to their local data. This method is particularly useful in domains where there is insufficient labeled training data[60][61].

Federated reinforcement learning combines reinforcement learning with federated learning, allowing cooperative model training across decentralized data sources while maintaining data security and privacy. In reinforcement learning, an agent learns to make a series of decisions to maximize a cumulative reward. Many entities or parties maintain their local datasets and model parameters as part of the Federated Reinforcement Learning process. Using its own data, each entity trains its local reinforcement learning agent to simulate sequential decision-making in that particular environment. By interacting with their local data, these agents pick up knowledge while adhering to confidentiality and privacy regulations[62][63].

Federated Proximal (FedProx) is an optimization method that builds on the fundamental ideas of Federated Averaging (FedAvg) by including proximal terms. The objective of this algorithmic modification is to improve the convergence, robustness, and performance of the model in federated learning environments. The main idea behind FedProx is to include a regularization term, also known as the proximal term, in the optimization objective. During training, this term is used to impose restrictions or penalties on the model parameters. FedProx is especially helpful in situations where convergence speed or model accuracy are crucial because it introduces these constraints, which promote quicker convergence and better model performance[64].

3.3 Pain detection using federated learning

The following is a review of some of the projects that have used federated learning for automatic pain recognition and will serve as a reference for this project.

Rudovic et al.[65] make a significant contribution using Personalized Federated Deep Learning (PFDL) for pain estimation to the UNBC-McMaster Shoulder Pain Database. This way each patient's information remains private but all clients feed the global model. In their study, they demonstrate that PFDL performs comparably or even outperforms traditional centralized and FL algorithms while simultaneously enhancing data privacy. In their study, they use several models in order to make a comparison between them. The base model is the basic centralized deep learning (BCDL) model, trained on non-target subjects. Then, they use a centralized deep learning model on target subjects (CDL) to evaluate the impact of centralized training specifically on target subjects.

Also, traditional federated deep learning (FDL) is used so that all parties can learn in a collaborative manner. Locally trained models (LDL) are adapted to each subject and, the results of a randomly initialized CNN (RND) model provide a point of comparison without model training.

Similarly, N. Tobis[66] in his study also uses the UNBC-McMaster database to detect pain automatically in a federated manner. In his study he uses a CNN architecture to classify the dataset and uses ResNet-50 and VGG16 architectures to find the best model architecture. The author conducts several experiments. In one of them he tries to start the model with random parameters, in another one he tries to use pre-trained parameters and in another one he tries to release the data per session. That is, he does not assume that all the information is available from the beginning but that it is generated sequentially in regular therapy sessions. Like the previous author, the model that achieves the best accuracy is federated penalisation.

These last two studies will be used as the basis for this project due to their similarity in the tasks and results to be obtained.

3.4 Differential Privacy

Differential privacy (DP) provides a fundamental framework in the field of data privacy, designed to allow the analysis of datasets while safeguarding the privacy of individual contributions [67][68]. This concept has become increasingly important considering growing concerns about data privacy, where the need to use data to obtain information must be balanced with the individual's right to privacy.

The beginnings of DP date back to the pioneering work of Dwork et al.

[67], who laid the foundations for a formal approach to privacy-preserving data analysis. This approach has since evolved, becoming the cornerstone of modern privacy-preserving techniques.

The foundation of distributed learning (DL), the -differential privacy model, introduces a parameter to assess privacy loss, guaranteeing that an algorithm's result does not heavily rely on the data of a single individual. Moreover, -differential privacy provides a subtle extension that permits a very low likelihood of privacy infringement. The Laplace and Gaussian processes, which modify query results on datasets to conceal specific data points while preserving the overall utility of the data, are the foundation of the mathematical architecture of DP [69][70][71].

3.4.1 Differential Privacy in a Federated Learning system

A major step forward in resolving privacy issues in decentralised machine learning models is the integration of differential privacy (DP) in a federated learning (FL) system. The goal of this integration is to improve privacy without sacrificing the ability to learn collaboratively across several servers or devices.

Federated Learning with Differential Privacy (DP) is a privacy-preserving approach that combines federated learning and the principles of differential privacy. Within this framework, decentralized entities use their own data for local training, which results in model updates that enhance performance.

To maintain the privacy and anonymity of individual contributions, controlled noise is introduced during model updates through the application of differential privacy mechanisms. The privacy of each data source is preserved as these perturbed updates are then safely combined to produce a global model. Federated Learning with Differential Privacy is especially useful in situations where data privacy is a top priority, like healthcare or financial data applications, because of its iterative process, which permits collaborative model development while maintaining strong privacy guarantees [72].

Through the use of federated learning (FL), conventional centralised machine learning may be transferred to a distributed architecture, enabling devices to cooperatively learn a common model while maintaining localised training data. Because raw data is not sent to a central server, this method naturally resolves some privacy concerns. Nevertheless, FL alone cannot fully protect against all privacy risks, such as inference attacks on modifications to shared models. This calls for the use of strong privacy-preserving techniques, of which differential privacy (DP) is a standout example [69].

The application of DP in FL intends to mitigate privacy risks by ensuring that shared updates to the model do not reveal sensitive information about the data of any of the participating devices. The application of DP in FL intends to mitigate privacy risks by ensuring that shared updates to the model do not reveal sensitive information about the data of any of the participating devices. The challenge is to implement DP in a way that balances privacy with model accuracy and learning efficiency [70].

This involves adapting DP mechanisms, such as Laplace or Gaussian, to the federated context, by adding noise to model updates before they are aggregated. In the context of this thesis, the Gaussian mechanism will be the one used since it is the most commonly used [69][71].

There are special difficulties with integrating DP into FL, especially when it comes to balancing privacy guarantees with model performance. When privacy is protected by adding noise, the accuracy of the learnt model may suffer, particularly in situations where the data is highly heterogeneous and scattered [70].

Research in this field involves the development of optimisation techniques to minimise this impact, such as adaptive noise [73] scaling and secure aggregation protocols that improve privacy without significantly compromising model quality.

4 Data

A major challenge in the development of robust automatic facial expression detection systems is the limited amount of representative data[74]. Researchers and other specialists have worked to develop some databases such as the UNBC-McMaster Shoulder Pain Expression Archive database, the BioVid Heat Pain database and the EmoPain database. By doing this, they aim to overcome data scarcity to improve the availability of relevant data for model development. These three databases will be used in this project. A description of each of them is given below.

4.1 UNBC-McMaster shoulder pain expression

The UNBC-McMaster Shoulder Pain Expression Dataset[74] is a publicly available dataset designed for research in pain expression recognition. It has been created to facilitate the development and evaluation of computer vision and machine learning algorithms for automatic recognition of pain expressions. This database has been developed by researchers at McMaster University and the University of Northern British Columbia.

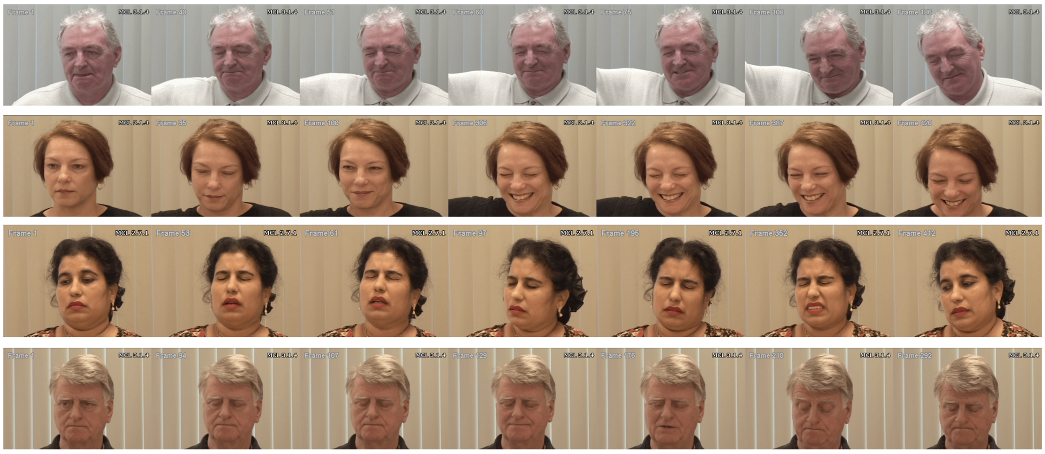


Figure 3. *Examples of some of the sequences from the UNBC-McMaster Pain Shoulder Archive.[74]*

To create this database, they videotaped the faces of participants (suffering from shoulder pain) as they performed a series of active and passive range-of-motion tests on their affected and unaffected limbs on two separate occasions. Each frame of this data has been encoded in AU by certified FACS coders, they self-reported and observer measures have been also taken at the sequence level. The investigators have made publicly available a portion of the database, which includes: 1) 200 video sequences containing spontaneous facial expressions, 2) 48,398 FACS coded frames, 3) associated pain frame-by-frame scores and sequence-level self-report and observer measures, and 4) 66-point AAM landmarks[74].

A total of 129 participants (63 men, 66 women) with self-identified shoulder pain problems were recruited from three physiotherapy clinics and through advertisements on the McMaster University campus. This varied group included students and community members with a variety of shoulder pain diagnoses, such as arthritis, bursitis or tendinitis, and more than half of them were taking pain relievers. However, there are FACS annotations for only 25 participants, therefore, for this project, only these 25 participants will be used.

Participants were subjected to eight standardized range-of-motion tests, which included abduction, flexion, and internal and external rotation of each arm, both actively and passively. These tests were performed in a laboratory setting, recorded by two Sony digital cameras that captured facial expressions. Verbal pain descriptors and visual analog scales (VAS) were available to help participants rate the pain experienced during each test. Independent observers also rated pain intensity from the recorded videos, demonstrating high inter-observer reliability and concurrent validity.

Each of the tests in this study involved video data extraction and subsequent Facial Action Coding System (FACS) coding[75]. FACS categorizes facial expressions into 44 individual action units (AUs). This study particularly focused on actions potentially related to pain, including brow-lowering (AU4), cheek-raising (AU6), eyelid tightening (AU7), nose wrinkling (AU9), upper-lip raising (AU10), oblique lip raising (AU12), horizontal lip stretch (AU20), lips parting (AU25), jaw dropping (AU26), mouth stretching (AU27), and eye-closure (AU43).

These actions were coded with five intensity levels (A-E) by certified FACS coders, and each action has been coded frame by frame, with a fourth certified FACS coder reviewing the coding. The system uses an Active Appearance Model (AAM) approach[76][77], which employs AAM to track facial features and extract visual information.

In the data distribution provided, there are 66 landmarks identified by the AAM for each image. Active Appearance Models (AAMs) have been proven to be a good method for aligning a predefined linear shape model that also has a linear variation in appearance, with a previously unseen source image containing the object of interest. In general, AAMs adjust their shape and appearance components using a gradient-descent search, although other optimization methods have been employed with similar results[78].

This dataset is relevant because it focuses on the specific domain of shoulder pain, allowing researchers to explore pain-related facial expressions in a controlled and standardized context.

