Detection of AVPU Scale Consciousness Levels on Emergency Triage Videos

A computer vision-based approach using action recognition and analysis of facial landmarks

Marco Delgado Esperança

Supervisor/s:

João Carlos Amaro Ferreira, Assistant Professor with Habilitation, ISCTE

Tomás Gomes da Silva Serpa Brandão, Assistant Professor, ISCTE

Date: February, 2026

Co-funded by the European Union

Detection of AVPU Scale Consciousness Levels on Emergency Triage Videos

A computer vision-based approach using action recognition and analysis of facial landmarks

Marco Delgado Esperança

Supervisor/s:

João Carlos Amaro Ferreira, Assistant Professor with Habilitation, ISCTE

Tomás Gomes da Silva Serpa Brandão, Assistant Professor, ISCTE

Date: February, 2026

Co-funded by the European Union

Dedication / Acknowledgements

This thesis marks a culmination of a long and challenging but very enriching learning process that would never have been possible without the help and support of a multitude of individuals and institutions.

This research stemmed from a genuine clinical need. I am especially thankful to Dr. Filipa Barros for this. She currently works as a physician with INEM (Instituto Nacional de Emergência Médica in Portugal) and has personal experiences with consciousness assessment in emergency medicine. This directly influenced a concept that forms the genesis of this particular Dissertation. This particular concept has a specific and pertinent goal.

I would also like to extend my gratitude to Pedro Faria (Nostra Prium), for words of advice and entrepreneurial vision that convinced me to look beyond this academic exercise and consider that any research must end up with something with practical application in the real world.

I would like to extend my most sincere gratitude to Professor João Carlos Ferreira for supervising this work and always being available for direction and advice. I would also like to extend a special acknowledgement to Professor Tomás Brandão for his expertise in Image & Video Processing. This work has certainly benefited much from his technical input. Even more importantly for me personally, I shall always be grateful for his commitment to teaching.

This thesis also owes much to the stimulating academic environment that has been encouraged and brought about through the ManagiDiTH program. I would like to thank ISCTE and ISTAR-Iscte for their support in enabling my principal research and Laurea University of Applied Sciences and Aristotle University of Thessaloniki for opening my eyes to new horizons, whether technical or cultural. The chance to pursue studies in this manner has taught me to look at problems in a number of different ways.

Finally, I would personally like to thank my family for being there for me. To my dad: Thank you for being such a source of strength and for being with me when I doubted myself. To my mom: Thank you for being such a source of love and patience throughout this time. To my sister: Thank you for being such a source of kindness and for always being someone I can turn to for a little relief whenever life gets tough. This research may be mine, but it would never have come about without a multitude of contributions—of hearts, minds, and helping hands. I would be deeply grateful to each of you who has joined me on this incredible adventure.

Abstract

Assessment of patient consciousness is a key part of emergency triage, enabling rapid identification of patients at risk of deterioration. The AVPU scale (Alert, Verbal, Pain, Unresponsive) is widely used due to its speed and simplicity, but it remains subjective and prone to variability. Recent advances in computer vision and machine learning enable automated video-based approaches that may improve consistency in assessment.

In this study, 132 simulated triage videos from 52 volunteers were collected and annotated at frame level for microactions such as examiner–patient interaction, eye state, and pain-related facial expressions. A three-stage pipeline was developed to detect interactions, extract microactions, and classify consciousness (conscious vs unconscious). Pose-based temporal models performed best for interaction recognition, with Long Short-Term Memory (LSTM) and Temporal Convolutional Network (TCN) achieving up to 0.90 accuracy. A Convolutional Neural Network (CNN) achieved 0.92 accuracy for eye-state classification, outperforming a traditional baseline, while pain expression recognition showed moderate performance (F1-score of 0.54–0.60). The final TCN-based model achieved 0.92 accuracy for overall consciousness classification. Overall, the results show that microaction-driven video analysis is a promising approach to support automated consciousness assessment in simulated triage settings, while further validation in larger and real-world clinical datasets is required.

Keywords: AVPU Scale, Computer Vision, Action Recognition, Emergency Triage, Deep Learning, Healthcare | JEL Codes: D83, D89

Index

Dedication / Acknowledgements i
Abstract iii
Chapter 1. Introduction 1
1.1 Objectives 4
1.2 Contributions 5
1.3 Methodology 6
1.4 Outline of the Dissertation 7
Chapter 2. State of the art 9
2.1 AVPU-oriented literature review 9
2.2 PRISMA-based review methodology 11
2.3 Related work 13
2.3.1 Action Recognition 13
2.3.2 Facial Expression and Pain Detection 15
Chapter 3. Dataset creation 21
3.1 Video recording setup 21
3.2 Video annotation 23
3.3 Population characterization 24
3.4 Summary 25
Chapter 4. System Development 27
4.1 System requirements 27
4.2 System architecture 27
4.2.1 Data Acquisition Module 28
4.2.2 Feature Extraction Module 29
4.2.3 Microaction Detection Module 29
4.2.4 Clinical Inference Module 30
4.2.5 Output and Visualization Module 32
4.3 Technological implementation 32
4.4 Microaction algorithms 33
4.4.1 Touch / No Touch 34
4.4.2 Pain induction 36
4.4.3 Combined Gesture Recognition: Touch and Pain Induction 37
4.4.4 Eyes open/closed 39

4.4.5 Pain expression detection 40
Chapter 5. Evaluation and Discussion 41
Chapter 6. Conclusion 49
Bibliographic References and Sources 53
Annex A – Additional data on Touch vs. No Touch detection 57
Annex A.1 – Models parametrization 57
Annex A.2 – Confusion matrixes 58
Annex B – Additional data for Pain Induction vs. No Pain Induction detection 59
Annex B.1 – Models parametrization 59
Annex B.2 – Confusion matrixes 60
Annex C – Additional data for combined Touch and Pain Induction recognition 61
Annex C.1 – Models parametrization 61
Annex C.2 – Confusion matrixes 62
Annex D – Additional data for Pain Expression Detection 63
Annex D.1 – Models parametrization 63
Annex D.2 – Confusion matrixes 64
Annex E – Additional data for Eyes Open/Closed Detection 65
Annex E.1 – Models parametrization 65
Annex E.2 – Confusion matrixes 66
Annex F – Additional data for the final prediction of conscious or unconscious 67
Annex F.1 – Models parametrization 67
Annex F.2 – Confusion matrixes 68

Index of Figures

Figure 1.1. Physiological parameters and scoring criteria in the NEWS system. 3
Figure 1.2. DSRM methodology 6
Figure 2.1. Alignment between the AVPU scale and the systematic review 10
Figure 2.2. Prisma diagram for systematic literature review. 13
Figure 2.3. Pain intensity scales. (a) VAS (b) NRS (c) FPS-R. 18
Figure 2.4. Example of a painful face with the PSPI-related AUs. 19
Figure 3.1. Touch interaction. 22
Figure 3.2. Pain induction interaction. 23

Figure 3.3. Example of Touch annotation in Label Studio. 23
Figure 3.4. Example of Pain induction annotation in Label Studio. 24
Figure 4.1. Architecture of the proposed system for AVPU-based assessment. 28
Figure 4.2. Flowchart illustrating the automated AVPU classification process. 31
Figure 4.3. Keypoints of EAR. 40
Figure 5.1. Demonstration of the prototype with a "Conscious" as final prediction. 45
Figure 5.2. Demonstration of the prototype with a "Unconscious" as final prediction. 46

Index of Tables

Table 2.1. Search strategy and eligibility for Action Recognition. 12
Table 2.2. Search strategy and eligibility for Facial Expression and Pain Analysis. 12
Table 2.3. Summary of key studies on action and movement recognition. 16
Table 2.4. Main studies on deep learning models for automatic facial pain assessment. . 19
Table 3.1. Age distribution of the 52 volunteers included in the AVPU dataset. 25
Table 4.1. Comparison of models for Touch vs. No Touch detection. 35
Table 4.2. Comparison of models for Pain induction vs. No Pain induction. 37
Table 4.3. Comparison of models for combined Touch and Pain Induction recognition.. 39
Table 5.1. Model performance for detecting touch-only interactions. 42
Table 5.2. Model performance for detecting pain induction only interactions. 42
Table 5.3. Model performance for detecting both interactions. 43
Table 5.4. Performance comparison of the models for pain expression detection. 44
Table 5.5. Performance comparison of the models for eye open/closed detection. 44
Table 5.6. Final prediction of conscious or unconscious after workflow. 44

Glossary

A&E – Accident & Emergency

AI – Artificial Intelligence

AU – Action Unit

AVPU – Alert, Verbal, Pain, Unresponsive

CNN – Convolutional Neural Network

DSRM – Design Science Research Methodology

EAR – Eye Aspect Ratio

ED – Emergency Department

FACS – Facial Action Coding System

FN – False Negative

FP – False Positive

GAN - Generative Adversial Networks

GCS - Glasgow Coma Scale

INEM – Instituto Nacional de Emergência Médica

LSTM – Long Short-Term Memory

NEWS – National Early Warning Score

NRS – Numeric Rating Scale

PRISMA - Preferred Reporting Items for Systematic reviews and Meta-Analyses

PSPI – Prkachin and Solomon Pain Intensity

RGB – Red, Green and Blue

ROI - Region-of-Interest

TCN – Temporal Convolutional Network

TP - True Positive

TN – True Negative

VAS – Visual Analog Scale

VGG - Visual Geometry Group

Chapter 1. Introduction

Medical emergencies are time-sensitive, high-stress experiences for health professionals, providers and systems, as they often include making life or death decisions in seconds. Over the past few years, the demand for emergency services has continued to grow significantly. For example, in the United States, there were over 140 million emergency department (ED) visits in 2021, which represents 43 visits per every 100 individuals (Ohaiba et al., 2024). In England, the hospital EDs experienced unprecedented demand in 2024, with 27.42 million attendances, corresponding to a 7.1% increase from the previous year. The Royal College of Emergency Medicine estimates that prolonged Accident & Emergency (A&E) waits were linked to over 16,600 excess deaths in 2024, and that in 2023 there were almost 14,000 associated deaths from patients waiting 12 hours or longer (~268 per week) (RCEM, 2024, 2025). Ambulance delays while transferring patients to hospitals compounded the impact of delays in A&E, with "severe potential harm" to more than 44,000 patients during the same time as the reports of deaths, including preventing permanent disability or avoiding death (NHS England, 2025).

Although triage systems exist to prioritize care based on clinical urgency, their effectiveness remains limited. A systematic review found that while major triage systems demonstrate high sensitivity (>90%) for identifying patients at risk of ED mortality, their ability to predict critical illness or post-discharge mortality drops below 80% (Hinson et al., 2019). Historically, 3%-45% of inpatients were under-triaged which is the exact opposite acuity to be (Hinson et al., 2019). Additionally, triage validation faces challenges due to inconsistencies and variations in inter-rater agreement (Moll, 2010).

In standard emergency care, multiple established triage systems have been integrated to differentiate patient care needs based on their clinical urgency. This typically includes the Manchester Triage System, developed in the United Kingdom, widely used across Europe, which uses complaint-specific flowcharts and discriminators to assign one of five urgency levels, each linked to maximum waiting-time targets (Brouns et al., 2019; Zachariasse et al., 2017). The Emergency Severity Index, which is largely used in the United States, combines patient acuity with anticipated resource needs, categorizing patients from level 1 (immediate) to level 5 (least urgent) (Travers et al., 2009). In Canada, the Canadian Triage and Acuity Scale also employs five categories, based on standardized complaint lists and presentation modifiers. It is used to set time-to-physician benchmarks, with adaptations applied in both hospital and prehospital settings (Bullard et al., 2017). Across Australia and New Zealand, the Australasian Triage Scale uses a five category triage supported on maximum waiting time (for each

category), with explicit linkages to risk of physiological decline, and urgency for intervention per each category (ACEM, 2023). Across many low- and middle-income settings, the South African Triage Scale uses both clinical discriminators and a Triage Early Warning Score; patients will be assigned to the final category based on the highest urgency of the components, where validated versions for both adults and children can be found (Markussen et al., 2023).

Beyond triage tools, the Glasgow Coma Scale (GCS) plays a central role in the assessment of neurological status. GCS is widely used for trauma and critical care assessment, measuring three vital components, which include eye opening, verbal response, and motor response to create a score between 3 and 15, generating replicable results. GCS assessment takes time and precise scoring, which leads to the possibility of scoring inconsistencies. Therefore, to fulfill the need for faster assessments in high-pressure environments, simpler scales such as AVPU (Alert, Verbal, Pain, Unresponsive) are usually preferred in triage. Across these systems, a shared underlying principle is the prioritisation of patients according to physiological instability and risk of deterioration. However, they differ in how explicitly neurological status and consciousness are incorporated into the assessment process. This is particularly relevant in the National Early Warning Score (NEWS), a clinical deterioration scoring system based on routinely measured physiological parameters such as heart rate, blood pressure, respiratory rate, oxygen saturation, temperature, and level of consciousness. In NEWS, consciousness is assessed using the AVPU scale, meaning that responsiveness is directly embedded as one of the contributing variables to the overall risk score.

The AVPU is a quick way of evaluating patient consciousness, which is used in systems like NEWS for initial assessments before healthcare professionals can perform a full GCS evaluation where possible. The combination of triage systems with consciousness evaluation tools demonstrates the need for emergency care to achieve both fast response times and accurate assessments, while maintaining objective documentation standards.

One widely adopted tool for assessing clinical deterioration is the NEWS, which uses physiological parameters to evaluate risk (Kolic et al., 2015). However, its major limitation lies in assessing the patient's level of consciousness, the only subjective parameter in the scale. Under high-pressure conditions, this assessment is often incomplete, inconsistently recorded, or subject to significant variability among healthcare providers. Such inconsistencies can lead to misclassification of patient severity, delayed interventions, and undetected deterioration, ultimately impacting patient outcomes.

Figure 1.1 illustrates the physiological parameters and corresponding scoring criteria of the NEWS system, including respiration rate, oxygen saturation, supplemental oxygen use,

temperature, systolic blood pressure, heart rate, and level of consciousness, which is assessed using the AVPU scale (indicated by the arrow in the figure).



PHYSIOLOGICAL PARAMETERS	3	2	1	0	1	2	3
Respiration Rate	≤8		9 - 11	12 - 20		21 - 24	≥25
Oxygen Saturations	≤91	92 - 93	94 - 95	≥96			
Any Supplemental Oxygen		Yes		No			
Temperature	≤35.0		35.1 - 36.0	36.1 - 38.0	38.1 - 39.0	≥39.1	
Systolic BP	≤90	91 - 100	101 - 110	111 - 219			≥220
Heart Rate	≤40		41 - 50	51 - 90	91 - 110	111 - 130	≥131
Level of Consciousness				A			V, P, or U

Figure 1.1. Physiological parameters and scoring criteria in the NEWS system.

Note. Adapted from "Factors affecting response to National Early Warning Score (NEWS)," by I. Kolic, S. Crane, S. McCartney, Z. Perkins, and A. Taylor, 2015, Resuscitation. Copyright 2015 by Elsevier Ireland Ltd.

Early recognition of at-risk patients experiencing clinical deterioration is a foundation of efficient and safe emergency care. NEWS has been adopted throughout Europe as a standardized system to support timely decision-making in acute and prehospital settings (Kolic et al., 2015). NEWS enables clinicians to rapidly quantify severity by aggregating physiological observations into a composite score, enabling them to streamline healthcare resource utilization and dictate patient management pathways (Kolic et al., 2015).

In Portugal, the system of prehospital emergency medicine is already benefited by having an electronic clinical record that automatically integrates information on vital signs directly from portable monitoring devices. However, a significant limitation remains: the qualitative assessment of consciousness through the AVPU scale must still be entered manually by emergency medical technicians. Omission of this data is problematic, as without the AVPU score the NEWS calculation cannot be established, leading to under-recognition of high-risk patients. At the same time, emergency personnel are faced with additional administrative burdens that heighten prehospital delays, and the failure to classify patient deterioration nullifies the NEWS system's main purpose.

Addressing this gap is essential to unlock the full potential of NEWS in the prehospital environment. A solution that enables automatic documentation of consciousness level, along with already automated vital signs, would allow a complete, accurate, and real-time NEWS

score. This would not only improve early recognition of deterioration but also improve decision-making, patient prioritization, and continuity of care from the field to hospital.

This challenge is even more pronounced in prehospital care settings, as those experienced by INEM (Instituto Nacional de Emergência Médica) in Portugal. The AVPU scale is an expedient and simple instrument for determining level of consciousness. However, since it relies on subjective assessment, it is subject to inconsistency and omission when used in real-time triage.

1.1 Objectives

The primary objective of this Dissertation is to research and develop a computer vision-based automated system inspired by the AVPU assessment protocol for evaluating patient consciousness in emergency triage scenarios. Although the AVPU scale categorizes responsiveness into four levels, the proposed system does not perform a direct four-class AVPU classification. Instead, it focuses on the automated detection of clinically relevant microactions associated with different stages of the AVPU assessment process, such as examiner–patient touch interactions commonly associated with verbal responsiveness assessment and pain-inducing stimuli used in reduced-responsiveness evaluation, as well as patient-related cues including eye state and facial expressions associated with pain. These detected microactions are then temporally aggregated to produce a final binary classification of the patient as conscious or unconscious, with the goal of supporting medical personnel in time-critical emergency and pre-hospital environments.

For this purpose, the research seeks to identify the clinical application and working practices of AVPU testing, particularly in pre-hospital and emergency settings. Such environments tend to place a lot of pressure on medical professionals, where manual verification of a patient's responsiveness can result in inconsistency and delay in treatment. At the heart of this initiative is thereby an inquiry into the possibility of addressing the limitations of manual AVPU testing through automation.

Since there are no existing publicly accessible datasets for the assessment using AVPU, one of the major objectives of this study is the creation and design of an AVPU assessment custom video dataset consisting of emergency response simulations. This includes constructing clinically relevant scenarios from AVPU protocols, constructing an intensive annotation protocol directed by medical experts to annotate microactions and reactions with precise temporal boundaries, performing quality control checks such as multi-rater annotation and

inter-annotator agreement analysis to supply guarantees of ground truth reliability. Ethical and data-protection statutes are adhered to at all times.

The Dissertation also describes the development of a prototype system that processes video data to classify the responsiveness of a patient. It is achieved via the combination of facial expression analysis and body movement recognition—two complementary computer vision approaches that when combined can record important behavior cues regarding all AVPU categories.

The prototype's functionality is verified by means of quantitative and qualitative testing and evaluation. This includes evaluating the system on unseen video data in terms of robustness and generalization capacity. Finally, the Dissertation presents a reflective consideration of the broader implications for the suggested method, e.g., clinical utility, ethical problem, and potential for future development in real emergency and pre-hospital scenarios.

Keeping these goals in mind, the Dissertation offers an end-to-end, clinically informed solution from dataset development and labeling, model training and evaluation to decision-making and enhancing consciousness assessment consistency when time and accuracy are of the essence. This leads to the research question:

How can computer vision techniques support automated video-based binary consciousness classification of patients as conscious or unconscious, inspired by the AVPU protocol, in pre-hospital emergency triage environments?

