Automated Equine Pain Recognition from Facial Videos Using I3D Feature Extraction and Bi-LSTM Temporal Fusion

Rimsha Mahmood
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
rmahmood.bese20seecs@seecs.edu.pk

***Abstract*—Effective equine pain management requires continuous monitoring, yet current methods rely heavily on subjective manual scoring or intermittent observations. This paper presents an automated pipeline for recognizing mild and moderate pain in horses using facial video analysis. Utilizing a dataset of 12 horses annotated with the Equine Facial Action Coding System (EquiFACS), we propose a region-aware spatiotemporal framework. Our approach leverages YOLOv8 for robust Region of Interest (ROI) detection (eyes and chin), followed by I3D feature extraction to capture subtle micro-expressions. To address the temporal dynamics of pain, we integrate a Bi-directional Long Short-Term Memory (Bi-LSTM) network with temporal attention. Experimental results demonstrate that the chin region contains the most discriminative pain cues. While data scarcity remains a challenge for end-to-end training, our region-specific I3D models achieve a video-level accuracy of 66.7%, highlighting the potential of automated systems to assist in objective veterinary welfare assessment.**

***Index Terms*—Equine pain detection, facial expression recognition, I3D, bi-directional LSTM, temporal attention, video classification, animal welfare.**

I. Introduction

A. Motivation and Background

Horses play a major economic and social role in sport, leisure, and agriculture, which makes effective welfare monitoring and timely pain management clinically and ethically important [1], [2]. In practice, equine pain assessment still relies heavily on behavioural observation and composite scoring systems applied by trained veterinarians or caretakers, which are labour-intensive, require substantial expertise, and are difficult to apply continuously in real-world settings [2], [3].

Facial-expression-based tools such as the Horse Grimace Scale (HGS), EquiFACS, and related "equine pain face" descriptions formalise how specific configurations of ears, eyes, muzzle, and nostrils relate to pain, providing interpretable region-level scores validated for acute pain in horses [4]–[7]. These scales have established that facial expressions are a rich and clinically meaningful source of pain information, forming the semantic backbone for current automatic equine pain recognition systems [2], [8].

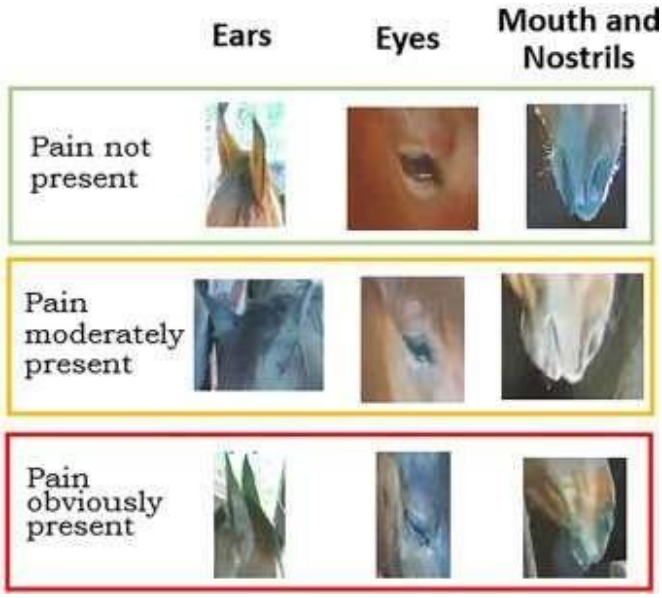

Fig. 1: Sample frames from Lencioni *et al.* [3], showing mild, moderate and obvious pain with visible changes in ear, eye, and mouth/nostril regions.

B. Challenges in Equine Pain Recognition

Despite their clinical value, manual grimace scales are constrained by observer training requirements, inter-rater variability, and the fact that prey species such as horses may suppress overt pain behaviours in the presence of humans [1]–[3]. Frequent manual scoring is ill-suited to continuous tracking of pain trajectories before and after interventions.

Automatic, camera-based systems offer a route towards unobtrusive, long-duration monitoring, but they face several domain-specific challenges: facial pain datasets are small and imbalanced, severe pain examples are rare, head pose and coat colour vary widely, and the most informative cues may lie in subtle temporal changes rather than single frames [8]–[10]. These factors make it difficult for purely spatial, image-based deep models to generalise across individuals.

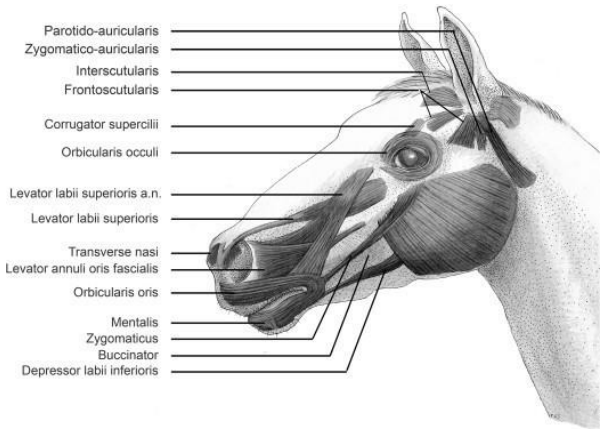

Fig. 2: EquiFACS-based schematic of the equine face showing anatomical regions and example facial action units (AUs) relevant for pain assessment, adapted from Wathan *et al.* [5].