4.2 The BioVid Heat Pain

The BioVid Heat Pain is an available dataset designed for research in pain expression recognition. It has been created to improve automatic pain monitoring, improve treatment and address the inherent subjectivity in the perception of pain. This database has been developed by researchers at Otto-von-Guericke University and University of Ulm[79].

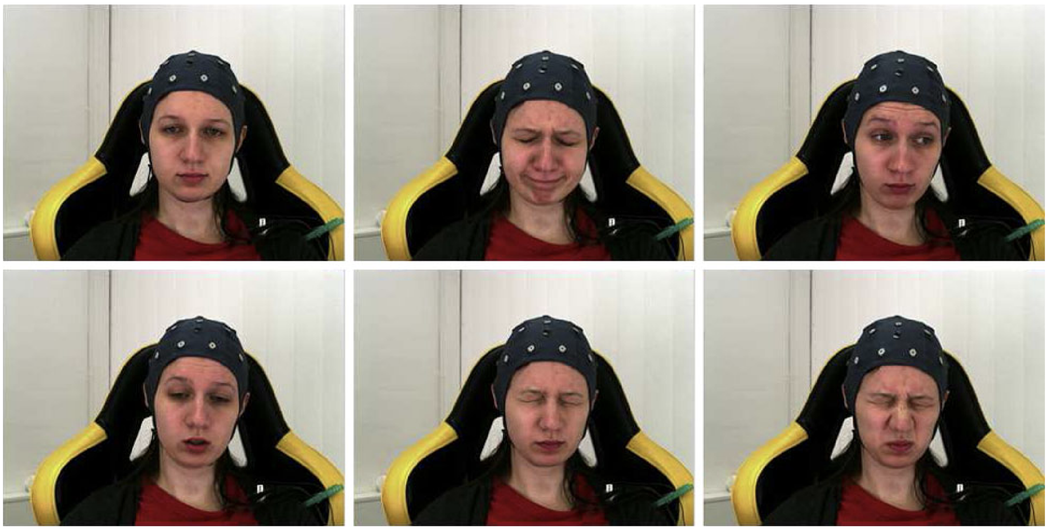


Figure 4. *Face samples from the BioVid Heat Pain database.[80]*

The BioVid heat pain database has been created based on a study with 90 participants in three age groups (18-35, 36-50 and 51-65 years), each consisting of 15 men and 15 women. Pain was induced experimentally by means of a thermode applied to the right arm. Experiments were recorded with video cameras and physiological sensors, including synchronised AVT Pike F145C cameras (1388 x 1038 pixels, 25 Hz) placed in front of and to the sides of the participants. They captured depth information with a Microsoft Kinect camera (640 x 480 pixels, approximately 30 Hz). The physiological data included skin conductance level (SCL), electrocardiogram (ECG), electromyogram (EMG) of three pain-related muscles and electroencephalogram (EEG).

Individual pain thresholds and pain tolerance levels of each participant were determined prior to recording. Throughout the experiment, there were four levels of pain, including the lowest and the highest. In the first part, each pain level was stimulated 20 times in random order. In the second part, the participants expressed pain and basic emotions, and they were shown pictures and videos to trigger spontaneous emotions. In the third part, they repeated the pain stimulation of the first part, with EMG facial electrodes attached[81].

Since this project is based on analyzing pain expressions, only data from the first part will be used. Their facial expression analysis is based on a set of landmarks, which are automatically extracted. For each image in the video stream, the face is first found using the Haar-like feature detector cascade of Lienhart et al.[82]. Within the face region the eye detector cascades trained by Castrillón et al.[83] and the mouth corner detector cascades of Panning et al.[84] are applied.

To identify false detection, the candidates are compared with an estimate given by a generic face model that is placed inside the bounding box of the face. Based on the points obtained for the centre of the eyes and the corner of the mouth, the remaining reference points are found, as described by Niese et al[85]. The upper and lower lip points are determined using a colour-based lip segmentation approach. It is based on the normalised green channel histogram for the mouth region of interest. The segmentation contour is also used to redefine the points of the corners of the mouth, as the results are more accurate than the detector in most cases.

Each of the eyebrow points is selected from a line segment perpendicular to the eye axis by finding the maximum peak of the vertical gradient. The eye axis is also used to compensate for head rotation in the head rotation plane by rotating the detector input image of the next frame. For each image frame a set of distance and gradient features are extracted. They are selected to capture several pain-related facial actions that have been identified by several previous studies[86].

These actions include lowering the eyebrows, squeezing the eyelids, closing the eyes, closing the eyelid, closing the eyes, lifting the cheeks and upper lip, wrinkling the nose, and stretching and opening the mouth. In the BioVid heat pain database, several pain levels have been considered, ranging from no pain to pain levels up to 4. However, within this project, only two different categories will be considered: pain and non-pain. This will simplify the pain assessment process. Taking into account the categories included in this database, all pain ranges PA1 to PA4 will be considered as ”pain” in a binary decision.

As with the previous database, this one allows progress in the field of automatic pain detection and assignment to help those who cannot utter[81].

5 Methodology

The methodology section describes the overall approach adopted in this study. After preparing the datasets in a pre-processing phase, several experimental setups have been conducted to evaluate the results with these different setups.

5.1 Preprocessing

This phase consists of pre-processing the data from the UNBC and Biovid databases and ensuring a common, standardised framework for both. This phase involves four steps: (1) data selection, in which the dataset has been selected and refined; (2) feature extraction, during which OpenFace has been used to extract the relevant facial features; (3) data relabelling, standardising the labels across all datasets to maintain consistency; and (4) data modifications, adjusting the image formats to be compatible with the models. This step is necessary in order to conduct the various experiments and obtain reliable and accurate results.

5.1.1 Data Selection

- **UNBC Database:** For the UNBC database, all 25 available subjects and all images have been used. No sub-selection of images has been done.

- **Biovid Database:** Originally, the Biovid database consisted of 87 subjects. However, as recommended in the database description, 20 participants who never showed an expression of pain have been excluded [87]. In addition 13 participants have withdrawn their consent to use their images. Considering these two exclusions and that some participants are in both groups, the final number of subjects is 60. This database has four levels of pain in addition to the baseline level.

  In the description of the database it is also stated that the first levels of pain never produce a pain expression [88]. For this reason, the lower pain levels have been eliminated as well. This leaves only the baseline level and the two higher levels of pain. Finally, it can be seen that the pain expression is triggered around the fourth second of the video [88]. So a selection of the images has been made to leave only those corresponding to the second 4 to 5.5. This corresponds to frames between 88 to 138. This selection allows the labels to be more accurate, since not all of the frames in the whole video show an expression of pain.

This sub-selection has been done in order to reduce the amount of data, since it initially affected the performance of the model and to balance the number of pain and non-pain images.

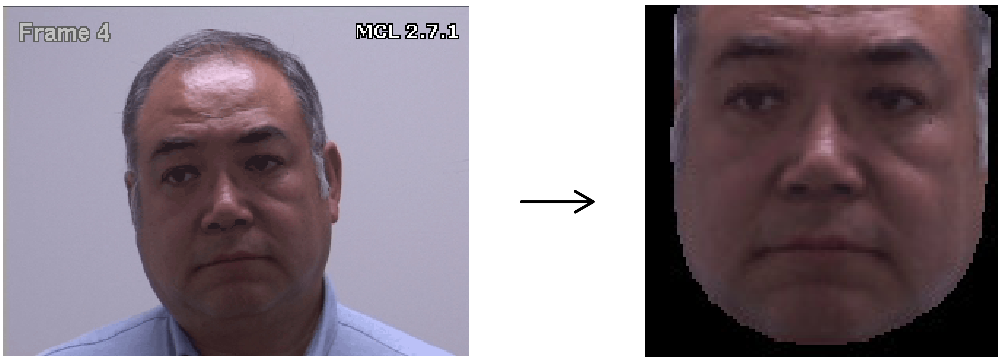
5.1.2 Feature Extraction

The OpenFace tool [89] has been used, in order to extract the faces from the images and thus have only the relevant part of the images. OpenFace is a facial behavior analysis toolkit. The use of this tool is due to its previous use in other similar studies as well as OpenFace's strong ability to isolate and capture facial expressions with high accuracy [90][91][92].

- **UNBC Database:** For the UNBC database, all images have been processed by OpenFace and an equivalent number of colour images containing only the face have been obtained. Analysing the output, it has been observed however that in some cases other elements of the images such as hair or part of the hands were detected as an additional face in the same image. Therefore, a manual inspection of all the images has been made to eliminate the incorrect ones.

- **Biovid Database:** For the Biovid database, the process has been similar, only in this case the input are videos (5.5 seconds long) instead of images. Each video generated 138 images. In this case no errors have been detected.

Figure 5 shows an example from the UNBC database of how the original image has changed after extracting the face with OpenFace.



(a) Raw image

(b) Processed image

Figure 5. *Feature Extraction with OpenFace*

5.1.3 Data Relabeling

The original data sets used different methods for labeling pain, so a standardized approach is necessary to ensure consistency throughout the study. Therefore, a binary labeling system has been adopted: "0" for no pain and "1" for pain.

- **UNBC Database:** For the UNBC Database, images have been categorized according to the presence of facial expressions indicative of pain, utilizing the Prkachin and Solomon Pain Intensity Scale (PSPI) [93].

  $$PSPI_{\text{pain}} = AU4 + \max(AU6, AU7) + \max(AU9, AU10) + AU43$$

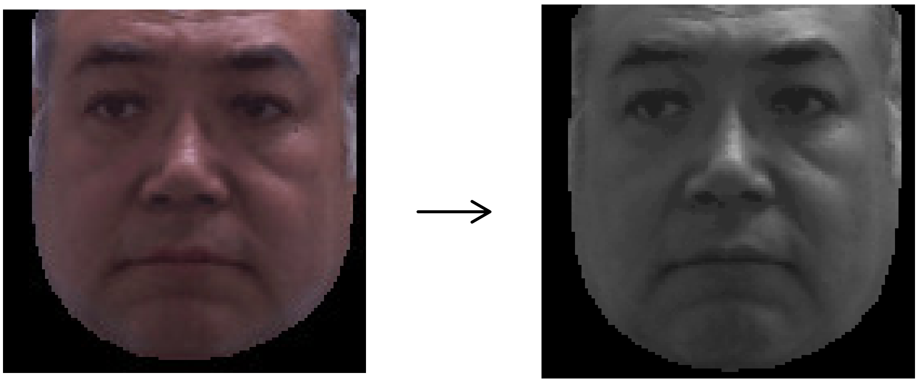
  This method involves this specific equation for assessing pain expressions, where images demonstrating any of the action units (AUs) associated with pain are classified as 'pain'. Conversely, images that do not display these particular action units are labeled as 'no pain.' This approach ensures a consistent methodology for labeling, based on the visible manifestation of pain through facial expressions.

- **Biovid Database:** In the case of the Biovid database, the relabeling process has been done based previous studies, which categorized the baseline and levels 1 and 2 as “no pain” [88]. Although, levels 1 and 2 have been later excluded from the analysis [88], levels 3 and 4 have been labeled as “pain”. Therefore the baseline level has been relabeled with '0' and the levels 3 and 4 with '1'. It should also be taken into account the previous data selection where only the last second and a half of the video has been used for analysis. That is why it is safe to relabel the whole level 3 and 4 with a '1'.