1.2 Contributions

This Dissertation offers relevant and innovative contributions with application in the combined fields of emergency medicine, computer vision, and e-health in addressing the problem of consciousness classification. It presents a computer vision-based system that integrates facial expression analysis, body pose estimation, and temporal modeling to automatically estimate patient responsiveness using the AVPU scale. It also introduces a dataset that was created to enable training and evaluation of these tasks. This dataset comprises simulated emergency conditions recorded with volunteers under clinical monitoring and annotated by medical staff to ensure robust medical criteria. Furthermore, the work enriches the field by introducing action and movement detection through pose estimation and sequence modelling.

Thus, this facilitates the identification of goal-directed and reflexive microactions, such as responses to speech input or painful stimuli, which are essential to distinguish AVPU states. A real-time prototype is developed to offer emergency medical personnel objective and

standardized support for assisting in making quick triage decisions. The system is qualitatively tested for reliability, clinical utility, and usability by medical professionals and quantitatively tested for performance, reflecting its effectiveness, clinical utility, and usefulness. Together, these contributions strengthen the automation of consciousness assessment in emergency settings and provide a clinically guided and innovative basis for further research and use in real-life scenarios.

1.3 Methodology

The methodology adopted in this Dissertation is the Design Science Research Methodology (DSRM), as documented by Figure 1.2 (Peffers et al., 2007). This approach was selected because it is well-suited for research that aims not only to understand a problem but also to design and evaluate an artifact as a solution for such problem. In this context, the artifact is a prototype capable of automatically performing AVPU consciousness assessments from emergency triage videos using computer vision.

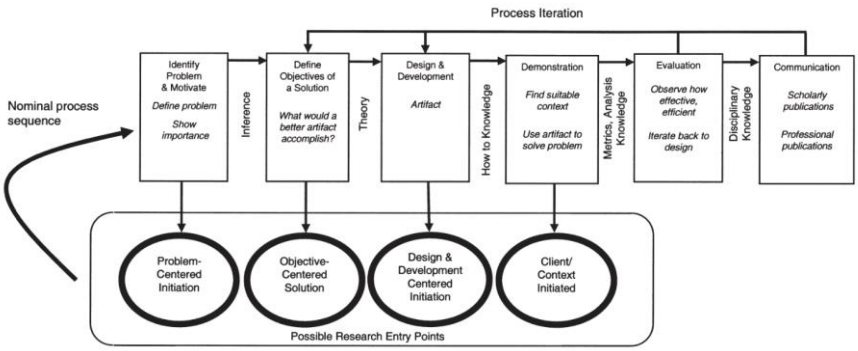


Figure 1.2. DSRM methodology

Note. Adapted from "A Design Science Research Methodology for Information Systems Research," by K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, 2007, *Journal of Management Information Systems, 24*(3), 45–77. Copyright 2007 by M. E. Sharpe, Inc.

DSRM is distinct from behavioral Information Systems research due to its emphasis on intentional design and innovation. While behavioral research focuses on observing and theorizing existing practices, DSRM aims to transform these practices through the creation of novel and useful artifacts. In the context of health informatics and digital transformation, where technology must adapt to delicate, real-time clinical environments, DSRM offers a rigorous yet flexible methodology from problem identification to pragmatic evaluation.

This process entails six core activities: problem identification and research motivation; definition of solution objectives; design and development of the artifact; demonstration of the

artifact in situ; evaluation of performance; and communication of results. These processes are not necessarily linear and often iterate, particularly between design and evaluation, refining the artifact based on empirical evidence.

The entry point for this research was problem-initiation-centered. The initiative was conceived from a clinical observation: the AVPU scale, although widespread throughout prehospital care environments, is marred by inconsistent and subjective interpretation, particularly in high-pressure scenarios such as those encountered by emergency responders. The motivation for this research was to reduce such subjectivity by offering an automated classification method based on facial expressions and bodily movement captured through video.

Afterwards, a prototype system was developed to analyze and classify simulated emergency triage videos according to the AVPU scale. This entailed combining facial Action Unit detection (AU), in the sense of the Facial Action Coding System (FACS) and ratings of pain, with pose estimation and short-sequence action recognition. The system was trained on a new dataset created for this task, consisting of several videos containing four annotated states of AVPU. The implementation was done under test conditions to reduce the unpredictability of live clinical environments and verify that the initial iteration of the system operated as designed and could be tested.

Following the development phase, the system was tested to see how the prototype performed when dealing with novel sequences of videos, simulating real-time classification. In the second release, exploratory evaluation was conducted with the involvement of clinical professionals, who compared the output of the system with their own ratings. This provided technical and domain feedback, crucial to making a judgment about the validity and possible utility of the artifact under real-world conditions.

1.4 Outline of the Dissertation

With the research objectives and methodology established, the Dissertation is structured into six chapters, as outlined below:

- Chapter 1: Introduces the context and motivation behind the Dissertation, addressing the clinical and technical challenges in emergency triage, particularly in assessing the level of consciousness using the AVPU scale. It outlines the research problem, the main goals of the study, and provides an overview of the structure of the Dissertation.
- Chapter 2: Presents a systematic review of the literature on emergency triage systems, with focus on the AVPU scale and its limitations in clinical settings. It also explores the

current state-of-the-art in pain and consciousness assessment using computer vision and deep learning, including facial expression analysis, pose estimation, and human action recognition. The purpose is to identify existing gaps and to establish the relevance of an automated AVPU classification system.

- Chapter 3: Covers the dataset creation process using simulated emergency scenarios, the annotation strategy, and the development of the system's core components, including facial expression and pose analysis modules.
- Chapter 4: Demonstrates the developed prototype through the iterative steps of Design and Development, as well as Demonstration defined by the DSRM process. It also explains the user interaction flow, provides examples of video input for each AVPU level, and describes how the system processes facial and body movement data to generate classifications. Screenshots and visual outputs illustrate the key stages of the system in operation.
- Chapter 5: Evaluates and discuss the proposed prototype. The evaluation is conducted using unseen video data and incorporates both quantitative methods—such as precision, recall, and confusion matrices—and qualitative validation through feedback from healthcare professionals. Therefore, the purpose of this chapter is to examine the effectiveness, reliability, and responsiveness of the artifact while ensuring its relevance to emergency triage contexts. Beyond the evaluation, the chapter also offers a critical discussion of the broader implications of the system. This includes a reflection on its potential strengths and limitations, the ethical considerations associated with the use of video technologies in healthcare, and an assessment of the practical applicability of the system in both pre-hospital and clinical environments.
- Chapter 6: This final chapter summarizes the developed work, highlights the main contributions to the field of digital health and emergency care, and suggests directions for future research, including potential integration into real-world emergency services and scaling with larger clinical datasets.

Chapter 2. State of the art

The review of the state of the art starts by examining whether there is pre-existing Artificial Intelligence (AI) research that addresses the specific problem of detecting levels of consciousness through the AVPU scale. Identifying such work makes it possible to understand what has already been attempted, what contributions have been made, and what limitations remain. After clarifying this question, the review expands into the two main technical areas that underpin automated AVPU assessment: action recognition and facial expression analysis.

2.1 AVPU-oriented literature review

Literature review was organized around the two domains most relevant to AVPU automation. The first concerns the recognition of clinical interactions and motor responses, that are discrete and repeatable actions that can be automatically identified using action recognition methods, an area where advances in action recognition and pose estimation provide tools to detect gestures such as touch, pain induction, or spontaneous movement. The second focuses on facial expression and pain analysis, which are essential when overt motor responses are absent and subtle nociceptive cues like grimacing or eye tightening must be detected. Together, these two strands of research form the conceptual foundation for the automated assessment of consciousness levels in emergency triage. In addition, they directly inform the methodology adopted in this Dissertation, where the creation of a bespoke dataset with fine-grained annotations of clinical interactions and patient responses provides the necessary ground truth to develop and evaluate such models.

The AVPU scale is structured as a sequence of progressively applied assessments. After observing whether the patient is spontaneously alert, the clinician initiates two key forms of interaction: first through verbal contact combined with light touch (V), and then, if necessary, through the application of a painful stimulus (P) such as a trapezius squeeze or sternal rub. If neither of these interactions produces a reaction, the patient is classified as Unresponsive.

This two-part structure, beginning with interaction recognition and moving to patient responsiveness, directly reflects the clinical logic of AVPU and underpins the organization of the review. This alignment of the AVPU protocol and the systematic review focus is shown in Figure 2.1, where it is possible to observe how the key clinical interactions (touch and pain induction) and their related domains of analysis (action recognition and facial expression analysis) are integrated into the review framework.

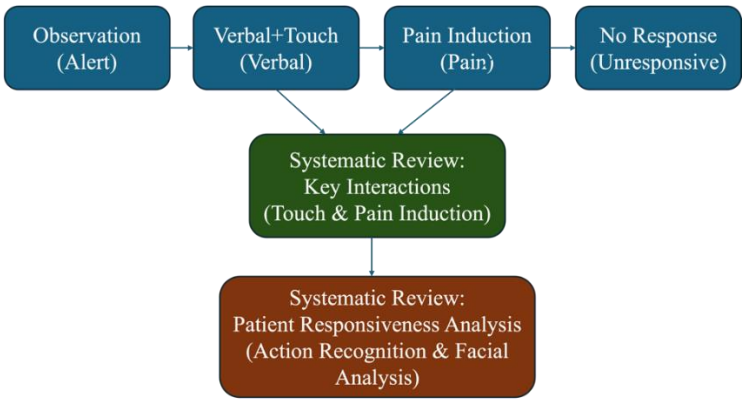


Figure 2.1. Alignment between the AVPU scale and the systematic review

Furthermore, advances in human activity recognition and pose estimation are particularly relevant in this context. Detecting whether a clinician applies a tactile stimulus or performs a painful maneuver, and whether the patient responds with movement, can be framed as a sequence of actions unfolding over time. Video-based action recognition methods provide tools for capturing these interactions, not only by analyzing the patient's body movements but also by recognizing the clinician's initiating gestures. This is especially important because the meaning of a patient's response depends on its temporal relationship to the applied stimulus: a reflexive withdrawal after a painful stimulus has a different diagnostic significance than a spontaneous gesture in the absence of stimulation. By reviewing the literature on action recognition in healthcare, it becomes possible to identify techniques capable of modeling these stimulus–response dynamics, which lie at the heart of automated AVPU assessment.

However, it is important to recognize that not all patient responses to pain are expressed through gross motor actions. In many cases, particularly where the patients are physically restrained, weakened, or neurologically compromised, motor responses are absent or trivial. Thus, facial reactions are usually the most reliable indicators of unimpaired consciousness. Grimacing, eye closing, or minimal movements around the brow and mouth are generally the only visible responses that demonstrate the painful stimulus has been perceived.

Therefore, studies in facial expression analysis and pain detection provide complementary methods to action recognition by focusing on micro-level cues rather than full-body movements. Techniques based on the FACS and pain intensity scores, such as the Prkachin and Solomon Pain Intensity (PSPI) measure, have demonstrated the ability to detect nociceptive responses from visual features alone. These approaches are directly applicable to the "Pain" stage of AVPU, since they provide objective means to identify reactions that may otherwise go unnoticed in the absence of overt motor activity. Integrating insights from this body of research alongside action recognition therefore ensures a more complete coverage of the AVPU process:

body movement analysis addresses spontaneous activity and responses to touch, while facial expression analysis captures subtler responses to painful stimuli. Together, they form a coherent technological framework that mirrors the clinical sequence of AVPU assessment.

2.2 PRISMA-based review methodology

The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) methodology helps authors to perform literature reviews and meta-analyses more effectively (Moher et al., 2009). It is an essential method to present systematic reviews of other research forms, including observational studies, diagnostic test accuracy studies, and qualitative research, ensuring transparency, completeness, and reproducibility in reporting.

Two complementary searches, i.e., for facial expression and action and pain analysis and pose estimation, were conducted in Scopus, Web of Science and PubMed. The searches were limited to the last ten years and to English language publications, such as peer-reviewed journal articles, reviews, and conference papers. Since vision-based testing is highlighted in this Dissertation, only studies using video as first modality were considered; multimodal studies were just included if central position of visual stream was in approach or result.

To isolate result of construction and search, the process is described in two levels.

Building directly upon the Concept-Context-Limitations schema summarized in Table 1 and Table 2, two query search strings were instantiated. The Concept column supplied the core keywords for each domain, the Context column captured the target phenomenon and visual scope (i.e., hand-focused action recognition and facial pain analysis), while the Limitations column defined the temporal window, language, study type, modality (video-based), and the requirement for quantitative outcomes. These query results were processed with the same inclusion/exclusion rules and then screened according to PRISMA (see Figure 2.2).

Table 2.1 initially lays out the action-recognition topic search plan, what was searched for, what limits were placed on the search to confine it to emergency and triage settings, and what was included and excluded. It's a form table: it tells us what was being looked for and under what parameters records were included and excluded but doesn't tell us about outcomes.

Table 2.2 has the same purpose for the facial pain-analysis group and shows an open register of corresponding terms, scopes of scope, and inclusion criteria. Following screening and eligibility determination based on PRISMA is the summary of studies included above said methods.

Table 2.1. Search strategy and eligibility for Action Recognition.

Concept	Context	Limitations
Action Recognition Pose Estimation Skeleton training Hand	Hand / upper-limb focus Video-based action recognition General-domain (not restricted to a clinical setting)	Last 10 years Peer-reviewed journals, articles, reviews and news Written in English Vision-based (video) only Quantitative results required (e.g., accuracy, F1)

Table 2.2. Search strategy and eligibility for Facial Expression and Pain Analysis.

Concept	Context	Limitations
Facial expression FACS Pain detection / assessment Deep learning / Machine learning / “Computer vision”	Pain-related facial analysis Pain detection / assessment General-domain pain analysis (not restricted to triage) Video/image studies (when applicable)	Last 10 years Peer-reviewed journals, articles, reviews and news Written in English Quantitative results required

The first query - "action recognition" AND ("pose estimation" OR "skeleton tracking") AND ("hand") realizes the Action Recognition strategy in Table 1 by combining motion-centric terminology with pose-based representations relevant to fine-grained upper-limb interactions, which are central to clinician–patient contact and stimulus–response assessment in emergency triage scenarios.

The second query - ("facial expression" OR "facial action coding system" OR "FACS") AND ("pain detection" OR "pain assessment") AND ("deep learning" OR "machine learning" OR "computer vision"); the search strategy underlying the Facial Expression and Pain Analysis plan in Table 2 targeted nociception-related facial cues in triage contexts. We listed each study according to its year of publication, application domain-namely movement versus facial pain, methodological family, and use case.

In addition to the mentioned structured queries, complementary ad-hoc searches were conducted to ensure coverage of the most recent and emerging work. These included manual screening of references from relevant systematic reviews and targeted searches of preprint repositories (e.g., arXiv) to identify recent studies not yet published in peer-reviewed journals. All records identified through these alternative sources were subjected to the same screening and eligibility criteria as database-derived records and are reported separately in the PRISMA 2020 flow diagram.

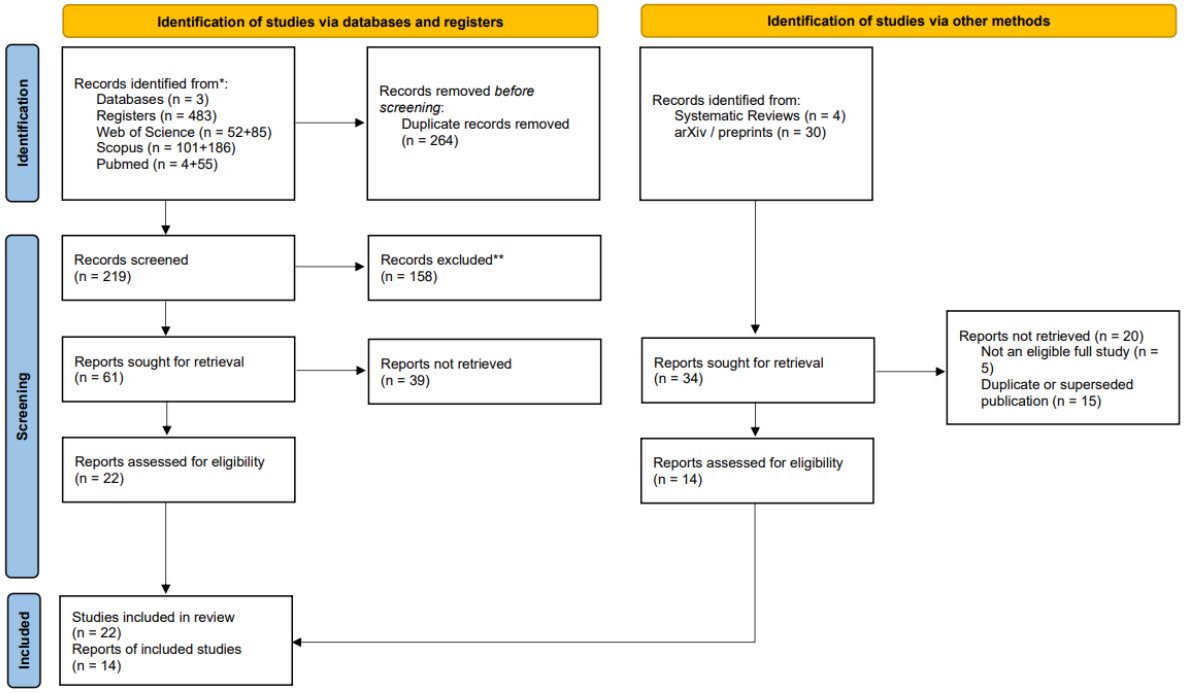


Figure 2.2. Prisma diagram for systematic literature review.

2.3 Related work

2.3.1 Action Recognition

Automated detection of human action and movement is poised to become central to contactless, real-time patient monitoring. Consequently, vision-based systems are able to classify human behavior without the need for intrusive sensors. Therefore, they enable applications such as fall detection, postural monitoring, rehabilitation assessment, and early detection of responsiveness changes at the bedside. In emergency and critical care, determining whether a patient shows purposeful movement when responding to stimuli—the "P" in the AVPU scale—is an important clinical assessment. However, whereas earlier work on automated testing has focused on pain recognition through facial analysis, comparatively less effort has gone into recognizing short and simple body movements as indicators of consciousness. A methodological gap today exists that can be plugged by cutting-edge action recognition.

A study by Choi et al. directly tackles AVPU classification with computer vision and speech technologies (Choi et al., 2023). Their system combined body and face video with speech recognition to identify whether a simulated patient could be categorized as alert, responsive to verbal or painful stimulus, or unresponsive. The results were promising, with accurate performance under controlled conditions, and the research demonstrated for the first time that AVPU can be operationalized via automated video analysis. Nevertheless, the study also

revealed considerable shortcomings. It was limited to simulation scenarios with a small group of participants, depended heavily on language-specific speech recognition, and relied on rule-based mapping between detected events and AVPU states. These constraints reduce generalizability to real prehospital environments, where conditions are far less predictable, and the critical temporal relationship between a clinician's stimulus and the patient's response is not considered. This gap defines the "drawer" that motivates the present Dissertation: the need for methods capable of modeling stimulus–response dynamics in a robust, data-driven way.

