OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

EIT

FAKULTÄT FÜR
ELEKTROTECHNIK UND
INFORMATIONSTECHNIK

AN AUTOMATIC AND MULTI-MODAL SYSTEM FOR CONTINUOUS PAIN INTENSITY MONITORING BASED ON ANALYZING DATA FROM FIVE SENSOR MODALITIES

Dissertation

zur Erlangung des akademischen Grades

Doktoringenieur

(Dr.-Ing.)

von **M.Sc. Ehsan Abdulraheem Mohammed Othman**

geb. am 02. Juli 1983 in Taiz, Yemen

genehmigt durch die Fakultät für Elektrotechnik und Informationstechnik

der Otto-von-Guericke-Universität Magdeburg

Gutachter:

Prof. Dr.-Ing. habil. Ayoub Al-Hamadi
Prof. Dr. rer. nat. Andreas Wendemuth
Prof. Dr. Aly Farag

Promotionskolloquium am: 20. April. 2023

Without the help of my deceased sister *Nasim*, this thesis would not have been feasible, and I would want to dedicate this work to her soul with the deepest love and gratitude.

Ehsan

Abstract

Pain is an unpleasant feeling associated with tissue damage or psychological factors (non-physical). It is a reliable indicator of health issues. Pain assessment needs to be done for vulnerable patients who cannot self-report their pain, such as intensive care patients, people with dementia, or adults with cognitive impairment. So far, the current methods in the clinical application are subjected to biases and errors; moreover, such methods do not facilitate continuous pain monitoring. Therefore, recent studies have proposed and developed automatic pain assessment methodologies due to their possibility to objectively and robustly measure and monitor pain.

Regarding medical evidence, facial expressions, vocalizations, and physiological signals are valid indicators of pain. Hence, this thesis presents an automatic system for continuous monitoring of pain intensity based on analyzing data from five sensor modalities (frontal RGB video [frontal faces], audio, Electrocardiogram [ECG], Electromyogram [EMG], and Electrodermal Activity [EDA]) in the X-ITE Pain Database. Further, due to the promising multi-modality fusion, two and all modalities are fused using a model and a late fusion to produce more reliable information with less uncertainty.

In recent years, other authors have proposed automated methods for pain recognition by using features that were extracted independently per time series of a given sequence. However, the obtained results were poor due to the lack of representation of movement dynamics. Therefore, in this work, three distinct real-time methods are proposed to solve this problem for more reliable monitoring of continuous pain intensity. These methods are based on classifying descriptors of facial expressions from frontal faces, audio, and physiological time series and using a sliding-window strategy to obtain 10s-length input samples. The first proposed method is a Random Forest [RF] baseline method (Random Forest classifier [RFc] and Random Forest regression [RFr]), the second is Long-Short Term Memory Network (LSTM), and the third is LSTM using a sample weighting method (called LSTM-SW). The sample weighting method is used to reduce the weight of misclassified samples with less facial response in the training set to improve the performance.

In regard to classification and regression, several experiments using the three proposed methods (RF, LSTM, and LSTM-SW) are conducted in order to gain insights into monitoring continuous pain intensity using data from single, two, and all modalities from the X-ITE Pain Database. Before data is fed into models, it is pre-processed; 11 datasets were obtained to simplify the imbalanced database problem, improve the results, and generalize the capability of the proposed system. This system recognizes pain intensity (low, moderate, and severe) for two pain stimulus types (phasic and tonic) and two qualities (heat and electrical stimuli).

The results: (1) report that regression performs better than classification in imbalanced datasets, (2) show that LSTM and LSTM-SW methods outperform significantly guessing (majority of votes = no pain) and RF, (3) emphasize that EDA is the best single modality, (4) confirm that model fusion using multiple modalities ([facial expressions and EDA] or [EMG and EDA] or [facial expressions, audio, ECG, EMG, and EDA]) overcomes the limitations of single modalities, the performance improves significantly in 10 out of 11 datasets, (5) present that the models using all modalities (facial expressions, audio, ECG, EMG, and EDA) outperform those models using two modalities ([facial expressions and EDA] or [EMG and EDA]) with imbalanced tonic datasets, and (6) show that most model fusion models using EMG and EDA modalities are the best when using phasic imbalanced datasets and Heat Tonic Dataset [HTD] (almost balanced dataset). These findings are the baseline results for future research related to real-time pain intensity monitoring systems using single or multiple sensor modalities in the X-ITE Pain Database.

Index Terms—Continuous pain intensity monitoring; Electrocardiogram; Electrodermal Activity; Electromyogram; facial expressions; late fusion;Long-Short Term Memory Network; model fusion; modalities; Random Forest; sample weighting.

Zusammenfassung

Schmerz ist ein unangenehmes Gefühl, welches mit Gewebeschäden oder psychologischen (nicht-körperlichen) Faktoren verbunden ist. Zudem ist es ein zuverlässiger Indikator für Gesundheitsprobleme. Eine Schmerzbewertung muss für gefährdete Patienten durchgeführt werden, die ihre Schmerzen nicht selbst angeben können, wie z. B. Intensivpatienten, Menschen mit Demenz oder Erwachsene mit kognitiven Beeinträchtigungen. Bisher sind die aktuellen Methoden in der klinischen Anwendung mit Verzerrungen und Fehlern behaftet; darüber hinaus ermöglichen solche Verfahren keine kontinuierliche Schmerzüberwachung. Daher haben neuere Studien aufgrund ihrer Möglichkeit einer objektiven und robusten Messung sowie Überwachung von Schmerzen automatische Schmerzbewertungsmethoden vorgeschlagen und entwickelt.

Aus medizinischer Sicht sind Gesichtsausdrücke, Lautäußerungen und physiologische Signale gültige Schmerzindikatoren. Daher präsentiert diese Arbeit ein automatisches System zur kontinuierlichen Überwachung der Schmerzintensität basierend auf der Analyse von Daten von fünf Sensormodalitäten (frontales RGB-Video [Frontalgesichter], Audio, Elektrokardiogramm [EKG], Elektromyogramm [EMG] und Elektrodermale Aktivität [EDA]) in der X-ITE-Schmerzdatenbank. Darüber hinaus werden aufgrund der vielversprechenden Fusion mehrerer Modalitäten zwei und alle Modalitäten unter Verwendung von Modell- und Entscheidungsfusion fusioniert, um zuverlässigere Informationen mit weniger Unsicherheit zu erzeugen.

In den letzten Jahren haben andere Autoren automatisierte Verfahren zur Schmerzerkennung unter Verwendung von Merkmalen vorgeschlagen, die unabhängig pro Zeitreihe aus einer bestimmten Sequenz extrahiert wurden. Die erzielten Ergebnisse waren jedoch aufgrund der fehlenden Darstellung der Bewegungsdynamik unzuverlässig. Daher werden in dieser Arbeit drei unterschiedliche Echtzeitmethoden vorgeschlagen, um dieses Problem für eine zuverlässigere Überwachung der kontinuierlichen Schmerzintensität zu lösen. Diese Methoden basieren auf der

Klassifizierung von Deskriptoren von Gesichtsausdrücken aus frontalen Gesichtern, Audio- und physiologischen Zeitreihen und der Verwendung einer Sliding-Window-Strategie, um Eingabeproben mit einer Länge von 10 Sekunden zu erhalten. Die erste vorgeschlagene Methode ist eine Random Forest [RF]-Basismethode (Random Forest Classifier [RFc] und Random Forest Regression [RFr]), die zweite ist ein Long-Short Term Memory Network (LSTM) und die dritte ist ein LSTM mit Stichprobengewichtung (genannt LSTM-SW). Die Probengewichtungsmethode wird verwendet, um das Gewicht falsch klassifizierter Proben mit geringerer Gesichtsreaktion im Trainingssatz zu reduzieren und dadurch die Leistung zu verbessern.

In Bezug auf Klassifizierung und Regression wurden mehrere Experimente mit den drei vorgeschlagenen Methoden (RF, LSTM und LSTM-SW) durchgeführt, um Einblicke in die Überwachung der kontinuierlichen Schmerzintensität unter Verwendung von Daten von einzelnen, zwei und allen Modalitäten aus der X-ITE-Schmerzdatenbank zu gewinnen. Bevor Daten in Modelle eingespeist werden, werden sie vorverarbeitet; es wurden 11 Datensätze erhalten, um das Problem der unausgeglichenen Datenbank zu vereinfachen, die Ergebnisse zu verbessern und die Leistungsfähigkeit des vorgeschlagenen Systems zu verallgemeinern. Dieses System erkennt die Schmerzintensität (niedrig, mittel und stark) für zwei Arten von Schmerzreizen (phasisch und tonisch) und für zwei Qualitäten (Wärme und elektrische Reize).

Die Ergebnisse zeigen (1), dass die Regression in unausgewogenen Datensätzen besser abschneidet als die Klassifizierung; (2), dass LSTM- und LSTM-SW-Methoden das Schätzen und RF deutlich übertreffen; (3), dass EDA die beste Einzelmodalität ist, (4), dass Modellfusion mit mehreren Modalitäten ([Gesichtsausdruck und EDA] oder [EMG und EDA] oder [Gesichtsausdruck, Audio, EKG, EMG und EDA]) die Einschränkungen einzelner Modalitäten überwindet (es verbessert sich die Leistung in 10 von 11 Datensätzen signifikant); (5), dass die Modelle, die alle Modalitäten (Gesichtsausdruck, Audio, EKG, EMG und EDA) verwenden, die Modelle mit zwei Modalitäten ([Gesichtsausdruck und EDA] oder [EMG und EDA]) mit unausgewogenen tonischen Datensätzen übertreffen und (6), dass die meisten Modellfusionsmodelle, die EMG- und EDA-Modalitäten verwenden, die besten sind, wenn phasisch unausgeglichene Datensätze und Heat Tonic Dataset [HTD] (fast ausgeglichener Datensatz) verwendet werden. Diese Ergebnisse sind die Basisergebnisse für die zukünftige Forschung im Zusammenhang mit Echtzeit-Schmerzintensitätsüberwachungssystemen unter Verwendung von Einzel- oder Multisensormodalitäten in der X-ITE-Schmerzdatenbank.

Table of Contents

List of Figures

List of Tables

List of Equations

Acronyms and Abbreviations

Nomenclature	Description
AAM	Active Appearance Model
Audio	Audio Descriptor
AUs	Action Units
BCE	Binary Cross Entropy
biLSTM	bidirectional Long Short-Term Memory network
Bi-modality	[facial expressions and EDA] or [EMG and EDA]
BioVid	BioVid Heat Pain Database
BP4D	BP4D-Spontaneous Database
BSIF	Binarized Statistical Image Features
BVP	blood volume pulse
CCE	Categrical Cross Entropy
CERT	Computer Expression Recognition Toolbox
CLNF	Conditional Local Neural Fields
CNN	convolutional neural network
CNNs	Convolutional Neural Networks
COPE	Classification of Pain Expressions database
corrugator superscillii	electrical properties of the muscle close to eyebrows
DDP	delivered duty paid
DT	detection threshold
ECG	Electrocardiogram
ECG	ECG Descriptor
EDA	Electrodermal Activity
EDA	EDA Descriptor
EEG	Electroencephalography

Nomenclature	Description
ELBP	Elongated Binary Pattern
ELTP	Elongated Ternary Pattern
EMG	Electromyogram
EMG-D	EMG Descriptor
EPD	Electrical Phasic Dataset
ETD	Electrical Tonic Dataset
FACS	Facial Action Coding System
FAD	Facial Activity Descriptor
GSR	Galvanic Skin Response
HMP	head movements and postures
HNR	logarithmic Harmonics to Noise Ratio
HOG	Histogram of Oriented Gradients
HPD	Heat Phasic Dataset
HR	Heart Rate
HRV	Heart Rate Variability
HTD	Heat Tonic Dataset
HTD	Heat Tonic Dataset
HTK	Hidden-Markov Toolkit
IASP	International Association for the Study of Pain
ICC	Intraclass Correlation Coefficient
KNN	k-nearest neighbors
LBP	Local Binary Pattern
LDA	Discriminant Analysis
LLD	low-level descriptors
LPQ	Local Phase Quantization
LSTM	Long-Short Term Memory Network
LSTM-SW	LSTM using sample weighting method
LTP	Local Ternary Pattern
Meas.	Measure
MFCCs	Mel Frequency Cepstral Coefficients
Micro avg. F1-score	Micro average F1-score
Micro avg. precision	Micro average precision
Micro avg. recall	Micro average recall
MSE	Mean Squared Error

Nomenclature	Description
Multi-modality	facial expressions , audio, ECG, EMG, and EDA
NPL	noise peak level
PCA	Principal Component Analysis
PD	Phasic Dataset
PDM	joint Point Distribution Model
PLP	Perceptual Linear Predictive Coefficients
PPG	photoplethysmogram
PSPI	Prkachin and Solomon Pain Intensity
QRS	The combination of three waves in ECG (Q, R, and S waves)
RCNN	Recurrent Convolutional Neural Network
Red. Subset	Reduced Subset
REPD	Reduced Electrical Phasic Dataset
RETD	Reduced Electrical Tonic Dataset
RF	Random Forest
RFc	Random Forest classifier
RFr	Random Forest regression
SCL	Skin Conductance Level
SNS	sympathetic nervous system
SVM	Support Vector Machine
SVR	Support Vector Regression
TC	threshold coefficient
TD	Tonic Dataset
TOP	Three Orthogonal Planes
trapezius	electrical properties of the muscle at neck/shoulder
Triv.	Trivial
UNBC	UNBC-McMaster Shoulder Pain Database
Uni-modality	facial expressions, audio, ECG, EMG, or EDA
USF-MNPAD-I	Multimodal Neonatal Pain Assessment Dataset
X-ITE Pain Database	Experimentally Induced Thermal and Electrical (X-ITE) Pain Databas
zygomaticus	electrical properties of the muscle at the cheeks

Related Publications

Most of the material contained in this dissertation is partly based on the following refereed papers and journals published in a variety of peer-reviewed journals and international conferences.

Journal Articles

1) E. Othman, P. Werner, F. Saxen , A. Al-Hamadi, S. Gruss, and S. Walter, "Automated Electrodermal Activity and Facial Expression Analysis for Continuous Pain Intensity Monitoring on the X-ITE Pain Database," *Life* ***(IF 3.2)***, vol. 13, no. 9, 2023.

2) E. Othman, P. Werner, F. Saxen , A. Al-Hamadi, S. Gruss, and S. Walter, "Classification networks for continuous automatic pain intensity recognition in video using facial expression on the X-ITE Pain Database," *J. Vis. Commun. Image Represent* ***(IF 2.678)***, vol. 91, 2023.

3) E. Othman, P. Werner, F. Saxen , M. Fiedler, and A. Al-Hamadi, "An Automatic System for Continuous Pain Intensity Monitoring based on Analyzing Data from Uni-, Bi-, and Multi-modality," *Sensors* ***(IF 3.576)***, vol. 22, no. 13, 2022.

4) E. Othman, P. Werner, F. Saxen , A. Al-Hamadi, S. Gruss, and S. Walter, "Automatic vs. human recognition of pain intensity from facial expression on the X-ITE Pain Database," *Sensors* ***(IF 3.275)***, vol. 21, no.9, 2021.

5) E. Othman, F. Saxen , D. Bershadskyy, P. Werner, A. Al-Hamadi, and J. Weimann, "Predicting group contribution behaviour in a public goods game from face-to-face communication," in *Sensors* ***(IF 3.031)***, vol. 19, no.12, 2019.

6) E. Othman, and A. Al-Hamadi, "Automatic arabic document classification based on the HRWiTD algorithm," in *JSEA Journal* ***(IF 0.99)***, vol. 11, no.4, 2018.

Conference Proceedings & Workshops

1) E. Othman, P. Werner, F. Saxen , A. Al-Hamadi, and S. Walter, "Regression networks for automatic pain intensity recognition in video using facial expression on the X-ITE Pain Database," *In the 25th int'l conf on image processing, computervision and pattern recognition (IPCV'21)* las Vegas, USA, 26-29 July 2021.

2) E. Othman, P. Werner, F. Saxen , A. Al-Hamadi, and S. Walter, "Cross-database evaluation of pain recognition from facial video. In International symposium on image and signal processing and analysis," in *in the 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Dubrovnik, Croatia,23-25 September 2019.

3) D. Bershadskyy, E. Othman, and F. Saxen "Predicting free-riding in a public goods game: Analysis of content and dynamic facial expressions in face-to-face communication," in *In Halle institute for economic research (IWH) discussion papers*, 2019.

4) F. Saxen , P. Werner, S. Handrich, E. Othman, L. Dinges, and A. Al-Hamadi, "Face attribute detection with MobileNetv2 and NasNet-mobile. In International symposium on image and signal processing and analysis," in *In the 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Dubrovnik, Croatia, 23-25 September 2019.

5) F. Saxen , P. Werner, S. Handrich, E. Othman, and A. Al-Hamadi, "Detecting arbitrarily rotated faces for face analysis," in *In International conference on image processing (ICIP)*, Taipei, Taiwan, 22-25 September 2019.

CHAPTER 1

Introduction

THE International Association for the Study of Pain (IASP) defines pain as an unpleasant sensory and emotional experience associated with, or resembling that is associated with, actual or potential tissue damage [1]. Pain is a complex phenomenon and not well understood yet. It includes several components [2]: an affective component (emotions or feelings associated with pain), a cognitive component (all thought processes or intellectual activity related to pain) [3], a social component, and a sensory component (pain intensity, quality, location, and duration) [4]. Due to the subjective nature of pain, all those components influence how individuals experience pain in their lives. Further, pain is considered as a warning mechanism that gives a solid and reliable message about body health condition; it is an indicator of something within the body that needs to be cured. Therefore, it is necessary to indicate people to pay serious medical attention and respond quickly.

Pain is classified into many categories, depending on different characteristics. It is commonly classified based on duration into acute or chronic: acute pain comes on suddenly for a short period (less than three months), and chronic pain comes on quickly or slowly, continuously or intermittently, and sometimes it is intense for a long period (longer than three months) [5]. It is also classified regarding intensity or severity as mild, moderate, or severe. In addition, pain is classified as either: nociceptive (pain caused by body tissue injury), neuropathic (pain caused by nerve injury), or psychogenic (pain that is influenced by psychological factors such as mental, emotional, and behavioral disorders).

Reliable and robust pain measurement and monitoring contribute significantly to diagnose and treat pain at the right time and monitoring the success of the ongoing treatment. Pain is either expressed physically through visual cues (facial expressions and body movements), vocalization cues (verbally and non-verbally), and physiological cues (bio-signals and brain activity); these cues play an important role in assessing pain with individuals [6].

Facial expressions aid in predicting human behaviors [7,8], and emotions [9,10]. It is one of the main indicators of how a human expresses pain [9,11], and it is expressed similarly across different nationalities, genders, cultures, ages, and genders. Facial expressions are important signals that are widely used in nonverbal communication [12]. There are distinct facial expressions associated with pain experiences; moreover, rising intensities of noxious stimulation increase the intensity of facial expressions [13,14]. Ekman and Friesen [15] decompose facial expressions into individual facial Action Units (AUs) with the Facial Action Coding System (FACS). Fig. 1.1 shows an example of facial response during pain described by involved AUs.

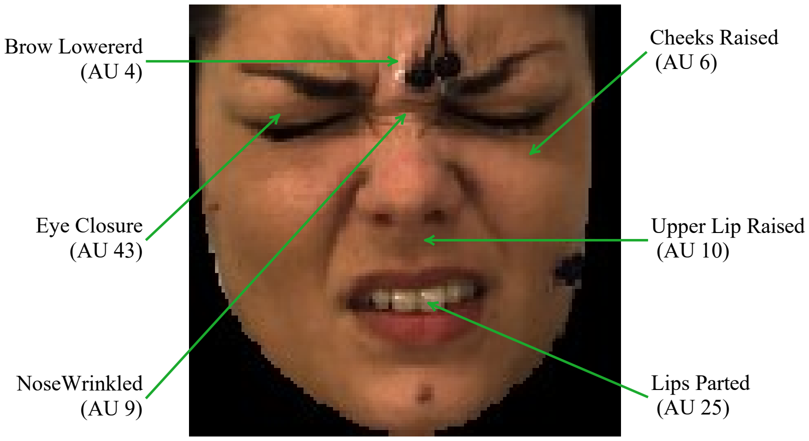


FIG. 1.1. An example of facial expressions of pain and associated AUs.

Further, body movement is considered an indicator for pain assessment. Individuals often display many body movements as behavioral responses to pain, such as protective reflexes, rubbing, and writhing. Additionally, some individuals express pain verbally by mentioning pain or using offensive words, and some express their pain non-verbally by moaning, crying, groaning, and sighing (vocalization cues) [16,17].

In clinical practice, observation and patient self-report are used to measure pain intensity, location, and duration. However, it is impossible to monitor the pain intensity of patients constantly with those standard measurements of pain assessment. Hence, it is beneficial to introduce reliable automated pain recognition systems based on these behavioral (visual and para-linguistic) and physiological pain responses to complement current methods for better pain management. In this regard, several artificial intelligence (AI) methods have been developed over the last years to automatically detect pain and estimate its intensity based on analysis of behavioral or physiological pain indicators or a combination of these indicators.

In order to develop and evaluate automatic pain intensity assessment methods, the objective data (pain-relevant and pain-irrelevant information) are usually recorded using non-invasive sensor technology, which captures data on the body responses of the individual in pain. This is achieved with cameras that capture facial expressions, gestures, or posture, while microphones record para-linguistic features (vocalizations). Physiological information such as heart rate and muscle tone are collected via biopotential sensors (electrodes) [18]. Then, machine learning methods build a model based on the collected data to identify semantic patterns for accurate and objective pain assessment.

1.1 Problem Statement

Pain compromises patients' quality of life when not well managed [19]; they might suffer from anxiety, sleep disturbance, and inability to concentrate on the activity of daily living [20, 21]. Further, patients face difficulties during hospitalization as care is influenced by a lack of knowledge about pain. Therefore, a reliable and trustworthy pain assessment is necessary for adequate pain management, changing analgesic dose, and additional interventions if required. Additionally, good care requires more than one pain intensity measure (self-report, external observations, or physiological tests [22]).

The self-report (numeric pain scale or interview) is efficient for patients who are able to communicate. However, the behavioral observation by medical staff is the preferred method for pain assessment compared to the self-report [23], especially with patients with limited communication abilities (e.g., intensive care patients, people with moderate to severe cognitive impairment/dementia, paralyzed patients, patients on oxygen, or normal patients who suddenly lose the ability to express their pain due to a weakness or discomfort). Nevertheless, medical staff could not measure pain intensity every minute or second to notice such emergencies and act immediately; they are often busy, especially in crowded hospitals and healthcare centers. Further, the human observer may be influenced by personal factors, such as the relationship to the sufferer [24] and the patient's attractiveness [25]. Thus, those problems confirm the need for objective and robust automatic systems for effective pain management.

1.2 Motivations and Application

This work is motivated by two main problems in hospitals and healthcare centers: the great challenge for pain intensity recognition through non-verbal communication and the need for more medical staff to look after patients, as mentioned previously (see Section 1.1 above). Since pain is usually followed by visual, vocalization, and physiological cues, automatic systems rely on them to continuously monitor pain intensity. Facial expressions are very informative for pain detection [26, 27]. Further, in critical emergencies with patients, such as apneas, the facial expressions are generally very expressive. Additionally, vocalizations that are extracted from audio signal are used for pain assessment by analyzing moan, cry, or groan [16]; moreover, physiological signals are effective indicators of pain assessment [28]. That is why the combination of facial expressions, vocalizations, and physiological signals can be a great addition to pain assessment means. Hence, the availability of automatic systems, that use this combination to provide an objective and robust pain intensity measurement and monitoring, would be a helpful solution for the problems mentioned above.

This work is also motivated by potential future directions presented by Werner et al. [29]. They used the X-ITE Pain Database, which comprised reactions to pain stimulus intensities in four different qualities: phasic (short) and tonic (long) variants of each heat and electrical stimuli. As the database plays an important role in refining, improving, and evaluating automated recognition systems, the X-ITE Pain Database has been introduced for an objective pain assessment. Werner et al. [29] and Walter et al. [30] reported a promising finding to automatically recognize pain intensity when using multiple sensors: frontal RGB camera and audio (for external observation); and ECG, EMG, and EDA (for physiological tests). In their experiments, they used parts of complete data. Each part was cut out from the continuous recording of data for 7 seconds. In line with their studies, data from the same sensors was used in this work, but with advancement (most of the data was used from the continuous recording of the main stimulus phase). Using the same data (balanced dataset) as in [29,30] was not suitable for continuous monitoring of pain intensity; dealing with most of the data (imbalanced dataset) obtained better results for such task. Further, as long as automated systems at low-cost are preferred, this study shows possibilities for good continuous monitoring of pain intensity when not all sensors are used.

Due to the problem of how individuals differ in showing facial expressions [31] and pain response threshold [32], it is found plenty of labels (which represent the pain stimulus) that do not match the pain stimulation intensity. Fig. 1.2 shows some confusing samples. Some subjects show a lack of facial responses to pain: some have low pain sensitivity resulting in a high tolerance threshold requiring a temperature cutoff to avoid burns; others show a low tolerance threshold intentionally or unintentionally during stimulus calibration, possibly because they do not want to feel severe pain. Such inconsistencies between the label and the video may be considered outliers or label noise. Data cleaning by removing such outlier samples may be used to improve the facial expression-based recognition performance. However,

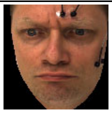
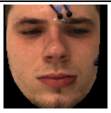
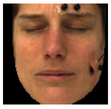
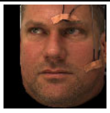
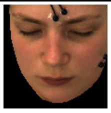
Modalities	no pain	Phasic Heat Pain			Phasic Electrical Pain		
Ground truth	BL	PH1	PH2	PH3	PE1	PE2	PE3
Samples							

FIG. 1.2. Some examples of difficult samples to recognise phasic pain intensity. Pain intensity on a scale of 1 to 3, 1 = low, 2= moderate, and 3 = severe.

this may remove some samples, which are useful for improving multimodal pain recognition because there may be pain responses in other modalities like EMG or EDA, although no facial responses show up. This is enough motivation to propose a method (sample weighting method) that is useful for improving the performance of the single and fused modalities models for pain assessment by downweighting noisy samples rather than eliminating them as in the cleaning up strategy.

The X-ITE Pain Database is extremely imbalanced; the samples without pain stimulus are the vast majority. This problem is hard to solve when using one balanced dataset for training due to many reasons: (1) the models would fail to recognize more no pain samples in an imbalanced testing set, and their performance would be poor; (2) the sample size of four different pain qualities (phasic [short] and tonic [long] variants of each, heat and electrical stimuli) in three intensities (low, moderate, and severe) is also imbalanced, and solving this problem by balancing the database would bring some risks such as duplication of the outlier samples when increasing the sample size of the minority class; moreover, eliminating some samples from majorities classes to balance the database would decrease the performance of models because these removed samples would be informative and useful for improving pain recognition models. Further, the pain recognition model using all samples without a balancing database would be biased towards the majority and fail to recognize pain intensity in samples of minority classes. The size of the tonic samples is very small compared to that of the phasic samples, and the size of the heat samples size is less than the size of electrical samples for both pain qualities (phasic and tonic). Hence, this is the motivation to propose 11 datasets from the X-ITE Pain Database to reduce the impact of the imbalanced problem and generalize the capability of machine learning methods when using pain stimulus intensities in different qualities. The proposed datasets refer to use phasic data, heat phasic data, electrical phasic data, tonic data, heat tonic data, electrical tonic data, and reduced datasets after using a reduction strategy on the previously mentioned imbalanced datasets.

Deep learning methods provide a lot of promise for time series, such as Long

Short-Term Memory (LSTM). It is used to significantly improve the performance of the continuous pain intensity monitoring task. Further, LSTM units are well-trained in big data. Nevertheless, machine learning methods such as Random Forest (RF) generally work well with high-dimensional data. Thus, another motivation is introduced in this work, which is a comparison between LSTM and RF for monitoring continuous pain intensity in the X-ITE Pain Database.

Due to the COVID-19 pandemic, there is an unmanageable number of patients every day in hospitals; the medical staff can become very busy and fail to monitor all patients effectively, especially those who need more attention and care. Additionally, the physical distance between patients and others is necessary to avoid contamination. Therefore, many countries such as China [33] found the perfect opportunity to switch to automated systems, which can help effectively to manage patients with COVID-19. Thus, these issues have motivated the creation of automatic continuous monitoring of pain intensity system based on facial expressions, vocalizations, and physiological signals. The automated pain monitoring system in hospitals and healthcare centers is proposed to (1) provide an alert when a pain event is detected; (2) allow faster response; (3) make the correct diagnoses at the right time; (4) reduce monitoring time; (5) reduces the risk of the patient; (6) reduce stress on the workers; (7) give peace of mind to the families; and (8) avoid hiring more medical staff.

1.3 Goal and Contributions

This work aims to create an objective and reliable automatic system for continuous monitoring of pain intensity by analyzing facial expressions, vocalization, and physiological cues. Several experiments were carried out to achieve this aim, including three automatic methods: (1) Random Forest (RF) baseline methods [Random Forest classifier (RFc) and Random Forest regression (RFr)], (2) a Long-Short Term Memory (LSTM) method, and (3) a LSTM using sample weighting method (LSTM-SW). These methods were applied on 11 datasets, which were suggested to reduce the impact of huge imbalanced datasets. The proposed system is the first to monitor continuous pain intensity based on analyzing data from five sensor modalities (frontal RGB video [frontal faces], audio, and physiological signals [ECG, EMG, and EDA]) in the X-ITE Pain Database. This work goes beyond the Werner et al. [29] and Walter et al. [30] studies: most of the data was used from the continuous recording of the main stimulation phase in the X-ITE Pain Database; whereas they used phasic (short) and tonic (long) samples, which have been cut out from the continuous recording of the main stimulation phase and were temporally aligned with the stimuli.

Many contributions were made to achieve the aim, including (1) an investigation of the most efficient loss function with the LSTM method by comparing the results achieved when using frontal RGB video sensor modality; the selective functions were Mean Squared Error (MSE) and Binary Cross-Entropy (BCE), (2) a comparison between classification and regression methods when using RF, LSTM, and LSTM-SW, (3) a performance comparison of the proposed methods when using data from single sensors (modalities) to those when using data from two or more fused modalities, (4) an emphasis that the proposed 11 datasets from the X-ITE Pain Database help to simplify the imbalanced database problem and improve the results, and (5) a confirmation that deep learning methods (LSTM and LSTM-SW) are the best for time series tasks when using big data.

1.4 Methodology

Fig. 1.3 shows the methodology of the proposed automatic system for continuous monitoring of pain intensity. First, the input data from the five modalities (facial expressions, audio, ECG, EMG, and EDA) were described and pre-processed to extract useful features; for more details, see Section 4.1 and 4.2, respectively. Second, temporal integration features were calculated from each time series data coming from the five sensors (see Section 4.3). Third, the experimental data was prepared by further processing the obtained data from the temporal integration process; such processing was suggested to overcome imbalanced database and outliers problems, see Section 4.4. Fourth, three methods were used to analyze the experimental data for continuous monitoring pain intensity, which were Random Forest (RF) as baseline methods (Random Forest classifier [RFc] and Random Forest regression [RFr]), Long-Short Term Memory (LSTM), and LSTM using sample weighting method (LSTM-SW); see Sections 5.1, 5.2, and 5.3, respectively. In regard to classification and regression, three types of experiments were introduced using the proposed methods: (1) Uni-modality (data from single sensors) experiments, (2) Bi-modality (fusing data from two modalities) experiments, and (3) Multi-modality (fusing data from five modalities) experiments. The reason for suggesting those experiments is to introduce the best automatic system for continuous monitoring of pain intensity after analyzing facial expressions that extracted from frontal faces, audio, and physiological signals [ECG, EMG, and EDA]. For more details about the conducted experiments, see Section 5.4. Finally, for each type of experiment, the experiments were run on 11 datasets from the experimental data for performance evaluation; for more details about evaluation results, see Chapter 6.

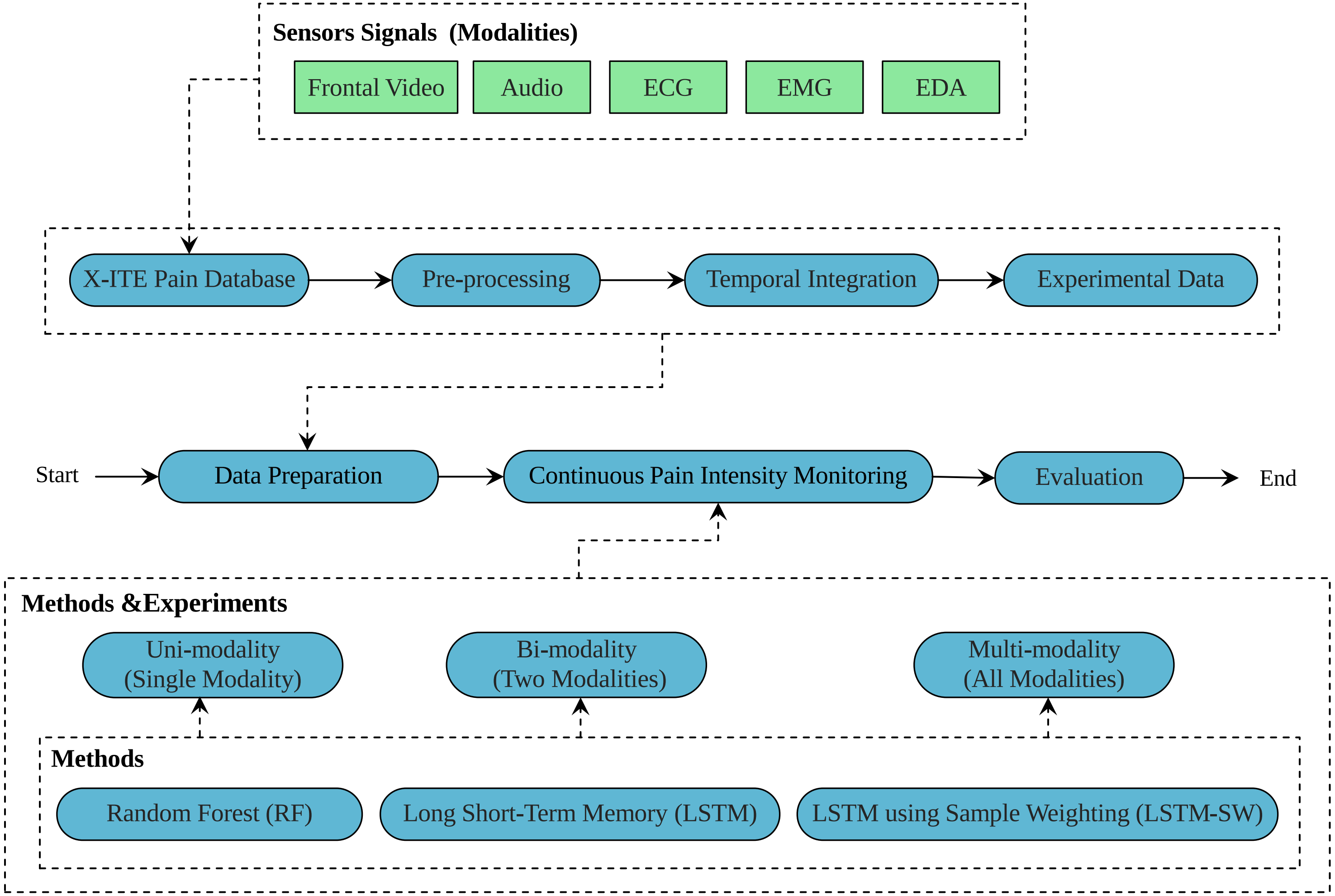


FIG. 1.3. The general pipeline of the proposed automatic system for continuous monitoring of pain intensity.

1.5 Thesis Outline

The thesis is organized into seven major chapters, including this introductory chapter. Chapter 2 delivers an overview of pain recognition methods based on data from frontal RGB video, audio, and physiological signals, then describes their relevance to this thesis. Chapter 3 describes the necessary fundamentals of the employed tools and methods. Chapter 4 summarizes the X-ITE Pain Database and feature extraction preprocessing; it also presents the temporal integration process and experimental data. Chapter 5 introduces the proposed methods for continuous monitoring of pain intensity followed by the experimental setup regarding classification and regression. Chapter 6 presents the evaluation of the experiments' results. Finally, Chapter 7 discusses the results; it also concludes the contributions, investigations, and experiments that are discussed in the thesis, as well as directions for future research on automatic continuous pain intensity monitoring. Appendix A details the results of the proposed methods (RF, LSTM, and LSTM-SW) using all Uni-modality models regarding classification and regression. B details the results of RF, LSTM, and LSTM-SW using Bi-modality models regarding classification. C details the results of RF, LSTM, and LSTM-SW using Multi-modality regarding classification and regression.

CHAPTER 2

State-of-the-Art

THIS chapter overviews the important topics related to automatic pain recognition, which has been studied extensively in the last years. First, the existing methods that detect pain and recognize pain intensity in terms of the subject of pain from behavioral, physiological, and fusion perspectives are presented in Section 2.1. This knowledge is required in later chapters for the development and validation of automated methods for pain assessment. Then, the available pain databases for research use are displayed in Section 2.2. Finally, the general training techniques for data recognition tasks are described in Section 2.3, followed by a conclusion in Section 2.4.

2.1 Automatic Pain Assessment

Many studies have focused on various possible objective indicators for pain. The behavioral and physiological pain indicators are commonly used for pain assessment. Facial expressions, verbal and non-verbal vocalizations, and body movements are considered behavioral pain indicators; heart rate variability, muscle activity, electrodermal activity, brain activity, blood pressure, and respiration rate are considered physiological pain indicators. A rich variety of automatic pain recognition methods have been proposed based on analysis behavior and physiological data in the last years. To date, machine learning and deep learning are preferable due to their good predictive power. This section reviews the pain recognition methods using: behavior pain indicators (see Section 2.1.1), physiological pain indicators (see Section 2.1.2), fusion of behavioral and physiological indicators (Section 2.4).

2.1.1 Behavior Pain Indicators

The focus of this section is on introducing state-of-the-art automated methods for pain assessment when using behavioral responses to pain. Such methods were used to extract and analyze pain-relevant features from behavioral pain indicators such as facial expressions (see Section 2.1.1.1), vocalizations (see Section 2.1.1.2), and body movements (see Section 2.1.1.3).

2.1.1.1 Facial Expressions

Facial expressions are the most reliable indicator of pain [25, 27, 34–36]. Ekman and Friesen [25] decomposed facial expressions into individual facial Action Units [AUs] with the Facial Action Coding System [FACS]. Examples of facial changes associated with pain are brow lower, cheek raise, lids tight, nose wrinkle, nasolabial deepen, upper lip raise, lip corner pull, lip stretch, lips apart, jaw drop, lids droop, eyes closed, blink [13, 27, 37–40]. Fig. 2.1 shows a list of the AUs that occur in a painful experience; only a combination of them expresses pain, but not all. Such combination of AUs is often combined with head pose towards head movements and postures as pain behaviors [41].

Prkachin and Solomon Pain Intensity [PSPI] [14] is a metric that measures pain as a linear combination of the intensities (1-15) of facial action units associated with pain. PSPI scores are assigned to images on a frame-by-frame basis using the metric in Equation 2.1. A couple of studies have attempted pain recognition mainly on the frame level and at the sequence level based on this metric. AU4 and AU43 must be present in pain, one of AU6 and AU7 and one of AU9 and AU10 must be present too (the highest intensity is selected if both are present).

$$PSPI = AU4 + \max(A \cup 6 \ \ or \ \ AU7) + \max(A \cup 9 \ \ or \ \ AU10) + AU43 \qquad (2.1)$$

The automatic pain recognition from facial expressions consists of three main steps: (1) face detection and registration, (2) feature extraction, and (3) pain expression recognition. Several automatic systems analyzed AUs and their combinations for recognizing frame-level and sequence(video)-level pain intensity. Feature reduction based methods are the most common type of frame-level methods. In a simple feature reduction based method, the static images are manually rotated, and then the face is cropped out with an elliptical bounding box. Colour information is discarded in the extracted face; the resulting grey-scale image is row concatenated to form a single feature vector of H×W dimensions, where H and W represent the image's height and width, respectively. Then, to reduce the vector's dimensionality, feature reduction methods are applied, such as Principal Component Analysis [PCA] [42] and Sequential Floating Forward Selection (SFFS) [43].

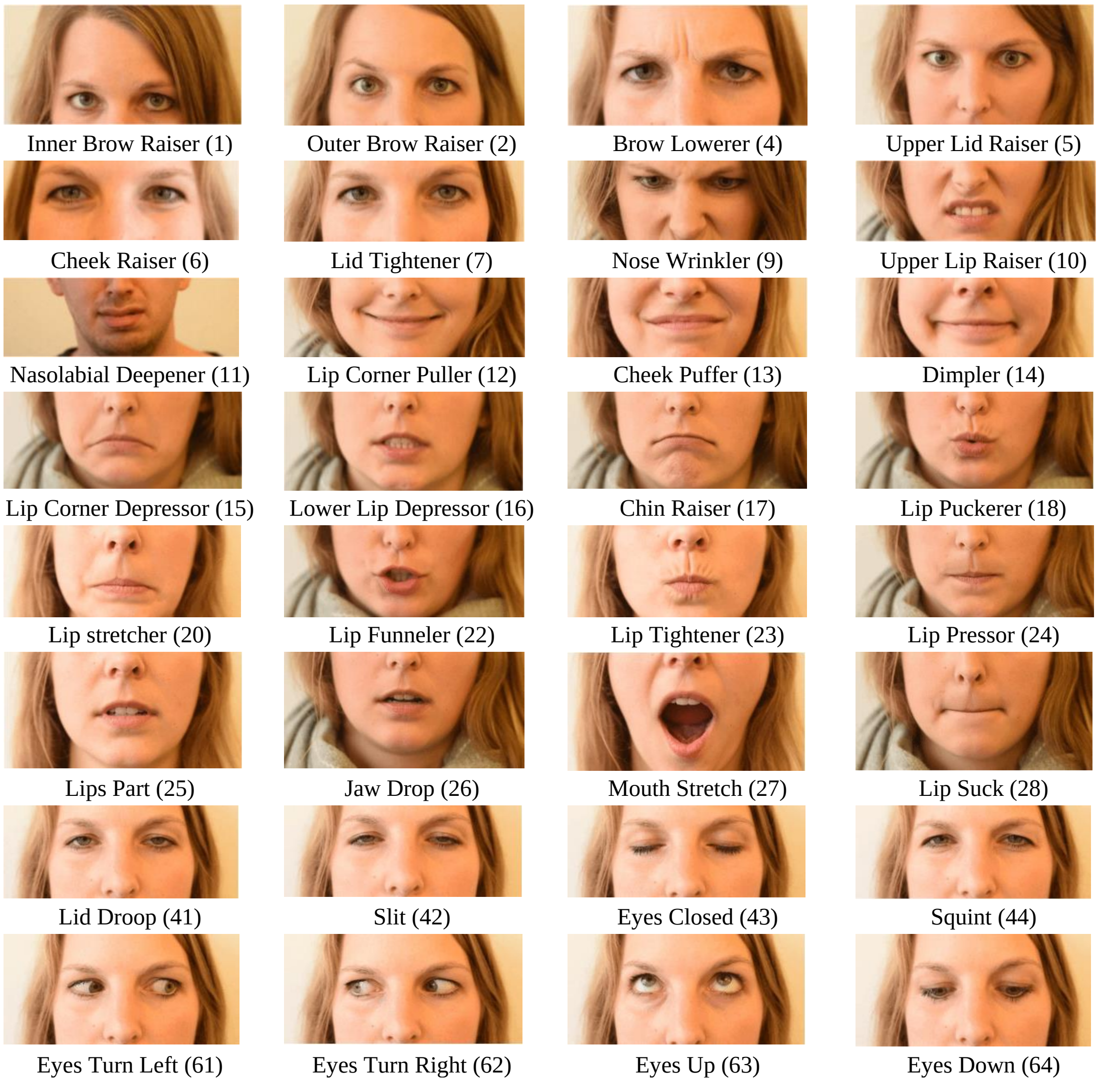


FIG. 2.1. Description of AUs that associated with pain. A combination of some of these AUs expresses pain, but not all. The number in practices indicates the AU's number [40]. Images are modified and taken from [44] after getting permission.

Brahnam et al. [45, 46] presented one of the first research in automatic pain recognition; they proposed and applied feature reduction based method on the Classification of Pain Expressions [COPE] dataset [46] followed by the application of distance-based methods (PCA [42] and Discriminant Analysis [LDA] [47]) and Support Vector Machine [SVM] [48]. The results showed that SVM was significantly outperformed distance-based methods in classifying pain versus no pain (accuracy = 88.00%). This work was extended in [49], the results showed Network Simultaneous Optimization Algorithm [NSOA] achieved the highest accuracy of 92.20% compared to SVM (82.35%), PCA (80.39%), and LDA (76.9%). Gholami, et al. [50] presented binary and multi-class classification. To estimate the intensity level of the detected pain expression, they introduced Relevance Vector Machine [RVM], which was a Bayesian version of SVM that provides the posterior probabilities for the class memberships.

Nanni et al. [51] introduced a method to detect pain expressions using some common texture descriptor based methods (second most common type of frame-level methods), which were Local Binary Pattern [LBP] [52] algorithm or other variants of LBP such as Local Ternary Pattern [LTP] [53], Elongated Binary Pattern [ELBP] [54], and Elongated Ternary Pattern [ELTP] [55]. Then, they selected the most discriminant features using SFFS on the training set, and an ensemble of Radial Bias SVMs was used for binary classification (no pain/pain). Celona et al. [56] applied the Histogram of Oriented Gradients (HOG) descriptor [57] on the COPE database, then SVM was used for classification (obtained accuracy = 81.75%). Recently, Convolutional Neural Networks [CNNs] extracted deep features showed good performance in several classification tasks [56,58,59].

Frame-level methods ignore temporal information and are thus limited in describing relevant dynamic information that is beneficial for pain intensity recognition. Further, occlusion, such as self-occlusion, oxygen mask, and pacifier, is another limitation of using such methods. Thus, many recent works focus on video-level pain recognition because it is more effective in describing such information [60–62]. It often uses temporal integration of frame-level features. For example, the video content can be condensed to high-level features by using a time series statistics descriptor that consists of several statistical measures of each individual frame.

Several facial feature descriptors have been proposed to analyze the spatio-temporal texture of facial videos, such as Local Phase Quantization (LPQ) [63] and Binarized Statistical Image Features (BSIF) [64]. LBP-TOP [65][64], LPQTOP [66], BSIF-TOP [67], HOG-TOP [68], and LGBP-TOP [69] were extended descriptors that used the Three Orthogonal Planes (TOP). Further, Werner et al. [60] and Kächele et al. [61] proposed the spatio-temporal descriptors based on appearance- and geometry-based facial features and head pose; the pain levels were classified using Random Forest (RF) [70] with those descriptors.

Other methods that have been used for pain recognition to detect pain expression from videos, such as a motion based method [71] (directly estimating the pixel's velocity over consecutive frames), a model based method [72,73] (search for the optimal parameters of a learning model that best match the model and the input data such as Active Appearance Model [AAM]), a FACS based method [74] (extract useful features from the videos using any Toolbox such as Computer Expression Recognition Toolbox [CERT], then apply classifier).

According to the ability of Random Forest (RF) for pain detection using facial expression [29,61,75–77], Othman et al. [76] introduced RFc as a baseline method and compared its performance to the proposed deep learning methods that analyze a RGB image encoding temporal information. RFc was used with time series statistics descriptor that was calculated from 16 statistical measures with their first and second derivatives per time series. RFc with Facial Activity Descriptor

(FAD) performed well compared to reduced MobileNetV2 (using transfer learning with the first five inverted residual blocks) and performed similarly to simple Convolutional Neural Network (CNN).

In [75], the performance of CNN with frontal RGB images was improved compared to RFc with FAD by about 1% when using the sample weighting method. Downweighting misclassified samples during training improves the performance; these samples often contain low or no facial responses to pain (see [78] for details of this phenomenon). Some training samples with more facial responses based on the classification score (score above 0.3) were duplicated. The performance improvement of the CNN model was not very high to classify seven classes (no pain and three phasic pain intensities for heat and electrical modalities). Nevertheless, the performance of deep models, to a large extent, depends on their size and the amount of training data. Different sorts of network architectures have been developed to increase the capacity of deep models.

In [59, 79, 80], the authors used various neural networks for pain recognition, including Convolutional Neural Networks (CNNs) [81] and Long-Short Term Memory (LSTM) [82]. Zhou et al. [83] introduced a Recurrent Convolutional Neural Network (RCNN) for predicting pain intensity in video sequences. A sliding-window strategy was used to extract features for obtaining fixed-length input samples for the recurrent network. However, the structural information of the face was missing due to the spatial conversion. Rodriguez et al. [59] considered a temporal relationship between video frames by integrating the extracted features from CNN to address this issue. Then they fed these features to a Long-Short Term Memory (LSTM) to exploit the temporal information.

Further, several hybrid deep learning methods have been proposed for pain recognition by combining CNN with LSTM [59, 84, 85] or CNN with Bidirectional LSTM [86, 87]. Tavakolian [40] presented a cross-architecture transfer learning to leverage the knowledge of pre-trained models to train other methods' architectures. They formulated pain intensity estimation as a self-supervised learning problem for the first time to exploit the abundant information of unlabeled data; they also introduced a video distillation method to encode the appearance and dynamic of the facial video into one RBG image map.

2.1.1.2 Vocalisations

Vocalizations are another indicators of pain, including verbal, non-verbal, and breathing behaviors. Vocalizations during pain experience are defined as the utterance of sounds, noises, and words using the vocal apparatus. Verbal vocalizations include protests and complaints by mentioning pain or using offensive words [88]. Behavioral pain indicators also include non-verbal vocalizations such as moaning, crying, groaning, and gasps [16]. The changes in breathing patterns have also been considered as vocalizations in regard to Waters et al. [89] study, such as sighing. The automatic recognition of pain using vocalizations consists of three main stages: (1) preprocessing, (2) feature extraction, and (3) pain recognition. In [90–96], very encouraging results were obtained showing that this indicator should get attention, which also plays an important role in related applications of affective computing [97]. The background noises are a major challenge in the analysis of audio data; the sounds of interest separation is required for better pain assessment by identifying the background noises and then removing them, which may originate from medical devices, other people, or events.

2.1.1.3 Body movements

Body movements such as bracing (holding onto an object, the fist, or the affected area during movement), knee bending, shoulder to front movements, rubbing (massaging the affected area), and restlessness (i.e., constant shifting in position) are behavioral responses to pain [88, 98]. Pain can also be expressed by moving the head towards the pain location with different speeds and ranges compared to other conditions [40]. The automatic pain recognition from body movement consists of three main steps: (1) preprocessing and body tracking, (2) feature extraction, and (3) pain recognition. In [91, 99–102], several body movements' methods were introduced for pain assessment. During pain experience, Zamzmi et al. [90, 103] measured body movement of neonates, whereas Werner et al. [41] analyzed head movements and postures [HMP] of adults in three pain datasets (BioVid, UNBC, and BP4D). HMP tends to be oriented downwards or towards the pain site and differs in the movement speed and range compared to other conditions.

2.1.2 Physiological Signals

Clinical studies [28, 104–107] have provided strong empirical evidence for the correlation between individual physiological signals and pain. The physiological signals are capable of indicating the state of the autonomic nervous system, and pain is one function of such systems. The pain process starts from the sensory receptors (also called nociceptors) by noxious thermal, chemical, or mechanical stimuli, which can be activated to the body from an external or internal source. The information regarding detecting harmful stimuli and converting these into electrical signals is transduced via nociceptors and transmitted through the spinal cord to the brain. Then, specific parts of the brain are responsible for responding to pain signals, which are the prosencephalon, mesencephalon, and cortex [108–111]. In this process, pain indicators cause alterations in tissues and organs (e.g., skin, heart, muscles' electrical properties). Physiological signals such as electrocardiogram [ECG], facial electromyography [EMG], and Electrodermal Activity [EDA] are most widely used in pain assessment [112, 113]. However, the changes in the physiological signals are also indicative of other pathological conditions unrelated to pain [114]. Thus, the combination of multiple behavioral and physiological pain indicators is potentially good for developing objective pain assessment in regard to Odhner et al. [115] and Hinduja et al. [116] study.

ECG captures the changes in the electrical activity of the heartbeats (low-frequency / high-frequency ratio) and the heart rate interval. The Heart Rate Variability [HRV] is calculated on ECG data; the changes of HRV in the low-frequency power increase during painful stimulation [117, 118]. EMG measures the changes in electrical properties of the muscle; EMG activity is often measured at the zygomaticus (mouth corner raiser), trapezius (back of the neck), and corrugator superscillii (brow lowerer) muscles. EMG has been used as an indicator for pain assessment [22]. EDA records the changes in the electrical activity of the skin when using two electrodes connected to the index and ring fingers. In response to a pain stimulus, EDA is a good measure because of intense body activity after experiencing pain; when a painful stimulus is applied, the sympathetic nervous system [SNS] activates the finger's sweat glands to produce more sweat, and this, in turn, increases skin conductance [119–121].

During the last years, researchers have shown great interest in investigating physiological signals and machine learning models for objective pain intensity assessment. Treister et al. [122] used a linear combination of multiple physiological sensors, including ECG, photoplethysmogram [PPG], and EDA to successfully differentiate between four categories of pain (no pain and three pain intensity induced by heat stimulator $P < 0.01$). Chu et al. [123] applied linear discriminant analysis [LDA] on blood volume pulse [BVP], ECG, and EDA signals to classify pain into no pain pre stimulate (calm), four different pain intensities induced by an electrical stimulator, and post-stimulate (post). They extended their study in [124], LDA, k-nearest neighbors (KNN), and support vector machines (SVMs) were applied to the same dataset with about 84.28%, 83.94%, and 96.47%, respectively. Walter et al. [125] presented the BioVid database to facilitate advances in robust recognition of pain and its intensity based on multiple physiological signals, including ECG, EMG, electroencephalography [EEG], and EDA signals. One hundred thirty-five features were extracted to train SVM to classify each of the four levels of pain intensity against no pain. Those features capture (1) amplitude, (2) frequency, (3) stationary, (4) entropy, (5) linearity, and (6) variability.

