
Intelligent Sight and Sound: A Chronic Cancer Pain Dataset

Catherine Ordun^{1,2}, Alexandra N. Cha¹, Edward Raff^{1,2},
Byron Gaskin¹, Alex Hanson¹, Mason Rule³, Sanjay Purushotham², James L. Gulley³

¹ Booz Allen Hamilton

²University of Maryland, Baltimore County

³Center for Cancer Research, National Cancer Institute, National Institutes of Health

Abstract

1 Cancer patients experience high rates of chronic pain throughout the treatment
2 process. Assessing pain for this patient population is a vital component of psycho-
3 logical and functional well-being, as it can cause a rapid deterioration of quality
4 of life. Existing work in facial pain detection often have deficiencies in labeling
5 or methodology that prevent them from being clinically relevant. This paper in-
6 troduces the first chronic cancer pain dataset, collected as part of the Intelligent
7 Sight and Sound (ISS) clinical trial, guided by clinicians to help ensure that model
8 findings yield clinically relevant results. The data collected to date consists of 29
9 patients, 509 smartphone videos, 189,999 frames, and self-reported affective and
10 activity pain scores adopted from the Brief Pain Inventory (BPI). Using static im-
11 ages and multi-modal data to predict self-reported pain levels, early models show
12 significant gaps between current methods available to predict pain today, with room
13 for improvement. Due to the especially sensitive nature of the inherent Personally
14 Identifiable Information (PII) of facial images, the dataset will be released under
15 the guidance and control of the National Institutes of Health (NIH).

16 1 Introduction

17 The prevalence of chronic pain in cancer patients is high, with an estimated prevalence of 59% in
18 those undergoing anticancer treatment, 64% of whom have advanced stage disease and 33% who
19 continue to experience pain following completion of curative treatment [1]. Despite advances in pain
20 management, prompt assessment and management of cancer pain remains a challenge and a large
21 proportion of patients continue to experience moderate to severe pain.

22 Sub-optimal pain management can block patient recovery and improvement, making the already
23 difficult cancer experience, worse, for both patient and family [2, 3]. Manual clinical assessment
24 requires accounting for a landscape of complex emotions and beliefs that clinicians must regularly
25 take into account when assessing cancer patient pain - physical, psychological, social, and spiritual
26 elements combined with severe distress for future outlook [2, 4]. For example, patients undergoing
27 chemotherapy are more likely to believe that “good patients” do not complain about pain which they
28 believe can be distracting to clinicians and become non-communicative [2]. Further, few patients are
29 actually screened for pain at each clinical visit [3], and pain is under-reported in patient populations
30 such as nursing home patients [3]. Due to the variety of complex conditions affecting cancer pain,
31 experts recommend repeated, regular pain assessment, which can be difficult and impractical for
32 manual assessment by clinicians.

33 Currently, no facial pain datasets exists for chronic cancer pain and little research overall has been
34 conducted into machine learning for the identification and evaluation of chronic pain. For example, in

35 a review by [4], only seven out of 52 machine learning papers evaluated pain in a non-acute context
 36 such as chronic fatigue, fibromyalgia, chronic pancreatitis, migraines, and genetic pain. Existing
 37 facial pain research focuses on acute, musculoskeletal pain such as chronic lower back pain [5]
 38 and shoulder pain [6, 7] or simulated pain induced by heat or electrical stimuli [8, 9] where painful
 39 expressions are obvious through grimaces and eye raises. Such datasets are manually labeled by
 40 trained observers with Facial Action Coding Units [10], making the labeling procedure prohibitively
 41 expensive and impractical for clinical use. Further, external pain labels may be biased towards an
 42 outside observer’s impression of a patient’s pain, not the patient themselves. Research also shows that
 43 typical pain facial expressions that correlate with physical pain are less frequently observed among
 44 chronic pain cancer patients who exhibit subdued and placid expressions [11].

45 Given the limitations of existing facial expression pain data, the U.S. National Institutes of Health
 46 (NIH) National Cancer Institute (NCI) initiated “A Feasibility Study Investigating the Use of Machine
 47 Learning to Analyze Facial Imaging, Voice and Spoken Language for the Capture and Classification
 48 of Cancer Pain” [12], or “Intelligent Sight and Sound” (ISS). Details of the protocol are available
 49 publicly at <https://clinicaltrials.gov/ct2/show/NCT04442425>. This is an observational,
 50 non-interventional clinical study that aims to address the following problem statement [12, 13]: “To
 51 determine if a new observational based pain prediction algorithm can be produced that is accurate to
 52 standard, patient-reported pain measures and is generalizable for a diverse set of individuals, across
 53 sexes and skin types.” The study has two objectives: 1) investigate facial image data, and 2) analyze
 54 text and audio, as modalities for predicting self-reported chronic cancer pain.

55 The study is ongoing and aims to recruit 112 patients. We report the initial dataset, which is less
 56 than a quarter of the final data consisting of 29 patients. Data include multimodal extracts from
 57 video submitted in a spontaneous, home setting, and in a few cases of in-clinic capture at the NIH. It
 58 includes visual spectrum (RGB) video frames, facial images resulting from face detection models,
 59 facial landmarks from Active Appearance Models (AAMs) [14, 15], audio files, Mel spectrograms,
 60 audio features, and self-reported pain scores adopted from the Brief Pain Inventory (BPI) [16–18]. We
 61 will present details of the study design, data distribution, and storage procedures to ensure patient
 62 privacy. We also provide initial baseline results for pain classification using simple, traditional,
 63 machine learning models and neural networks.

64 2 Related Works

65 Automatically detecting pain from facial expressions has been extensively published following
 66 methods of facial emotion recognition (FER). The majority of these works have focused on acute or
 67 musculoskeletal physical pain [4, 19–27]. Primary pain datasets based upon facial imaging include
 68 UNBC-McMaster Shoulder Pain Expression Archive [6, 7], the Biopotential and Video Heat Pain
 69 (BioVid) Database using controlled, simulated heat to induce pain [8], Multimodal Intensity Pain
 70 (MIntPAIN) database using pain resulting from electrical stimulation [9], the Experimentally Induced
 71 Thermal and Electrical (X-ITE) Pain Database [28, 29], and the EmoPain for chronic, musculoskeletal
 72 pain [30]. These datasets traditionally contain video sequences since video enables continuous clinical
 73 monitoring of pain response [18]. These datasets also contain extensive offline annotations of pain
 74 ratings by external observers, and sometimes include additional modalities such as thermal and depth
 75 data. Additional video facial expression pain datasets exist that focus on different patient populations,
 76 but primarily focus on physical pain. These include multimodal behavioral and physiological data for
 77 neonatal pain [31, 32] and the University of Regina (UofR) Pain in Severe Dementia dataset [33, 34].
 78 A summary of the pain datasets is provided in Table 1.