Early and mid-cycle surveys set a path from hand-crafted features towards deep models that all jointly learn spatiotemporal patterns and combine Red, Green and Blue (RGB), depth, and skeleton streams (Zhang et al., 2016). This path is promising for AVPU-related scenarios because motions are brief, context-dependent, and often partially occluded. Even so, most of the existing work is tested on entertainment or daily-activity benchmarks rather than on clinical footage; hence, direct transferability to triage is restricted.

Here, pose-based approaches are particularly well suited since they are structure-focused rather than appearance-focused and, as a result, can better preserve privacy. For instance, Rajasegaran et al. demonstrated that recovering 3D joint coordinates from monocular video with Tracking and encoding them using a transformer (PoseFormer) can rival or surpass RGB baselines on large datasets such as Kinetics-400 and AVA with up to 72.8% top-1 on Kinetics-400 using pose alone (Rajasegaran et al., 2023). Hence, for AVPU-like micro-movements, this approach has some appeal: skeleton sequences are generally insensitive to clothing, lighting, and background variation prevalent in prehospital scenarios. However, a major limitation remains: since these benefits are tested on general-domain corpora, it remains unclear how far they generalize under restricted clinical perspectives, heavy occlusions due to blankets or caregivers, and unusual postures on stretchers.

Efficiency and latency constraints also apply to deployment; for this purpose, neuromorphic deployments seek low-power designs. As an example, Layher, Brosch, and Neumann project a DCNN onto IBM TrueNorth in order to classify chosen key poses based on motion-energy extrema and pose "extendedness" that combines optical-flow and gradient cues (Layher et al., 2017). Their findings confirm that low-power real-time action cues are feasible and therefore applicable to ambulances' compute- and power-limited embedded devices. However, the analyses were typically done on well-curated actions with cooperative subjects, so it is difficult to state how well the key-pose heuristic works on quiet reflexes, tremors, or clinician–patient conversations common to triage videos.

In parallel, hybrid pipelines which restrict attention to semantically meaningful motion offer a dual route toward interpretability and efficiency. As an example, Camarena et al. replace dense sampling with the utilization of "Key Trajectories" which are connected to pose keypoints, thereby reducing trajectories by up to 90% and accelerating inference by up to 8 times with comparable accuracy on KTH, UCF11, UCF50, and HMDB51 (Camarena et al., 2022). In the process, such efficiency might make real-time triage support possible on commodity hardware. But the trade-off is substantial: because such data represent stylized, full-body motion, performance over short, localized responses—hand removal, flinchy shoulder motions, or small postural adjustments—is still due for testing on clinical video with mess, PPE, and varying camera placement.

Furthermore, a more general class of deep architectures—two-stream Convolutional Neural Networks (CNNs) using RGB and optical flow (Feichtenhofer et al., 2016), hierarchical 2D/3D CNN with temporal recurrent heads (Li et al., 2016), and motion-vector CNNs for inference efficiency (Zhang et al., 2016) —continuously achieves state-of-the-art performance on public benchmarks and thus provides implementation blueprints. Nonetheless, because these models are heavily rooted in appearance cues and longer, better-segmented activities, their application to AVPU is not suitable, as diagnostically significant windows are brief and must be time-locked to the stimulus. In general, the literature suggests that pose-based or pose-anchored methods yield most practical trade-offs between robustness, privacy, and efficiency. However, there is a strong lack of evidence on healthcare-specific data.

Table 2.3 summarizes key studies on action and movement recognition, emphasizing the continued disparity between technical progress and clinical applicability.

2.3.2 Facial Expression and Pain Detection

Pain detection through facial expression analysis is a forefront of research in computer vision, particularly when verbal communication is impaired or impossible. Among the pioneering tools in this area is the FACS that identifies individual facial movements—called Aus —that are associated with affective and physiological state, including pain (Ben Aoun, 2024). The AUs, such as brow lowering (AU4), orbital constriction (AU6 and AU7), and lip stretch (AU20), are all closely linked with nociceptive response and underlie much of the automated pain detection technology (Ben Aoun, 2024).

Table 2.3. Summary of key studies on action and movement recognition.

Author(s)	Deep Learning Model	Main Contribution
Fu et al. (2023)	Improved YOLOv7-Pose + ConvNeXt + Coordinate Attention	Enhanced fitness pose/action recognition with attention & better pooling
Li et al. (2016)	Multi-stream 2D/3D CNN + LSTM	Hierarchical spatio-temporal representation (frame, motion, clip, full video)
Feichtenhofer et al. (2016)	Two-Stream ConvNet Fusion	Spatial & temporal fusion for improved recognition
Layher et al. (2017)	Deep CNN (Eedn) on Neuromorphic hardware	Real-time key pose-based recognition with low energy on IBM TrueNorth
Camarena et al. (2022)	Key Trajectories + Dense Trajectories + Pose Estimation	Faster hand-crafted + DL hybrid for efficient recognition
Hachaj et al. (2018)	Neural Network, Random Forest	MoCap-based position-invariant classification with multiple algorithms
Lim et al. (2024)	CNN using time-warped joint motion images	Based Taekwondo UA recognition
Zhang et al. (2016)	Motion Vector CNN + Optical Flow CNN (Knowledge Transfer)	Real-time recognition using motion vectors, enhanced by pre-trained optical flow CNNs
Gong & Luo (2025)	CNN-based models (2D CNNs, 3D CNNs, Two-Stream, Fusion architectures)	Survey

Ground truth labelling in automated facial pain analysis is a crucial step on the way towards supervised learning algorithm training and testing. Clinical pain assessment still relies traditionally on self-report measures such as the Visual Analog Scale (VAS) and the Numeric Rating Scale (NRS). The VAS requires patients to report subjective pain on a continuous scale, usually "no pain" to "worst pain imaginable," while the NRS uses a discrete number scale to rate the intensity of pain (usually 0–10). Both scales are appropriate for communicative patients but inappropriate for unconscious, non-verbal patients or those with altered mental status.

To avoid the limitation mentioned above, objective scoring systems based on observable facial expressions are most frequently used. From these, the most frequently utilized measure of facial expression measurement is Prkachin and PSPI score (Prkachin & Solomon, 2008). The PSPI score is a summation of intensities of certain AUs coded in FACS, viz., AU4 (brow lowering), AU6 and AU7 (orbital tightening), AU9 (nose wrinkling), and AU10 (upward movement of the upper lip) (Ben Aoun, 2024). The composite score measures the degree of pain-related facial activity, independent of verbal complaint or subjective report. The PSPI is traditionally used as ground truth facial pain databases such as the Master Shoulder Pain Expression Archive Database (Werner et al., 2022). This dataset enables researchers to train machine learning and deep learning algorithms to detect pain from visual facial expressions

only, which can translate into the creation of automated systems capable of estimating pain in non-verbal patients.

A spur to progress in the field has been the availability of open-access pain expression datasets. The most frequently applied dataset is the Master Shoulder Pain Expression Archive Database containing over 200 video sequences of 25 patients. It contains frame-by-frame pain annotations with detailed labelling on the Visual Analog Scale (VAS) and FACS-coded AUs. The dataset is the gold standard for facial pain expression detection under both classification and regression algorithms. Another appropriate database is the BioVid Heat Pain Database capturing thermal-induced pain reactions (Werner et al., 2022). In addition to facial videos, BioVid also includes physiological signals such as ECG and skin conductance, making it appropriate for multimodal analysis. However, BioVid is less popular in deep learning studies since it contains less facial information and is comparatively smaller. Other datasets exist but are proprietary or insufficiently labeled to meaningfully train or benchmark supervised models. To that degree, dataset constraints still place a large-scale burden on generalizability and deployment.

Building on the availability of these open-access datasets—particularly McMaster and, to a lesser extent, BioVid—the methodological focus of facial pain analysis has progressively shifted from hand-crafted representations to deep, spatiotemporal models. Early work relied on geometric and appearance-based features combined with conventional classifiers such as Random Forests. As larger, consistently annotated facial video corpora became available, deep CNN architectures such as ResNet emerged as the dominant approach, generally achieving superior performance and robustness in pain versus no-pain classification. More recent CNN hybrids further extend this line of work by explicitly modeling temporal micro-dynamics in facial expressions, thereby improving the estimation of pain intensity over video sequences.

Spatiotemporal models, such as 3D CNNs and attention-based transformers, pushed the accuracy even higher by learning joint space-time features, with 3D-ResNet-like models exceeding 89% in transformer versions having better correlations for subtle and transient expressions (Tavakolian et al., 2020; Zhou et al., 2016). Attention mechanisms that either focus on pain-salient areas or impose identity-aware regularization improve robustness by disentangling person-specific appearance from expression dynamics (Cui et al., 2021; Xin et al., 2020, 2021). Reported accuracy values—e.g., 96–99% with pre-processing, transfer learning, and ensemble combination tailored to them (Alghamdi & Alaghband, 2022; Karamitsos et al., 2021; Semwal & Londhe, 2021b)—must be approached with caution: the majority of reports often feature subject-overlapping folds, over-aggressive pre-processing, or

highly restricted evaluation protocols, which can result in biased performance. Cross-subject, cross-dataset generalization takes a steep drop in most cases, in favor of identity-disentangling training, additional regularization, and demographic diversity.

Facial analysis is most applicable to the "Pain" step from an AVPU perspective because it can detect nociceptive response in cases where limb movement is prevented, confused, or absent. Figure 2.3 positions conventional clinical pain scales, and Figure 2.4 demonstrates PSPI calculation from AU intensities. Clinical application, however, is faced with three inherent limitations. First, emergency video streams bear little resemblance, frontal facial views; head pose change, occlusion due to oxygen masks or hands, and non-uniform lighting undermine face tracking and AU estimation. Second, the majority of facial responses in critically ill patients carry risks of false negatives (FNs), which suggest conservative thresholds or confirmation by other modalities. Third, there are concerns regarding equity because AU detectors can be poor performers on all skin types, age, and face morphologies if training sets are not balanced. These concerns justify the multimodal strategy followed here, in which facial information is augmented by pose-based micro-behaviors and, where feasible, synchronized to the recorded onset of an incidental vocal or pain stimulus to further strengthen the evidentiary value of responses elicited.

Table 2.4 overviews deep learning works on automatic facial pain estimation and summarizes the general trend: state-of-the-art spatial–temporal models and attention models perform optimally, multimodal extensions are promising, and the primary challenge to AVPU-specific application remains external validity in the real-world triage environment.

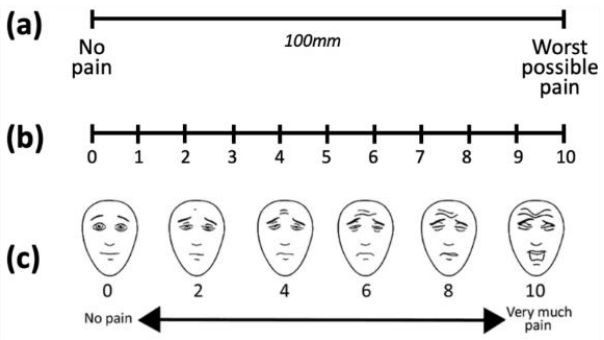


Figure 2.3. Pain intensity scales. (a) VAS (b) NRS (c) FPS-R.

Note. Adapted from "A Review of Automatic Pain Assessment from Facial Information Using Machine Learning,"by N. Ben Aoun, 2024, *Technologies, 12*(6), Article 92. Copyright 2024 by the author, distributed under the Creative Commons Attribution (CC BY 4.0) license.

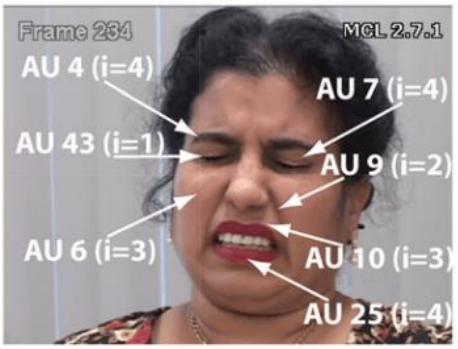


Figure 2.4. Example of a painful face with the PSPI-related AUs.

Note. Adapted from "A Review of Automatic Pain Assessment from Facial Information Using Machine Learning," by N. Ben Aoun, 2024, *Technologies, 12*(6), Article 92. Copyright 2024 by the author, distributed under the Creative Commons Attribution (CC BY 4.0) license.

Table 2.4. Main studies on deep learning models for automatic facial pain assessment.

Author(s)	Deep Learning Model	Main Contribution
Casti et al. (2019)	Pre-trained AlexNet	Re-annotated dataset with 7 experts; illumination-invariant 3D frame transformation
Dragomir et al. (2020)	ResNet	Hyper-parameter optimization and data augmentation
Xin et al. (2020)	9-layer CNN	Attention mechanism focused on region-related pain
Xin et al. (2021)	CNN	Locality-aware and identity-aware modules
Rathee et al. (2022)	CNN	Efficient CNN parameter initialization using MAML++
Cui et al. (2021)	Multi-scale regional attention	Self-attention and relation modules to emphasize pain-related face parts and inter-relationships
Karamitsos et al. (2021)	Customized and deeper Visual Geometry Group (VGG16)	Pre-processing: grayscale, equalization, cropping, filtering, face detection
Semwal & Londhe (2021a)	SPANET	Shallow CNN with FP reduction strategy
Alghamdi & Alaghband (2022)	InceptionV3 + shallow CNN	Frozen conv block; shallow CNN as classifier
Zhou et al. (2016)	3D CNN	Spatiotemporal pain detection from facial video
Egede et al. (2020)	CNN (frame-based)	Frame-based pain classification on Master Shoulder Dataset
Aung et al. (2016)	CNN	CNN features combined with Relevance Vector Regression for intensity estimation
Werner et al. (2017)	Deep Neural Network	Combined handcrafted and learned features for pain intensity regression
Semwal & Londhe (2021b)	Fusion of VGG and CNN	Combined geometric and appearance features; combined distinct feature types (e.g., RGB, entropy-based texture)
Huang et al. (2019)	Generative Adversial Networks (GAN) + CNN	Used GAN-based augmentation to improve CNN training on limited pain data
Tavakolian et al. (2020)	Transformer-based	Vision Transformer model for pain recognition in video sequences
Wang et al. (2017)	Graph CNN	Used facial landmark graphs for pain estimation

Chapter 3. Dataset creation

An important aspect of the prototype design was the development of a dataset tailored to the needs of automated AVPU classification. Since no public datasets are available for such a task, a bespoke dataset of simulated emergency situations was established. Beyond simply recording video content, the dataset was annotated meticulously at the microaction level—documenting the exact onset and duration of clinically significant events such as light touch, pain stimulation, eye opening, or head movement. These granular annotations, therefore, are required for model development, as they provide the temporal and semantic precision to train algorithms to recognize subtle stimulus–response patterns that define each AVPU state. The final dataset consisted of approximately 132 video sequences with a total duration of approximately two hours, recorded using fifty-two volunteers across the age groups of 18–65.

3.1 Video recording setup

Data recording was conducted in a simulated environment rather than in ambulances. This decision was largely driven by pragmatism and ethics. Ambulances are precious resources dedicated to real emergency care and cannot be occupied by research without taking away from their availability to the needy. In addition, video capture in ambulances during live emergencies poses serious problems in terms of patient consent, privacy and safety, as well as practical constraints in deploying and configuring the required recording equipment quickly and safely without interfering with clinical care. Due to these concerns, the dataset was created under controlled conditions that nonetheless strove to ensure a high degree of clinical realism. The scenarios were built and acted out by health care practitioners whose skill level ensured that the simulated interaction was close to the reality of pre-hospital emergency inspections.

The development of the dataset was subject to a process designed to deliver clinical realism and methodological rigor while maintaining the scenarios clinically consistent and reproducible:

1) Healthcare professionals first examined typical pre-hospital environments and determined the set of scenarios required to span the whole variety of AVPU states. These scenarios were carefully developed to reflect clinically tested processes and to provide adequate coverage of the different levels of consciousness assessment.
2) In the second stage, pilot recordings of interactions between the professionals were conducted. This allowed them to pilot and refine the simulation process so that the procedures were replicable and represented clinical practice.

3) Finally, the final dataset videos were recorded, consisting of interactions between healthcare professionals and a group of volunteers who simulated varying levels of responsiveness on the AVPU scale.

This task was performed in collaboration with an emergency specialist, who designed the simulated scenario and developed the dataset. Under her supervision, the recording sessions were organized based on clinically tested interactions that are representative of the AVPU scale. These scenarios were based on microactions most relevant to the determination of responsiveness: gentle touch to secure verbal-level response, pain stimulation to evaluate responses at the pain threshold, spontaneous or elicited head movement, and eye opening or closing as salient indicators of alertness or unresponsiveness. Medical specialists were consulted in the development of these scenarios to ensure that they were consistent with existing triage procedures and were representative of diagnostic behavior.

Figure 3.1 illustrates an example of a touch interaction, where a clinician is applying a gentle tactile stimulus to the shoulder of a volunteer simulating impaired responsiveness. This type of interaction is relevant to distinguishing the "Verbal" level of the AVPU scale, where a patient is not spontaneously alert but does respond to light stimulation. Figure 3.2 illustrates a pain induction interaction, where a more severe physical stimulus is being applied. This is the clinical step to assess whether a patient who did not respond to voice or touch can make any reaction to pain, which corresponds to the "Pain" level of the AVPU score. Both examples illustrate the stepped clinical process of going from less invasive to more severe stimuli to assess reducing levels of consciousness.



Figure 3.1. Touch interaction.



Figure 3.2. Pain induction interaction.

3.2 Video annotation

Once recorded, all video material was imported into the Label Studio platform for detailed annotation. The annotation process was carried out by clinical experts, who identified and labeled the precise moments where microactions occurred, such as the onset of a touch or the response to a painful stimulus. This process enabled the dataset to be segmented into temporal intervals with medically meaningful ground truth labels. Figure 3.3 presents a screenshot of a touch annotation within Label Studio, where the interaction is highlighted in green against the surrounding context. This visual segmentation allows the learning models to focus on the exact time window of interest rather than the entire video sequence. Figure 3.4 shows the annotation of a pain induction event which was labeled in red to differentiate it from intervals which do not have such stimulation. Each annotation includes its exact start and end times. Consequently, the dataset captures both the type of microaction and its duration. This temporal information is crucial for training models that recognize patterns over consecutive frames.



Figure 3.3. Example of Touch annotation in Label Studio.

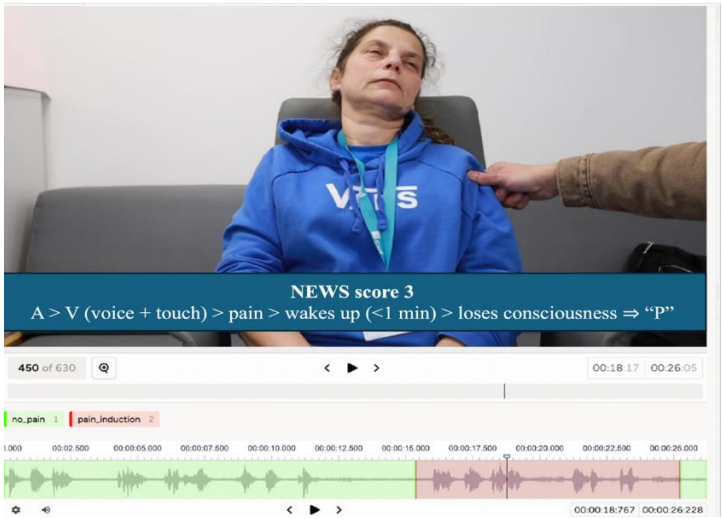


Figure 3.4. Example of Pain induction annotation in Label Studio.