Gruss et al. [126] extended this work by extracting 159 features from the EDA, ECG, and EMG signals from the same BioVid dataset with the respective person-specific mean baseline signal to recognize induced head pain. They trained SVM to classify no pain and pain tolerance threshold (about 90.94%) or no pain and pain threshold (about 79.29%). Kächele et al. [127] also used the same physiological signals (EDA, ECG, and EMG) from the same database (BioVid), together with meta-information and similarity. A random forest classifier was trained to classify no pain and a specific pain level. Several recent studies focus on deep learning due to its success in various domains; researchers have recently investigated its application in pain recognition. Lopez-Martinez et al. [128] implemented a multi-task learning method based on neural networks that accounted for individual differences in pain responses and achieved a classification accuracy of 82.75% for baseline vs. pain tolerance threshold and 54.22% for baseline vs. pain threshold using Skin Conductance Level [SCL] and ECG features. Kächele et al. [129] advanced pain assessment task, their work focused on continuous pain estimation by training adaptive Random Forest; they treated pain intensity as a continuous variable instead of as an ordinal variable with fixed categories.

Recent studies have drawn attention to the EDA signal due to its significant correlation with pain intensity ratings. It quite consistently performs best in terms of the automatic system compared to other single physiological signals [22,29,30,113,124,129–132]; moreover, it is easy to use. Lopez-Martinez et al. [133] also presented a recurrent neural network method to continuously estimate pain intensity with EDA signal from the BioVid dataset. Thiam et al. [134] improved the binary pain classification when using 1D convolutional neural networks on raw EDA signals from the BioVid dataset with minimum preprocessing. Posada et al. [135] presented classification and regression machine learning models to estimate pain sensation in healthy subjects using EDA. They computed the extracted features of EDA based on time-domain decomposition, spectral analysis, and differential features. The maximum macro averaged geometric mean scores of models were 69.7% and 69.2%, respectively. Kong et al. [136] analyzed the spectral of EDA to obtain reliable performance because it is more sensitive and reproducible for the assessment of sympathetic arousal than traditional indices (tonic and phasic signals). Bhatkar et al. [137] reported a successful novel method to discriminate the reduction of pain with clinically effective analgesics by combining self-report with continuous physiological data in a structured and specific-to-pain protocol.

2.1.3 Fusion

There is a growing number of researches investigating the fusion of the data. Werner et al. [22] and Walter et al. [138] combined (early and late fusion) visual features (facial expression) with physiological signals (Galvanic Skin Response [GSR, also called SCL], EMG, and ECG) from the BioVid dataset, and Random Forest classifier (RFc) was used to increase the accuracy of distinguishing baseline vs. pain tolerance threshold and baseline vs. pain threshold. Other studies [29,30] also proposed a multi-modal information fusion approach based on RFc using video, audio, and physiological features from X-ITE Pain Database. The late fusion (decision fusion) improves results further. Zamzmi et al. [90, 139] proposed KNN, SVM, and RF to assess pain in infants based on analysis of facial expression, body movement, cry sound, and vital signs (HR: heart rate, SpO2: blood oxygenation, BP: blood pressure). Three pain levels (no pain, moderate pain, and severe pain) were recognized using single and combined indicators. The multi-modal results outperformed the single model significantly.

Thiam et al. [77] proposed several fusion architectures to develop a multi-modal pain intensity classification. The estimation is based on the SenseEmotion Database, and the accuracy improved in 2-class, 3-class, and 4-class pain intensity classification tasks. Thiam et al. [6] advanced the binary pain classification task by proposing a new multi-modal information fusion method based on deep denoising convolutional autoencoders on EDA, EMG, and ECG signals from the BioVid dataset. They extended their work in [140] by introducing multi-modal methods (supervised deep learning method and self-supervised method) for recognizing pain intensity based on physiological signals. The self-supervised method automatically generated physiological data and simultaneously performed a fine-tuning of the deep learning model, which had been previously trained on a significantly smaller amount of data. Thus, they were able to significantly improve the data efficiency.

Yu et al. [141] proposed a diverse frequency band-based ConvNets for tonic cold pain states classification using EEG signal from their database, which provides higher accuracy than state-of-the-art techniques. First various feature representations were extracted from different frequency bands. Then these features were concatenated and fed into a fully connected network that classified pain states with no pain and two tonic pain levels (moderate and severe). Wang et al. [142] introduced a bidirectional Long Short-Term Memory network (biLSTM) to learn the temporal dynamics from physiological signals in the BioVid dataset. The RNN-generated features with a set of hand-crafted features were fused for binary pain classification tasks.

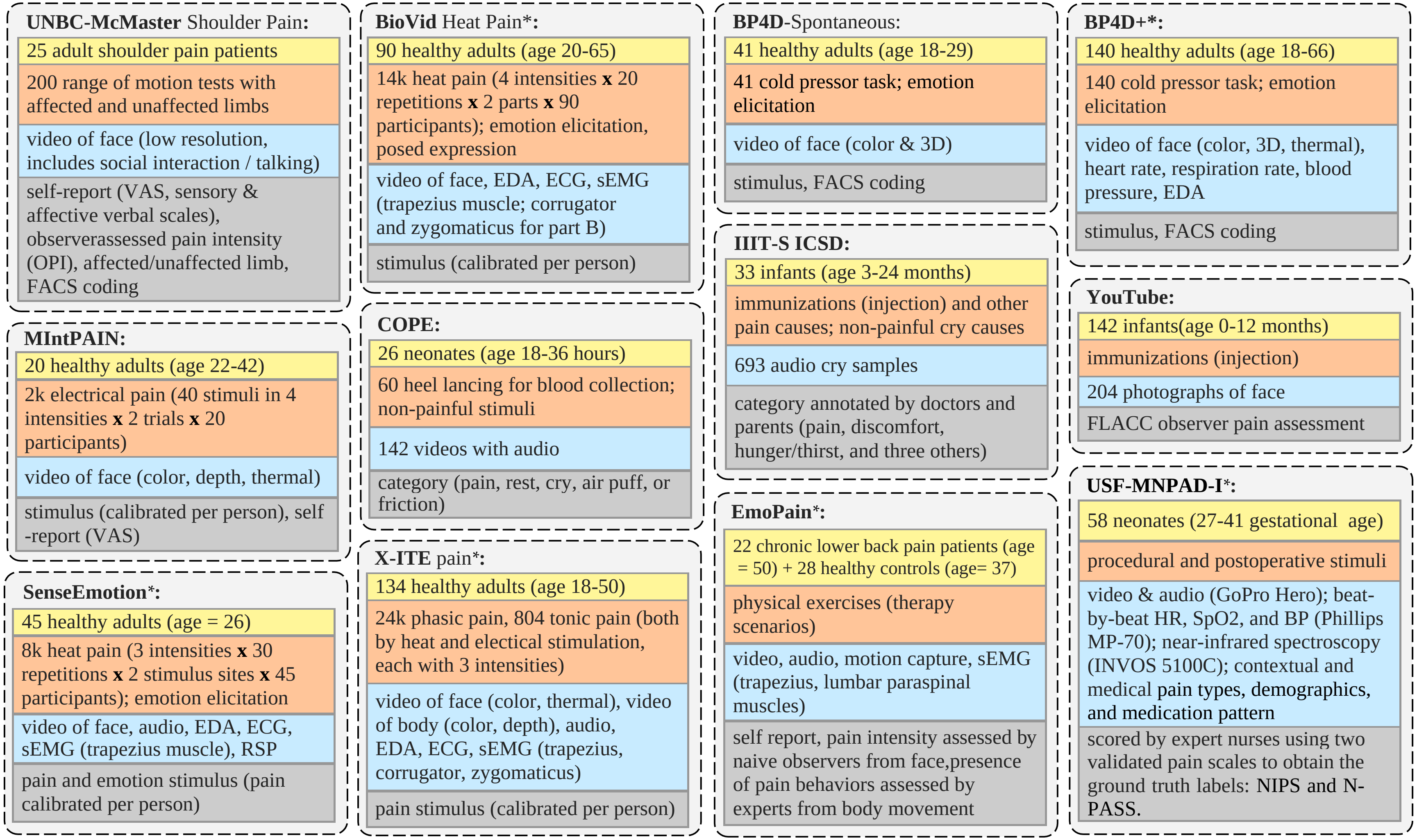
Subramaniam et al. [132] proposed a multi-modal hybrid Deep Learning network (CNN-LSTM) using physiological signals (ECG and EDA) from BioVid dataset for binary pain classification. The obtained results outperformed the unimodal results. Hinduja et al. [116] introduced a multi-modal method for pain recognition by fusing facial expressions and physiological signals (HR, respiration, BP, and EDA) from BP4D+ database. The fusion improved accuracy when evaluation included all subjects or same gender compared to using only one modality (facial expressions or physiological signals). Pouromran et al. [113] focused on pain intensity estimation using the BioVid dataset; they built different machine learning models for continuous pain estimation: Linear Regression, Support Vector Regression (SVR), Neural Networks, KNN, Random Forest, and XGBoost. They used the extracted features from a single sensor and the combination features from multiple sensors. The EDA single model outperforms multi-modal results for pain intensity recognition, and SVR gave the best predictive performance across different sensors.

2.2 Pain Databases

Several databases are designed and released for automatic pain recognition methods in computer vision and machine learning domains, which are from oldest to newest: COPE Database [46], UNBC-McMaster Shoulder Pain Data-base [143], BioVid Heat Pain Database [112], BP4D-Spontaneous Database [144], YouTube Database [145], BP4D+ Database [146], IIIT-S ICSD [147], SenseEmotion Database [148], Multimodal EmoPain Database [91], Mint PAIN Database [79], X-ITE Pain Database [149], and Multi-modal Neonatal Pain Assessment Dataset [USF-MNPAD-I] [150]. Most approaches of pain recognition use a single sensor modality: [60, 69] use video, [93, 151] use audio signal, and [124, 125, 133] use physiological signals, but recent approaches use multiple sensor modalities [29, 130, 134] that can improve the performance and flexibility of pain recognition. Fig. 2.2 shows more details about those databases.

2.3 Training Techniques

When training machine learning models, the input data must be split regarding image level or subject level separation into training, validation, and testing sets. The training set is used for training the model; the validation set is used for tuning the hyper-parameters and selecting models, whereas the testing set is used for performance evaluation. Each split cannot contain any of the same samples; otherwise, the samples in multiple splits will always almost classified correctly, giving falsely high performance. Most recent databases have moved towards larger training sets, e.g., 80% of the data set would be in the training set, 10% in the validation, and 10% in the testing. Image level separation means all splits often contain images that belong to the same subject. In the real scenario, the pain recognition systems should work well with subjects that have never been seen before; therefore, image level separation is generally not preferred due to bias with known subjects. Thus, subject level separation is suitable for pain recognition because the validation and testing subjects are completely ignored in the training set and never seen by the system before validation and testing. This method is used in this thesis as it provides a performance to unknown subjects, which is expected in field trials. Cross-validation is also another helpful technique for assessing the effectiveness of models to mitigate overfitting. There are several types of cross-validation techniques, and all of them have similar steps. First, the dataset is divided into training and test parts (sets).



UNBC-McMaster Shoulder Pain:

25 adult shoulder pain patients

200 range of motion tests with affected and unaffected limbs

video of face (low resolution, includes social interaction / talking)

self-report (VAS, sensory & affective verbal scales), observerassessed pain intensity (OPI), affected/unaffected limb, FACS coding

MIntPAIN:

20 healthy adults (age 22-42)

2k electrical pain (40 stimuli in 4 intensities **x** 2 trials **x** 20 participants)

video of face (color, depth, thermal)

stimulus (calibrated per person), self -report (VAS)

SenseEmotion*:

45 healthy adults (age = 26)

8k heat pain (3 intensities **x** 30 repetitions **x** 2 stimulus sites **x** 45 participants); emotion elicitation

video of face, audio, EDA, ECG, sEMG (trapezius muscle), RSP

pain and emotion stimulus (pain calibrated per person)

BioVid Heat Pain*:

90 healthy adults (age 20-65)

14k heat pain (4 intensities **x** 20 repetitions **x** 2 parts **x** 90 participants); emotion elicitation, posed expression

video of face, EDA, ECG, sEMG (trapezius muscle; corrugator and zygomaticus for part B)

stimulus (calibrated per person)

COPE:

26 neonates (age 18-36 hours)

60 heel lancing for blood collection; non-painful stimuli

142 videos with audio

category (pain, rest, cry, air puff, or friction)

X-ITE pain*:

134 healthy adults (age 18-50)

24k phasic pain, 804 tonic pain (both by heat and electical stimulation, each with 3 intensities)

video of face (color, thermal), video of body (color, depth), audio, EDA, ECG, sEMG (trapezius, corrugator, zygomaticus)

pain stimulus (calibrated per person)

BP4D-Spontaneous:

41 healthy adults (age 18-29)

41 cold pressor task; emotion elicitation

video of face (color & 3D)

stimulus, FACS coding

IIIT-S ICSD:

33 infants (age 3-24 months)

immunizations (injection) and other pain causes; non-painful cry causes

693 audio cry samples

category annotated by doctors and parents (pain, discomfort, hunger/thirst, and three others)

EmoPain*:

22 chronic lower back pain patients (age = 50) + 28 healthy controls (age= 37)

physical exercises (therapy scenarios)

video, audio, motion capture, sEMG (trapezius, lumbar paraspinal muscles)

self report, pain intensity assessed by naive observers from face,presence of pain behaviors assessed by experts from body movement

BP4D+*:

140 healthy adults (age 18-66)

140 cold pressor task; emotion elicitation

video of face (color, 3D, thermal), heart rate, respiration rate, blood pressure, EDA

stimulus, FACS coding

YouTube:

142 infants(age 0-12 months)

immunizations (injection)

204 photographs of face

FLACC observer pain assessment

USF-MNPAD-I*:

58 neonates (27-41 gestational age)

procedural and postoperative stimuli

video & audio (GoPro Hero); beat-by-beat HR, SpO2, and BP (Phillips MP-70); near-infrared spectroscopy (INVOS 5100C); contextual and medical pain types, demographics, and medication pattern

scored by expert nurses using two validated pain scales to obtain the ground truth labels: NIPS and N-PASS.

FIG. 2.2. Characteristics of publicly available benchmark datasets for automatic pain assessment. ECG: electrocardiogram, EDA: electrodermal activity, sEMG: surface electromyography, FACS: Facial Action Coding System, RSP: Respiration, HR: heart rate, SpO2: blood oxygenation, and BP: blood pressure. In USF-MNPAD-I: procedural and postoperative stimuli such as immunizations and after surgical procedure, respectively; the two validated pain scale are NIPS [152] (procedural) and N-PASS [153] (postoperative). The different background color indicates the characteristics of the databases: yellow for subjects, orange for stimuli, and blue for data modalities, and grey for annotation.* multimodal database. Modified, with permission, from [26].

Second, the model is trained on the training set. Third, the model is validated on the test set. Those steps are repeated a couple of times; this number depends on the selective cross-validation technique. K-Fold cross-validation is commonly used. The data set is broken into K sets; each time, a different set forms the test set, and the training set is the rest sets. The final performance is the average performance of models obtained from K sets. Further, leave-one-subject-out validation is a subset of cross-validation, which avoids subject overlap between the training and the test sets and approximates generalization performance with unseen subjects. The cross-validation process is then applied N times, with each subject being used exactly once as the test set. In total, N subjects provide N training sets and N test sets. Finally, the results of test sets are averaged to form the final performance. To train N models, the total number of sets increases, and that is expensive. Hence, this technique was not used in this work.

2.4 Conclusion

This chapter summarizes the important background for the state of automated methods for pain assessment, which have not yielded estimation accuracies acceptable in clinical settings due to several limitations. These limitations can be summarized as follows:

- Half of the available pain databases for research purposes contain response data from a single modality (e.g., facial expressions); a large number of the current methods assess pain using a single modality. However, studies have shown that it is better to use a combination of behavioral and physiological signals to obtain a reliable system for pain assessment.

- The multimodal pain databases have a significant impact on the performance of automatic pain assessment systems, many databases have been used in recent researches. To the common belief that the quality and duration of pain provide additional valuable information for more advanced discriminating pain or pain intensities versus no pain, the X-ITE Pain Database has been made to complement existing databases and the analysis of pain regarding quality and length.

- There are some studies to assess pain based on using more than three modalities. The results from those studies are promising; they show that it is possible to obtain a reliable pain assessment system by analyzing pain and detecting valid pain patterns from multiple modalities, including both behavioral and physiological signals.

- Few of the current pain assessment methods focus on continuous monitoring, which is more necessary for pain assessment for prompt pain detection and immediate intervention.

This thesis addresses the above-mentioned limitations and proposes an automatic and multimodal system for continuous monitoring of pain intensity using five sensor modalities (frontal RGB camera, audio, ECG, EMG, EDA) from the X-ITE Pain Database. A combination of behavioral and physiological signals was used with appropriate machine learning methods. RF was utilized with a single or multiple fused modalities [facial expression, audio, ECG, EMG, EDA, (facial expression and EDA), (EMG and EDA), or all modalities] as a baseline method to predict continuous phasic or tonic pain intensity versus no pain. Further, the previously mentioned modalities were used with the proposed LSTM and LSTM-SW for better handling time series prediction as an advanced continuous recognition method. So this thesis advanced over [75] by investigating a more complex problem (classify phasic and tonic pain in sequence level) with a single modality and multiple fused modalities. In addition, a comparison between classification and regression was presented of monitoring continuous pain intensity.

CHAPTER 3

Fundamentals

THIS chapter describes the necessary tools, fundamentals of algorithms, and methods exploited throughout this thesis for monitoring continuous pain intensity. The chapter is divided into three sections. The first section presents a detailed description of the tools and algorithms used to analyze five modalities' signals; see Section 3.1. In the second section, the time series statistics descriptor is described since it is employed with all proposed machine learning methods, see Section 3.2. The third section provides a brief overview of two machine learning methods, which are exploited for the classification and regression purposes within the thesis (see Section 3.3). Finally, the activation and loss functions, which were used, were described in Section 3.4 and Section 3.5, respectively.

3.1 Sensor Signal Processing

The tools and algorithms for sensor signal processing are:

3.1.1 OpenFace

Fig. 3.1 shows the OpenFace facial behavior analysis pipeline. Baltrušaitis et al. [154] introduced an easy and first open source toolbox that use to analyze facial behavior. OpenFace tool extracts facial landmark motion, head pose (orientation and motion), facial expressions, and eye gaze as important models to understand human behavior. It comprises several technologies, including facial landmark detection and tracking, head pose and eye gaze estimation, and action unit detection.

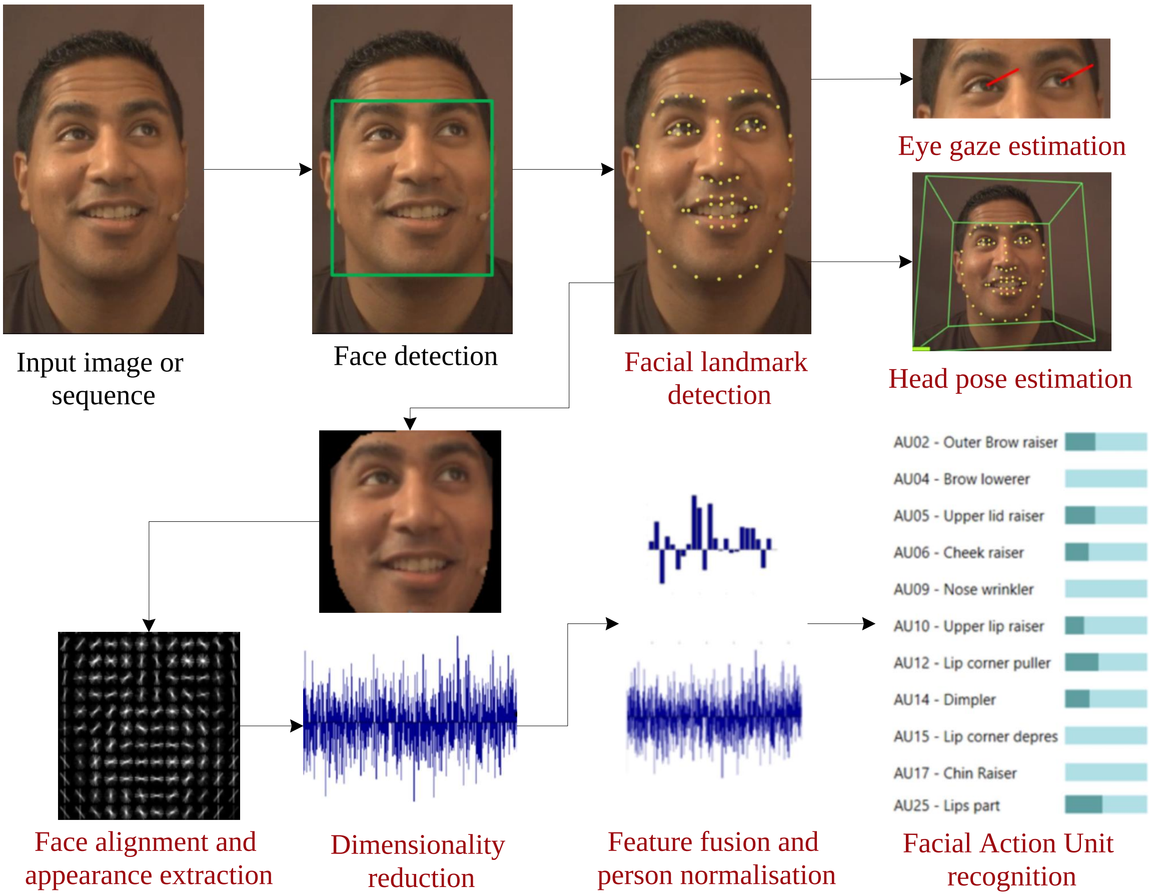


FIG. 3.1. The pipeline of the OpenFace facial behavior analysis, with permission from [154]. This system includes facial landmark detection, head pose & eye gaze estimation, and facial action unit recognition. The outputs (indicated by red) are saved to send over a network.

3.1.1.1 Facial Landmark Detection and Tracking

The face is detected in an image and tracked in the video. The landmarks are detected with individual models using Conditional Local Neural Fields (CLNF) [155] by training separate sets of point distribution and patch expert models for nose, jaw, eyes, lips, and eyebrows. The CLNF patch experts are trained on three training sets: Multi-PIE [156], LFPW [157], and Helen [158]. Each separate set of patch experts is trained for seven views and four scales, which (1) allows for accurate detection of landmarks on a low- and high-resolution face image and (2) tracks faces with out of plane motion and to model self-occlusion caused by head rotation. These detected landmarks fit to a joint Point Distribution Model (PDM) [159], which provides 40 shape parameters. For training, the PDM, LFPW, and Helen training sets are used.

To initialize the CLNF model, the dlib face detector in [160, 161] is used to draw a simple linear mapping from the bounding box surrounding the detected 68 facial landmarks. Further, to track the landmarks in the video, the CLNF model is initialized based on landmark detections in the previous frame. In order to deal with tracking drift, a simple three-layer convolutional neural network (CNN), which gives a face aligned using a piecewise affine warp, is trained with correct and randomly offset landmark locations to predict the expected landmark detection error. In case of landmark detection fails to track according to the validation module reports from CNN, the model needs to be reinitialized. For more details about landmark detection and tracking algorithm, see Baltrušaitis et al. [154, 155].

3.1.1.2 Head Pose and Eye Gaze Estimation

The CLNF internally is applied for head pose estimation, including 3D representation of facial landmarks. These landmarks are projected to the image using orthographic camera projection. Once the landmarks are detected, the head pose (translation and orientation) is estimated. Further, CLNF is used for eye-region landmark detection, including registration of eyelids, iris, and the pupil. The eye gaze vector, which is calculated individually for each eye, is computed once the location of the eye, and the pupil are detected using the CLNF model. The gaze vector is provided from the center of the 3D eyeball to the pupil in the image plane. The PDM and CLNF patch experts are trained on the SynthesEyes training dataset [162]. Both methods are fast and accurate for a person independent of head pose and gaze estimation in webcam images. In this work, three head pose features were used, which were yaw, pitch, and roll. For more details about head pose and gaze estimation algorithm, see Baltrušaitis et al. [154].

3.1.1.3 Action Unit Recognition

The Action Units (AUs) are extracted from webcam images based on geometry features (shape parameters and landmark locations) and appearance (Histograms of Oriented Gradients), including 33 facial features referring to the occurrence (presence) and the intensity of AUs (see Table 3.1). To analyze the texture of the face, the similarity transformation is used in representing frontal landmarks from a neutral expression (a projection of mean shape from a 3D PDM). Then, appearance features are extracted using Histograms of Oriented Gradients (HOGs) [163] from the obtained results (112 ×112 pixel image of the frontal face with 45 pixels interpupillary distance). The dlib [160] implementation of HOGs is used with blocks of 2 ×2 cells of 8 ×8 pixels, leading to 12×12 blocks of 31 dimensional histograms (4464 dimensional vector describing the face). The Principal Component Analysis (PCA) is used to reduce the feature dimensionality keeping 95% of explained variability on a number of facial expression datasets (CK+ [164], DIS-FA [165], AVEC 2011 [166], FERA 2011 [167], and FERA 2015 [168]). This led to a reduced basis of 1391 dimensions, which allows for a generic basis, model train, and no need to recompute the PCA to unseen datasets. Further, the geometry features are extracted using the non-rigid shape parameters and landmark locations in object space inferred during CLNF model tracking, which led to a 227 dimensional vector. In regards to classification and regression using OpenFace, the linear kernel SVM and linear kernel SVR [169] are used to predict AU presence and AU intensity in this sequence, respectively. Table 3.1 shows the list of OpenFace recognizing the AU (I- intensity, P-presence).

Table 3.1. List of extracted facial features in OpenFace (I-intensity, P-presence).

AU	AU Full Name	Prediction		AU	AU Full Name	Prediction	
AU1	Inner brow raiser	I	P	AU14	Dimpler	I	P
AU2	Outer brow raiser	I	P	AU15	Lip corner depresser	I	P
AU4	Brow lowerer	I	P	AU17	Chin raiser	I	P
AU5	Upper lid raiser	I	P	AU20	Lip stretched	I	P
AU6	Cheek raiser	I	P	AU23	Lip tightener	I	P
AU7	Lid tightener	I	P	AU26	Jaw frop	I	P
AU9	Nose wrinkler	I	P	AU28	Lip suck	-	P
AU10	Upper lip raiser	I	P	AU45	Blink	I	P
AU12	Lip corner puller	I	P				

Both the dimensionality reduced HOGs, and the facial shape features (from CLNF) are combined and used as the feature vector of the appearance of the face. For estimating a neutral expression, the median value of the features in a video sequence of a person is computed, then extracted value subtracted from the estimates in the current frame leading to a normalized feature. This cheap and effective method to increase model performance is presented in [170].

3.1.2 OpenSMILE

Eyben et al. [171] introduced a novel open-source feature extractor toolkit (OpenSMILE) for audio signal processing. OpenSMILE can be used to extract acoustic features from speech, music, and general sound events. There are three main components involved in OpenSMILE architecture: data memory for reading the data from external sources, data processors for reading and modifying data from the data memory, and data sinks for reading the data from data memory and writing it to external files or perform classification, see Fig. 3.2.

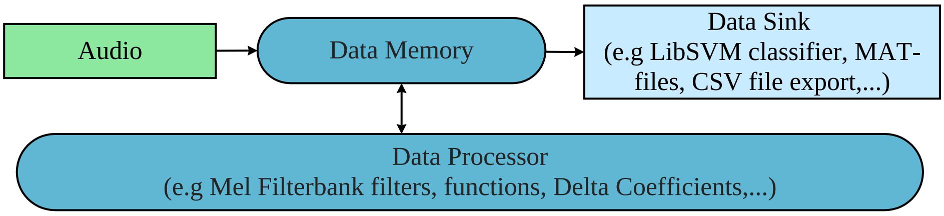

Audio
Data Memory
Data Sink
(e.g LibSVM classifier, MAT-files, CSV file export,...)
Data Processor
(e.g Mel Filterbank filters, functions, Delta Coefficients,...)

FIG. 3.2. Overview of openSMILE's architecture.

The audio signal processing is presented in three levels, which are wave (sampled at 16–44.1 kHz), frames (overlapping frames of 50 ms length at 20 ms frame rate [50 fps]), and pitch (extracting pitch features from the frames). The components process the framed audio vectors after converting them of size 16 kHz × 50 ms = 800. Various filters, functionals, and transformations are applied to the obtained low-level features. Fig. 3.3 shows the group of features that use for robust speech emotion recognition. The Mel-frequency features, Mel-Spectrum and Mel-Frequency Cepstral Coefficients (MFCC), as well as the Perceptual Linear Predictive Coefficients (PLP) are common features computed in full accordance with the popular Hidden-Markov Toolkit (HTK) [172].

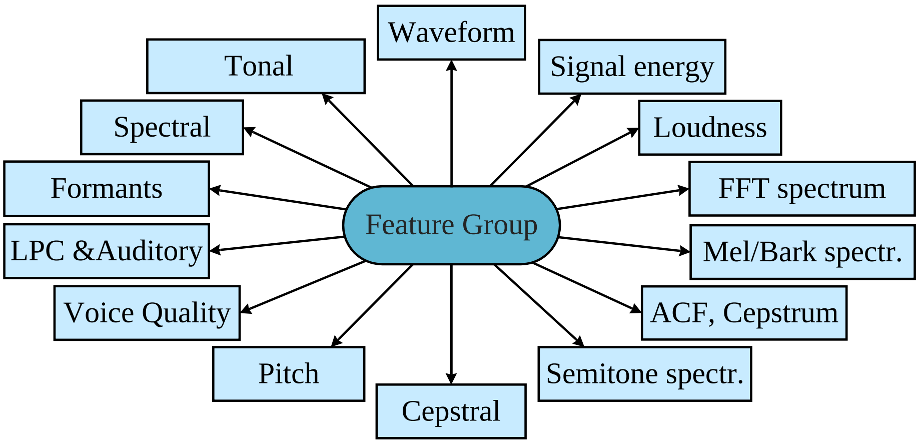


FIG. 3.3. The openSMILE's low-Level descriptors.

An overview of the 65 low-level descriptors (LLD), those provided in the COMPARE acoustic feature set, is given in Table 3.2; see [171, 173, 174] for full details. Delta regression coefficients are computed from LLD, and a moving average filter is applied to smooth LLD contours. Next, the functionals, which belong to the ComParE (Computational Paralinguistics Evaluation) set, are used for the LLD contours, including mean, standard deviation, percentiles and quartiles, linear regression functionals, and local minima/maxima related functionals. ComParE comprises high dimensional brute-force acoustic feature sets (6373 features) of LLD contours.

3.1.3 QRS-detection algorithm, R-to-R intervals, and Linear Interpolation

Pan et al. [175] and Hamilton et al. [176] developed the QRS detection algorithm using the optimized decision rules. The QRS detector is divided into preprocessor and decision rule sections, see Fig. 3.4.

The preprocessor section involves three stages: filtration, peak detection, and fiducial mark location. In the filtration stage, the ECG signal is processed using a

Table 3.2. The prosodic, spectral, cepstral, and voice quality LLD's provided in the ComParE acoustic feature set. See [173].

4 energy related LLD	**Group**
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0-8 k Hz)	spectral
MFCC 1–14 cepstral	cepstral
Spectral energy 250–650 Hz, 1 k–4 k Hz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice quality
log. HNR, Jitter (local & DDP), Shimmer (local)	voice quality

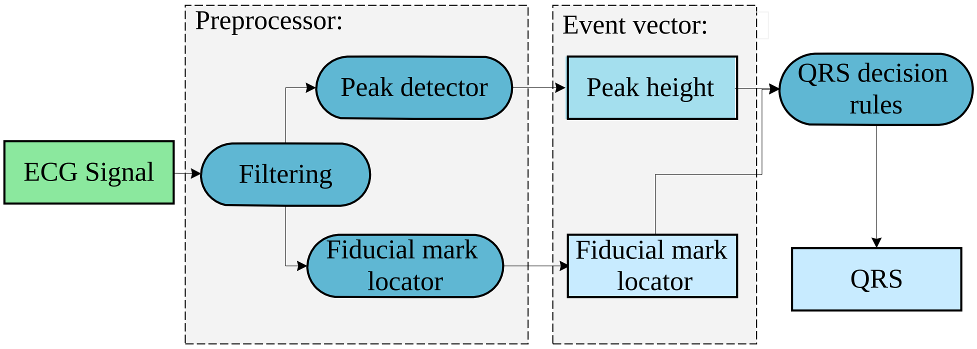


FIG. 3.4. Block diagram of QRS-detector algorithm.

band-pass filter to reduce the influence of muscle noise, 60 Hz interference, baseline wander, and T-wave interference; the desired passband is 5-15 Hz. The band-pass filter is obtained with a low-pass filter and a high-pass filter to reduce ripples and multiple peaks before peak detection. After filtering, the signal is differentiated to provide the QRS-complex slope information. Afterward, the nonlinear squaring is implemented on the processed signal. Finally, time averaging is done by calculating the mean of 32 most recent values from the squaring function. A separate derivative of the original ECG signal is used for the sampling period wave discrimination.

In peak detection, the peaks with time occurrence are detected using a detection algorithm on the final output from time averaging. The peak detection algorithm is developed to eliminate multiple ripples on large waves and also very small noise peaks: (1) the maximal levels in the processed signal since the last detected peak are stored, (2) the new peak is determined if the height of its level is less than half of the maximum level. The peak detection may be delayed relative to the wavelength duration. The valid peaks occur until the middle of the falling slope when the level drop below half the distance from the maximal value to the base point. To avoid this delay, in the fiducial mark location stage, the time of occurrence of the wave's peak is located with a fixed delay back in time from the point of peak detection. The reason is the time between the middle of the rising slope and the middle of the falling slope in the wave was equal to the duration of the averaging time.

Further, the peak detection occurs with a wave in the band-passed signal in intervals 225 to 125 ms preceding a peak detection in the time average signal. The valid fiducial mark is identified from the location of the largest peaks; the three-point scheme is used to detect the peaks in this signal for a more consistent location. For the late detection, the fiducial mark is set from the long wave in the band-passed signal in intervals 250 to 150 ms preceding a peak detection. After preprocessing the signal, the decision rule section operates different rules to discriminate the QRS events from the noise events. A two-dimensional event vector is determined and saved for each detected peak, including the peak signal level of the preprocessed waveform and the elapsed time from the last fiducial mark. Further, the status of the resulting event vector, whether from noise or QRS complex, is saved in a flag. In the decision rule section, three peak-level estimators are applied to the peaks derived from the time-averaged signal: mean a specified number of past peaks, median peak level, and iterative peak level.

The mean square prediction errors are calculated from these different estimators to evaluate the performance of the three predictors applied to QRS peaks. Next, the best method for estimating peak levels (median) is determined to set the detection threshold between the noise level estimate and the QRS peak level estimate. This method is tested for calculating detection thresholds with median peak level estimators.

After detecting QRS with median peak level estimation and a threshold between the noise and peak estimate, 200 ms refractory blanking is used to eliminate false detections on the sample period of wave and multiple detections of the QRS complex. Afterward, the reverse search and the optimization of the relative reverse search and normal thresholds are applied. The mean, median, and iterative predictions of the RR interval are tested, and the median method with eight-point performs the best. For more details about the QRS-detector algorithm, see [175, 176].

In this work, after detecting the R-peaks in the processed ECG signal using the QRS-detector algorithm, the heart rate is calculated from the R-to-R intervals, see Fig. 3.5 and see Eq. 3.1. Then, the linear interpolation is applied to replace missing data in the heart rate signal, Fig. 3.6 and see Eq. 3.2.

$$HeartRate(HR) = \frac{60000}{\text{R-to-R interval } (ms)} \tag{3.1}$$

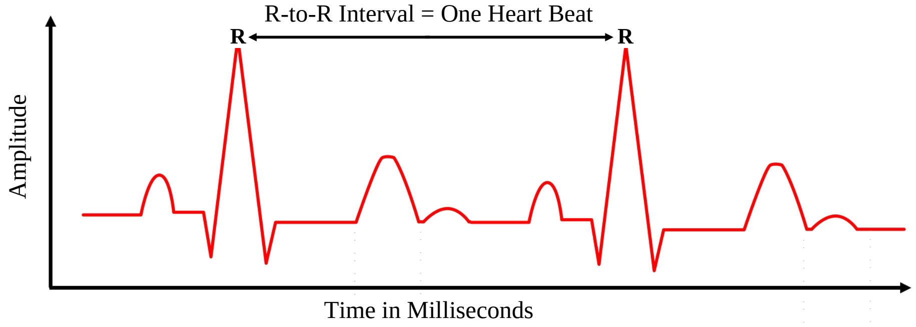


FIG. 3.5. R-to-R intervals.

$$LinearInterpolation = \frac{(x - x_1)(y_2 - y_1)}{(x_2 - x_1)} + y_1 \tag{3.2}$$

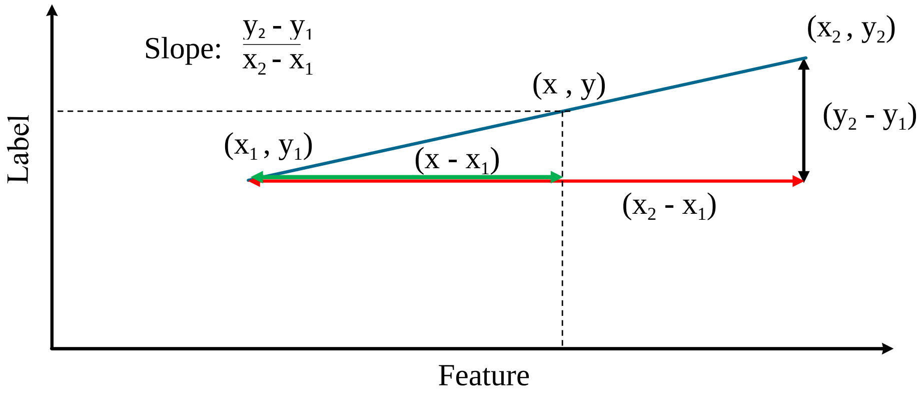


FIG. 3.6. Linear interpolation.

3.1.4 Zero-phase 3rd-order Butterworth Band-pass Filter

This section describes Zero-phase 3rd order butterworth band-pass filter, which is used to remove unwanted frequencies and reduce background noise from surface electromyography (EMG) signal. This filter passes all frequency signals in a specific cut-off range and rejects signals in other frequency ranges. The pass-band and zero rolled off response in the stop-band have maximally flat (no ripples) frequency responses. All frequencies until the cut-off range of high pass frequency from the inputted signal were allowed to pass, and then these frequencies are rolled off based on the rate of the 3rd order filter. Further, all frequencies lower than the cut-off range of low pass frequencies are removed (see Fig. 3.7). For more details, see [177].

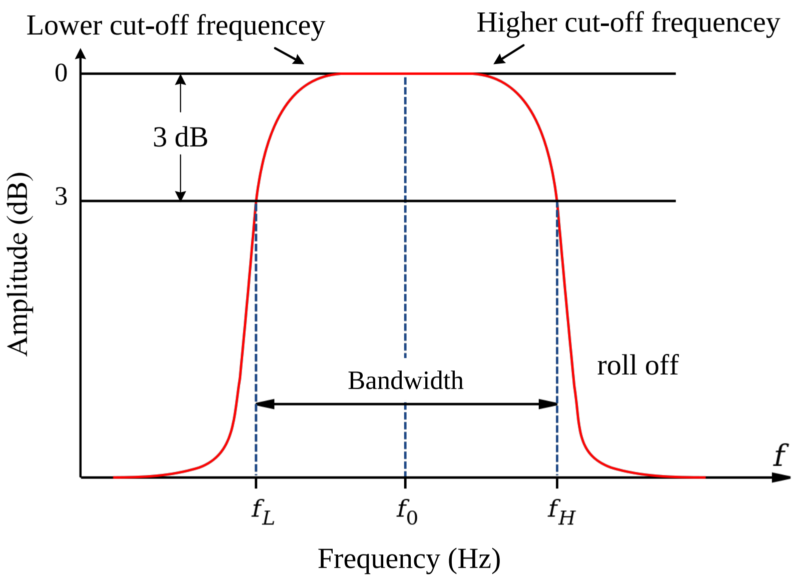


FIG. 3.7. The frequency characteristics of Zero-phase 3rd-order Butterworth Band-pass Filter.

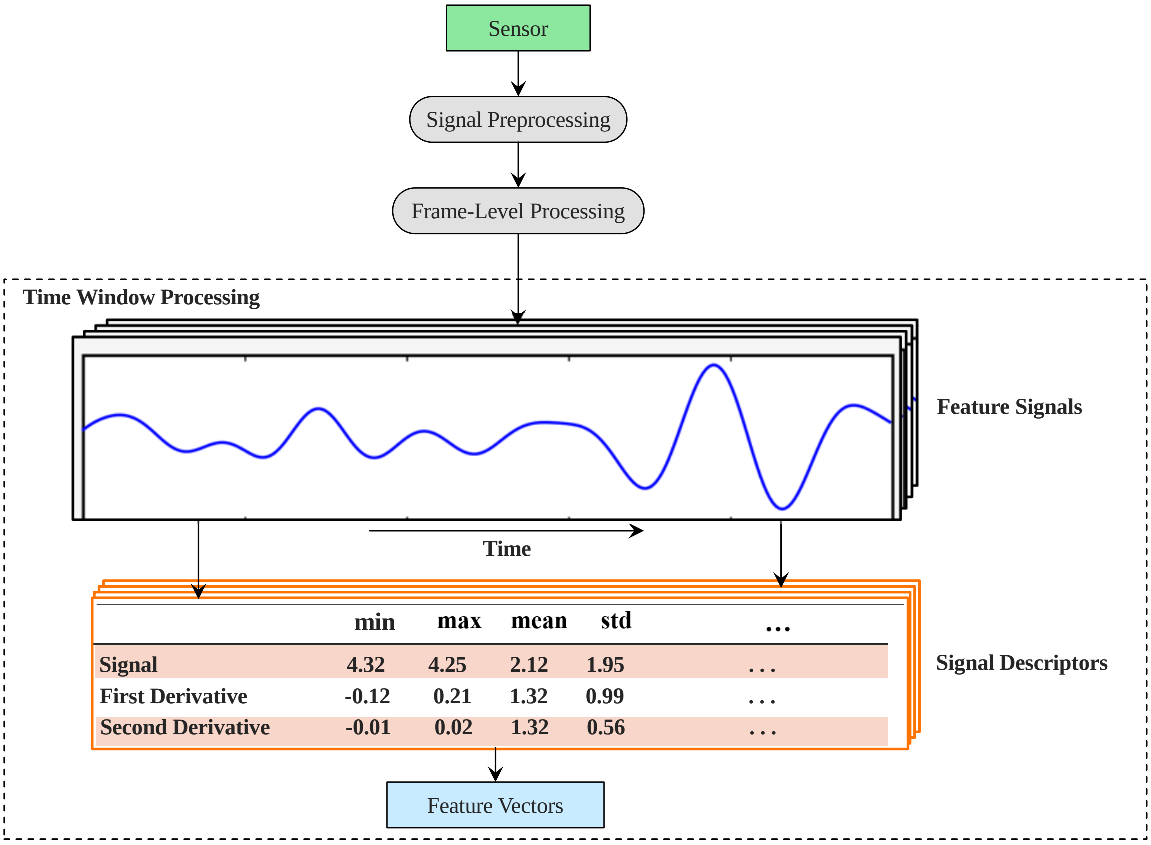
3.2 Time series Descriptors

The temporal information is well determined in the time window of the extracted feature from signals by using a time series statistics descriptor. In [60], the feature signals in descriptors are obtained by gathering the frame-level features per time series. In this thesis, only four statistical measures in Table 3.3 were used: min, max, mean, and SD. Then, the feature signal is smoothed using a first order Butterworth filter with a cutoff of 1 Hz.

Table 3.3. Signal descriptor methods.

Variable	**Description**	**Domain**
mean median min max	mean value of signal median value of signal minimum value of signal maximum value of signal	value
range SD IQR IDR MAD	range of signal standard deviation inter-quartile range of signal inter-decile range of signal median absolute deviation of signal	Value variability
tmax	instant of time when signal is its maximum	time
TGM TGA	duration the signal is greater than mean duration the signal is greater than average of mean and min	duration
SGM SGA	number of segments where the signal is greater than mean number of segments where the signal is greater than mean and min	count
area areaR	Area between signal and its minimum Quotient of area and area between max and min	Value × duration

Next, several statistical measures are calculated from the smoothed signal with their first and second derivatives for a given time period. These methods are used to extract many features from each signal, such as the state, speed, variability, duration, and acceleration signal. 3 × 16 (number of measures) is the descriptor dimension per time series. Afterward, the mean and SD for each subject (participant) is calculated from the feature signal due to the different pain sensitivity between subjects. The feature signal is subtracted from the mean and divided by SD, which are computed from the same subject. This process is called a person-specific standardization of the feature signal, see [60]. Finally, after concatenating the signal descriptors, the obtained feature vector was fed into some classification or regression methods in machine learning. Fig. 3.8 shows the overview of the signal descriptors. For more details, see [60].



	min	max	mean	std	...
Signal	4.32	4.25	2.12	1.95	...
First Derivative	-0.12	0.21	1.32	0.99	...
Second Derivative	-0.01	0.02	1.32	0.56	...

FIG. 3.8. Overview of the signal descriptors.

3.3 Machine Learning

Machine learning methods such as supervised learning methods are used for classification and regression problems. They are capable of performing the learning and testing to discover patterns in labeled datasets. These methods are used to learn the mapping function from the input to the output of datasets and then use them to predict the output for unlabeled datasets. Throughout this work, two types of supervised machine learning methods were used: classification and regression. The classification methods categorize the data by predicting discrete labels without using information about the order of classes. In contrast, the regression methods distinguish data into continuous real values instead of using classes or discrete values. Such methods predict continuous labels and exploit their ordinal relationship. For more details about the methods that were used in this study, see the Random Forest (RF) in Section 3.3.1 and Long-Short Term Memory Network (LSTM) in Section 3.3.2.

3.3.1 Random Forest

Random forest (RF) is constructed from decision trees that are introduced in [178]. A decision tree consists of three components: a first node (root), internal nodes, and end nodes (leaves). The training dataset is used to build a tree structure that recursively divides the space into regions with similar labels. The decision tree grows, including the decision on which features to choose, where each node except the leaves splits into two or more branches outputs (subspaces). The last branches are located on the leaves when it is no longer possible to break further. Three common stopping criteria of the decision tree growth are the maximum depth, minimum number of samples in a node, and a purity node. The pure node includes the data from a single class only. Such splitting minimizes the node impurity I(N). To decide the best feature split, entropy, gini, and misclassification impurity are the most common methods regarding classification for measuring node impurity using decision trees, the gini impurity is the default for both RF and decision tree, see Eq. 3.3. C is the total classes and $P(i)$ of picking a datapoint with class i.

$$I(N) = \sum_{i=1}^{(C)} P(i) * (1 - P(i)) \tag{3.3}$$

For regression, RF and decision tree calculate varience reduction using Mean Square Error (MSE), Eq. 3.4 shows the regression impurity. y_i is the node value, t_i is the target value corresponding to the sample i.

$$I(N) = \sum_{i=1}^{N} (y_j - t_i)^2 \tag{3.4}$$

Decision trees are built from the root to the leaves and used for both classification and regression: decision tree classifier and decision tree regression. Leaves in the decision tree classifier are where the classes are assigned by the majority vote, and leaves in decision tree regression are where the average (mean) predictions are set as the final value for the target. Breiman et al. [70] invented Random Forest (RF) as an ensemble version of the decision trees. Increasing the number of trees improved the accuracy significantly. The bagging or bootstrap aggregation used by RF is shown in Fig. 3.9.

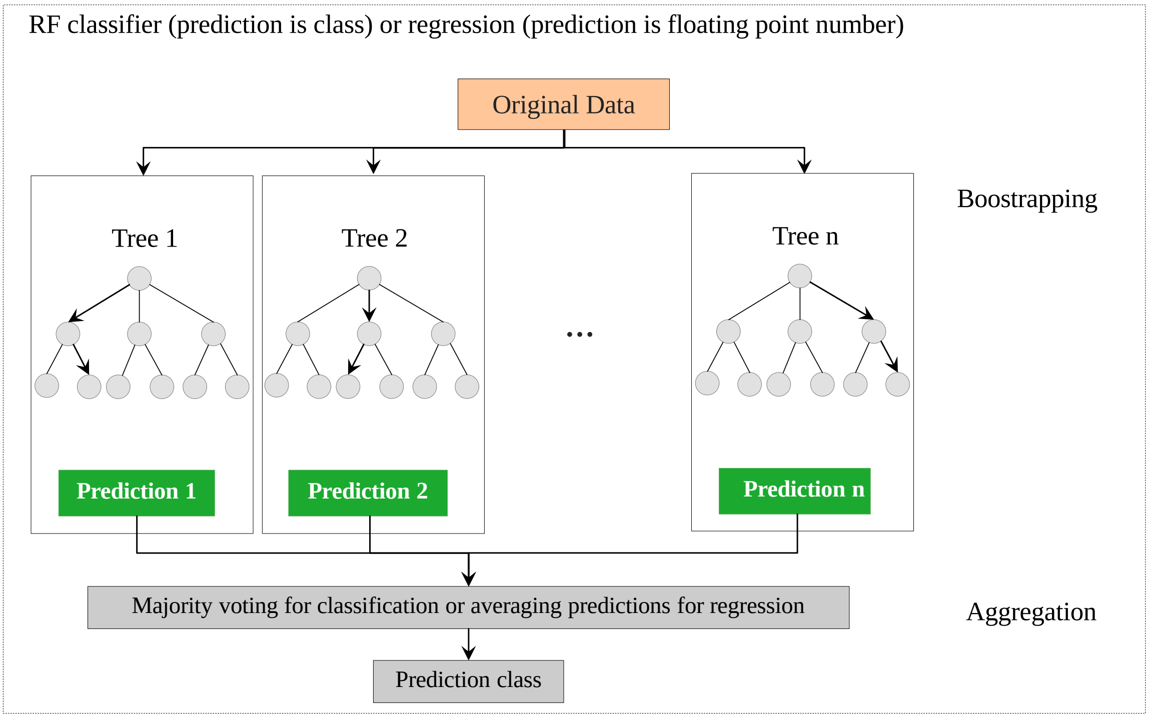

RF classifier (prediction is class) or regression (prediction is floating point number)
Original Data
Tree 1
Tree 2
...
Tree n
Boostrapping
Prediction 1
Prediction 2
Prediction n
Majority voting for classification or averaging predictions for regression
Aggregation
Prediction class

FIG. 3.9. Overview on the Random Forest (RF).

The bootstrap sample is taken from the original training dataset with a replacement for each decision tree. The decision tree grows using the bootstrap sample, where each node split is built upon randomly selected features (random sampling). Random sampling is used to strengthen the independence among the trees. The decision tree models are trained independently and evaluated using out-of-bag (OOB) data to generalize the capability of the RF. The final output is obtained by taking the majority vote of the generated results from decision tree models when using the Random Forest classifier (RFc) and mean predictions when using Random Forest regression (RFr); this step is called aggregation. RFc and RFr are defined in Eq. 3.5 and Eq. 3.6, respectively. $Ti(x)$ the output of the i tree, and j is the class.

$$f_{RFc}(x) = \arg\max \ _{j=1,2,\ldots,c} \sum_{i=1}^{N_t} (T_i(x) = j) \tag{3.5}$$

$$f_{RFr}(x) = \frac{1}{N_t} \sum_{i=1}^{N_t} T_i(x) \tag{3.6}$$

3.3.2 Long-Short Term Memory (LSTM)

Long-Short Term Memory (LSTM) is a type of Recurrent Neural Networks (RNN) [179, 180] that uses their feedback connection to store and maintain information in 'memory' over extended time intervals. Both LSTM and RNN are powerful networks designed to recognize patterns in sequences of data. RNN produces output and then copies and loops it back into the network. Thus, all the inputs are dependent on each other by taking the output of the previous input; each output obtains when the input is fed into a hidden layer with sigmoid or tanh activations. The hidden neurons are a kind of "memory" of the previous inputs because their outputs are passed through the delay block, and the output feeds it back to them as an input, see Fig. 3.10. RNN adjusts the weights for both the current and also to the previous input through gradient descent [181] (see Eq. 3.7) or backpropagation [182] (see Eq. 3.8) through time.

$$W_{x_{t+1}} = W_{x_t} - lr * \frac{d}{dW_{x_t}} \text{loss} \tag{3.7}$$

Where $W_{x_{t+1}}$ is the new weight of input x, W_{x_t} is the current weight,loss is the cost function (for more details see section 3.5), and lr is learning rate.

$$* W_{x_t} = W_{x_t} - lr * \frac{d}{dW_{x_t}} f\left(W_{x_t}\right) \tag{3.8}$$

Where $*W_{x_t}$ is the update weight of input x, W_{x_t} is the old weight of input x, $f(W_{x_t})$ is the output based on W_{x_t}, and lr is learning rate.