5.1.4 Data Modifications

Aligned with the requirements of the centralized model and various federated learning scenarios, it has been necessary to convert the images to grayscale. This process, has been essential to ensure compatibility with the single-channel input expectation of the model, and also contributed to computational efficiency by reducing the complexity of image processing. The exception has been when using the convnext(tiny) model, which apart from the grey-scale conversion also required resizing the images to 224x224 pixels.

Figure 6 shows an example from the UNBC database of how the processed image looks like after it is converted to grayscale.



(a) Processed image

(b) Grey-scaled image

Figure 6. *Converting RGB image into grayscale*

5.2 Experimental setup

The experimental setup of this study is based on a series of scenarios designed to investigate pain detection based on facial expression, taking advantage of advanced machine learning techniques. As a starting point, a codebase developed by AshwinRJ [94], which has been made publicly available on GitHub [94], has been used. This code includes among other things a centralized module and a federated learning module.

To ensure reliability and generalisability of results across diverse subsets of data, all models including the centralized and federated learning configurations have been executed with 5-fold cross-validation. This methodological approach allows for rigorous evaluation and benchmarking. In addition, the F1 score has been chosen as the primary metric for all evaluations.

This is because it effectively captures both precision and recall and because it is a widely used metric in similar studies.

Following the same reasoning, according to previous studies, Stochastic Gradient Descent (SGD) has been used as the optimiser [65][66]. Other parameters, such as the learning rate (0.01) or the number of local epochs (10 epochs) have also been chosen according to the literature and have been left fixed in order to compare the performance of the models [95][96].

5.2.1 Centralized model

In this study, the centralized model serves as a benchmark to assess the performance of the federated learning model. In the context of this study, it would be the equivalent of both clients training their models individually. This allows an assessment of whether the federated model has any advantages. This model has been implemented using a convolutional neural network (CNN) architecture from a publicly available code base[94]. The choice of a CNN for the centralized model is in line with its proven effectiveness in image classification tasks, providing a simple but powerful approach for initial performance evaluation.

Figure 7 shows an illustrative diagram of what this model looks like. In this case, two medical centres are using only their own data and training their model individually.

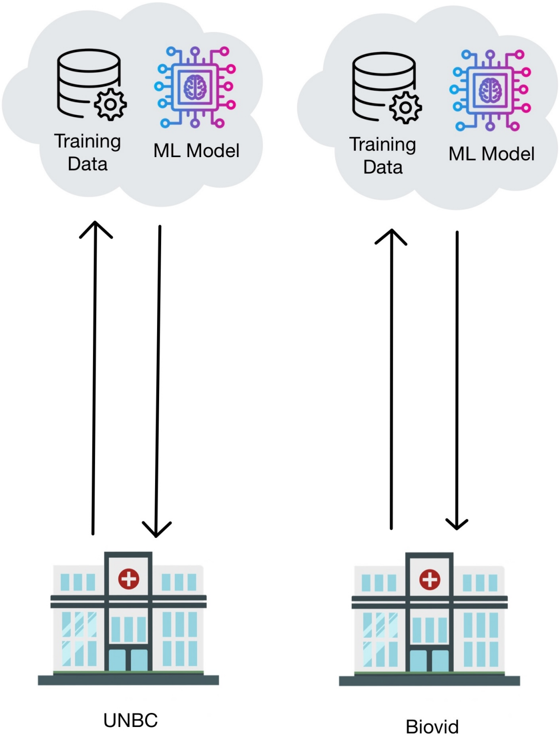


Figure 7. *Illustration of Centralized Model*

This architecture is built using PyTorch's neural network module (nn.Module). The network is initialised with two convolutional layers: conv1 and conv2. To overcome overfitting, conv2 is followed by a dropout layer. There are then two fully connected layers: fc1 converts the convolved output into a 50-unit vector and fc2 maps it to the specified class outputs. In the forward method, data goes through ReLU activations and maximizes the grouping of similar features after each convolution, with a final log softmax function applied after fc2 to produce a class probabilities distribution, ensuring non-linear processing throughout the network.

5.2.2 Local model

To further explore the effectiveness of the centralized model, individual evaluations within the UNBC and Biovid datasets have been conducted. In this case, each subset of data has been tested separately to get a deeper insight into the performance of the model. Specifically, five subgroups with five patients each have been created. The model has been trained on four of them and tested on the remaining one. This experiment has been carried out in each of the five subgroups. This scenario allows to know what is the performance at the individual level of each patient with the centralised model.

Figure 8 shows an illustrative diagram of what this model would look like. In this case, two medical centres are using only their own data and training their model individually. However, they are only using part of their data for training and testing.

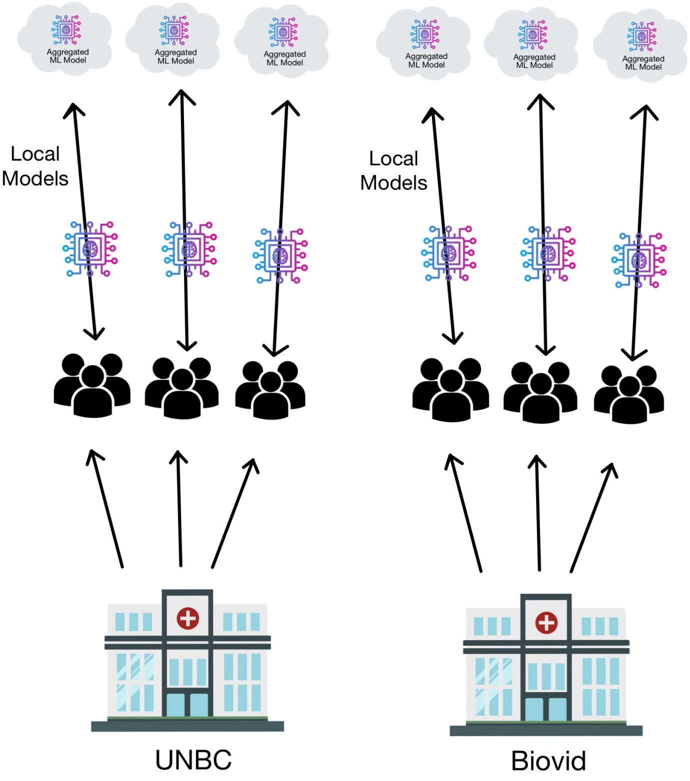


Figure 8. *Illustration of Local models*

5.2.3 Federated learning model using a single database

For the federated learning scenario, the same source code and CNN architecture given [94] has been used as a starting point[94]. The basis of this setup has been kept the same and is therefore very similar to the centralized model. This allows the comparison between the two scenarios to be more accurate.

However, a significant modification has been introduced in the user group assignment mechanism compared to the source code. User groups refer to the different clients that are part of the federated scenario. For example, if there are two medical centres, each medical centre would be a different user group. Within the context of a single medical centre, one or more patients would be part of a user group. In this source code, the allocation of user groups is random. This is because it is intended to be used in generic databases such as CIFAR [97] or MNIST [98] where it is not relevant whether certain images belonged to one specific user group.

In the case of this study, each patient's data had to remain in the same user group, either individually or by grouping several patients under one user group. It could not happen that the images of the same patient were in more than one user group. This strategy has been vital for two main reasons: first, it preserved the authenticity and consistency of the learning process by preventing the model from being trained with overlapping or redundant data from the same patients in different user groups. Second, it respected the principles of patient privacy and data security, which are fundamental considerations in medical and health research.

To achieve this, an additional field has been added to the input data, a patient id. This way, when creating user groups, it could be done on the basis of these ids and thus avoid an id being in more than one user group.

This model has been trained using Federated Learning Averaging. As mentioned in the literature, this technique has been chosen because it is the most commonly used among other similar studies. For detailed information on how this model works, see section 2.2.

Figure 9 shows an illustrative diagram of what this model would look like. In this case, two medical centres are using only their own data and training their model individually. They are also dividing the patients either in groups or on an individual basis. In other words, we are simulating federated learning setup within database.

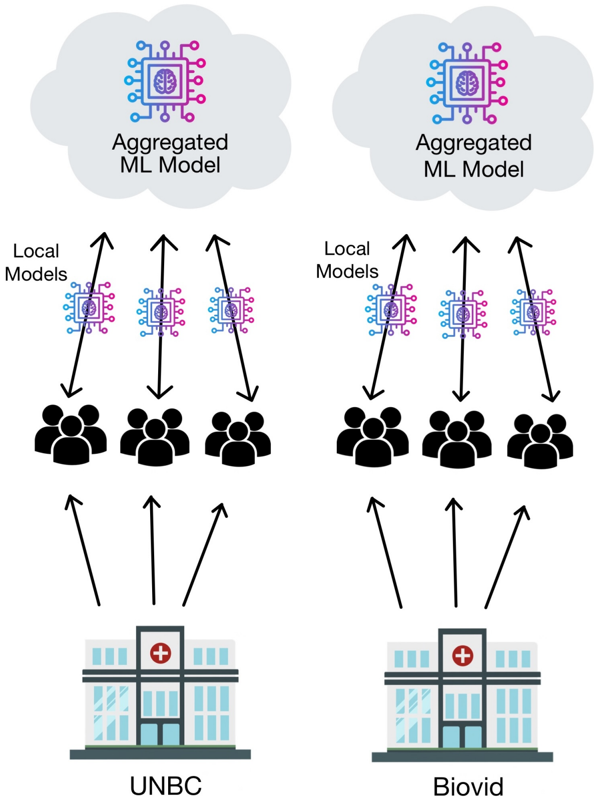


Figure 9. Illustration of Individual federated learning model

5.2.4 Federated learning model with two databases

This scenario has been conducted based on the existing federated learning approach used in the previous scenario. In this case, the aim is to use both databases in such a way that both can benefit from each other without compromising data privacy.

In this scenario, there are two distinct domains although both are pain expression databases. One dataset captures participants experiencing shoulder pain, while the other refers to pain induced by electrical stimulation. These datasets were recorded with different cameras, include different age groups, etc.

The goal of this experiment is to understand whether using the information from another, unfamiliar dataset can further improve the performance. For this purpose, each database has been assigned to a different user group. Therefore, using only two groups of users. This ensures that within the execution environment the raw data of each database is never mixed with the raw data of the other database.

Figure 10 shows an illustrative diagram of what this model would look like. In this case, two medical centres are sharing the aggregated ML model while training their local model. Their training data is kept within each organisation and only the output of the local model is shared.