To ensure reliability, annotations were performed independently by more than one clinical expert and later compared. Inter-annotator agreement was measured using Cohen's kappa, which confirmed a high level of consistency in identifying the key microactions across raters. This strengthened the dataset's validity as a clinical ground truth.

The resulting dataset therefore provides a structured, clinically informed foundation for model training. It includes the basic microactions corresponding to AVPU classes, is temporally partitioned for learning spatio-temporal dynamics, and presents visual samples illustrating both the clinical interactions and the corresponding annotation procedure. Thus, the dataset not only enables the training of automatic models but also integrates clinical knowledge in the annotation process, closing the loop between medical knowledge and computational modeling.

The study was ethically cleared by Comissão de Ética da Escola de Tecnologias e Arquitetura, Iscte (ref. CE-ISTA/2025.08, project Proj-ISTA-2025.07, 21 July 2025 decision).

The approved protocol covers simulated sessions with healthy adult volunteers, immediate pseudonymization, secure institutional storage with access control, and deletion of identifiable source videos within six months after data collection concludes; written informed consent is provided using the committee's model.

3.3 Population characterization

The distribution for the population of volunteers was defined to ensure heterogeneity in terms of age, sex, and physical appearance, thereby replicating the heterogeneity of patients one would typically encounter in pre-hospital emergency settings. Fifty-two healthy adult volunteers participated in the study, with no known neurological or chronic conditions likely to

hamper the assessment of responsiveness. This was required in order to avoid systematic biases in the simulation of responses to verbal or painful stimuli.

The group was made up of twenty-nine males and twenty-three females to ensure a balanced gender representation. The subjects' age ranged from eighteen to sixty-five years, with the average age at about thirty-seven years. The young adults were included to gather rapid and expressive motor responses, while older adults were intentionally represented so that slower or less apparent tendencies of reaction would be possible. This heterogeneity is particularly relevant as emergency triage systems must be able to generalize to a wide variety of patients differing in physiological and expressive characteristics. The age distribution of the dataset is presented in Table 3.1.

Table 3.1. Age distribution of the 52 volunteers included in the AVPU dataset.

Age group (years)	Number of participants	Percentage (%)
18–25	10	19.2
26–35	12	23.1
36–45	13	25.0
46–55	10	19.2
56–65	7	13.5
Total	52	100.0

This distribution shows that more than half of the subjects were over the age of thirty-five years old, ensuring that the dataset will include a wide range of slower or more subdued responses typically included within older participants. At the same time, many younger subjects were also included, providing samples of more emotive and physically agile reactions.

Thus, this level of demographic variation ensures that the dataset provides a solid foundation for training and testing computer vision models for AVPU classification. It encompasses variability in facial composition, movement patterns, and expressiveness across different ages and genders, making it more likely that the models generalize beyond the controlled simulation setting and perform well in real emergency scenarios.

3.4 Summary

This chapter described the creation of a custom video dataset to support automated AVPU classification, from preparation to annotation. First, we motivated the need for a bespoke resource and recorded approximately 132 video sequences (~2 hours) in a controlled simulation setting, balancing realism with ethical and operational constraints. The scenario design

followed standard pre-hospital procedures and targeted clinically meaningful microactions-light touch, pain stimulation, eye/head movements-such that stimulus-response dynamics central to the AVPU could be observed. Using Label Studio, clinical experts annotated each sequence at the microaction level with precise onset and duration, thus enabling temporally aligned ground truth usable by spatiotemporal learning. Inter-annotator agreement by Cohen's kappa confirmed the reliability of the annotations, while the entire study proceeded under approved ethical protocols, including consent, pseudonymization, access control, and limits on retention. Finally, population characterization ensured demographic heterogeneity, improving the dataset's representativity and model generalization potential: 52 healthy adult volunteers aged between 18 and 65 years, well balanced by sex and age groups. The established dataset was curated and ethically approved, reliably annotated, and forms the empirical basis for the prototype development described in the next chapter.

Chapter 4. System Development

Since prototype development involves creating an artifact that addresses a specific clinical issue—timely and accurate monitoring of the patient's state of awareness based on video observation—it adhered to the DSRM methodology guidelines. The primary objective is to help medical personnel automate the observation of microactions related to the AVPU scale such as reaction to touch, pain behavior, or eye movement.

4.1 System requirements

Functional/non-functional requirements have been set before implementation by consulting with experts from emergency medicine to ensure the system is clinically acceptable, technically feasible, and of high ethical standards.

Operationally, the system is required to recognize the discriminative microactions associated with each responsiveness phase, representing them as temporally bounded events, and integrate pose, facial, and contextual features extractable from video. Based on these observations, the system can then infer the corresponding AVPU score.

In addition to these requirements, the prototype was conceived with a view to future real-time or near real-time operation, although no formal latency measurements were performed in the scope of this Dissertation. The system was also designed to comply with the privacy obligations established by the ethics committee and to adopt a modular architecture, enabling each component to be developed, evaluated, and improved independently.

4.2 System architecture

The design employs a modular processing architecture where each module performs different computational tasks that cumulatively work towards a broader goal. The use of such architecture improves scalability, allowing the system to have modules easily replaced, modified, or even removed with no significant impact on the system architecture.

Figure 4.1 shows a schematic representation of the system’s architecture, showing the flow of information among the different modules, as well as the path that leads to the final result. The four major stages of the pipeline are identified. The first is where the input is acquired in the form of video sequences showcasing interactions during an emergency scenario. The second is the microaction detection stage, where facial features, body pose, and microactions such as touch, pain, eye opening or closure, are scrutinized by trained deep learning-based computer

vision models. The third is where the microactions are combined for the final result, which is an estimation of the AVPU score, along with logs that are meant for analysis.

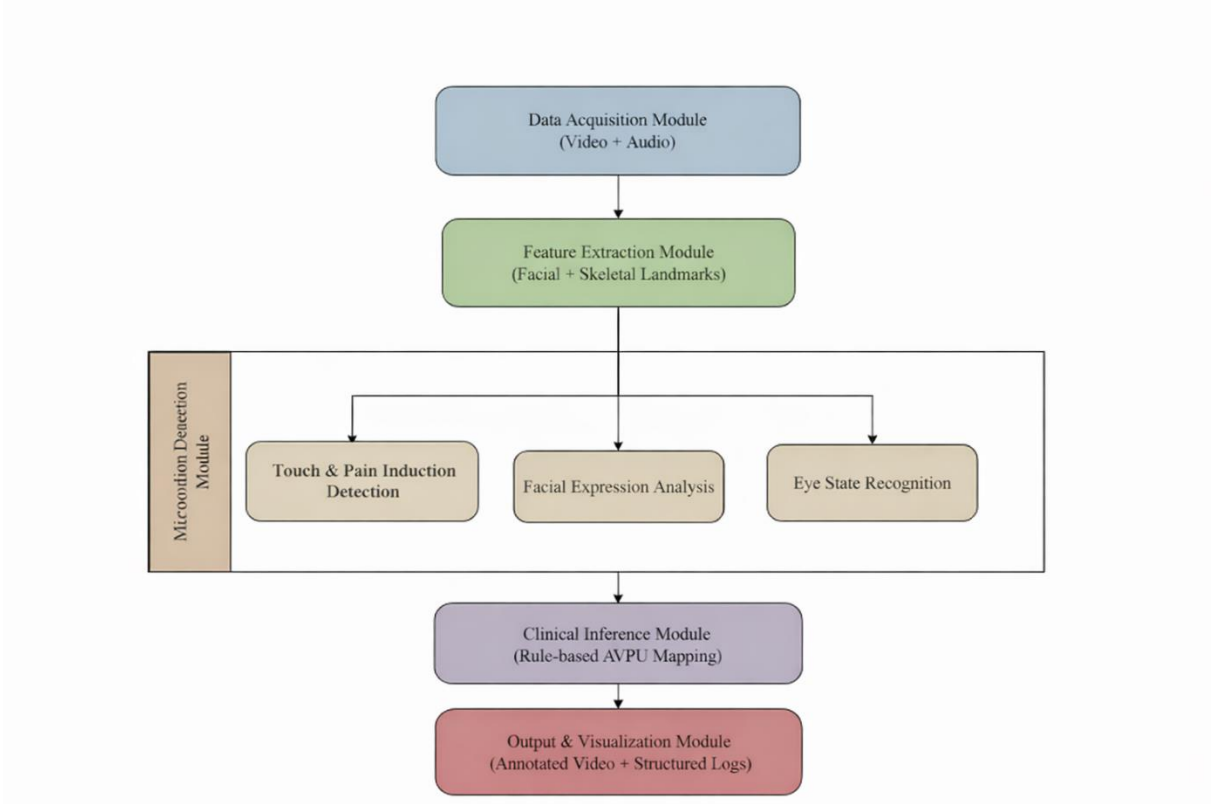


Figure 4.1. Architecture of the proposed system for AVPU-based assessment.

Finally, the clinical inference and visualization stage aggregates the detected microactions to produce a final binary classification of the patient as conscious or unconscious. The stage also generates annotated video outputs synchronized with the original audio and formatted logs for offline analysis. The audio stream is preserved exclusively for visualization and review purposes and is not used as an input to the classification models. Consequently, although the proposed approach is inspired by the AVPU assessment process, it does not implement a complete AVPU evaluation, particularly because verbal responses are not analysed directly from the audio signal.

4.2.1 Data Acquisition Module

The first stage of the pipeline is the acquisition process, where the scenarios are acquired in the form of video. Since a temporally dependent assessment such as AVPU is based on the response

of the patient, the video frames are processed one by one to preserve the continuity of behavior observations as well as the temporal relationship of patient responses.

Audio track preservation is also important, as vocalization and speech attributes (e.g., patient statements, tone, and voice) are useful in interpretation. However, the primary video processing component used by this prototype does not support audio recording while video outputs are annotated. To address this issue, the approach that will be detailed later in Section 4.3 was used, allowing all simulated interactions within a triage scenario to be replayed with full audio support. The Data Acquisition Module therefore serves as the entry point of the overall system pipeline, as illustrated in Figure 4.1.

4.2.2 Feature Extraction Module

After the raw video is acquired, the resulting data is processed by a Feature Extraction Module to extract the visual information that can be used for further analysis. Using the MediaPipe library, the Feature Extraction Module extracts keypoints corresponding to the body skeleton, as well as landmarks on the face, from the video frames.

This is a critical component in ensuring that the inputs are properly structured for model inference. The spatial features, which include the facial structure, are supplemented with temporal descriptions that convey the behaviors, such as head and arm movements, as well as facial expressions. Additionally, ROI (Region-of-Interest) segmentation, such as the eyes, face, and torso, helps reduce variations in the background and focuses on significant visual elements.

The result of this module consists of sequences of landmarks, which are then fed into the Microaction Detection Module.

4.2.3 Microaction Detection Module

The core of the system is the microaction detection module, which is responsible for detecting small but significant movements that are essential in AVPU scoring. This module is based on the hypothesis that a set of observable microactions can quantitate patient responsiveness.

For this purpose, the module is divided into three self-contained tasks that use deep learning models trained on labeled data.

The initial task, touch detection, which includes the induction of pain, analyzes cases where a healthcare professional delivers a gentle touch stimulus or a stronger gesture with the intention of inducing pain. The detection is performed through two different mechanisms: (i) an algorithm trained to identify temporal patterns of skeletal points to detect touch, and (ii) an

algorithm that identifies spatio-temporal patterns in local regions to detect a gesture that induces pain.

The second task, interpretation of facial expressions, decodes patient responses in facial behavior signals. The solution utilizes a model that analyzes temporal sequences of facial landmarks to decode subtle expressions such as frowning, grimacing, which are indicators of distress or pain.

The third task is the eye state recognition, which identifies whether the patient's eyes are open or closed, an important sign regarding the patient’s consciousness.

4.2.4 Clinical Inference Module

The outputs produced by the Microaction Detection Module are not sufficient, on their own, to determine a patient’s level of consciousness. To be clinically meaningful, they must be interpreted within a structured decision process that reflects how AVPU is applied in practice. This interpretation is performed by the Clinical Inference Module, which maps the system’s technical predictions to AVPU categories by applying a ruleset derived from written triage protocol guidelines.

The module combines the detected microactions and their temporal relationships to generate one AVPU label for each time interval. The underlying logic follows the stimulus–response sequence of AVPU and prioritizes the earliest reliable indicator of responsiveness. If spontaneous or purposeful movement is detected without any preceding stimulation, the patient is classified as Alert. If a response such as head movement or eye opening occurs after a verbal prompt and/or a light touch, the patient is classified as Verbal. If, following a pain-inducing stimulus, the system detects a pain-related reaction (for example, a pain facial expression and/or a movement response), the patient is classified as Pain. If no meaningful response is detected after stimulation and the absence of response persists beyond a predefined time threshold (set to five seconds in the prototype), the patient is classified as Unresponsive.

As illustrated in Figure 4.2, the AVPU assessment implemented in this module follows a stepwise procedure. The process begins with an observation of the patient’s baseline state, verifying whether the patient appears awake without any external stimulus. When wakefulness is sustained continuously for at least one minute, the assessment can be concluded and the patient is recorded as Alert. If this condition is not met, the system proceeds to subsequent stages of the protocol, in which progressively stronger stimuli are applied and the corresponding responses are evaluated. The AVPU label is then assigned based on the first stage in which a consistent, clinically meaningful response is detected within the corresponding time window.

In case the patient is not conscious, the following action is the use of verbal or touch stimuli. If the patient shows a response, the same time-based criterion is applied again. The patient must be conscious for one minute. If this is not the case or there is no reaction, it proceeds to the next level of assessment.

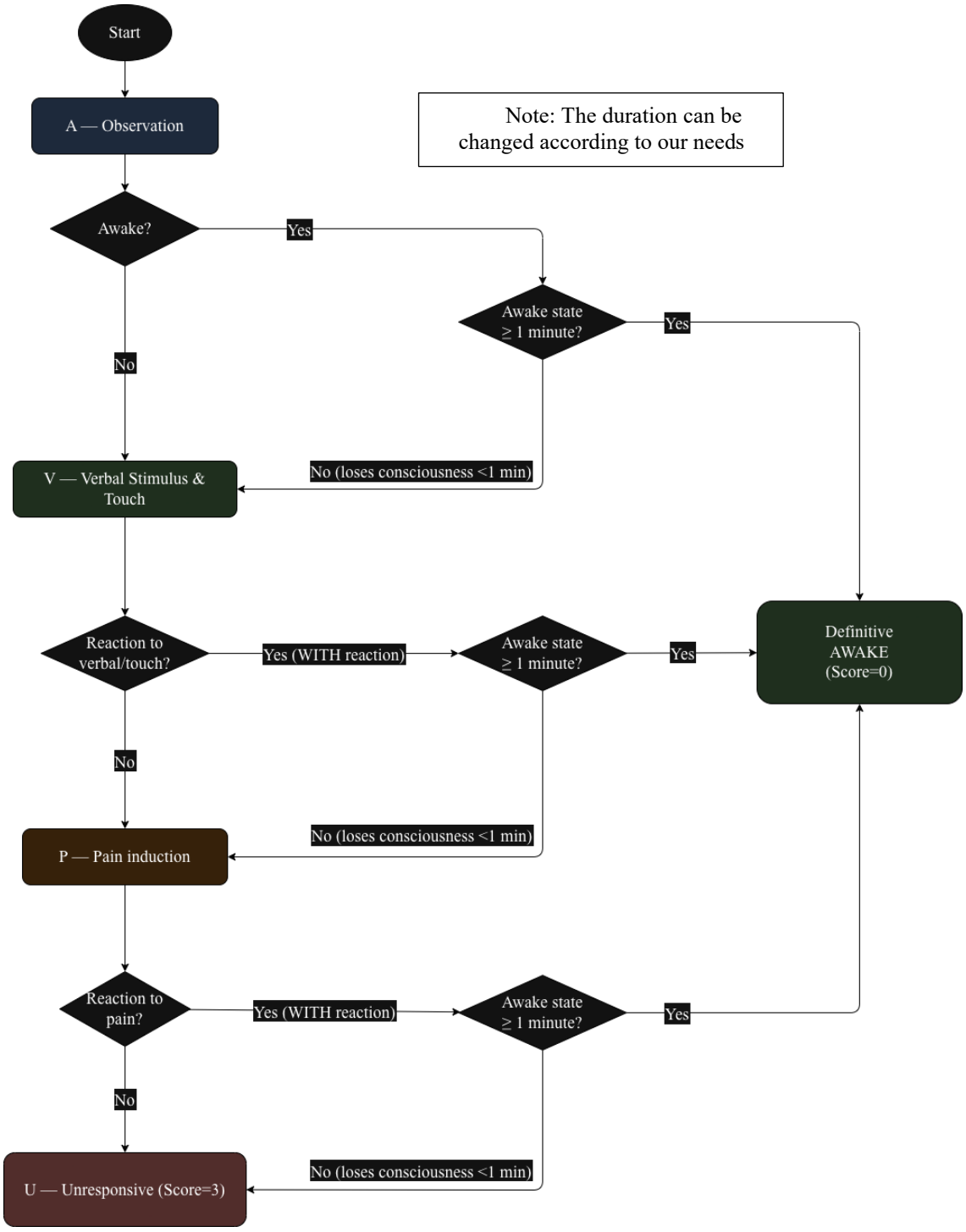


Figure 4.2. Flowchart illustrating the automated AVPU classification process.

In this phase, painful stimuli are used, preferably through valid clinical maneuvers such as trapezius squeeze or sternal rub. The patient must remain conscious for a predefined period of time. If the patient does not react or the reaction is so brief that keeping consciousness becomes difficult, the level will be marked as "unresponsive." This level corresponds to the lowest level of the AVPU scale.

The whole logic of this flow chart is sequential and progressive. As far as the patient does not meet the criterion of being awake for a predefined period, the level of assessment moves to the next level of stimuli. Therefore, all levels of AVPU could be assessed systematically with this logic. This logic guarantees a reliable medical approach to the determination of the level of awareness.

4.2.5 Output and Visualization Module

The final stage of the pipeline provides the system's result in a clinically relevant, transparent manner. The Output and Visualization Module produces labeled videos where the microactions identified by the system, as well as the corresponding AVPU classification, are marked upon the original video. Thus, the system provides not only the classification result but also the microactions that support it.

The preservation of the original audio is useful additionally from the perspective of interpretability, which helps a clinician confirm the correctness of the microaction classification generated by the system on the basis of vocal responses, as well as sound cues delivered by the patient. The system also has the advantage of providing high-resolution logs that keep track of microactions identified by the system with a high degree of precision.

4.3 Technological implementation

Following the specification of system requirements and architecture, the prototype was developed using OpenCV as the main framework for processing the video. OpenCV allowed frame-by-frame processing of clinical simulation videos, from file-reading stages to output stages. The incorporation of OpenCV with MediaPipe allowed the system to extract both body pose key points as well as facial key points, which constitute the foundation of the Microaction Detection Module.