Table 1: **Related Pain Datasets.**

Dataset	Stimulus	Subjects	Frames	Sequences	Seq. Duration	Modality
UNBC-McMaster [6, 7]	Shoulder pain	25	48,398	200	10 - 30 sec., per	Unimodal
BioVid [8]	Heat stimulus	90	8700	87	5.5 sec.	Multimoda
MIntPAIN [9]	Electronic stimulus	20	187,939	9366	1 - 10 sec.	Multimodal
EmoPain [30]	Chronic lower back pain	22	44,820	35	3 sec.	Multimodal
Neonatal Pain, USF [32, 35, 36]	Heel lancing	31	3026	200	9 sec.	Multimodal
UofR [33]	Physical, painful movements	102	162,629	95	Unknown	Multimodal
X-ITE [28, 29, 37]	Heat and electronic stimuli	134	26,454	N/A	7 sec.	Multimodal
ISS (Dec. 2020 - Jul. 2021)	Chronic cancer pain	29	189,999	509	3.52 - 135.59	Multimodal

79 **3 ISS Dataset**

80 The ISS protocol is a single site study with a goal of enrolling a total of 112 patients (90 adult and
 81 22 pediatric) who are actively receiving treatment for advanced malignancies and/or tumors at the
 82 NIH Clinical Center or treated with standard of care in the community. The study is overseen by
 83 the NIH Institutional Review Board (IRB), and the protocol was also reviewed by NCI’s Center for
 84 Cancer Research (CCR) Scientific Review Committee. New patient enrollment was paused during
 85 the Covid-19 pandemic due to initially unknown risks, but has resumed with vaccine availability and
 86 clinician guidance.

87 **3.1 Sample and Study Design**

88 To obtain as representative a sample as possible within the constraints of a feasibility study with an
 89 overall small sample size, the sample consists of twelve cohort groups of seven patients each. Patients
 90 represent a breadth of age, sex, skin tone (as a proxy for ethnicity), and pain experience. The current
 91 ISS dataset consists of 29 adult patients ages 18 years and over who have consented to participate in
 92 the study; no pediatric patients (ages 12-17) have been enrolled yet.

93 The goal is to evenly
 94 split the sample by
 95 i) sex (Male or Fe-
 96 male), ii) Fitzpatrick
 97 Skin Type [38], a
 98 self-reported, visual
 99 method of skin tone
 100 classification, where
 101 patients are asked to
 102 type themselves into
 103 one of two groups:
 104 "light" skin tones in
 105 types I-III or "dark"
 106 skin tones in types
 107 IV-VI, and iii) a self-
 108 reported "worst" pain
 109 score reported on a 0
 110 – 10 Numerical Rating
 111 Scale (NRS) [39]. The

Table 2: ISS: Twelve Patient Cohorts.

Number	Pain Target	Skin Types	Sex	Pain Class	Goal	Current
1A	0	I - III	Male	None	7	7
1B	0	I - III	Female	None	7	2
1C	0	IV - VI	Male	None	7	4
1D	0	IV - VI	Female	None	7	0
2A	1-3	I - III	Male	Low	7	1
2B	1-3	I - III	Female	Low	7	1
2C	1-3	IV - VI	Male	Low	7	3
2D	1-3	IV - VI	Female	Low	7	0
3A	4-6	I - III	Male	Moderate	7	2
3B	4-6	I - III	Female	Moderate	7	2
3C	4-6	IV - VI	Male	Moderate	7	0
3D	4-6	IV - VI	Female	Moderate	7	0
4A	7-10	I - III	Male	Severe	7	1
4B	7-10	I - III	Female	Severe	7	2
4C	7-10	IV - VI	Male	Severe	7	2
4D	7-10	IV - VI	Female	Severe	7	3
					112	30

112 self-reported pain score is referred to as the "Pain Target" and are grouped into levels 0, 1-3, 4-6, and
 113 7-10. It represents the worst pain the patient has experienced in the past thirty days prior to the start
 114 of the study. It is fixed throughout the patient’s enrollment and does not change. As a result, there is
 115 no variance for the "Pain Target" score. The "Pain Target" is the classification target which is later
 116 used for our baseline tasks.

117 The twelve different cohorts are
 118 shown in Table 2, along with the goal
 119 of seven patients to be enrolled per
 120 cohort, and the current distribution
 121 of patients enrolled. Note that data
 122 from one patient (0009) in Cohort 2B
 123 was unusable. As a result our analy-
 124 sis reports across 29 patients. Clin-
 125 ical inclusion criteria include indi-
 126 viduals with a diagnosis of a cancer
 127 or tumor who are under active treat-
 128 ment for this condition at NIH/NCI.
 129 Patients must also have access to a
 130 smart phone or computer with cam-
 131 era, microphone, and internet access.
 132 Several clinical exclusion criteria ap-
 133 ply. Excluded are patients with active

Table 3: ISS Study Design Overview. * Note that due to the global Covid-19 pandemic, the majority of patient videos were submitted in the remote setting.

Study Duration	3 months
# Remote Submission	3 / week; Max. 1 /day
# In-Clinic Submissions	1 - 4 /week*
Survey Tool	Smartphone (iPhone or Android) or computer (camera/mic.)
Time / Submission	Average total time: 3 min
Time / Question	Q 1-9: 1 min, Q10: 15 sec, Q11: 15 sec. - 3 min.
Self-Reported Pain	Q 1-9: Questions with Likert-scale responses
Voice/Video: Prompt	Q10: Read and record one of 3 randomized nursery rhymes.
Voice/Video: Narrative	Q11: Record respond to "Describe how you feel right now."
Compensation:	Min. 3 / week, they earn \$15.

central nervous system (CNS) metastases, with the exception

134 of those who have completed curative intent radiotherapy or surgery and have been asymptomatic for
 135 three months prior to consent, patients with Parkinson’s disease, and any psychiatric condition that
 136 would prohibit the understanding or rendering of informed consent. Additional exclusion criteria
 137 those who are non-English speaking or have known current alcohol or drug abuse. Each patient is
 138 enrolled for a three-month period and are financially incentivized to complete three check-ins per
 139 week remotely and up to four in-clinic check-ins. The study design is summarized in Table 3. Patients
 140 engage using an electronic questionnaire and through video recording using a custom developed
 141 mobile or web application, using an Android, iPhone, or computer with camera and microphone.

142 **3.2 Patient Protocol**

143 Figure 1 provides a series of screenshots showing the patient at-home or in-clinic check-in using
 144 the ISS application. For each approximately 3-minute check-in, patients respond to a nine element
 145 questionnaire based on the Brief Pain Inventory (BPI, licensed from MD Anderson) [16–18] and two
 146 prompts to record videos of themselves.

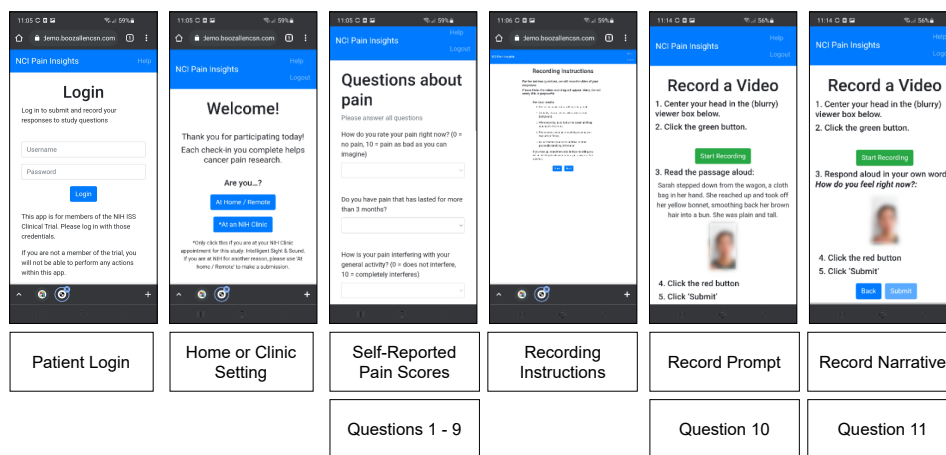


Figure 1: Submitting a Video through the ISS Mobile Application.