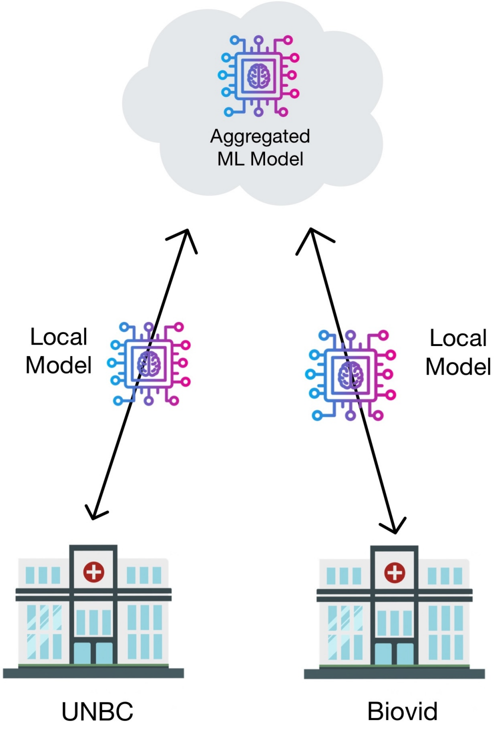


Figure 10. *Illustration of individual federated learning model with two databases*

This scenario required two main changes in the implementation. First difference is in the data ingestion. In this case, the dataloader had to be prepared to accept two databases and to keep the training data, test data and user groups separate. On the other hand, each database required an individualised test process, so the test process has been duplicated in such a way that this phase has been completed first for one database and then for the other, thus being able to return individualised learning results for each of the databases.

5.2.5 Federated learning model with differential privacy

As discussed in the literature review (Section 2.4), the federated learning setup may not be private enough. This is the reason why differential privacy has been used within the context of federated learning.

This method consists of adding noise after training the local model but before training the global model. The Gaussian mechanism is the one most commonly used in other studies, which is why it has been decided to use it to create this setup [69][71].

The Gaussian mechanism is a key technique in differential privacy, it adds noise to the weights of the model in a way that preserves privacy. However, it should be noted that there is a trade-off between privacy and performance. Adding more noise for stronger privacy protection reduces the accuracy of the model.

Gradient clipping, privacy budget, and noise are important parameters in differential privacy, ensuring that the privacy guarantees are maintained while preserving the utility of the data or model.

Gradient clipping limits the magnitude of gradients during training to control their sensitivity. The privacy budget represents the aggregate privacy loss incurred over multiple iterations and measures the level of privacy protection provided by each iteration. The noise level represents the scale of Gaussian noise added to gradients to achieve differential privacy. The formula for adding Gaussian noise to gradients is:

$$\text{noisy\_gradient} = \text{gradient} + \text{noise\_level} \times \text{random\_noise}$$

Where noisy_gradient is the resulting noisy gradient, gradient is the original gradient of the model parameters, noise_level is the scale of the Gaussian noise, and random_noise is a sample from a Gaussian distribution with mean 0 and standard deviation 1. Adjusting the noise level allows for balancing privacy protection and model utility [99].

In federated learning, the local models are updated at each client and the updates are aggregated by a central server to update the global model. Let w_t represent the global model weights at iteration t. Each client i computes the updated weights w_t^i based on its local data. The central server aggregates these updates using Federated Averaging (FedAvg) as follows:

$$w_{t+1} = \frac{1}{N} \sum_{i=1}^{N} w_t^i$$

where N is the number of participating clients [71].

To integrate differential privacy into the federated learning framework, there are several steps to follow.

First, each client i computes its local model update (gradient) Δw_i based on its local dataset. This involves performing standard gradient descent or another optimization method on the client's local data to obtain the gradient update.

Then, the gradients need to be clipped. To bound the sensitivity of the gradients and limit their magnitude, the local gradients are clipped to a predefined norm C:

$$\Delta w_i \leftarrow \Delta w_i \cdot \min\left(1, \frac{C}{\|\Delta w_i\|}\right)$$

This step ensures that no single update can have an unbounded influence on the aggregated model. The parameter C determines the maximum allowed gradient norm. If the norm of Δw_i exceeds C, it is scaled down to C; otherwise, it remains unchanged.

Afterwards, the noise scale is calculated. For different values of a noise scale, the Renyi Differential Privacy (RDP)[100] value is calculated to ensure the desired level of privacy. The formula for RDP is given by:

$$\mathrm{RDP}(\alpha, q, \sigma) = \frac{1}{\alpha - 1} \log\left(\sum_{k=0}^{\alpha} \binom{\alpha}{k} (1-q)^{\alpha-k} q^k e^{\frac{k(k-1)}{2\sigma^2}}\right)$$

Where:
- α is the order of RDP, controlling the trade-off between privacy and accuracy.
- q is the subsampling rate, representing the probability of selecting a given client's data in each round.
- σ is the noise scale, influencing the amount of Gaussian noise added to the

gradients.

Then it is chosen the maximum σ within the specified range that satisfies the condition RDP $> \epsilon$, where ϵ is a predefined privacy parameter, indicating the level of privacy desired.

After determining the appropriate noise scale σ, each client adds Gaussian noise to its clipped gradient before sending it to the server:

$$\Delta w_i^{DP} = \Delta w_i + \mathcal{N}(0, \sigma^2 I)$$

Where:
- Δw_i^{DP} is the differentially private gradient.
- $\mathcal{N}(0, \sigma^2 I)$ is the Gaussian noise with mean zero and variance σ^2.

This step ensures that the gradients shared by clients are differentially private, preventing any single data point from having a significant impact on the model.

Then, clients send their noisy gradients Δw_i^{DP} to the central server. In this communication step, the updates (not the raw data) are shared with the server, preserving data locality and privacy.

Finally, the server aggregates the received noisy gradients to update the global model:

$$\Delta w^{DP} = \frac{1}{N} \sum_{i=1}^{N} \Delta w_i^{DP}$$

where N is the number of participating clients. This aggregation step combines the noisy updates from all clients, averaging them to form the new global model update [71][100].

Figure 11 shows an illustrative diagram of what this model looks like. In this case, two medical centres are sharing the aggregated ML model while training their local model. Their training data is kept within each organisation and only the output of the local model is shared. However, after updating the local model and before updating the global model Gaussian noise is added to enhance the privacy of the patients.

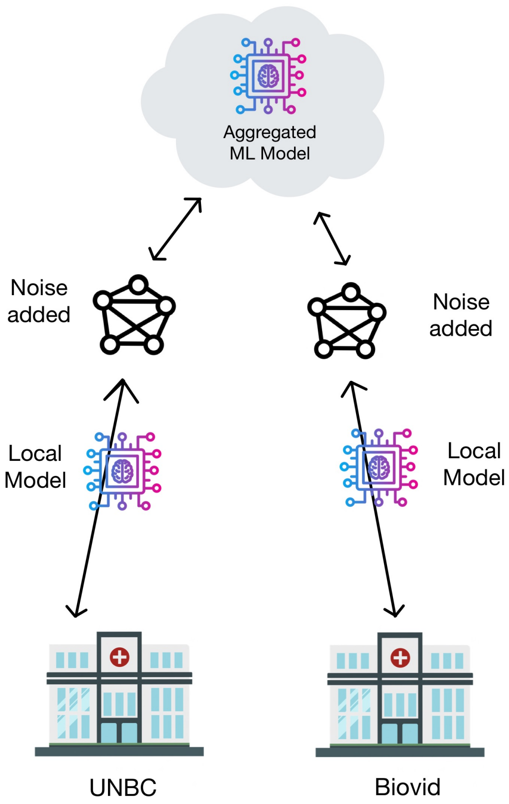


Figure 11. *Illustration of federated learning model with differential privacy*

6 Results

In this section, the findings of the research into the performance of the different models are presented.

6.1 Centralized model

The Centralized model has been executed using an early stopping mechanism to determine the optimal number of epochs, preventing overfitting and unnecessary computations. For the UNBC dataset, the early stopping criterion has been met at 10 epochs, indicating that this is the point at which the model's performance on the validation set is maximized without further generalization improvements.

Similarly, for the Biovid dataset, the best number of epochs before the model ceased to show performance gains has been identified to be 6. This approach not only optimized the training time but also ensured that each model is trained just to the point of maximal efficacy, as evidenced by the validation data.

In addition to the CNN presented in the methodology (Section 5.2.1), other architectures such as ResNet50 [101] and ConvNext(tiny) [102] have been used to try to achieve the best performance in the centralized model.

For ResNet50, the default pretrained weights have been utilized [101]. ResNet50, despite its deep structure designed for feature learning through residual connections, underperformed. This suggests that the model may be too complex for the specific features of the datasets used.

On the other hand, ConvNext(tiny) showed slightly lower results than the base CNN, although comparable, indicating that, for the datasets in question, a very complex architecture does not equate to a significant performance improvement.

For ConvNext(tiny), the default pretrained weights have been utilized [102]. Due to the relatively inferior results of these models and the long execution time, it has been decided not to extend the tests to the Biovid database. Collectively, these results reinforce the baseline CNN as the model of choice for the rest of this study, balancing effectiveness and efficiency for facial expression-based pain detection tasks.

Table 1 shows a summary of the results. This table include the results of all the experiment with each of the architectures and for both databases. It can be seen that the best result for the UNBC database is obtained at epoch 10 with a f1 score of 0,64. For the biovid database, the best result has been obtained at epoch 6 with a f1 score of 0,59.

Database	**Model**	**Best Epoch**	**F1 Score**
UNBC	Baseline CNN	10	**0,64**
UNBC	ResNet50	5	0,45
UNBC	ConvNext_tiny	8	0,58
Biovid	Baseline CNN	6	**0,59**

Table 1. *Model Performance on Facial Expression-Based Pain Detection*

In order to better understand the decision-making process of the model, Gradient-weighted Class Activation Mapping (Grad-CAM) [103][104] has been used. This provides a “heat map” overlaid on the input image. This heat map highlights regions that significantly influence the model's predictions, effectively visualising areas of interest within the image that lead to a particular classification result.

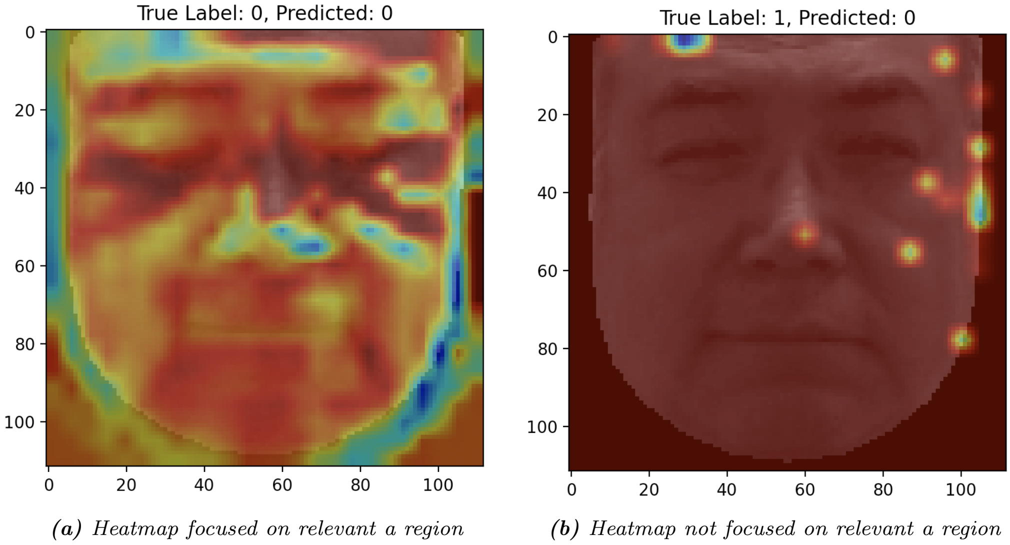
Grad-CAM uses the gradients going into the final convolutional layer of the model to capture the importance of each neuron in the prediction of the target class. By applying a ReLU function to the linear combination of these gradients and activation maps, it is ensured that only positive influences on the class prediction are visualised.