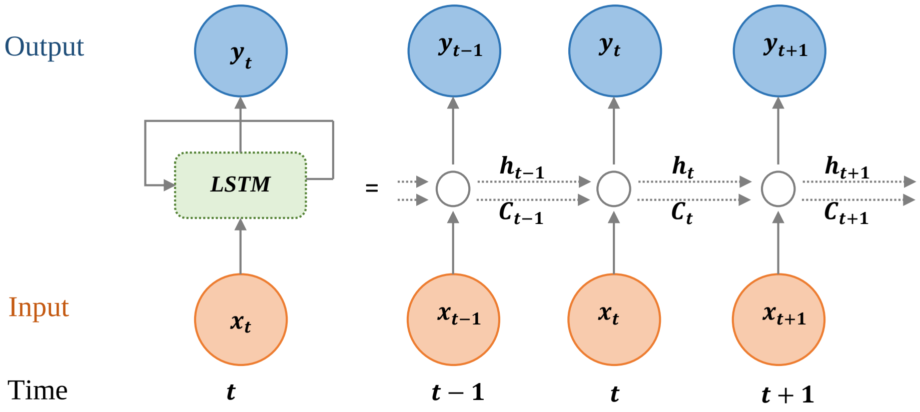
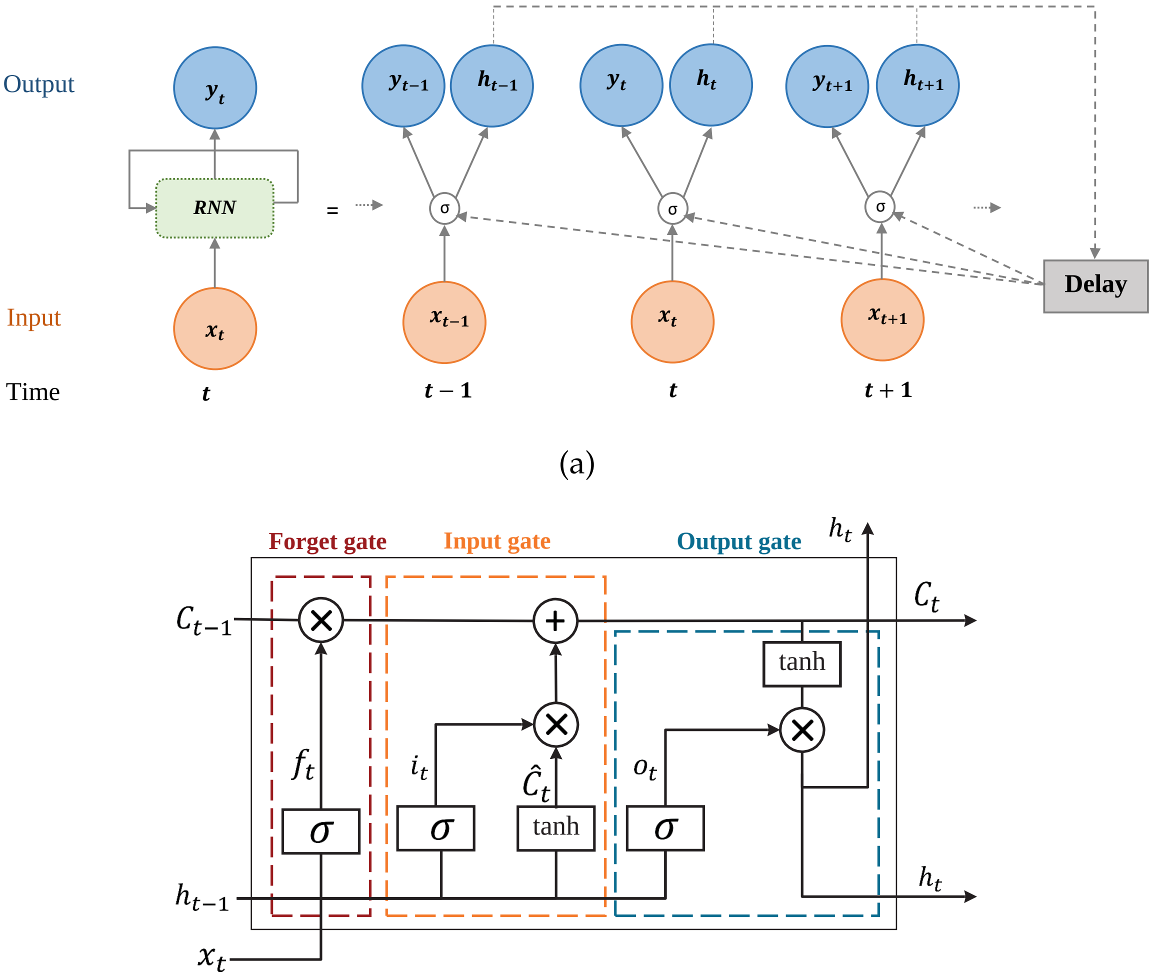


FIG. 3.10. The structure of series prediction RNN.

The problem with standard RNN: the backpropagated error signal (gradient) tends to be unstable or explode or vanish (called the vanishing gradient problem). Dealing with learning long-term temporal dependencies in RNN is a difficult task. Thus, LSTM [82] is designed based on extending the memory of RNN; this helps to store more information for an even longer time. It is an effective network for better handling time series prediction. Fig. 3.11 shows the LSTM structure.



(a)

(b)

FIG. 3.11. (a) the structure of series prediction LSTM and (b) basic structure of LSTM unit [175].

Each LSTM cell (mid row) for a given time t consists of: (1) three inputs h_{t-1}, C_{t-1}, and X_t, and (2) two outputs which are the hidden state h_t and the cell state or memory C_t. The output is usually in the range [0,1] where '0' means 'reject all' and '1' means 'include all'.

LSTM units include three kinds of gates and one cell state, which are activated using different activation functions, section 3.4 shows the description of these functions. The input gate decides which new information in the cell state to be stored by passing the input and the previous cell state (C_t) through sigmoid activation, this is shown in Eq. 3.9. The forget gate chooses when the existing information needs to be thrown away in the cell state (C_t); if the output value is closer to 0 means forget, and the closer to 1 means to keep. The intermediate cell state ($\hat{C}_t$) is calculated by passing the input and the previous state through tanh activation, which is the cell state in Eq. 3.10. Next, the element wise multiplication is performed, the forget gate is calculated and multiply it with old state C_{t-1}, this is shown in Eq. 3.11 and Eq. 3.12. The output gate determined the final output by passing the input and the previous cell state (C_t) through sigmoid activation and multiplying it with cell state passed through the tanh activation, these are shown in Eq. 3.13 and Eq. 3.14.

- Input gate:

$$i_t = \sigma\left(w_i\left[h_{t-1}, x_t\right] + b_i\right) \tag{3.9}$$

- Intermediate cell state:

$$\hat{C}_t = \tanh\left(w_c - \left[h_{t-1}, x_t\right]\right) + b_c \tag{3.10}$$

- Cell state:

$$C_t = \left(f_t * C_{t-1}\right) + \left(i_t * \hat{C}_{\mathrm{t}}\right) \tag{3.11}$$

- Forget gate:

$$f_t = \sigma\left(w_f\left[h_{t-1}, x_t\right]\right) + b_f \tag{3.12}$$

- Output gate:

$$o_t = \sigma\left(w_t\left[h_{t-1}, x_t\right]\right) + b_t \tag{3.13}$$

- New state:

$$h_t = o_t * \tanh\left(C_t\right) \tag{3.14}$$

Where t is the timestep, i_t is input gate at t, f_t is forget gate at t, out_t is output gate at t, $x_{(t)}$ is the current input, $h_{t-1}$is previous hidden state, w_i is weight matrix of sigmoid operator between input gate and output gate, w_c is weight matrix of tanh operator between input gate and output gate, w_f is weight matrix of sigmoid operator between forget gate and input gate, w_o is weight matrix of output gate. b_i bais vector at t, b_c bais vector at t and w_c, b_f connection bias at t, b_t connection bias at w_c. C_t is cell state information, ($\hat{C}_t$) is value generated by tanh, C_{t-1} is previous timestep, and h_t is LSTM output.

3.4 Activation Functions

Activation functions, also known as threshold functions or transfer functions, help in learning and facilitating non-linear and complicated mappings between the inputs and desired outputs. In neural networks: (1) the sum of inputs and their corresponding weights is calculated, (2) the activation function activates the neuron to transform the output of that particular layer via nodes in the next layer of the network. The choice of activation function in the hidden layer influences how perfectly the network model learns the training dataset. In output layers, the activation function is determined based on the type of prediction. For more details, see [183].

Some common activation functions for deep learning throughout this work are described below:

x indicates input feature, x_i indicates element in input vector (one-hot encoded matrix), k indicates the total number of classes in multiclasses.

- Linear Function:

$$g(x) = x \tag{3.15}$$

  This activation function is also known as Identity Function where the activation is the input.

- Sigmoid:

$$g(x) = \frac{1}{1 + e^{-x}} \tag{3.16}$$

  This activation function transforms the values in the range 0 to 1; these values can be treated as probabilities for binary and multi-label classification tasks.

- Tanh:

$$g(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3.17}$$

  This activation function transforms the values in the range -1 to 1. It has gradients that are not restricted to vary in a certain direction.

- ReLU:

$$g(x) = \max(0, x) \tag{3.18}$$

  This activation function is most widely used in Conventional Neural Networks (CNNs); all the negative values are converted into zero. ReLU ranges from 0 to infinity, and the deactivated neuron is produced when the output of linear transformation is zero.

- Softmax:

$$g\left(x_i\right)=\frac{e^{x_i}}{\sum_{j=0}^{k} e^{x_j}} \text{ for } i=0,..,k-1 \tag{3.19}$$

This activation function is a combination of multiple sigmoid functions. It gives the probability of each class, and the sum of these probabilities is eventually one. The target class will have a high probability. They can be used for multiclass classification tasks.

These activation functions except ReLU are continuous and differentiable so that gradient descent and backpropagation can be used to optimize the loss function. ReLU is differentiable at all the points except 0.

3.5 Loss Functions

Loss functions help to determine the model performance by calculating the distance between the predicted output with the expected output of the machine learning networks. The model weights are updated using those functions during training until getting the best result. The model's performance is maximized by minimizing the loss.

The loss functions used in deep learning throughout this work are described below:

- Binary Cross Entropy:

$$BCE=-\frac{1}{\mathrm{N}} \sum_{i=0}^{N}\left[y_i \log \left(\hat{y}_i\right)+\left(1-y_i\right) \log \left(1-\hat{y}_i\right)\right] \tag{3.20}$$

This can be used for binary single- or multiple-label classification with sigmoid activation in the last layer. It compares each of the predicted probabilities to the expected class output, which can be between 0 and 1. It is used for predicting only two classes (classification tasks) or continuous values between 0 and 1 (regression tasks).

- Categorical Cross Entropy:

$$CCE=-\sum_{i=0}^{N} y_i \log \left(\hat{y}_i\right) \tag{3.21}$$

This can be used for multiclass classification with softmax activation in the last layer. It is designed to quantify the difference between two probability distributions. It uses where a sample can only belong to one out of many possible categories, and the model must decide which one. It is used for predicting multiple labels in multi-class classification tasks.

- Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{3.22}$$

This can be used for regression with linear activation in the last layer. This loss function is responsible for computing the average squared difference between the predicted probabilities and expected class output. It is used for predicting continuous output in regression tasks.

CHAPTER 4

Data Preparation

THIS chapter describes the X-ITE Pain Database. It also presents the database pre-processing and temporal integration processes in the proposed automatic system for continuous monitoring of pain intensity using the X-ITE Pain Database; Fig. 4.1 shows the data preparation steps in the pipeline of the suggested system. The first step was selecting pain response data (from participants) during applied pain stimulation from the five modalities: frontal RGB camera, audio, electrocardiogram [ECG], facial electromyography [EMG], electrodermal activity [EDA]. This data have been collected by Gruss et al. [149]. Section 4.1 describes the data collection and selection process using the X-ITE Pain Database. The second step was (1) processing the frontal facial RGB video using OpenFace [154] for detecting the face from each frame for each participant (subject) and for extracting Facial Features (FF) & head pose, (2) processing the audio signal using openSMILE [171], (3) applying the QRS-detection algorithm by Hamilton et al. [176] with the ECG signals, and (4) processing the three EMG channels with a zero-phase 3rd-order Butterworth band-pass filter. More details about sensor modalities and signal processing were described in Section 4.2. The third step was representing the time window which includes temporal integration of frame-level features by a time series statistics descriptor on the processed data from each modality individually. Five descriptors were provided: Facial Activity Descriptor [FAD], Audio Descriptor [Audio-D], ECG Descriptor [ECG-D], EMG Descriptor [EMG-D], and EDA Descriptor [EDA-D]. The labels three seconds were moved forward and then used a sliding window with a time length of ten seconds ago. See Section 4.3for more details about the temporal integration process. Finally, the data were further processed in order to ensure and improve the performance of automatic methods for monitoring continuous pain intensity, see experimental data in section 4.4.

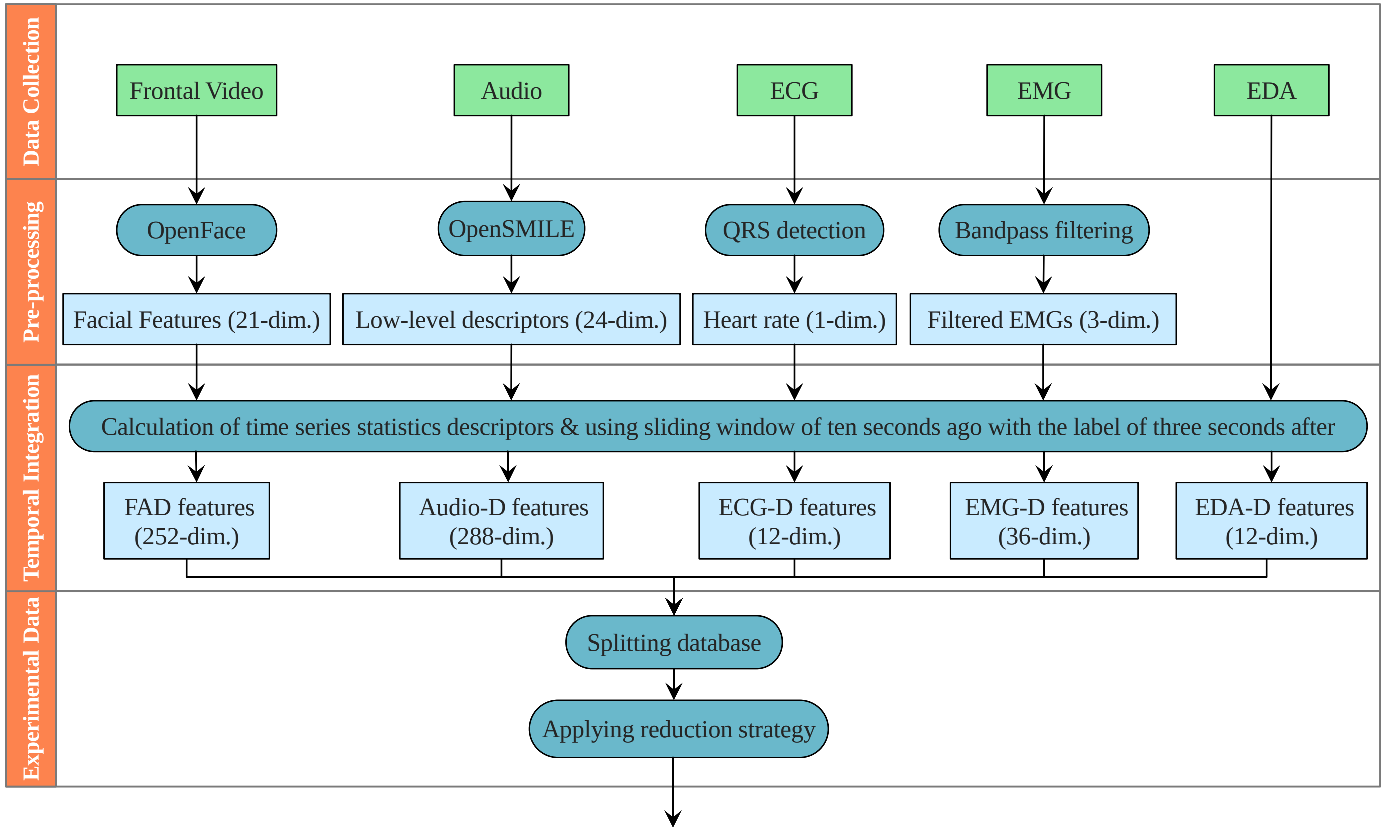


FIG. 4.1. The data preparation pipeline in continuous pain intensity monitoring system. The input consists of response data collected from five modalities (frontal video, audio, ECG, EMG, and EDA) when participants were exposed to pain stimuli. The data is processed using different methods and filters, and extracted features were used to determine temporal integration by using a time series statistics descriptor.

4.1 X-ITE Pain Database

In this thesis, a multimodal Experimentally Induced Thermal and Electrical (X-ITE) Pain Database [149] was used to validate the performance of different automatic methods for continuous intensity pain monitoring. This database was selected because it is made to complement existing databases, including behavioral and physiological data that was recorded when healthy participants (subjects) were exposed to different qualities (heat/electric) and duration (5 s /1 min) of pain stimuli. This diversity in the database provides additional valuable information to advance the discrimination between pain or intensity of pain versus no pain.

In this database, a total of 134 human healthy participants aged between 18 and 50 years were subjected to two types of pain modalities (heat and electricity) in three intensities (low, medium, and high) and two different stimuli durations (phasic and tonic). The heat pain stimulus was stimulated at participants' forearm using a thermal stimulator (Medoc PATHWAY Model ATS). The electrical pain stimulus was stimulated at participants when electrodes were attached to participants' index and middle fingers using an electrical stimulator (Digitimer DS7A). Fig. 4.2 shows the pain response data derived from multiple sensor modalities.

The intensities of both pain stimuli (heat and electricity) were selected individually based on participants' personal pain sensitivity (tolerances). For this purpose, there was a person-calibration procedure before the main stimulation phase, in which the participant self-reported the pain experienced during several stimuli using the numeric rating scale.

Six phasic (short) and six tonic (long) stimulus types were applied to each participant based on their pain thresholds and tolerances. For each phasic stimulus, the three pain intensities (times two pain modalities) were repeated 30 times for five seconds duration, applied in randomized order with pauses of 8-12 seconds. The one-minute tonic stimuli were applied once per intensity, followed by a pause of five minutes. There were three phases of how tonic heat and electrical pain intensity stimuli were applied: the two lower intensities were applied randomly during the phasic stimulus period, and the highest intensity was applied at the end of the experiment. The entire experiment (preparation and actual experiment) took about 3 hours per participant. For more details see Gruss et al. [149].

The facial expression, head pose, body gestures, and facial skin temperature were analyzed from video; para-linguistic responses (vocalizations) were analyzed from the recorded audio signal; heart rate and its variability were analyzed from the measured electrocardiogram (ECG); surface electromyography (EMG) has been recorded for measuring the activity of trapezius (neck/shoulder), corrugator supercilii (close to eyebrows), and zygomaticus major (at the cheeks) muscles; electrodermal activity (EDA) has been recorded for measuring sweating.

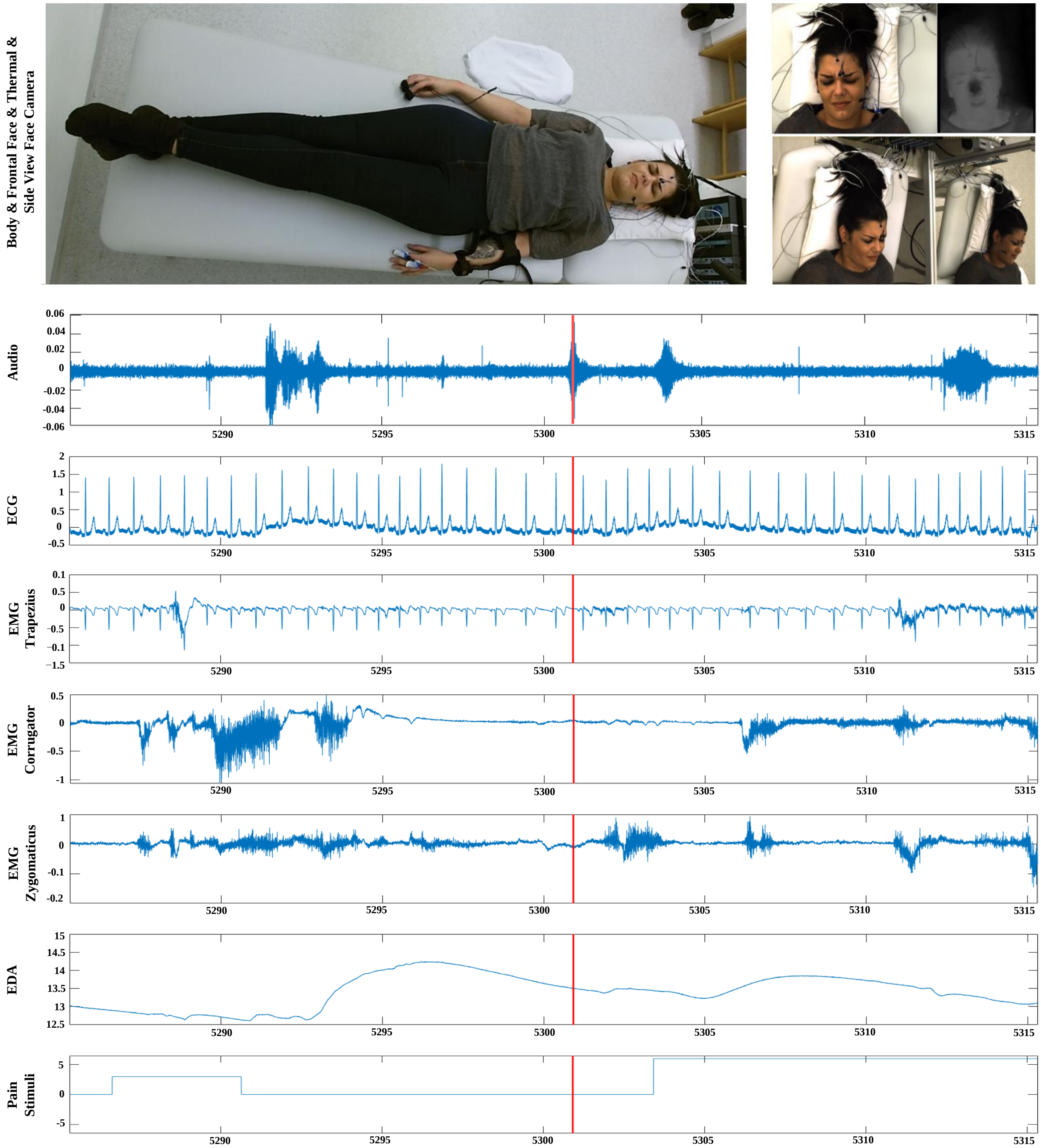


FIG. 4.2. The pain response data from the X-ITE Pain Database. The representative screenshots of the video signals (top) show the reactions when one of an intense pain stimuli was applied. The figure depicts plots of the recorded signals (middle part) before, during, and after the application of a pain stimulus (bottom plot). (EMG = Electromyography, EDA = Electrodermal Activity, ECG = Electrocardiogram).

In line with [29] and Othman et al. [75], the same 127 participants (subjects) subset was selected, including samples only, for which data were available from five sensors (frontal RGB camera, audio, ECG, EMG, EDA). Approximately one and a half hours was the duration of the actual experiment for each participant, which was used in this work. Five sensor modalities were analyzed in this work to objectively monitor the phasic and tonic pain intensity during the application of the thermal and electrical pain stimuli and no pain.

4.2 Pre-processing

This section explains the steps in pre-processing on X-ITE Pain Database when using the pain response data from five sensor modalities. Participants were subjected to painful stimuli that differ in intensity, duration, and modality; for more details about the database, see Section 4.1. The following is a description of the types of sensors used as well as the sensor-specific signal processing.

4.2.1 Frontal Video

Fig. 4.3 shows the processing pipeline for RGB frontal video. For more details about the OpenFace tool, see Section 3.1.1.

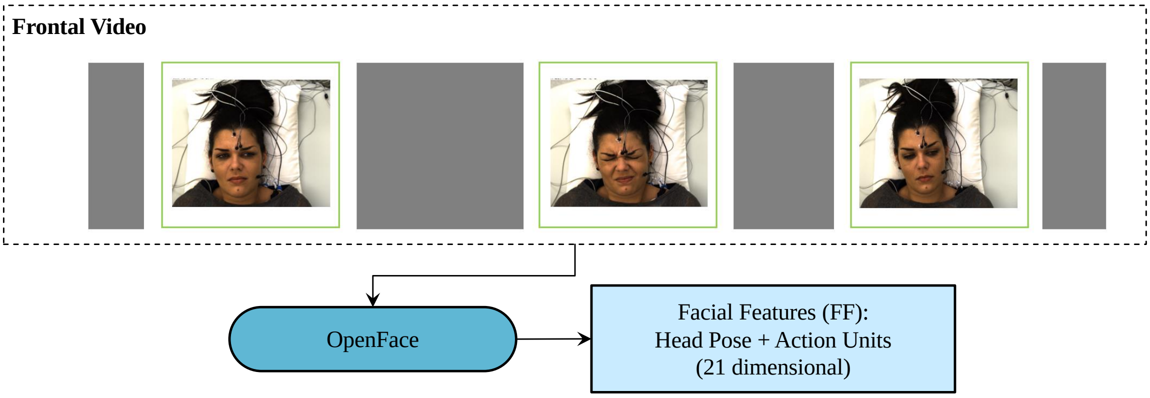


FIG. 4.3. The pipeline for processing RGB fronal video.

The RGB frontal face from videos were used for analyzing facial expressions and head pose information using OpenFace [154]. For each frame of each video, the OpenFace tool extracted Facial Features (FF) through the following steps: (1) it detects the face and facial landmarks, (2) it extracts Action Units(AUs), and (3) it estimates head pose. As frame-level expression features, the FF that are used includes 21 features: 3 head poses (Yaw, Pitch, and Roll), AU1 (binary occurrence output), and 17 AU intensity outputs of OpenFace, which are AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, and AU45. The 21-dimensional facial expression time series were recorded at a frame rate of 25 frames per second (fps).

4.2.2 Audio

The audio signal was used for analyzing para-linguistic responses using openSMILE [171]. For each frame, a 24-dimensional low-level descriptor [LLD] was extracted, comprising four energy features (Sum of auditory spectrum [loudness], Sum of RASTA-filtered auditory spectrum, Root-Mean Square [RMS] Energy, and Zero-Crossing Rate), 6 voicing features (F0 (Subharmonic-Summation [SHS] & Viterbi smoothing), probability of voicing, logarithmic Harmonics to Noise Ratio [HNR], Jitter (local delivered duty paid [DDP]), Shimmer [local]), and 14 spectral features [Mel Frequency Cepstral Coefficients (MFCCs)]. Further, the 24-dimensional LLD audio time series were extracted at the same time series sampling rate as the FF time series frame rate (1/25 seconds). Fig. 4.4 shows the processing pipeline for Audio. For more details about the openSMILE tool, see Section 3.1.2.

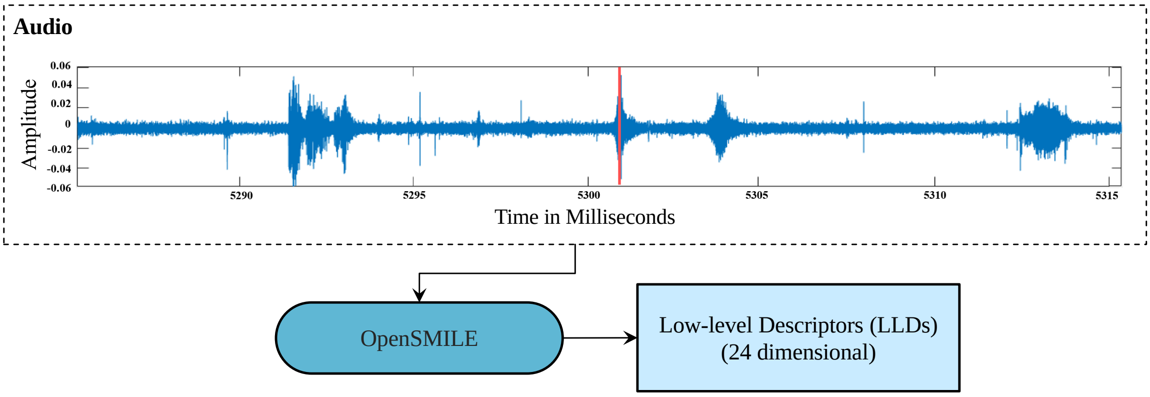


FIG. 4.4. The pipeline for processing audio signal.

4.2.3 ECG

The electrocardiogram (ECG) was used to analyze heart rate and its variability. The QRS-detection algorithm by Hamilton et al. [176] was applied in order to find the R-peaks in the ECG signal. Then, R-to-R intervals were used to determine heart rate. Afterward, we interpolated the heart rate signal linearly to match the sampling of the EMG and EDA (1000 Hz). Finally, only the ECG data (1-dimensional) was used at the same time series sampling rate as the FF time series frame rate (1/25 seconds).

Fig. 4.5shows the processing pipeline for the ECG signal. For more details about QRS-detection algorithm and methods, see Section 3.1.3.

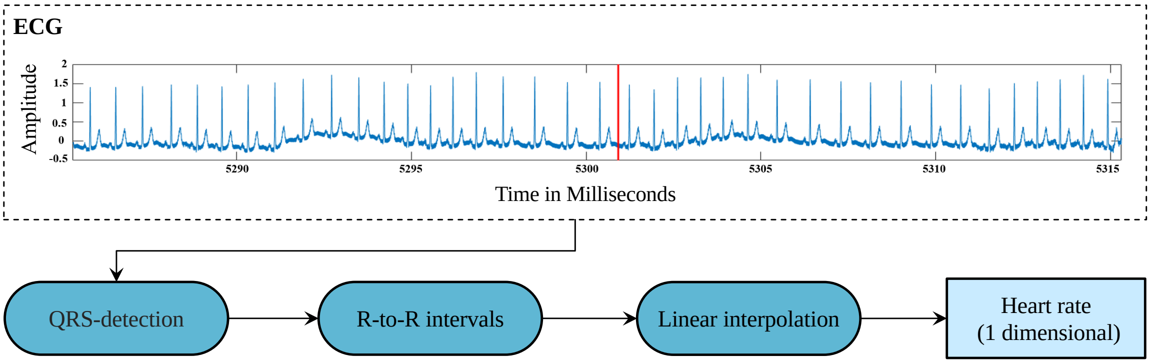


FIG. 4.5. The pipeline for processing ECG signal.

4.2.4 EMG

Fig. 4.6 shows the processing pipeline for the EMG channel signals.

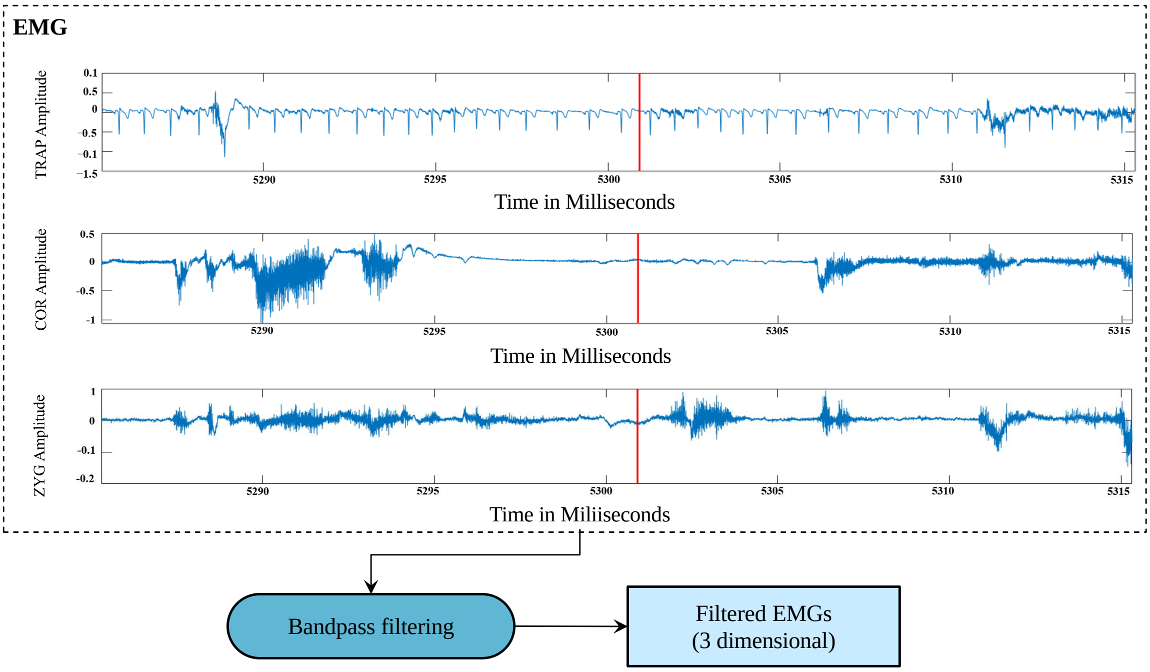


FIG. 4.6. The pipeline for processing EMG channel signal. COR: corrugator supercilii, ZYG: zygomaticus major, TRAP: trapezius.

The surface electromyography (EMG) was used to measure the activity of three muscles, corrugator supercilii (close to eyebrows), zygomaticus major (at the cheeks), and the trapezius (neck/shoulder). A zero-phase 3rd-order Butterworth band-pass filter with cut-off frequencies of 20 and 250 Hz was used to process the three channel EMG signals. Further, 3-dimensional EMG time series was used at the same time series sampling rate as the FF time series frame rate (1/25 seconds). For more details about this filter, see Section 3.1.4.

4.2.5 EDA

Electrodermal activity (EDA) was used to measure sweating. The 1-dimensional EDA time series was used a frame rate of 25 frames per second (fps) same as FF frames. Fig. 4.7 shows the EDA signal without filtering.

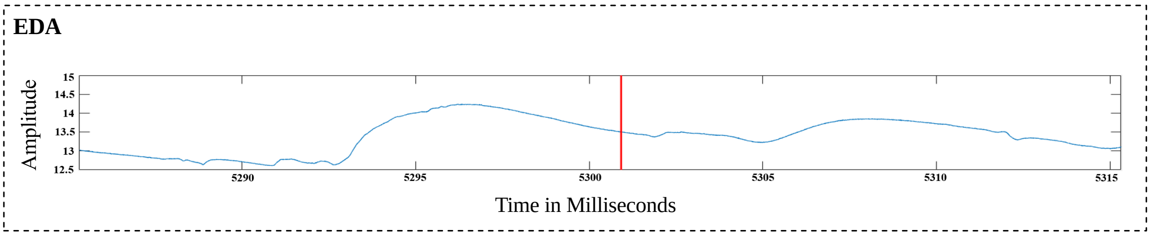


FIG. 4.7. The pipeline for processing EDA signal.

4.3 Temporal Integration

Temporal integration of frame-level features (also called time series) were calculated from the 21-dimensional facial expression time series, 24-dimensional audio time series, 1-dimensional ECG time series, 3-dimensional EMG time series, and 1-dimensional EDA time series, see Fig. 4.8. For more details about signal processing, see Section 4.2. The temporal integration for each sensor modality was represented by a time series statistics descriptor [7,60] to describe the changes of features, which are called Facial Activity Descriptor [FAD], Audio Descriptor [Audio-D], ECG Descriptor [ECG-D], EMG Descriptor [EMG-D], and EDA Descriptor [EDA]. Each second in each descriptor was summarized by four statistics of the time series itself and its first and second derivative, including minimum, maximum, mean, and standard deviation, yielding a 12×21-dimensional, 12×24-dimensional, 12×1-dimensional, 12×3-dimensional, and 12×1-dimensional descriptor for facial, audio, ECG, EMG, and EDA features, respectively. A person-specific standardization of the features [60] was applied with all descriptors in order to focus on the within-subject response variation rather than the differences between subjects. For each subject, the mean and standard deviation were calculated, then each feature value was subtracted by the mean and divided by the standard deviation that belonged to the same subject.

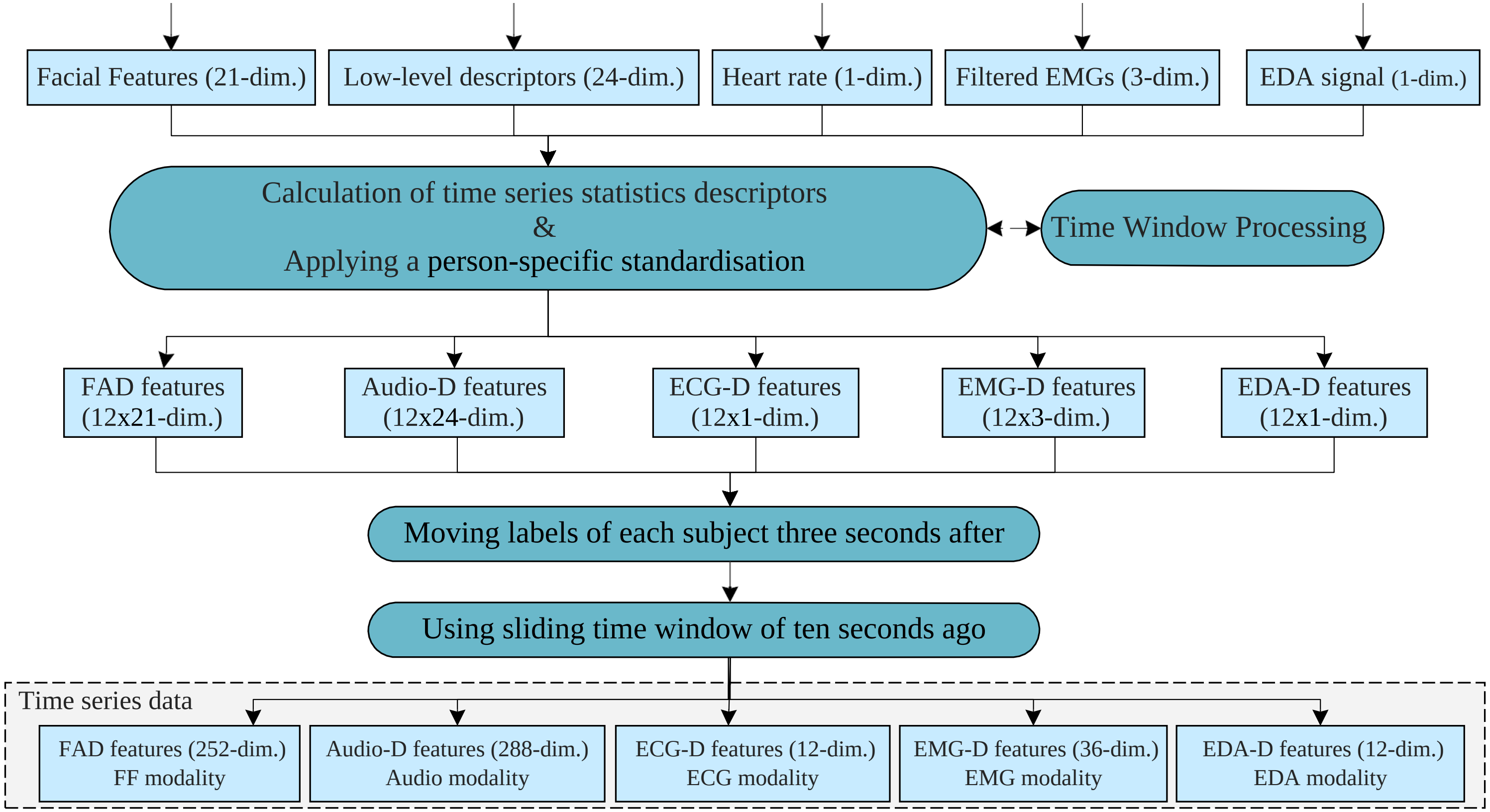

Facial Features (21-dim.)
Low-level descriptors (24-dim.)
Heart rate (1-dim.)
Filtered EMGs (3-dim.)
EDA signal (1-dim.)
Calculation of time series statistics descriptors
&
Applying a person-specific standardisation
Time Window Processing
FAD features (12x21-dim.)
Audio-D features (12x24-dim.)
ECG-D features (12x1-dim.)
EMG-D features (12x3-dim.)
EDA-D features (12x1-dim.)
Moving labels of each subject three seconds after
Using sliding time window of ten seconds ago
Time series data
FAD features (252-dim.) FF modality
Audio-D features (288-dim.) Audio modality
ECG-D features (12-dim.) ECG modality
EMG-D features (36-dim.) EMG modality
EDA-D features (12-dim.) EDA modality

FIG. 4.8. Temporal Integration pipeline for processed data from all sensor modalities. Time Window Processing were explained in Fig. 3.8.

The labels of each subject were moved three seconds after because the facial pain responses typically are delayed by 2-3 seconds compared to stimulus. Further, a sliding window was applied by combining the FAD, Audio-D, ECG-D, EMG-D, or EDA-D ten seconds ago once per second to predict the next time step of the pain intensity label. The data for the first ten seconds are removed because there are no prior observations to use.

4.4 Experimental Data

The average length of each sequence in X-ITE Pain Database was about one and a half hours. The imbalanced data distribution of the pain intensity was shown in Fig. 4.9.

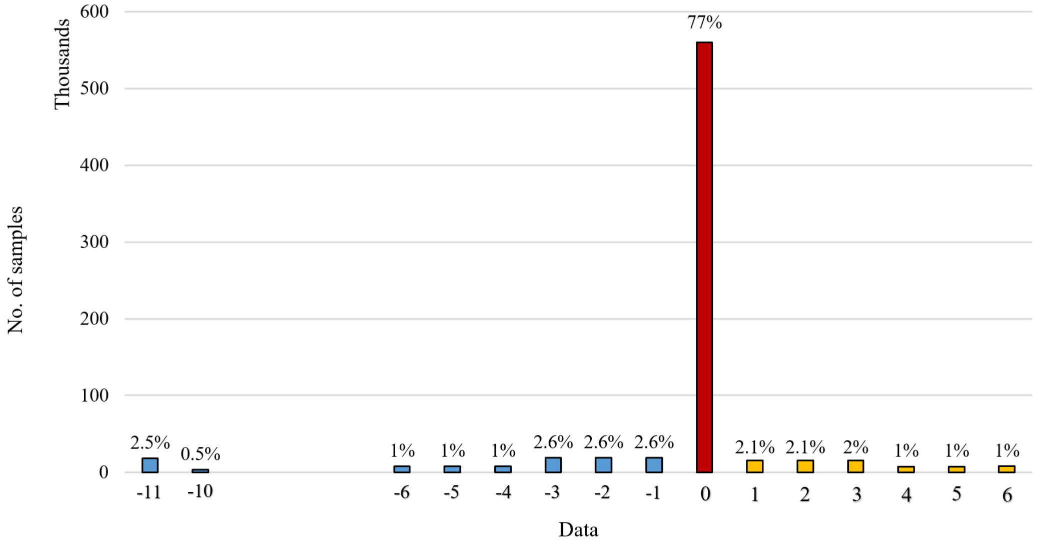


FIG. 4.9. Sample distribution based on labels.

The notations in the conducted experiments were summarized in Table 4.1. In the X-ITE Pain Database, 0 indicates the samples in which subjects experience no pain. Phasic and tonic pain levels were represented by positive and negative labels 1 to 3 and 4 to 6, respectively. Samples with labels 3 & -3 (phasic pain stimulus) and 6 & -6 (tonic pain stimulus) indicate severe heat and electrical stimuli, respectively; samples with labels 2 & -2 (phasic pain stimuli) and 5 & -5 (tonic pain stimulus) indicate moderate heat and electrical stimulus, respectively; samples with labels 1 & -1 (phasic pain stimuli) and 4 & -4 (tonic pain stimuli) indicate low heat and electrical stimulus, respectively. -10 indicates samples with problems such as false start and restart of the stimuli, overlapping between heat or electrical stimulation, unbalanced phasic estimation, short pause, short tonic electrical stimulus, single heat stimulus in front, or additional stimulus. -11 indicates the samples when the subject speaks or interacts during the experiment (the beginning and after the first & second tonic stimuli of the experiment.

Table 4.1. List of abbreviations of pain stimuli type, modalities, intensities, and numerical class labels with the percentage samples distribution.

Type	**Modality**	**Intensities**			
		severe	**moderate**	**low**	**no pain (77%)**
Phasic	H	PH3 = 3 (2%)	PH2 = 2 (2.1%)	PH1 = 1 (2.1%)	BL= 0
	E	PE3 = -3 (2.6%)	PE2 = -2 (2.6%)	PE1 = -1 (2.6%)	
Tonic	H	TH3 = 6 (1%)	TH2 = 5 (1%)	TH1 = 4 (1%)	BL= 0
	E	TE3 = -6 (1%)	TE2 = -5 (1%)	TE1 = -4 (1%)	
E: Electrical pain stimulus, H: Heat pain stimulus					
-10 & -11 Labels not used in the experiments: -10 (0.5%) and -11(2.5%)					

Several pre-processing steps were proposed on the X-ITE Pain Database to reduce the impact of the extremely imbalanced database problem:

First, the intensities of facial expressions for most samples when expressing pain intensity were investigated, then all subjects into four categories were assigned based on how they expressed pain intensity.

Second, Splitting the database into three splits was suggested: training set, validation, and testing set. A randomly subjects for each split from each category were selected based on 80% of data for training (100 subjects = 572696 samples), 10% for validation (13 subjects = 75537 samples), and 10% for testing (14 subjects = 79485 samples). Each split contained subjects from all intensity categories, see Fig. 4.10.

Third, the obtained splits from the database were processed: (a) all sequences of samples with labels -10, -11, and no pain samples sequence before and after these samples were excluded to simplify the problem and reduce the impact of imbalance in the database; (b) the obtained dataset was split into 6 subsets to evaluate the proposed methods (see the Subsets which are the first six datasets in Table 4.2); (c) each obtained dataset was reduced by removing some no pain samples prior to pain intensity samples in a time series for each subject to evaluate the proposed methods, these datasets are called Reduced Subsets; see the Reduced Subsets which were the last six datasets in Table 4.2.

Intensity: 1

Training Subjects:
S002, S010, S019, S020, S022, S033, S042, S065, S082, S093, S048, S050, S051

Validation Subject:
S113

Testing Subjects:
S079, S107

Intensity: 2

Training Subjects:
S011, S016, S026, S027, S028, S037, S041, S044, S074, S077, S080, S089, S091, S092, S095, S102, S103, S118, S120, S122, S123, S124, S130, S132

Validation Subjects:
S046, S100, S128

Testing Subjects:
S007, S047, S056, S081

Intensity: 3

Training Subjects:
S003, S004, S005, S006, S008, S012, S015, S017, S029, S032, S034, S035, S036, S038, S039, S040, S045, S052, S053, S054, S055, S057, S058, S061, S063, S064, S066, S067, S068, S069, S070, S072, S073, S075, S076, S078, S083, S084, S085, S086, S087, S090, S096, S098, S099, S101, S105, S106, S109, S110, S111, S112, S114, S116, S117, S119, S125, S129, S133

Validation Subjects:
S009, S013, S018, S043, S088, S094, S097, S126

Testing Subjects:
S031, S049, S058, S062, S071, S127, S131

Intensity: 4

Training Subjects: S060, S108, S115, S134

Validation Subject: S104

Testing Subject: S021

FIG. 4.10. Assignment of subjects to categories of facial response intensities [184]. Intensity 1= lack of facial responses to pain, Intensity 2, 3 = moderate intensity of facial responses to pain, and intensity 4 = intensive facial responses to pain.

Table 4.2. No. of samples in each dataset for each splits to evaluate proposed methods before & after applying sample weighting method.

Datasets		Description	Training set	Validation set	Test set	Applying sample weighting	
						Training set	Increased
Subsets	PD	Phasic Dataset	352,133	46,476	50,362	405,060	52,927 (20%)
	HPD	Heat Phasic Dataset	159,998	21,441	23,019	190,178	30,180 (20%)
	EPD	Electrical Phasic Dataset	316,939	41,794	45,325	353,560	36,621 (10%)
	TD	Tonic Dataset	117,646	14,885	16,689	142,667	25,021 (20%)
	HTD	Heat Tonic Dataset	21,198	2,755	3,103	37,087	15,889 (70%)
	ETD	Electrical Tonic Dataset	95,458	12,000	13,446	109,644	14,186 (10%)
Reduced Subsets	RPD	Reduced Phasic Dataset	158,472	20,897	22,501	237,735	79,263 (50%)
	RHPD	Reduced Heat Phasic Dataset	69,390	9,233	9,933	119,780	50,390 (70%)
	REPD	Reduced Electrical phasic Dataset	88,148	11,548	12,438	150,937	62,789 (70%)
	RTD	Reduced Tonic Dataset	55,804	7,041	7,983	99,455	43,651 (80%)
	RETD	Reduced Electrical Tonic Dataset	33,826	4,156	4,740	62,799	28,973 (90%)

Table 4.3. The percentage samples distribution of pain stimuli type, modalities, intensities for each dataset after database preprocessing. P = phasic and T = tonic indicate the two types of pain stimuli, H = heat and E = electrical indicate the modalities, 1 = low, 2 = moderate, and 3 = severe indicate the three intensity.

Datasets		Phasic Pain Intensities							No Pain = 0	Datasets	Tonic Pain Intensities							No Pain = 0
		Severe		Moderate		low		mean			Severe		Moderate		low		mean	
		PH3 = 3	PE3 = -3	PH2 = 2	PE2= -2	PH1 = 1	PE1 = -1				TH3 = 6	TE3 = -6	TH2 = 5	TE2 = -5	TH1 = 1	TE2 = -1		
Subsets	PD	3.3	4.2	3.3	4.2	3.2	4.2	3.7	77.7	TD	5.0	4.4	5.0	5.0	5.0	5.0	4.9	70.4
	HPD	7.2	-	7.3	-	7	-	7.17	78.5	HTD	25.7	-	27.1	-	27.1	-	26.7	20.1
	EPD	-	4.6	-	4.6	-	4.6	4.6	86.2	ETD	-	5.5	-	6.2	-	6.2	5.9	82
Reduced Subsets	RPD	7.4	9.3	7.2	9.3	7.2	9.3	8.3	50	RTD	10.5	9.3	10.5	10.5	10.5	10.5	10.3	38.1
	RHPD	16.7	-	16.8	-	16.3	-	16.6	50.1	-	-	-	-	-	-	-	-	-
	REPD	-	16.6	-	16.8	-	16.6	16.7	50	RETD	-	15.6	-	17.7	-	17.7	17	49

Table 4.3 shows the distribution of samples in each proposed dataset. The applied Subsets were (1) Phasic Dataset [PD]: Excluded tonic samples (labeled 4, 5, 6, -4, -5, -6, -10, -11) and no-pain samples before these samples and also after samples with -10, -11 labeled, (2) Heat Phasic Dataset [HPD]: Excluded electrical samples (labeled -1, -2, -3) from PD and no-pain samples before these frames, (3) Electrical Phasic Dataset [EPD]: Excluded heat samples (labeled 1, 2, 3) from PD and no pain frames before these frames, (4) Tonic Dataset [TD]: Excluded phasic samples (labeled 1, 2, 3, -1, -2, -3, -10, -11) and no pain samples before these samples and also after samples with -10, -11 labeled, (5) Heat Tonic Dataset [HTD]: Excluded electrical samples (labeled -1, -2, -3) from TD and no pain frames before these frames, and (6) Electrical Tonic Dataset [ETD]: Excluded heat samples (labeled 1, 2, 3) from TD and no pain frames before these frames. The Reduced Subsets were (7) Reduced Phasic Dataset [RPD]: Reduced the no pain frames in PD to about 50%, (8) Reduced Heat Phasic Dataset [RHPD]: Reduced the no pain frames in HPD to about 50%, (9) Reduced Electrical Phasic Dataset [REPD]: Reduced the no pain

frames in EPD to about 50%, (10) Reduced Tonic Dataset [RTD]: Reduce the no pain frames in TD to about 38%, (11) Reduced Electrical Tonic Dataset [RETD]: Reduced the no pain frames in ETD to about 49%.

7-Class pain recognition was considered for two types of pain stimuli (P = phasic and T = tonic) variants of each modality (Heat = H and Electrical = E) in three intensity (1 = low, 2 = moderate, and 3 = severe): (1) BL, PH1, PH2, PH3, PE1, PE2, and PE3 for the phasic recognition task, and (2) BL, TH1, TH2, TH3, TE1, TE2, and TE3 for the tonic recognition task. Further, 4-Class pain recognition was considered for one type of pain stimulus (P / T) variants of one modality (H / E) in three intensities: (1) BL, PH1, PH2, and PH3 for the phasic heat recognition task, (2) BL, PE1, PE2, and PE3 for the phasic electrical recognition task, (3) BL, TH1, TH2, and TH3 for tonic heat recognition task, (4) BL, TE1, TE2, and TE3 for the tonic electrical recognition task.

Fig. 4.11 shows the suggested reduction strategy, which focuses on reducing some no pain samples prior to each pain intensity sequence by preserving different numbers of no pain samples that are directly adjacent to each pain intensity sequence. This number was assigned based on the number of samples in each pain intensity sequence, e.g., for a sequence of phasic electrical moderate pain intensity that contains five samples; the previous five no pain samples were kept, and the rest before were deleted. Thus, five additional datasets (Reduced Subsets) were obtained, and the Heat Tonic Dataset (HTD) was not reduced because it is nearly balanced.

The proposed automatic models, which were introduced in Chapter 5, were trained on all 11 datasets. The models that were trained on the database before the splitting performed poorly due to the huge imbalanced class distribution. The pain intensity labels were conditioned into the right format by: (1) converting the negative labels (-1, -2, -3) to positive (4, 5, 6) in Phasic Dataset [PD], the obtained labels are 1, 2, 3, 4, 5, 6, (2) converting the labels 4 , 5 , 6, -4, -5, -6 to 1, 2, 3, 4, 5, 6 in Tonic Dataset [TD], (3) converting the negative labels (-1, -2, -3) to positive (1, 2, 3) in Electrical Phasic Dataset [EPD], (4) converting labels 4, 5, 6 to 1, 2, 3 in Heat Tonic Dataset [HTD], and (5) converting the labels -4, -5, -6 to 1, 2, 3 in Electrical Tonic Dataset [ETD]. With regression models, no pain and pain intensity labels were normalized to bring them in the range of [0,1].

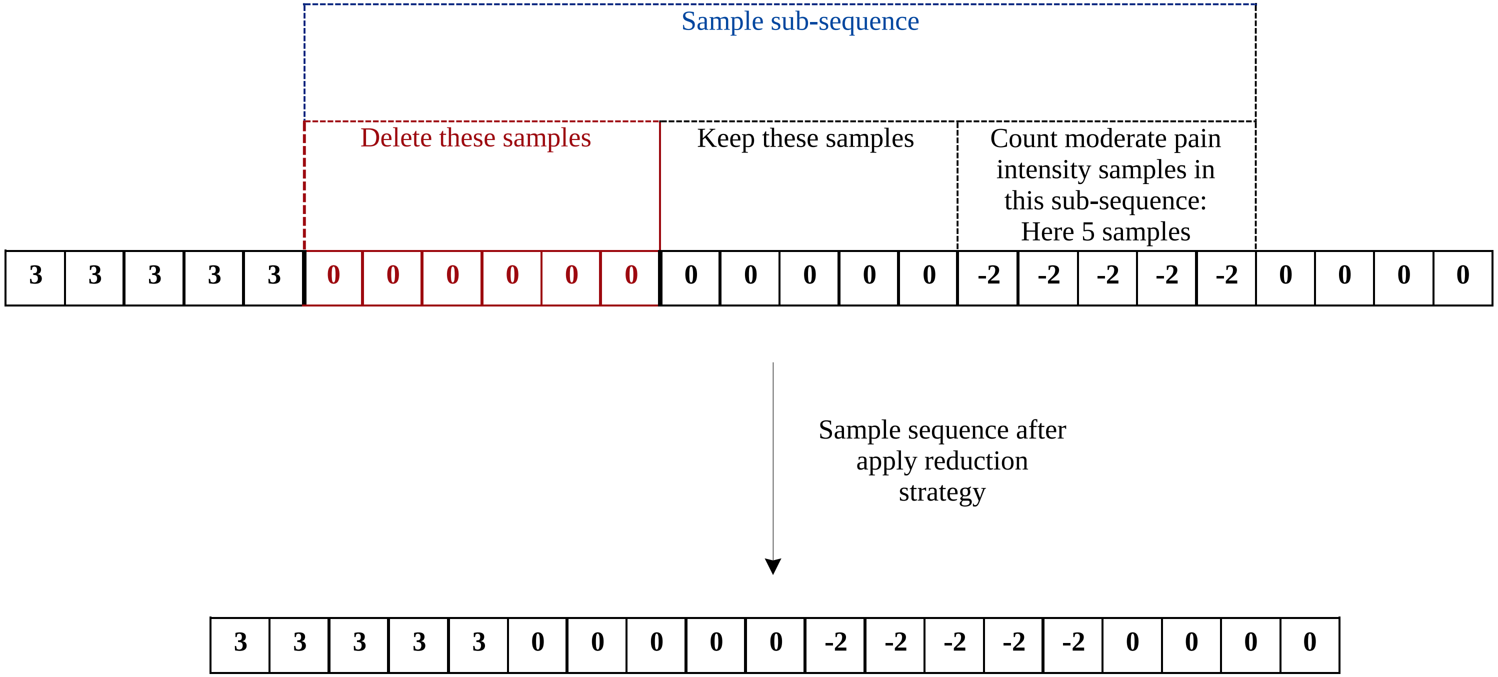


FIG. 4.11. Overview of reduction strategy on sample sequences.