147 **3.2.1 Questions 1 - 9: Self-Reported Pain Scores**

148 In addition to the self-reported “Pain Target” which was assigned to each patient upon enrollment
 149 shown in Table 2, there are nine additional self-reported pain scores. We capture these self-reported
 150 pain scores based on the cancer pain literature which indicates that cancer patients experience complex
 151 emotions and beliefs that can influence their perception of pain and as a result, clinical treatment
 152 [2, 3]. These nine pain scores are submitted at the time of video submission and change at each
 153 submission. They are distinct and unrelated to the “Pain Target” which is used for the baseline
 154 classification tasks. There is no formula that relates the nine self-reported scores among themselves
 155 or to the “Pain Target”. In the below, Question 1 captures current pain intensity scored on an 11-point
 156 Likert Scale (0 No Pain - 10 Worst Possible Pain) followed by Question 2 which, when answered
 157 affirmatively, indicates the presence of chronic pain. Questions 3-9 utilize an 11-point Likert Scale (0
 158 Does not Interfere - 10 Completely Interferes) to measure the interference of pain in an individual’s
 159 activity (3, 5, 6) and an individual’s affect or mood (4, 7, 8, 9).

- 160 1. How do you rate your pain right now? (0 No Pain – 10 Worst Possible Pain on Likert-scale).
- 161 2. Do you have pain, related to your cancer, that has lasted for more than 3 months? (Yes/No)
- 162 3. How is your pain interfering with your General Activity?
- 163 4. How is your pain interfering with your Mood?
- 164 5. How is your pain interfering with your Walking Ability?
- 165 6. How is your pain interfering with your Normal Work (both work outside the home and housework)?
- 166 7. How is your pain interfering with your Relationships with other people?
- 167 8. How is your pain interfering with your Sleep?
- 168 9. How is your pain interfering with your Enjoyment of life?

169 **3.2.2 Questions 10 and 11: Prompt and Narrative**

170 Following the questionnaire, Question 10 is a prompt to record a video where the patient reads a 10-
 171 15 second passage of text at a grade 3 reading level selected at random from three different passages.
 172 The use of this sort of prompt is common practice in mood induction or conditioning trials where a
 173 neutral, non-emotion inducing prompt is used as a control versus a potentially, emotionally charged
 174 response related to the experimental condition [40–42]. The neutral passage options are:

- 175 • “Sarah stepped down from the wagon, a cloth bag in her hand. She reached up and took off her
 176 yellow bonnet, smoothing back her brown hair into a bun. She was plain and tall.” From Sarah Plain
 177 and Tall by Patricia MacLachlan [43]
- 178 • “And then the dog came running around the corner. He was a big dog. And ugly. And he looked
 179 like he was having a real good time. His tongue was hanging out and he was wagging his tail.” From
 180 Because of Winn Dixie by Kate DiCamillo [44]
- 181 • “You have brains in your head. You have feet in your shoes. You can steer yourself any direction
 182 you choose. You’re on your own. And you know what you know. And YOU are the one who’ll
 183 decide where to go.” From Oh the Places You Will Go by Dr. Seuss [45]

184 Finally, in Question 11, the patient records a video responding to the prompt *“Please describe how*
 185 *you feel right now.”* Narratives include discussion of medical conditions, mood, daily activities,
 186 current beliefs and attitudes about their pain. The allowable video length can range from 15 seconds
 187 to 3 minutes, with recording instructions shown prior to each video prompt. For “at-home” check-ins,
 188 patients are instructed to complete the submission alone, in a quiet and brightly lit room, preferably
 189 with a white wall or background. In addition, patients are asked not to reveal personal information
 190 such as their name or address. In Figure 1, the application screens for Questions 10 and 11 include a
 191 live video image to help the patient keep their face centered in the frame, but the application blurs the
 192 video. The blur effect is to prevent the patient from manipulating their facial expression and minimize
 193 self-conscious alteration of their appearance, allowing them to focus on their responses.

194 **3.3 Data Description**

195 A high level summary of the ISS dataset is provided in Table 4. The ISS dataset is comprised of
 196 29 patients submitting videos in a spontaneous, non-posed, home setting through a smartphone or
 197 computer. Patients are adults over the age of 18 y.o. and consist of the following demographics:
 198 20 Male, 9 Female, 17 Skin Type I-III, 12 Skin Type IV - VI. All patients were enrolled between
 199 December 2020 and July 2021. There are 189,999 total video frames. After facial detection, we
 200 extracted 173,011 facial images. After landmark detection on the facial images, the dataset was
 201 reduced by 2.86% to 168,063 facial images with landmarks, since landmarks could not be detected for
 202 some faces. We show the ratio of data imbalance across four pain levels using the total frames in Table
 203 4 where the "None" label is the majority class. The dataset also contains self-reported pain scores from
 204 Questions 1 - 9, described in detail in the Study Design section, along with sex and skin type labels
 205 assigned upon enrollment. Additional descriptive analysis is provided in Supplementary Materials.

Table 4: ISS Data Summary.

ISS Data Summary				Ratios of Total Frames by Pain Levels							
Total Patients	29	20 M, 9 F, 17 Skin Type I-III, 12 Skin Type IV-VI		4 Pain Levels	Frames	Ratio	No. Patients	2 Pain Levels	Frames	Ratio	No. Patients
Total Videos	509	Avg. Videos per Patient	17.55	None	100984	1.00	13	No Pain	100984	1.00	13
Total Frames	189,999	Avg. Frames per Patient	6551	Low	11784	8.57	4	Pain	89015	1.13	16
Total Duration	316 min.	Avg. Duration per Patient	655 sec.	Mod.	25999	3.88	4				
Avg. Duration per Video	37.32 sec.	Range of Duration per Video	3.52 - 135.79 sec.	Severe	51232	1.97	8				

206 A notional depiction of ISS data types is shown in Figure 3 to provide context for the data types.
 207 Due to the sensitivity of Personal Identifiable Information (PII) in the clinical study protocol, we are
 208 unable to display actual facial images from the dataset at this time.