This approach not only improves the interpretability of the inner functioning of the model, but also serves as an analysis tool to verify that the model is focusing on the relevant features of the images. It is especially useful for identifying cases where the model may be making decisions based on spurious patterns or noise, rather than on meaningful content [103][104].

Across the experiments, Grad-CAM successfully generated distinct heatmaps for a significant portion of the dataset, demonstrating the model's attention to key features, while some images showed less pronounced heatmaps, indicating areas where the model's confidence or focus is not as strong. These visualisations help to understand and trust the model, as they provide a window into the neural networks' prediction process.

Some examples of images obtained are shown in figure 12. Figure 12.a shows a case where the heatmap has correctly focused on regions significant to detect pain and has made a correct prediction. In figure 12.b, an example is shown in which the attention has been drawn to a non-significant region and therefore the prediction is incorrect.

Figure 13 shows the overall heatmap obtained with the centralized model. This heatmap has been obtained by averaging all the heatmaps for this patient. So it reflects where the model has been looking when making decisions, whether the expression is pain or not.



(a) Heatmap focused on relevant a region *(b) Heatmap not focused on relevant a region*

Figure 12. *Sample heatmaps obtained from the centralized model*

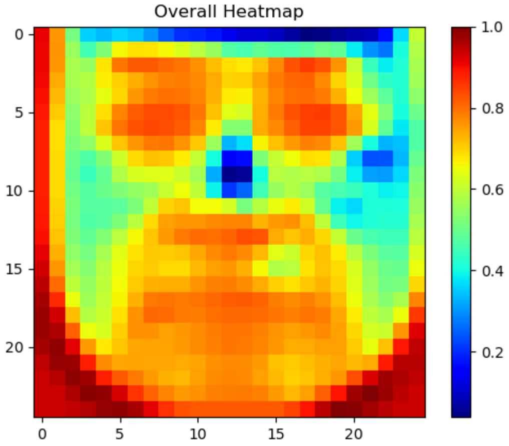


Figure 13. *Overall heatmap of Centralized Model*

In this heatmap it can be seen that the model has focused on important regions of the face. When relabeling the database, the Prkachin and Solomon Pain Intensity Scale (PSPI) [93] has been used to determine which images show pain expressions. Specifically, it has been determined that the Action Units showing a pain expression are AU4: brow-lowering, AU6: cheek-raising, AU7: eyelid tightening, AU9: nose wrinkling, AU10: upper-lip raising and AU43: eye-closure. All these Action Unitis can be seen marked in red or orange on this heatmap. The only Action Unit that is less present among the relevant areas is AU9, which is marked in blue and green.

Although in this case it can be seen that there is a focus on these important regions and has avoided some non-relevant regions, there is still a lot of influence from the region under the face, which is just noise. This suggests that the model, for this patient, has not learned to look only at the important regions for detecting pain. These results also reinforce the need to extract the face from the rest of the image. Since the model may focus on areas outside the face, if the image has more noise, the regions outside the face may be weighted more heavily than the face itself.

6.2 Local model

The local model has been executed using the same number of epochs as when the entire database was available in order to be able to compare the results better. These are 10 epochs in the case of the UNBC databse and 6 in the case of the Biovid database.

As expected the results have been worse than when using the whole database. This is manly due to the significantly reduced volume of training data. This result underlines the importance of the volume of training data for optimal model performance.

As explained in Section 5.2.2, the data has been divided into 5 subgroups, and for each subgroup four users have been used for training and one for testing. A cross-validation has been done with all users in each subgroup so that all have been used as a train user and as a test user. Table 2 shows the average results obtained for each subgroup.

Training Fold	**F1 Score**
Subgroup 1	0,49 (0,015)
Subgroup 2	0,53 (0,09)
Subgroup 3	0,43 (0,069)
Subgroup 4	0,67 (0,093)
Subgroup 5	0,41 (0,052)
Average	**0,50**

Table 2. *Individual results on UNBC Dataset (10 Epochs)*

For the UNBC database, the f1 score of the model with the whole database is 0.64. Although in the case of subgroup 4, the results are better, the f1 score of the rest of the subgroups and the mean is lower. Upon further analysis of the data for this subgroup, it has been founded that in the case of two patients (106-nm106 and 108-th108), there are fewer images showing pain expressions. With fewer images of pain expressions available for these patients, the model may have overfitted the prevalent class, which explains its higher accuracy in classifying expressions, making the average higher.

The same process has been done with the Biovid dataset. Table 3 shows the average results obtained for each subgroup.

Training Fold	F1 Score
Subgroup 1	0,52 (0,035)
Subgroup 2	0,53 (0,078)
Subgroup 3	0,49 (0,048)
Subgroup 4	0,47 (0,047)
Subgroup 5	0,51 (0,072)
Average	**0,50**

Table 3. *Individual results on Biovid Dataset (8 Epochs)*

For the Biovid database, the f1 score of the model with the whole database is 0.59. In this case, no subgroup achieves better results than those obtained when the model was trained on the entire database. This demonstrates the importance of having access to a larger amount of data when training models.

In the next section the federated approach is examined, where a similar case is studied by dividing the patients into user groups. In this way, the same privacy setup will be maintained but the user groups will be able to benefit from each other.

6.3 Federated learning model using a single database

Several tests have been conducted with the federated learning model within each database in order to see how medical entities could benefit from federated learning within their own organisation. Some of the experiments include separating each subject into a user group so that individual patient data within each organisation would remain private.

Other tests have also been done by combining several users within the same group, in an attempt to improve prediction accuracy. This way, although some patients' data will be combined with each other, will be in a limited way.

In addition, as in previous cases, the early stopping mechanism has been used to determine the optimal number of epochs. In this case, depending on the number of users, it varies between 3 and 6 epochs. Using more epochs does not lead to an improvement in learning.

Table 4 shows the best results obtained for each of the setups for the UNBC database.

Number of users	3 epochs	5 epochs
4	0,56	**0,60**
5	0,58	0,57
20	0,55	0,55

Table 4. *Model Performance on UNBC database*

The best result is obtained when users are divided into 4 groups and after 5 epochs. The f1 score in this case is 0,6. This result is a bit lower than when using the centralised model (f1 score 0,64), but the difference is not very large. Considering the privacy gained by training the models in this way, this approach may still be beneficial.

However, these data can also be compared with the local model discussed in section 6.2. In that case the data available for training were also being limited. In this scenario the results are higher than with the local model. So if within the same medical centre, the data cannot be freely shared across the whole centre and has to be divided for example by doctors, this approach is certainly better, allowing better results to be obtained by limiting the data sharing.

Table 5 shows the best results obtained for each of the setups for the Biovid database.

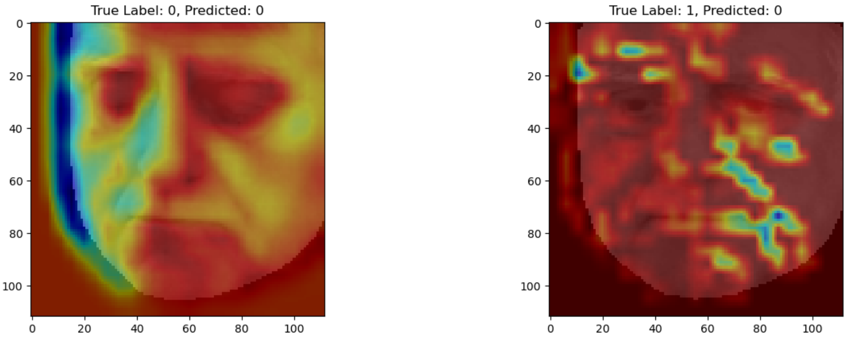
Number of users	3 epochs	4 epochs	5 epochs	6 epochs
4	0,54	0,54	0,53	0,55
6	**0,57**	0,5	0,56	0,55
48	0,53	0,55	0,52	0,52

Table 5. *Model Performance on Biovid database*

The best result is obtained when users are divided into 6 groups and after 3 epochs. The f1 score in this case is 0,57. This result is a slightly lower than when using the centralised model (f1 score 0,59), but the difference is not significant. In this case it is definitely worth the privacy gained compared to the performance lost.

Moreover, if local models are once more taken into account, this is another added advantage. Again, if data could not be shared freely within a medical centre, individual patients would still benefit from this setup. Furthermore, looking at the setup where each patient has been added to a single user group, the accuracy has hardly decreased (f1 score 0.55) compared to the centralised model (f1 score 0,59) or the federated model with 6 patients per user (f1 score 0,57). In this case the data for each patient has remained separate from the rest offering a high level of privacy for each patient without decreasing the predictive capacity of the model.

Once again, to make this data more explainable, a heatmap has been obtained to identify the regions on which the model has focused to make decisions. Figure 14.a shows a case where the heatmap has correctly focused on the significant regions and has made a correct prediction. In figure 14.b, an example is shown in which the attention has been drawn to a other regions or ignored some key ones and therefore the prediction is incorrect.



(a) *Heatmap focused on relevant regions*

(b) *Heatmap not focused on relevant regions*

Figure 14. *Heatmap obtained from the centralized model*

Figure 15 shows the overall heatmap obtained with the federated model.

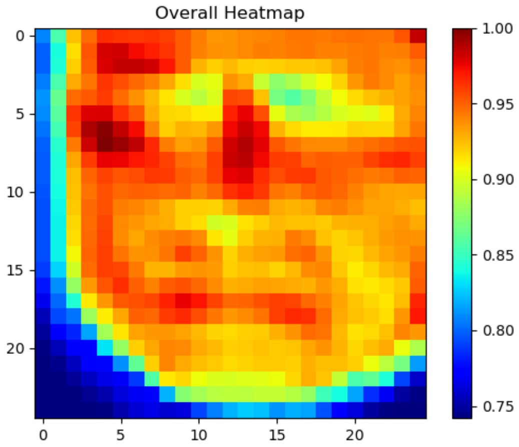


Figure 15. *Overall heatmap of Federated Model*

In this case, in contrast to the centralized model, the focus is primarily on relevant facial regions, disregarding the off-face area. This indicates a more effective learning process. However, when examining the Action Units used to identify pain expressions [93] (specifically, AU4: brow-lowering, AU6: cheek-raising, AU7: eyelid tightening, AU9: nose wrinkling, AU10: upper-lip raising, and AU43: eye-closure), they appear less pronounced here, especially those related to the eye region (AU4, AU7, and AU43). On the other hand, AU9 and AU10, which had less significance in the heatmap from the centralized model, have more weight in this case.