Each video frame was marked with the information resulting from the inference models, layering visual cues on top of it, and each identified activity was represented by its corresponding AVPU classification. Additional care was taken to preserve the original audio track, as it contributes to the realism and interpretability of the recordings. Since OpenCV does

not retain audio when writing processed video files, a complementary Python library, MoviePy, was integrated into the pipeline to reattach the original audio to the annotated video stream. In practice, OpenCV generates the annotated, silent video, and MoviePy then combines this video with the corresponding audio track, ensuring synchronized playback for subsequent analysis and demonstration.

This modular and transparent design helped ensure that every technological decision was directly aligned with the clinical and methodological needs established earlier. The result is that the system design and development comply with the tenets of DSRM, where the technological artifact functions as both proof of concept and tool for exploring viability of AVPU assessment by AI.

To ensure a rigorous evaluation and minimize the risk of identity and temporal leakage, the dataset was partitioned at the subject level. All videos belonging to a given volunteer were assigned exclusively to either the training or testing subset, ensuring that no individual appeared in both sets. The split was performed before the creation of temporal windows and derived samples, preventing overlapping segments from being distributed across subsets. A total of 70% of the subjects were allocated to the training set and the remaining 30% to the testing set. Class proportions were preserved as closely as possible during the split. No data augmentation was used. All preprocessing, feature extraction, and model training procedures were conducted independently within each subset. This evaluation protocol was designed to assess the ability of the models to generalize to previously unseen individuals.

4.4 Microaction algorithms

Identifying which actions will provide the most reliable indicators of a patient's level of awareness is a central part of the prototype in building. The key microactions selected to capture this include touching, pain induction, head turning, eye opening-even if this is subtle-frowning-with or without overt pain responses-and movement, in some cases, of other body parts like arms.

These microactions were selected based on the recommendation of experts and due to their direct relevance for AVPU assessment, ensuring that the system is trained to recognize behavioral patterns that are clinically meaningful for classification. Through focusing on these key moments, the system prioritizes the informative visual cues and reduces sensitivity to irrelevant motion that would introduce noise and reduce the accuracy otherwise. Besides,

temporal localization of these events supports learning of the characteristic stimulus-response patterns associated with different responsiveness states.

In order to facilitate an efficient recognition of these microactions, three deep learning models are developed, with each being trained on a set of different representations of the input. In the first model, touch actions are identified based on a series of skeletal pose landmarks. In the second, actions that induce pain are identified from pose-based motion dynamics. In the third, expressions that are linked with pain are identified from a series of facial landm

The inference process is controlled by a set of rules which are clinically guided, assigning a particular AVPU level to a set of microactions identified. For example, the presence of a touch event with a facial expression indicating pain, resulting from a pain-inducing maneuver, is considered a pain response, assigning the patient to level “P” (Pain). The presence of a touch event but no corresponding pain response indicates verbal or light touch stimulation responsiveness, assigning the patient to level “V" (Verbal). The presence of a pain-inducing maneuver indicates responsiveness to pain-related stimuli. If there was no significant activity recorded for longer than a defined period of time, the patient would be rated at level “U” (Unresponsive). On the other hand, presence of spontaneous events/activities without external stimulation would imply alertness, which would rate the patient at level “A” (Alert).

The prototype also has a real-time annotation component that produces videos marked with the labels of the performed actions and the corresponding AVPU status. The annotated videos not only offer an image of the assessment performed by the system, but they also have audio. The system also records the AVPU status of each frame of the video as an array of time series data, which can then be evaluated. This development stage shows how computer vision, machine learning, and medical knowledge come together in the production of an artifact that provides technological functionality as well as medical utility. Through aligning microaction detection with the AVPU scale by embedding this logic into the real-time video processing system, the prototype provides an important foundation that can be built upon in improving the practice of emergency assessment and triage. The system represents an instantiation of the design artifact according to the DSRM framework, which applies both to medical theory-building as well as medical innovation.

4.4.1 Touch / No Touch

In clinical assessment, light tactile contact is typically the first physical interaction used to evaluate a patient’s responsiveness. This contact is intended as a gentle initial cue, commonly applied to the patient’s shoulder, to establish engagement and assess basic reactivity. As

illustrated in Figure 7, this contact is performed using an open hand with a light touch applied to the upper body of a subject positioned in a reduced state of alertness. Thus, accurate recognition of this contact in the video is essential for automating the medical assessment of a patient's level of consciousness.

To detect and distinguish between "Touch" and "No Touch" events, the models listed and briefly described in Table 4.1 were tested. Each model exploits a different representation of motion and the spatial information present in the video frames.

Table 4.1. Comparison of models for Touch vs. No Touch detection.

Model	Architecture	Main Advantages	Main Limitations
CNN + LSTM	3D feature extraction using CNN followed by LSTM	Captures both spatial and temporal cues; strong visual interpretability	Sensitive to illumination and background variation; computationally demanding
Pose-Based	Landmark extraction via MediaPipe	Fast inference; invariant to lighting and appearance	Ignores temporal continuity; less accurate for subtle movements
Pose + LSTM	LSTM applied to temporal sequences of pose vectors	Robust to visual noise; generalizes across subjects	Requires accurate landmark tracking; limited by pose estimation errors
TSM	CNN backbone with temporal shift operation	Lightweight and efficient; suitable for real-time deployment	Limited ability to model long-term motion dependencies

The first approach combines the use of Convolutional Neural Network (CNNs) and Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997). A CNN extracts frame-level spatial embeddings (e.g., position, touch area, body posture) from short video sequences, and an LSTM models their temporal dynamics to recognize "Touch" motion stimulus effectively over a few frames.(Núñez et al., 2018).

The second technique hinges solely on human pose estimation (Fu et al., 2023). Using MediaPipe Pose, the method extracts per-frame 3D (x, y, z) keypoints for joints (e.g., wrists, shoulders, elbows) to form a compact kinematic representation that reduces sensitivity to background/lighting and isolates the core "hand approaching the body" motion.

Based on the pose representation, a sequential learning element is incorporated through an LSTM network, which captures the consistency of human motion (Li et al., 2016). Feeding the temporally ordered 33 pose landmarks into an LSTM enables it to learn repetitive arm-displacement dynamics—smooth touching and retracting cycles—thereby robustly recognizing generalized "Touch" patterns.

Finally, a lightweight Temporal Shift Module (TSM) model was implemented to evaluate an alternative temporal modeling paradigm. Instead of explicit recurrence, the TSM introduces temporal reasoning directly within the convolutional layers by shifting part of the feature maps along the temporal dimension (Lin et al., 2018).

4.4.2 Pain induction

As a follow-up of the physical contact, the next level of AVPU is the induction of pain, which determines whether a patient does not react to verbal or touch stimuli by finding out if the patient responds to painful stimuli.

Figure 8 indicates how the examiner applies the localized pinch of short duration on the trapezius region by employing the 'pincer-shaped' position of the hands. This task corresponds to the Pain level of the AVPU scales. The activation of the least possible reflex or voluntary reaction is the intention of this level. The determination of this automated reaction in the recorded video is considered indispensable in distinguishing among unconscious responses and those of painful perceptions.

For this task, three different pose models for capturing temporal information have been considered. These represent distinct approaches for learning motion dynamics from skeleton data: Pose + LSTM, Pose + Transformer, and Temporal Convolutional Network (TCN).

The Pose + LSTM network is an extension of the network design used in touch detection in order to tackle more abrupt localized motion that corresponds with the infliction of pain (Fu et al., 2023; Li et al., 2016). The MediaPipe's Pose solution is applied on every frame of the video, which provides the 3D location of the 33 landmarks on the entire body (Lugaresi et al., 2019). Pose landmarks at joints (e.g., wrists, elbows, shoulders) provide a compact, texture-invariant representation of body location, and frame sequences are segmented into fixed windows (typically ~1 second ≈ 10 frames) for input to a two-layer LSTM. The LSTM processes the sequence step by step to learn the full pinch dynamics—from arm motion into a pincer grip through contraction—capturing both the acceleration before contact and the relaxation after contact via temporal recurrence.

In the Pose + Transformer network, an attention mechanism makes use of the frame relevance in a more flexible or 'parallel' fashion (Rajasegaran et al., 2023). Each frame's 99-dimensional landmark vector is projected into a higher-dimensional embedding space and processed by a lightweight transformer encoder whose self-attention emphasizes the most informative frames, particularly around the pincer motion and contact. A learnable

classification token aggregates global temporal context for low-cost binary classification (pain vs. no pain), improving cross-subject and cross-view generalization while providing more interpretable temporal salience than an LSTM.

The third approach, TCN, uses temporal convolutions instead of recurrence (Bai et al., 2018). The network relies on the same pose sequences as it did on the input Data, using one-dimensional convolutions with progressively larger dilation values. The TCN network is thus able to capture both short-term dependencies as well as long-term dependencies on the input data. The dilations of the convolutions can be seen as 'receptive fields' on the time axis. The TCN network can successfully capture the 'fast closing and release motion' corresponding to the induction of pain. The TCN's convolutional structure has many benefits. It's computationally efficient, can be computed in parallel, and is stable during training. Also, its hierarchical structure automatically captures variations of motion's frequency as well as its amplitude in a non-sequential fashion. This is useful for finding the small amplitude motion with large energy corresponding to a pinch among more continuous or slower motions. Table 4.2 presents a comparative summary of the three approaches analyzed for differentiating Pain Induction from No Pain Induction interactions using various input representations and temporal modeling techniques.

Table 4.2. Comparison of models for Pain induction vs. No Pain induction.

Model	Architecture	Main Advantages	Main Limitations
Pose+LSTM	Two-layer LSTM followed by dense layer	Captures detailed motion dynamics; interpretable temporal flow	Sensitive to sequence length and frame alignment
Pose+Transformer	Transformer Encoder (2 layers, 4 heads)	Models global dependencies; robust to subject variability	Higher computational cost; needs more data for stability
TCN	1D Dilated Convolutional Network	Fast, parallel, and stable; suitable for short, intense actions	Limited modeling of very long temporal dependencies

4.4.3 Combined Gesture Recognition: Touch and Pain Induction

After validating each gesture detection model independently, a final set of experiments was conducted to evaluate the joint recognition of Touch and Pain Induction gestures within a unified computational framework. The task was aligned with the clinical process of AVPU responsiveness assessment. As indicated earlier, the AVPU process begins with checking the eye response of the patient. The individual is then stimulated through touch. Painful stimuli can then follow. The ultimate requirement was therefore to identify the best approach in modeling

time series data from Touch and Pain Induction gestures learned jointly or in a hierarchical fashion.

The four supporting architectures considered will be the TCN, Bidirectional LSTM, the Transformer Encoder, and a two-stage Cascade LSTM. Each of these approaches was deployed on the same dataset of tagged clinical interactions. For this dataset, landmarks based on MediaPipe Pose (33 keypoints × 3 coordinates) on each frame of the video presented Touch, Pain Induction, or Nothing.

The TCN network was the benchmark network used for the task of unified multi-class gesture recognition (Bai et al., 2018). The network used a stack of one-dimensional convolutions with an incremental list of dilation factors that automatically captured both short-term and long-term dependencies in motion. The flattened pose sequences underwent convolutions followed by dense layers for classification into the three classes of gestures. The proposed architecture is able to capture both bursty motion involved in the induction of pain as well as smoother motion patterns involved in touch. It is thus able to perform efficient feedforward processing of motion features.

The Bidirectional LSTM (BiLSTM) architecture incorporates explicit recurrence on the sequence level by processing pose vectors in both forward and reverse time (Li et al., 2016; Schuster & Paliwal, 1997). This allows BiLSTM to capture temporal transitions, such as approaching touch and sharp contractions linked to pain induction, making motion-phase changes easier to interpret. Although more computationally intensive than a TCN, it offers flexible modeling of temporal continuity and dynamics for finer-grained transition distinctions.

The Transformer Encoder uses a self-attention mechanism to capture global temporal dependencies (Rajasegaran et al., 2023). Each landmark vector is projected into a higher-dimensional space and processed through transformer encoder layers with sinusoidal positional embeddings. Attention assigns relative importance to frames, highlighting key moments such as motion onset or contraction during pain induction. This global sequence modeling improves generalization across individuals and motion variability, albeit with higher computational cost.

Also, a two-stage Cascade LSTM network was designed to emulate the hierarchical logic of the AVPU protocol (Kolic et al., 2015), following the broader trend of hierarchical recurrent modeling in action recognition (Li et al., 2016). For this setup, one LSTM classifier was employed exclusively to identify Touch, while another binary LSTM was used to identify Pain Induction in sequences where there was no touch. The structure of this sequence of decisions corresponds to the step-by-step process of medical diagnosis. The cascade model of decision making breaks down the task into two simpler ones, allowing it to consider the processing

characteristics of human stimuli. Table 4.3 provides a comparative summary of the four approaches evaluated for combined gesture recognition, integrating Touch and Pain Induction Detection through different input modalities and temporal modeling methods.

4.4.4 Eyes open/closed

For establishing the level of the patient's consciousness using the AVPU scale, the state of the eyes plays an important role. The fact that the eyes can be opened spontaneously or in response to stimuli can be considered an important sign of the level of awareness or responsiveness of the nervous system. Patients considered alert or responsive to verbal stimuli can display eye opening, either spontaneously or in response to stimuli. However, this does not apply to unresponsive patients, who will retain their eyes closed with no reaction to stimuli.

Table 4.3. Comparison of models for combined Touch and Pain Induction recognition.

Model	Architecture	Main Advantages	Main Limitations
TCN	Dilated 1D temporal convolutions + MLP head	Efficient, stable training; good at localized motion bursts; real-time friendly	Less expressive for very long-range dependencies; fixed receptive field per stack
BiLSTM	Bidirectional LSTM + linear classifier	Captures gradual transitions and sequence order; interpretable temporal phases	Slower training/inference; may overfit on small datasets; sensitive to sequence length
Transformer Encoder	Multi-head self-attention encoder + classifier	Models long-range dependencies; attention offers interpretability; parallelizable	Higher memory/compute; benefits from larger datasets and tuning
Cascade LSTM (Touch → Pain)	Two LSTM classifiers in sequence (binary→binary)	Aligns with clinical workflow; modular and debuggable per stage	Error propagation between stages; tuning thresholds adds complexity; slightly higher latency end-to-end

Two methodologies were experimented for tackling this problem: Eye Aspect Ratio (EAR) calculation and classification using CNNs. The EAR is a geometric approach that computes the ratio of key distances from eye landmarks to ascertain the state of the eye (open or closed). The EAR is efficient, interpretable, and thus useful for real-time processing. However, this method is highly susceptible to variations in lighting levels, head position, or obstructions, which can be encountered frequently in triage settings. Hence, its effectiveness could potentially be compromised in real-world triage settings, which could pose unpredictable movements by the patients or conditions of the video. EAR can be computed as:

$$EAR = \frac{\|p_2 - p_6\| + \|p_3 - p_5\|}{2\|p_1 - p_4\|}, \tag{4.1}$$

where the distance $\| p1 - p4 \|$ corresponds to the horizontal width of the eye, used as normalization to account for different eye sizes; $\| p2 - p6 \|$ represents the vertical opening at the left side of the eye; and $\| p3 - p5 \|$ represents the vertical opening of the eye at its right side. These are illustrated in Figure 4.3.



Figure 4.3. Keypoints of EAR.

The EAR calculates the average vertical lid distances divided by the eye wdth, where an opened eye has a larger EAR value, while a closed eye has a smaller EAR value. If the eye is opened, then the vertical distances will be larger, resulting in the EAR value being higher. As the eye is closed, the value of the vertical distances tends closer to zero. Hence, the value of EAR will decrease. The purpose of the EAR is thus achieved by enabling a clear distinction to be made, not only between an open eye and a closed eye, but also an almost closed eye.

Therefore, the CNN learns discriminative eye-state features directly from raw eye images, capturing spatial detail beyond geometric measurements. Trained on an eye-region ROI to classify eyes as open or closed, it can detect subtle cues such as partially closed eyelids and expression-driven variations.

4.4.5 Pain expression detection

Complementary approaches were developed to detect pain from facial video data, combining deep learning techniques with feature-based methods.Deep spatiotemporal models-R(2+1)D, R3D, Swin3D, and TimeSformer-operate directly on short sequences of frames to learn appearance and motion cues like brow contraction, eyelid tightening, and other microexpressions without explicit feature engineering, returning a probability of pain for each video segment (Bertasius et al., 2021; Gomutbutra et al., 2022; Hara et al., 2017; Liu et al., 2021; Othman et al., 2023; Tran et al., 2017). An alternative, more interpretable approach uses facial Action Units (AUs) extracted with OpenFace and classified using models such as logistic regression or random forests, focusing on those linked to pain responses. When AU extraction is unreliable, geometric features from MediaPipe landmarks, such as eye aspect ratio (EAR) for eyelid closure, are used as proxies. Overall, these methods complement deep learning approaches by relying on explicit clinical and anatomical indicators of pain.

Chapter 5. Evaluation and Discussion

In this chapter, the performance of the proposed prototype is judged based on quantitative metrics and feedback from healthcare experts. The quantitative component is mainly the accuracy of the microactions detection models, namely touch induction, pain induction, as well as the overall end-to-end classification of whether a patient is conscious or not based on unseen videos from the self-created dataset. Train/test was conducted with 70%/30%.

For each binary classification problem, True Positives (TP) refers to the cases where the system has correctly identified the presence of the event, such as when it indicates the induction of pain because it has actually been administered. True Negatives (TN) refer to cases that have been identified correctly by the system, such as when it indicates no pain because no pain has been administered. The False Positives (FP) refer to false identifications when the system indicates the event has happened when it has not actually happened. The FN refer to cases when the event has occurred but has not been identified.

Accuracy provides an overall view of whether the prototype correctly classified the presence or absence of the target event. It can be computed as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \tag{5.1}$$

Precision measures how often detected positives are correct. High precision is clinically relevant because false alarms may consume attention and resources, particularly in time-sensitive environments. On the other hand, recall measures the proportion of true events that were detected. They can be computed as:

$$Precision = \frac{TP}{TP+FP}; \tag{5.2}$$

$$Recall = \frac{TP}{TP+FN}. \tag{5.3}$$

Because precision and recall typically trade off against each other, the F1-score is used to provide a balanced summary of both metrics computing their harmonic mean. This is particularly relevant in triage settings, where the system must simultaneously avoid excessive false alarms while also minimizing missed clinically meaningful responses. For that reason, this evaluation reports all four metrics (accuracy, precision, recall, and F1-score). F1-score can be computed as:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{5.4}$$

On top of this set of evaluation metrics, the evaluation of the prototype has been carried out on a hold-out subset of the custom dataset, which accounted for around 30% of the total number of 132 videos, to assess robustness. This assessment can facilitate the evaluation of correctness

in conjunction with the cost implications of incorrect microactions, as well as the incorrect determination of the final state of consciousness.

The interaction recognition component was evaluated across three scenarios: touch-only (light stimuli for Verbal-level assessment), pain-only (intensive stimuli for Pain-level assessment), and combined touch–pain interactions simulating real triage conditions. Model parameterizations and confusion matrices for touch-only are provided in Annex A.

Table 5.1. Model performance for detecting touch-only interactions.