CHAPTER 5

Continuous Pain Intensity Monitoring

THIS chapter introduces three automatic methods suggested to predict continuous pain intensity using the time series data, which are (1) Random Forest (RF) as baseline methods [Random Forest classifier (RFc) and Random Forest regression (RFr)], (2) Long-Short Term Memory (LSTM), and (3) LSTM using sample weighting method (called LSTM-SW); see Section 5.1, Section 5.2, and Section 5.3, respectively. After data preparation (see Chapter 4), the obtained time series data (experimental data), including FAD, Audio-D, ECG-D, EMG-D, and EDA-D, was used to train the proposed methods, see Fig. 5.1. The reason for using different methods with the experimental data was to explore the generalizability of continuous pain intensity monitoring models. The experimental data included several datasets for each modality that were used individually and combined in terms of modalities. All obtained models were examined to provide the most reliable model (system) that fits with the continuous pain data type.

In Section 5.4, an overview of the conducted experiments was provided to recognize continuous pain intensity regarding classification and regression. First, automatic models were trained using the features from each modality individually (Uni-modality experiments); see Section 5.4.1. Second, Decision Fusion [DF] method using fusion mapping, in which individual RF, LSTM, and LSTM-SW, was trained with two modalities (Bi-modality using DF experiments) and all modalities (Multi-modality using DF experiments); the two modalities were FAD/EMG-D and EDA-D, see Section 5.4.2. Third, two LSTM/LSTM-SW were combined by concatenating the last layer (Model Fusion [MF]); these experiments were called Bi-modality using MF. Each LSTM/LSTM-SW was used to handle data from a single modality, FAD/EMG-D for training and testing one model and EDA-D for training and testing the other model (see Section 5.4.4).

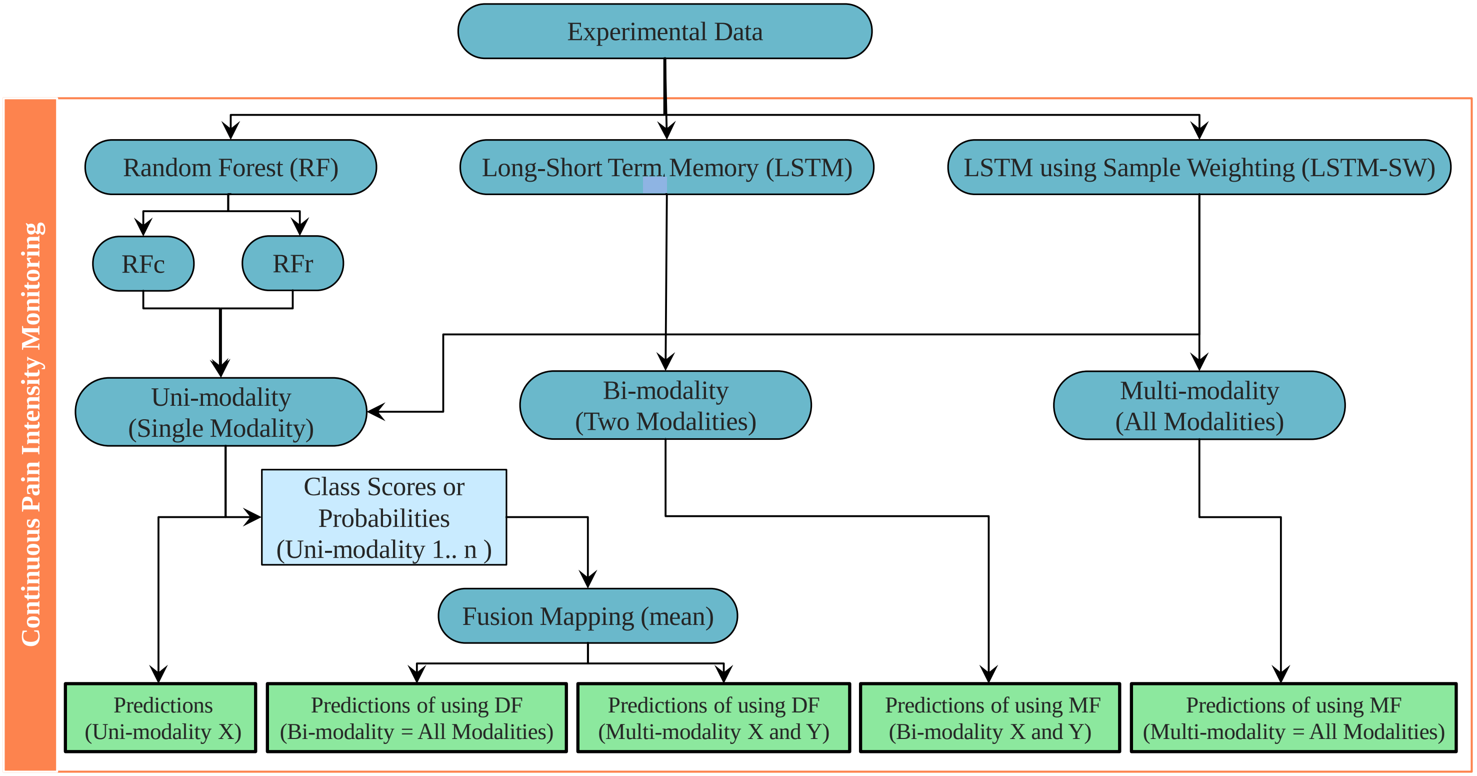


FIG. 5.1. The monitoring pipeline in a continuous pain intensity monitoring system. Three automatic methods were used to train and test the Uni-modality model, Bi-modality model using DF or MF, and Multi-modality model using DF or MF. X is FAD/EMG, Y is EDA. RFc: Random Forest classifier, RFr: Random Forest regression, DF: Decision Fusion, MF: Model Fusion.

In line with Bi-modality using MF experiment, similar LSTM/LSTM-SW were used, but five instead of two LSTM/LSTM-SW were individually trained and tested with FAD, Audio-D, ECG-D, EMG-D, and EDA-D. These experiments were called Multi-modality using Model Fusion (MF); see Section 5.4.5.

5.1 Random Forest Baseline Method

Random Forest [RF] was used because it is an applicable method regarding classification and regression tasks. RF is parallelizable method, which means that the process can be split into multiple machines to run and this leads to a faster computation time (faster to train and predict). In contrast, the Boosting is a sequential ensemble method, which takes longer to compute. Further, RF is good with high dimensionality data, robust to outliers and non-linear data, good to handle imbalanced data, it has also low bias and moderate variance. Alongside Werner et al. [29] and Othman et al. [75], Random Forest classifier (RFc) and Random Forest regression (RFr) were trained with 100 trees and a maximum depth of 10 nodes for classification and regression tasks. RFc method showed good results in predicting pain intensity and no pain from the time windows [29] of samples that were cut out from the continuous recording of the main stimulation phase [29,75]. In this thesis, both RFc and RFr are the baseline methods to compare them with other deep learning methods (LSTM and LSTM-SW). For more details about the comparison results, see Section 6.2.1.2, 6.3.1.2, and 6.4.1.2.

5.2 Long-Short Term Memory

Long-Short Term Memory [LSTM] is an effective method for better handling time series prediction compared to other time series methods because it has a memory cell that can maintain information in memory for long periods of time. It is more accurate on datasets using large sequences. Table 5.1 shows the six Long-Short Term Memory (LSTM) architectures used for classification and regression. The learning rate [lr] is the most important hyperparameter; it is tuned to control how quickly the model is adapted to the problem and how much to change the model in response to the estimated error. The lr range is often between 0.0 and 1.0. Multiple lr were tested to avoid too large or too small lr problems. The experimental data (time series data = samples) from each modality provided after applying the data preparation process were inserted into LSTMs one by one in sequence.

The architectures A(c), B(c), C(c), and D(c) for classification and A(r), B(r) for regression all have input size of 10 × 252/288/36/12, the number of features was variant according to the used modality. 10 indicates timesteps (25 Hz time series were reduced to one Hz after applying tempral integration process), 252 indicates facial features [FAD], 288 indicates audio features [Audio-D], 12 indicates ECG features [ECG-D] or EDA features [EDA-D], and 36 indicates EMG features [EMD-D]. A(c) and C(c) classification architectures comprised a single LSTM layer with 4 units activated by ReLU followed by a flatten layer, and then one dense layer with 128 neurons activated by ReLU. The final output layer had 7 neurons in A(c) and 4 neurons in C(c). B(c) and D(c) classification architectures comprised a single LSTM layer with 8 units activated by ReLU and followed by flatten layer, and then one dense layer with 64 neurons activated by ReLU. The final dense output layer had 7 neurons in B(c) and 4 neurons in D(c). The Softmax was used as the activation function in the output layer, and the Categorical Cross-Entropy [CCE] was the used as the loss function. The configurations of A(r) regression architecture were similar to A(c) and C(c), and the configurations of B(r) regression architecture were similar to B(c) and D(c), except the final output layer with 1 neuron was activated using a sigmoid function. The used loss function was the Binary Cross-Entropy [BCE]. Linear activation function and MSE loss function were also used with the FAD single modality experiment. The obtained models were trained for 2000 epochs with different lr when setting up the batch size equal to 512 and using adam optimizer. chapter 6 presents the lr which results best results for each dataset, which were 10^{-4} or 10^{-5} or 10^{-6}.

Table 5.1. A summary of the LSTM architectures' configurations using data from FAD single modality (FAD Uni-Modality).

Layer type	**Attribute**	**Classification**				**Regression**	
		A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
Input	Size:	10 × 252	10 × 252	10 × 252	10 × 252	10 × 252	10 × 252
	Timestep:	10	10	10	10	10	10
	Features:	252	252	252	252	252	252
LSTM	Activation:	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
	No. of units:	4	8	4	8	4	8
Dropout	with p:	0.5	0.5	0.5	0.5	0.5	0.5
Flatten	Output:	80	40	80	40	80	40
Dense1	Activation:	ReLU	ReLU	ReLU	ReLU	ReLU	ReLU
	No. of units:	128	64	128	64	128	64
Dense2	Activation:	Softmax	Softmax	Softmax	Softmax	Linear/Sigmoid	
	No. of units:	7	7	4	4	1	1
Output	Continuous	-	-	-	-	√	√
	Discrete	√ 7 levels	√ 7 levels	√ 4 levels	√ 4 levels	-	-

The LSTM predicted one output with time period by using several adjacent periods, which kept the estimation line stable, smooth, and closed to the ground-truth labels. Section 3.4 and Section 3.5 describe the activation (Softmax, Linear, and Sigmoid) and loss functions (MSE, BCE, and CCE) that have been used in

LSTM architectures. This work focuses on LSTM/LSTM-SW using BCE regarding regression task due to its better performance compared to those using MSE based on using FAD single modality results, see Table 6.6.

By combining these hyper-parameters, a total of 528 models were trained in a PC " Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz, NVIDIA GeForce RTX 2080 Ti 32 GB RAM". The software libraries and frameworks used were: Python 3.6.10, Tensorflow-GPU 1.14.0, Numpy 1.19.0, and OpenCV-python 4.3.0.36.

5.3 Long-Short Term Memory using Sample Weighting

After observing the highly imbalanced database (see Fig. 4.9), Long-Short Term Memory (LSTM) again was used after increasing the weight of the training samples with more facial responses, called LSTM using Sample Weighting (LSTM-SW). The sample weighting method was based on duplicating some samples with high scores. Fig. 5.2 shows an overview of the methodology sample weighting method with LSTM.

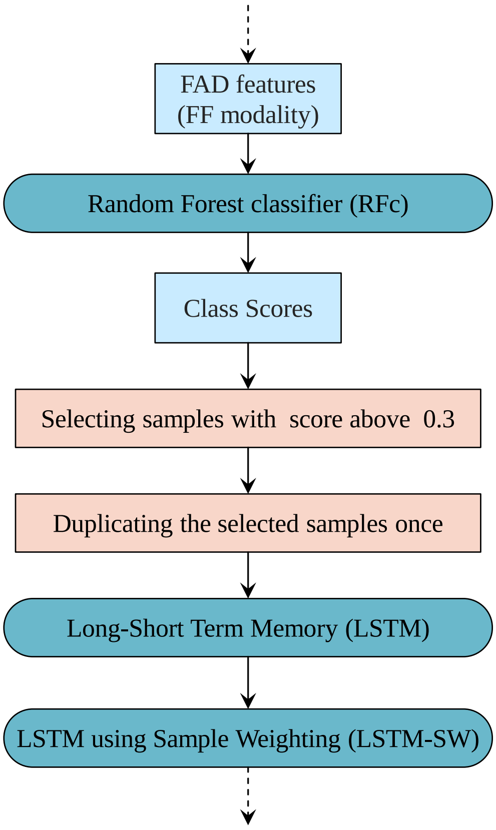

FAD features
(FF modality)
Random Forest classifier (RFc)
Class Scores
Selecting samples with score above 0.3
Duplicating the selected samples once
Long-Short Term Memory (LSTM)
LSTM using Sample Weighting (LSTM-SW)

FIG. 5.2. Overview of LSTM using Sample Weighting method (LSTM-SW).

The samples with prediction scores higher than 0.3 in training data when using the RFc with FAD modality (see RF in Section 5.1) were determined, and then these samples were replicated once. The duplicates are desirable, as some single images could appear multiple times per epoch because the LSTM model puts more weight on getting these samples (with observable pain reaction) correct and less focuses on samples without an observable pain reaction. The samples after increasing were trained on the suggested LSTM (see Section 5.2) for classification and regression. To ensure comparability of test results, samples were never duplicated in the test data.

5.4 Experiments

Here several experiments are described to gain insights into automatic continuous pain intensity monitoring and compare the performance of deep learning models (LSTM and LSTM-SW) to the performance of a baseline model (RF) regarding classification and regression tasks. All LSTM classification models were optimized using the loss function Categorical Cross-Entropy [CCE], and LSTMs regression models were optimized using the loss function Binary Cross-Entropy [BCE]. LSTMs regression models with FAD were also optimized using the loss function Mean Squared Error [MSE]. For reference, a Trivial classifier and regressor were calculated, which always votes for the majority class of the dataset (no pain in our experiments). This section presents three categories of experiments for monitoring continuous pain intensity in the X-ITE Pain Database: Uni-modality experiments using data from single modalities (see Section 5.4.1), Bi-modality experiments using data from two modalities, and Multi-modality experiments using data from multiple modalities. Section 5.4.2 shows the experiments of Bi-modality and Multi-modality using Decision Fusion [DF], and Section 5.4.3 presents the experiments of Bi-modality and Multi-modality when using Model Fusion [MF]. Each model in experiments was trained regarding classification and regression; the discrete predictions indicate classification task, and continuous predictions indicate regression task.

5.4.1 Uni-modality Experiments

Fig. 5.3 shows the Uni-modality experiments when using RF, LSTM, and LSTM-SW. In order to be able to know which modality is best for monitoring continuous pain intensity, the suggested automatic methods were trained with the time series data from each single modality. In these experiments, each time series data (FAD, Audio-D, ECG-D, EMG-D, and EDA-D) was used individually to predict pain intensity using RF, LSTM, and LSTM-SW for classification (discrete predictions) and regression (continuous predictions). In regard to classification and regression, section 6.2 presents the Uni-modality experiments' results for continuous pain intensity monitoring with the X-ITE Pain Database.

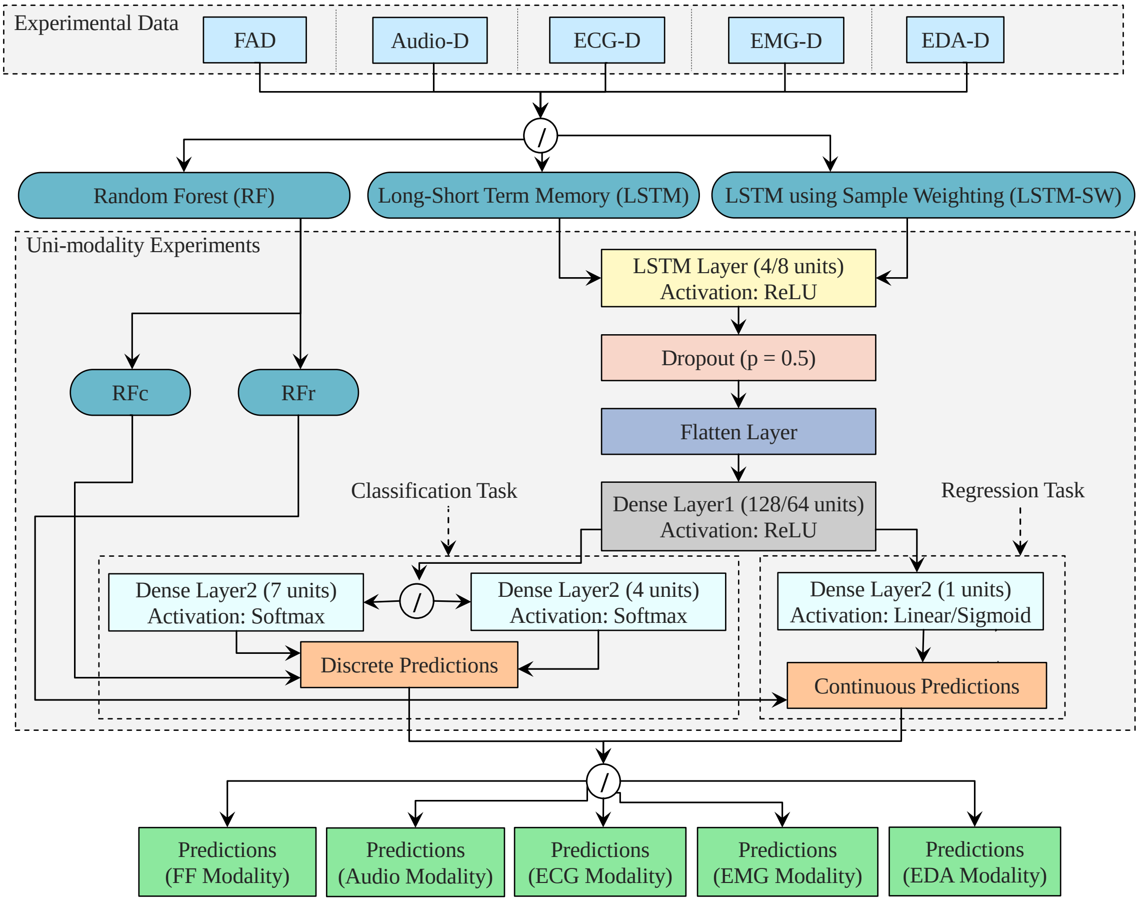


FIG. 5.3. Overview of Uni-modality Experiments. RFc: Random Forest classifier, RFr: Random Forest regression, /: OR.

5.4.2 Decision Fusion Experiments

After observing the performance of individually trained models when using RF, LSTM, and LSTM-SW with Uni-modality experiments, Decision Fusion (DF) was used on obtained predictions; the predictions of two modalities (FAD/EMG-D and EDA-D) or all modalities (FAD, Audio-D, ECG-D, EMG-D, and EDA-D) were used, see Fig. 5.4. In Bi-modality experiments regarding classification and regression: EDA-D (the best performing single modality) was fused once with FAD and once with EMG-D. FAD and EMG-D were the second and the third best performing single modalities; thus, they were used in Bi-modality experiments. The classification models (RFc, LSTM, and LSTM-SW) yield a score for each possible class, and the regression models (RFr, LSTM, and LSTM-SW) predict a continuous value. The classifier scores and regression outputs were aggregated individually into a final decision using a fixed mapping approach. Regarding classification, DF was implemented by calculating the mean of output scores per class of both models using (FAD and EDA-D) or (EMG-D and EDA-D) and selecting the class with the highest score. Regarding regression, all RFr, LSTM, and LSTM-SW predictions were averaged individually in terms of calculating DF. The results of Bi-modality and

Multi-modality models using DF for continuous pain intensity monitoring with the X-ITE Pain Database are presented in Section 6.3 and Section 6.4.

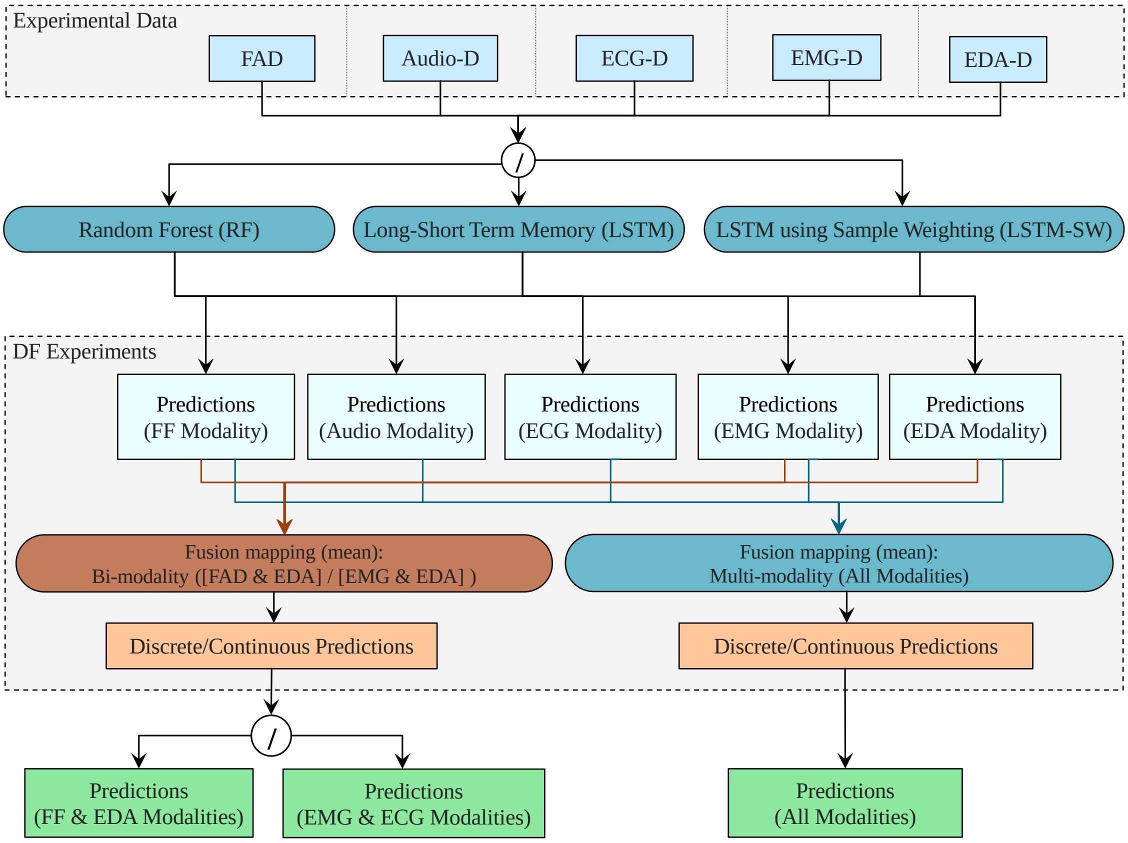

Experimental Data
FAD
Audio-D
ECG-D
EMG-D
EDA-D
/
Random Forest (RF)
Long-Short Term Memory (LSTM)
LSTM using Sample Weighting (LSTM-SW)
DF Experiments
Predictions (FF Modality)
Predictions (Audio Modality)
Predictions (ECG Modality)
Predictions (EMG Modality)
Predictions (EDA Modality)
Fusion mapping (mean): Bi-modality ([FAD & EDA] / [EMG & EDA])
Fusion mapping (mean): Multi-modality (All Modalities)
Discrete/Continuous Predictions
Discrete/Continuous Predictions
/
Predictions (FF & EDA Modalities)
Predictions (EMG & ECG Modalities)
Predictions (All Modalities)

FIG. 5.4. Overview of Bi-modality and Multi-modality using Decision Fusion (DF) experiments. Discrete is the output from the classification task and continuous is the output from the regression task. /: OR.

5.4.3 Model Fusion Experiments

Several experiments were conducted using LSTM and LSTM-SW with time series data in order to increase the performance of continuous pain intensity monitoring. This section describes Bi- and Multi-modality using Model Fusion [MF] experiments, including the combination of two or five LSTM/LSTM-SW models. Alongside DF experiments, two types of experiments were applied: Bi-modality and Multi-modality using MF experiments.

5.4.4 Bi-modality Experiments

Two Uni-modality architectures (FAD/EMG-D and EDA-D) were combined using LSTM/LSTM-SW by merging their final dense layers using a concatenate layer (see Table 5.2 and Fig. 5.5.

Table 5.2. A summary of the LSTM architectures' configurations using data from two modalities (FAD/EMG-D and EDA-D Bi-modality). A(c), B(c), C(c), D(c), A(r), and B(r) are LSTM architectures using Uni-modality (see Table 5.1).

Layer type	**Attribute**	**Architectures Configurations (Bi-modality)**					
		Classification				**Regression**	
		A-Bi(c)	B-Bi(c)	C-Bi(c)	D-Bi(c)	A-Bi(r)	B-Bi(r)
Concatenate (after dense1)	Modality X	A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
	+	+	+	+	+	+	+
	Modality Y	A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
Dense2	Activation:	Softmax	Softmax	Softmax	Softmax	Sigmoid	Sigmoid
	No. of units:	7	7	4	4	1	1
Output	Continuous	-	-	-	-	√	√
	Discrete	√	√	√	√	-	-
		7 levels	7 levels	4 levels	4 levels		

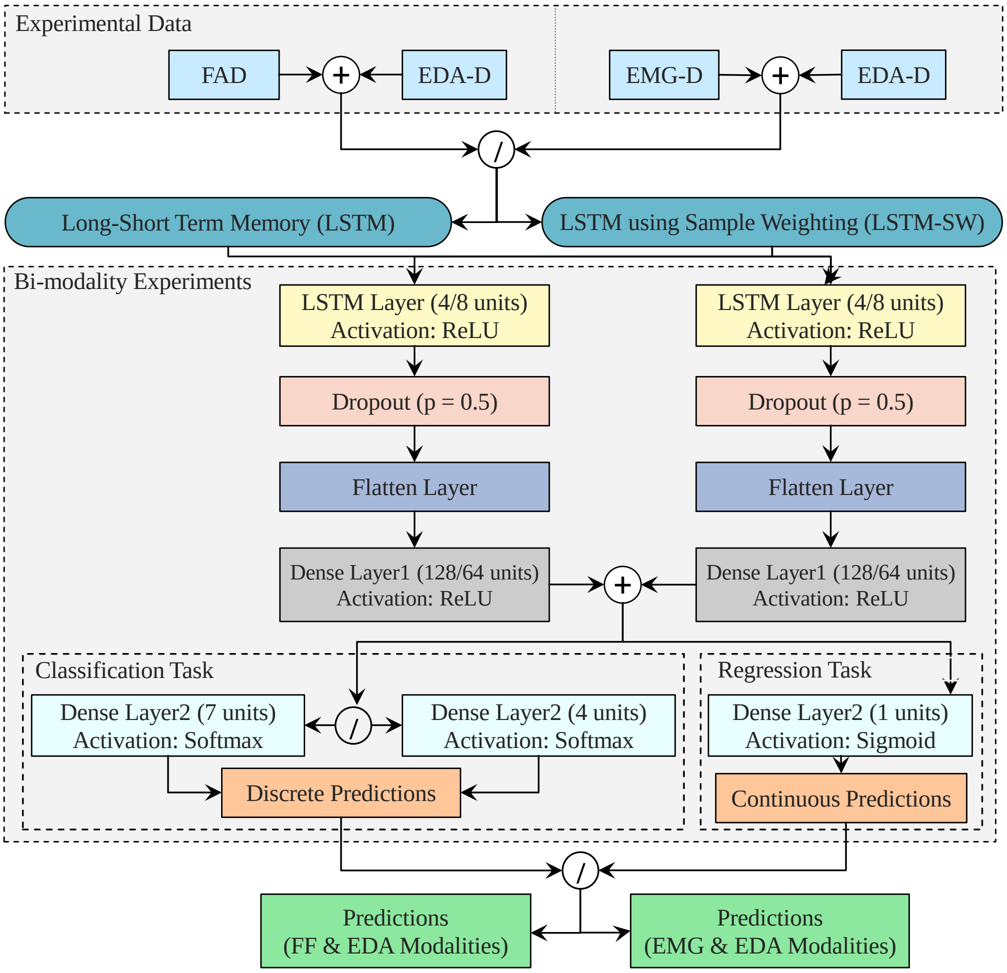


FIG. 5.5. Overview of Bi-modality using Model Fusion (MF) experiments with LSTMs (LSTM and LSTM-SW). MF: Model Fusion, /: OR, +: Concatenate layer.

A-Bi(c) and C-Bi(c) classification architectures of (FAD and EDA-D) Bi-modality or (EMG-D and EDA-D) Bi-modality, including A(c) and C(c), comprised a single LSTM layer with 4 units activated by ReLU and followed by a flatten layer, and then one dense layer with 128 neurons activated by ReLU. The output of dense layer (dense1) from X (EDA-D) modality architecture was concatenated with the output of dense layer (dense1) from Y (FAD/EMG-D) modality. The final layer (dense2) had 7 neurons in A-Bi(c) and 4 neurons in C-Bi(c). B-Bi(c) and D-Bi(c) classification architectures, including B(c) and D(c), comprised a single LSTM layer with 8 units activated by ReLU and followed by flatten layer, and then one dense layer with 64 neurons activated by ReLU. The output of dense layer (dense1) from X (EDA-D) modality architecture was concatenated with the output of dense layer (dense1) from Y (FAD/EMG-D) modality. The final layer (dense2) had 7 neurons in B-Bi(c) and 4 neurons in D-Bi(c). The configurations of A-Bi(r) regression architecture were similar to A-Bi(c) and C-Bi(c), and the configurations of B-Bi(r) regression architecture were similar to B-Bi(c) and D-Bi(c) except the final output layer with 1 neuron.

5.4.5 Multi-modality Experiments

Table 5.3 and Fig. 5.6 show the overview of Multi-modality experiments using MF. All single modalities' architectures using LSTMs were combined by concatenating the outputs from the dense1 layers using concatenate layer.

Table 5.3. A summary of the LSTM architectures' configurations using all modalities. A(c), B(c), C(c), D(c), A(r), and B(r) are LSTM architectures using Uni-modality (see Table 5.1).

Layer type	**Attribute**	**Architectures Configurations (Multi-modality)**					
		Classification				**Regression**	
		A-Mu(c)	B-Mu(c)	C-Mu(c)	D-Mu(c)	A-Mu(r)	B-Mu(r)
Concatenate (after dense1)	Modality 1	A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
	+	+	+	+	+	+	+
	Modality 2	A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
	+	+	+	+	+	+	+
	Modality 3	A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
	+	+	+	+	+	+	+
	Modality 4	A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
	+	+	+	+	+	+	+
	Modality 5	A(c)	B(c)	C(c)	D(c)	A(r)	B(r)
Dense2	Activation:	Softmax	Softmax	Softmax	Softmax	Sigmoid	Sigmoid
	No. of units:	7	7	4	4	1	1
Output	Continuous	-	-	-	-	√	√
	Discrete	√ 7 levels	√ 7 levels	√ 4 levels	√ 4 levels	-	-

The Multi-modality experiments were similar to Bi-modality experiments when using MF, except the input were the time series data from all modalities (FAD, Audio-D, ECG-D, EMG-D, and EDA-D). A-Mu(c) and C-Mu(c) classification architectures of Multi-modality, including A(c) and C(c), comprised a single LSTM layer with 4 units activated by ReLU and followed by a flatten layer, and then one dense layer with 128 neurons activated by ReLU. The dense2 layer had 7 neurons in A-Mu(c) and 4 neurons in C-Mu(c). B-Mu(c) and D-Mu(c) classification architectures, including B(c) and D(c), comprised a single LSTM layer with 8 units activated by ReLU and followed by flatten layer, and then one dense layer with 64 neurons activated by ReLU. The dense2 layer has 7 neurons in B-Mu(c) and 4 neurons in D-Mu(c). The configurations of A-Mu(r) regression architecture were similar to A-Mu(c) and C-Mu(c), and the configurations of B-Mu(r) regression architecture were similar to B-Mu(c) and D-Mu(c) except the final dense out-put layer with 1 neuron. The results of both classification (discrete predictions) and regression (continuous predictions) are presented in Section 6.3 and Section 6.4.

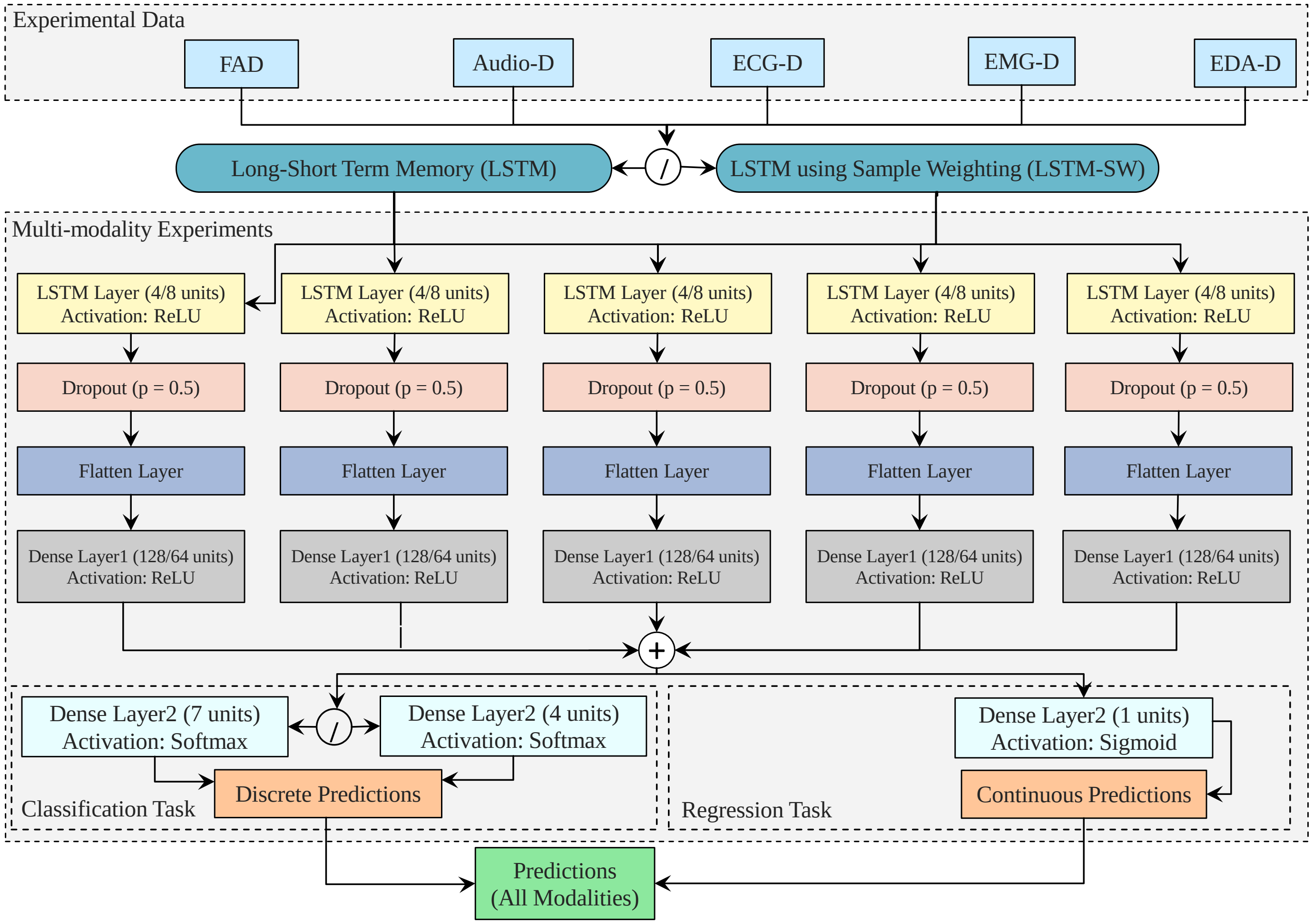

Experimental Data
FAD
Audio-D
ECG-D
EMG-D
EDA-D
Long-Short Term Memory (LSTM)
/
LSTM using Sample Weighting (LSTM-SW)
Multi-modality Experiments
LSTM Layer (4/8 units) Activation: ReLU
Dropout (p = 0.5)
Flatten Layer
Dense Layer1 (128/64 units) Activation: ReLU
+
Dense Layer2 (7 units) Activation: Softmax
/
Dense Layer2 (4 units) Activation: Softmax
Discrete Predictions
Classification Task
Dense Layer2 (1 units) Activation: Sigmoid
Continuous Predictions
Regression Task
Predictions (All Modalities)

FIG. 5.6. Overview of Mutli-modality using Model Fusion (MF) experiments with LSTMs (LSTM and LSTM-SW). MF: Model Fusion, /: OR, +: Concatenate layer.

CHAPTER 6

Evaluation

THIS chapter provides a detailed description of experimental results regarding classification and regression. Many experiments were conducted to evaluate the three proposed methods (RF, LSTM, and LSTM-SW) for continuous monitoring of pain intensity with 11 proposed datasets (7-Class and 4-Class datasets) from the X-ITE Pain Database (experimental data, see Section 4.4); for more details about these experiments, see Section 5.4. This section describes steps to evaluate those methods in all experiments. Mean Squared Error [MSE] and the Intraclass Correlation Coefficient [ICC] were used on the test set to measure the performance of classification models versus regression models; the best performances were determined when MSE got the smallest values and ICC got the highest values. Further, the best classification models that outperformed the regression models were further evaluated regarding classification using different classification measures on the test set. For more details about the measures, see Section 6.1. Additionally, paired t-test was used to calculate the p-value for evaluating if the performances of the LSTM and LSTM-SW classification models were significantly better than the baseline models (Random Forest classifier [RFc]). The findings from the Uni-modality (single modality) models were presented in Section 6.2, followed by the findings from the Bi-modality (two fused modalities) models that were shown in Section 6.3. Finally, in Section 6.4, the findings from the Multi-modality (all fused modalities) models were summarized.

6.1 Evaluation Measures

Mean Squared Error [MSE] and the intraclass correlation coefficient [ICC] [185] were calculated on the test set to compare the performances of classification versus regression models after normalizing the output between 0 and 1. Further, the classification models were also evaluated using Micro average precision (Micro avg. precision), Micro average recall (Micro avg. recall), and Micro average F1-score (Micro avg. F1-score), which are useful when datasets vary in size (aggregate the contributions of all classes to compute the average metric). Additionally, accuracy was calculated, but the previously mentioned measures were better when the class sizes were unbalanced. Both classification and regression measures that use in the applied experiments are shown below.

- Classification:

$$Accuracy = \frac{True\ Postive\ (TP) + True\ Negative\ (TN)}{n} \tag{6.1}$$

Percentage of correctly classified samples.

$$Micro\ avg.\ precision = \frac{Sum\ the\ TP\ of\ all\ classes}{Sum\ the\ TP\ of\ all\ classes + Sum\ the\ FP\ of\ all\ classes} \tag{6.2}$$

An average per-class agreement of the data class labels with those of a classifier.

$$Micro\ avg.\ recall = \frac{Sum\ the\ TP\ of\ all\ classes}{Sum\ the\ TP\ of\ all\ classes + Sum\ the\ FN\ of\ all\ classes} \tag{6.3}$$

An average per-class effectiveness of a classifier to identify class labels.

$$Micro\ avg.\ F1\text{-}score = 2 * \frac{Micro\ avg\ Precision * Micro\ avg\ recall}{Micro\ avg\ Precision + Micro\ avg\ Recall} \tag{6.4}$$

Relations between data's positive labels and those given by a classifier based on a per-class average.

- Regression:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{3.22}$$

The average squared difference between the predicted values $y_i - \hat{y}$ and the actual value y_i. n is the total number of samples.

$$ICC = \frac{BMS - EMS}{BMS + (k-1)EMS} \tag{6.5}$$

Intraclass correlation coefficent. ICC (3,1) [185] used to assess measure reliability based on average of k measurements (conditions, raters). BMS: Between-targets means square, EMS: Within-targets means square.

6.2 Uni-modality Results

This section shows the comparison between the Trivial (majority of vote = no pain) and proposed methods (RF, LSTM, and LSTM-SW) with individual modalities from the experimental data, including the best three single modalities (see Section 6.2.1). Further evaluation measures were applied to the classification models that outperformed the regression models when conducting Uni-modality experiments (see Section 6.2.2). For more details about the results of all Uni-modality models, see Appendix A. Finally, the discussion of the results from Uni-modality models is summarized in Section 6.2.3.

6.2.1 Classification vs Regression

The results of applying the single modalities models (Uni-modality models) when applying RF, LSTM, and LSTM-SW with a single modality from experimental data were provided in Section 6.2.1.1. The comparison between the best classification and regression models were shown in Section 6.2.1.2.

6.2.1.1 Modeling Methods

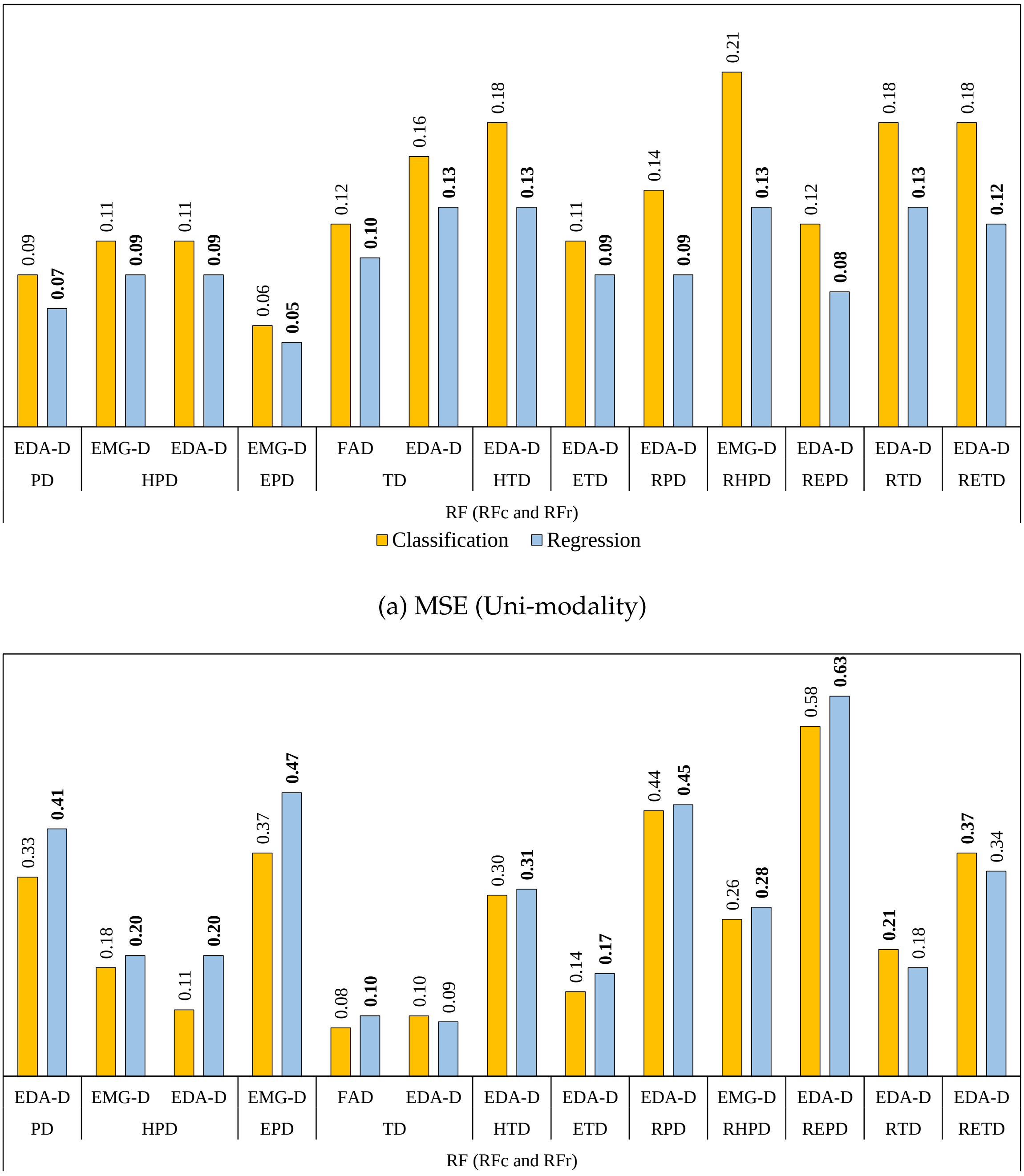
Table 6.1 shows that all EDA-D Uni-modality baseline models (Random Forest classifier [RFc] and Random Forest regression [RFr]) are superior to those with FAD except with Tonic Dataset [TD]. FAD Uni-modality model with TD when applying RFr performed the best; it got the MSE of 0.10 and the ICC of 0.10. EMG-D Uni-modality models got the highest ICC values and smallest recognition error when applying RFc and RFr with Heat Phasic Dataset [HPD] and Reduced Heat

Phasic Dataset [RHPD]; RFc models got the MSE of 0.11, 0.21 and the ICC of 0.18, 0.26; RFr models with HPD and RHPD got the MSE of 0.09, 0.13 and the ICC of 0.20, 0.28, respectively.

Table 6.1. Comparison of the best Uni-modality models when applying RF (RFc and RFr) regarding classification and regression tasks with MSE and ICC measures. Triv.: Trivial, Red. Subsets: Reduced Subsets. The cells with a light grey background indicate the best results regarding classification and regression tasks. The **bold** font indicates the best results.

Measure	Model			Triv.	RFc (Classification)			RFr (Regression)		
	Dataset		n-Class	-	FAD	EMG-D	EDA-D	FAD	EMG-D	EDA-D
MSE	Subsets	PD	7	0.10	0.10	0.10	0.09	0.09	0.08	**0.07**
		HPD	4	0.11	0.11	0.11	0.11	0.09	**0.09**	**0.09**
		EPD	4	0.07	0.07	0.07	0.06	0.06	0.06	**0.05**
		TD	7	0.12	0.12	0.16	0.16	**0.10**	0.7	0.13
		HTD	4	0.41	0.25	0.19	0.18	0.13	0.14	**0.13**
		ETD	4	0.09	0.09	0.17	0.11	0.09	0.14	**0.09**
	Red. Subsets	RPD	7	0.23	0.20	0.19	0.14	0.12	0.11	0.09
		RHPD	4	0.26	0.23	0.21	0.22	0.13	**0.13**	0.13
		REPD	4	0.26	0.21	0.18	0.12	0.12	0.12	**0.08**
		RTD	7	0.25	0.23	0.21	0.18	0.14	0.13	**0.13**
		RETD	4	0.25	0.24	0.21	0.18	0.15	0.13	**0.12**
ICC	Subsets	PD	7	0	0.10	0.17	0.33	0.13	0.19	**0.41**
		HPD	4	0	0.16	0.18	0.11	0.19	**0.20**	**0.20**
		EPD	4	0	0.16	0.23	0.37	0.18	0.24	**0.47**
		TD	7	0	0.08	0.09	0.10	**0.10**	0.04	0.09
		HTD	4	0	0.11	0.29	0.30	0.17	0.22	**0.31**
		ETD	4	0	0.07	0.06	0.14	0.09	0.07	**0.17**
	Red. Subsets	RPD	7	0	0.19	0.28	0.44	0.23	0.30	**0.45**
		RHPD	4	0	0.21	0.26	0.23	0.26	**0.28**	0.24
		REPD	4	0	0.27	0.38	0.58	0.32	0.38	**0.63**
		RTD	7	0	0.09	0.17	**0.21**	0.05	0.14	0.18
		RETD	4	0	0.12	0.26	**0.37**	0.15	0.28	0.34

Fig. 6.1 shows the best results from RFc and RFr models. Most regression models (RFr) got the highest ICC values and smallest recognition error. EDA-D Uni-modality models got the highest ICC values when applying RFc with only Reduced Tonic Dataset [RTD] and Reduced Electrical Tonic Dataset [RETD]. They got the ICC of 0.21, 0.37 and the MSE of 0.18, 0.18, when those EDA-D Uni-modality models using RFr got the ICC of 0.18, 0.34 and smallest recognition error, the MSE of 0.13, 0.12, respectively. Regarding classification and regression, the EDA-D Uni-modality models, when applying LSTM, performed the best except with those models with TD and Electrical Tonic Dataset [ETD], EMG-D Uni-modality models performed better.



(a) MSE (Uni-modality)

(b) ICC (Uni-modality)

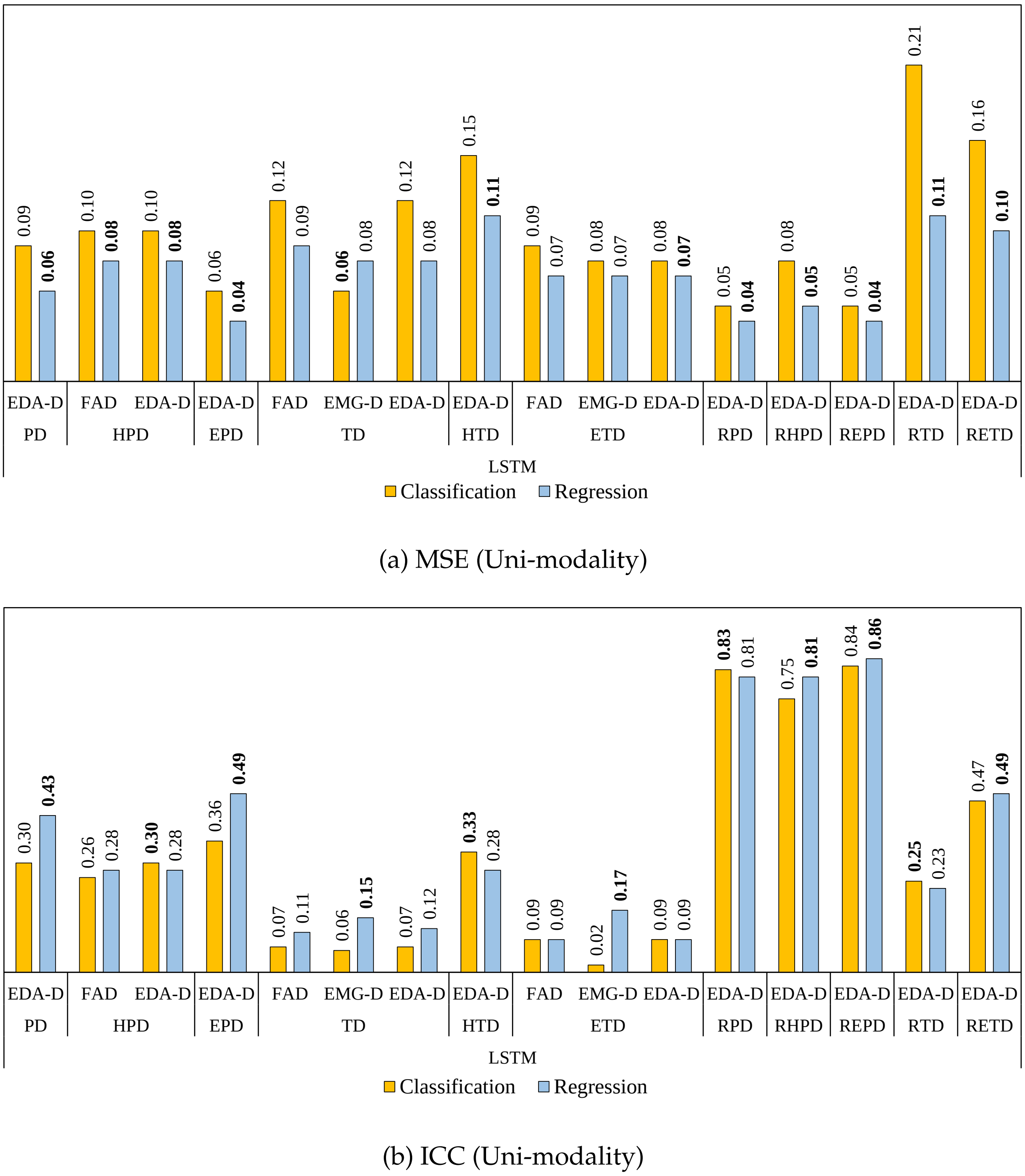
FIG. 6.1. Comparison of the best Uni-modality models when applying RFc and RFr regarding classification and regression tasks with MSE and ICC measures. The **bold** font indicates the best results.

Table 6.2 shows that EMG-D Uni-modality regression models with TD and ETD got the MSE of 0.08, 0.07, and the ICC of 0.15, 0.17, respectively. Further, EDA-D Uni-modality models obtained similar results to FAD Uni-modality models with HPD regarding regression, TD, and ETD regarding classification.

Table 6.2. Comparison of the best Uni-modality models when applying LSTM regarding classification and regression tasks with MSE and ICC measures. Triv.: Trivial, Red. Subsets: Reduced Subsets. The cells with a light grey background indicate the best results regarding classification and regression tasks. The **bold** font indicates the best results.