6.4 Federated learning model with two databases

The Federated learning model has been executed using an early stopping mechanism to determine the optimal number of epochs. In this case it has been observed that after 5 epochs, learning does not improve. In order to compare the results with the centralised model, the CNN architecture described in section 5.2.1 has been used, and the rest of the parameters have also been maintained for the same reason.

Table 6 shows the results obtained. In this case, only two users have been used, so that each database has been assigned to a single user group without the data being mixed between them.

	UNBC	**Biovid**
Fold 1	0,53	0,59
Fold 2	0,61	0,55
Fold 3	0,44	0,61
Fold 4	0,66	0,49
Fold 5	0,58	0,54
Average	**0,56**	**0,56**

Table 6. *Model Performance on both databases*

The results show an average f1 score of 0.56 for both databases. In the case of UNBC, the results are worse compared to the centralised model (f1 score 0.64) but in the case of Biovid the difference in performance is not so large (f1 score 0.59). These results are in line with expectations, since not all data is available a drop in performance was expected.

However, looking again at the results of the local model, this process is still beneficial. In a potential scenario where complete data from a medical center is not accessible, but a portion wishes to contribute to a global model for enhanced predictions, segments of data from various medical centers can be combined to improve overall results.

6.5 Federated learning model with differential privacy

Within this scenario, the most complex issue is to find the best balance between adding enough noise so that the data remains as private as possible while making the model as accurate as possible.

In the literature there is no consensus on particular values for the gradient clipping and the privacy budget [69][71]. The best values should be found through experimentation but, by default, some experiments take as a starting point a gradient clipping of '1' and privacy budget '1'. That is why in this study this value has been used as a starting point and then other tests have been carried out with different values. However, it should be noted that with values in this range no significant differences have been obtained.

There have been two rounds of experiments. First, a fixed value of noise has been added to the gradients to test their effect. The noise values varied between 0.01 and 2. After adding a noise of 2, the accuracy drops regardless of the other parameters. With lower noise and values of gradient clipping and the privacy budget between 0 and 1, the values are very similar. However, it still achieves reasonably good values considering that extra noise is being added. So for certain tasks it could be a trade off that could be assumed.

Table 7 show the results obtained. This table show the different experiments carried out with trying different parameters.

Noise	Gradient Clipping	F1 Biovid	F1 UNBC
0.01	1	0,52	0,51
0.5	1	0,45	0,47
1	1	0,31	0,45
1	0,5	0,37	0,45
2	1	0,19	0,37

Table 7. *Model Performance with Differential Privacy*

Initially a very low noise level (0.01) has been used to determine how it changes in this case. The results show that the accuracy drops slightly but not significantly. Subsequently a much higher noise level (2) has been added. In this case the accuracy dropped significantly. This is why the following experiments took place with values in this range.

In a second round of experiments, the noise has been calculated dynamically based on the gradient clipping and the privacy budget [105]. The noise values vary between 0,25 and 1.99 depending on the Gradient Clipping and the Privacy Budget. Since the noise range is the same as when it was a fixed value, the results obtained are also very similar.

Table 8 shows the results obtained.

Noise	Gradient Clipping	Privacy Budge	F1 Biovid	F1 UNBC
0,96 - 1.96	1	0,5	0,29	0,41
1,05 - 1.99	1	1	0,39	0,41
0,51 - 0,97	0,5	0,5	0,36	0,45
0,25 - 0,47	0,5	1	0,45	0,46

Table 8. *Model Performance with Differential Privacy*

The experiments have been conducted with varying values of gradient clipping, ranging from 0.5 to 1, and privacy budget, ranging also from 0.5 to 1. As the results obtained are very similar to the previous experiments, no further experiments have been carried out.The results indicate that a higher privacy budget corresponds to a lower addition of noise, which translates into lower privacy guarantees.

Alternatively, a lower privacy budget leads to more noise being added, which improves privacy protection. Gradient clipping limits the magnitude of the gradients during training, primarily to mitigate the exploding gradient problem. While gradient clipping itself does not directly dictate the amount of noise added, it can indirectly influence noise levels by affecting sensitivity calculations. From the results, it can be seen that higher gradient clipping values result in more noise being added to the gradients.

These two rounds of experiments show relatively low F1 score values. Only in the case where the noise is set to 0.01 the result does not drop as much. So, as expected, there is indeed a trade off between performance and privacy. Furthermore, as discussed in other sections the data is very unbalanced so it is possible that the results are better than expected as the model has a tendency to classify according to the dominant class.

6.6 Summary of the results

In this study, the goal has been to investigate the suitability of federated learning to improve the performance of automated pain detection systems. A summary of the results obtained is shown in Table 9.

	UNBC	**Biovid**
Centralised model	**0,64**	**0,59**
Local models	0,50	0,50
Federated (1DB)	0,60	0,57
Federated (2DB)	0,56	0,56
Federated (DP)	0,52	0,51

Table 9. *Summary of the results obtained*

Table 9 shows that the best result is obtained by the centralized model for both the UNBC and Biovid databases; however, the federated model with one or two databases obtains better results than the local models, which reinforces the idea that more data generally obtain better results. While the model with differential privacy does not exhibit notably inferior performance, this result shows the case with less noise added, as the noise or privacy budget increases the results drop drastically. For more detailed data see Table 7 and Table 8.

7 Discussion and Conclusion

The subsequent analysis delves into a detailed exploration of the research questions and their corresponding findings according to the results shown in Table 9.

1. When a single database is used, how does the model trained using federated learning compare to centralized and local models?

   Intuitively, centralised learning is expected to outperform federated learning approaches. Centralised learning works with a consolidated dataset, allowing for more thorough model training and optimisation compared to federated learning, which relies on decentralised data sources with limited access to global information. However, the literature presents mixed results [65] on the comparative performance of centralised and federated learning.

   Some experiments show that federated learning outperforms centralised approaches [66][106], while others show the opposite [107] [108]. This lack of consensus underlines the complexity of the comparison and the influence of various factors on model performance.

   The disparity in performance between centralised and federated learning can be due to several factors such as the characteristics of the dataset and the model architectures used. In particular, it is shown in the literature that smaller databases tend to favour centralised learning [107].

   In addition, the balance of the data distribution within the dataset plays an important role, as more balanced datasets tend to give better results with federated learning [107][109]. In this case the data is considerably unbalanced to the benefit of the “no pain” label, so this may have affected the results.

This research shows that centralized learning exhibited superior performance for both the UNBC and BioVid databases compared to the federated learning model (For UNBC, F1 score of 0,64 for centralized model compared to F1 score of 0,64 for federated learning model. For Biovid, F1 score of 0,59 for centralized model compared to F1 score of 0,57 for federated learning model).

The difference in performance has been more pronounced for the UNBC database, which is considerably smaller compared to the BioVid dataset. This discrepancy can be attributed to the limited size of the UNBC dataset, which inherently restricts the effectiveness of federated learning. Furthermore, the unbalanced nature of the data may have exacerbated the performance gap, as federated learning thrives on more evenly distributed data.

Regarding the local models, compared to the centralized and the federated model with a single database, it can be seen that both the centralized and the federated models obtain better results.

The local models achieved an F1-score of 0.5 for both databases. In contrast, the centralized model obtained an F1-score of 0.64 for UNBC and 0.59 for Biovid. Furthermore, the federated model with a single database obtained an F1-score of 0.6 for UNBC and 0.57 for Biovid.

In both cases, it is demonstrated that having access to a larger number of data helps to improve performance. For both local model (Section 5.2.2) and federated learning model (Section 5.2.3), the data remained private at the individual level or within small groups of patients.

However, the federated learning framework, which divides the data into user groups, allows for more personalized model training while preserving data privacy. As a result, federated learning outperforms local approaches in this scenario, offering improved performance at the individual level.

The advantages of federated learning become particularly evident in organizations where not all data can be freely shared across the entire organization.

As demonstrated in Table 9, the performance of federated learning models surpasses that of local models when data availability is limited within a medical center. By dividing the data into smaller groups, such as by doctors or patient groups, federated learning enables effective model training while respecting data privacy constraints.

This approach allows organizations to leverage their data more efficiently, obtaining better results while minimizing the need for extensive data sharing across the organization.

2. How do the models trained using multiple databases (UNBC and BioVid) in a federated learning setup compare to the centralized models trained on these models separately?

   In evaluating the performance of models trained using multiple databases (UNBC and BioVid) within a federated learning framework against centralized models trained separately on each database, some key observations can be made.

   Firstly, the centralized model exhibits superior performance compared to the federated model utilizing both databases. Specifically, the centralized model achieves an F1-score of 0.64 for UNBC and 0.59 for Biovid, surpassing the corresponding scores of the federated model with two databases, which are 0.56 and 0.56, respectively.

   The federated model with a single database obtains higher F1-scores compared to the federated model with two databases. For instance, for UNBC, the federated model with one database achieves an F1-score of 0.60, outperforming the federated model with two databases (F1-score of 0.56). Similarly, for BioVid, the F1-score of the federated model with one database (0.57) surpasses that of the federated model with two databases (0.56).

Despite the advantage of centralized models in terms of overall performance, it's important to note that federated learning with two databases still outperforms local models (F1-score of 0,5 for both databases). This suggests that while centralized models may offer higher performance, federated learning facilitates better utilization of data resources compared to local models.

This becomes particularly relevant when considering the practical application of such models in real-world scenarios. For example, in a healthcare environment, where different departments or units within a hospital may have varying degrees of willingness or ability to collaborate, federated learning offers a flexible solution. Rather than requiring centralization of data from all units, which can pose logistical or privacy issues, federated learning enables collaboration between specific units, departments or medical centers while preserving data privacy.

This means that even if only some parts of a medical center decide to collaborate with other parts of another medical center, federated learning can still be applied effectively. This decentralized approach not only improves the scalability and efficiency of model learning, but also ensures compliance with data privacy regulations and addresses potential data privacy concerns.

3. How does performance vary when using differential privacy in the models compared to a less private model?

   When incorporating differential privacy into models, a key consideration is how performance varies compared to less private models. It was anticipated that privacy-preserving techniques, by introducing additional noise, might lead to lower performance [69][71]. Furthermore, as discussed in Section 6.5, the results could have been influenced by data imbalance and the model's tendency to favor the dominant class.

   When examining the performance variation associated with the incorporation of differential privacy (DP) into the models, the results from the Table 9 shows the impact of privacy-preserving techniques. The introduction of differential privacy with the least noise leads to a slight decrease in performance compared to the non-private federated models.

   Specifically, for the UNBC database, the F1-score decreases to 0.52, while for the Biovid database, it decreases to 0.51. As mentioned in the results section this result shows the case with less noise added, as the noise or privacy budget increases the results drop drastically.

Interestingly, the decrease in performance with very low noise levels is not significantly pronounced. This suggests that, in scenarios where privacy is very important, adding minimal noise may be a viable strategy to improve privacy without substantially compromising performance.

The decision to balance privacy with performance is context dependent and should be studied on a case-by-case basis. Through iterative experimentation, models can be adapted to achieve the optimal balance between privacy preservation and performance optimization.

In summary, this thesis has investigated the performance of centralized, federated and local models in the context of medical data analysis. The results show that while centralized models show superior overall performance, federated learning offers a flexible and efficient solution, especially in scenarios where data sharing is restricted or privacy is paramount.