209 **3.3.1 Data Extraction**

210 We use the patient narrative (Question 11) video files (.mp4) and extract frames at 10 frames-per-
 211 second. We decide to use the narrative versus the prompt since it may contain greater signals of pain
 212 and emotion, compared to the neutral baseline recording. An audio .wav file of the patient narrative
 213 is simultaneously extracted using the ffmpeg library. We use the PyTorch FaceNet library that

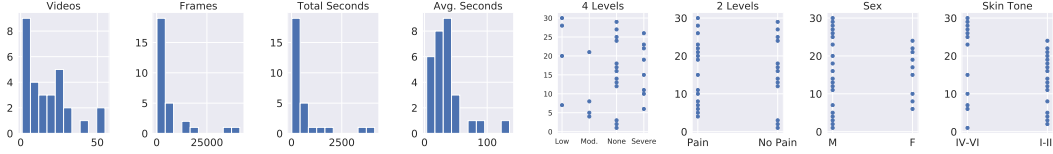


Figure 2: **Distribution of ISS Data.** Histograms for the total videos, frames, seconds, and average seconds per video, for the ISS dataset are in the four left-most plots. The four plots on the right illustrate the distribution of patients (y axis) by the four pain levels, when combined into two pain levels, by sex, and by skin type.

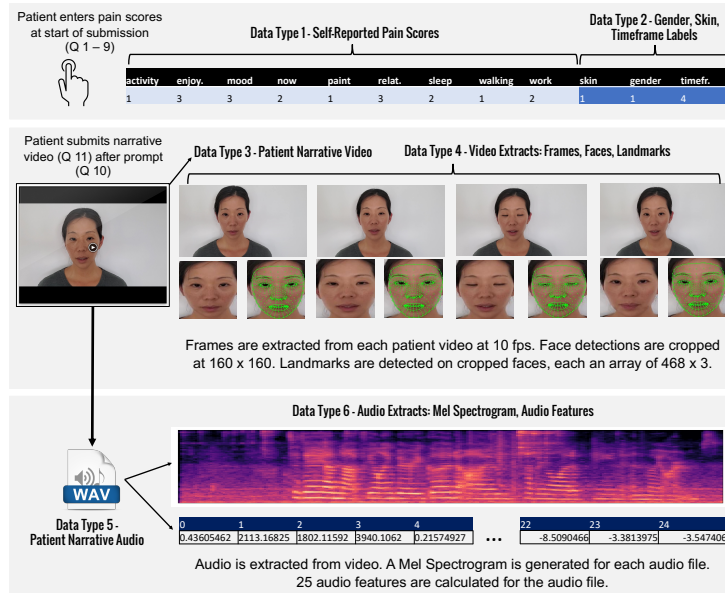


Figure 3: **ISS Data Types.** Facial images shown are *not* actual patients from the ISS dataset due to privacy restrictions. The ISS Dataset currently consists of six types of data: 1) Nine Self-Reported Pain Scores, 2) Labels for Sex, Skin Type, and Timeframe, 3) Patient Narrative Video, 4) Video Extracts: Frames, Faces, Landmarks, and 5) Patient Audio, and 6) Audio Extracts: Mel Spectrogram, Audio Features.

214 implements a fast and CUDA-enabled version of the Multi-task Cascaded Convolutional Networks
 215 (MTCNN) algorithm [46] using an InceptionResnetV1 model pre-trained on VGGFace2 for face
 216 detection and cropping faces from frames. All patient faces were recorded in a frontally aligned
 217 position so no realignment was implemented. Similar to [26], we extract features using AAMs.
 218 Specifically, we use the Google MediaPipe [47] Face Mesh AAM model based on 3D Morphable
 219 Models [15] to detect facial landmarks where each face returns an array of 468 points for three
 220 coordinates. From the audio .wav file, we use the Librosa [48] library to generate a Mel Spectrogram
 221 ($n_fft=2048$, $hop_length=512$, $n_mels=128$), and apply signal processing to capture audio features
 222 about the .wav file to include Mel-frequency cepstral coefficients (MFCCs), chromogram, spectral
 223 centroid, spectral bandwidth, roll-off frequency, and zero crossing rate, leading to 25 audio features.
 224 We further break up the original video into 4-second chunks leading to 40 frames per video chunk,
 225 extracting its respective .wav file, spectrogram, and audio features.

226 3.3.2 No External Labels

227 In contrast to existing acute pain datasets [7–9], the ISS dataset lacks external offline labeling
 228 traditionally completed using the Facial Action Coding System (FACS) [10]. Per the ISS problem
 229 statement, the goal is to predict patient (self)-reported pain, as opposed to observations made by non-
 230 patients via offline pain coders. There are three reasons for not externally encoding ISS video frames
 231 using FACS. First, researchers agree that FACS is expensive due to the need for a trained coder to
 232 annotate each video frame, making the process time-consuming and clinically infeasible [18, 49].
 233 Second, ethicists and psychologists argue that there is limited evidence that facial expressions

234 are reliably and specifically mapped to emotion production [50, 51]. Emotion production is not
235 necessarily tied to a single set of facial expressions, but relies on the context of the situation and
236 human culture [50]. Third, cancer patients with chronic pain may not display the typical set of facial
237 action units (AUs) commonly associated with acute pain. For example [11] collected video data
238 from 43 outpatient lung cancer patients obtained in a spontaneous home setting [11]. They found that
239 the cancer patients were more subdued in expression, and displayed fewer AUs such as grimaces or
240 clenched teeth, commonly found in facial pain images. As a result, AU labels associated with pain
241 such as brow lowering (AU4), orbital tightening (AU6, AU7), levator contraction (AU9, AU10), and
242 eye closure (AU43) may not be applicable to chronic cancer facial pain detection [52]. However, when
243 the ISS dataset is released, there are no prohibitions on researchers attempting to annotate using FACS.

244 3.4 Data Storage and Access

245 A secure cloud-based environment receives mobile and web-based submissions of patients' video,
246 audio, and survey (nine self-reported pain scores) data. No PII such as names or date of birth is stored,
247 with the exception of face, voice, and sex information. The environment is AWS GovCloud FedRAMP
248 Moderate, with Federal Information Security Management Act (FISMA) moderate Authority-to-
249 Operate (ATO) credentials.

250 The ISS dataset consists of cancer patients discussing their medical conditions. The very nature of the
251 images and videos make the data Protected Health Information (PHI) due to the NIH/NCI not being
252 classified as a "covered entity". Extreme care must be exercised to ensure patient privacy and rights
253 are not violated. As a result, we plan to ensure proper patient protections by placing the collected
254 data in restricted access repositories under the stewardship of the NIH. Members of the scientific
255 community will be able to request access to the data and code which may be granted on a per-case
256 basis. This requirement is necessary to ensure legal requirements are met, avoid public spillage of
257 PII data, and ensure patient trust that their data is used within the scope of the intended scientific
258 use. In return researchers receive access to a dataset with numerous modalities and potential clinical
259 relevance of results.

260 4 Baselines

261 We conduct seven baseline experiments for a classification task to predict each patient's self-reported
262 "Pain Target" level assigned at the start of their enrollment shown in Table 2. These levels are fixed
263 upon enrollment for cohort assignment and remain unchanged throughout the study. As a result, one
264 patient represents a single pain level throughout the study. All experiments are static models, which
265 return predictions on a frame-by-frame level. Given how we are in the initial phase of the ISS study,
266 we train models using facial images, landmarks, and the additional nine self-reported pain scores
267 for emotion and activity. However, we do provide baseline results on 4-second chunks of audio via
268 spectrograms and audio features. These are meant to be representative of common approaches to
269 similar work, and establish the careful curation results in a task more difficult than prior literature
270 with simpler labeling or collection. More details on all results are in the Supplementary Materials.