Measure	Model			Triv.	LSTM (Classification)			LSTM (Regression)				loss		Learning rate
	Dataset		n-Class	-	FAD CCE	EMG-D CCE	EDA-D CCE	FAD MSE	FAD BCE	EMG-D BCE	EDA-D BCE	CCE	MSE or BCE	
												Architecture		
MSE	Subsets	PD	7	0.10	0.09	0.09	0.09	0.08	0.08	0.07	**0.06**	A(c)	A(r)	10^{-5}
		HPD	4	0.11	0.10	0.10	0.10	0.08	0.08	0.08	0.08	C(c)	C(r)	
		EPD	4	0.07	0.07	0.06	0.06	0.05	0.05	0.05	**0.04**	C(c)	C(r)	
		TD	7	0.12	0.12	0.12	0.12	0.09	0.09	**0.08**	0.08	A(c)	A(r)	10^{-6}
		HTD	4	0.41	0.18	0.16	0.15	0.13	0.12	0.11	0.11	C(c)	C(r)	
		ETD	4	0.09	0.09	0.09	0.08	0.08	0.07	**0.07**	0.07	C(c)	C(r)	
	Red. Subsets	RPD	7	0.23	0.12	0.16	0.05	0.10	0.08	0.09	0.04	A(c)	A(r)	10^{-4}
		RHPD	4	0.26	0.13	0.16	0.08	0.10	0.09	0.10	**0.05**	C(c)	C(r)	
		REPD	4	0.26	0.13	0.16	0.05	0.11	0.10	0.09	**0.04**	C(c)	C(r)	
		RTD	7	0.25	0.25	0.21	**0.21**	0.13	0.13	0.12	0.11	B(c)	B(r)	10^{-6}
		RETD	4	0.25	0.24	0.19	0.16	0.14	0.14	0.12	**0.10**	D(c)	D(r)	
ICC	Subsets	PD	7	0	0.18	0.20	0.30	0.18	0.20	0.26	**0.43**	A(c)	A(r)	10^{-5}
		HPD	4	0	0.26	0.21	**0.30**	0.27	0.28	0.24	0.28	C(c)	C(r)	
		EPD	4	0	0.25	0.27	0.36	0.24	0.27	0.32	**0.49**	C(c)	C(r)	
		TD	7	0	0.07	0.06	0.07	0.14	0.11	**0.15**	0.12	A(c)	A(r)	10^{-6}
		HTD	4	0	0.19	0.29	**0.33**	0.13	0.15	0.18	0.28	C(c)	C(r)	
		ETD	4	0	0.09	0.02	0.09	0.13	0.09	**0.17**	0.09	C(c)	C(r)	
	Red. Subsets	RPD	7	0	0.57	0.44	**0.83**	0.49	0.56	0.45	0.81	A(c)	A(r)	10^{-4}
		RHPD	4	0	0.56	0.48	0.75	0.58	0.62	0.5	**0.81**	C(c)	C(r)	
		REPD	4	0	0.55	0.49	0.84	0.50	0.52	0.56	**0.86**	C(c)	C(r)	
		RTD	7	0	0.05	0.21	**0.25**	0.08	0.04	0.16	0.23	B(c)	B(r)	10^{-6}
		RETD	4	0	0.15	0.33	0.47	0.14	0.09	0.28	**0.49**	D(c)	D(r)	

Fig. 6.2 shows that the EDA-D Uni-modality classification models got the highest ICC values when applying LSTM with HPD, Heat Tonic Dataset [HTD], Reduced Phasic Datasets [RPD], and RTD. The ICC values of HPD, HTD, RPD and RTD are 0.30, 0.33, 0.83 and 0.25, respectively. However, the EDA-D Uni-modality regression models, when applying LSTM, got the smallest recognition error on the same 4 datasets. Further, the Uni-modality regression models got the highest ICC values and smallest recognition error on the rest 7 datasets.



(a) MSE (Uni-modality)

(b) ICC (Uni-modality)

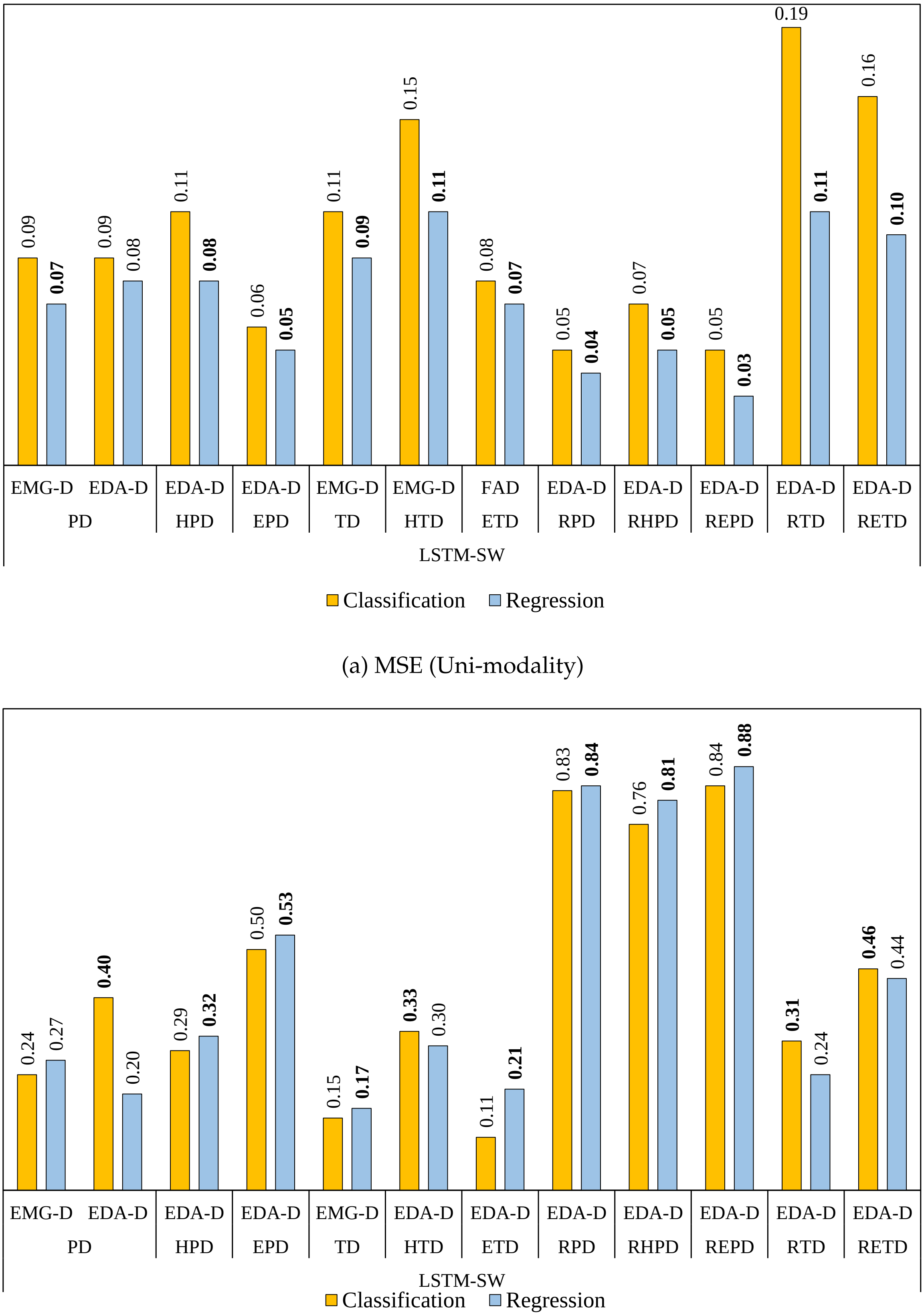
FIG. 6.2. Comparison of the best Uni-modality models when applying LSTM regarding classification and regression tasks with MSE and ICC measures. The **bold** font indicates the best results.

Most Uni-modality classification and regression models, when applying LSTM-SW with EDA-D, performed the best, see Table 6.3. EMG-D Uni-modality models with TD were better than FAD Uni-modality and EDA-D Uni-modality models; they got the MSE of 0.11, 0.09, and the ICC of 0.15, 0.17 regarding classification and regression, respectively. Further, the EMG-D Uni-modality classification model with HTD yielded the highest ICC (0.33) and the MSE of 0.15; the EDA-D Uni-modality regression model got the ICC of 0.30 and the MSE of 0.11.

Table 6.3. Comparison of the best Uni-modality models when applying LSTM-SW regarding classification and regression tasks with MSE and ICC measures. Triv.: Trivial, Red. Subsets: Reduced Subsets. The cells with a light grey background indicate the best results regarding classification and regression tasks. The **bold** font indicates the best results.

Measure	Model			Triv.	LSTM-SW (Classification)			LSTM-SW (Regression)			loss		Learning rate
		Dataset	n-Class	-	FAD CCE	EMG-D CCE	EDA-D CCE	FAD BCE	EMG-D BCE	EDA-D BCE	CCE	BCE	
											Architecture		
MSE	Subsets	PD	7	0.10	0.10	0.09	0.09	0.08	0.07	0.08	A(c)	A(r)	10^{-5}
		HPD	4	0.11	0.11	0.11	0.11	0.09	0.09	**0.08**	C(c)	C(r)	
		EPD	4	0.07	0.08	0.07	0.06	0.06	0.06	**0.05**	C(c)	C(r)	
		TD	7	0.12	0.12	0.11	0.12	0.09	**0.09**	0.09	A(c)	A(r)	10^{-6}
		HTD	4	0.41	0.16	0.15	0.15	0.16	0.11	0.11	C(c)	C(r)	
		ETD	4	0.09	0.08	0.09	0.08	0.08	0.07	**0.07**	C(c)	C(r)	
	Red. Subsets	RPD	7	0.23	0.14	0.15	0.05	0.09	0.09	**0.04**	A(c)	A(r)	10^{-4}
		RHPD	4	0.26	0.14	0.17	0.07	0.09	0.10	**0.05**	C(c)	C(r)	
		REPD	4	0.26	0.14	0.15	0.05	0.11	0.10	**0.03**	C(c)	C(r)	
		RTD	7	0.25	0.24	0.22	0.19	0.13	0.12	0.11	B(c)	B(r)	10^{-6}
		RETD	4	0.25	0.26	0.19	0.16	0.13	0.12	0.10	D(c)	D(r)	
ICC	Subsets	PD	7	0	0.20	0.24	**0.40**	0.22	0.27	0.20	A(c)	A(r)	10^{-5}
		HPD	4	0	0.26	0.25	0.29	0.27	0.26	**0.32**	C(c)	C(r)	
		EPD	4	0	0.24	0.32	0.50	0.28	0.34	**0.53**	C(c)	C(r)	
		TD	7	0	0.08	0.15	0.11	0.12	**0.17**	0.11	A(c)	A(r)	10^{-6}
		HTD	4	0	0.2	**0.33**	0.31	0.15	0.27	0.30	C(c)	C(r)	
		ETD	4	0	0.11	0.06	0.09	0.11	0.19	**0.21**	C(c)	C(r)	
	Red. Subsets	RPD	7	0	0.49	0.44	0.83	0.54	0.45	**0.84**	A(c)	A(r)	10^{-4}
		RHPD	4	0	0.55	0.45	0.76	0.62	0.53	**0.81**	C(c)	C(r)	
		REPD	4	0	0.52	0.52	0.84	0.51	0.53	**0.88**	C(c)	C(r)	
		RTD	7	0	0.08	0.18	**0.31**	0.07	0.17	0.24	B(c)	B(r)	10^{-6}
		RETD	4	0	0.10	0.30	**0.46**	0.20	0.28	0.44	D(c)	D(r)	

Fig. 6.3 shows that Uni-modality classification models, when applying LSTM-SW, got the highest ICC values with 4 datasets: Phasic Dataset [PD], HTD, RTD, and Reduced Electrical Tonic Dataset [RETD]. The ICC values were 0.40, 0.33, 0.31 and 0.46, respectively. However, the Uni-modality regression models, when applying LSTM-SW, got the smallest recognition error on the same 4 datasets and the highest ICC values and smallest recognition error on the rest 7 datasets.



(a) MSE (Uni-modality)

(b) ICC (Uni-modality)

FIG. 6.3. Comparison of the best Uni-modality models when applying LSTM-SW regarding classification and regression tasks with MSE and ICC measures. The **bold** font indicates the best results.

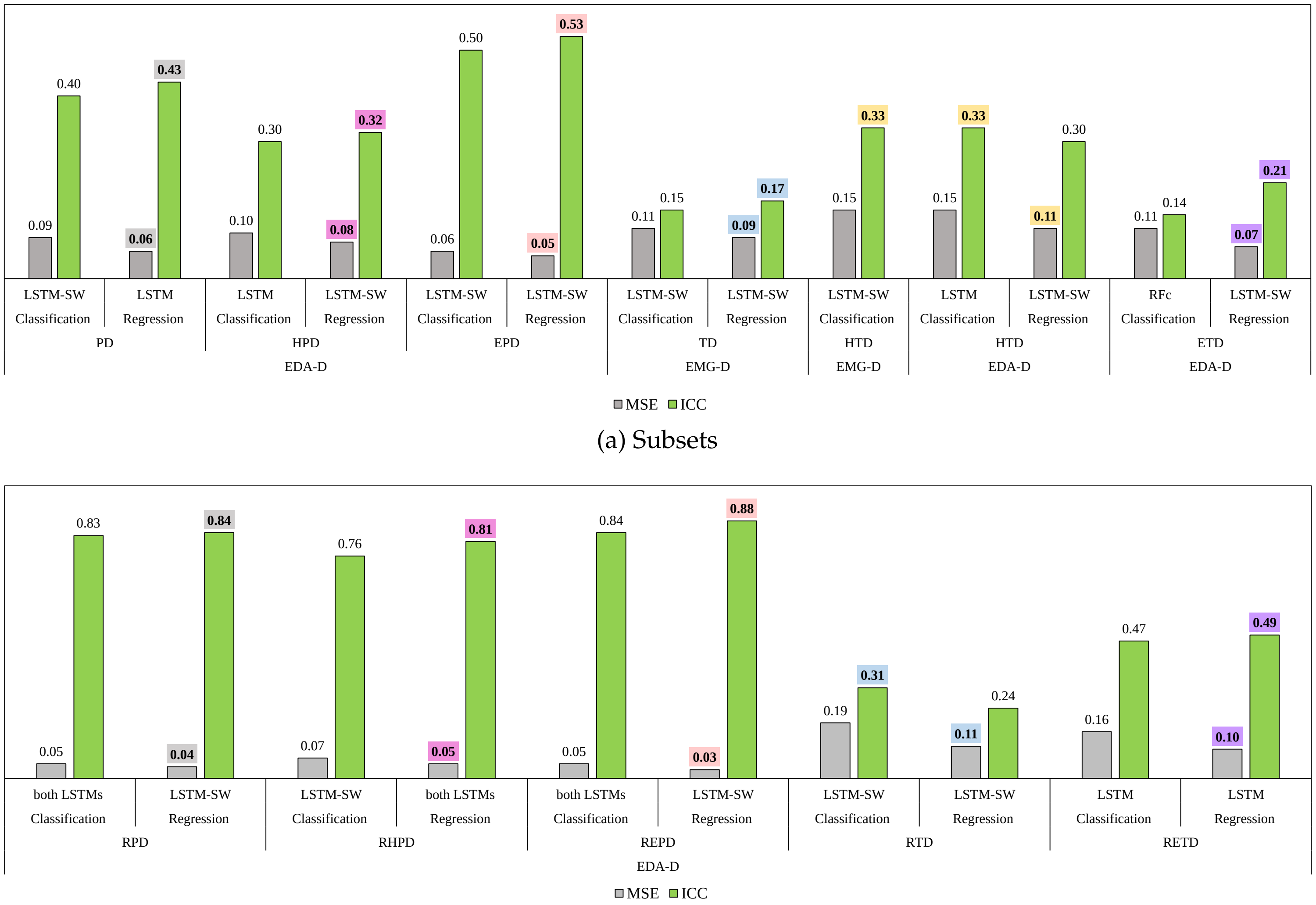
6.2.1.2 Comparison of Modeling Methods

This section provides the comparison between the Uni-modality models when applying baseline methods (RFc and RFr), LSTM and LSTM-SW with 11 datasets from the experimental data, see Table 6.4 and Fig. 6.4.

Table 6.4. Comparison of the best Uni-modality models regarding classification and regression tasks with MSE and ICC measures. Meas.: Measure. The cells with numbers only indicate models with EDA-D results, the **bold** font indicates the best results.

Meas.	Task		Classification			Regression		
	Dataset		RFc	LSTM	LSTM-SW	RFr	LSTM	LSTM-SW
MSE	Subsets	PD	0.09	0.09	0.09	0.07	**0.06**	0.07 (EMG-D)
		HPD	0.11 (EMG-D)	0.10	0.11	0.09 (EMG-D) (EDA-D)	0.08 (FAD) (EDA-D)	**0.08**
		EPD	0.06	0.06	0.06	0.05	0.04	0.05
		TD	0.16	0.12 (FAD) (EDA-D)	0.11 (EMG-D)	0.10 (FAD)	**0.08** (EMG-D)	**0.09** (EMG-D)
		HTD	0.18	0.15	0.15 (EMG-D)	0.13	0.11	0.11
		ETD	0.11	0.08	0.08 (FAD)	0.09	0.07 (EMG-D)	**0.07**
	Red. Subsets	RPD	0.14	0.05	0.05	0.09	0.04	**0.04**
		RHPD	0.21 (EMG-D)	0.08	0.07	0.13	**0.05**	**0.05**
		REPD	0.12	0.05	0.05	0.08	0.04	**0.03**
		RTD	0.18	0.21	0.19	0.13	0.11	0.11
		RETD	0.18	0.16	0.16	0.12	**0.10**	0.10
ICC	Subsets	PD	0.33	0.30	0.40	0.41	**0.43**	0.27 (EMG-D)
		HPD	0.18 (EMG-D)	0.30	0.29	0.20 (EMG-D) (EDA-D)	0.28 (FAD) (EDA-D)	**0.32**
		EPD	0.37	0.36	0.50	0.47	0.49	**0.53**
		TD	0.10	0.07 (FAD) (EDA-D)	0.15 (EMG-D)	0.10 (FAD)	0.15 (EMG-D)	**0.17** (EMG-D)
		HTD	0.3	**0.33**	**0.33** (EMG-D)	**0.31**	0.28	0.30
		ETD	0.14	0.09	0.11 (FAD)	0.17	0.17 (EMG-D)	**0.21**
	Red. Subsets	RPD	0.44	0.83	0.83	0.45	0.81	**0.84**
		RHPD	0.26 (EMG-D)	0.75	0.76	0.28	**0.81**	**0.81**
		REPD	0.58	0.84	0.84	0.63	0.86	**0.88**
		RTD	0.21	0.25	**0.31**	0.18	0.23	0.24
		RETD	0.37	0.47	0.46	0.34	**0.49**	0.44

Most EDA-D Uni-modality models are better than those with FAD Uni-modality and EMG Uni-modality models; see cells with numbers only in Table 6.4. The **bold** font indicates the best LSTM results.



(a) Subsets

(b) Reduced Subsets

FIG. 6.4. Comparison of the best Uni-modality models regarding classification and regression tasks with MSE and ICC measures. The **bold** and font and colored background indicates the best results.

The cells with a light grey background in Table 6.4 indicate the best results regarding classification and regression tasks. Both LSTMs (LSTM and LSTM-SW) obtained similar results with HTD, RPD, and RETD regarding classification. EMG-D Uni-modality and EDA-D Uni-modality models, when applying RFr with HPD, obtained similar results. FAD Uni-modality and EDA-D Uni-modality models, when applying LSTM, obtained similar results when using TD regarding classification and HPD regarding regression. Most regression models outperformed classification models. The best results for each dataset were summarized in Fig. 6.4. The regression models outperformed classification models except with HTD and RTD. The highest ICC with HTD was 0.33 when using EDA-D Uni-modality and EMG-D Uni-modality classification models when applying LSTM and LSTM-SW. The highest ICC with RTD was 0.31 when using the EDA-D Uni-modality classification model when applying LSTM-SW. Further, the performance was improved when using the reduced datasets.

6.2.2 Classification

Table 6.5 and Fig. 6.5 shows how the best Uni-modality classification models, when applying RFC, LSTM, and LSTM-SW with HTD and RTD, successfully predict discrete pain intensity levels in sequences compared to Trivial. The best results were obtained from models when using EDA-D modality except some models when using EMG-D performed the best when using: LSTM with HTD in terms of Micro avg. recall (99.6%), RFc with RTD in terms of accuracy (35%), and LSTM with RTD in terms of Micro avg. recall (10.4%), and Micro avg. F1-score (15.2%). However, these results were not the best; LSTM and LSTM-SW models Uni-modality classification EDA-D modality performed the best (outperform RFc) with HTD. RFc when using EDA-D modality achieved the best Micro avg. recall and Micro avg. F1-Score results when using RTD (about 38% and 25%), whereas LSTM and LSTM-SW model results were the best in terms of accuracy and Micro avg. precision with the same dataset (RTD).

Table 6.5. Comparison of the best Uni-modality models with HTD and RTD regarding classification task and in terms of classification measures. The **bold** font indicates the best results. Triv.: Trivial. * $p < 0.05$ when using paired t-test between RFc and LSTMs (LSTM and LSTM-SW). The architectures of LSTM and LSTM-SW are C(c) and B(c) with 10^{-6} learning rate for HTD and RTD, respectively.

Dataset	Measurement	Uni-modality	Triv.	RFc	LSTM	LSTM-SW
HTD	Accuracy %	FAD	20	29.1	32.81	33.7
		EMG-D	20	35.2	39.3	39.9
		EDA-D	20	41.0	**48.4**	**47.7**
	Micro avg. precision	FAD	0	31.1	33.28	33.7
		EMG-D	0	38.0	39.3	39.9
		EDA-D	0	42.7	**48.2**	**47.7**
	Micro avg. recall	FAD	0	49.4	92.42*	**100***
		EMG-D	0	65.5	**99.6***	**100***
		EDA-D	0	71.0	94.6*	**100***
	Micro avg. F1-score	FAD	0	37.6	46.6	48.2*
		EMG-D	0	46.0	55.6*	56.3
		EDA-D	0	52.9	**62.3***	**62.5***
RTD	Accuracy %	FAD	38.1	33.9	39.2*	37.4
		EMG-D	38.1	35	41.6*	40.5*
		EDA-D	38.1	30.4	**42.2***	**42.7***
	Micro avg. precision	FAD	0	15.8	30.7	37.7*
		EMG-D	0	20.2	35.4*	32.4*
		EDA-D	0	19.10	**44.6***	**40.8***
	Micro avg. recall	FAD	0	11.7	5.55	9.48
		EMG-D	0	**19.9**	10.4	9.70
		EDA-D	0	**38.2**	8.90	10.5
	Micro avg. F1-score	FAD	0	12.8	8.26	12.2
		EMG-D	0	**19.6**	15.2	13.9
		EDA-D	0	**25.1**	14.3	16.2

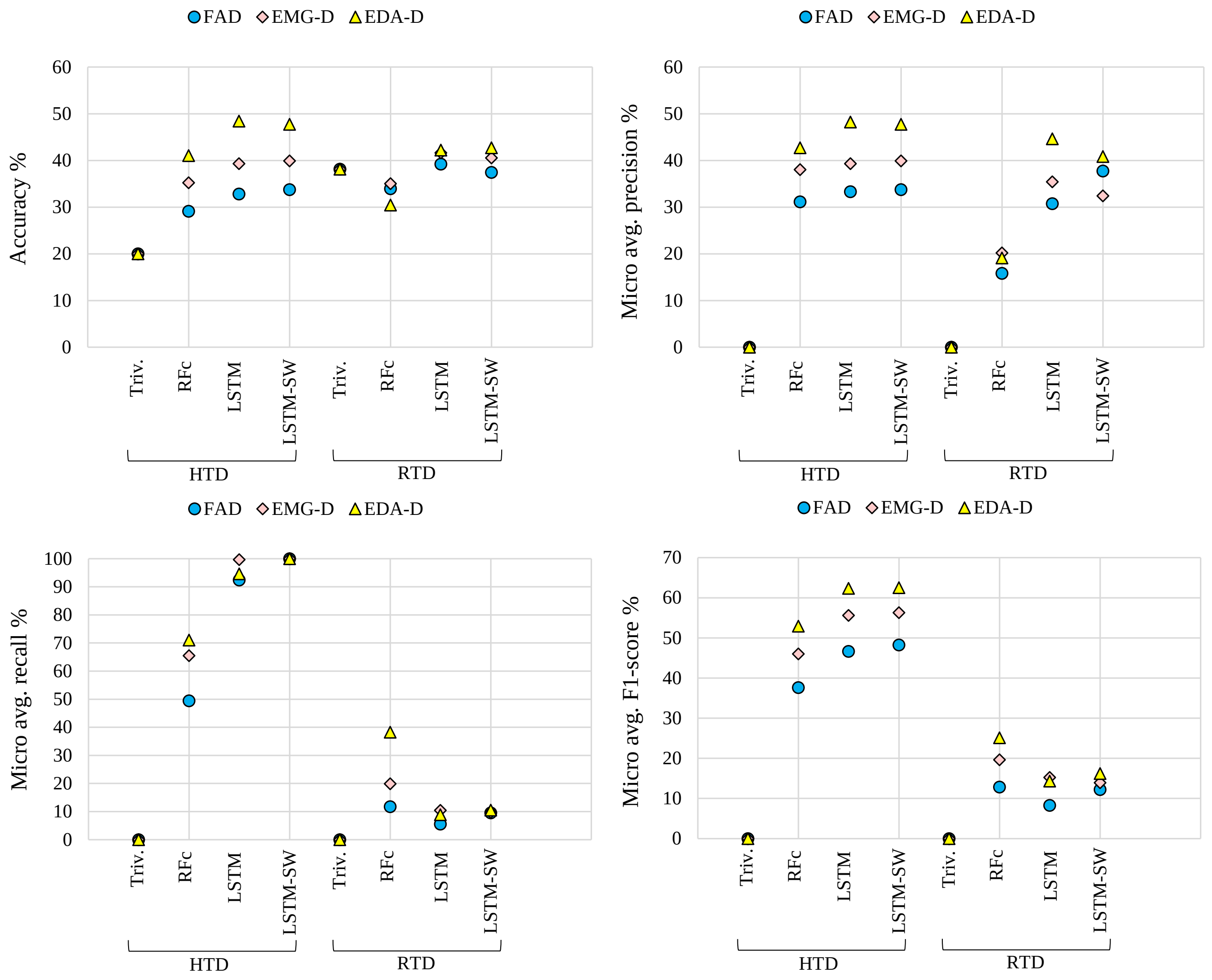


FIG. 6.5. Comparison of Uni-modality models when applying Trivial, RFcs, LSTM, and LSTM-SW) with HTD and RTD regarding classification task. Triv.: Trivial.

6.2.3 Discussion

In this section, the obtained results of Uni-modality models were summarized when applying RFc, LSTM, and LSTM-SW on FAD, EMG-D, and EDA-D modalities; each model was trained and tested on 11 datasets from the experimental data. Regarding the regression tasks, FAD Uni-modality models, when applying LSTM, were evaluated to decide which loss function is better: MSE and BCE; the obtained results show that the BCE was the best in 7 out of 11 datasets (see Table 6.6 and Fig. 6.6). Thus, the LSTM and LSTM-SW were used in the rest experiments with BCE loss. After comparing the performance between Uni-modality models for continuous monitoring of pain intensity, most regression models perform better than classification models, see Table 6.4. Further, all results of the EDA-D Uni-modality models were superior to FAD Uni-modality and EMG-D Uni-modality models, except with TD, EMG-D Uni-modality models were better. The EMG-D Uni-modality classification models performed the best with HTD and RTD (see Table 6.5).

The **bold** font indicates the best results of LSTM models regarding regression task. The proposed reduction strategy on Subsets datasets improves the performance further; see Reduced Subsets results in Fig. 6.4. Additionally, almost all LSTM and LSTM-SW models' results are significantly better than the baseline models (RFc and RFr). RFc model with RTD outperforms LSTM and LSTM-SW regarding classification (see Fig. 6.5). Finally, using the sample weighting method with LSTM improves the performance of several Uni-modality models compared to LSTM performance.

Table 6.6. Comparison of the best LSTM Uni-modality models with MSE and BCE loss regarding regression task. Reduced Subsets: Red. Subsets.

Measure			MSE			ICC			Architecture	Learning rate
Model			Trivial	LSTM		Triv.	LSTM			
	Dataset	n-Class	-	FAD MSE	FAD BCE	-	FAD MSE	FAD BCE		
Subsets	PD	7	0.10	0.08	**0.08**	0	0.18	**0.20**	A(r)	10^{-5}
	HPD	4	0.11	0.08	**0.08**	0	0.27	**0.28**	C(r)	
	EPD	4	0.07	0.05	**0.05**	0	0.24	**0.27**	C(r)	
	TD	7	0.12	**0.09**	0.09	0	**0.14**	0.11	A(r)	10^{-6}
	HTD	4	0.41	0.13	**0.12**	0	0.13	**0.15**	C(r)	
	ETD	4	0.09	**0.08**	0.07	0	**0.13**	0.09	C(r)	
Red. Subsets	RPD	7	0.23	0.10	**0.08**	0	0.49	**0.56**	A(r)	10^{-4}
	RHPD	4	0.26	0.10	**0.09**	0	0.58	**0.62**	C(r)	
	REPD	4	0.26	0.11	**0.10**	0	0.50	**0.52**	C(r)	
	RTD	7	0.25	**0.13**	0.13	0	**0.08**	0.04	B(r)	10^{-6}
	RETD	4	0.25	**0.14**	0.14	0	**0.14**	0.09	D(r)	

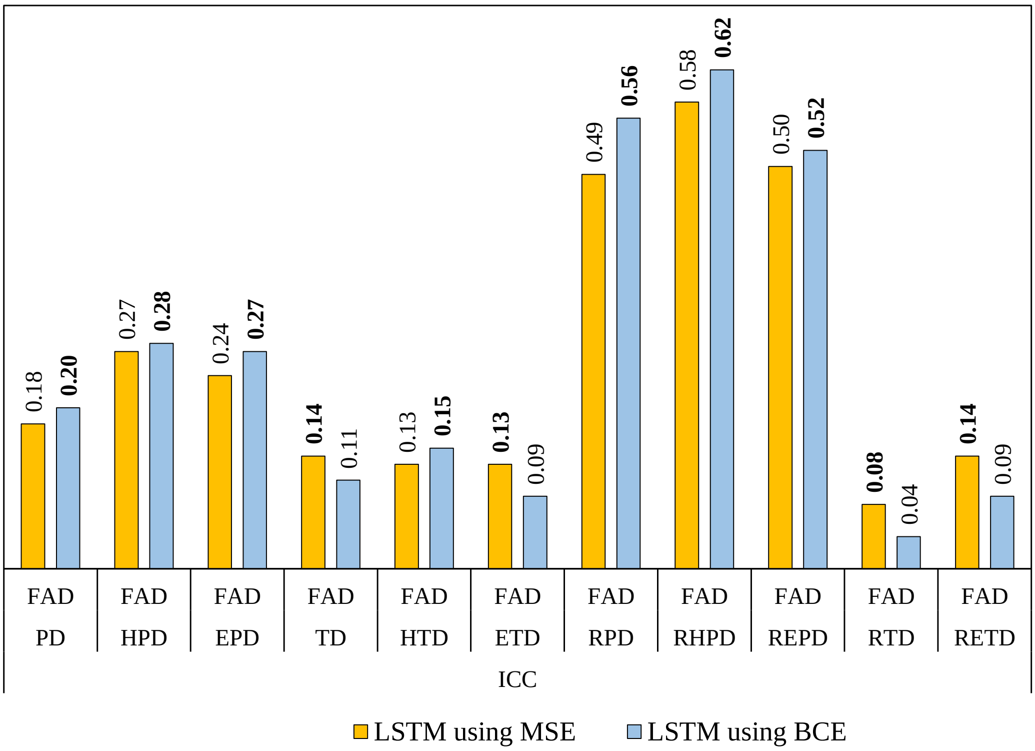


FIG. 6.6. Comparison of the best LSTM Uni-modality models with MSE and BCE loss regarding regression task with ICC measure.

6.3 Bi-modality Results

This Section shows the comparison between the Trivial and proposed methods (RF, LSTM, and LSTM-SW) when focusing on fusing two modalities from experimental data, including the FAD & EDA-D modalities or EMG-D & EDA-D modalities (see Section 6.3.1). More detailed results of the classification models that outperform the regression model when using two fused modalities of data were presented in Section 6.3.2. Finally, the discussion of the model results of two combined modalities was summarized in Section 6.3.3.

6.3.1 Classification vs Regression

The results of models that use FAD & EDA-D Bi-modality or EMG-D & EDA-D Bi-modality for continuous monitoring of pain intensity were provided in Section 6.3.1.1. The comparison between the best classification and regression models was shown in Section 6.3.1.2.

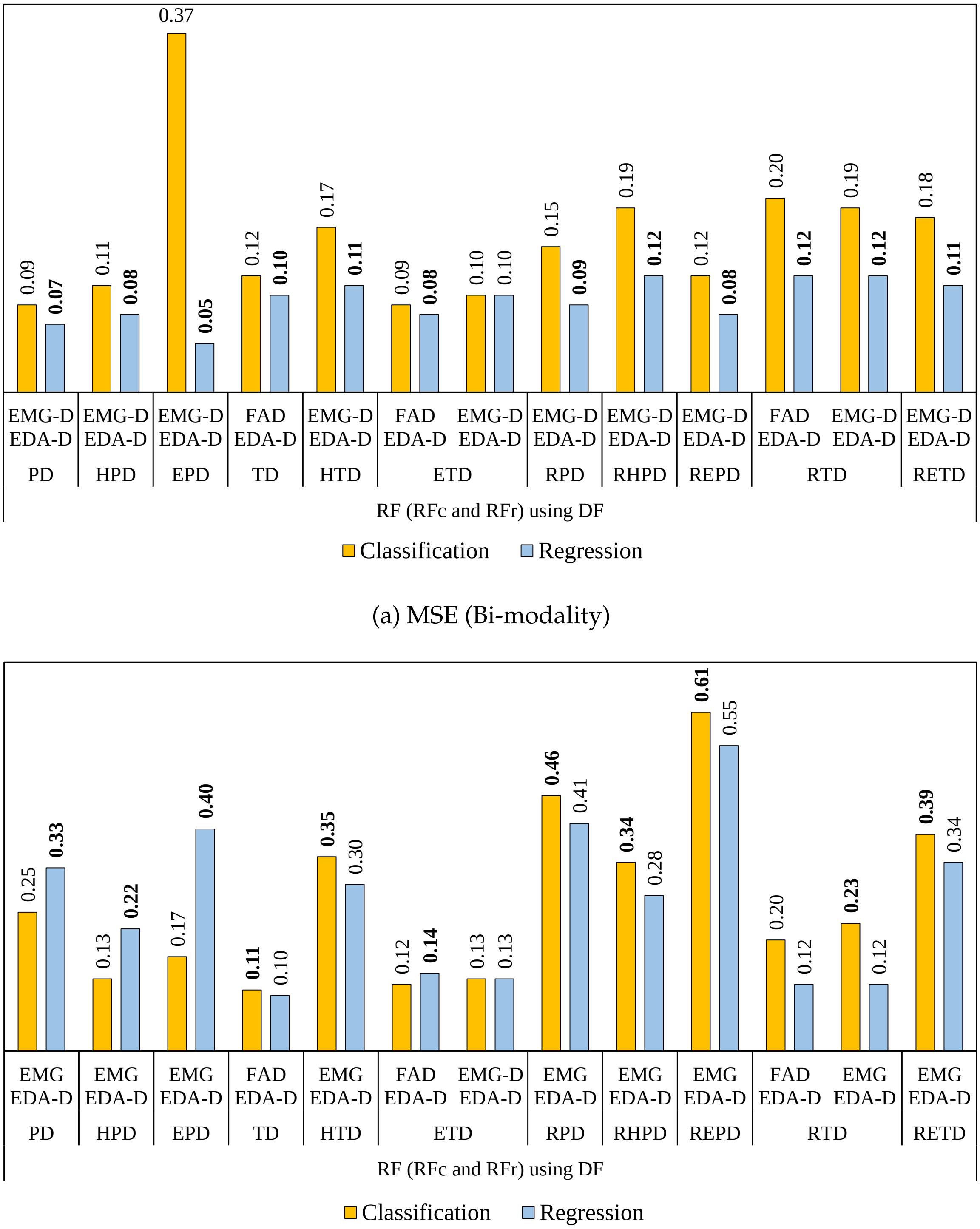
6.3.1.1 Modeling Methods

Table 6.7 and Fig. 6.7 show that most EMG-D & EDA-D Bi-modality models are better than FAD & EDA-D Bi-modality models regarding classification and regression. In Table 6.7, the cells with a light grey background indicate the best results regarding classification and regression tasks. The **bold** font indicates the best results of RF (RFc and RFr) models with ICC measure. FAD & EDA-D Bi-modality models performed the best when using Tonic Dataset [TD]; they got the MSE of 0.12, 0.10 and the ICC of 0.11, 0.10 (both perform almost similarly). Further, FAD & EDA-D Bi-modality models yielded similar results as EMG-D & EDA-D Bi-modality models regarding regression (RFr) with Electrical Tonic Dataset [ETD] and RTD.

Table 6.7. Comparison of the best Bi-modality models when applying RF (RFc and RFr) regarding classification and regression tasks with MSE and ICC measures. Triv.: Trivial, Red. Sub-sets.: Reduced Subsets. DF: Decision Fusion.

Measure			MSE					ICC				
Model			Triv.	RFc with DF		RFr with DF		Triv.	RFc with DF		RFr with DF	
Dataset		n-Class	-	FAD EDA-D	EMG-D EAD-D	FAD EDA-D	EMG-D EAD-D	-	FAD EDA-D	EMG-D EAD-D	FAD EDA-D	EMG-D EAD-D
Subsets	PD	7	0.10	0.10	0.09	0.07	**0.07**	0	0.16	0.25	0.30	**0.33**
	HPD	4	0.11	0.11	0.11	0.08	**0.08**	0	0.07	0.13	0.21	**0.22**
	EPD	4	0.07	0.45	0.37	0.05	**0.05**	0	0.16	0.17	0.37	**0.40**
	TD	7	0.12	0.12	0.14	0.10	0.12	0	**0.11**	0.12	0.10	0.08
	HTD	4	0.41	0.18	0.17	0.11	0.11	0	0.30	**0.35**	0.27	0.30
	ETD	4	0.09	0.09	0.10	**0.08**	0.10	0	0.12	0.13	**0.14**	0.13
Red. Subsets	RPD	7	0.23	0.16	0.15	0.10	0.09	0	0.40	**0.46**	0.37	0.41
	RHPD	4	0.26	0.20	0.19	0.12	0.12	0	0.31	**0.34**	0.27	0.28
	REPD	4	0.26	0.13	0.12	0.08	0.08	0	0.56	**0.61**	0.53	0.55
	RTD	7	0.25	0.20	0.19	0.12	0.12	0	0.20	**0.23**	0.12	0.12
	RETD	4	0.25	0.19	0.18	0.12	0.11	0	0.32	**0.39**	0.27	0.34

Fig. 6.7 shows the best results of the baseline Bi-modality models using Decision Fusion [DF] regarding classification (RFc) and regression (RFr). The EMG-D & EDA-D Bi-modality classification models got the highest ICC values on 6 datasets: Heat Tonic Dataset [HTD], Reduced Phasic Dataset [RPD], Reduced Heat Tonic Dataset [RHTD], Reduced Electrical Tonic Dataset [RETD], Reduced Tonic Dataset [RTD] and Reduced Electrical Tonic Dataset [RETD], they got ICC of 0.35, 0.46, 0.34, 0.61, 0.23, 0.39, respectively. However, the EMG-D & EDA-D Bi-modality regression models got the smallest recognition error on the same three datasets. They have the highest ICC values and smallest recognition error on phasic subsets: Phasic Dataset [PD], Heat Phasic Dataset [HPD], and Electrical Phasic Dataset [EPD]; they got an ICC of 0.33, 0.22, 0.40 and MSE of 0.07, 0.08, 0.05.



(a) MSE (Bi-modality)

(b) ICC (Bi-modality)

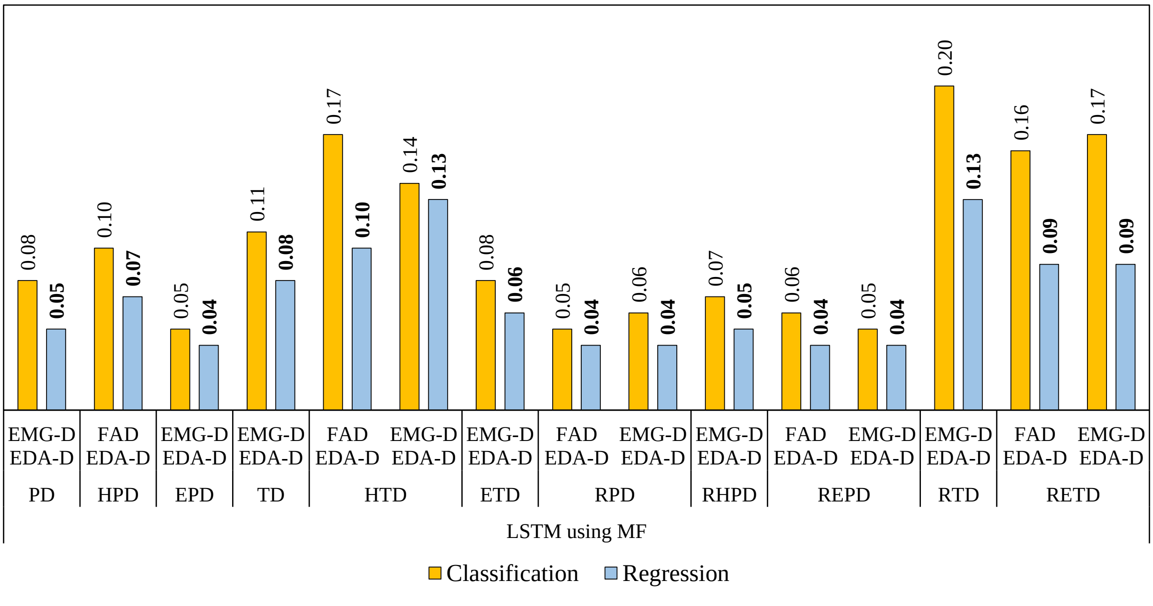
FIG. 6.7. Comparison of the best Bi-modality models when using RF (RFc and RFr) regarding classification and regression tasks with MSE and ICC measures. DF: Decision Fusion. The **bold** font indicates the best results.

Regarding classification and regression, the Bi-modality models using Model Fusion [MF] when applying LSTM were superior to those using Decision Fusion [DF], see Table 6.8 and Fig. 6.8.

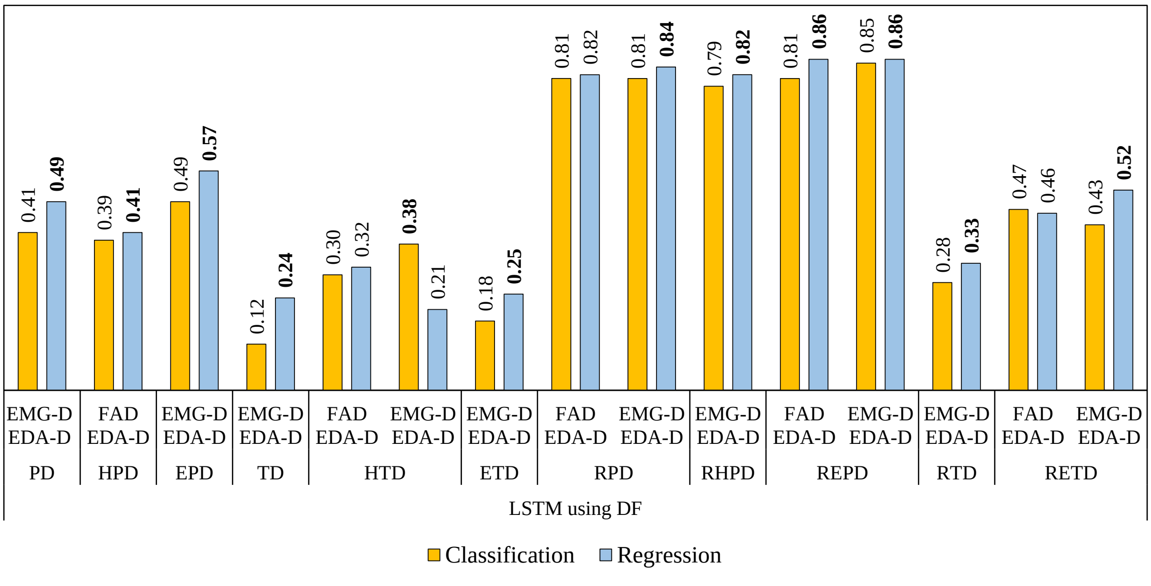
Table 6.8. Comparison of the best Bi-modality models when applying LSTM regarding classification and regression tasks with MSE and ICC measures. Trivial:Triv., Reduced Subsets: Red. Subsets. DF: Decision Fusion, MF: Model Fusion.

Measure	Model			Triv.	LSTM (Classification)				LSTM (Regression)				Loss		Learning rate
					DF	MF	DF	MF	DF	MF	DF	MF	CCE	BCE	
	Dataset		n-Class	-	FAD EDA-D	FAD EDA-D	EMG-D EAD-D	EMG-D EAD-D	FAD EDA-D	FAD EDA-D	EMG-D EAD-D	EMG-D EAD-D	Architecture of MF		
MSE	Subsets	PD	7	0.1	0.09	0.08	0.09	0.08	0.06	0.06	0.06	**0.05**	A-Bi(c)	A-Bi(r)	10^{-5}
		HPD	4	0.11	0.10	0.10	0.21	0.09	0.07	**0.07**	0.07	0.07	C-Bi(c)	C-Bi(r)	
		EPD	4	0.07	0.06	0.05	0.06	0.05	0.04	0.04	0.04	**0.04**	C-Bi(c)	C-Bi(r)	
		TD	7	0.12	0.12	0.12	0.12	0.11	0.08	0.08	0.08	**0.08**	A-Bi(c)	A-Bi(r)	10^{-6}
		HTD	4	0.41	0.16	0.17	0.15	**0.14**	0.11	0.10	0.11	0.13	C-Bi(c)	C-Bi(r)	
		ETD	4	0.09	0.08	0.08	0.09	0.08	0.07	0.07	0.07	**0.06**	C-Bi(c)	C-Bi(r)	
	Red. Subsets	RPD	7	0.23	0.07	0.05	0.08	0.06	0.05	0.04	0.05	**0.04**	A-Bi(c)	A-Bi(r)	10^{-4}
		RHPD	4	0.26	0.09	0.08	0.10	0.07	0.05	0.06	0.06	**0.05**	C-Bi(c)	C-Bi(r)	
		REPD	4	0.26	0.06	0.06	0.07	0.05	0.05	**0.04**	0.04	**0.04**	C-Bi(c)	C-Bi(r)	
		RTD	7	0.25	0.23	0.21	0.22	0.20	0.11	0.11	0.11	0.13	B-Bi(c)	B-Bi(r)	10^{-6}
		RETD	4	0.25	0.19	0.16	0.17	0.17	0.10	0.09	0.10	**0.09**	D-Bi(c)	D-Bi(r)	
ICC	Subsets	PD	7	0	0.17	0.36	0.20	0.41	0.34	0.47	0.37	**0.49**	A-Bi(c)	A-Bi(r)	10^{-5}
		HPD	4	0	0.25	0.39	0.20	0.37	0.31	**0.41**	0.28	0.39	C-Bi(c)	C-Bi(r)	
		EPD	4	0	0.26	0.45	0.27	0.49	0.42	0.56	0.44	**0.57**	C-Bi(c)	C-Bi(r)	
		TD	7	0	0.01	0.08	0.03	0.12	0.12	0.15	0.14	**0.24**	A-Bi(c)	A-Bi(r)	10^{-6}
		HTD	4	0	0.35	0.30	0.36	**0.38**	0.23	0.32	0.24	0.21	C-Bi(c)	C-Bi(r)	
		ETD	4	0	0.06	0.11	0.03	0.18	0.09	0.20	0.14	**0.25**	C-Bi(c)	C-Bi(r)	
	Red. Subsets	RPD	7	0	0.76	0.81	0.72	0.81	0.75	0.82	0.71	**0.84**	A-Bi(c)	A-Bi(r)	10^{-4}
		RHPD	4	0	0.71	0.76	0.69	0.79	0.77	0.79	0.73	**0.82**	C-Bi(c)	C-Bi(r)	
		REPD	4	0	0.81	0.81	0.79	0.85	0.79	**0.86**	0.80	**0.86**	C-Bi(c)	C-Bi(r)	
		RTD	7	0	0.16	0.26	0.2	0.28	0.14	0.25	0.20	**0.33**	B-Bi(c)	B-Bi(r)	10^{-6}
		RETD	4	0	0.34	0.47	0.41	0.43	0.33	0.46	0.42	**0.52**	D-Bi(c)	D-Bi(r)	

The architecture of MF in detail was described in Section 5.4.4. The cells with a light grey background indicate the best results regarding classification and regression tasks. The **bold** font indicates the best results of LSTM-SW models according to ICC measure. Fig. 6.8 shows that EMG-D & EDA-D Bi-modality regression models got the highest ICC values and smallest recognition error on all datasets except HPD and HTD. FAD & EDA-D Bi-modality regression model performed the best with HPD; it got an ICC of 0.41 and MSE of 0.07. EMG-D & EDA-D Bi-modality classification model got the highest ICC value on HTD (0.38). Most EMG-D & EDA-D Bi-modality models performed the best based on the highest ICC and smallest MSE values, see Table 6.8. FAD & EDA-D Bi-modality classification models were the best when using HPD, RPD, and RETD; they got an ICC of 0.39, 0.81, 0.47, and MSE of 0.10, 0.05, 0.16.



(a) MSE (Bi-modality)



(b) ICC (Bi-modality)

FIG. 6.8. Comparison of the best Bi-modality models when using LSTM regarding classification and regression tasks with MSE and ICC measures. MF: Model Fusion. The **bold** font indicates the best results.

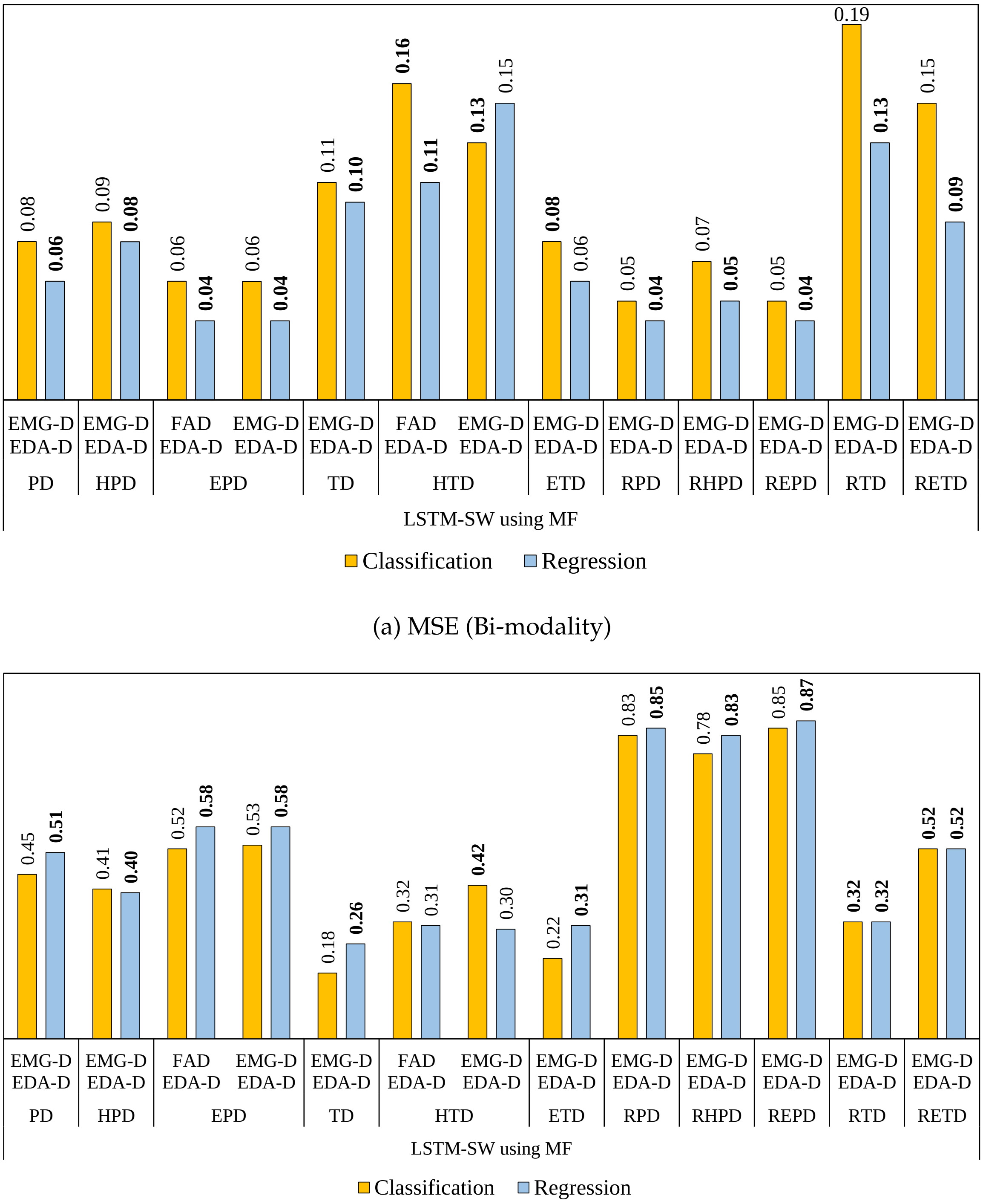
Further, FAD & EDA-D Bi-modality regression models were the best when using HPD, HTD, REPD, and RETD, they got the ICC of 0.41, 0.32, 0.86, 0.25 and the MSE of 0.07, 0.10, 0.04, 0.11.

The Bi-modality models performed better regarding classification than those using DF; see Table 6.9 and Fig. 6.9. The architecture of MF in detail was described in Section 5.4.4. The cells with a light grey background indicate the best results regarding classification and regression tasks. The **bold** font indicates the best results of LSTM-SW models.

Table 6.9. Comparison of the best Bi-modality models when applying LSTM-SW regarding classification and regression tasks with MSE and ICC measures. Trivial:Triv., Reduced Subsets: Red. Subsets. DF: Decision Fusion, MF: Model Fusion.