Furthermore, it has been observed that federated learning with multiple databases can outperform local models, highlighting the potential of federated approaches to effectively leverage distributed data resources.

Furthermore, the exploration of differential privacy revealed that although the incorporation of privacy-preserving techniques may result in a slight decrease in performance, the impact is relatively small, especially with minimal noise levels. This suggests that differential privacy can be a valuable tool for improving data privacy without significantly compromising model performance.

Overall, the study underscores the importance of considering various factors, such as data distribution, model architecture, and privacy requirements, when designing and implementing machine learning models for healthcare applications. By carefully balancing performance and privacy considerations, organizations can develop robust and effective models that meet both technical and ethical standards.

8 Limitations and Future work

In future work, it would be beneficial to explore alternative federated learning techniques beyond the conventional federated averaging approach such as federated transfer learning [60], federated reinforcement [62] or federated proximal [64] which offer promising avenues for improving model performance and convergence in decentralized environments [110]. Investigating these techniques could provide insights into novel approaches for federated learning across diverse application domains.

Although two databases have been used in this study, it would be interesting to use more to see how this affects the results and make the models more generalisable. Intuitively it can be said that the more databases there are, the more likely it is that the model will be able to make better predictions. So future work could do these same experiments using more databases.

Addressing the challenge of unbalanced data distribution within federated learning settings is important for improving model performance and fairness. Future research efforts could focus on developing robust techniques for managing unbalanced data, such as data re-sampling, class weighting, or specialized loss functions. These techniques could mitigate the impact of data imbalances on model training and ensure equitable representation of all classes or categories.

Finally, further research should be done on other methods of enhancing the privacy of the models. This could entail further exploration of differential privacy, investigating alternative privacy mechanisms beyond Gaussian noise addition, and exploring innovative approaches like collaborative learning [111][112]. By exploring these alternative privacy techniques and comparing their effectiveness in federated learning settings, research can contribute to the development of more robust and privacy-preserving machine learning solutions.

References

[1] Kyle Vader Srinivasa N. Raja; Daniel B. Carr; Milton Cohen; Nanna B. Finnerup; Herta Flor; Stephen Gibson; Francis J. Keefe; Jeffrey S. Mogil; Matthias Ringkamp; Kathleen A. Sluka; Xue-Jun Song; Bonnie Stevens; Mark D. Sullivan; Perri R. Tutelman; Takahiro Ushida. The revised international association for the study of pain definition of pain: concepts, challenges, and compromises. *PAIN*, 161(9):1976–1982, 2020.

[2] J Takala P Mäntyselkä; E Kumpusalo; R Ahonen; A Kumpusalo; J Kauhanen; H Viinamäki; P Halonen. Pain as a reason to visit the doctor: a study in finnish primary health care. *PAIN*, 89(2-3):175–180, 2001.

[3] Brizendine EJ. Cordell WH; Keene KK; Giles BK; Jones JB; Jones JH. The high prevalence of pain in emergency medical care. *The American Journal of Emergency Medicine*, 20(3):165–169, 2002.

[4] J. Gregory and L. Mcgowan. An examination of the prevalence of acute pain for hospitalised adult patients: A systematic review. *Journal of Clinical Nursing*, 25(5-6):583–598, 2016.

[5] Sigridur Gunnarsdottir Sigridur Zoëga; Herdis Sveinsdottir; Gisli H. Sigurdsson; Thor Aspelund; Sandra E. Ward. Quality pain management in the hospital setting from the patient's perspective. *Pain Practice*, 15(3):236–246, 2015.

[6] A. Mitchell and B. J. Boss. Adverse effects off pain on the nervous systems of newborns and young children: a review of the literature. *Journal of Neuroscience Nursing*, 34(5), 2002.

[7] F. Tennant. The physiologic effects of pain on the endocrine system. *Pain and Therapy*, 2(2), 2013.

[8] M. E. Lynch. The need for a canadian pain strategy. *Pain research management*, 16(2), 2011.

[9] Andrew Moore Henry McQuay and Douglas Justins. Treating acute pain in hospital. *British Medical Journal*, 314(7093), 1997.

[10] Anjan Kumar Ghosh Sourav Dey Roy; Mrinal Kanti Bhowmik; Priya Saha. An approach for automatic pain detection through facial expression. *7th International conference on Intelligent Human Computer Interaction, IHCI 2015*, 84:99–106, 2016.

[11] Sandra Merkel Keela Herr; Patrick J Coyne; Margo McCaffery; Renee Manworren. Pain assessment in the patient unable to self-report: position statement with clinical practice recommendations. *Pain Manag Nurs*, 12(4):230–250, 2011.

[12] Stefanos Gkikas Manolis Tsiknakis. Automatic assessment of pain based on deep learning methods: A systematic review. *Computer Methods and Programs in Biomedicine*, 231), 2023.

[13] He Bingsheng Li Qinbin; Wen Zeyi Wen; Wu Zhaomin Wu; Hu Sixu; Wang Naibo; Li Yuan; Liu Xu. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3347–3366, 2021.

[14] K. D. Craig. The social communication model of pain. *Canadian Psychology / Psychologie canadienne*, 50(1):22–32, 2009.

[15] Reza Kharghanian, Ali Peiravi, and Farshad Moradi. Pain detection from facial images using unsupervised feature learning approach. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 419–422, 2016.

[16] Pau Rodriguez, Guillem Cucurull, Jordi Gonzàlez, Josep M. Gonfaus, Kamal Nasrollahi, Thomas B. Moeslund, and F. Xavier Roca. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE Transactions on Cybernetics*, 52(5):3314–3324, 2022.

[17] Philipp Werner, Ayoub Al-Hamadi, Kerstin Limbrecht-Ecklundt, Steffen Walter, Sascha Gruss, and Harald C. Traue. Automatic pain assessment with facial activity descriptors. *IEEE Transactions on Affective Computing*, 8(3):286–299, 2017.

[18] Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald Traue. Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. *Proceedings of the British Machine Vision Conference*, 09 2013.

[19] Joy Egede, Michel Valstar, and Brais Martinez. Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation. *IEEE International Conference on Automatic Face Gesture Recognition*, 01 2017.

[20] Karan Sikka. Facial expression analysis for estimating pain in clinical settings. *ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction*, 11 2014.

[21] Domenica Delgado, Bradley Lambert, Patrick McCulloch, Michael Moreno, Joshua Harris, and Andrew Robbins. Validation of digital visual analog scale pain scoring with a traditional paper-based visual analog scale in adults. *Journal of the American Academy of Orthopaedic Surgeons. Global Research Reviews*, 2, 05 2018.

[22] Mathias Haefeli and Achim Elfering. Pain assesment. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, 15 Suppl 1:S17–24, 02 2006.

[23] Kenneth Prkachin and Patricia Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139:267–74, 06 2008.

[24] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, pages 1–1, 10 2019.

[25] Prkachin K. M. Grunau R. E. Craig, K. D. The facial expression of pain. *Handbook of pain assessment*, page 117–133, 2011.

[26] Kenneth D. Craig. The facial expression of pain better than a thousand words? *APS Journal*, 1(3):153–162, 1992.

[27] Marcella Saccò, Michele Meschi, Giuseppe Regolisti, Simona Detrenis, Laura Bianchi, Marcello Bertorelli, Sarah Pioli, Andrea Magnano, Francesca Spagnoli, Pasquale Gianluca Giuri, Enrico Fiaccadori, and Alberto Caiazza. The relationship between blood pressure and pain. *The Journal of Clinical Hypertension*, 15(8):600–605, 2013.

[28] Eulália Pinheiro, Fernanda Campbell, Pedro Montoya, Cleber Luz, Marion Nascimento, Clara Ito, Manuela Silva, David Santos, Silvia Benevides, José Garcia Miranda, Katia Sá, and Abrahão Baptista. Electroencephalographic patterns in chronic pain: A systematic review of the literature. *PloS one*, 11:e0149085, 02 2016.

[29] Astrid Terkelsen, Henning Mølgaard, John Hansen, Ole Andersen, and Troels Jensen. Acute pain increases heart rate: Differential mechanisms

during rest and mental stress. *Autonomic neuroscience : basic clinical*, 121:101–9, 09 2005.

[30] Henrik Pedersen. Learning appearance features for pain detection using the unbc-mcmaster shoulder pain expression archive database. *International Conference on Virtual Storytelling*, 07 2015.

[31] Robert Gatchel, Yuan Peng, Madelon Peters, Perry Fuchs, and Dennis Turk. The biopsychosocial approach to chronic pain: Scientific advances and future directions. *Psychological bulletin*, 133:581–624, 07 2007.

[32] Barry Kussman, Christopher Aasted, Meryem Yücel, Sarah Steele, Mark Alexander, David Boas, David Borsook, and Lino Becerra. Capturing pain in the cortex during general anesthesia: Near infrared spectroscopy measures in patients undergoing catheter ablation of arrhythmias. *PLOS ONE*, 11:e0158975, 07 2016.

[33] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[34] Joy O. Egede, Siyang Song, Temitayo A. Olugbade, Chongyang Wang, Amanda C. De C. Williams, Hongying Meng, Min Aung, Nicholas D. Lane, Michel Valstar, and Nadia Bianchi-Berthouze. Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 849–856, 2020.

[35] Saandeep Aathreya Sidhapur Lakshminarayan, Saurabh Hinduja, and Shaun J. Canavan. Three-level training of multi-head architecture for pain detection. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 839–843, 2020.

[36] Ruijing Yang, Xiaopeng Hong, Jinye Peng, Xiaoyi Feng, and Guoying Zhao. Incorporating high-level and low-level cues for pain intensity estimation. In *International Conference on Pattern Recognition*, pages 3495–3500, 08 2018.

[37] Dong Huang, Zhaoqiang Xia, Lei Li, Kunwei Wang, and Xiaoyi Feng. Pain-awareness multistream convolutional neural network for pain estimation. *Journal of Electronic Imaging*, 28:043008, July 2019.

[38] Xuwu Xin, Xiaoyan Lin, Shengfu Yang, and Xin Zheng. Pain intensity estimation based on a spatial transformation and attention cnn. *PLOS ONE*, 15:e0232412, 08 2020.

[39] Conghui Li, Zhaocheng Zhu, and Yuming Zhao. Saliency supervision: An intuitive and effective approach for pain intensity regression. *Lecture Notes in Computer Science*, 2018.

[40] Ashish Semwal and Narendra Londhe. S-panet: A shallow convolutional neural network for pain severity assessment in uncontrolled environment. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0800–0806, 01 2021.

[41] Ashish Semwal and Narendra Londhe. Mvfnet: A multi-view fusion network for pain intensity assessment in unconstrained environment. *Biomedical Signal Processing and Control*, 67:102537, 03 2021.

[42] Jiann-Shu Lee and Chuan-Wei Wang. Facial pain intensity estimation for icu patient with partial occlusion coming from treatment. In *BIBE 2019; The Third International Conference on Biological Information and Biomedical Engineering*, pages 1–4, 2019.