271 **Training Details** All experiments are trained using 10-fold cross validation where three test patients
272 are withheld in the test set for nine splits and two patients set aside for the tenth split. There is no
273 overlap between training and test sets for each split. Please refer to the Supplementary Materials
274 Appendix Section F.1. Table 10 that shows the "10-fold-CV details - Test Patients per Split." For
275 neural networks in Experiments 1 and 3 - 7, we use a batch size of 16, Adam optimizer with 1e-4
276 learning rate, and cross entropy as the loss function, training for 10 epochs, for all experiments. The
277 batch size of 16 was selected empirically based on cross validation accuracy, after running several
278 experiments varying batch size from 4, 8, 16, 32, and 64. We selected Adam optimizer since it has
279 been used in recent facial pain detection studies such as [24, 53]. We fine-tune ResNet50 as the
280 convolutional neural network (CNN) backbone for all multimodal experiments, which is pretrained
281 on ImageNet. We use PyTorch for model training and train on four NVIDIA Tesla T4 GPUs.
282 Experiments 2 and 3 are trained using the Scikit-Learn library for the Random Forest Classifier, using
283 100 estimators, gini criterion, min_samples_split=2, and min_samples_leaf=1.

284 **Experiment 1: Pain Prediction using Static Face Images** The first set of experiments only uses
285 static, facial images. We fine-tune ResNet-50 [54] pretrained on ImageNet [55] to predict four and
286 two levels of pain. Four levels are "None" (Self-Reported Pain Level 0), "Low" (1-3), "Moderate"

287 (4-6), or "Severe" (7-10), and two levels combine "Low", "Moderate", and "Severe" pain levels into
288 a single "Pain" class. Training binary classifiers for "No Pain"/"Pain" prediction is similar to many
289 existing facial pain detection works [7, 22]. We found that the binary classifier leads to better test
290 patient accuracy scores, and continued Experiments 2 through 7 using only two pain levels.

291 **Experiments 2, 3: Pain Prediction using Static Landmarks or Pain** In these experiments, we
292 use only one modality to train two separate models and use traditional machine learning models,
293 specifically the Random Forest algorithm [56]. Experiment 2 uses the landmark arrays detected for
294 each facial image and Experiment 3 uses the nine self-reported pain scores explained in Section 3.2.1
295 that represent how pain interferes with the patient's emotions and activity, plus labels for sex, skin
296 type, and timeframe. The timeframe label is categorical and is extracted from the video submission
297 timestamp representing what time of day (early AM, late PM, etc.) the video was submitted. For
298 both Experiments 2 and 3, we train a Random Forest Classifier. Note, that the target "Pain Target" is
299 not in the set of the nine self-reported pain scores, which are distinct and separate.

300 **Experiments 4 - 6: Pain Prediction using Static Multimodal Data** We train three multimodal
301 networks using an early, joint fusion strategy as proposed by [57]. For Experiment 4 ("Fusion 1"),
302 we concatenate the fully connected outputs of ResNet50 with raw landmarks. The feature vector is
303 then inputted to a feedforward neural network for binary pain prediction. Experiment 5 ("Fusion 2")
304 concatenates the fully connected outputs of ResNet50 with raw landmarks, in addition to the nine
305 pain scores, skin, sex, and timeframe labels. Similarly, the feature vector is inputted to the same
306 feedforward network architecture for binary pain prediction. Experiment 6 ("Fusion 3") concatenates
307 three vectors: the feature map from `layer-4-conv2D-1`, the landmark features outputted from a
308 landmark-specific feedforward network, and the nine pain scores, sex, skin, and timeframe features
309 outputted from a pain-specific feedforward network. The resulting feature vector is inputted to a
310 CNN for binary pain prediction.

311 **Experiment 7: Audio Models** Experiment 7 is a binary pain prediction model that uses the Mel
312 spectrogram image and 25 audio features from 4-second chunks of audio extracted from each patient
313 video. A feature vector resulting from the concatenation of the spectrogram feature map from
314 `layer-4-conv2D-1` and audio features learned by a feedforward network, are inputted to the same
315 CNN architecture as used in Experiment 6. Diagrams for all experimental architectures are provided
316 in the Supplementary Materials.

317 5 Results

318 **Accuracy Calculation** The accuracy of each model is evaluated for each test patient using the tenth
319 model checkpoint. Using the checkpoint, we evaluate each test patient individually. We only evaluate
320 test patients using their respective, assigned split per 10-fold cross-validation (See Supplementary
321 Materials Section F.1. Table 10 "10-fold-CV details - Test Patients per Split" for details). For example,
322 test patients 0002, 0029, and 0021 are only evaluated using the trained model from Split 1, not Split 2
323 which would have included these three patients in its training set. We evaluate each test patient using
324 a batch size of 1, predicting the target pain score for each patient image. We then calculate accuracy
325 for the test patient in question as simply $accuracy_score(y_true, y_pred)$ where y_true is the set
326 of true "Pain Target" labels and y_pred is the set of predicted "Pain Target" labels.

327 As a result, in Table 5, we show the mean accuracy computed for each "Pain Target" level across all
328 test patients ("No Pain" or "Pain" for two levels, and "None", "Low", "Moderate", or "Severe" for
329 four levels of pain). For example, in Experiment 1 "ResNet50-4-static", the accuracy scores for all
330 patients with ground truth pain labels of "None", were averaged together to calculate the result of
331 0.583. In Figures 4 and 5, the bars are color-coded by the ground truth "Pain Target" level for each
332 patient. The y-axis is the accuracy predicted for the patient. For example, upon zooming into Figure
333 4a, Patient 0029's (8 marks from the right of the x-axis) ground truth "Pain Target" level is "No Pain".
334 However, the Experiment 1 static binary model only predicts it with 0.309 accuracy.

335 **Experiment Results** The Experiment 6 multimodal network combining multiple features from
336 the facial images, landmarks, pain scores, sex, skin, and timeframe labels performs the best for
337 overall pain classification. Compared to training on a single modality alone (Experiments 1, 2, 3, 7),
338 Experiment 6 (Fusion 3) shows the best overall class accuracy of 0.657 shown in Table 5. Fusion 3 also
339 shows the highest accuracy for the "Pain" level at 0.717. Experiment 6 (Fusion 3) led to 72.4% of test
340 patients exceeding 50% accuracy per frame as noted in Figure 5b. However, it ties with Experiment

341 5 (Fusion2) and Experiment 2 (Random Forest PM) for 51.7% of test patients achieving over 75%
 342 accuracy per Figure 5a and Figure 4d. While the Random Forest pain model (Figure 5d) shows greater
 343 “No Pain” accuracy, using only the self-reported nine pain scores does not detect the original “Low”
 344 pain levels as well as the multimodal Fusion 3 model visualized in Figure 5c shown in blue bars.