Measure	Model			Triv.	LSTM-SW (Classification)				LSTM-SW (Regression)				Loss		Learning rate
					DF	MF	DF	MF	DF	MF	DF	MF	CCE	BCE	
	Dataset		n-Class	-	FAD EDA-D	FAD EDA-D	EMG-D EAD-D	EMG-D EAD-D	FAD EDA-D	FAD EDA-D	EMG-D EAD-D	EMG-D EAD-D	Architecture of MF		
MSE	Subsets	PD	7	0.10	0.09	0.08	0.09	0.08	0.06	0.06	0.07	**0.06**	A-Bi(c)	A-Bi(r)	10^{-5}
		HPD	4	0.11	0.10	0.10	0.10	0.09	0.08	0.08	0.08	0.08	C-Bi(c)	C-Bi(r)	
		EPD	4	0.07	0.06	0.06	0.05	0.06	0.04	**0.04**	0.04	**0.04**	C-Bi(c)	C-Bi(r)	
		TD	7	0.12	0.12	0.11	0.11	0.11	0.09	0.09	0.09	0.10	A-Bi(c)	A-Bi(r)	10^{-6}
		HTD	4	0.41	0.16	0.16	0.14	0.13	0.12	0.11	0.10	0.15	C-Bi(c)	C-Bi(r)	
		ETD	4	0.09	0.10	0.08	0.10	0.08	0.07	0.07	0.07	**0.06**	C-Bi(c)	C-Bi(r)	
	Red. Subsets	RPD	7	0.23	0.07	0.05	0.07	0.05	0.04	0.04	0.05	**0.04**	A-Bi(c)	A-Bi(r)	10^{-4}
		RHPD	4	0.26	0.09	0.08	0.09	0.07	0.05	0.06	0.05	**0.05**	C-Bi(c)	C-Bi(r)	
		REPD	4	0.26	0.06	0.06	0.06	0.05	0.05	0.05	0.04	**0.04**	C-Bi(c)	C-Bi(r)	
		RTD	7	0.25	0.22	0.20	0.21	0.19	0.11	0.11	0.11	0.13	B-Bi(c)	B-Bi(r)	10^{-6}
		RETD	4	0.25	0.2	0.17	0.18	0.15	0.10	0.09	0.10	**0.09**	D-Bi(c)	D-Bi(r)	
ICC	Subsets	PD	7	0	0.24	0.44	0.30	0.45	0.37	0.49	0.25	**0.51**	A-Bi(c)	A-Bi(r)	10^{-5}
		HPD	4	0	0.3	0.41	0.28	**0.41**	0.33	0.39	0.32	0.40	C-Bi(c)	C-Bi(r)	
		EPD	4	0	0.33	0.52	0.41	0.53	0.47	**0.58**	0.49	**0.58**	C-Bi(c)	C-Bi(r)	
		TD	7	0	0.07	0.15	0.12	0.18	0.11	0.19	0.14	**0.26**	A-Bi(c)	A-Bi(r)	10^{-6}
		HTD	4	0	0.29	0.32	0.39	**0.42**	0.24	0.31	0.30	0.23	C-Bi(c)	C-Bi(r)	
		ETD	4	0	0.01	0.17	0.01	0.22	0.17	0.28	0.21	**0.31**	C-Bi(c)	C-Bi(r)	
	Red. Subsets	RPD	7	0	0.76	0.82	0.77	0.83	0.77	0.80	0.75	**0.85**	A-Bi(c)	A-Bi(r)	10^{-4}
		RHPD	4	0	0.71	0.77	0.72	0.78	0.76	0.79	0.74	**0.83**	C-Bi(c)	C-Bi(r)	
		REPD	4	0	0.8	0.82	0.83	0.85	0.79	0.84	0.80	**0.87**	C-Bi(c)	C-Bi(r)	
		RTD	7	0	0.21	0.24	0.24	**0.32**	0.16	0.25	0.21	**0.32**	B-Bi(c)	B-Bi(r)	10^{-6}
		RETD	4	0	0.31	0.43	0.41	**0.52**	0.35	0.50	0.40	**0.52**	D-Bi(c)	D-Bi(r)	

All EMG-D & EDA-D Bi-modality models performed the best regarding regression except with EPD and HTD, see Table 6.9 and Fig. 6.9. FAD & EDA-D Bi-modality models or EMG-D & EDA-D Bi-modality models performed similarly with EPD; they got an ICC of 0.58 and MSE of 0.04. Additionally, FAD & EDA-D Bi-modality models performed similarly to EMG-D & EDA-D Bi-modality models when applying LSTM-SW using Decision fusion (DF); they got an ICC of 0.31, 0.30, and MSE of 0.11, 0.10.



(a) MSE (Bi-modality)

(b) ICC (Bi-modality)

FIG. 6.9. Comparison of the best Bi-modality models when using LSTM-SW regarding classification and regression tasks with MSE and ICC measures. MF: Model Fusion. The **bold** font indicates the best results.

Figure 6.9 shows that Bi-modality models using MF when applying LSTM-SW got the highest ICC values with HPD, RTD, and RETD in both tasks. The remaining EMG-D & EDA-D Bi-modality regression models performed the best except with HTD; the Bi-modality classification model got the highest ICC (0.42).

6.3.1.2 Comparison of Modeling Methods

This section provides the comparison between the baseline method (RFc and RFr) and LSTM and LSTM-SW methods when using data from single and two modalities with all proposed subsets from the experimental data, Bi-modality models using Model Fusion (MF) outperformed those using Decision Fusion [DF]. Table 6.10 shows that most EMG-D & EDA-D Bi-modality models were better than FAD & EDA-D Bi-modality models when using MF. Further, most Bi-modality classification models, when using LSTM-SW, performed the best except RHPD, REPD, and RTD; they performed similarly to RFc and LSTM models. Further, the Bi-modality regression models when using LSTM-SW performed the best except with HPD and RTD; they performed similarly to LSTM. The LSTM and LSTM-SW regression models performed almost similarly to HPD.

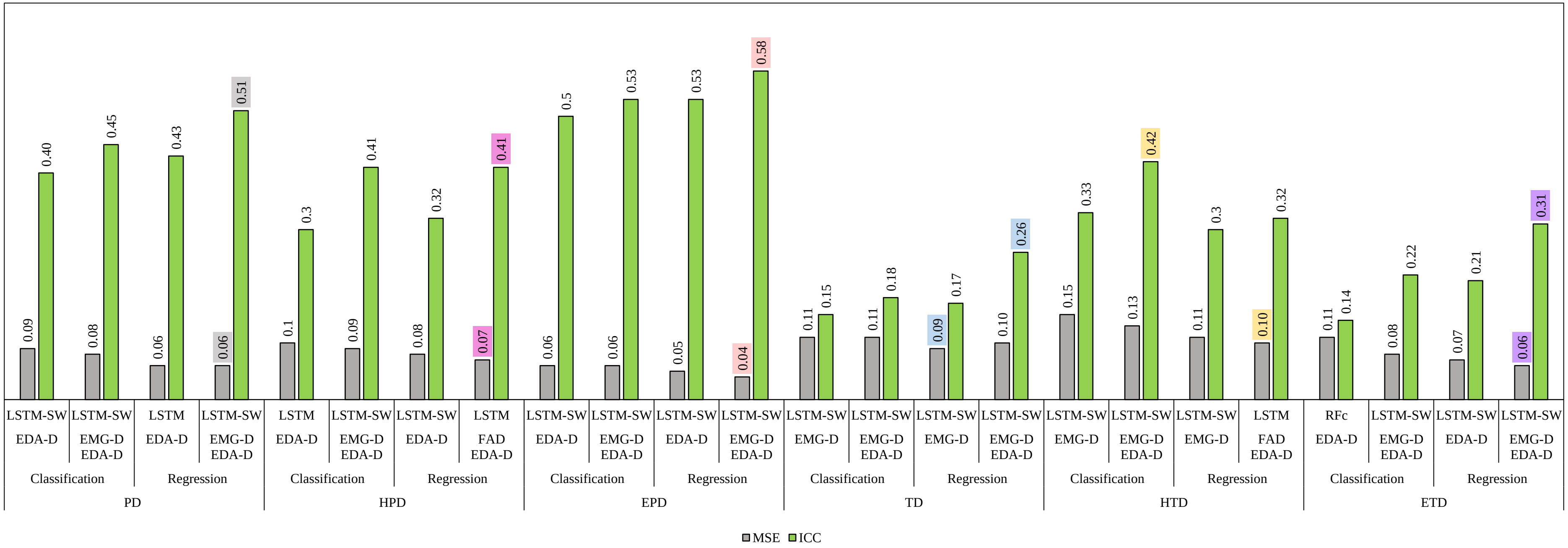
Fig. 6.10 summarizes the best results of both Uni-modality and Bi-modality models with each dataset. Only one Bi-modality model did not improve the Uni-modality model with REPD. The regression models outperformed classification models except with HTD. The highest ICC with HTD was 0.42 when using EMG-D and EDA-D Bi-modality classification model using MF when applying LSTM-SW. Additionally, the performance was improved when using the reduced datasets. In regards to ICC values, the performances of the best Bi-modality models improved: from 0.51 with PD to 0.85 with RPD, from 0.41 with HPD to 0.83 with RHPD, from 0.58 with EPD to 0.85 with REPD, from 0.26 with TD to 0.33 with RTD, and from 0.31 with ETD to 0.52 with RETD.

6.3.2 Classification

This section introduces more results of comparing Uni-modality and Bi-modality classification models that use MF when applying RFc, LSTM, and LSTM-SW on the Heat Tonic Dataset [HTD] due to their superior performance to regression results. Additional classification measures were Accuracy, Micro avg. precision, Micro avg. recall, and Micro avg. F1-Score. The architecture of LSTM and LSTM-SW is C-Bi(c) learning rate of with 10^{-6}. Table 6.11 and Fig. 6.11 shows how the LSTM and LSTM-SW models with HTD successfully predict discrete pain intensity level in sequences compared to Trivial and RFc. The EDA-D Uni- and Bi-modality models performed better than those Uni-modality models with FAD and EMG-D modalities.

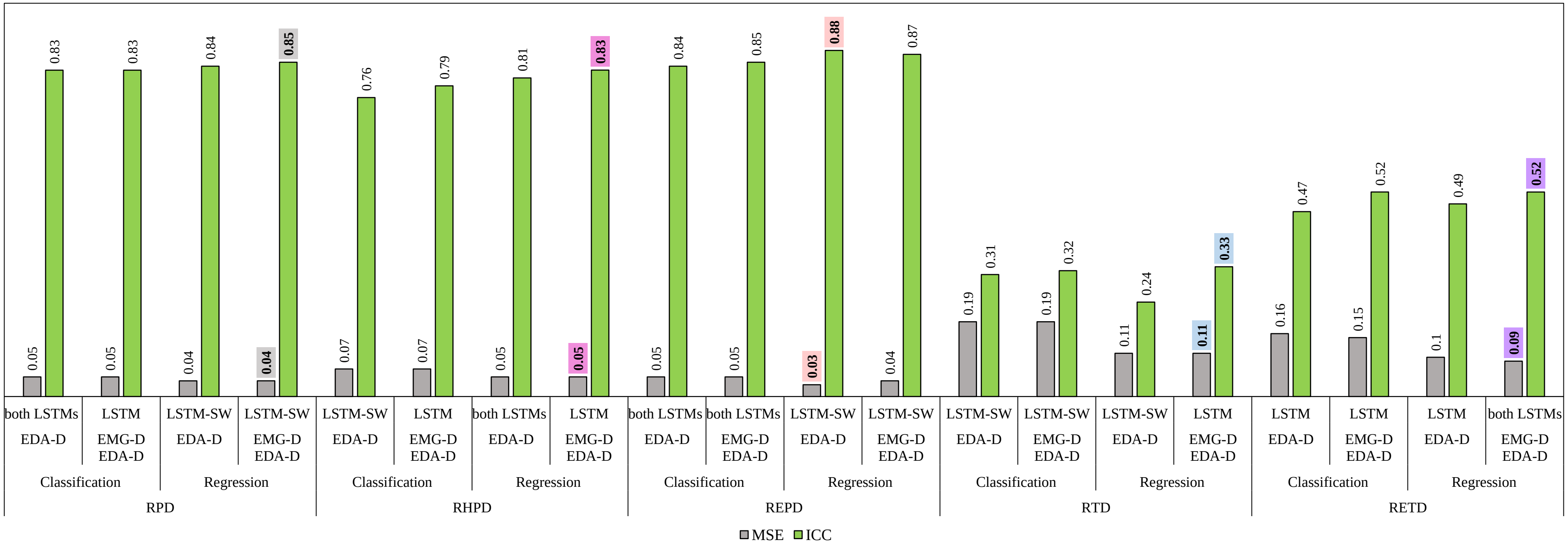
Table 6.10. Comparison of the best Bi-modality models using MF regarding classification and regression tasks with MSE and ICC measures. Meas.: Measure, MF: Model Fusion. The cells with numbers only indicate models with EMG & EDA-D results, the cells with a light grey background indicate the best results regarding classification and regression tasks. The **bold** font indicates the best results.

Meas.	Task		Classification			Regression		
	Dataset		RFc	LSTM	LSTM-SW	RFr	LSTM	LSTM-SW
MSE	Subsets	PD	0.09	0.08	**0.08**	0.07	**0.05**	**0.06**
		HPD	0.11	0.09 FAD EDA-D	**0.09**	0.08	**0.07** **FAD** **EDA-D**	0.08
		EPD	0.37	**0.05**	**0.06**	0.05	0.04	**0.04**
		TD	0.12 FAD EDA-D	0.11	**0.11**	0.10 FAD EDA-D	**0.08**	**0.10**
		HTD	0.17	0.14	**0.13**	0.11	**0.10** **FAD** **EDA-D**	0.11 FAD EDA-D
		ETD	0.10	0.08	**0.08**	0.08 FAD EDA-D	0.06	**0.06**
	Red. Subsets	RPD	0.15	0.05 FAD EDA-D	**0.05**	0.09	**0.04**	**0.04**
		RHPD	0.19	**0.07**	0.07	0.12	**0.05**	**0.05**
		REPD	0.12	**0.05**	**0.05**	0.08	**0.04**	**0.04**
		RTD	**0.19**	0.20	**0.19**	0.12	**0.11**	0.13
		RETD	0.18	0.16 FAD EDA-D	**0.15**	0.11	0.09	**0.09**
ICC	Subsets	PD	0.25	0.41	**0.45**	0.33	0.49	**0.51**
		HPD	0.16	0.39 FAD EDA-D	**0.41**	0.22	**0.41** **FAD** **EDA-D**	0.40
		EPD	0.17	0.49	**0.53**	0.40	0.57	**0.58**
		TD	0.11 FAD EDA-D	0.12	**0.18**	0.10 FAD EDA-D	0.24	**0.26**
		HTD	0.35	0.38	**0.42**	0.30	**0.32** **FAD** **EDA-D**	0.31 FAD EDA-D
		ETD	0.13	0.18	**0.22**	0.14 FAD EDA-D	0.25	**0.31**
	Red.Subsets	RPD	0.46	0.81	**0.83**	0.41	**0.84**	**0.85**
		RHPD	0.34	**0.79**	0.78	0.28	**0.82**	**0.83**
		REPD	0.61	**0.85**	**0.85**	0.55	**0.86**	**0.87**
		RTD	**0.32**	0.28	**0.32**	0.12	**0.33**	0.32
		RETD	0.39	0.47 FAD EDA-D	**0.52**	0.34	0.52	**0.52**



(a) Subsets

FIG. 6.10. Comparison of the best Uni- and Bi-modality models regarding classification and regression tasks with MSE and ICC measures. The Bi-modality models use MF. MF: Model Fusion. Both LSTMs: LSTM and LSTM-SW. The **bold** font with colored background indicates the best results.



(b) Reduced Subsets

FIG. 6.10. Comparison of the best Uni- and Bi-modality models regarding classification and regression tasks with MSE and ICC measures. The Bi-modality models use MF. MF: Model Fusion. Both LSTMs: LSTM and LSTM-SW. The **bold** font with colored background indicates the best results.

The best results were obtained from EMG-D & EDA-D Bi-modality model when applying LSTM-SW in terms of Accuracy (49.8%), Micro avg. precision (48.7%) and Micro avg. recall (63.3%). However, the EDA-D Uni-modality model, when applying LSTM-SW, achieved the best Micro avg. recall with about 100% when 47.7%, 47.7%, 62.5% were the Accuracy results, Micro avg. Precision and Micro avg. Recall (not the best), respectively.

Table 6.11. Comparison of the best Uni- and Bi-modality models with HTD regarding classification task. The Bi-modality models use MF. MF: Model Fusion. Triv.: Trivial. The **bold** font indicates the best models' results. * $p < 0.05$ when using paired t-test between RFc and LSTMs (LSTM and LSTM-SW).

Meas.	HTD									
	Model	Triv.	RFc	LSTM	LSTM-SW	Meas.	Triv.	RFc	LSTM	LSTM-SW
Accuracy %	FAD & EDA-D (Bi-modality)	20	-	44.6	45.6	Micro avg. recall %	0	-	87.9	97.4
	EMG-D & EDA-D (Bi-modality)	20	-	47.4	**49.8**		0	-	92.9	97*
	FAD (Uni-modality)	20	29.1	32.81	33.7		0	49.4	92.42*	**100***
	EMG-D (Uni-modality)	20	35.2	39.3	39.9		0	65.5	99.6*	**100***
	EDA-D(Uni-modality)	20	41.0	48.4	47.7		0	71.0	94.6*	**100***
Micro avg. precision %	FAD & EDA-D (Bi-modality)	0	-	44.0	47.2	Micro avg. F1-score %	0	-	57.8	60.3
	EMG-D & EDA-D (Bi-modality)	0	-	47.2	**48.7**		0	-	60.6	**63.3**
	FAD (Uni-modality)	0	31.1	33.28	33.7		0	37.6	46.63	48.2*
	EMG-D (Uni-modality)	0	38.0	39.3	39.9		0	46.0	55.6*	56.3
	EDA-D (Uni-modality)	0	42.7	48.2	47.7		0	52.9	62.3*	62.5*

6.3.3 Discussion

This section summarizes the results of comparing Uni-modality models and Bi-modality models that use MF when applying RFc, LSTM, and LSTM-SW on 11 datasets from the experimental data. In line with results from Uni-modality models (see Section 6.2): (1) Regression models were superior to classification models except when using HTD, and the reduction strategy on Subsets datasets, which were obtained Reduced Subsets, improved the performance (see Table 6.10 and Fig. 6.10), (2) Uni- and Bi-modality models when applying LSTM and LSTM-SW were significantly better than the baseline model (RFc and RFr), and (3) LSTM using sample weighting method (LSTM-SW) improved the performance of several LSTM models. Further, the most results of EMG-D & EDA-D Bi-modality models using MF were superior to FAD & EDA-D Bi-modality models and Uni-modality models with FAD, EMG-D, and EDA-D modalities. However, only one Uni-modality model performed the best: the LSTM-SW model with the EDA-D modality from REPD.

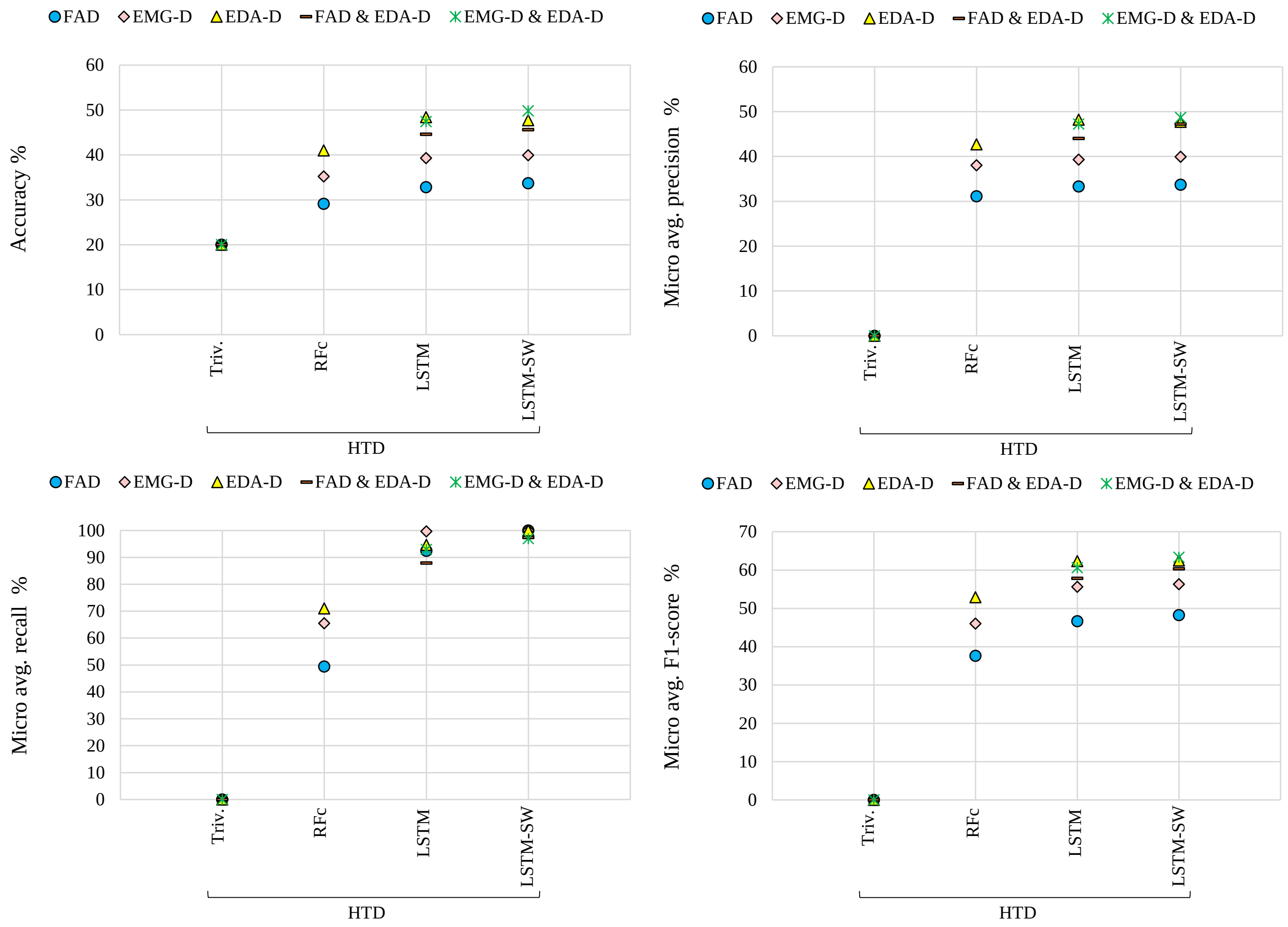


FIG. 6.11. Comparison of Uni-modality models when applying Trivial, RFc, LSTM, and LSTM-SW with HTD regarding classification task. Triv.: Trivial. The Bi-modality models use MF. MF: Model Fusion.

6.4 Multi-modality Results

This section shows the comparison between the Trivial and proposed methods (RF, LSTM, and LSTM-SW) when focusing on fusing all modalities from the experimental data, including the FAD, Audio-D, ECG-D, EMG-D, and EDA-D modalities (see Section 6.4.1). More detailed results of the classification model that outperform the regression model when using the five fused modalities data (Multi-modality) were presented in Section 6.4.2. Finally, the discussion of Multi-modality results was summarized in Section 6.4.3.

6.4.1 Classification vs Regression

The results of Multi-modality models for continuous monitoring of pain intensity were provided in Section 6.4.1.1. The comparison between the best classification and regression models in combined modalities was shown in Section 6.4.1.2.

6.4.1.1 Modeling Methods

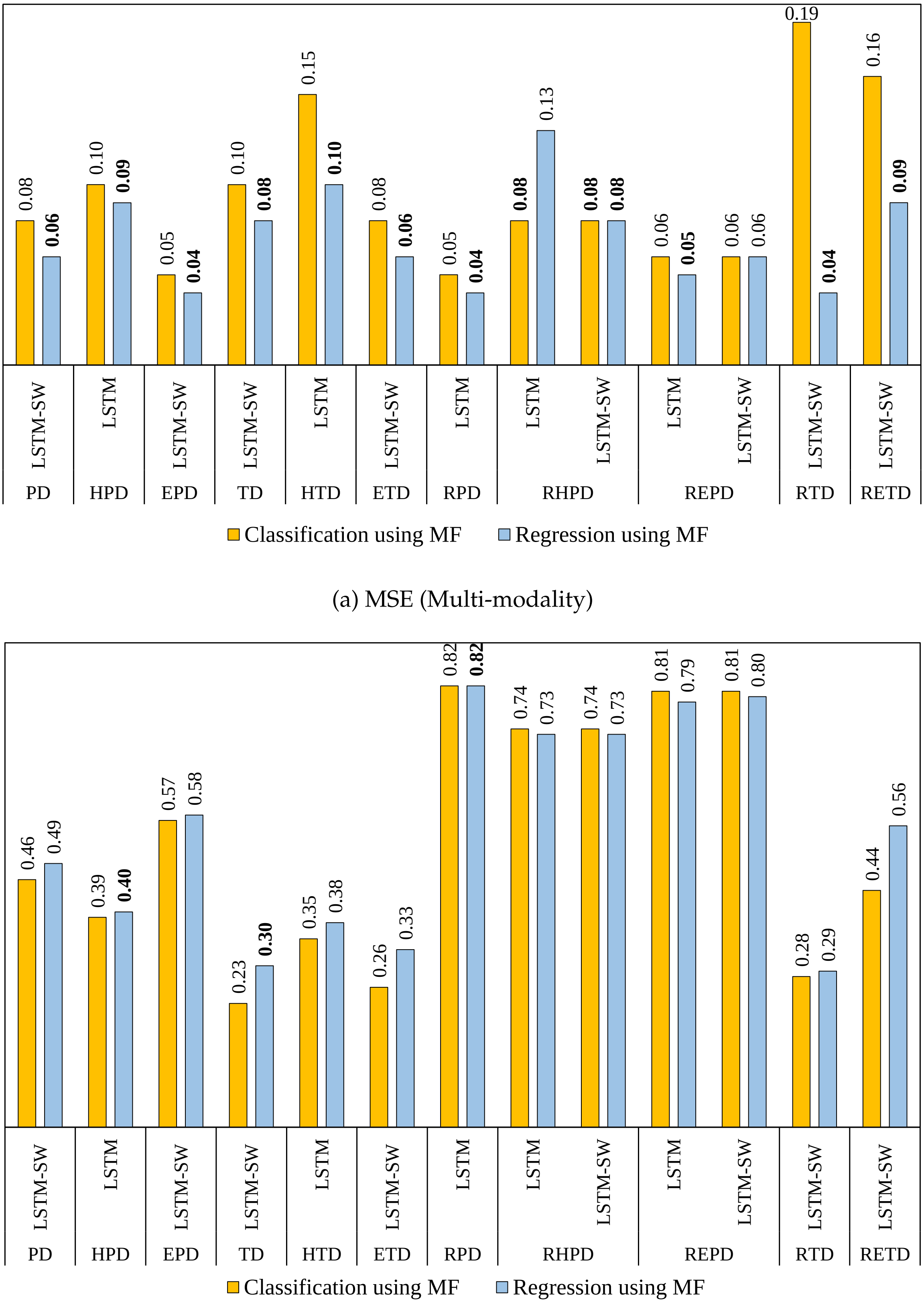
Table 6.12 shows that all Multi-modality models using MF were better than those using decision fusion [DF]. The architecture of MF in detail is described in Section 5.4.5. Fig. 6.12 shows that the performances of the most Multi-modality regression models using Model Fusion [MF] when applying LSTM were the best except with Reduced Heat Phasic Dataset [RHPD] and Reduced Electrical Phasic Dataset [REPD]. Multi-modality classification models, when applying LSTM and LSTM-SW, performed the best based on the highest ICC (0.74, 0.81) and the MSE of 0.08, 0.06, respectively. However, the Multi-modality regression model when applying LSTM with REPD got an ICC of 0.79 and the smallest MSE (0.05). Further, both classification and regression models performed similarly with the Reduced Phasic Dataset [RPD]; they got the highest ICC (0.82).

6.4.1.2 Comparison of Modeling Methods

This section provides the comparison between the best Uni-, Bi-, and Multi-modality models when applying LSTM and LSTM-SW on the 11 proposed datasets from the experimental data. The best Uni-modality models use the EDA-D modality, and the best Bi-modality models use both EMG-D and EDA-D modalities. The Bi- and Multi-modality models use MF. Table 6.13 and 6.14 and Fig. 6.13 show that only one Uni-modality model performed better than Bi- and Multi-modality models, which was the EDA-D Uni-modality model when applying LSTM-SW with REPD.

Table 6.12. Comparison of the best Bi-modality models using MF regarding classification and regression tasks with MSE and ICC measures. Meas.: Measure, MF: Model Fusion. The cells with a light grey background indicate the best results regarding classification and regression tasks. The **bold** font indicates the best results.

Meas.	Task				Classification					Regression					loss		Learning rate
	Model		n-Class	Triv.	RFc	LSTM		LSTM-SW		RFr	LSTM		LSTM-SW		CCE	BCE	
	Dataset			-	DF	DF	MF	DF	MF	DF	DF	MF	DF	MF	Architecture of MF		
MSE	Subsets	PD	7	0.10	0.10	0.10	0.08	0.10	0.08	0.07	0.07	**0.06**	0.07	0.06	A-Mu(c)	A-Mu(r)	10^{-5}
		HPD	4	0.11	0.11	0.11	0.10	0.10	0.11	0.08	0.08	0.08	0.08	**0.09**	C-Mu(c)	C-Mu(r)	
		EPD	4	0.07	0.26	0.07	0.05	0.06	0.05	0.06	0.05	**0.04**	0.05	0.05	C-Mu(c)	C-Mu(r)	
		TD	7	0.12	0.12	0.12	0.11	0.12	0.10	0.10	0.08	0.09	0.09	**0.08**	A-Mu(c)	A-Mu(r)	10^{-6}
		HTD	4	0.41	0.20	0.17	0.15	0.17	0.14	0.11	0.11	**0.10**	0.12	0.11	C-Mu(c)	C-Mu(r)	
		ETD	4	0.09	0.09	0.09	0.08	0.09	0.08	0.08	0.07	0.06	0.07	**0.06**	C-Mu(c)	C-Mu(r)	
	Red. Subsets	RPD	7	0.23	0.19	0.11	0.05	0.09	0.05	0.1	0.06	**0.04**	0.06	0.06	A-Mu(c)	A-Mu(r)	10^{-4}
		RHPD	4	0.26	0.22	0.1	**0.08**	0.09	**0.08**	0.12	0.06	0.13	0.06	0.08	C-Mu(c)	C-Mu(r)	
		REPD	4	0.26	0.16	0.09	**0.06**	0.08	**0.06**	0.1	0.06	0.05	0.06	0.06	C-Mu(c)	C-Mu(r)	
		RTD	7	0.25	0.21	0.23	0.20	0.22	0.19	0.12	0.12	**0.04**	0.12	0.12	B-Mu(c)	B-Mu(r)	10^{-6}
		RETD	4	0.25	0.2	0.2	0.17	0.19	0.16	0.12	0.11	0.09	0.12	**0.09**	D-Mu(c)	D-Mu(r)	
ICC	Subsets	PD	7	0	0.05	0.04	0.45	0.12	0.46	0.19	0.23	**0.49**	0.18	0.48	A-Mu(c)	A-Mu(r)	10^{-5}
		HPD	4	0	0.03	0.1	0.39	0.19	0.38	0.16	0.22	0.38	0.25	**0.40**	C-Mu(c)	C-Mu(r)	
		EPD	4	0	0.15	0.11	0.49	0.22	0.57	0.24	0.29	**0.58**	0.33	**0.58**	C-Mu(c)	C-Mu(r)	
		TD	7	0	0.05	0	0.20	0.02	0.23	0.08	0.09	0.23	0.10	**0.30**	A-Mu(c)	A-Mu(r)	10^{-6}
		HTD	4	0	0.27	0.31	0.35	0.32	0.33	0.22	0.16	**0.38**	0.18	0.35	C-Mu(c)	C-Mu(r)	
		ETD	4	0	0.05	0.01	0.20	0	0.26	0.1	0.09	0.31	0.12	**0.33**	C-Mu(c)	C-Mu(r)	
	Red. Subsets	RPD	7	0	0.23	0.61	0.82	0.68	0.81	0.28	0.66	**0.82**	0.66	0.78	A-Mu(c)	A-Mu(r)	10^{-4}
		RHPD	4	0	0.23	0.67	**0.74**	0.71	**0.74**	0.23	0.71	0.73	0.69	0.73	C-Mu(c)	C-Mu(r)	
		REPD	4	0	0.46	0.73	**0.81**	0.77	**0.81**	0.38	0.71	0.79	0.71	0.80	C-Mu(c)	C-Mu(r)	
		RTD	7	0	0.19	0.14	0.27	0.18	0.28	0.08	0.11	**0.29**	0.12	0.26	B-Mu(c)	B-Mu(r)	10^{-6}
		RETD	4	0	0.28	0.27	0.42	0.30	0.44	0.21	0.24	0.50	0.24	**0.56**	D-Mu(c)	D-Mu(r)	



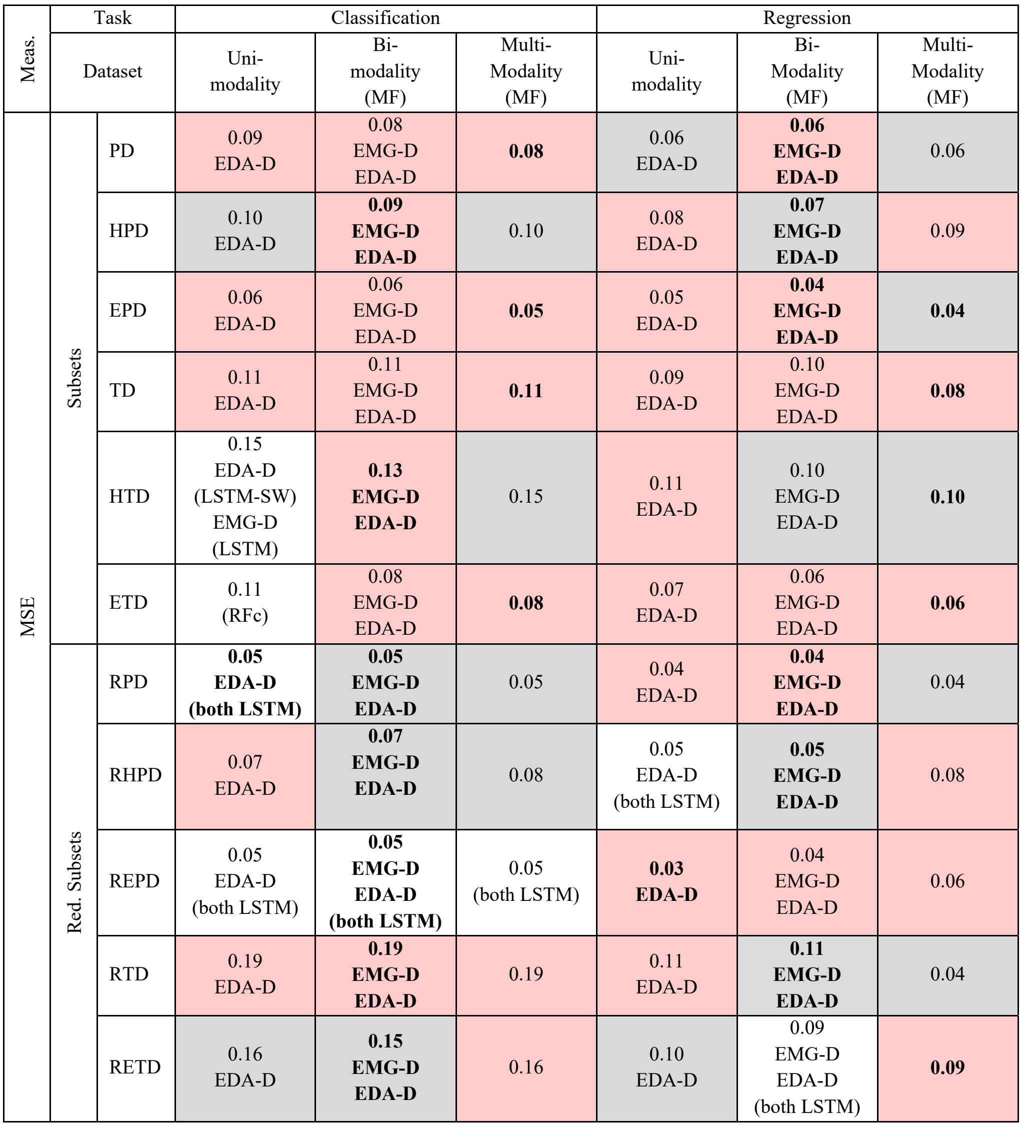
(a) MSE (Multi-modality)

(b) ICC (Multi-modality)

FIG. 6.12. Comparison of the best Multi-modality models when applying LSTM regarding classification and regression tasks with MSE and ICC measures. The Multi-modality models use MF. MF: Model Fusion. The **bold** font indicates the best results.

The remaining models, when applying EMG-D & EDA-D Bi-modality model that uses MF, were the best compared to Multi-modality models that use MF, except with Tonic Dataset [TD], Electrical Tonic Dataset [ETD], and Reduced Electrical Tonic Dataset [RETD], Multi-modality models are the best; they got MSE 0.08, 0.06, 0.09 and the ICC of 0.30, 0.33, 0.56, respectively. Further, Bi-and Multi-modality models that use MF when applying LSTM and LSTM-SW with EPD dataset performed similarly and got a highest ICC (0.58) and smallest MSE (0.04).

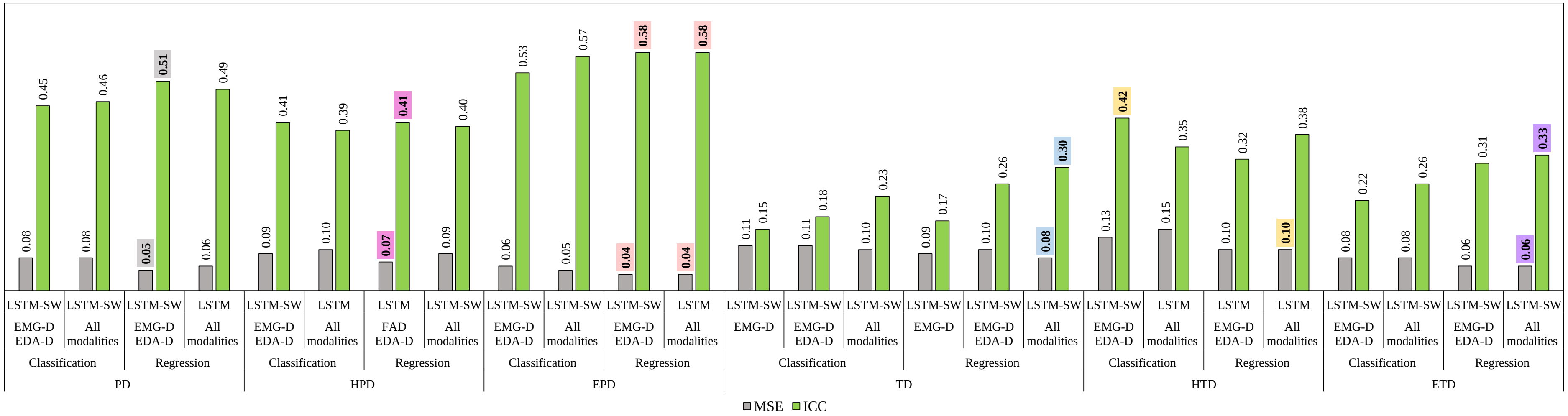
Table 6.13. Comparison of the best Uni-, Bi-modality, and Multi-modality models using MF regarding classification and regression tasks with MSE measure [186]. Meas.: Measure, MF: Model Fusion. The cells with lightgrey background indicate the models using LSTM and Cells with pink background indicate the models using LSTM-SW. The **bold** font indicates the best results.



Meas.	Task		Classification			Regression		
	Dataset		Uni-modality	Bi-modality (MF)	Multi-Modality (MF)	Uni-modality	Bi-Modality (MF)	Multi-Modality (MF)
MSE	Subsets	PD	0.09 EDA-D	0.08 EMG-D EDA-D	**0.08**	0.06 EDA-D	**0.06** **EMG-D** **EDA-D**	0.06
		HPD	0.10 EDA-D	**0.09** **EMG-D** **EDA-D**	0.10	0.08 EDA-D	**0.07** **EMG-D** **EDA-D**	0.09
		EPD	0.06 EDA-D	0.06 EMG-D EDA-D	**0.05**	0.05 EDA-D	**0.04** **EMG-D** **EDA-D**	**0.04**
		TD	0.11 EDA-D	0.11 EMG-D EDA-D	**0.11**	0.09 EDA-D	0.10 EMG-D EDA-D	**0.08**
		HTD	0.15 EDA-D (LSTM-SW) EMG-D (LSTM)	**0.13** **EMG-D** **EDA-D**	0.15	0.11 EDA-D	0.10 EMG-D EDA-D	**0.10**
		ETD	0.11 (RFc)	0.08 EMG-D EDA-D	**0.08**	0.07 EDA-D	0.06 EMG-D EDA-D	**0.06**
	Red. Subsets	RPD	**0.05** **EDA-D** **(both LSTM)**	**0.05** **EMG-D** **EDA-D**	0.05	0.04 EDA-D	**0.04** **EMG-D** **EDA-D**	0.04
		RHPD	0.07 EDA-D	**0.07** **EMG-D** **EDA-D**	0.08	0.05 EDA-D (both LSTM)	**0.05** **EMG-D** **EDA-D**	0.08
		REPD	0.05 EDA-D (both LSTM)	**0.05** **EMG-D** **EDA-D** **(both LSTM)**	0.05 (both LSTM)	**0.03** **EDA-D**	0.04 EMG-D EDA-D	0.06
		RTD	0.19 EDA-D	**0.19** **EMG-D** **EDA-D**	0.19	0.11 EDA-D	**0.11** **EMG-D** **EDA-D**	0.04
		RETD	0.16 EDA-D	**0.15** **EMG-D** **EDA-D**	0.16	0.10 EDA-D	0.09 EMG-D EDA-D (both LSTM)	**0.09**

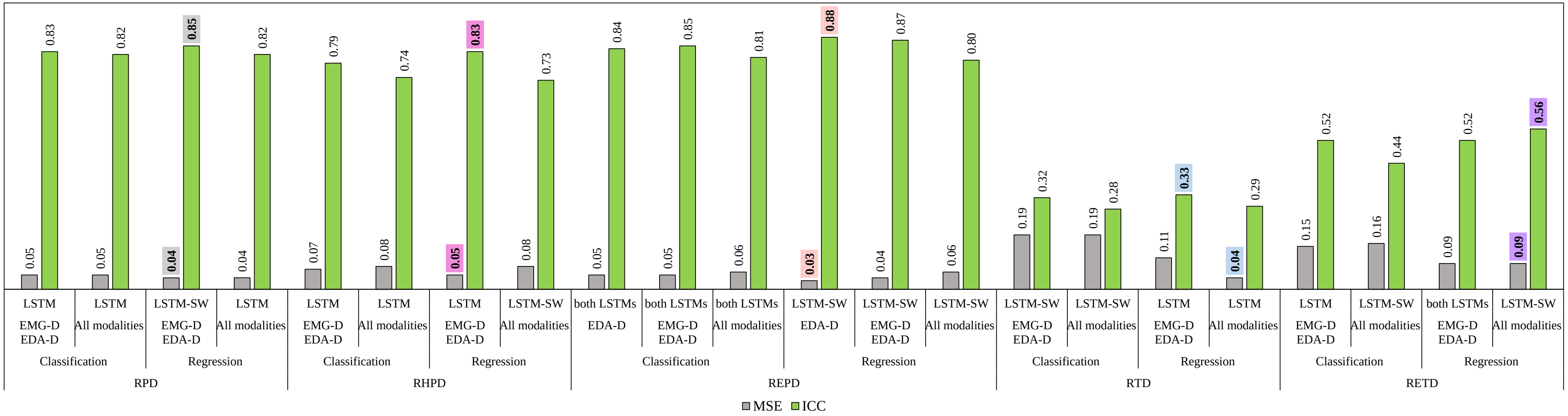
Table 6.14. Comparison of the best Uni-, Bi-modality, and Multi-modality models using MF regarding classification and regression tasks with ICC measure [186]. Meas.: Measure, MF: Model Fusion. The cells with lightgrey background indicate the models using LSTM and Cells with pink background indicate the models using LSTM-SW. The **bold** font indicates the best results.

Meas.	Task		Classification			Regression		
	Dataset		Uni-modality	Bi-Modality (MF)	Multi-Modality (MF)	Uni-modality	Bi-Modality (MF)	Multi-Modality (MF)
ICC	Subsets	PD	0.40 EDA-D	0.45 EMG-D EDA-D	**0.46**	0.43 EDA-D	**0.51** **EMG-D** **EDA-D**	0.49
		HPD	0.30 EDA-D	**0.41** **EMG-D** **EDA-D**	0.39	0.32 EDA-D	**0.41** **EMG-D** **EDA-D**	0.40
		EPD	0.50 EDA-D	0.53 EMG-D EDA-D	**0.57**	0.53 EDA-D	**0.58** **EMG-D** **EDA-D**	**0.58**
		TD	0.15 EDA-D	0.18 EMG-D EDA-D	**0.23**	0.17 EDA-D	0.26 EMG-D EDA-D	**0.30**
		HTD	0.33 EDA-D (LSTM-SW) EMG-D (LSTM)	**0.42** **EMG-D** **EDA-D**	0.35	0.30 EDA-D	0.32 EMG-D EDA-D	**0.38**
		ETD	0.14 (RFc)	0.22 EMG-D EDA-D	**0.26**	0.21 EDA-D	0.31 EMG-D EDA-D	**0.33**
	Red.Subsets	RPD	**0.83** **EDA-D** **(both LSTM)**	**0.83** **EMG-D** **EDA-D**	0.82	0.84 EDA-D	**0.85** **EMG-D** **EDA-D**	0.82
		RHPD	0.76 EDA-D	**0.79** **EMG-D** **EDA-D**	0.74	0.81 EDA-D (both LSTM)	**0.83** **EMG-D** **EDA-D**	0.73
		REPD	0.84 EDA-D (both LSTM)	**0.85** **EMG-D** **EDA-D** **(both LSTM)**	0.81 (both LSTM)	**0.88** **EDA-D**	0.87 EMG-D EDA-D	0.80
		RTD	0.31 EDA-D	**0.32** **EMG-D** **EDA-D**	0.28	0.24 (EDA-D)	**0.33** **EMG-D** **EDA-D**	0.29
		RETD	0.47 EDA-D	**0.52** **EMG-D** **EDA-D**	0.44	0.49 EDA-D	0.52 EMG-D EDA-D (both LSTM)	**0.56**



(a) Subsets

FIG. 6.13. Comparison of the best Uni-, Bi-modality, and Multi-modality models regarding classification and regression tasks with MSE and ICC measures. The Bi-modality and Multi-modality models use MF. MF: Model Fusion. Both LSTMs: LSTM and LSTM-SW. The **bold** font indicates the best results.



(b) Reduced Subsets

FIG. 6.13. Comparison of the best Uni-, Bi-modality, and Multi-modality models regarding classification and regression tasks with MSE and ICC measures. The Bi-modality and Multi-modality models use MF. MF: Model Fusion. Both LSTMs: LSTM and LSTM-SW. The **bold** font indicates the best results.

6.4.2 Classification

This section introduces more results by comparing the best Uni-, Bi, and Multi-modality classification models when applying RFc, LSTM, and LSTM-SW due to their superior performance than regression models with the Heat Phasic Dataset [HPD] and Heat Tonic Dataset [HTD], see Table 6.15 and Fig. 6.14.

Table 6.15. Comparison of the best Uni-, Bi-, and Multi-modality models with HPD and HTD regarding classification task. Triv.: Trivial. The bold font indicates the best models' results. * $p < 0.05$ when using paired t-test between RFc and LSTMs (LSTM and LSTM-SW). The **bold** font indicates the best models' results.

Measurement	Datasets	HPD				HTD			
	Model	Triv.	RFc	LSTM	LSTM-SW	Triv.	RFc	LSTM	LSTM-SW
Accuracy %	EDA-D (Uni-modality)	78.5	78.1	79.8*	79*	20	41.0	48.4	47.7
	FAD & EDA-D (Bi-modality)	78.5	-	**80.5***	80.2*	20	-	47.4*	**49.8***
	Multi-modality	78.5	-	79.3	77.6	20	-	41.6	42.2
Micro avg. precision%	EDA-D (Uni-modality)	0	24.6	36.6*	32.2*	0	42.7	48.2	47.7
	FAD & EDA-D (Bi-modality)	0	-	**42.8**	40.1*	0	-	47.2	**48.7**
	Multi-modality	0	-	34.9*	29.2	0	-	41.8	42.0
Micro avg. recall%	EDA-D (Uni-modality)	0	3.4	9.9*	10.9*	0	71.0	94.6*	**100***
	FAD & EDA-D (Bi-modality)	0	-	16.3*	21.4*	0	-	92.9*	97*
	Multi-modality	0	-	19.8*	**22.3***	0	-	90.8*	**99.9***
Micro avg. F1-Score%	EDA-D (Uni-modality)	0	5.9	15.2*	15.5*	0	52.9	62.3*	62.5*
	FAD & EDA-D (Bi-modality)	0	-	22.3*	**26.3***	0	-	60.7*	**63.3***
	Multi-modality	0	-	23.6*	24*	0	-	56	57.9*

The Bi- and Multi-modality models use MF. The architecture of LSTM and LSTM-SW is C-Mu(c) with a learning rate of $10^{-}6$. Bi-modality models, when applying both LSTM and LSTM-SW with EMG-D and EDA-D modalities, performed the best with HPD and HTD in terms of the 4 measures. LSTM models with HPD and HTD successfully predict discrete pain intensity levels in sequences compared to Trivial and RFc in terms of Accuracy, Micro avg. precision, Micro avg. recall, and Micro avg. F1-Score. The Bi- and Multi-modality models when applying LSTM-SW with HPD and HTD were better than those models using LSTM in terms of Micro avg. recall, and Micro avg. F1-Score. Further, Bi-modality models, when applying LSTM-SW, outperformed LSTM models with HTD in terms of Accuracy (49.8%) and Micro avg. precision (48.7%). EDA-D Uni-modality model, when applying LSTM-SW with HTD, performed the best in terms of Micro avg. recall (100%); however, the Bi-modality model, when applying LSTM-SW with the same dataset, performed excellent (97%).

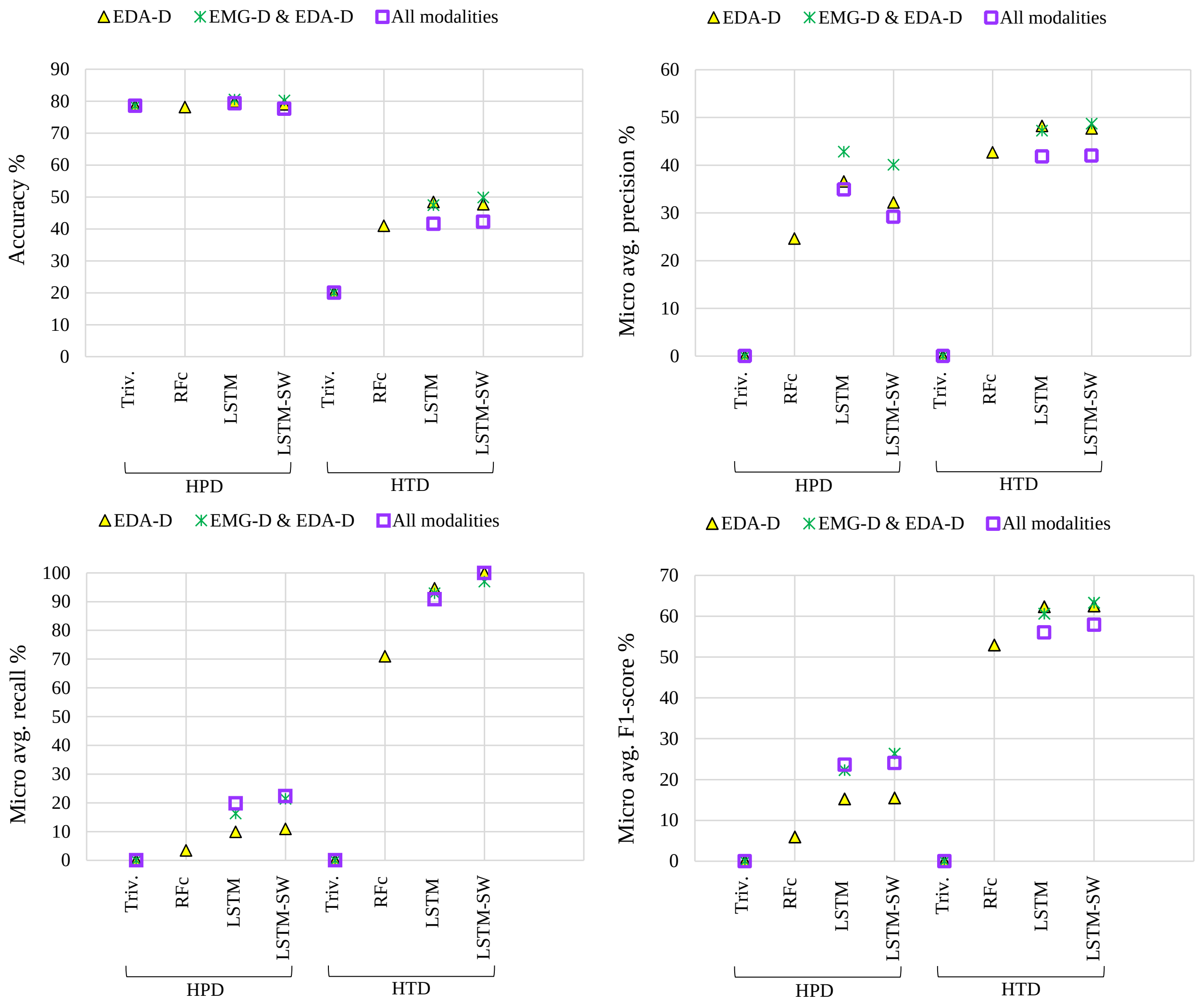


FIG. 6.14. Comparison of Uni-, Bi, and Multi-modality models when applying Trivial, RFc, LSTM, and LSTM-SW with HPD & HTD regarding classification task with different measures. The Bi- and Multi-modality models use MF. MF: Model Fusion, Triv.: Trivial.

Additionally, the Multi-modality model, when applying LSTM-SW with HPD, performed the best in terms of Micro avg. recall (22.3%). In terms of F1-score, Bi-modality models, when applying LSTM-SW with HPD and HTD, performed the best, about 26.3% and 63.3%.