Model	Accuracy	Precision	Recall	F1-score
CNN+LSTM	0.67	0.67	0.66	0.66
Pose	0.77	0.76	0.61	0.61
Pose+LSTM	0.84	0.84	0.84	0.83
TSM	0.80	0.40	0.50	0.44

The results for the detection of touch only interactions are presented in Table 5.1. Pose+LSTM showed the better results, achieving 84% accuracy by effectively capturing temporal dynamics in skeletal landmarks. This suggests that combining pose estimation with recurrent networks is suitable for detecting subtle tactile interactions. The lower performance achieved in TSM (Temporal Shift Module) may stem from its sensitivity to short-sequence noise in simulated data. Model parameterizations and confusion matrices are presented in Annex B.

Table 5.2. Model performance for detecting pain induction only interactions.

Model	Accuracy	Precision	Recall	F1-score
Pose+LSTM	0.74	0.37	0.50	0.43
Pose+Transformer	0.75	0.38	0.50	0.43
TCN	0.74	0.37	0.50	0.43

As shown in Table 5.2, the performance when detecting pain induction only interactions, was moderate across all models, with accuracy values in the range 74-75%. Precision and recall were lower, indicating challenges in distinguishing pain induction from background motion. The Transformer variant showed marginal improvement, likely due to its attention mechanism handling spatio-temporal variations better, but overall results highlight the need for more diverse training data to reduce FPs.

Table 5.3. Model performance for detecting both interactions.

Model	Accuracy	Precision	Recall	F1-score
TCN	0.90	0.90	0.90	0.90
BiLSTM	0.79	0.78	0.77	0.78
Transformer	0.77	0.75	0.76	0.75
Simple cascade Touch -> Pain induction with LSTM	0.86	0.78	0.77	0.77

The incorporation of the touch detection and pain stimulation process into a single temporal sequence model showed significant improvement in performance, as depicted in Table 5.3. TCN performed best with overall accuracy, precision, recall, and F1-score values of 0.90, which confirms that the temporal convolution mechanism is able to identify short-range and localized interactions that exist in clinician-patient interactions when touch and pain are close together in time. The simple cascading system (Touch -> Pain with LSTM) also performed well (accuracy = 0.86), which indicates that although a clinically motivated sequence-based architecture is useful, the precision and recall are reduced compared to that of the best-performing TCN, which might be because of error accumulation from the previous to the latter stages. BiLSTM and Transformer produced moderate results with accuracies of 0.79 and 0.77, respectively, which indicates that, for this particular problem, capturing the fine, temporal, local dependencies is more essential than learning the overall temporal dependencies of a sequence. The observations clearly demonstrate that a single overall temporal sequence modelling approach provides a clearer vision regarding the triage process when compared to distinct models for Touch and Pain Induction Detection. Model parameterizations and confusion matrices are presented in Annexes D and E.

In parallel, additional complementary cues were evaluated to support the workflow in cases where gross motor responses are weak or ambiguous, namely facial pain expression and eye state. Table 5.4 and Table 5.5 present the results for the microaction classifications (pain expression detection and eye open/closed detection).

The results show moderate and relatively similar performance across all pain expression detection methods, with accuracy and F1-scores between 0.54 and 0.60. Among the deep models, Swin3D achieved the best overall results, followed closely by TimeSformer, suggesting that transformer-based spatiotemporal modeling is well suited to capturing subtle pain-related facial dynamics. OpenFace-based classifiers performed slightly worse but remain competitive, with Random Forest outperforming Logistic Regression, while the MediaPipe EAR-based geometric approach achieved the lowest scores, as expected given its reliance on a single eyelid-related feature. Overall, the table highlights that facial pain cues alone are

informative but not highly discriminative, supporting their use as a complementary rather than standalone modality.

Table 5.4. Performance comparison of the models for pain expression detection.

Model	Accuracy	Precision	Recall	F1-score
R(2+1)D	0.59	0.58	0.57	0.57
R3D	0.56	0.55	0.56	0.55
Swin3D	0.60	0.60	0.59	0.59
TimeSformer	0.58	0.57	0.58	0.57
OpenFace + Logistic Regression	0.55	0.56	0.54	0.55
OpenFace + Random Forest	0.57	0.58	0.56	0.57
MediaPipe (EAR-based geometry)	0.54	0.53	0.55	0.54

Table 5.5. Performance comparison of the models for eye open/closed detection.

Model	Accuracy	Precision	Recall	F1-score
CNN	0.92	0.91	0.93	0.92
EAR	0.81	0.80	0.83	0.81

The results for the eye openness detection show that the CNN-based approach achieves significantly higher performance than the EAR method in detecting whether the eyes are open or closed, indicating superior robustness to variations in head pose, illumination, and partial occlusions.

Having established the performance of the individual microaction components and these complementary cues, the next step is to evaluate the end-to-end performance of the full AVPU-inspired workflow. Table 5.6 therefore summarizes the final classification results (conscious versus unconscious), comparing alternative sequence-modelling architectures used to aggregate the detected microactions into a single decision. Model parameterizations and confusion matrices are presented in Annex F.

Table 5.6. Final prediction of conscious or unconscious after workflow.

Model	Accuracy	Precision	Recall	F1-score
TCN	0.92	0.93	0.91	0.92
BiLSTM	0.81	0.80	0.79	0.80
Transformer	0.78	0.76	0.77	0.76
Simple cascade Touch -> Pain induction with LSTM	0.87	0.79	0.78	0.78

The TCN performed better in the overall integrated environment with 90% accuracy. This is because the convolutional layers of the network are capable of handling localized motion,

which is a significant advantage that served the entire process quite well. In the entire process, the best overall performance was yielded by the TCN in the final classification of the state of consciousness (Accuracy=0.92, Precision=0.93, Recall=0.91, F1-Score=0.92). The outcome has suggested that the process led to a slight improvement in the overall recall rate for the unconscious state of the patient, relying on microactions within the overall AVPU-inspired logic. The cascade approach also performed well (accuracy=0.87), proving that hierarchical inference, involving touch detection before pain, can be used within a clinical setting. It is, however, worth noting that although the accuracy is high, the recall rate (0.78), F1-score, and slightly improved alignment with the ground truth within the conscious category, as shown in the table captions, are constrained by potential errors that occur in the early stages of interaction detection. On BiLSTM, the model performed moderately (accuracy=0.81, F1=0.80), with the workflow helping to result in a small improvement in the FNs. The Transformer models reported a very small improvement (Accuracy=0.78, F1=0.76), which agrees with the findings that the workflow's effect on this model is small, because the sequence modeling of this particular architecture might not have harnessed the cues of microactions as effectively as the convolutional architecture.

To illustrate the prototype's behavior, two representative scenarios are considered: one in which the patient is ultimately classified as unconscious; and another in which the patient is classified as conscious, demonstrating step by step how the detected microactions are integrated into the final AVPU-based decision.

For each scenario, four components are presented: the applied stimulus, the state of the patient's eyes, the detected pain response, and the final classification of the patient as conscious or unconscious.



Figure 5.1. Demonstration of the prototype with a "Conscious" as final prediction.



Figure 5.2. Demonstration of the prototype with a "Unconscious" as final prediction.

The prototype testing demonstrated that automated AVPU-related microaction detection using computer vision is theoretically feasible and operationally promising - although also susceptible to several methodological and practical limitations. Unseen video analysis demonstrated that models could identify clinically relevant interactions with reasonable accuracy, whereas temporal pose modelling produced the most robust results, especially for touch-related events. These findings confirm that incorporating skeletal dynamics with recurrent sequence modelling is particularly well-suited for the detection of light motor responses that characterize the early stages of AVPU assessment.

The dominance of the Pose+LSTM model in the touch-only setting suggests that temporal continuity is more important than raw visual appearance when interpreting light, targeted stimuli. This is clinically intuitive: responses to verbal prompts or light touch are often minimal and thus demand models capable of interpreting short, low-amplitude motion patterns. The performance, which is comparably lower in the pain-induction scenario, underlines important challenges with the detection of stronger but also more variable actions. While accuracy remained in the mid-seventies across models, precision and recall were markedly lower. The models seemed to confuse pain-related movements with background motion or non-stimulus-related activity, which indicates that the dataset does not yet cover enough variability in how pain is induced or expressed. In clinical practice, reactions to painful stimuli can be very different between individuals depending on age, pain tolerance, neurological status, and specifics of the stimulation method. The homogeneous conditions of the simulated recordings while necessary for consistency, might have inadvertently limited the diversity of pain expressions available for training. Consequently, models tended to adopt cautious behavior, missing true pain events while generating false alarms. A marginal improvement observed for the Transformer-based model suggests an advantage of attention mechanisms in capturing subtle spatiotemporal cues; however, it was not sufficient to overcome the inherent limitations of the dataset.

The combined interaction case clearly demonstrates the potential of the system. As compared to the other scenarios, the TCN achieved the highest accuracy, and the most balanced precision, recall, and F1-scores. This ability to model local temporal dependencies proved particularly useful in sequences where both a touch and a painful stimulus were induced in close proximity to each other. In clinical terms, this mirrors the way the AVPU protocol unfolds through an escalating series of stimuli and/or combinations of stimuli. These results suggest that convolutional temporal models may be well-suited to handle the multi-interaction environments found in such clinical protocols. The results from the cascade architecture were also promising; architectures that follow the clinical examination process may lead to more accurate inference, as they would enforce the logic of examining the patient sequentially. However, even with the TCN and the cascade architecture, there were still many classification errors made in borderline cases that are inherently difficult for both human observers and automated systems to distinguish due to the ambiguity associated with very weak or ambiguous reactions.

Although preliminary, the overall results are encouraging and confirm the viability of the system in supporting triage processes. The prototype was reliable in determining whether the patient was clearly alert or clearly unresponsive. The result is consistent with existing evidence indicating that gross cognitive states are more easily identified by algorithms, since they represent pronounced behavioral differences such as spontaneous eye opening or a complete lack of response. In contrast, distinguishing between verbal and pain responsiveness was more challenging. These categories depend on subtle cues – minor facial movements, partial withdrawal from stimuli, momentary tightening of the eyes – which may not have a strong, consistent presence in the dataset. The challenges in differentiating these interim states are indicative of the need for richer annotation frameworks and multimodal input.

Furthermore, these findings should be interpreted considering limitations of both the dataset and the simulated environment. Although designed to approximate prehospital triage conditions, real emergency scenarios cannot be fully reproduced, including factors such as ambulance motion, variable lighting, occlusions, and unpredictable patient behaviour. The limited number of volunteers and short recording duration also restrict the diversity of motion patterns available for learning. In addition, the model may partially capture structured aspects of the examiner's protocol, since examiner actions can be correlated with AVPU outcomes, meaning that part of the predictive signal may come from procedural context rather than only patient responses; this highlights the importance of future work exploring the contribution of different input sources, such as patient-only cues, facial and eye features, and pose information,

as well as comparisons between rule-based aggregation and more flexible temporal models. Finally, while the rule-based AVPU aggregation is clinically informed, it introduces structural constraints that could be refined in future work using more adaptive probabilistic or end-to-end approaches. Despite these limitations, the results still provide a strong proof-of-concept for microaction-based video analysis in simulated emergency triage settings.

The prototype has proven a promising capability with respect to the detection of clinically significant stimuli, as well as patient responses, that the automated assistance with AVPU assessment is a technically drivable premise within a controlled setting. This potential has implications regarding the assistance that computer vision systems might provide emergency services personnel during triage, with respect to improved responsiveness assessment, as well as a reduced likelihood of missed AVPU assessments, although real-world conditions would need to mitigate the identified issues with the dataset used to support this premise. Accordingly, future iterations should incorporate a more diverse dataset, enhanced modeling of patient facial responses, and consideration of multiple stimulus modalities—such as auditory cues—which are now recognized components of assessment systems used in prehospital care.

Chapter 6. Conclusion

Following the DSRM, the study addressed a clinically relevant problem: the subjectivity and variability associated with manual consciousness assessment. The resulting prototype mirrors the logic of the AVPU workflow by first identifying examiner–patient interactions, such as touch and pain-inducing stimuli, then detecting patient responses, including eye state and facial expressions associated with pain, and finally aggregating these observations temporally to generate a binary consciousness classification. This decomposition provides a structured and interpretable representation of the assessment process and enables the contribution of each detected microaction to be examined.

Several contributions support these findings. First, a custom dataset of 132 simulated triage videos involving 52 healthy volunteers was created and annotated at frame level by domain experts for clinically relevant microactions. Second, a prototype system integrating multiple computer vision and deep learning components was developed and evaluated. Third, the experimental results demonstrated promising technical performance, with strong results for interaction recognition and eye-state classification, and a final TCN-based model achieving an accuracy of 0.92 for binary consciousness classification.

These results should be interpreted with appropriate caution. The study was conducted using a relatively small dataset composed of simulated scenarios involving healthy volunteers rather than real patients. In addition, some component tasks, particularly facial pain-expression recognition, achieved only moderate performance. Although the proposed approach demonstrated encouraging results, further validation using larger and more diverse clinical datasets, subject-independent evaluation protocols, and more robust statistical analyses is necessary before any clinical application can be considered.

Nevertheless, this Dissertation provides preliminary evidence that a microaction-driven approach based on computer vision can support automated consciousness assessment in simulated emergency triage contexts. By structuring the assessment process around clinically interpretable actions and responses, the proposed framework establishes a foundation for future research aimed at extending and validating this approach in real-world emergency and pre-hospital environments.

Bibliographic References and Sources

ACEM. (2023). *Guidelines on the Implementation of the Australasian Triage Scale in Emergency Departments*. https://acem.org.au/getmedia/51dc74f7-9ff0-42ce-872a-0437f3db640a/G24_04_Guidelines_on_Implementation_of_ATS_Jul-16.aspx

Alghamdi, T., & Alaghband, G. (2022). Facial Expressions Based Automatic Pain Assessment System. *Applied Sciences*, *12*(13), 6423. https://doi.org/10.3390/app12136423

Aung, M. S. H., Kaltwang, S., Romera-Paredes, B., Martinez, B., Singh, A., Cella, M., Valstar, M., Meng, H., Kemp, A., Shafizadeh, M., Elkins, A. C., Kanakam, N., De Rothschild, A., Tyler, N., Watson, P. J., Williams, A. C. D. C., Pantic, M., & Bianchi-Berthouze, N. (2016). The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset. *IEEE Transactions on Affective Computing*, *7*(4), 435–451. https://doi.org/10.1109/TAFFC.2015.2462830

Bai, S., Kolter, J. Z., & Koltun, V. (2018). *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1803.01271

Ben Aoun, N. (2024). A Review of Automatic Pain Assessment from Facial Information Using Machine Learning. *Technologies*, *12*(6), 92. https://doi.org/10.3390/technologies12060092

Bertasius, G., Wang, H., & Torresani, L. (2021). *Is Space-Time Attention All You Need for Video Understanding?* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2102.05095

Brouns, S. H. A., Mignot-Evers, L., Derkx, F., Lambooij, S. L., Dieleman, J. P., & Haak, H. R. (2019). Performance of the Manchester triage system in older emergency department patients: A retrospective cohort study. *BMC Emergency Medicine*, *19*(1), 3. https://doi.org/10.1186/s12873-018-0217-y

Bullard, M. J., Musgrave, E., Warren, D., Unger, B., Skeldon, T., Grierson, R., Van Der Linde, E., & Swain, J. (2017). Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) Guidelines 2016. *CJEM*, *19*(S2), S18–S27. https://doi.org/10.1017/cem.2017.365

Camarena, F., Chang, L., Gonzalez-Mendoza, M., & Cuevas-Ascencio, R. J. (2022). Action recognition by key trajectories. *Pattern Analysis and Applications*, *25*(2), 409–423. https://doi.org/10.1007/s10044-021-01054-z

Casti, P., Mencattini, A., Comes, M. C., Callari, G., Di Giuseppe, D., Natoli, S., Dauri, M., Daprati, E., & Martinelli, E. (2019). Calibration of Vision-Based Measurement of Pain Intensity With Multiple Expert Observers. *IEEE Transactions on Instrumentation and Measurement*, *68*(7), 2442–2450. https://doi.org/10.1109/TIM.2019.2909603

Choi, D. H., Hong, K. J., Do Shin, S., Kim, S., Chung, M., Kim, K. H., Song, K. J., Cho, M., Yoon, D., & Lee, J. (2023). Measurement of level of consciousness by AVPU scale assessment system based on automated video and speech recognition technology. *The American Journal of Emergency Medicine*, *74*, 112–118. https://doi.org/10.1016/j.ajem.2023.09.044

Cui, S., Huang, D., Ni, Y., & Feng, X. (2021). Multi-Scale Regional Attention Networks for Pain Estimation. *2021 13th International Conference on Bioinformatics and Biomedical Technology*, 1–8. https://doi.org/10.1145/3473258.3473259

Dragomir, M.-C., Florea, C., & Pupezescu, V. (2020). Automatic Subject Independent Pain Intensity Estimation using a Deep Learning Approach. *2020 International Conference on E-Health and Bioengineering (EHB)*, 1–4. https://doi.org/10.1109/EHB50910.2020.9280190

Egede, J. O., Song, S., Olugbade, T. A., Wang, C., Williams, A. C. D. C., Meng, H., Aung, M., Lane, N. D., Valstar, M., & Bianchi-Berthouze, N. (2020). EMOPAIN Challenge 2020: Multimodal Pain Evaluation from Facial and Bodily Expressions. *2020 15th IEEE*

International Conference on Automatic Face and Gesture Recognition (FG 2020), 849–856. https://doi.org/10.1109/FG47880.2020.00078

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). Convolutional Two-Stream Network Fusion for Video Action Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1933–1941. https://doi.org/10.1109/CVPR.2016.213

Fu, H., Gao, J., & Liu, H. (2023). Human pose estimation and action recognition for fitness movements. *Computers & Graphics*, *116*, 418–426. https://doi.org/10.1016/j.cag.2023.09.008

Gomutbutra, P., Kittisares, A., Sanguansri, A., Choosri, N., Sawaddiruk, P., Fakfum, P., Lerttrakarnnon, P., & Saralamba, S. (2022). Classification of elderly pain severity from automated video clip facial action unit analysis: A study from a Thai data repository. *Frontiers in Artificial Intelligence*, *5*, 942248. https://doi.org/10.3389/frai.2022.942248

Gong, P., & Luo, X. (2025). A Survey of Video Action Recognition Based on Deep Learning. *Knowledge-Based Systems*, *320*, 113594. https://doi.org/10.1016/j.knosys.2025.113594

Hachaj, T., Ogiela, M. R., & Koptyra, K. (2018). Human actions recognition from motion capture recordings using signal resampling and pattern recognition methods. *Annals of Operations Research*, *265*(2), 223–239. https://doi.org/10.1007/s10479-016-2308-z

Hara, K., Kataoka, H., & Satoh, Y. (2017). *Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1708.07632

Hinson, J. S., Martinez, D. A., Cabral, S., George, K., Whalen, M., Hansoti, B., & Levin, S. (2019). Triage Performance in Emergency Medicine: A Systematic Review. *Annals of Emergency Medicine*, *74*(1), 140–152. https://doi.org/10.1016/j.annemergmed.2018.09.022

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, D., Xia, Z., Li, L., Wang, K., & Feng, X. (2019). Pain-awareness multistream convolutional neural network for pain estimation. *Journal of Electronic Imaging*, *28*(04), 1. https://doi.org/10.1117/1.JEI.28.4.043008