345 Experiment 3 (Random Forest Pain) shows the highest “No Pain” accuracy scores at 0.706 per Table
 346 5. Adding the nine self-reported pain scores appears to boost accuracy, compared to training only
 347 on faces and landmarks per Experiment 4 (Fusion 1, 0.513) in Table 5. This is likely due to high
 348 correlations between the nine reported pain scores. Analysis shows strong Pearson correlation values
 349 exceeding 0.89 among activity, mood, work, enjoyment, and relationship scores. Continued analysis
 350 as more patients enroll in the study is required to understand the effect of the nine pain scores across
 351 all patients. The facial landmarks perform the worst in Experiment 2 (Random Forest LM) with only
 352 37.9% of test patients exceeding better than random at over 50% accuracy per Figure 4b. However,
 353 when adding landmarks to facial images in Experiment 4 (Fusion 1), several test patients completely
 354 fail to be detected (1, 2, 16, 13, 29, 3, 28, 25) per Figure 4d. This may be consistent with recent
 355 research by [34] who show that landmark detection declines when comparing different populations,
 356 such as older patients with dementia, to healthy adults.

Table 5: Experiment Results by Pain Level Accuracy. “LM” indicates facial landmarks.

Experiment	4-Class Model	Data	All Classes	None	Low	Moderate	Severe
Exp. 1	ResNet50-4-static	Faces , only	0.378	0.583	0.168	0.252	0.213
			All Classes	No Pain	Pain		
Exp. 1	ResNet50-2-static	Faces, only	0.568	0.513	0.612		
Exp. 2	Random Forest LM	Landmarks, only	0.373	0.479	0.287		
Exp. 3	Random Forest Pain	Pain Scores, only	0.650	0.706	0.602		
Exp. 4	Fusion 1	Faces + Landmarks	0.513	0.304	0.683		
Exp. 5	Fusion 2	Faces + Landmarks + Pain Scores	0.631	0.563	0.687		
Exp. 6	Fusion 3	Faces + Landmarks + Pain Scores	0.657	0.582	0.717		
Exp. 7	Static Audio	Audio, only	0.456	0.645	0.303		

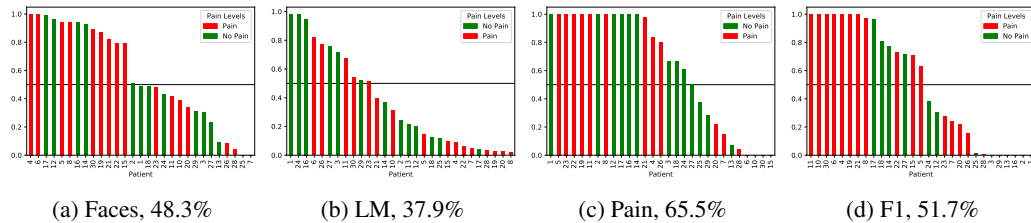


Figure 4: Accuracy Scores per Test Patient by Model: Faces, Landmarks, Pain, and Fusion 1. We show the resulting scores per test patient for the binary pain classifiers. Horizontal bar indicates 50% accuracy. Percentages in sub-captions indicates the number of patients exceeding 50% test accuracy. Notation: Faces=ResNet50-2-static; LM=Random Forest LM (landmarks); Pain=Random Forest Pain; F1=Fusion1. Best viewed in color and zoomed in.

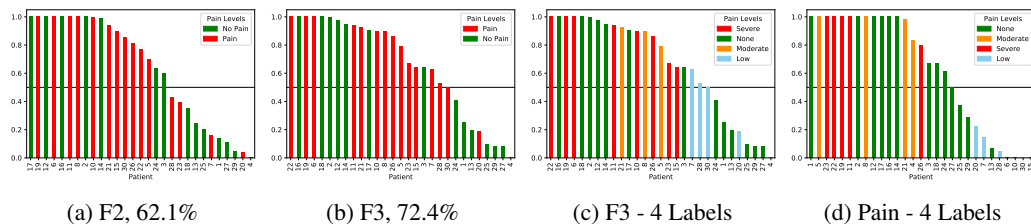


Figure 5: Accuracy Scores per Test Patient by Model - Fusion 2, 3, and Pain, Visualized with 4 Original Labels. We show the resulting scores per test patient for the binary pain classifiers. Horizontal bar indicates 50% accuracy. Percentages in sub-captions indicates the number of patients exceeding 50% test accuracy. Notation: Pain=Random Forest Pain; F1=Fusion1; F2=Fusion2; F3=Fusion3. Best viewed in color and zoomed in.

357 6 Discussion and Future Work

358 Due to a variety of state-of-the-art techniques, we sought to implement simple models to demonstrate
359 baseline results using fairly minimal preprocessing, transformations, and architectures. The results of
360 our models show the dataset’s difficulty. For comparison, acute pain detection studies have shown
361 accuracy scores up to 82.4% (hit rate) [26] using the UNBC-McMaster Shoulder Dataset and 95%
362 for multimodal infant pain detection using a custom dataset by [31]. Chronic pain detection using
363 psychological inventories have achieved 86.5% (cross-validated balanced accuracy) using a support
364 vector machine [58].

365 **Limitations** The first limitation of the dataset is the low number of currently enrolled patients at only
366 29 patients and the imbalance across pain levels. However, we observe that two new patients enroll
367 into the study every month. As the number of patients grow, we expect a more balanced distribution of
368 pain levels, sex, skin type, and increased volume of data, consistent with the cohort design indicated in
369 Table 2. However, medical datasets using active patient populations for major diseases such as cancer,
370 are extremely scarce due to the time and review required for medical privacy and ethics. This differs
371 greatly from current pain datasets that have recruited fairly healthy patients, who are not actively
372 undergoing disease treatment. Due to the special sensitivity of the ISS study population, we believe
373 that our current initial results offers important insights currently missing in the medical AI community.

374 Next, despite the patient instructions to complete the submission in a quiet, brightly lit room with a
375 white wall or background, many videos submitted varied in quality and resolution. The following
376 examples observed in the dataset present challenges to machine learning: 1) Patient sitting in front of
377 a door with signage in the background showing letters and numbers; 2) Patient occasionally wears a
378 mask in some videos (due to Covid-19); 3) Patient records video in area of intense sunshine and glare
379 causing reflection from various surfaces; 4) Patient records in a dark, shady room, leading to grainy
380 resolution and video quality; 5) Patient speaks very quietly or muffled, making it difficult to hear the
381 patient narrative; 6) Missing data as is the case of Patient 0009 and absent self-reported nine pain
382 scores from Patient 0015.

383 **Ethics** Publicly available acute pain datasets have lacked ethnic diversity. For example, the UNBC-
384 McMaster Database [6] uses ethnicity as a demographic indicator where out of the original 129
385 patients (63 Male, 66 Female), a minimum of 13.2% (17 patients) consisted of non-Caucasian
386 ethnicity (refer to Table 1 of [6]). It may be less given how studies using the UNBC-McMaster dataset
387 have access to data from only 25 out of 129 patients [20–22, 24]. The BioVid and MIntPAIN datasets
388 provide no information about ethnicity and race [8, 9]. EmoPAIN contains 22 patients (18 Caucasian,
389 3 African-American, 1 South-Asian) who are majority white [30]. As a result, we sought to increase
390 the diversity of enrolled patients by using cohorts that include sex and skin type specifications. While
391 the Fitzpatrick Skin Type scale was originally developed for dermatological use, it has recently
392 been criticized for its conflation with race and ethnicity [59]. It has been found to overestimate the
393 prevalence of Type IV skin classification in African Americans [60]. The visual grouping of patients
394 into lighter tones (Skin Types I - III) or darker tones (Skin Types IV - VI) may be too restrictive and
395 biased in terms of broadening our diversity of patients. As a result, the ISS dataset requires careful
396 monitoring and a regular ethics review.