6.4.3 Discussion

In this section, the obtained results of comparing Uni-, Bi, and Multi-modality models were summarized when applying RFc, LSTM, and LSTM-SW with 11 datasets from the experimental data. For evaluation, the focus in this work was on true and false positive samples when the models predict pain intensity. As true positive samples increase, the ICC will increase; as false positive samples decrease, the ICC will increase. Thus, the best results regarding regression were selected based on the highest ICC values rather than the lowest MSE values. The best Uni-modality models use EDA-D modality, and the best Bi-modality models use both EMG-D and EDA-D modalities. The Multi-modality models with reduced datasets improved the performance significantly compared to those with huge imbalanced datasets except with the regression model that uses RTD. The possible reason is that some useful samples are removed when applying the reduction strategy. Further, the regression models were superior to classification models with all datasets except HTD. The EMG-D & EDA-D Bi-modality classification models, when applying LSTM-SW with HTD, performed the best because this dataset was almost balanced. Additionally, the Multi-modality models performed the best with Tonic Dataset [TD], Electrical Tonic Dataset [ETD], and Reduced Electrical Tonic Dataset [RETD]. The Bi-modality models performed the best with the remaining 7 datasets except the Reduced Electrical Phasic Dataset [REPD]. The Bi- and Muti-modality models with REPD do not improve the performance, possibly because the outlier rate is increased when using more than one modality. Finally, both LSTM and LSTM-SW models outperformed the baseline model (RFc and RFr).

The correctly predicted samples regarding classification were closely investigated when applying RFc, LSTM, and LSTM-SW. The models with HPD performed the best in recognizing no pain and the highest pain intensity, whereas models with HTD performed the best in recognizing intermediate pain (low and moderate), see Table 6.16. The reason is that models may have difficulty in recognizing intermediate pain intensity stimulation in large imbalanced datasets (HPD) when HTD is less imbalanced (no pain [20%], two intermediate pain [27%], and the highest pain [26%]). Additionally, Bi- and Multi-modality models, when applying both LSTM and LSTM-SW with the imbalanced HPD, improved the performance of Uni-modality models. Finally, the EMG-D and EDA-D Bi-modality models using MF, when applying LSTM-SW with HPD and HTD, performed the best based on calculating the average of the 4-Class performance, which were 37.5% and 48.2%.

Table 6.16. Recall% result of 4-Class continuous pain intensity recognition tasks of HPD and HTD on testing set. Uni-modality refer to EDA-D Uni-modality, Bi-modality refer to EMG & EDA-D Bi-modality. The **bold** font indicates the best results.

Model		Dataset									
		HPD					HTD				
		BL	PH1	PH2	PH3	mean	BL	TH1	TH2	TH3	mean
Uni-modality	Trivial	100	0	0	0	25	100	0	0	0	25
	RFc	98.7	1.5	2	6.3	27.1	34.4	27.3	47	54.5	40.8
	LSTM	99.3	5.2	5.4	15.2	31.3	9.4	73	35.1	67.8	46.3
	LSTM-SW	98	3.1	1.7	23.3	31.5	0.30	72	39.8	68	45
Bi-modality	Trivial	100	0	0	0	25	100	0	0	0	25
	RFc	-	-	-	-	-	-	-	-	-	-
	LSTM	98.7	6.5	5.2	29.5	35	20.3	45.4	56.1	62.5	46.1
	LSTM-SW	97.4	7.4	7.2	37.9	**37.5**	18.6	45.8	62.7	65.5	**48.2**
Multi-modality	Trivial	100	0	0	0	25	100	0	0	0	25
	RFc	-	-	-	-	-	-	-	-	-	-
	LSTM	96.8	6.2	4.6	36.2	36	11.2	56.3	28.9	63.5	40
	LSTM-SW	94.1	5.9	6.3	39.5	36.5	2.7	40.4	55.2	61.8	40

CHAPTER 7

Discussion and Conclusion

THE current methods in the clinical application do not always allow for objective and robust measurement for pain diagnosis; moreover, they do not facilitate continuous monitoring of pain, especially for vulnerable patients. Automatic systems can be reliable and economical to solve these issues compared to human observers. This thesis shows the investigation of five sensor modalities for automatically monitoring continuous pain intensity on the X-ITE Pain Database, which are frontal RGB video, audio, electrocardiogram [ECG], surface electromyography [EMG], electrodermal activity [EDA]. Three distinct methods were proposed regarding classification and regression: A Random Forest (RF) baseline method ([Random Forest classifier (RFc) and Random Forest regression (RFr)]), Long-Short Term Memory (LSTM) method, and LSTM using the sample weighting method (LSTM-SW). This chapter has two sections, the first explains the findings from Uni-modality, Bi-modality, and Multi-modality experiments (see Section 7.1), and the second summarizes the key contributions of this thesis presents some future directions in research within this area (see Section 7.2).

7.1 Discussion

Implementing a reliable automatic model which relies on an imbalanced database is a major challenge for continuous monitoring of pain intensity. Another challenge is that there are a lot of outliers or label noise. The facial expressions and vocalizations responses, which were obtained from frontal RGB video and audio, were investigated for most samples in the X-ITE Pain Database. It turns out that plenty of labels do not match the observed facial pain expressions or expected vocalizations of pain due to individual differences in pain sensitivity and expressiveness. Some participants (subjects) show a lack of facial responses to pain (see Fig. 1.2 in Section 1.2): some have low pain sensitivity resulting in a high tolerance threshold requiring a temperature cutoff to avoid burns; others show a low tolerance threshold intentionally or unintentionally during stimulus calibration, possibly because to avoid severe pain. Further, some subjects talk or make a sound when no pain is experienced during the experiment. Such inconsistencies between the label and both video & audio were considered outliers. To address the extremely imbalanced database and outliers problems, it is preferable to use machine learning models with two types of dataset categories: (1) Subsets (6 datasets) and (2) Reduced Subsets (5 datasets) that were obtained after applying the reduction strategy (see Section 4.4); the Heat Tonic Dataset [HTD] was not reduced because it is nearly balanced [no pain (20%), two intermediate pain (27%), and the highest pain (26%)].

The reduction strategy, which was based on facial expressions analysis, reduced the influence of outliers (unimportant samples) by reducing some no pain samples prior to pain intensity sequences in a time series for each subject. An additional reason for using these different datasets was to explore the generalization capability of continuous pain intensity monitoring models. The results of comparing between Mean Squared Error [MSE] and Binary Cross-Entropy [BCE] loss functions, when applying LSTM with FAD modality, show that BCE is the best on most of 11 datasets; see the results of PD, HPD, EPD, HTD, RPD, RHPD, and REPD in Table 6.6 and Fig. 6.6. The results showed that the performance of pain intensity recognition in sequences was increased when using reduced datasets, except when using all modalities (frontal RGB video [frontal faces], audio, ECG, EMG, and EDA) with Tonic Dataset [TD]. See each dataset in Reduced Subsets compared to their equivalent datasets in Subsets in Fig. 6.4, 6.10, and 6.13; Section 6.2.3, 6.3.3, and 6.4.3 explain the results in detail. It seems more important to reduce the noise in imbalanced data than to keep very hard samples. The all modalities (Multi-modality) model with Reduced Tonic Dataset [RTD] does not outperform that with TD, probably due to more outliers (responses to heat stimuli) of audio and ECG modalities and high feature space dimensionality compared to the quite few training samples.

The regression models were superior to classification models when using huge imbalanced datasets, whereas classification models were the best with the almost balanced dataset (Heat Tonic Dataset [HTD]); both perform similarly with Heat Phasic Dataset [HPD]. For example, see the distribution of datasets in Table 4.3; HTD has the highest percentage distribution of pain intensity samples (showing the subjects while experiencing pain) in huge imbalanced datasets (Subsets) based on the mean value, which is about 26.7%, followed by HPD (7.17%). The results show that continuous monitoring of pain intensity in huge imbalanced datasets before and after reduction is a regression task. Thus, it is better to deal with the output of pain intensity as continuous values.

Alongside to prior works on the X-ITE Pain Database [29,30], the BioVid, and SenseEmotion databases [61,129,130], most EDA modality [EDA-D Uni-modality] models outperformed the other single modalities models. EDA is very sensitive [130] and less person-specific than other modalities [22]. EMG modality models were the second best single modality [EMG-D Uni-modality] models. With the TD and HTD, EMG outperformed EDA, in line with Werner et al. [22] and Kächele et al. [187] when they used the BioVid database. It seems the changes in muscle activity [EMG] to tonic stimuli tend to be more intense than the changes of the superficial muscles of the skin of the hand [EDA]. However, EDA-D Uni-modality models performed better than EMG-D Uni-modality models with reduced tonic datasets (RTD and RETD); it may be because the reduction strategy is more successful in reducing noise in EDA than in EMG. The facial expressions were the third best single modality [FAD Uni-modality] models and performed well, but not better than EDA due to low facial responses to pain with some subjects: some people have low sensitivity to pain (see [75,78] for more details about this phenomenon) or vice versa (some people are extremely expressive of their pain). Fig. 1.2 shows some difficult samples to recognize pain intensity based on facial features. The performances of ECG modality [ECG-D Uni-modality] models followed by audio modality [Audio-D Uni-modality] models were the worst single modalities, probably because of varying levels of noise. The ECG is sensitive to miscellaneous mixed noises. The audio signal includes many label noises, possibly due to the vocalizations responses when some subjects are stimulated with different pain intensities or when some subjects talk or produce other vocal responses during no pain is experienced.

To find good options for low-cost pain monitoring (using only two sensor modalities = Bi-modality models), the best three single modalities were chosen and then fused. The best modality (EDA-D Uni-modality) fused once with FAD Uni-modality and once with EMG-D Uni-modality. The Bi- and Multi-modality models help improve the performance of continuous pain intensity monitoring compared to the Uni-modality models. All Bi- and Multi-modality models using Model Fusion [MF] with LSTMs (LSTM and LSTM-SW) outperformed those using Decision Fusion [DF]. Models using MF succeeded in focusing on information from more reliable modalities while reducing the emphasis on the less reliable modalities. The Bi- and Multi-modality models using DF by applying mean-score mapping method performed well only when using HTD (almost balanced dataset); these results are in line with the results of Werner et al. [29] and Walter et al. [29,30]. Thus, facial expressions and EMG significantly benefit from the DF method, while it is found that EDA is a superior candidate for generic pain assessment models that use

DF. The possible reason for the poor performance of the DF method is the extremely imbalanced data; moreover, there are a lot of outliers in each single modality. The impact of such problems may be increased after calculating the average aggregation of classifier scores or regression probabilities when using two or more modalities, which tends to decrease the performance compared to the Uni-modality models.

The EMG-D & EDA-D Bi-modality models outperformed Multi-modality models when using MF with LSTMs on 6 datasets: (1) Phasic Dataset [PD], Heat Reduced Dataset [HTD], Reduced Phasic Dataset [RPD], Reduced Heat Phasic Dataset [RHPD], and RTD Reduced Tonic Dataset [RTD], and (2) [FAD and EDA-D] Bi-modality model was the best with HPD [Heat Phasic Dataset]. The possible reason is that the Multi-modality models include conflicts between modalities in these datasets, especially because of the outliers in worse modalities (Audio and ECG). Both EMG-D & EDA-D Bi-modality and Multi-modality models performed similarly with Electrical Phasic Database [EPD]. The Multi-modality models performed the best with three datasets: TD, Electrical Tonic Database [ETD], and Reduced Electrical Tonic Database [RETD]. Using Multi-modality models improved the performance with only tonic datasets because the responses to tonic stimuli are more intense than the response to phasic stimuli for each modality. Further, the response to electrical tonic stimuli tends to start earlier and more rapidly than the response to heat for each modality. Thus, all modalities could significantly benefit from this Multi-modality model that uses MF when the responses in all modalities are more intense. The EDA-D Uni-modality model performed the best with the Reduced Electrical Phasic Database [REPD]; it got an ICC of 0.88 and MSE of 0.03; both Bi- and Multi-modality models do not improve the performance; the large noise and conflict between Bi- and Multi-modality may be the reason.

According to the obtained results from Uni-, Bi-, and Multi-modality models, LSTMs outperformed Trivial (majority of vote = no pain) and baseline methods (RFc and RFr) on the 11 proposed datasets regarding the classification and regression (see Fig. 6.4, 6.10, and 6.13. Fig. 6.4 shows that the result of RFc model using EDA-D with ETD dataset is the best regarding classification in only Uni-modality experiments. With too small datasets for automated pain recognition, RFc outperformed deep learning methods [75]. However, after comparing LSTMs regression models to RFc using EDA-D with the same dataset [ETD], LSTM-SW models using EDA-D provided the best results. Further, several LSTM-SW models were better than LSTM models due to downweighting noisy samples by applying the sample weighting method [75]. The training samples with more facial responses were duplicated; these samples were identified using RFc (samples with classification scores above 0.3). The validation and test data were kept unmodified to ensure comparability of validation and test results. This weighting method was beneficial in several experiments because the dataset contains many pain samples without observable facial pain reactions that were impossible to classify correctly based on the facial features modality. Additionally, regarding best classification models compared to regression models in terms of micro avg. recall measurement, EMG-D & EDA-D Bi-modality models using MF with LSTM-SW on HPD and HTD datasets performed the best based on calculating the average of the 4-Class performance (see Table 6.15), which are 37.5% and 48.2%, respectively.

7.2 Conclusion

Due to the prior work in automatically recognizing pain intensity, the results showed that machines were much better than humans at recognizing pain intensity when analyzing the frontal faces in videos [75]. Thus, this thesis aims to introduce a promising automatic system for continuous monitoring of pain intensity based on analyzing facial expressions in frontal faces and other informative signals, which are audio and physiological signals. Fig. 7.1 shows how the proposed system monitors continuous pain intensity using two separate types of pain stimuli: phasic and tonic. It confirms that it is possible to recognize pain intensity using machine learning models with frontal RGB video, audio, and physiological signals (ECG, EMG, and EDA).

In this work, several experiments were conducted in order to (1) gain insights into continuous monitoring of pain intensity using frontal RGB video, audio, and physiological (ECG, EMG, and EDA) signals from X-ITE Pain Database, (2) to compare the performance between different proposed automatic methods regarding classification and regression, (3) to find the best machine learning model when analyzing time series features from each single modality and from a combination of two or all modalities; the modalities that were selected with the combination of two modalities were the first best single modality [EDA] with the second best single modality [EMG] once and the third best single modality (facial expressions) once, and (4) to introduce the baseline results for further research related to recognize continuous pain intensity in the X-ITE Pain Database.

In Chapter 4, the data from each modality was preprocessed to extract features for continuous monitoring of pain intensity; then the data was split into 80% of data for training, 10% for validation, and 10% for testing. The proposed machine learning methods were evaluated on 11 datasets (6 Subsets and 5 Reduced Subsets) from testing split to reduce the impact of an extremely imbalanced database and the impact of a lot of outliers or label noise. The Reduced Subsets were obtained after applying our reduction strategy on all datasets in Subsets except one (Heat Tonic Dataset [HTD]), which is almost balanced (only 20% of samples experience no pain). The reduction strategy focuses on reducing some no pain samples prior to each pain intensity sequence. Further, the labels of each participant (subject) were moved three seconds after (the facial pain responses typically are delayed by 2-3 seconds compared to stimulus), then used a sliding window with a time length of ten seconds ago. For more details about the proposed datasets, see Section 4.4.

Three machine learning methods (RF, LSTM, LSTM-SW) regarding classification and regression were introduced to monitor continuous pain intensity, see Chapter 5. The sample weighting method was suggested to reduce the weight of misclassified samples during training to improve the pain intensity recognition performance; these samples often contain low or no facial responses to pain. Further, three types of experiments were conducted using RF, LSTM, and LSTM-SW, which are: Uni-modality experiments (models were trained on data from single modality sensors), Bi-modality experiments (models were trained on fusing data from two sensors), and Multi-modality (models were trained on fusing data from five sensors) experiments. The reason for applying these multiple experiments is to provide the best automatic system for continuous monitoring of pain intensity after analyzing frontal faces from frontal RGB video, audio, physiological signals (ECG, EMG, and EDA), and their combination.

The quantitative results of Chapter 6 confirmed that automatically monitoring continuous pain intensity is possible regarding regression or classification. In general, among all obtained models, LSTMs (LSTM and LSTM-SW) gave the best predictive performance across different models compared to baseline methods (RFc and RFr); the Trivial failed to recognize pain intensity, whereas the proposed methods were significantly better. Further, the performance of LSTM with FAD when using Binary Cross-Entropy [BCE] loss was better than MSE. Therefore, LSTM using BCE loss was used with other Uni-modality (audio-D, ECG-D, EMG-D, and EDA-D) experiments and Bi- and Multi-modality using Model Fusion [MF] experiments. The RFc performed well with the very small size of data such as Electrical Tonic Dataset [ETD] when using to train the single modality of facial expressions [FAD Uni-modality]; however, its result was not superior Uni-, Bi-, and Multi-modality models that using LSTM regarding regression. In LSTM-SW, downweighting misclassified samples during training often increased the performance of LSTM. Further, reducing no pain samples in imbalanced datasets using reduction strategy on Subsets improved the performance (see Reduced Subsets result in Fig. 6.4), 6.10), and 6.13). Thus, it was demonstrated that the X-ITE Pain Database contains a lot of outliers which result lower predictive modeling performance. Additionally, the results indicated that for both classification and regression, EDA was the best single modality for monitoring continuous pain intensity, then EMG followed by facial expressions modality. These results were in line with prior works on the X-ITE Pain Database [29,30], the BioVid, and SenseEmotion databases [61,129,130]. ECG and audio were the worst, probably due to a lot of outliers in both modalities.

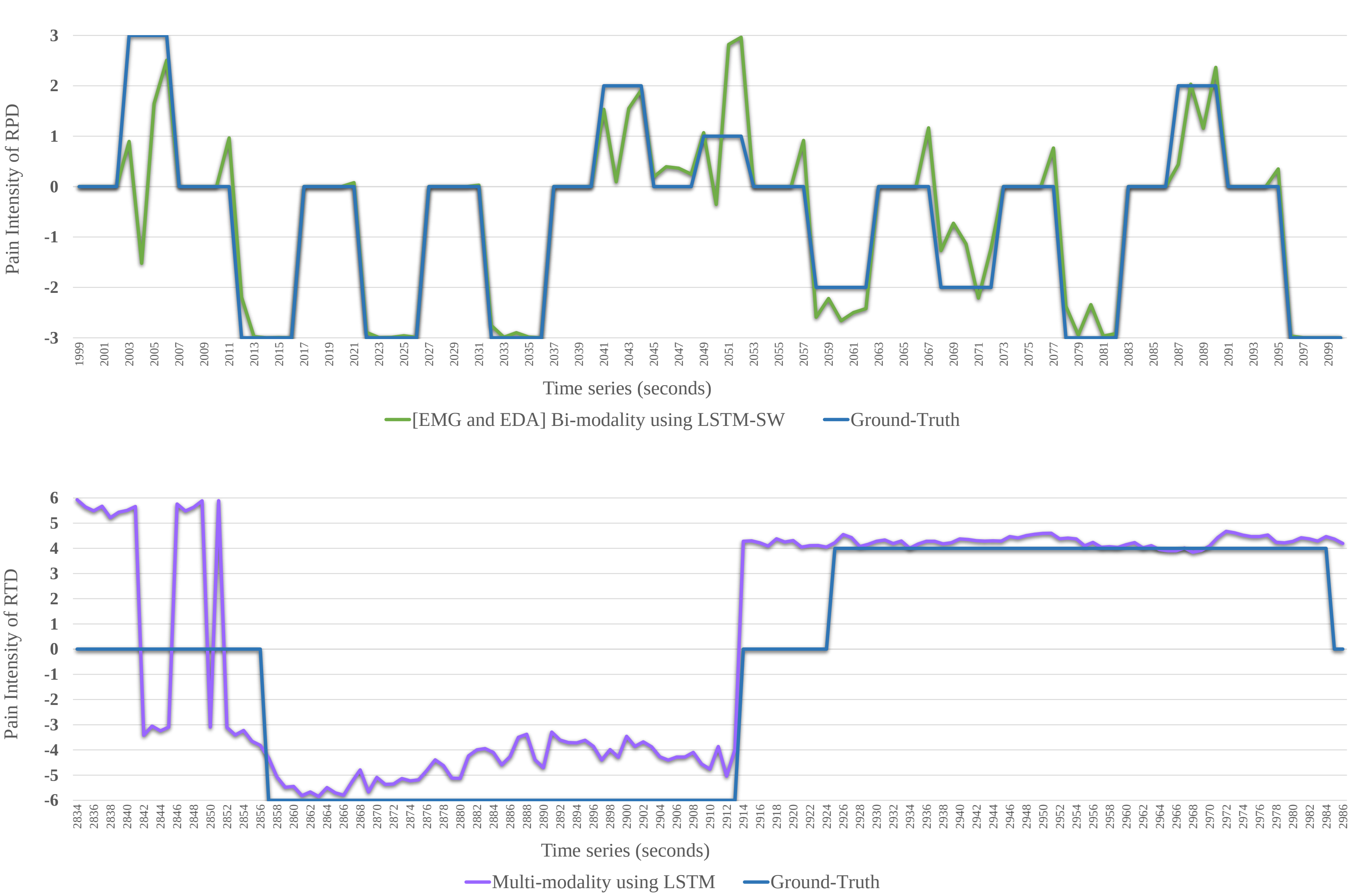


FIG. 7.1. A performance example of the proposed automated system for continuous monitoring pain intensity from test set, including results of the best models with RPD and RTD compared to Ground-Truth. RPD: Reduced Phasic Dataset. RTD: Reduced Tonic Dataset.

The combinations of two or all modalities were suggested to improve continuous monitoring of pain intensity compared to the best single modality models [EDA Uni-modality models]. In this case, it was a good idea to have a fused deep learning model (LSTMs using MF) to benefit from the advantages of two or all sensor modalities, see Section 6.3 and 6.4. Only the performance of one single modality model does not improve, which is EDA when applying LSTM-SW with Reduced Electrical Phasic Dataset [REPD]. The possible reason is the conflicts between modalities in this dataset, especially because of the outliers; it seems that outliers detection methods are required for further improvements when using the fused modality. The models where all modalities were fused (Multi-modality) were the best when using most tonic datasets (Tonic Dataset [TD], Electrical Tonic Dataset [ETD], and Reduced Electrical Tonic Dataset [RETD]). The responses to tonic stimuli and especially the electrical stimuli tended to be more intense and rapid than responses to phasic stimuli for each single modality. The Multi-modality models with MF significantly benefited from all modalities when the responses were more intense within all modalities. Clearly, the fused models of EMG & EDA Bi-modality models were the best when using phasic balanced dataset [HTD] compared to EDA Uni-modality and Multi-modality models; these phasic datasets were Phasic Dataset [PD], Heat Phasic Dataset [HPD], Electrical Phasic Dataset [EPD], Reduced Phasic Dataset [RPD], Reduced Heat Phasic Dataset [RHPD], and Reduced Electrical Phasic Dataset [REPD]. The possible reason for the poor performance of using the Multi-modality models is the conflicts between modalities because of the outliers, especially in the worst modalities (Audio and ECG). This result shows that EMG and EDA are very good options for cost-effective pain monitoring. There is no need to use all modalities.

Most results of applying Bi- and multi-modality experiments when using Decision Fusion (DF) with RF, LSTM, and LSTM-SW were not superior to EDA Uni-modality. A major drawback of DF is the possibility of impeding the overall system performance by combining several single classifiers; the impact of the outliers may be increased after calculating the average aggregation of classifier scores or regression probabilities. Further, the classification models outperformed regression models with the almost balanced dataset (HTD) and performed similarly with HPD. The HTD result is consistent with Werner et al. [29] and Walter et al. [30] findings, who used a balanced database to classify the pre-segmented time windows. The results showed that automatically monitoring continuous pain intensity is a regression task when using imbalanced datasets. Regression reduces the effect of confounding variables by isolating the effect of each variable by allowing the role of each independent variable to learn without worrying about other variables in the model; such analysis is done by estimating the effect of changing one independent variable on the dependent variable while keeping all other independent variables constant. Although separating tonic datasets from phasic datasets because of the huge difference in size [models would be biased towards the majority (phasic datasets)], the performance was worse compared to use phasic datasets due to the small size of the data. Additionally, the recognition of electrical pain stimuli worked better than the recognition of heat pain stimuli in imbalanced datasets sequences. It may probably be the responses to electrical stimuli tend to be more intense, start earlier, and more rapidly than responses to heat because the electrical pain is instantly felt with full intensity. In contrast, the heat pain is building up slowly. However, this finding can not be generalized because the size of electrical pain data is larger than heat pain data; see the number of samples for each dataset in Table 4.2.

The results of this work are promising. However, the following limitations should be addressed to advance this system. These limitations can be summarized as follows:

- The X-ITE Pain Database is based on healthy participants. The database does not contain a vulnerable group; however, this system can help to predict pain particularly in vulnerable patients, but it has not yet been implemented to them. Such a system should be applied to those patients before it is ready for clinical studies.

- The imbalanced database problem. Although applying the proposed reduction strategy to the imbalanced datasets helps to improve the performance, there are still plenty of outliers that limit further performance improvement. The possible solution for this problem: the outliers in each modality should be handled by (1) identifying them (i.e., those based on distances, density, or clustering, etc.) based on the calculation of distances and then downweighting them in all modalities, (2) training the deep learning networks with the minkowski error [188], or (3) using any other methods that reduce the impact of outliers but do not eliminate them. Data cleaning by removing such outlier samples may be used for improving the individual modality-based recognition performance. However, this may remove some samples, which are useful for improving multimodal pain recognition because there may be pain responses in other modalities.

- The small size of training data. A larger dataset with more pain intensities is necessary for more reliable automatic monitoring of continuous pain intensity. An alternative solution is to apply self-supervised techniques, which can be beneficial to improve the efficiency of the small amount of data.

- The need for extracting informative features. In this work, each time series (second) is summarized using four statistical measures (minimum, maximum, mean, and standard deviation) calculated from the time series itself, its first and second derivative. The proposed methods for automatically monitoring continuous pain intensity are based on combining the temporal aspects of preceding ten-seconds of descriptors for each modality. It is possible that using more statistical measures of the time series will improve the performance; using the remaining of the statistical measures in the Werner et al. [189] study is suggested to improve system performance.

- The need for efficient deep learning methods. The used methods are built based on the frontal faces [facial expressions] modality; these are not good enough with audio and ECG modalities. It is highly recommended to research the best method for each modality individually, then combine them to improve system performance.

APPENDIX A

Uni-modality Results

This chapter is dedicated to provide the detailed results of performing the proposed methods [RF and LSTMs (LSTM, and LSTM-SW)] with single modalities, two fused modalities, and all fused modalities regarding classification and regression for continuous monitoring pain intensity from the 11 proposed datasets on the X-ITE Pain Database.

Appendix A shows all results of using all single modalities [Uni-modality] models (FAD, Audio-D, ECG-D, EMG-D, and EDA-D). Regarding classification and regression and in terms of MSE and ICC measures, Table A.1 and A.2 present that all Uni-modality regression models outperform those using classification when the datasets are imbalanced. Heat Tonic Dataset [HTD] is almost balanced, and the classification model is the best (ICC = 0.33). Table A.3 and A.4 show the results of performing Accuracy, Micro avg. precision, Micro avg. recall, and Micro avg. F1-score measures, which confirmed that EDA is the best single signal for recognizing continuous pain intensity followed by EMG and then facial expressions, ECG and audio are the worst. RFc classification models performed the best in terms of F1-score with imbalanced tonic datasets (Tonic Dataset [TD], Electrical Tonic Dataset [ETD], Reduced Tonic Dataset [RTD], and Reduced Electrical Tonic Dataset [RETD]), about 12.2%, 14.1%, 25%, and 36.8%, respectively. However, these results were not the best due to the comparison between classification and regression models results, EDA Uni-modality regression models using LSTM and LSTM-SW were the best except with RTD (ICC = 0.31) (see Table 6.4 in Section 6.2.1.2).

Table A.1. Comparison of the Uni-modality models when applying Trivial, RF (RFc and RFr), and LSTMs (LSTM and LSTM-SW) regarding classification and regression tasks with MSE and ICC measures. Red. Subsets: Reduced Subsets. The grey cells indicate the best results in each model. The **bold** font indicate the best results versus regression in Table A.2.

Task	Measure	Model		n-Class	Trivial	RFc					Architecture (Loss =CCE)	LSTM					LSTM-SW					Learning rate
		Dataset				FAD	Audio-D	ECG-D	EMG-D	EDA-D		FAD	Audio-D	ECG-D	EMG-D	EDA-D	FAD	Audio-D	ECG-D	EMG-D	EDA-D	
Classification	MSE	Subsets	PD	7	0.10	0.10	0.11	0.17	0.10	0.09	A(c)	0.09	0.1	0.1	0.09	0.09	0.10	0.1	0.1	0.09	0.09	10^{-5}
			HPD	4	0.11	0.11	0.12	0.11	0.11	0.11	C(c)	0.1	0.12	0.11	0.10	0.1	0.11	0.12	0.11	0.11	0.11	
			EPD	4	0.07	0.07	0.07	0.25	0.07	0.06	C(c)	0.07	0.08	0.07	0.06	0.06	0.08	0.08	0.07	0.07	0.06	
			TD	7	0.12	0.12	0.12	0.18	0.16	0.16	A(c)	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.11	0.12	10^{-6}
			HTD	4	0.41	0.25	0.23	0.3	0.19	0.18	C(c)	0.18	0.24	0.16	0.16	**0.15**	0.16	0.24	0.18	**0.15**	0.15	
			ETD	4	0.09	0.09	0.09	0.11	0.17	0.11	C(c)	0.09	0.09	0.09	0.09	0.08	0.08	0.09	0.09	0.09	0.08	
		Red. Subsets	RPD	7	0.23	0.20	0.23	0.22	0.19	0.14	A(c)	0.12	0.21	0.13	0.16	0.05	0.14	0.2	0.12	0.15	0.05	10^{-4}
			RHPD	4	0.26	0.23	0.26	0.25	0.21	0.22	C(c)	0.13	0.21	0.13	0.16	0.08	0.14	0.21	0.12	0.17	0.07	
			REPD	4	0.26	0.21	0.25	0.25	0.18	0.12	C(c)	0.13	0.24	0.14	0.16	0.05	0.14	0.23	0.13	0.15	0.05	
			RTD	7	0.25	0.23	0.24	0.26	0.21	0.18	B(c)	0.25	0.25	0.25	0.21	0.21	0.24	0.25	0.25	0.22	0.19	10^{-6}
			RETD	4	0.25	0.24	0.24	0.35	0.21	0.18	D(c)	0.24	0.24	0.26	0.19	0.16	0.26	0.27	0.25	0.19	0.16	
	ICC	Subsets	PD	7	0	0.1	0.01	0.01	0.17	0.33	A(c)	0.18	0.02	0	0.20	0.30	0.2	0.04	0.02	0.24	0.40	10^{-5}
			HPD	4	0	0.16	0.01	0.02	0.18	0.11	C(c)	0.26	0.06	0.03	0.21	0.30	0.26	0.08	0.06	0.25	0.29	
			EPD	4	0	0.16	0.01	0	0.23	0.37	C(c)	0.25	0.06	0.01	0.27	0.36	0.24	0.06	0.05	0.32	0.50	
			TD	7	0	0.08	0.02	0.04	0.09	0.10	A(c)	0.07	0.04	0	0.06	0.07	0.08	0.05	0	0.15	0.11	10^{-6}
			HTD	4	0	0.11	0.08	0.01	0.29	0.30	C(c)	0.19	0.12	0.11	0.29	**0.33**	0.20	0.12	0.10	**0.33**	0.31	
			ETD	4	0	0.07	0.04	0.05	0.06	0.14	C(c)	0.09	0.06	0	0.02	0.09	0.11	0	0.01	0.06	0.09	
		Red. Subsets	RPD	7	0	0.19	0.07	0.12	0.28	0.44	A(c)	0.57	0.21	0.51	0.44	0.83	0.49	0.25	0.53	0.44	0.83	10^{-4}
			RHPD	4	0	0.21	0.07	0.1	0.26	0.23	C(c)	0.56	0.3	0.57	0.48	0.75	0.55	0.31	0.62	0.45	0.76	
			REPD	4	0	0.27	0.11	0.11	0.38	0.58	C(c)	0.55	0.23	0.52	0.49	0.84	0.52	0.24	0.55	0.52	0.84	
			RTD	7	0	0.09	0.06	0.04	0.17	0.21	B(c)	0.05	0.06	0.03	0.21	0.25	0.08	0.1	0.03	0.18	**0.31**	10^{-6}
			RETD	4	0	0.12	0.11	0	0.26	0.37	D(c)	0.15	0.17	0.02	0.33	0.47	0.1	0.15	0.01	0.33	0.46	

Table A.2. Comparison of the Uni-modality models when applying Trivial, RF (RFc and RFr), and LSTMs (LSTM and LSTM-SW) regarding regression task with MSE and ICC measures. Red. Subsets: Reduced Subsets. The grey cells indicate the best results in each model. The **bold** font indicate the best results versus classification in Table A.1.

Task	Measure	Model		n-Class	Trivial	RFr					Architecture (Loss =BCE)	LSTM					LSTM-SW					Learning rate
		Dataset				FAD	Audio-D	ECG-D	EMG-D	EDA-D		FAD	Audio-D	ECG-D	EMG-D	EDA-D	FAD	Audio-D	ECG-D	EMG-D	EDA-D	
Regression	MSE	Subsets	PD	7	0.10	0.09	0.09	0.12	0.08	0.07	A(r)	0.08	0.08	0.08	0.07	**0.06**	0.08	0.09	0.09	0.07	0.08	10^{-5}
			HPD	4	0.11	0.09	0.10	0.10	0.09	0.09	C(r)	0.08	0.09	0.09	0.08	0.08	0.09	0.10	0.09	0.09	**0.08**	
			EPD	4	0.07	0.06	0.07	0.14	0.06	0.05	C(r)	0.05	0.10	0.06	0.05	0.04	0.06	0.07	0.07	0.06	**0.05**	
			TD	7	0.12	0.10	0.10	0.11	0.70	0.13	A(r)	0.09	0.07	0.09	0.08	0.08	0.09	0.10	0.10	0.09	0.09	10^{-6}
			HTD	4	0.41	0.13	0.14	0.14	0.14	0.13	C(r)	0.12	0.14	0.13	0.11	0.11	0.16	0.14	0.13	0.11	0.11	
			ETD	4	0.09	0.09	0.09	0.08	0.14	0.09	C(r)	0.07	0.07	0.08	0.07	0.07	0.08	0.09	0.08	0.07	**0.07**	
		Red. Subsets	RPD	7	0.23	0.12	0.14	0.13	0.11	0.09	A(r)	0.08	0.14	0.07	0.09	0.04	0.09	0.14	0.08	0.09	**0.04**	10^{-4}
			RHPD	4	0.26	0.13	0.15	0.14	0.13	0.13	C(r)	0.09	0.16	0.07	0.10	**0.05**	0.09	0.16	0.08	0.10	**0.05**	
			REPD	4	0.26	0.12	0.15	0.15	0.12	0.08	C(r)	0.10	0.17	0.08	0.09	0.04	0.11	0.16	0.08	0.10	**0.03**	
			RTD	7	0.25	0.14	0.13	0.16	0.13	0.13	B(r)	0.13	0.13	0.13	0.12	0.11	0.13	0.13	0.13	0.12	0.11	10^{-6}
			RETD	4	0.25	0.15	0.14	0.2	0.13	0.12	D(r)	0.14	0.17	0.14	0.12	**0.10**	0.13	0.15	0.15	0.12	0.10	
	ICC	Subsets	PD	7	0	0.13	0.04	0.03	0.19	0.41	A(r)	0.20	0.03	0.04	0.26	0.43	0.22	0.06	0.05	0.27	0.20	10^{-5}
			HPD	4	0	0.19	0.04	0.05	0.20	0.20	C(r)	0.28	0.08	0.06	0.24	0.28	0.27	0.08	0.12	0.26	0.32	
			EPD	4	0	0.18	0.04	0.03	0.24	0.47	C(r)	0.27	0.06	0.05	0.32	0.49	0.28	0.06	0.06	0.34	0.53	
			TD	7	0	0.10	0.05	0.05	0.04	0.09	A(r)	0.11	0.05	0.02	0.15	0.12	0.12	0.05	0.02	**0.17**	0.11	10^{-6}
			HTD	4	0	0.17	0.09	0.05	0.22	0.31	C(r)	0.15	0.08	0.03	0.18	0.28	0.15	0.08	0.03	0.27	0.30	
			ETD	4	0	0.09	0.05	0.06	0.07	0.17	C(r)	0.09	0.06	0	0.17	0.09	0.11	0.04	0.01	0.19	**0.21**	
		Red. Subsets	RPD	7	0	0.23	0.08	0.14	0.3	0.45	A(r)	0.56	0.24	0.62	0.45	0.81	0.54	0.22	0.58	0.45	**0.84**	10^{-4}
			RHPD	4	0	0.26	0.12	0.12	0.28	0.24	C(r)	0.62	0.3	0.68	0.5	**0.81**	0.62	0.28	0.63	0.53	**0.81**	
			REPD	4	0	0.32	0.11	0.12	0.38	0.63	C(r)	0.52	0.22	0.62	0.56	0.86	0.51	0.25	0.62	0.53	**0.88**	
			RTD	7	0	0.05	0.06	0.01	0.14	0.18	B(r)	0.04	0.06	0.02	0.16	0.23	0.07	0.06	0.02	0.17	0.24	10^{-6}
			RETD	4	0	0.15	0.13	0.01	0.28	0.34	D(r)	0.09	0.13	0.01	0.28	**0.49**	0.2	0.25	0	0.28	0.44	

Table A.3. Comparison of the Uni-modality models when applying Trivial, RF (RFc and RFr), and LSTMs (LSTM and LSTM-SW) regarding classification task with Accuracy and Micro average precision measures. Reduced Subsets: Red. Subsets. The grey cells indicate the best results in each model. The **bold** font indicate the best results. * p-value<0.05 when using paired t-test between RFc and LSTMs (LSTM and LSTM-SW).

Task	Measure	Model		n-Class	Trivial	RFc					Architecture (Loss =CCE)	LSTM					LSTM-SW					Learning rate
		Dataset				FAD	Audio-D	ECG-D	EMG-D	EDA-D		FAD	Audio-D	ECG-D	EMG-D	EDA-D	FAD	Audio-D	ECG-D	EMG-D	EDA-D	
Classification	Accuracy %	Subsets	PD	7	77.7	76.6	76.8	60.9	75.9	75.8	A(c)	78.2*	77.4*	77.6*	78.5*	**78.6***	77.5*	77.2	77.5*	78*	78.4*	10^{-5}
			HPD	4	78.5	77.8	77.7	78.1	76.3	78.1	C(c)	78.9*	0.12	78.5*	79.1*	**79.8***	77.9	77.4	78.4	78.7*	79*	
			EPD	4	86.1	85.5	85.5	58.2	84.9	86.1	C(c)	86.6*	0.12	86.1*	87.1*	**87.1***	85.4	85.5	86*	86.6*	87*	
			TD	7	70.3	68.5	69	62.9	63.2	50	A(c)	70.7*	70.3*	70.3	70.7*	70.7*	69.9*	70.2*	70.3	**71.3***	70.6*	10^{-6}
			HTD	4	20	29.1	31	27.8	35.2	41	C(c)	32.8	0.08	31.4	39.3	**48.4**	33.7	33.8	31.6	39.9	47.7	
			ETD	4	82	80.7	80.9	78.8	71	74.3	C(c)	81.7	82*	82	82.1*	**82.4***	**82.4***	81.5	82	81.8*	**82.4***	
		Red. Subsets	RPD	7	50	47.6	45.7	48.5	48.5	43.2	A(c)	57.4*	47.6*	57.3*	55.1*	**67.3***	54.7*	46.9	57*	53.8*	66.9*	10^{-4}
			RHPD	4	50.1	49.0	44.6	48.7	49.4	45.1	C(c)	60.9*	49.4*	61.6*	58.8*	**68.6***	58.7*	49.5*	62.1*	56.2*	**68.6***	
			REPD	4	50	49.3	44.8	37.5	52.2	55.5	C(c)	61.5*	47.8*	61.3*	58.9*	**75.8***	58.9*	49.2*	62*	58.9*	75.1*	
			RTD	7	38.1	33.9	32.2	24.3	35	30.4	B(c)	39.2*	36*	37.7*	41.6*	42.2*	37.4	34.9	37.9*	40.5*	**42.7***	10^{-6}
			RETD	4	49	43.1	42.3	33.7	43.4	48.7	D(c)	50.1*	49.8*	48.2*	53.7*	**57.7***	47.5*	47*	49.1*	53.7*	57.5*	
	Micro avg. recision %	Subsets	PD	7	0	16.8	11.1	4.9	16.8	24.9	A(c)	27.5*	17.9	8.30	34.5*	**53.3***	24.7*	13.7	13.2	35.3*	43.7*	10^{-5}
			HPD	4	0	24.3	10.4	13.9	21.9	24.6	C(c)	32.4*	20.2*	8.10	34.2*	**36.6***	26.6	18.9*	21.8	33.5*	32.2*	
			EPD	4	0	27.7	11	4.8	24.5	37.5	C(c)	38.7*	24.3	21.2	49.1*	**56.3***	29.7	24.3	14.3	41.1*	47.9*	
			TD	7	0	14.3	8.4	8.3	11.3	10.2	A(c)	18.5	17.4	0	27.1	23.2*	13.3	12.2	0	**36.7***	25.8	10^{-6}
			HTD	4	0	31.1	31.8	29.2	38	42.7	C(c)	33.3	32	31.4	39.3	**48.2**	33.7	33.8	31.6	39.9	47.7	
			ETD	4	0	14.5	19.8	14.3	9.2	15.2	C(c)	12.4	15.3	0	0	28.6	24.4	0.9	5	15.1	**35.7**	
		Red. Subsets	RPD	7	0	20.1	14.1	14	23.1	20.1	A(c)	28.5*	17.7	26*	32.1*	**41.8***	26.6*	17.7	25.7*	29.8*	40.5*	10^{-4}
			RHPD	4	0	27.7	21.1	20.7	28.8	24.2	C(c)	36.8*	26.1*	34.9*	38.4*	**40.2***	27.8	27.1*	32.5*	34.8*	39.6*	
			REPD	4	0	31.5	24.9	21.6	35.3	35.6	C(c)	40.7*	25.6	38.7*	44.9*	**55.2***	33.5*	28	35.3*	42*	53.9*	
			RTD	7	0	15.8	14.6	16.9	20.2	19.1	B(c)	30.7	17.3*	21	35.4*	**44.6***	37.7*	16.3	8.9	32.4*	40.8*	10^{-6}
			RETD	4	0	20.2	23.8	16	25.2	35.5	D(c)	27.7	28.6	13.2	38.1	61.6*	19.1	22.4	7.1	38.1	**61.9***	

Table A.4. Comparison of the Uni-modality models when applying Trivial, RF (RFc and RFr), and LSTMs (LSTM and LSTM-SW) regarding classification task with the Micro average recall and Micro average F1-score measures. Reduced Subsets: Red. Subsets. The grey cells indicate the best results in each model. The **bold** font indicate the best results. * p-value < 0.05 when using paired t-test between RFc and LSTMs (LSTM and LSTM-SW).

Task	Measure	Model		n-Class	Trivial	RFc					Architecture (Loss =CCE)	LSTM					LSTM-SW					Learning rate
		Dataset				FAD	Audio-D	ECG-D	EMG-D	EDA-D		FAD	Audio-D	ECG-D	EMG-D	EDA-D	FAD	Audio-D	ECG-D	EMG-D	EDA-D	
Classification	Micro avg. recall %	Subsets	PD	7	0	3.5	0.5	**15.4**	6.6	10.6	A(c)	7.15*	0.7	0.1	6.6	8	9.42*	1.4	0.5	9.4*	12.9*	10^{-5}
			HPD	4	0	6.48	0.7	0.7	8.8	3.4	C(c)	11.8*	2.2*	0.5	8.8	9.9*	**13.5***	3.7*	1.5	11.8	10.9*	
			EPD	4	0	6.29	0.6	17	11.6	14.8	C(c)	9.93*	2.8	0.4	10.8	15.6	12.1*	2.8	2	14.9*	**27***	
			TD	7	0	3.51	0.9	7.2	6.1	**16.7**	A(c)	2.01	1	0	1.8*	1.6	3.74	1.3	0	4.4	2.9	10^{-6}
			HTD	4	0	49.4	71.7	69.2	65.5	71	C(c)	92.4*	99.5*	**100***	99.6*	94.6*	**100***	**100***	92.9*	**100***	**100***	
			ETD	4	0	2.6	2.3	1.6	7.9*	**13.8**	C(c)	4.78	2.3	0	0	3.3	4.39	0.4	0.3	2.1	3.1*	
		Red. Subsets	RPD	7	0	11.9	5.2	4.9	18.2	37.7	A(c)	44.6*	12.9*	31.2*	23.5*	73.3*	39.3*	17.5*	40.4*	28.9*	**78.4***	10^{-4}
			RHPD	4	0	21.6	12.7	7.7	27.7	36.2	C(c)	57.6*	33.5*	57.6*	39.8*	88*	62.4*	37.2*	68.8*	40.2*	**93.1***	
			REPD	4	0	25.2	16.9	38.4	37.3	59.5	C(c)	53.8*	26.8*	43.9	35.6	92.7*	59.2*	26.3*	53.4	48.2*	**96.8***	
			RTD	7	0	11.7	9.1	22.2	19.9	**38.2**	B(c)	5.55	4.8	0.8	10.4	8.9	9.48	7.3	1.4	9.7	10.5	10^{-6}
			RETD	4	0	15.1	20.1	23.9	26.4	**41**	D(c)	8.97	10.6	1.6	16.6	20.9	8.09	16.8	0.6	16.6	20.6	
	Micro avg. F1-score %	Subsets	PD	7	0	5.7	1	3.9	8.7	14.2	A(c)	10.2*	1.3	0.2	10	13	12.1*	2.4*	0.9	12.9*	**18.4***	10^{-5}
			HPD	4	0	9.9	1.2	1.4	11.9	5.9	C(c)	15.5*	3.9*	1	12.7	15.2*	**16.2***	5.8*	2.7	15.8*	15.5*	
			EPD	4	0	10	1.1	6.6	14.6	20.5	C(c)	14.7*	4.4*	0.7	15.6	21.3	16.3*	4.4*	3.4	19.8*	**31.2***	
			TD	7	0	5.5	1.6	3	7.4	**12.2**	A(c)	3.37	1.8	0	3.2	2.9	5.63	2.3	0	7.4	5.1	10^{-6}
			HTD	4	0	37.6	43.7	40.7	46	52.9	C(c)	46.6	47.9	46.2	55.6*	62.3*	48.2*	49.5*	45.5	56.3	**62.5***	
			ETD	4	0	4.2	4	2.7	7.5*	**14.1**	C(c)	6.21	3.6	0	0	5.6	6.13	0.5	0.5	3.2	5.2*	
		Red. Subsets	RPD	7	0	14.6	7.4	7.2	20	25.7	A(c)	34.6*	14.2*	26.6*	26.2*	53.2*	31.5*	17.1*	28.8*	28.6*	**53.4***	10^{-4}
			RHPD	4	0	23.8	15.7	11	27.6	28.8	C(c)	44.8*	29*	38.5*	38.5*	55*	43.5*	31.1*	41.2*	37*	**55.4***	
			REPD	4	0	27.8	19.9	25.7	36.1	44.3	C(c)	46.2*	25.9*	37.8*	38.6	**69***	45.9*	27*	40.5*	44.5*	**69***	
			RTD	7	0	12.8	10.8	15.2	19.6	**25.1**	B(c)	8.26	6.3	1.4	15.2	14.3	12.2	7.4	2.2	13.9	16.2	10^{-6}
			RETD	4	0	16.8	21.5	19.1	26.4	**36.8**	D(c)	12.0	14.6	2.7	21.5	30.2	10.7	17.4	1	21.5	29.7	

APPENDIX B

Bi-modality Results

This chapter is dedicated to provide the detailed results of performing the proposed methods [RF and LSTMs (LSTM, and LSTM-SW)] with single modalities, two fused modalities, and all fused modalities regarding classification and regression for continuous monitoring pain intensity from the 11 proposed datasets on the X-ITE Pain Database.

Two Uni-Modality ([FAD & EDA-D] or [EMG & EDA]) were fused to improve Uni-modality results, which were called Bi-Modality. Section 6.3 and Section 6.4 introduces all Bi-Modality models' results when using Model Fusion [MF] and Decision Fusion [DF] regarding classification and regression. Appendix B presents more detailed results regarding classification using Accuracy, Micro avg. precision, Micro avg. recall, and Micro avg. F1-score measures, see Table B.1 and B.2. In line with the MSE and ICC measures' results, Bi-modality models using MF with LSTMs outperformed those using DF except when using TD and RTD, Bi-modality models using DF with RFc performed the best (F1-score about 10.2% and 25.2%, respectively). The possible reason is that RFc performs well with a small size of data. However, this result was not the best due to the comparison of the results between classification and regression models; EMG & EDA Bi-modality regression models using MF with LSTMs were the best (see Table 6.10 in Section 6.3.1.2).

Table B.1. Comparison of the best Bi-modality models when applying Trivial, RF (RFc and RFr), and LSTMs (LSTM and LSTM-SW) regarding classification task with Accuracy and Micro average precision measures. Reduced Subsets: Red. Subsets. DF: Decision Fusion, MF: Model Fusion. The grey cells indicate the best results in each model. The **bold** font indicate the best results. * p-value<0.05 when using paired t-test between RFc and LSTMs (LSTM and LSTM-SW).

Task	Measure	Model		n-Class	Trivial	RFc		Architecture (Loss =CCE)	LSTM				LSTM-SW				Learning rate
		Dataset				DF			DF		MF		DF		MF		
						FAD EDA-D	EMG-D EAD-D		FAD EDA-D	EMG-D EAD-D	FAD EDA-D	EMG-D EAD-D	FAD EDA-D	EMG-D EAD-D	FAD EDA-D	EMG-D EAD-D	
Classification	Accuracy %	Subsets	PD	7	77.7	78.1	78.2	A-Bi(c)	78.5*	78.7*	78.7	**79.1***	78.8*	78.9*	77.8	78.3	10^{-5}
			HPD	4	78.5	47.7	55.2	C-Bi(c)	79.9*	78.2	79.7	**80.5**	79.9*	79.8*	79.5	80.2*	
			EPD	4	86.1	78.2	78.9	C-Bi(c)	87.2*	87.4*	87.7	**87.8***	87.2*	87.7*	87	87.2	
			TD	7	70.3	67.5	64.3	A-Bi(c)	70.4*	70.6*	70.7	71.2	70.9*	71.3*	71.1	**71.6**	10^{-6}
			HTD	4	20	40.9	40.8	C-Bi(c)	45.5	48.1	44.6	47.4	43.2	47.5	45.6	**49.8**	
			ETD	4	82	81.9	79.2	C-Bi(c)	82.4	82.2*	82	**83.2**	81	81.1	82.4	**83.2**	
		Red. Subsets	RPD	7	50	51.7	52.1	A-Bi(c)	66.2*	64.6*	66.2	66.7	65.8*	65.7*	66.9*	**67.3**	10^{-4}
			RHPD	4	50.1	52	52.3	C-Bi(c)	68*	67.3*	68.4	**70.5***	67.2*	68.3*	68.4	70.6	
			REPD	4	50	58.9	60.2	C-Bi(c)	74.9*	73.2*	73.4	**75.3**	73.9*	74.8*	73.9	75.2	
			RTD	7	38.1	36.7	36.7	B-Bi(c)	41.6*	42.6*	43.9*	44.4	42.2*	42.6*	42.2	**45.2**	10^{-6}
			RETD	4	49	50.4	49.7	D-Bi(c)	54.8*	55.9*	55.7	**57.2**	54.3*	55.8*	55.9	56.6	
	Micro avg. Precision %	Subsets	PD	7	0	41	41.5	A-Bi(c)	47.7	**68.6***	41	48.1	48.5	54.8*	31.4	38.4	10^{-5}
			HPD	4	0	34.3	34.8	C-Bi(c)	46.4	1.5	35	42.8	38	**46.5***	35.2	40.1*	
			EPD	4	0	7.3	8.1	C-Bi(c)	62*	**69.5***	54.3	58.3	54.8*	64.3*	46.1	49.7	
			TD	7	0	17.2	16	A-Bi(c)	25	25.7	27.7	26.6	26.8	30.1	37.6	**46.4***	10^{-6}
			HTD	4	0	41.7	41.7	C-Bi(c)	45.2	47.7	44	47.2	43.2	47.5	44.5	**48.7**	
			ETD	4	0	32.3	19	C-Bi(c)	12.3	14.3	35.3	35.7	1.8	1.3	**45.8***	38.1*	
		Red. Subsets	RPD	7	0	29.6	29.5	A-Bi(c)	43.1*	**46.3***	40.1	41.4	42.7*	43.9*	40.8*	41.6	10^{-4}
			RHPD	4	0	31.8	31.4	C-Bi(c)	42.5*	43.7*	39.8	43.3*	40.2*	**43.9***	39.8	43.1	
			REPD	4	0	42.1	43	C-Bi(c)	58.3*	**58.6***	52.4	54.4	55.1*	56.6*	51.7	53.4	
			RTD	7	0	23.2	23.4	B-Bi(c)	44.6*	**52.8***	52.7	41.9	46.2*	44.1*	38.9	40.3	10^{-6}
			RETD	4	0	36.8	35	D-Bi(c)	62.1*	55.3*	49.5	**66**	51.1	54.3	60.9	55.3	

Table B.2. Comparison of the best Bi-modality models when applying Trivial, RF (RFc and RFr), and LSTMs (LSTM and LSTM-SW) regarding classification task with Micro average recall and Micro average F1-score measures. Reduced Subsets: Red. Subsets. DF: Decision Fusion. MF: Model Fusion. The colored cells indicate the best results in each model. MF: Model Fusion. The **bold** font indicate the best results. * p-value<0.05 when using paired t-test between RFc and LSTMs (LSTM and LSTM-SW).