Karamitsos, I., Seladji, Ii., & Modak, S. (2021). A Modified CNN Network for Automatic Pain Identification Using Facial Expressions. *Journal of Software Engineering and Applications*, *14*(08), 400–417. https://doi.org/10.4236/jsea.2021.148024

Kolic, I., Crane, S., McCartney, S., Perkins, Z., & Taylor, A. (2015). Factors affecting response to National Early Warning Score (NEWS). *Resuscitation*, *90*, 85–90. https://doi.org/10.1016/j.resuscitation.2015.02.009

Layher, G., Brosch, T., & Neumann, H. (2017). Real-Time Biologically Inspired Action Recognition from Key Poses Using a Neuromorphic Architecture. *Frontiers in Neurorobotics*, *11*. https://doi.org/10.3389/fnbot.2017.00013

Li, Q., Qiu, Z., Yao, T., Mei, T., Rui, Y., & Luo, J. (2016). Action Recognition by Learning Deep Multi-Granular Spatio-Temporal Video Representation. *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 159–166. https://doi.org/10.1145/2911996.2912001

Lim, J., Luo, C., Lee, S., Song, Y. E., & Jung, H. (2024). Action Recognition of Taekwondo Unit Actions Using Action Images Constructed with Time-Warped Motion Profiles. *Sensors*, *24*(8), 2595. https://doi.org/10.3390/s24082595

Lin, J., Gan, C., & Han, S. (2018). *TSM: Temporal Shift Module for Efficient Video Understanding* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1811.08383

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2021). *Video Swin Transformer* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2106.13230

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). *MediaPipe: A Framework for Building Perception Pipelines* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1906.08172

Markussen, D. L., Brevik, H. S., Bjørneklett, R. O., & Engan, M. (2023). Validation of a modified South African triage scale in a high-resource setting: A retrospective cohort study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *31*(1), 13. https://doi.org/10.1186/s13049-023-01076-y

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & for the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, *339*(jul21 1), b2535–b2535. https://doi.org/10.1136/bmj.b2535

Moll, H. A. (2010). Challenges in the validation of triage systems at emergency departments. *Journal of Clinical Epidemiology*, *63*(4), 384–388. https://doi.org/10.1016/j.jclinepi.2009.07.009

NHS England. (2025). *Busiest year on record for emergency services as winter flu pressure rages on*. https://www.england.nhs.uk/2025/01/busiest-year-on-record-for-emergency-services-as-winter-flu-pressure-rages-on/

Núñez, J. C., Cabido, R., Pantrigo, J. J., Montemayor, A. S., & Vélez, J. F. (2018). Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, *76*, 80–94. https://doi.org/10.1016/j.patcog.2017.10.033

Ohaiba, M. M., Anamazobi, E. G., Okobi, O. E., Aguda, K., & Chukwu, V. U. (2024). Trends and Patterns in Emergency Department Visits: A Comprehensive Analysis of Adult Data From the National Center for Health Statistics (NCHS) Database. *Cureus*. https://doi.org/10.7759/cureus.66059

Othman, E., Werner, P., Saxen, F., Al-Hamadi, A., Gruss, S., & Walter, S. (2023). Classification networks for continuous automatic pain intensity monitoring in video using facial expression on the X-ITE Pain Database. *Journal of Visual Communication and Image Representation*, *91*, 103743. https://doi.org/10.1016/j.jvcir.2022.103743

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Prkachin, K. M., & Solomon, P. E. (2008). The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, *139*(2), 267–274. https://doi.org/10.1016/j.pain.2008.04.010

Rajasegaran, J., Pavlakos, G., Kanazawa, A., Feichtenhofer, C., & Malik, J. (2023). *On the Benefits of 3D Pose and Tracking for Human Action Recognition* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2304.01199

Rathee, N., Pahal, S., & Sheoran, P. (2022). Pain detection from facial expressions using domain adaptation technique. *Pattern Analysis and Applications*, *25*(3), 567–574. https://doi.org/10.1007/s10044-021-01025-4

RCEM. (2024). *Almost 300 deaths a week in 2023 associated with long A&E waits despite UEC Recovery Plan*. https://rcem.ac.uk/press-release/almost-300-deaths-a-week-in-2023-associated-with-long-ae-waits-despite-uec-recovery-plan/

RCEM. (2025). *Each a dearly loved family member' – Excess deaths linked to long A&E waits increased to over 16,600 last year*. https://rcem.ac.uk/news/each-a-dearly-loved-family-member-excess-deaths-linked-to-long-ae-waits-increased-to-over-16600-last-year/

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Semwal, A., & Londhe, N. D. (2021a). Computer aided pain detection and intensity estimation using compact CNN based fusion network. *Applied Soft Computing*, *112*, 107780. https://doi.org/10.1016/j.asoc.2021.107780

Semwal, A., & Londhe, N. D. (2021b). ECCNet: An Ensemble of Compact Convolution Neural Network for Pain Severity Assessment from Face images. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 761–766. https://doi.org/10.1109/Confluence51648.2021.9377197

Tavakolian, M., Bordallo Lopez, M., & Liu, L. (2020). Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation. *Pattern Recognition Letters*, *140*, 26–33. https://doi.org/10.1016/j.patrec.2020.09.012

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2017). *A Closer Look at Spatiotemporal Convolutions for Action Recognition* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.1711.11248

Travers, D. A., Waller, A. E., Katznelson, J., & Agans, R. (2009). Reliability and Validity of the Emergency Severity Index for Pediatric Triage. *Academic Emergency Medicine*, *16*(9), 843–849. https://doi.org/10.1111/j.1553-2712.2009.00494.x

Wang, F., Xiang, X., Liu, C., Tran, T. D., Reiter, A., Hager, G. D., Quon, H., Cheng, J., & Yuille, A. L. (2017). Regularizing face verification nets for pain intensity regression. *2017 IEEE International Conference on Image Processing (ICIP)*, 1087–1091. https://doi.org/10.1109/ICIP.2017.8296449

Werner, P., Al-Hamadi, A., Limbrecht-Ecklundt, K., Walter, S., Gruss, S., & Traue, H. C. (2017). Automatic Pain Assessment with Facial Activity Descriptors. *IEEE Transactions on Affective Computing*, *8*(3), 286–299. https://doi.org/10.1109/TAFFC.2016.2537327

Werner, P., Lopez-Martinez, D., Walter, S., Al-Hamadi, A., Gruss, S., & Picard, R. W. (2022). Automatic Recognition Methods Supporting Pain Assessment: A Survey. *IEEE Transactions on Affective Computing*, *13*(1), 530–552. https://doi.org/10.1109/TAFFC.2019.2946774

Xin, X., Li, X., Yang, S., Lin, X., & Zheng, X. (2021). Pain expression assessment based on a locality and identity aware network. *IET Image Processing*, *15*(12), 2948–2958. https://doi.org/10.1049/ipr2.12282

Xin, X., Lin, X., Yang, S., & Zheng, X. (2020). Pain intensity estimation based on a spatial transformation and attention CNN. *PLOS ONE*, *15*(8), e0232412. https://doi.org/10.1371/journal.pone.0232412

Zachariasse, J. M., Seiger, N., Rood, P. P. M., Alves, C. F., Freitas, P., Smit, F. J., Roukema, G. R., & Moll, H. A. (2017). Validity of the Manchester Triage System in emergency care: A prospective observational study. *PLOS ONE*, *12*(2), e0170811. https://doi.org/10.1371/journal.pone.0170811

Zhang, B., Wang, L., Wang, Z., Qiao, Y., & Wang, H. (2016). *Real-time Action Recognition with Enhanced Motion Vector CNNs* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1604.07669

Zhou, J., Hong, X., Su, F., & Zhao, G. (2016). Recurrent Convolutional Neural Network Regression for Continuous Pain Intensity Estimation in Video. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1535–1543. https://doi.org/10.1109/CVPRW.2016.191

Annex A – Additional data on Touch vs. No Touch detection

Annex A.1 – Models parametrization

Table A.1. Models parametrization for Touch vs. No Touch detection.

Parameter	CNN+LSTM	Pose	Pose+LSTM	TSM
Input representation	RGB video frames	MediaPipe Pose landmarks (33 × 3 = 99 features)	MediaPipe Pose landmark sequences	RGB video clips
Input size	10 frames, 224 × 224 pixels	99 features per frame	10 frames × 99 features	10 frames, 224 × 224 pixels
Architecture	3 CNN blocks + 1 LSTM + Dense	Multilayer Perceptron	2-layer LSTM + Dense	Temporal Shift Module backbone + Dense
Number of layers	3 convolutional layers, 1 LSTM layer	2 dense layers	2 LSTM layers	CNN backbone with temporal shift operations
Filters / Units	32, 64, 128 filters; 128 LSTM units	128 and 64 dense units	128 LSTM units per layer	Backbone-dependent
Kernel size	3 × 3	Not applicable	Not applicable	3 × 3
Activation function	ReLU; Sigmoid output	ReLU; Sigmoid output	tanh (internal LSTM), Sigmoid output	ReLU; Sigmoid output
Dropout	0.30	0.30	0.30	0.30
Loss function	Binary Cross-Entropy	Binary Cross-Entropy	Binary Cross-Entropy	Binary Cross-Entropy
Optimizer	Adam	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001	0.001
Batch size	32	32	32	32
Maximum epochs	50	50	50	50
Early stopping	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs
Transfer learning	No	No	No	Yes (ImageNet pretrained weights)

The four models were trained under the same experimental conditions to ensure a fair comparison. All models used binary cross-entropy loss, the Adam optimizer, a learning rate of 0.001, a batch size of 32, and early stopping based on validation loss with a patience of five epochs. The Pose+LSTM architecture achieved the best performance, indicating that temporal modeling of skeletal landmarks was the most effective strategy for detecting touch interactions while remaining robust to variations in appearance, illumination, and background.

Annex A.2 – Confusion matrixes

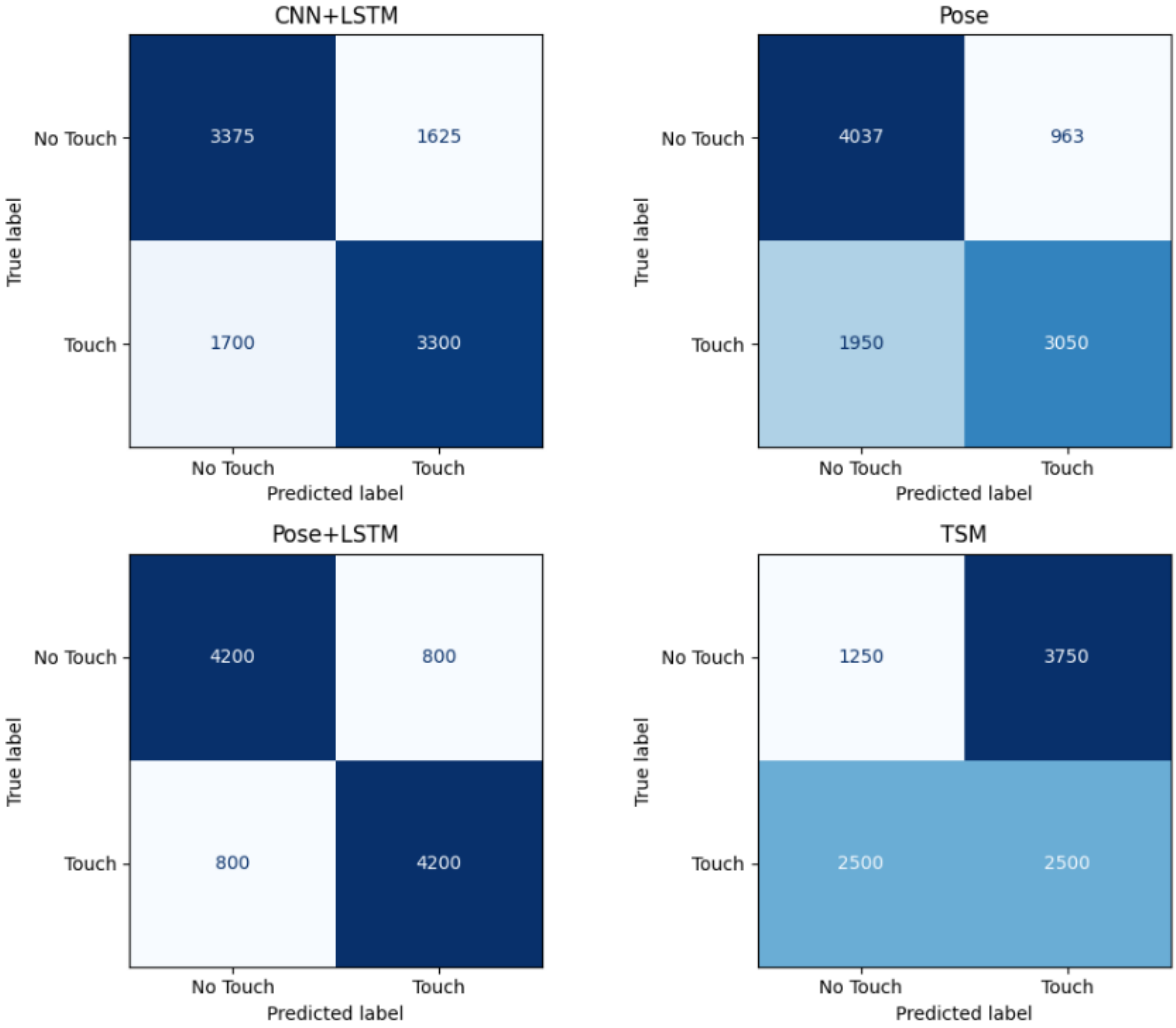


Figure A.1. Confusion Matrixes for Touch vs. No Touch Detection.

Annex B – Additional data for Pain Induction vs. No Pain Induction detection

Annex B.1 – Models parametrization

Table B.1. Models parametrization for Pain Induction vs. No Pain Induction detection.

Parameter	Pose+LSTM	Pose+Transformer	TCN
Input representation	MediaPipe Pose landmark sequences (33 × 3 = 99 features)	MediaPipe Pose landmark sequences (33 × 3 = 99 features)	MediaPipe Pose landmark sequences (33 × 3 = 99 features)
Input size	10 frames × 99 features	10 frames × 99 features	10 frames × 99 features
Architecture	2-layer LSTM + Dense	2-layer Transformer Encoder + Dense	Dilated 1D Convolutional Network + Dense
Number of layers	2 LSTM layers	2 encoder layers	3 temporal convolutional blocks
Units / Filters	128 LSTM units per layer	Embedding dimension 128, 4 attention heads	64, 128, and 128 filters
Kernel size	Not applicable	Not applicable	3
Dilation rates	Not applicable	Not applicable	1, 2, 4
Activation function	tanh (internal LSTM), Sigmoid output	ReLU in feed-forward blocks, Sigmoid output	ReLU; Sigmoid output
Dropout	0.30	0.30	0.30
Loss function	Binary Cross-Entropy	Binary Cross-Entropy	Binary Cross-Entropy
Optimizer	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001
Batch size	32	32	32
Maximum epochs	50	50	50
Early stopping	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs
Transfer learning	No	No	No

All three models were trained under identical experimental conditions to ensure a fair comparison. The input consisted of 10-frame sequences of MediaPipe Pose landmarks, and all models were optimized using binary cross-entropy loss and the Adam optimizer with a learning rate of 0.001. The results were broadly similar across architectures, with accuracy between 0.74 and 0.75 and F1-score of 0.43. The Pose+Transformer model achieved a marginally higher accuracy, suggesting that self-attention may provide a slight advantage in capturing localized motion patterns associated with pain-inducing gestures. However, the modest performance indicates that pain induction is a challenging task, likely due to the subtle and short-duration nature of the movement and the limited size and variability of the available dataset.

Annex B.2 – Confusion matrixes

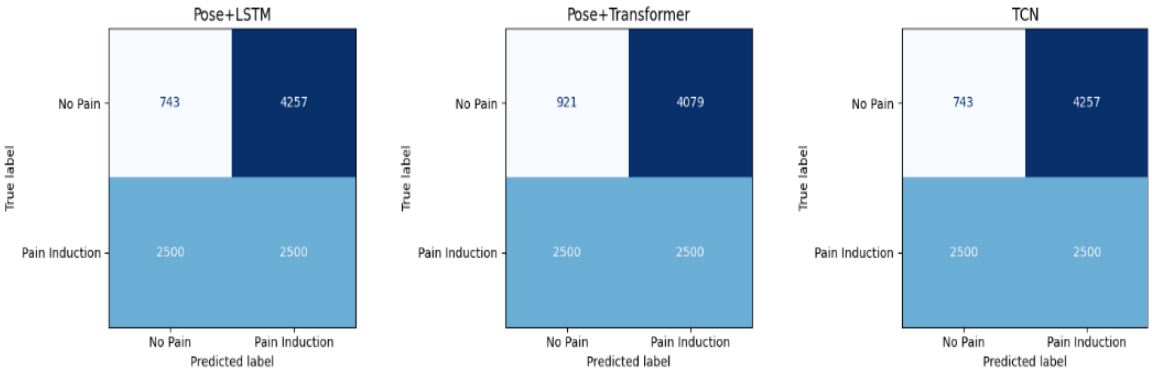


Figure B.1. Confusion Matrixes for Pain Induction vs. No Pain Induction detection.

Annex C – Additional data for combined Touch and Pain Induction recognition

Annex C.1 – Models parametrization

Table C.1. Models parametrization for combined Touch and Pain Induction recognition.

Parameter	TCN	BiLSTM	Transformer Encoder	Cascade LSTM (Touch → Pain)
Input representation	MediaPipe Pose landmark sequences (33 × 3 = 99 features)	MediaPipe Pose landmark sequences (33 × 3 = 99 features)	MediaPipe Pose landmark sequences (33 × 3 = 99 features)	MediaPipe Pose landmark sequences (33 × 3 = 99 features)
Input size	10 frames × 99 features	10 frames × 99 features	10 frames × 99 features	10 frames × 99 features
Output classes	3 (Touch, Pain Induction, No Interaction)	3	3	Two sequential binary classifiers
Architecture	Dilated 1D Convolutional Network + Dense	2-layer Bidirectional LSTM + Dense	2-layer Transformer Encoder + Dense	Two independent 2-layer LSTM classifiers
Number of layers	3 temporal convolutional blocks	2 BiLSTM layers	2 encoder layers	2 LSTM layers per stage
Units / Filters	64, 128, and 128 filters	128 units in each direction	Embedding dimension 128, 4 attention heads	128 LSTM units per layer
Kernel size	3	Not applicable	Not applicable	Not applicable
Dilation rates	1, 2, 4	Not applicable	Not applicable	Not applicable
Activation function	ReLU; Softmax output	tanh (internal LSTM); Softmax output	ReLU in feed-forward blocks; Softmax output	tanh (internal LSTM); Sigmoid output
Dropout	0.30	0.30	0.30	0.30
Loss function	Categorical Cross-Entropy	Categorical Cross-Entropy	Categorical Cross-Entropy	Binary Cross-Entropy (each stage)
Optimizer	Adam	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001	0.001
Batch size	32	32	32	32
Maximum epochs	50	50	50	50 per stage
Early stopping	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs
Transfer learning	No	No	No	No

All models were trained using the same subject-independent split and identical optimization settings to ensure comparability. The task consisted of jointly recognizing three classes: Touch, Pain Induction, and No Interaction. The TCN achieved the best overall performance, with 0.90 accuracy and 0.90 F1-score, substantially outperforming the BiLSTM, Transformer Encoder, and Cascade LSTM approaches. This result suggests that dilated temporal convolutions were particularly effective at capturing both the smoother motion patterns associated with touch and

the shorter, more abrupt gestures involved in pain induction. The strong performance of the TCN motivated its selection as the temporal aggregation model used in the final consciousness classification stage.