397 **Future Work** The second phase of the study will analyze more diverse modalities. First, we will
398 extract text from the audio files and explore its utility towards multimodal pain models. Next, since
399 patients were unable to conduct in-clinic visits, we were unable to gather thermal imagery captured
400 from a thermal camera stationed at the clinic. Thermal imagery offers insights into physiological
401 states that is unseen on visible images alone [61]. Our intent is to generate paired visible-thermal
402 datasets as collected by the Iris, Eurecom, and Equinox datasets [62–64]. Lastly, we estimate that
403 after enrolling 112 patients, the ISS dataset will contain an additional 1,456 videos, 543,733 frames,
404 and 3.8 hours of content.

405 References

- 406 [1] M. Van den Beuken-van Everdingen, J. De Rijke, A. Kessels, H. Schouten, M. Van Kleef, and J. Patijn,
407 “Prevalence of pain in patients with cancer: a systematic review of the past 40 years,” *Annals of oncology*,
408 vol. 18, no. 9, pp. 1437–1449, 2007.
- 409 [2] V. C.-Y. Sun, T. Borneman, B. Ferrell, B. Piper, M. Koczywas, and K. Choi, “Overcoming barriers to
410 cancer pain management: an institutional change model,” *Journal of pain and symptom management*,

- 411 vol. 34, no. 4, pp. 359–369, 2007.
- 412 [3] J. F. Cleary, “Cancer pain management,” *Cancer Control*, vol. 7, no. 2, pp. 120–131, 2000.
- 413 [4] J. Lötsch and A. Ultsch, “Machine learning in pain research,” *Pain*, vol. 159, no. 4, p. 623, 2018.
- 414 [5] B. Hu, C. Kim, X. Ning, and X. Xu, “Using a deep learning network to recognise low back pain in static
415 standing,” *Ergonomics*, vol. 61, no. 10, pp. 1374–1381, 2018.
- 416 [6] K. M. Prkachin and P. E. Solomon, “The structure, reliability and validity of pain expression: Evidence
417 from patients with shoulder pain,” *Pain*, vol. 139, no. 2, pp. 267–274, 2008.
- 418 [7] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, “Painful data: The unbc-mcmaster
419 shoulder pain expression archive database,” in *2011 IEEE International Conference on Automatic Face &
420 Gesture Recognition (FG)*. IEEE, 2011, pp. 57–64.
- 421 [8] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. O. Andrade,
422 and G. M. da Silva, “The biovid heat pain database data for the advancement and systematic validation of
423 an automated pain recognition system,” in *2013 IEEE international conference on cybernetics (CYBCO)*.
424 IEEE, 2013, pp. 128–131.
- 425 [9] M. A. Haque, R. B. Bautista, F. Noroozi, K. Kulkarni, C. B. Laursen, R. Irani, M. Bellantonio, S. Escalera,
426 G. Anbarjafari, K. Nasrollahi *et al.*, “Deep multimodal pain recognition: a database and comparison of
427 spatio-temporal visual modalities,” in *2018 13th IEEE International Conference on Automatic Face &
428 Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 250–257.
- 429 [10] P. Ekman and V. Friesen, “Facial action coding system manual,” 1978.
- 430 [11] D. J. Wilkie, “Facial expressions of pain in lung cancer,” *Analgesia*, vol. 1, no. 2, pp. 91–99, 1995.
- 431 [12] N. C. I. (NCI), “Machine learning to analyze facial imaging, voice and spoken language for the
432 capture and classification of cancer/tumor pain - full text view,” Jun 2020. [Online]. Available:
433 <https://clinicaltrials.gov/ct2/show/NCT04442425>
- 434 [13] I. J. Pomeranec, A. Cha, M. Rule, C. Ordun, M. Hirlinger, and J. Gulley, “Intelligent sight & sound: Ma-
435 chine learning to analyze facial imaging and voice mapping for classification of cancer pain,” *Neurosurgery*,
436 vol. 67, no. Supplement_1, p. nyaa447_524, 2020.
- 437 [14] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” *IEEE Transactions on pattern
438 analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.
- 439 [15] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, “Real-time facial surface geometry from
440 monocular video on mobile gpus,” *arXiv preprint arXiv:1907.06724*, 2019.
- 441 [16] C. S. Cleeland, “Pain assessment in cancer,” *Effect of cancer on quality of life*, vol. 293, p. 305, 1991.
- 442 [17] C. S. Cleeland and K. Ryan, “The brief pain inventory,” *Pain Research Group*, pp. 143–147, 1991.
- 443 [18] M. Kunz, D. Seuss, T. Hassan, J. U. Garbas, M. Siebers, U. Schmid, M. Schöberl, and S. Lautenbacher,
444 “Problems of video-based pain detection in patients with dementia: a road map to an interdisciplinary
445 solution,” *BMC geriatrics*, vol. 17, no. 1, pp. 1–8, 2017.
- 446 [19] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, “Automatically
447 detecting pain in video through facial action units,” *IEEE Transactions on Systems, Man, and Cybernetics,
448 Part B (Cybernetics)*, vol. 41, no. 3, pp. 664–674, 2010.
- 449 [20] J. Zhou, X. Hong, F. Su, and G. Zhao, “Recurrent convolutional neural network regression for continuous
450 pain intensity estimation in video,” in *Proceedings of the IEEE conference on computer vision and pattern
451 recognition workshops*, 2016, pp. 84–92.
- 452 [21] N. Neshov and A. Manolova, “Pain detection from facial characteristics using supervised descent method,”
453 in *2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing
454 Systems: Technology and Applications (IDAACS)*, vol. 1. IEEE, 2015, pp. 251–256.
- 455 [22] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, “Pain detection through shape and appearance features,”
456 in *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2013, pp. 1–6.
- 457 [23] L. Lo Presti and M. La Cascia, “Using hankel matrices for dynamics-based facial emotion recognition
458 and pain detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition
459 workshops*, 2015, pp. 26–33.