Task	Measure	Model		n-Class	Trivial	RFc		Architecture (Loss =CCE)	LSTM				LSTM-SW				Learning rate
		Dataset				DF			DF		EF		DF		EF		
						FAD EDA-D	EMG-D EAD-D		FAD EDA-D	EMG-D EAD-D	FAD EDA-D	EMG-D EAD-D	FAD EDA-D	EMG-D EAD-D	FAD EDA-D	EMG-D EAD-D	
Classification	Micro avg. recall %	Subsets	PD	7	0	4.1	6.7	A-Bi(c)	5	5.2	12.4	13*	8.4*	9.1*	16.6	**16.9**	10^{-5}
			HPD	4	0	2.1	4.1	C-Bi(c)	8.5*	0.1	18.3*	16.3*	11.4*	10.4*	**22.8***	21.4*	
			EPD	4	0	53.2	45.7	C-Bi(c)	8.9	10.1	22	23.6*	13.1	16.9	29.2	**30.9**	
			TD	7	0	4.9	**8**	A-Bi(c)	0.2	0.8	1.7	3.4	2.2	3.4	5.4	5.9*	10^{-6}
			HTD	4	0	75.7	78.6	C-Bi(c)	98.1*	99.9*	87.9	92.9	**100***	**100***	97.4	97*	
			ETD	4	0	4.8	7.9	C-Bi(c)	2.7	0.8	4.1	6.7	0.4	0.2	6.5*	**8.6***	
		Red. Subsets	RPD	7	0	21.1	26.5	A-Bi(c)	58.6*	46.3*	71.7	68.5*	60.4*	57.1*	73.3*	**74.2**	10^{-4}
			RHPD	4	0	28.2	33	C-Bi(c)	68.4*	62.5*	91.5	91.6*	70.3*	69.1*	92.9	**95.1**	
			REPD	4	0	45.7	53.8	C-Bi(c)	78.4*	71.1*	87.6*	92	81.2*	83.8*	92.3*	**95.2**	
			RTD	7	0	23.9	28	B-Bi(c)	6.8	8.9	11.8*	15.3*	9.8	10.2	15.7*	**18.7**	10^{-6}
			RETD	4	0	27.7	32.7	D-Bi(c)	13.8	16.9	21.6	20.3	13.4	16.8	20.1	**24.3**	
	Micro avg. F1-score %	Subsets	PD	7	0	7.2	10.8	A-Bi(c)	8.2	8.9	16.9	18.6*	12.6*	13.9*	19.9	**20.8**	10^{-5}
			HPD	4	0	3.8	6.9	C-Bi(c)	13*	0.10	22.5	22.3*	15.7*	15.3*	26*	**26.3***	
			EPD	4	0	12.7	13.6	C-Bi(c)	14.3	15.8	27.7	30.2*	19.6	24.3	33	**32.6**	
			TD	7	0	7.6	**10.2**	A-Bi(c)	0.5	1.6	3.1	5.5	4	5.9	9.2	10.1	10^{-6}
			HTD	4	0	53.2	53.5	C-Bi(c)	61.4*	62.8*	57.8	60.7	58.1	62.9*	60.3	**63.3**	
			ETD	4	0	7.9	10.9	C-Bi(c)	3.7	1.4	6.8	10.8	0.7	0.3	10.8*	**13.7***	
		Red. Subsets	RPD	7	0	24.2	27.5	A-Bi(c)	49.5*	45.6*	51.4	51.5	49.7*	49.2*	52.3*	**53.2**	10^{-4}
			RHPD	4	0	29.5	31.8	C-Bi(c)	52.4*	51.2*	55.2	58.5*	51.1*	53.5*	55.5	**58.8**	
			REPD	4	0	43.5	47.6	C-Bi(c)	66.6*	63.7*	65.3	68.1	65.5*	67.2*	66	**68.2**	
			RTD	7	0	22.9	**25.2**	B-Bi(c)	11	14.1	18.6*	21.8*	15.3*	15.4	21*	24.8	10^{-6}
			RETD	4	0	30.4	31.9	D-Bi(c)	20.4	24.2	28.6	28.9	20.7	24.1	28.2	**34**	

APPENDIX C

Multi-modality Results

This chapter is dedicated to provide the detailed results of performing the proposed methods [RF and LSTMs (LSTM, and LSTM-SW)] with single modalities, two fused modalities, and all fused modalities regarding classification and regression for continuous monitoring pain intensity from the 11 proposed datasets on the X-ITE Pain Database.

Table C.1 and C.2 in Appendix C shows the comparison between the best models from Uni-, Bi-, and Multi-modality (models obtained after fusing all modalities). The results in Table C.1 emphasized that regression models were the best with imbalanced datasets, except with Heat Phasic Dataset [HPD], classification and regression models of EMG EDA Bi-modality that using LSTM-SW and regression model of FAD EDA Bi-modality perform similar (ICC = 0.41) (see Table 6.13 in Section 6.4.1.2). For more detailed results about classification versus regression, see Section 6.4.1. Multi-modality using LSTMs performed the best in terms of recall except with Reduced Heat Phasic Dataset [RHPD] and Reduced Electrical Phasic Dataset [REPD]; the EMG EDA Bi-modality and EDA Uni-modality models were the best, respectively. Further, in terms of F1-score, EDA Uni-modality models using LSTMs were the best when using Reduced Phasic Dataset [RPD] and [REPD], about 53.4% and 69%; RTD and RETD using RFc got 25.1%, and 36.8%. EMG EDA Bi-modality model using LSTMs was the best with PD, Heat Phasic Dataset [HPD], HTD, and RHPD, about 20.8%, 26.8%, 63.3%, and 58.8%. The Multi-modality models were the best with Electrical Phasic Dataset [EPD], TD, and ETD, about 34.8%, 11.6%, and 16.7%.

Table C.1. Comparison of the best Uni-, Bi-, and Multi-modality models using MF regarding classification and regression tasks with MSE and ICC measures, Bi- and Multi-modalities models using Model Fusion (MF). MF: Model Fusion. The grey cells indicate the best results.

Measure	Model			Trivial	RFc	LSTM			LSTM-SW			Regression	Trivial	RFr	LSTM			LSTM-SW			Loss			Learning rate
					Uni-modality	Uni-modality	Bi-modality	Multi-modality	Uni-modality	Bi-modality	Multi-modality			Uni-modality	Uni-modality	Bi-modality	Multi-modality	Uni-modality	Bi-modality	Multi-modality	CCE/BCE			
	Dataset		n-Class	-	EDA-D	EDA-D	EMG-D EDA-D	All modalities	EDA-D	EMG-D EDA-D	All modalities	Measure	-	EDA-D	EDA-D	EMG-D EDA-D	All modalities	EDA-D	EMG-D EDA-D	All modalities	Modality model architecture			
																					Uni-	Bi-	Multi-	
																					-	MF	MF	
MSE	Subsets	PD	7	0.10	0.09	0.09	0.08	0.08	0.09	0.08	0.08	MSE	0.10	0.07	0.06	**0.05**	0.06	0.08	0.06	0.06	A(c/r)	A-Bi(c/r)	A-Mu(c/r)	10^{-5}
		HPD	4	0.11	0.11	0.10	0.09	0.10	0.11	0.09	0.11		0.11	0.09	0.08	**0.07**	0.08	0.08	0.08	0.09	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		EPD	4	0.07	0.06	0.06	0.05	0.05	0.06	0.06	0.05		0.07	0.05	0.04	0.04	**0.04**	0.05	**0.04**	0.05	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		TD	7	0.12	0.16	0.12	0.11	0.11	0.12	0.11	0.10		0.12	0.13	0.08	0.08	0.09	0.09	0.10	**0.08**	A(c/r)	A-Bi(c/r)	A-Mu(c/r)	10^{-6}
		HTD	4	0.41	0.18	0.15	0.14	0.15	0.15	0.13	0.14		0.41	0.13	0.11	0.13	**0.10**	0.11	0.15	0.11	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		ETD	4	0.09	0.11	0.08	0.08	0.08	0.08	0.08	0.08		0.09	0.09	0.07	0.06	0.06	0.07	0.06	**0.06**	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
	Red. Subsets	RPD	7	0.23	0.14	0.05	0.06	0.05	0.05	0.05	0.05		0.23	0.09	0.04	0.04	0.04	0.04	**0.04**	0.06	A(c/r)	A-Bi(c/r)	A-Mu(c/r)	10^{-4}
		RHPD	4	0.26	0.22	0.08	0.07	0.08	0.07	0.07	0.08		0.26	0.13	0.05	0.05	0.13	0.05	**0.05**	0.08	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		REPD	4	0.26	0.12	0.05	0.05	0.06	0.05	0.05	0.06		0.26	0.08	0.04	0.04	0.05	**0.03**	0.04	0.06	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		RTD	7	0.25	0.18	0.21	0.20	0.20	0.19	0.19	0.19		0.25	0.13	0.11	0.13	**0.04**	0.11	0.13	0.12	B(c/r)	B-Bi(c/r)	B-Mu(c/r)	10^{-6}
		RETD	4	0.25	0.18	0.16	0.17	0.17	0.16	0.15	0.16		0.25	0.12	0.1	0.09	0.09	0.1	0.09	**0.09**	D(c/r)	D-Bi(c/r)	D-Mu(c/r)	
ICC	Subsets	PD	7	0	0.33	0.30	0.41	0.45	0.40	0.45	0.46	ICC	0	0.41	0.43	0.49	0.49	0.20	**0.51**	0.48	A(c/r)	A-Bi(c/r)	A-Mu(c/r)	10^{-5}
		HPD	4	0	0.11	0.30	0.37	0.39	0.29	**0.41**	0.38		0	0.20	0.28	0.39	0.38	0.32	0.40	0.40	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		EPD	4	0	0.37	0.36	0.49	0.49	0.50	0.53	0.57		0	0.47	0.49	0.57	**0.58**	0.53	**0.58**	**0.58**	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		TD	7	0	0.10	0.07	0.12	0.2	0.11	0.18	0.23		0	0.09	0.12	0.24	0.23	0.11	0.26	**0.30**	A(c/r)	A-Bi(c/r)	A-Mu(c/r)	10^{-6}
		HTD	4	0	0.3	0.33	0.38	0.35	0.31	**0.42**	0.33		0	0.31	0.28	0.21	0.38	0.30	0.23	0.35	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		ETD	4	0	0.14	0.09	0.18	0.2	0.09	0.22	0.26		0	0.17	0.09	0.25	0.31	0.21	0.31	**0.33**	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
	Red. Subsets	RPD	7	0	0.44	0.83	0.81	0.82	0.83	0.83	0.81		0	0.45	0.81	0.84	0.82	0.84	**0.85**	0.78	A(c/r)	A-Bi(c/r)	A-Mu(c/r)	10^{-4}
		RHPD	4	0	0.23	0.75	0.79	0.74	0.76	0.78	0.74		0	0.24	0.81	0.82	0.73	0.81	**0.83**	0.73	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		REPD	4	0	0.58	0.84	0.85	0.81	0.84	0.85	0.81		0	0.63	0.86	0.86	0.79	**0.88**	0.87	0.80	C(c/r)	C-Bi(c/r)	C-Mu(c/r)	
		RTD	7	0	0.21	0.25	0.28	0.27	0.31	0.32	0.28		0	0.18	0.23	**0.33**	0.29	0.24	0.32	0.26	B(c/r)	B-Bi(c/r)	B-Mu(c/r)	10^{-6}
		RETD	4	0	0.37	0.47	0.43	0.42	0.46	0.52	0.44		0	0.34	0.49	0.52	0.5	0.44	0.52	**0.56**	D(c/r)	D-Bi(c/r)	D-Mu(c/r)	

Table C.2. Comparison of the best Uni-, Bi-, and Multi-modality models using MF regarding classification, Bi- and Multi-modalities models using Model Fusion (MF). The grey cells indicate the best results. * p-value<0.05 when using paired t-test between RFc and LSTMs (LSTM and LSTM-SW).

Measure	Model			Trivial	RFc	LSTM			LSTM-SW			Measure	Trivial	RFc	LSTM			LSTM-SW			Loss			Learning rate
					Uni-modality	Uni-modality	Bi-modality	Multi-modality	Uni-modality	Bi-modality	Multi-modality			Uni-modality	Uni-modality	Bi-modality	Multi-modality	Uni-modality	Bi-modality	Multi-modality	CCE			
	Dataset		n-Class	-	EDA-D	EDA-D	EMG-D EDA-D	All modalities	EDA-D	EMG-D EDA-D	All modalities		-	EDA-D	EDA-D	EMG-D EDA-D	All modalities	EDA-D	EMG-D EDA-D	All modalities	Modality model architecture			
																					Uni-	Bi-	Multi-	
																					-	MF	MF	
Accuracy %	Subsets	PD	7	77.7	75.8	78.6*	79.1*	78.3	78.4*	78.3	77.2*	Micro avg. recall %	0	10.6	8	13*	16.2*	12.9*	16.9	17.8	A(c)	A-Bi(c)	A-Mu(c)	10^{-5}
		HPD	4	78.5	78.1	79.8*	80.5	79.3	79*	80.2*	77.6		0	3.4	9.9*	16.3*	19.8*	10.9*	21.4*	22.3*	C(c)	C-Bi(c)	C-Mu(c)	
		EPD	4	86.1	86.1	87.1*	87.8*	87.5	87*	87.2	87.3		0	14.8	15.6	23.6*	23.4	27*	30.9	32.1	C(c)	C-Bi(c)	C-Mu(c)	
		TD	7	70.3	50	70.7*	71.2	70.9	70.6*	71.6	71.6		0	16.7	1.6	3.4	6.5*	2.9	5.9*	7.5*	A(c)	A-Bi(c)	A-Mu(c)	10^{-6}
		HTD	4	20	41	48.4	47.4	41.6	47.7	49.8	42.2		0	71	94.6*	92.9	90.8	100*	97*	99.9	C(c)	C-Bi(c)	C-Mu(c)	
		ETD	4	82	74.3	82.4*	83.2	82.7	82.4*	83.2	83		0	13.8	3.3	6.7	9	3.1*	8.6*	11.6*	C(c)	C-Bi(c)	C-Mu(c)	
	Red. Subsets	RPD	7	50	43.2	67.3*	66.7	64.9	66.9*	67.3	64.1*		0	37.7	73.3*	68.5*	81*	78.4*	74.2	85.1*	A(c)	A-Bi(c)	A-Mu(c)	10^{-4}
		RHPD	4	50.1	45.1	68.6*	70.5*	67.5	68.6*	70.6	67.4		0	36.2	88*	91.6*	91.5	93.1*	95.1	94.1	C(c)	C-Bi(c)	C-Mu(c)	
		REPD	4	50	55.5	75.8*	75.3	71.4	75.1*	75.2	71.4*		0	59.5	92.7*	92	92.4	96.8*	95.2	92.4*	C(c)	C-Bi(c)	C-Mu(c)	
		RTD	7	38.1	30.4	42.2*	44.4	43.3	42.7*	45.2	41.2		0	38.2	8.9	15.3*	15.4*	10.5	18.7	20.8*	B(c)	B-Bi(c)	B-Mu(c)	10^{-6}
		RETD	4	49	48.7	57.7*	57.2	55	57.5*	56.6	54.7		0	41	20.9	20.3	20.6	20.6	24.3	29.4	D(c)	D-Bi(c)	D-Mu(c)	
Micro avg. precision %	Subsets	PD	7	0	24.9	53.3*	48.1	36*	43.7*	38.4	28.7*	Micro avg. F1-score %	0	14.2	13	18.6*	19.5*	18.4*	20.8	20	A(c)	A-Bi(c)	A-Mu(c)	10^{-5}
		HPD	4	0	24.6	36.6*	42.8	34.9	32.2*	40.1*	29.2		0	5.9	15.2*	22.3*	23.6*	15.5*	26.3*	24*	C(c)	C-Bi(c)	C-Mu(c)	
		EPD	4	0	37.5	56.3*	58.3	48.6	47.9*	49.7	44.5		0	20.5	21.3	30.2*	28.5	31.2*	32.6	34.8	C(c)	C-Bi(c)	C-Mu(c)	
		TD	7	0	10.2	23.2*	26.6	32.7	25.8	46.4*	36.3		0	12.2	2.9	5.5	10.2*	5.1	10.1*	11.6*	A(c)	A-Bi(c)	A-Mu(c)	10^{-6}
		HTD	4	0	42.7	48.2	47.2	41.8	47.7	48.7	42		0	52.9	62.3*	60.7	56	62.5*	63.3	57.9	C(c)	C-Bi(c)	C-Mu(c)	
		ETD	4	0	15.2	28.6	35.7	38.6	35.7	38.1*	41.7*		0	14.1	5.6	10.8	13	5.2*	13.7*	16.7*	C(c)	C-Bi(c)	C-Mu(c)	
	Red. Subsets	RPD	7	0	20.1	41.8*	41.4	36.3*	40.5*	41.6	35.4*		0	25.7	53.2*	51.5	50.1	53.4*	53.2	49.9*	A(c)	A-Bi(c)	A-Mu(c)	10^{-4}
		RHPD	4	0	24.2	40.2*	43.3*	38.8	39.6*	43.1	38.2		0	28.8	55*	58.5*	54.3	55.4*	58.8	54.1	C(c)	C-Bi(c)	C-Mu(c)	
		REPD	4	0	35.6	55.2*	54.4	47.1*	53.9*	53.4	47.1*		0	44.3	69*	68.1	62.2*	69*	68.2	62.2*	C(c)	C-Bi(c)	C-Mu(c)	
		RTD	7	0	19.1	44.6*	41.9	39.2	40.8*	40.3	28.8		0	25.1	14.3	21.8*	21*	16.2	24.8	23.3*	B(c)	B-Bi(c)	B-Mu(c)	10^{-6}
		RETD	4	0	35.5	61.6*	66	52.7	61.9*	55.3	48.2		0	36.8	30.2	28.9	26.6	29.7	34	33.4	D(c)	D-Bi(c)	D-Mu(c)	

Bibliography

[1] S. N. Raja, D. B. Carr, M. Cohen, N. B. Finnerup, H. Flor, S. Gibson, F. J. Keefe, J. S. Mogil, M. Ringkamp, K. A. Sluka, X. Song, B. Stevens, M. D. Sullivan, P. R. Tutelman, T. Ushida, and K. Vader, “The revised international association for the study of pain definition of pain: Concepts, challenges, and compromises,” *Pain*, vol. 161, no. 9, pp. 1976–1982, 2020.

[2] A. C. de C. Williams and K. D. Craig, “Updating the definition of pain,” *Pain*, vol. 157, no. 11, pp. 2420–2423, 2016.

[3] D. C. Turk, “Cognitive-behavioral perspective of pain,” *Encyclopedia of Pain, Springer Berlin Heidelberg*, p. 405–408, 2007.

[4] R. D. Treede, D. R. Kenshalo, R. H. Gracely, and A. K. P. Jones, “The cortical representation of pain,” vol. 79, no. 2-3, pp. 105–111, 1999.

[5] E. J. Dayoub and A. B. Jena, “Does pain lead to tachycardia revisiting the association between self-reported pain and heart rate in a national sample of urgent emergency department visits,” *Mayo Clinic Proceedings*, vol. 90, no. 8, pp. 1165–1166, 2015.

[6] P. Thiam, H. Kestler, and F. Schwenker, “Multimodal deep denoising convolutional autoencoders for pain intensity classification based on physiological signals,” in *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, (Valletta, Malta), p. 289–296, 22–24 February 2020.

[7] E. Othman, F. Saxen, D. Bershadskyy, P. Werner, A. Al-Hamadi, and J. Weimann, “Predicting group contribution behaviour in a public goods game from face-to-face communication,” *Sensors*, vol. 19, no. 12, p. 2786, 2019.

[8] S. Shakya, S. Sharma, and A. Basnet, “Human behavior prediction using facial expression analysis,” in *Proceedings of the International Conference on*

Computing, Communication and Automation (ICCCA), (Greater Noida, India), 29-30 April 2016.

[9] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.

[10] A. R. Dores, F. Barbosa, C. Queirós, I. P. Carvalho, and M. D. Griffiths, "Recognizing emotions through facial expressions: A largescale experimental study," *International Journal of Environmental Research and Public Health (Int J Environ Res Public Health)*, vol. 17, no. 20, p. 7420, 2020.

[11] J. M. Oosterman, S. Zwakhalen, E. L. Sampson, and M. Kunz, "The use of facial expressions for pain assessment purposes in dementia: A narrative review," *Neurodegenerative Disease Management*, vol. 6, no. 2, pp. 119–31, 2016.

[12] M. V. Ciolacu, "Facial expressions and non verbal comunication," *Procedia - Social and Behavioral Sciences*, vol. 127, pp. 878–882, 2014.

[13] K. D. Craig, "The facial expression of pain," *Science Direct*, 1992.

[14] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, 2008.

[15] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, vol. 1. CA, USA: Consulting Psychologists Press, 1978.

[16] A. Corbett, W. Achterberg, B. Husebo, F. Lobbezoo, H. de Vet, M. Kunz, L. Strand, M. Constantinou, C. Tudose, J. Kappesser, M. de Waal, and S. Lautenbacher, "An international road map to improve pain assessment in people with impaired cognition: The development of the pain assessment in impaired cognition (paic) meta-tool," *BMC Neurology*, vol. 14, no. 1, p. 229, 2014.

[17] K. G. Snoek, M. Timmers, R. Albertyn, and M. van Dijk, "Pain indicators for persisting pain in hospitalized infants in a south african setting: An explorative study," *Journal of Pain Palliative Care Pharmacotherapy*, vol. 29, no. 2, pp. 125–132, 2015.

[18] G. Anbarjafari, P. Rasti, F. Noroozi, J. Gorbova, and R. E. Haamer, *Machine Learning for Face, Emotion, and Pain Recognition*, vol. SL37. Washington: Society of Photo-optical Instrumentation Engineers, 2018.

[19] D. J. Cipher and A. Clifford, "Dementia, pain, depression, behavioral disturbances, and adls Toward a comprehensive conceptualization of quality of life in long-term care," *International Journal of Geriatric Psychiatry*, vol. 19, no. 8, pp. 741–748, 2004.

[20] M. D. nas, B. Ojeda, A. Salazar, J. A. Mico, and I. Failde, "A review of chronic pain impact on patients, their social environment and the health care system," *Journal of Pain Research*, vol. 9, no. 1, p. 457–467, 2016.

[21] M. T. Bobo, *Pain Assessment Strategies for People with Cognitive Impairment in Nursing Home Settings*. Thesis, 2020.

[22] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain recognition from video and biomedical signals," in *Proceedings of the 22nd International Conference on Pattern Recognition*, (Stockholm, Sweden), 24-28 August 2014.

[23] A. Ampt, J. Westbrook, N. Creswick, and N. Mallock, "A comparison of self-reported and observational work sampling techniques for measuring time in nursing tasks," *Journal of Health Services Research and Policy*, vol. 12, no. 1, pp. 18–24, 2007.

[24] K. D. Craig, "The social communication model of pain," *Canadian Psychology*, vol. 50, no. 1, pp. 22–32, 2009.

[25] K. D. Craig, "The facial expression of pain better than a thousand words," *APS Journal*, vol. 1, no. 3, pp. 153–162, 1992.

[26] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. W. Picard, "Automatic recognition methods supporting pain assessment: A survey," *IEEE Transactions on Affective Computing*, pp. 1–23, 2019.

[27] A. C. de C. Williams, "Facial expression of pain: An evolutionary account," *Behavioral and Brain Sciences*, vol. 25, no. 4, pp. 439–455, 2002.

[28] D. Naranjo-Hernández, J. Reina-Tosina, and L. M. Roa, "Sensor technologies to manage the physiological traits of chronic pain: A review," *Sensors*, vol. 20, no. 2, p. 365, 2020.

[29] P. Werner, A. Al-Hamadi, S. Gruss, and S. Walter, "Twofold-multimodal pain recognition with the X-ITE pain database," in *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, (Cambridge, UK), 3-6 September 2019.

[30] S. Walter, A. Al-Hamadi, S. Gruss, S. Frisch, H. C. Traue, and P. Werner, "Multimodale erkennung von schmerzintensität und-modalität mit maschinellen lernverfahren," *Der Schmerz*, vol. 34, no. 5, p. 400–409, 2020.

[31] D. L. Martinez, O. Rudovic, and R. Picard, "Personalized automatic estimation of self-reported pain intensity from facial expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Honolulu, HI, USA), pp. 2318–2327, 21-26 July 2017.

[32] K. M. Prkachion and K. D. Craig, "Expressing pain: The communication and interpretation of facial pain signals," *Journal of Nonverbal Behavior volume*, vol. 19, no. 4, p. pages191–205, 1995.

[33] Q. Ye, J. Zhou, and H. Wu, "Using information technology to manage the covid-19 pandemic: Development of a technical framework based on practical experience in china," *JMIR Medical Informatics*, vol. 8, no. 6, p. e19515, 2020.

[34] K. M. Prkachin, "Assessing pain by facial expression: Facial expression as nexus," *Pain Research and Management*, vol. 14, no. 1, p. 53–58, 2009.

[35] J. Thevenot, M. B. Lòpez, and A. Hadid, "A survey on computer vision for assistive medical diagnosis from faces," *IEEE Journal of Biomedical and Health*, vol. 22, no. 5, pp. 1497 – 1511, 2018.

[36] T. Hassan, D. Seus, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J. Garbas, and U. Schmid, "Automatic detection of pain from facial expressions: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, 2021.

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

[38] K. M. Prkachin and S. R. Mercer, "Pain expression in patients with shoulder pathology: Validity, properties and relationship to sickness impact," *Pain*, vol. 39, no. 3, pp. 257–265, 1989.

[39] M. Schiavenato, "Facial expression and pain assessment in the pediatric patient: the primal face of pain," *Journal for Specialists in Pediatric Nursing*, vol. 13, no. 2, pp. 89–97, 2008.

[40] M. Tavakolian, *Efficient Spatiotemporal Representation Learning for Pain Intensity Estimation from Facial Expressions*. Thesis, 2021.

[41] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, and H. C. Traue, "Head movements and postures as pain behavior," *PLoS ONE*, vol. 13, no. 2, p. e0192767, 2018.

[42] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[43] D. Ververidis and C. Kotropoulos, "Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition," *Signal Processing*, vol. 88, no. 12, pp. 2956–2970, 2008.

[44] B. Farnsworth, "Facial action coding system (FACS) – a visual guidebook. IMOTIONS. https://imotions.com/blog/facial-action-coding-system/," 2022.

[45] S. Brahnam, C. Chuang, F. Y. Shih, and M. R. Slack, "Machine recognition and representation of neonatal facial displays of acute pain," *Artificial Intelligence in Medicine*, vol. 36, no. 3, pp. 211–222, 2006.

[46] S. Brahnam, C. Chuang, F. Y. Shih, and M. R. S. Melinda, "Svm classification of neonatal facial images of pain," in *Proceedings of the 6th International Workshop on Fuzzy Logic and Applications (WILF)*, (Crema, Italy), 15-17 September 2005.

[47] S. Mikat, G. Rätsch, J. Weston, B. Scholkopft, and K. Müller, "Fisher discriminant analysis with kernels," in *Proceedings of the IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*, (Madison,WI,USA), p. 41–48, 25-25 August 1999.

[48] C. Cortes and V. Vapnik, "Support-vector networks machine learning," *Machine learning*, vol. 20, no. 3, p. 273–297, 1995.

[49] S. Brahnam, L. Nanni, and R. Sexton, *Introduction to Neonatal Facial Pain Detection using Common and Advanced Face Classication Techniques*, vol. 2, pp. 225–253. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[50] B. Gholami, W. M. Haddad, and A. R. Tannenbaum, "Relevance vector machine learning for neonate pain intensity assessment using digital imaging," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 6, pp. 1457 – 1466, 2010.

[51] L. Nanni, A. Lumini, and S. Brahnam, "Local binary patterns variants as texture descriptors for medical image analysis," *Artificial Intelligence in Medicine*, vol. 49, no. 2, pp. 117–125, 2010.

[52] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[53] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635 – 1650, 2010.

[54] S. Liao and A. C. S. Chung, "Face recognition by using elongated local binary patterns with average maximum distance gradient magnitude," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, (Tokyo, Japan), pp. 672–679, 18-22 November 2007.

[55] L. Nanni, S. Brahnam, and A. Lumini, "A local approach based on a local binary patterns variant texture descriptor for classifying pain states," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7888–7894, 2010.

[56] L. Celona and L. Manoni, "Neonatal facial pain assessment combining handcrafted and deep features," in *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, (Catania, Italy), pp. 197–204, 11-15 September 2017.

[57] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPŔ05)*, (San Diego, CA, USA), 20-25 June 2005.

[58] D. Liu, F. Peng, A. Shea, O. Rudovic, and R. Picard, "Deepfacelift: Interpretable personalized models for automatic estimation of self-reported pain," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Honolulu, HI, USA), 21-26 July 2017.

[59] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–11, 2017.

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

[61] M. Kächele, P. Thiam, M. Amirian, P. Werner, S. Walter, F. Schwenker, and G. Palm, "Multimodal data fusion for person-independent, continuous estimation of pain intensity," in *Engineering Applications of Neural Networks: 16th International Conference*, (Rhodes, Greece), 25-28 September 2015.

[62] D. Erekat, Z. Hammal, M. Siddiqui, and H. Dibeklioğlu, "Enforcing multilabel consistency for automatic spatio-temporal assessment of shoulder pain intensity," in *Proceedings of the International Conference on Multimodal Interaction*

(ICMI 2́0 Companion), (Utrecht, the Netherlands), p. 156–164, 25-29 October 2020.

[63] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proceedings of the 3rd International Conference on Image and Signal Processing*, (Octeville, France), 1-3 July 2008.

[64] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Proceedings of the 21st International Conference on Pattern Recognition*, (Tsukuba, Japan), pp. 1363–1366, 11-15 November 2012.

[65] R. Yang, S. Tong, M. Bordallo, E. Boutellaa, J. Peng, X. Feng, and A. Hadid, "On pain assessment from facial videos using spatio-temporal local descriptors," in *Proceedings of the 6th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, (Oulu, Finland), 12-15 December 2016.

[66] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *Proceedings of the 9th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, (Santa Barbara, CA, USA), p. 314–321, 21-23 March 2011.

[67] S. R. Arashloo and J. Kittler, "Dynamic texture recognition using multiscale binarized statistical image features," *IEEE Transactions on Multimedia*, vol. 16, no. 8, p. 2099–2109, 2014.

[68] J. Chen, Z. Chi, and H. Fu, "A new framework with multiple tasks for detecting and locating pain events in video," *Computer Vision and Image Understanding*, vol. 155, pp. 113–123, 2017.

[69] P. Thiam, V. Kessler, and F. Schwenker, "Hierarchical combination of video features for personalised pain level recognition," in *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, (Bruges, Belgium), p. 465–470, 26-28 April 2017.

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

[71] G. Zamzami, G. Ruiz, D. Goldgof, R. Kasturi, Y. Sun, and T. Ashmeade, "Pain assessment in infants: Towards spotting pain expression based on infants' facial strain," in *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (Ljubljana, Slovenia), 4-8 May 2015.

[72] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 1998.

[73] E. Fotiadou, S. Zinger, W. T. a. Ten, S. Bambang-Oetomo, and P. H. N. de With, "Video-based facial discomfort analysis for infants," in *Proceedings of the SPIE 9029, Visual Information Processing and Communication V*, vol. 9029, (San Francisco, CA, USA), 19 February 2014.

[74] K. Sikka, A. A. Ahmed, D. Diaz, M. S. Goodwin, K. D. Craig, M. S. Bartlett, and J. S. Huang, "Automated assessment of children's postoperative pain using computer vision," *Pediatrics*, vol. 136, no. 1, pp. 124–131, 2015.

[75] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, S. Gruss, and S. Walter, "Automatic vs. human recognition of pain intensity from facial expression on the X-ITE pain database," *Sensors*, vol. 21, no. 9, p. 3273, 2021.

[76] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, and S. Walter, "Cross-database evaluation of pain recognition from facial video," in *Proceedings of the 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, (Dubrovnik, Croatia), 23-25 September 2019.

[77] P. Thiam, V.Kessler, M. Amirian, P. Bellmann, G. Layher, Y. Zhang, M. Velana, S. Gruss, S. Walter, H. C. Traue, J. Kim, D. Schork, E. Andre, H. Neumann, and F. Schwenker, "Multi-modal pain intensity recognition based on the senseemotion database," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 743 – 760, 2019.

[78] P. Werner, A. Al-Hamadi, and S. Walter, "Analysis of facial expressiveness during experimentally induced heat pain," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, (San Antonio, TX, USA), 23-26 October 2017.

[79] M. A. Haque, R. B. Bautista, F. Noroozi, K. Kulkarni, C. B. Laursen, R. Irani, M. Bellantonio, S. Escalera, G. Anbarjafari, K. Nasrollahi, O. K. Andersen, E. G. Spaich, and T. B. Moeslund, "Deep multimodal pain recognition: A database and comparison of spatio-temporal visual modalities," in *Proceedings of the 3th IEEE International Conference on Automatic Face Gesture Recognition(FG)*, (Xián, China), 15-19 May 2018.

[80] F. Wang, X. Xiang, C. Liu, T. D. Tran, A. Reiter, G. D. Hager, H. Quon, J. Cheng, and A. L. Yuille, "Regularizing face verification nets for pain intensity regression," in *Proceedings of the 24th IEEE International Conference on Image Processing (ICIP)*, (Beijing, China), 17-20 September 2017.

[81] Y. Lecun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, (Paris, France), 30 May - 2 June 2010.

[82] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, p. 1735–1780, 1997.

[83] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Las Vegas, Nevada, USA), 26 June-1 July 2016.

[84] N. Kalischek, P. Thiam, P. Bellmann, and F. Schwenker, "Deep domain adaptation for facial expression analysis," in *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, (Cambridge, United Kingdom, UK), pp. 317–323, 23-6 September 2019.

[85] G. Bargshady, J. Soar, X. Zhou, R. C. Deo, F. Whittaker, and H. Wang, "A joint deep neural network model for pain recognition from face," in *Proceedings of the 4th IEEE International Conference on Computer and Communication Systems (ICCCS)*, (Singapore), 23-25 February 2019.

[86] J. Soar, G. Bargshady, X. Zhou, and F. Whittaker, "Deep learning model for detection of pain intensity from facial expression," in *Proceedings of the International Conference on Smart Homes and Health Telematics*, (Singapore), pp. 249–254, 10-12 July 2018.

[87] P. Thiam, H. A. Kestler, and F. Schwenker, "Two-stream attention network for pain recognition from video sequences," *Sensors*, vol. 20, no. 3, p. 839, 2020.

[88] K. S. Feldt, "The checklist of nonverbal pain indicators (cnpi)," *Pain Manag Nurs*, vol. 1, no. 1, pp. 13–21, 2000.

[89] S. J. Waters, P. A. Riordan, F. J. Keefe, and J. C. Lefebvre, "Pain behavior in rheumatoid arthritis patients: Identification of pain behavior subgroups," *Journal of Pain and Symptom Management*, vol. 36, no. 1, 2008.

[90] G. Zamzmi, *Automatic Multimodal Assessment of Neonatal Pain*. Thesis, 2018.

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

[92] G. Zamzmi, R. Kasturi, D. Goldgof, R. Zhi, T. Ashmeade, and Y. Sun, "A review of automated pain assessment in infants: Features, classification tasks, and databases," *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 77 – 96, 2017.

[93] F. Tsai, Y. Hsu, W. Chen, Y. Weng, C. Ng, and C. Lee, "Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, (San Francisco, CA, USA), 8–12 September 2016.

[94] F. Tsai, Y. Weng, C. Ng, and C. Lee, "Embedding stacked bottleneck vocal features in a lstm architecture for automatic pain level classification during emergency triage," in *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction (ACII)*, (San Antonio, Texas, USA), 23-26 October 2017.

[95] J. Li, Y. Weng, C. Ng, and C. Lee, "Learning conditional acoustic latent representation with gender and age attributes for automatic pain level recognition," in *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*, (Hyderabad, India), 2-6 September 2018.

[96] F. Anders, M. Hlawitschka, and M. Fuchs, "Automatic classification of infant vocalization sequences with convolutional neural networks," *Mirco Fuchs*, vol. 119, pp. 36–45, 2020.

[97] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[98] V. A. Howard and F. W. F. W. Thurber, "The interpretation of infant pain: Physiological and behavioral indicators used by nicu nurses," *Journal of Family Nursing*, vol. 13, no. 3, pp. 164–74, 1998.

[99] T. Olugbade, M. S. H. Aung, N. Bianchi-Berthouze, N. Marquardt, and A. C. de C. Williams, "Bi-modal detection of painful reaching for chronic pain rehabilitation systems," in *Proceedings of the International Conference on Multimodal Interaction (IACMI)*, (Istanbul, Turkey), 12-16 Novmber 2014.

[100] T. A. Olugbade, N. Bianchi-Berthouze, N. Marquardt, and A. C. Williams, "Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, (Xi'an, China), 21-24 September 2015.

[101] S. Ashouri, M. Abedi, M. Abdollahi, F. D. Manshadi, M. Parnianpour, and K. Khalaf, "A novel approach to spinal 3-d kinematic assessment using inertial sensors: Towards effective quantitative evaluation of low back pain in clinical settings," *Computers in Biology and Medicine*, vol. 89, 2017.

[102] C. Hung, T. Shen, C. Liang, and W. Wu, "Using surface electromyography (semg) to classify low back pain based on lifting capacity evaluation with principal component analysis neural network method," in *Proceedings of the IEEE Engineering in Medicine and Biology Society*, (Chicago, Illinois, USA), pp. 18–21, 26-30 August 2014.

[103] G. Zamzmi, C. Pai, D. Goldgof, R. Kasturi, Y. Sun, and T. Ashmeade, "Automated pain assessment in neonates," in *Proceedings of the Scandinavian Conference on Image Analysis*, (Tromsø, Norway), pp. 350–361, 12–14 June 2017.

[104] L. Brown, "Physiologic responses to cutaneous pain in neonates," *Neonatal Network*, vol. 6, no. 3, pp. 18–22, 1987.

[105] J. Greisen, C. B. Juhl, T. Grøfte, H. Vilstrup, T. S. Jensen, and O. Schmitz, "Acute pain induces insulin resistance in humans," *Anesthesiology*, vol. 95, no. 3, pp. 578–584, 2001.

[106] B. J. Stevens and C. C. Johnston, "Physiological responses of premature infants to a painful stimulus," *Nursing Research*, vol. 43, no. 4, pp. 226–231, 1994.

[107] S. Moscato, P. Cortelli, and C. Lorenzo, "Physiological responses to pain in cancer patients: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 2017, no. 4, 2022.

[108] M. Saccò, M. Meschi, G. Regolisti, S. Detrenis, L. Bianchi, M. Bertorelli, S. Pioli, A. Magnano, F. Spagnoli, P. G. Giuri, E. Fiaccadori, and A. Caiazza, "The relationship between blood pressure and pain," *Journal of Clinical Hypertension*, vol. 15, no. 8, p. 600–605, 2013.

[109] G. C. Littlewort, M. S. Bartletta, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1797–1803, 2009.

[110] P. Lucey, J. F. Cohn, K. M. Prkachind, P. E. Solomon, S. Chewf, and I. Matthews, "Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database," *Image and Vision Computing*, vol. 30, no. 3, pp. 197–205, 2012.

[111] D. Borsook, E. A. Moulton, K. F. Schmidt, and L. R. Becerra, "Neuroimaging revolutionizes therapeutic approaches to chronic pain," *Molecular Pain*, vol. 3, no. 1, p. 25, 2007.

[112] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade, and G. M. da Silva, "The biovid heat pain database: Data for the advancement and systematic validation of an automated pain recognition system," in *Proceedings of the IEEE International Conference on Cybernetics (CYBCO)*, (Lausanne, Switzerland), 13-15 June 2013.

[113] F. Pouromran, S. Radhakrishnan, and S. Kamarthi, "Exploration of physiological sensors, features, and machine learning models for pain intensity estimation," *PLoS One*, vol. 16, no. 7, p. e0254108, 2021.

[114] J. O. Egede, *Automatic Pain Assessment from Face Video (Continuous Pain Intensity Estimation in Adults and Newborns)*. Thesis, 2018.

[115] M. Odhner, D. Wegman, N. Freeland, A. Steinmetz, and G. L. Ingersoll, "Assessing pain control in nonverbal critically ill adults," *Dimensions of Critical Care Nursing*, vol. 22, no. 6, p. 260–267, 2003.

[116] S. Hinduja, S. Canavan, and G. Kaur, "Multimodal fusion of physiological signals and facial action units for pain recognition," in *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, (Buenos Aires, Argentina), 16-20 November 2020.

[117] P. M. Aslaksen, I. N. Myrbakk, R. S. Høifødt, and M. A. Flaten, "The effect of experimenter gender on autonomic and subjective responses to pain stimuli," *Pain*, vol. 129, no. 3, pp. 260–268, 2006.

[118] J. Koenig, M. N. Jarczok, R. J. Ellis, T. K. Hillecke, and J. F. Thayer, "Heart rate variability and experimentally induced pain in healthy adults: A systematic review," *European Journal of Pain*, vol. 18, no. 3, pp. 301–314, 2014.

[119] H. Storm, "Changes in skin conductance as a tool to monitor nociceptive stimulation and pain," *Current Opinion in Anaesthesiology*, vol. 12, no. 6, pp. 796–804, 2008.

[120] T. Ledowski, J. Bromilow, M. J. Paech, H. Storm, R. Hacking, and S. A. Schug, "Monitoring of skin conductance to assess postoperative pain intensity," *British Journal of Anaesthesia*, vol. 97, no. 6, pp. 862–865, 2006.

[121] M. L. Loggia, M. Juneau, and C. M. Bushnell, "Autonomic responses to heat pain: Heart rate, skin conductance, and their relation to verbal ratings and stimulus intensity," *Pain*, vol. 152, no. 3, 2011.

[122] R. Treister, M. Kliger, G. Zuckerman, I. G. Aryeh, and E. Eisenberg, "Differentiating between heat pain intensities: The combined effect of multiple autonomic parameters," *Pain*, vol. 153, no. 9, 2012.

[123] Y. Chu, X. Zhao, J. Yao, Y. Zhao, and Z. Wu, "Physiological signals based quantitative evaluation method of the pain," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 2981–2986, 2014.

[124] Y. Chu, X. Zhao, J. Han, and Y. Su, "Physiological signal-based method for measurement of pain intensity," *Frontiers in Neuroscience*, vol. 11, p. 279, 2017.

[125] S. Walter, S. Gruss, K. Limbrecht-Ecklundt, and H. C. Traue, "Automatic pain quantification using autonomic parameters," *Psychology and Neuroscience*, vol. 7, no. 3, pp. 363–380, 2014.

[126] S. Gruss, R. Treister, P. Werner, H. C. Traue, S. Crawcour, A. Andrade, and S. Walter, "Pain intensity recognition rates via biopotential feature patterns with support vector machines," *PLoS ONE*, vol. 10, no. 10, p. e0140330, 2015.

[127] M. Kächele, P. Thiam, M. Amirian, F. Schwenker, and G. Palm, "Methods for person-centered continuous pain intensity assessment from bio-physiological channels," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 5, pp. 854 – 864, 2016.

[128] D. Lopez-Martinez and R. Picard, "Multi-task neural networks for personalized pain recognition from physiological signals," in *Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, (San Antonio, TX, USA), 23-26 October 2017.

[129] M. Kächele, M. Amirian, P. Thiam, P. Werner, S. Walter, G. Palm, and F. Schwenker, "Adaptive confidence learning for the personalization of pain intensity estimation systems," *Evolving Systems*, vol. 8, no. 1, p. 71–83, 2017.

[130] P. Thiam and F. Schwenker, "Multi-modal data fusion for pain intensity assessment and classification," in *Proceedings of the 7th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, (Montreal, QC, Canada), 28 November -1 December 2017.

[131] M. Amirian, M. Kächele, and F. Schwenker, "Using radial basis function neural networks for continuous and discrete pain estimation from biophysiological signals," in *Proceedings of the IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, (Ulm, Germany), 28–30 September 2016.

[132] S. D. Subramaniam and B. Dass, "Automated nociceptive pain assessment using physiological signals and a hybrid deep learning network," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 3335 – 3343, 2021.

[133] D. Lopez-Martinez and R. Picard, "Continuous pain intensity estimation from autonomic signals with recurrent neural networks," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, (Honolulu, Hawaii, USA), pp. 5624–5627, 17-21 July 2018.

[134] P. Thiam, P. Bellmann, H. A. Kestler, and F. Schwenker, "Exploring deep physiological models for nociceptive pain recognition," *Sensors*, vol. 19, no. 20, p. 4503, 2019.

[135] H. F. Posada-Quintero, Y. Kong, and K. H. Chon, "Objective pain stimulation intensity and pain sensation assessment using machine learning classification and regression based on electrodermal activity," *Am J Physiol Regul Integr Comp Physiol*, vol. 321, no. 2, 2021.

[136] Y. Kong, H. F. Posada-Quintero, and K. H. Chon, "Sensitive physiological indices of pain based on differential characteristics of electrodermal activity," *IEEE Trans Biomed Eng*, vol. 68, no. 10, 2021.

[137] V. Bhatkar, R. Picard, and C. Staahl, "Combining electrodermal activity with the peak-pain time to quantify three temporal regions of pain experience," *Front. pain res. (Lausanne)*, vol. 3, no. 764128, 2022.

[138] S. Walter, S. Gruss, H. Traue, P. Werner, A. Al-Hamad, M. Kächele, F. Schwenker, A. Andrade, and G. Moreira, "Data fusion for automated pain recognition," in *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, (Istanbul, Turkey), 20-23 May 2015.

[139] G. Zamzmi, C. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "An approach for automated multimodal analysis of infants' pain," in *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, (Cancun, Mexico), 4-8 December 2016.

[140] P. Thiam, H. Hihn, D. A. Braun, H. A. Kestler, and F. Schwenker, "Multimodal pain intensity assessment based on physiological signals: A deep learning perspective," *Front Physiology*, vol. 1, no. 12, 2021.

[141] M. Yu, Y. Sun, B. Zhu, L. Zhu, Y. Lin, X. Tang, Y. Guo, G. Sun, and M. Dong, "Diverse frequency band-based convolutional neural networks for tonic cold pain assessment using eeg," *Neurocomputing*, vol. 378, no. 2020, pp. 270–282, 2020.

[142] R. Wang, K. Xu, H. Feng, and W. Chen, "Hybrid rnn-ann based deep physiological network for pain recognition," in *Proceedings of the 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, (Montreal, QC, Canada), 20-24 July 2020.

[143] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, (Santa Barbara, CA, USA), p. 57–64, 21-25 March 2011.

[144] X. Zhanga, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[145] D. Harrison, M. Sampson, J. Reszel, K. Abdulla, N. Barrowman, J. Cumber, A. Fuller, C. Li, S. Nicholls, and C. M. Pound, "Too many crying babies: A systematic review of pain management practices during immunizations on youtube," *BMC Pediatrics*, vol. 14, no. 134, 2014.

[146] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Las Vegas, Nevada, USA), 27-30 June 2016.

[147] V. K. Mittal, "Discriminating the infant cry sounds due to pain vs. discomfort towards assisted clinical diagnosis," in *Proceedings of the SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies*, (San Francisco, CA, USA), 13 September 2016.

[148] M. Velana, S. Gruss, G. Layher, P. Thiam, Y. Zhang, D. Schork, V. Kessler, S. Meudt, H. Neumann, J. Kim, F. Schwenker, E. Andrè, H. C. Traue, and S. Walter, "The senseemotion database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system," in *Proceedings of the IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*, (Cancun, Mexico), pp. 127–139, 4 December 2016.

[149] S. Gruss, M. Geiger, P. Werner, O. Wilhelm, H. C. Traue, A. Al-Hamadi, and S. Walter, "Multimodal signals for analyzing pain responses to thermal and electrical stimuli," *Journal of Visualized Experiments : JoVE (J Vis Exp.)*, vol. 146, p. e59057, 2019.

[150] M. S. Salekin, G. Zamzmi, J. Hausmann, D. Goldgof, R. Kasturi, M. Kneusel, T. Ashmeade, T. Ho, and Y. Suna, "Multimodal neonatal procedural and postoperative pain assessment dataset," *Data in Brief*, vol. 35, no. 3, p. 106796, 2021.

[151] P. Thiam and F. Schwenker, "Combining deep and hand-crafted features for audio-based pain intensity classification," in *Proceedings of the IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction*, (Bejing, China), pp. 49–58, 20 August 2018.

[152] D. Hudson-Barr, B. Capper-Michel, S. Lambert, T. M. Palermo, K. Morbeto, and S. Lombardo, "Validation of the pain assessment in neonates (pain) scale with the neonatal infant pain scale (nips)," *Neonatal Network*, vol. 21, no. 6, p. 15–22, 2002.

[153] P. Hummel, M. Puchalski, S. D.Creech, and M. C. Weiss, "Clinical reliability and validity of the npass: Neonatal pain, agitation and sedation scale with prolonged pain," *Journal of Perinatology*, vol. 28, no. 1, pp. 55–60, 2008.

[154] T. Baltrušaitis, P. Robinson, and L. Morency, "Openface: An open source facial behavior analysis toolkit," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, (Lake Placid, NY, USA), 7-10 March 2016.

[155] T. Baltrušaitis, P. Robinson, and L. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the International Conference on Computer Vision Workshops*, (Sydney, NSW, Australia), 2-8 December 2013.

[156] R. G. Matthews, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *Proceedings of the 8th IEEE International Conference on Automatic Face Gesture Recognition*, (Amsterdam, Netherlands), 17-19 September 2008.

[157] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. K. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[158] V. Le, J. Brandt, Z. Lin, L. D. Bourdev, and L. D. Bourdev, "Interactive facial feature localization," in *Proceedings of the European Conference on Computer Vision*, (Florence, Italy), 7-13 October 2012.

[159] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, p. 200–215, 2011.

[160] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal Machine Learning Research (JMLR)*, vol. 10, pp. 1755–1758, 2009.

[161] D. E. King, “Max-margin object detection,” *Computing Research Repository (CoRR)*, 2015.

[162] E. Wood, T. Baltrusaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, “Rendering of eyes for eye-shape registration and gaze estimation,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, (Araucano Park, Las Condes, Chile), 7-13 December 2015.

[163] P. F. Felzenszwalb, R. Girshick, D. A. Mcallester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[164] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *Proceedings of the 3rd International Workshop on Computer Vision and Pattern Recognition (CVPR) for Human Communicative Behavior Analysis (CVPR4HB)*, (San Francisco, CA, USA), 7-12 June 2010.

[165] M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, “Disfa: A spontaneous facial action intensity database,” *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151 – 160, 2013.

[166] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, “Avec 2011–the first international audio/visual emotion challenge,” *2011 Affective Computing and Intelligent Interaction (ACII)*, vol. 6975, pp. 415–424, 2011.

[167] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer, “The first facial expression recognition and analysis challenge,” in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*.

[168] M. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. F. Cohn, “Fera 2015 - second facial expression recognition and analysis challenge,” in *Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (Ljubljana, Slovenia), 4-8 May 2015.

[169] S.Abe, *Support Vector Machines for Pattern Classification*. Secaucus, NJ, USA: Springer, 2nd edition ed., 2010.

[170] T. Baltrušaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specic normalisation for automatic action unit detection," in *Proceedings of the the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (Ljubljana, Slovenia), 4-8 May 2015.

[171] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia (MM 10)*, (New York,New York, USA), p. 1459–1462, 25-29 October 2010.

[172] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Press, 2006.

[173] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the ACM international conference on Multimedia (MM Í7)*, (Mountain View, CA USA), p. 478–484, 23–27 October 2017.

[174] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the ACM International Conference on Multimedia (MM'13)*, (Barcelona, Spain), p. 835–838, 21-25 October 2013.

[175] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230 – 236, 1985.

[176] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database," *IEEE Transactions on Biomedical Engineering*, vol. BME-33, no. 12, pp. 1157 – 1165, 1986.

[177] T. W. Parks and C. S. Burrus, *Frequency Transformations*, book section 7.3.3, pp. 213–217. Topics in Digital Signal Processing, New York, USA: Wiley Sons, Incorporated, John, 1987.

[178] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA, USA: Chapman and HallCRC, 1st edition ed., 1984.

[179] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*, vol. 1, book section 8, p. 599–604. San Diego, California: Institute for Cognitive Science, University of California, USA: MIT Press, 1986.

[180] M. I. Jordan, “Serial order: A parallel distributed processing approach,” *Advances in Psychology*, vol. 121, pp. 471–495, 1997.

[181] P. Baldi, “Gradient descent learning algorithm overview: A general dynamical dystems perspective,” *IEEE Transactions on Neural Network*, vol. 6, no. 1, pp. 182–195, 1995.

[182] R. Rojas, *The Backpropagation Algorithm*, book section 7, pp. 149–182. New York, NY, USA: Springer-Verlag Inc., 1996.

[183] S. Sharma, S. Sharma, and A. Athaiya, “Activition functions in neural networks,” *International Journal of Engineering Applied Sciences and Technology (IJEAST)*, vol. 4, no. 12, pp. 310–316, 2020.

[184] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, S. Gruss, and S. Walter, “Facial expression and electrodermal activity analysis for continuous pain intensity monitoringon the X-ITE pain database,” *Life*, vol. 13, no. 9, 2023.

[185] P. E. Shrout and J. L. Fleiss, “Intraclass correlations: Uses in assessing rater reliability,” *Psychological Bulletin*, vol. 86, no. 2, p. 420–428, 1979.

[186] E. Othman, P. Werner, F. Saxen, M. Fiedler, and A. Al-Hamadi, “An automatic system for continuous pain intensity monitoring based on analyzing data from uni-, bi-, and multi-modality,” *Sensors*, vol. 22, no. 13, 2022.

[187] M. Kächele, P. Werner, A. Al-Hamadi, G. Palm, S. Walter, and F. Schwenker, “Bio-visual fusion for person-independent recognition of pain intensity,” in *Proceedings of the International Workshop on Multiple Classifier Systems*, (Günzburg, Germany), 29 June-1 July 2015.

[188] A. Quesada, R. Lopez, and Artelnics, “3 methods to treat outliers in machine learning,” 2022.

[189] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, “Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges,” in *Proceedings of the British Machine Vision Conference*, (Bristol, UK), pp. 119.1–119.13, 9-13 September 2013.

A Written Declaration of Honor

“I hereby declare that I prepared this thesis without the impermissible help of third parties and that none other than the aids indicated have been used; all sources of information are clearly marked, including my own publications. In particular I have not consciously:

- Fabricated data or rejected undesirable results.
- Misused statistical methods with the aim of drawing other conclusions than those. warranted by the available data.
- Plagiarized external data or publications.
- Presented the results of other researchers in a distorted way.

I am aware that violations of copyright may lead to injunction and damage claims by the author and also to prosecution by the law enforcement authorities. I hereby agree that the thesis may be electronically reviewed with the aim of identifying plagiarism. This work has not yet been submitted as a doctoral thesis in the same or a similar form in Germany, nor in any other country. It has not yet been published as a whole.”

Magdeburg, 01.09.2023
Ehsan Othman