Annex C.2 – Confusion matrixes

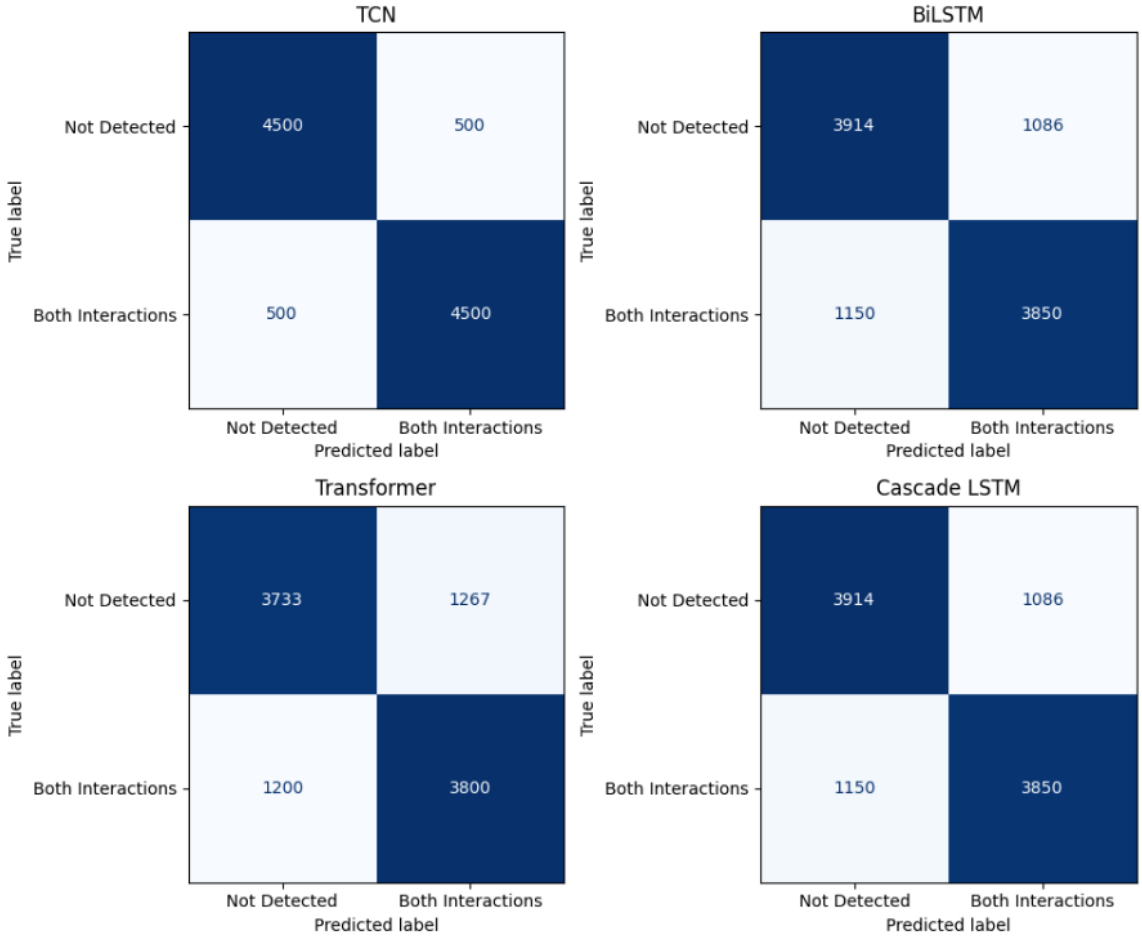


Figure C.1. Confusion Matrixes for combined Touch and Pain Induction recognition.

Annex D – Additional data for Pain Expression Detection

Annex D.1 – Models parametrization

Table D.1. Models parametrization for combined Pain Expression Detection.

Parameter	R(2+1)D	R3D	Swin3D	TimeSformer
Input representation	RGB facial video clips	RGB facial video clips	RGB facial video clips	RGB facial video clips
Input size	16 frames, 112 × 112 pixels	16 frames, 112 × 112 pixels	16 frames, 112 × 112 pixels	16 frames, 224 × 224 pixels
Architecture	Factorized 3D CNN	3D Residual CNN	Hierarchical Vision Transformer	Space-Time Transformer
Number of layers	18 residual layers	18 residual layers	Tiny configuration (4 stages)	8 transformer blocks
Units / Filters	Standard R(2+1)D-18 configuration	Standard R3D-18 configuration	Embedding dimension 96	Embedding dimension 768, 8 attention heads
Kernel size	3 × 3 × 3 (factorized spatial + temporal)	3 × 3 × 3	Window-based self-attention	Patch size 16 × 16
Dilation rates	Not applicable	Not applicable	Not applicable	Not applicable
Activation function	ReLU; Softmax output	ReLU; Softmax output	GELU; Softmax output	GELU; Softmax output
Dropout	0.30	0.30	0.30	0.30
Loss function	Categorical Cross-Entropy	Categorical Cross-Entropy	Categorical Cross-Entropy	Categorical Cross-Entropy
Optimizer	Adam	Adam	Adam	Adam
Learning rate	0.0001	0.0001	0.0001	0.0001
Batch size	8	8	8	4
Maximum epochs	30	30	30	30
Early stopping	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs
Transfer learning	Yes (Kinetics-400 pretrained weights)	Yes (Kinetics-400 pretrained weights)	Yes (Kinetics-400 pretrained weights)	Yes (Kinetics-400 pretrained weights)

Parameter	OpenFace + Logistic Regression	OpenFace + Random Forest	MediaPipe (EAR-based Geometry)
Input representation	OpenFace AUs	OpenFace AUs	EAR and geometric facial features
Input size	AU feature vector per clip	AU feature vector per clip	Low-dimensional geometric feature vector
Architecture	Linear classifier	Ensemble of decision trees	Threshold-based geometric classifier
Number of layers	1 linear model	100 decision trees	Not applicable
Units / Filters	Number of features = extracted AUs	100 trees, max depth unrestricted	Not applicable
Kernel size	Not applicable	Not applicable	Not applicable
Dilation rates	Not applicable	Not applicable	Not applicable
Activation function	Sigmoid	Gini impurity splits	Threshold-based decision
Dropout	None	None	None

Loss function	Binary Cross-Entropy (log loss)	Gini criterion	Not applicable
Optimizer			
Learning rate	Not applicable	Not applicable	Not applicable
Batch size	Not applicable	Not applicable	Not applicable
Maximum epochs	Until convergence	Until convergence	Not applicable
Early stopping	Not applicable	Not applicable	Not applicable
Transfer learning	No	No	No

All methods were evaluated under the same subject-independent train-test split to compare different strategies for recognizing pain-related facial expressions. The deep spatiotemporal models were fine-tuned from Kinetics-400 pretrained weights, while the OpenFace and MediaPipe approaches relied on engineered features derived from facial AUs or geometric measurements. Performance was modest across all methods, with accuracy and F1-scores ranging from 0.54 to 0.60. Swin3D achieved the highest overall accuracy (0.60), followed closely by R(2+1)D and TimeSformer. These results should be interpreted as preliminary, as the dataset consisted of simulated pain expressions from healthy volunteers rather than genuine nociceptive responses. The limited performance likely reflects the subtle nature of the task, the small sample size, variability in simulated expressions, and potential annotation uncertainty. Consequently, pain expression detection was used as a complementary input to the final classification pipeline rather than as a standalone indicator of consciousness.

Annex D.2 – Confusion matrixes

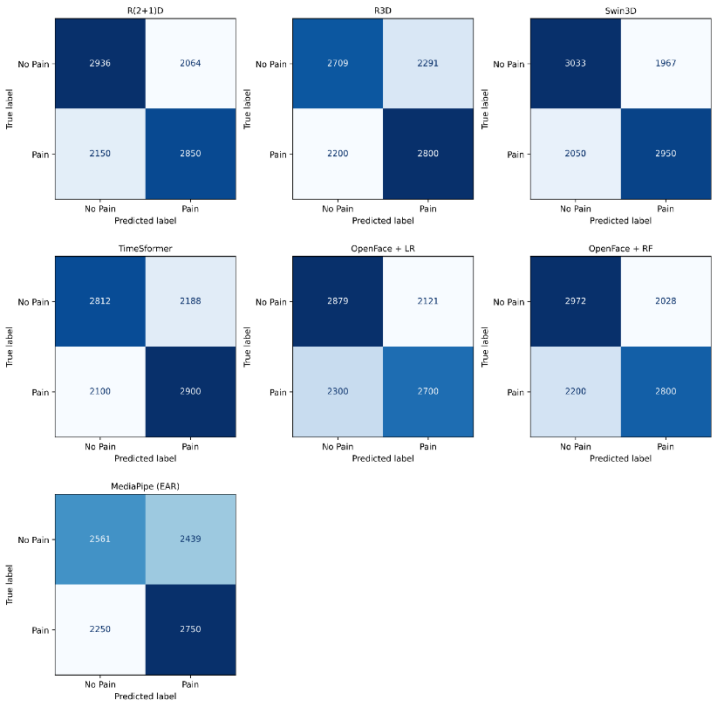


Figure D.1. Confusion Matrixes for Pain Expression detection.

Annex E – Additional data for Eyes Open/Closed Detection

Annex E.1 – Models parametrization

Table E.1. Models parametrization for combined Eyes Open/Closed Detection.

Parameter	CNN	EAR
Input representation	Cropped eye region images	MediaPipe facial landmarks
Input size	64 × 64 grayscale eye images	Six landmarks per eye
Architecture	Convolutional Neural Network	Geometric rule-based method
Number of layers	3 convolutional layers + 2 dense layers	Not applicable
Filters / Units	32, 64, and 128 filters; 128 dense units	Not applicable
Kernel size	3 × 3	Not applicable
Activation function	ReLU; Sigmoid output	Threshold-based decision
Dropout	0.30	None
Loss function	Binary Cross-Entropy	Not applicable
Optimizer	Adam	Not applicable
Learning rate	0.001	Not applicable
Batch size	32	Not applicable
Maximum epochs	30	Not applicable
Early stopping	Patience = 5 epochs	Not applicable
Transfer learning	No	No

Both approaches were evaluated for binary classification of the eye state as open or closed. The CNN model learned discriminative features directly from cropped eye images, while the EAR method estimated eye openness from geometric relationships between facial landmarks. The CNN substantially outperformed the EAR baseline, achieving an accuracy of 0.92 and an F1-score of 0.92, compared with 0.81 for the EAR method. These results indicate that image-based learning captures subtle eyelid configurations and partial closures more effectively than a fixed geometric threshold, particularly under variations in lighting, head pose, and facial expression.

Annex E.2 – Confusion matrixes

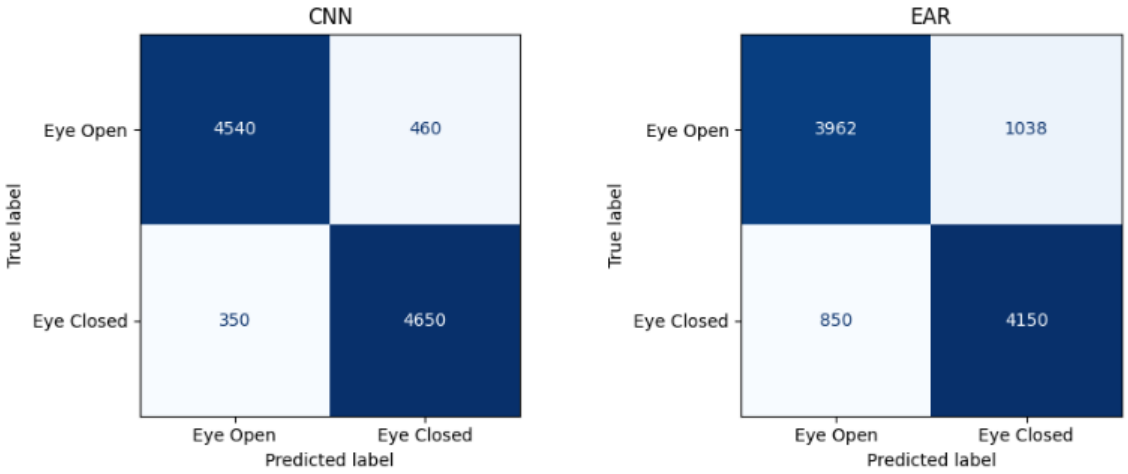


Figure E.1. Confusion Matrixes for combined Eyes Open/Closed Detection.

Annex F – Additional data for the final prediction of conscious or unconscious

Annex F.1 – Models parametrization

Table F.1. Models parametrization for the final prediction of conscious or unconscious.

Parameter	TCN	BiLSTM	Transformer Encoder	Cascade LSTM (Touch → Pain)
Input representation	Temporal sequences of detected microactions (touch, pain induction, eye state, pain expression)	Same as TCN	Same as TCN	Same as TCN
Input size	60 time steps × 4 features	60 time steps × 4 features	60 time steps × 4 features	60 time steps × 4 features
Output classes	2 (Conscious, Unconscious)	2	2	Two sequential binary classifiers
Architecture	Dilated 1D Convolutional Network + Dense	2-layer Bidirectional LSTM + Dense	2-layer Transformer Encoder + Dense	Two LSTM classifiers arranged sequentially
Number of layers	3 temporal convolutional blocks	2 BiLSTM layers	2 encoder layers	2 LSTM layers per stage
Units / Filters	64, 128, and 128 filters	128 units in each direction	Embedding dimension 128, 4 attention heads	128 LSTM units per layer
Kernel size	3	Not applicable	Not applicable	Not applicable
Dilation rates	1, 2, 4	Not applicable	Not applicable	Not applicable
Activation function	ReLU; Softmax output	tanh (internal LSTM); Softmax output	ReLU in feed-forward blocks; Softmax output	tanh (internal LSTM); Sigmoid output
Dropout	0.30	0.30	0.30	0.30
Loss function	Categorical Cross-Entropy	Categorical Cross-Entropy	Categorical Cross-Entropy	Binary Cross-Entropy (per stage)
Optimizer	Adam	Adam	Adam	Adam
Learning rate	0.001	0.001	0.001	0.001
Batch size	32	32	32	32
Maximum epochs	50	50	50	50 per stage
Early stopping	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs	Patience = 5 epochs
Transfer learning	No	No	No	No

The final classification stage received as input the time series of microactions detected by the preceding modules, including touch events, pain-inducing stimuli, eye state, and pain-related facial expressions. These features were aggregated over 60 time steps and used to predict whether the patient should be classified as conscious or unconscious. Among the evaluated approaches, the TCN achieved the highest performance, with an accuracy of 0.92 and an F1-score of 0.92, outperforming the BiLSTM, Transformer Encoder, and Cascade LSTM models.

This result indicates that dilated temporal convolutions were particularly effective at capturing the sequential stimulus-response patterns underlying consciousness assessment. Although these results are promising, they should be interpreted as preliminary because they were obtained using a relatively small dataset of simulated triage scenarios involving healthy volunteers rather than real patients.

Annex F.2 – Confusion matrixes

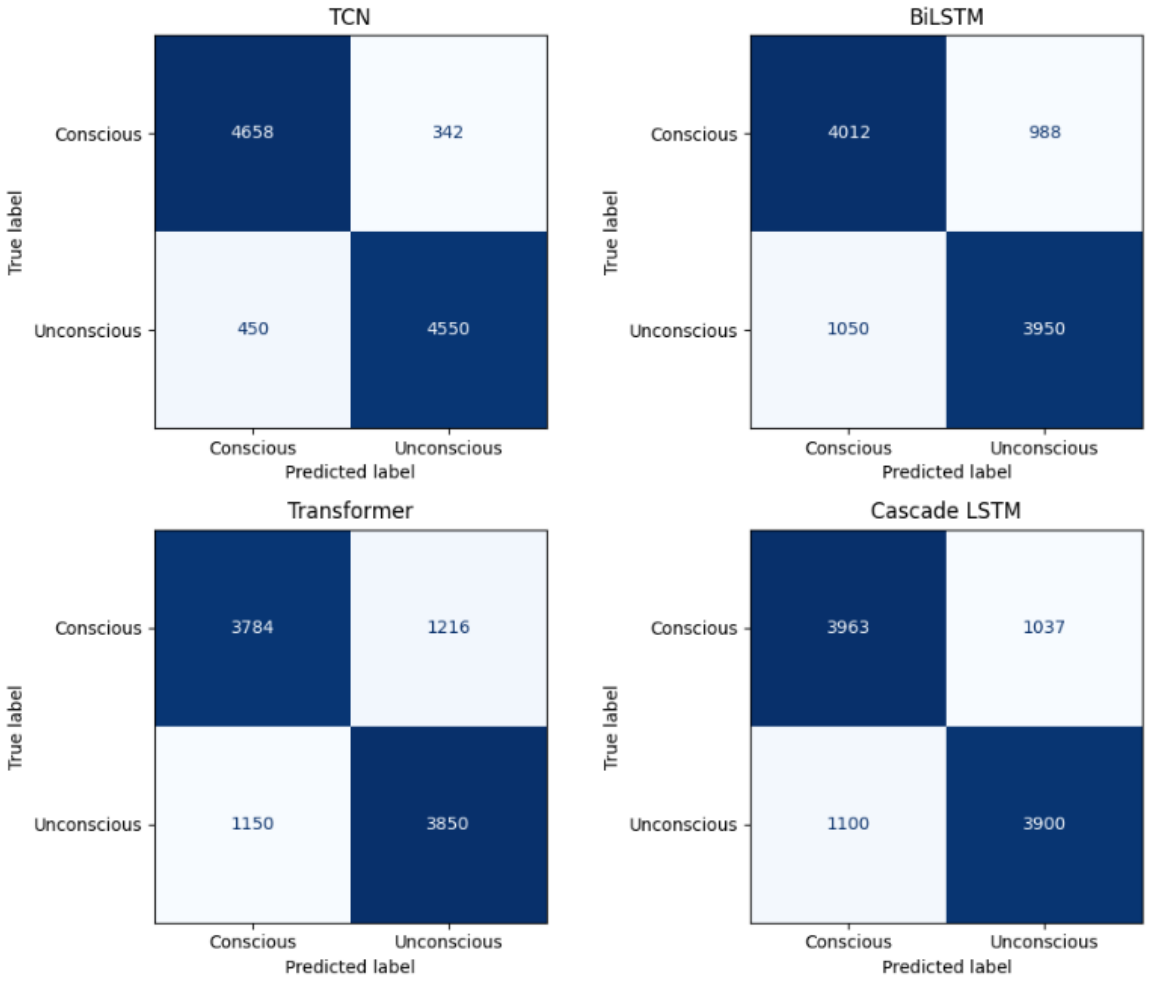


Figure F.1. Confusion Matrixes for the final prediction of conscious or unconscious.