- 460 [24] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B.
461 Moeslund, P. Rasti, and G. Anbarjafari, "Spatio-temporal pain recognition in cnn-based super-resolved
462 facial images," in *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*.
463 Springer, 2016, pp. 151–162.
- 464 [25] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid,
465 "Automatic detection of pain from facial expressions: a survey," *IEEE transactions on pattern analysis and*
466 *machine intelligence*, vol. 43, no. 6, pp. 1815–1831, 2019.
- 467 [26] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful
468 face–pain expression recognition using active appearance models," *Image and vision computing*, vol. 27,
469 no. 12, pp. 1788–1796, 2009.
- 470 [27] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, J. R. Movellan *et al.*, "Automatic
471 recognition of facial actions in spontaneous expressions." *J. Multim.*, vol. 1, no. 6, pp. 22–35, 2006.
- 472 [28] S. Gruss, M. Geiger, P. Werner, O. Wilhelm, H. C. Traue, A. Al-Hamadi, and S. Walter, "Multi-modal
473 signals for analyzing pain responses to thermal and electrical stimuli," *JoVE (Journal of Visualized*
474 *Experiments)*, no. 146, p. e59057, 2019.
- 475 [29] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic pain recognition from
476 video and biomedical signals," in *2014 22nd International Conference on Pattern Recognition*. IEEE,
477 2014, pp. 4582–4587.
- 478 [30] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng,
479 A. Kemp, M. Shafizadeh *et al.*, "The automatic detection of chronic pain-related expression: requirements,
480 challenges and the multimodal emopain dataset," *IEEE transactions on affective computing*, vol. 7, no. 4,
481 pp. 435–451, 2015.
- 482 [31] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, and Y. Sun, "An approach for automated
483 multimodal analysis of infants' pain," in *2016 23rd International Conference on Pattern Recognition*
484 *(ICPR)*. IEEE, 2016, pp. 4148–4153.
- 485 [32] G. Zamzmi, D. Goldgof, R. Kasturi, and Y. Sun, "Neonatal pain expression recognition using transfer
486 learning," *arXiv preprint arXiv:1807.01631*, 2018.
- 487 [33] S. Rezaei, A. Moturu, S. Zhao, K. M. Prkachin, T. Hadjistavropoulos, and B. Taati, "Unobtrusive pain
488 monitoring in older adults with dementia using pairwise and contrastive training," *IEEE Journal of*
489 *Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1450–1462, 2020.
- 490 [34] A. Asgarian, S. Zhao, A. B. Ashraf, M. E. Browne, K. M. Prkachin, A. Mihailidis, T. Hadjistavropoulos,
491 and B. Taati, "Limitations and biases in facial landmark detection d an empirical study on older adults with
492 dementia." in *CVPR Workshops*, 2019, pp. 28–36.
- 493 [35] M. S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun, "Multimodal spatio-temporal deep
494 learning approach for neonatal postoperative pain assessment," *Computers in Biology and Medicine*, vol.
495 129, p. 104150, 2021.
- 496 [36] ———, "Multi-channel neural network for assessing neonatal pain from videos," in *2019 IEEE International*
497 *Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 1551–1556.
- 498 [37] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, S. Gruss, and S. Walter, "Automatic vs. human recognition
499 of pain intensity from facial expression on the x-ite pain database," *Sensors*, vol. 21, no. 9, p. 3273, 2021.
- 500 [38] L. Goldsmith, S. Katz, B. Bilchrest, A. Paller, D. Leffel, and K. Wolff, "Fitzpatrick's dermatology in
501 general medicine, ed," *McGrawHill Medical*, pp. 2421–2429, 2012.
- 502 [39] M. Haefeli and A. Elfering, "Pain assessment," *European Spine Journal*, vol. 15, no. 1, pp. S17–S24, 2006.
- 503 [40] J. Apolinário-Hagen, L. Fritsche, C. Bierhals, and C. Salewski, "Improving attitudes toward e-mental health
504 services in the general population via psychoeducational information material: A randomized controlled
505 trial," *Internet interventions*, vol. 12, pp. 141–149, 2018.
- 506 [41] J. Fink-Lamotte, A. Widmann, J. Fader, and C. Exner, "Interpretation bias and contamination-based
507 obsessive-compulsive symptoms influence emotional intensity related to disgust and fear," *PloS one*,
508 vol. 15, no. 4, p. e0232362, 2020.
- 509 [42] J. R. Livesay and T. Porter, "Emg and cardiovascular responses to emotionally provocative photographs
510 and text," *Perceptual and motor skills*, vol. 79, no. 1, pp. 579–594, 1994.

- 511 [43] P. MacLachlan and L. Beech, *Sarah, plain and tall*. Scholastic Inc., 1997.
- 512 [44] K. DiCamillo, *Because of Winn-Dixie*. Candlewick Press, 2009.
- 513 [45] T. S. Geisel, *Oh, the places you'll go!* Random House Books for Young Readers, 1990.
- 514 [46] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded
515 convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
- 516 [47] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong,
517 J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*,
518 2019.
- 519 [48] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thom e, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee,
520 O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov,
521 K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Heref u, F.-R. St oter, P. Friesch,
522 A. Weiss, M. Vollrath, T. Kim, and Thassilo, "librosa/librosa: 0.8.1rc2," May 2021. [Online]. Available:
523 <https://doi.org/10.5281/zenodo.4792298>
- 524 [49] Z. Chen, R. Ansari, and D. Wilkie, "Automated pain detection from facial expressions using facts: A review,"
525 *arXiv preprint arXiv:1811.07988*, 2018.
- 526 [50] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered:
527 Challenges to inferring emotion from human facial movements," *Psychological science in the public*
528 *interest*, vol. 20, no. 1, pp. 1–68, 2019.
- 529 [51] K. Crawford, *The Atlas of AI*. Yale University Press, 2021.
- 530 [52] K. M. Prkachin, "The consistency of facial expressions of pain: a comparison across modalities," *Pain*,
531 vol. 51, no. 3, pp. 297–306, 1992.
- 532 [53] P. Rodriguez, G. Cucurull, J. Gonz alez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca,
533 "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE*
534 *transactions on cybernetics*, 2017.
- 535 [54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the*
536 *IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- 537 [55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image
538 database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- 539 [56] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- 540 [57] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and
541 electronic health records using deep learning: a systematic review and implementation guidelines," *NPJ*
542 *digital medicine*, vol. 3, no. 1, pp. 1–9, 2020.
- 543 [58] L. A. Antonucci, A. Taurino, D. Laera, P. Taurisano, J. Losole, S. Lutricuso, C. Abbatantuono, M. Giglio,
544 M. F. De Caro, G. Varrassi *et al.*, "An ensemble of psychological and physical health indices discriminates
545 between individuals with chronic pain and healthy controls with high reliability: A machine learning study,"
546 *Pain and Therapy*, vol. 9, no. 2, pp. 601–614, 2020.
- 547 [59] O. R. Ware, J. E. Dawson, M. M. Shinohara, and S. C. Taylor, "Racial limitations of fitzpatrick skin type,"
548 *Cutis*, vol. 105, no. 2, pp. 77–80, 2020.
- 549 [60] L. C. Pichon, H. Landrine, I. Corral, Y. Hao, J. A. Mayer, and K. D. Hoerster, "Measuring skin cancer risk
550 in african americans: is the fitzpatrick skin type classification scale culturally sensitive," *Ethn Dis*, vol. 20,
551 no. 2, pp. 174–179, 2010.
- 552 [61] C. Ordun, E. Raff, and S. Purushotham, "The use of ai for thermal emotion recognition: A review of
553 problems and limitations in standard design and data," *arXiv preprint arXiv:2009.10589*, 2020.
- 554 [62] A. Selinger *et al.*, "Appearance-based facial recognition using visible and thermal imagery: a comparative
555 study," Equinox Corp., Tech. Rep., 2006.
- 556 [63] OTCBVS benchmark dataset collection. [Online]. Available: <http://vcipl-okstate.org/pbvs/bench/>
- 557 [64] K. Mallat *et al.*, "A benchmark database of visible and thermal paired face images across multiple variations,"
558 in *BIOSIG*. IEEE, 2018, pp. 1–5.
- 559 [65] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender
560 classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.