[43] Danila Mamontov, Iana Polonskaia, Alina Skorokhod, Eugene Semenkin, Viktor Kessler, and Friedhelm Schwenker. *Evolutionary Algorithms for the Design of Neural Network Classifiers for the Classification of Pain Intensity*, pages 84–100. 05 2019.

[44] Antoni Mauricio, Fábio Cappabianco, Adriano Veloso, and Guillermo Cámara. A sequential approach for pain recognition based on facial representations. In Dimitrios Tzovaras, Dimitrios Giakoumis, Markus Vincze, and Antonis Argyros, editors, *Computer Vision Systems*, pages 295–304, Cham, 2019. Springer International Publishing.

[45] Diyala Erekat, Zakia Hammal, Maimoon Siddiqui, and Hamdi Dibeklioğlu. Enforcing multilabel consistency for automatic spatio-temporal assessment of shoulder pain intensity. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 156–164, New York, NY, USA, 2021. Association for Computing Machinery.

[46] Ghazal Bargshady, Jeffrey Soar, Xujuan Zhou, Ravinesh C Deo, Frank Whittaker, and Hua Wang. A joint deep neural network model for pain recognition from face. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, pages 52–56, 2019.

[47] Ghazal Bargshady, Xujuan Zhou, Ravinesh C. Deo, Jeffrey Soar, Frank Whittaker, and Hua Wang. Ensemble neural network approach detecting pain intensity from facial expressions. *Artificial Intelligence in Medicine*, 109:101954, 2020.

[48] Daniel Martinez, Ognjen, Ognjen Rudovic, and Rosalind Picard. Personalized automatic estimation of self-reported pain intensity from facial expressions. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 06 2017.

[49] Laduona Dai, Joost Broekens, and Khiet P. Truong. Real-time pain detection in facial expressions for health robotics. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 277–283, 2019.

[50] Vedhas Pandit, Maximilian Schmitt, Nicholas Cummins, and Björn Schuller. I see it in your eyes: Training the shallowest-possible cnn to recognise emotions and pain from muted web-assisted in-the-wild video-chats in real-time. *Information Processing Management*, 57(6):102347, 2020.

[51] Mohammad Tavakolian, Miguel Bordallo Lopez, and Li Liu. Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation. *Pattern Recognition Letters*, 140:26–33, 12 2020.

[52] H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016.

[53] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1), sep 2020.

[54] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 2020.

[55] Yogesh Kumar and Ruchika Singla. Federated learning systems for healthcare: Perspective and recent progress. *Federated Learning Systems*, 2021.

[56] Yongxin Tong Qiang Yang; Yang Liu; Tianjian Chen. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[57] Dashan Gao, Ce Ju, Xiguang Wei, Yang Liu, Tianjian Chen, and Qiang Yang. Hhhfl: Hierarchical heterogeneous horizontal federated learning for electroencephalography. *arXiv:1909.05784*, 2020.

[58] H. B. McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *ArXiv*, abs/1602.05629, 2016.

[59] Sudipan Saha and Tahir Ahmad. Federated transfer learning: concept and applications. *Intelligenza Artificiale*, 2021.

[60] Yang Liu, Yan Kang, Chaoping Xing, Tianjian Chen, and Qiang Yang. A secure federated transfer learning framework. *IEEE Intelligent Systems*, 35(4):70–82, 2020.

[61] Yiqiang Chen, Jindong Wang, Chaohui Yu, Wen Gao, and Xin Qin. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 2021.

[62] Hankz Hankui Zhuo, Wenfeng Feng, Yufeng Lin, Qian Xu, and Qiang Yang. Federated deep reinforcement learning. *arXiv:1901.08277*, 2020.

[63] Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. Federated reinforcement learning: techniques, applications, and open challenges. *Intelligence &amp Robotics*, 2021.

[64] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Conference on Machine Learning and Systems*, 2020.

[65] Ognjen Rudovic, Nicolas Tobis, Sebastian Kaltwang, Björn Schuller, Daniel Rueckert, Jeffrey F. Cohn, and Rosalind W. Picard. Personalized federated deep learning for pain estimation from face images. *arXiv:2101.04800*, 2021.

[66] Nicolas TOBIS. *Federated Machine Learning: A Distributed Approach to Pain Expression Recognition in Healthcare*. PhD thesis, PhD thesis, Imperial College London, 2019.

[67] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[68] Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation*, pages 1–19, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[69] Mohammed Adnan, Shivam Kalra, Jesse C. Cresswell, Graham W. Taylor, and Hamid R. Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific Reports*, 12(1), 2022.

[70] Adrien Banse, Jan Kreischer, and Xavier Oliva i Jürgens. Federated learning with differential privacy. *arXiv:2402.02230*, 2024.

[71] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farokhi Farhad, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 2019.

[72] Lingchen Zhao, Lihao Ni, Shengshan Hu, Yaniiao Chen, Pan Zhou, Fu Xiao, and Libing Wu. Inprivate digging: Enabling tree-based distributed data mining with differential privacy. In *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pages 2087–2095, 2018.

[73] Jie Fu, Zhili Chen, and Xiao Han. Adap dp-fl: Differentially private federated learning with adaptive noise. *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2022.

[74] Patrick Lucey, Jeffrey Cohn, Kenneth Prkachin, Patricia Solomon, and Iain Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, pages 57–64, 03 2011.

[75] Paul Ekman, W v Friesen, and J Hager. Facial action coding system: Research nexus. *Network Research Information, Salt Lake City, UT*, 1, 2002.

[76] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In Hans Burkhardt and Bernd Neumann, editors, *Computer Vision — ECCV'98*, pages 484–498, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[77] Xinbo Gao, Ya Su, Xuelong Li, and Dacheng Tao. A review of active appearance models. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 40:145–158, 03 2010.

[78] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[79] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C. Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O. Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *2013 IEEE International Conference on Cybernetics (CYBCO)*, pages 128–131, 2013.

[80] Mohammad Tavakolian and Abdenour Hadid. A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics. *International Journal of Computer Vision*, 127, 10 2019.

[81] Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald Traue. Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. *British Machine Vision Conference*, 09 2013.

[82] Rainer Lienhart, Alexander Kuranov, and Vadim Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. *Proceedings of the 25th Pattern Recognition Symposium; 10-12 September 2003*, 2781:297–304, 09 2003.

[83] Modesto Castrillón Santana, Oscar Deniz, Cayetano Guerra Artal, and Marycarmen Hernández. Encara2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of Visual Communication and Image Representation*, 18:130–140, 04 2007.

[84] Axel Panning, Ayoub Al-Hamadi, Robert Niese, and Bernd Michaelis. Facial expression recognition based on haar-like feature detection. *Pattern Recognition and Image Analysis*, 18:447–452, 09 2008.

[85] Robert Niese, Ayoub Al-Hamadi, Axel Panning, and Bernd Michaelis. Emotion recognition based on 2d-3d facial feature extraction from color image sequences. *Journal of Multimedia*, 5:488–500, 10 2010.

[86] Kenneth M. Prkachin. The consistency of facial expressions of pain: a comparison across modalities. *Pain*, 51:297–306, 1992.

[87] Philipp Werner, Ayoub Al-Hamadi, and Steffen Walter. Analysis of facial expressiveness during experimentally induced heat pain. *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 10 2017.

[88] Pooja Prajod, Tobias Huber, and Elisabeth André. Using explainable ai to identify differences between clinical and experimental pain detection models based on facial expressions. In Björn ór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Binh Huynh Thi Thanh, and Benoit Huet, editors, *MultiMedia Modeling*, pages 311–322, Cham, 2022. Springer International Publishing.

[89] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[90] Ehsan Othman, Philipp Werner, Frerk Saxen, Ayoub Al-Hamadi, Sascha Gruss, and Steffen Walter. Automatic vs. human recognition of pain intensity from facial expression on the x-ite pain database. *Sensors*, 21(9), 2021.

[91] Ayla İrem Aydın and Nurcan Özyazıcıoğlu. Assessment of postoperative pain in children with computer assisted facial expression analysis. *Journal of Pediatric Nursing*, 71:60–65, 2023.

[92] Patama Gomutbutra, Adisak Kittisares, Atigorn Sanguansri, Noppon Choosri, Passakorn Sawaddiruk, Puriwat Fakfum, Peerasak Lerttrakarnnon, and Sompob Saralamba. Classification of elderly pain severity from automated video clip facial action unit analysis: A study from a thai data repository. *Frontiers in Artificial Intelligence*, 5:942248, 2022.

[93] Zakia Hammal and Jeffrey Cohn. Automatic detection of pain intensity. *ICMI'12 - Proceedings of the ACM International Conference on Multimodal Interaction*, 2012:47–52, 10 2012.

[94] AshwinRJ. Federated-learning-pytorch. `https://github.com/AshwinRJ/Federated-Learning-PyTorch/tree/master`, 2020.

[95] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. *International Conference on Artificial Intelligence and Statistics*, 2023.

[96] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016.

[97] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[98] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[99] Alexander Dante Camuto. Understanding gaussian noise in neural networks. *ICLR 2020*, 2021.

[100] Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. 2019.

[101] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[102] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[103] Pietro Morbidelli, Diego Carrera, Beatrice Rossi, Pasqualina Fragneto, and Giacomo Boracchi. Augmented grad-cam: Heat-maps super resolution through augmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4067–4071, 2020.

[104] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[105] Huazheng Wang, David Zhao, and Hongning Wang. Dynamic global sensitivity for differentially private contextual bandits. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22. ACM, September 2022.

[106] Ittai Dayan, Holger Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Abidin, Andrew Liu, Anthony Costa, Bradford Wood, Chien-Sung Tsai, Chih-Hung Wang, Chun-Nan Hsu, C. Lee, Peiying Ruan, Daguang Xu, Dufan Wu, Eddie Huang, Felipe Kitamura, Griffin Lacey, and Quanzheng Li. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature Medicine*, 27:1–9, 10 2021.

[107] Viktorija Pruckovskaja, Axel Weissenfeld, Clemens Heistracher, Anita Graser, Julia Kafka, Peter Leputsch, Daniel Schall, and Jana Kemnitz. Federated learning for predictive maintenance and quality inspection in industrial applications. *IEEE Conference on Prognostics and Health Management*, 2023.

[108] Yan Sun, Li Shen, and Dacheng Tao. Which mode is better for federated learning? centralized or decentralized. *ICLR 2024*, 2024.

[109] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, May 2020.

[110] Sayaka Kamei and Sharareh Taghipour. A comparison study of centralized and decentralized federated learning approaches utilizing the transformer architecture for estimating remaining useful life. *Reliability Engineering System Safety*, 233:109130, 2023.

[111] Maarten G. Poirot, Praneeth Vepakomma, Ken Chang, Jayashree Kalpathy-Cramer, Rajiv Gupta, and Ramesh Raskar. Split learning for collaborative deep learning in healthcare. *AMIA Annual Symposium Proceedings*, 2019.

[112] Praneeth Vepakomma, Otkrist Gupta, Tristan Swedish, and Ramesh Raskar. Split learning for health: Distributed deep learning without sharing raw patient data. *ICLR AI for social good workshop 2019*, 2018.