TABLE I: Gap analysis of automatic equine facial pain assessment.

Study	Core Limitation / Gap
Dalla Costa *et al.* [4]	Manual HGS scoring only; no automation or temporal modelling.
Wathan *et al.* [5]	Detailed AU vocabulary but annotation is expert-intensive.
Lencioni *et al.* [3]	CNNs on ROIs; small cohort, frame-wise analysis only.
Pessanha *et al.* [8]	Landmark-based ROIs are brittle under extreme poses; mostly image-level.
Ruhof *et al.* [9]	Robust YOLO detection but limited temporal modelling.
Broome *et al.* [10]	Identifies temporal and data gaps but proposes no full model.

C. Related Work and Gaps

Early automatic approaches focused on transferring clinically validated facial scales into image-level classifiers. Lencioni *et al.* trained CNNs on HGS-inspired regions and achieved promising accuracy for three pain levels, but the study was limited to seven horses and treated frames largely independently [3]. Pessanha *et al.* introduced the UU Equine Pain Face dataset and a hierarchical pipeline utilizing head pose, landmarks, and SVMs [8]. Ruhof *et al.* replaced landmark-driven ROIs with YOLOv8-based patch detection and reported improved robustness to pose and breed variation [9].

Survey work emphasises that data scarcity, inter-rater disagreement, domain shift between pain aetiologies, and underutilisation of temporal information remain core bottlenecks [2], [10]. Most existing systems operate on still images or sparsely sampled frames, lacking explicit spatiotemporal modelling of facial action trajectories. A summary of key gaps is given in Table I.

TABLE II: Summary of experimental dataset derived from the EquiFACS pain corpus.

ID	Pain	Dur.(s)	Frames	Dominant EquiFACS (Eye / Chin)
S1	Mild	25–35	180–260	AU101, AU5, AU47 / AU10, AU17
S2	Mild	25–35	180–260	AU101, AU47 / AU10, AU24
S3	Mild	25–35	180–260	AU101, AU5 / AU10, AU18
S4	Mod.	25–35	180–260	AU101, AU5 / AU10, AU17, AU24
S5	Mod.	25–35	180–260	AU101, AU145 / AU10, AU18
S6	Mod.	25–35	180–260	AU101, AU5 / AU10, AU17
S7	Mod.	25–35	180–260	AU101, AU47 / AU10, AU24
S8	Mod.	25–35	180–260	AU101, AU5 / AU10, AU18, AU24
S9	Mild	25–35	180–260	AU101, AU5 / AU10, AU17
S10	Mod.	25–35	180–260	AU101, AU47 / AU10, AU18
S11	Mod.	25–35	180–260	AU101, AU5 / AU10, AU17, AU24
S12	Mod.	25–35	180–260	AU101, AU47 / AU10, AU18

D. Contributions

This work addresses these gaps by combining clinically grounded facial pain scales with explicit temporal modelling.

1) We utilize HGS and EquiFACS as a semantic backbone for defining pain-relevant facial regions.
2) We implement a robust ROI extraction pipeline using YOLOv8 to handle pose and appearance variation.
3) We introduce a temporally aware modelling strategy (I3D + Bi-LSTM) designed for low frame-rate equine video to capture evolving pain expressions that are often missed by purely spatial classifiers.

II. Methodology

A. Dataset and EquiFACS Annotation

1) Dataset characteristics: We use the EquiFACS pain dataset of Rashid *et al.*, which provides short video clips of horses annotated for global pain level and frame-wise EquiFACS codes [7]. From the full corpus we selected 12 subjects (S1–S12), each contributing one codeable segment labelled as mild or moderate pain. Four segments correspond to mild pain and eight to moderate pain, reflecting natural clinical class imbalance. Table II summarises the resulting experimental set.

2) EquiFACS-based region selection: We grouped EquiFACS codes into eye- and chin-focused sets to drive

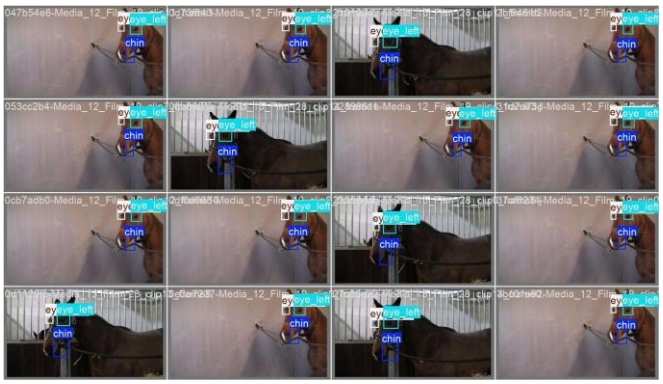

Fig. 3: Example YOLOv8 detections for chin and eye regions on different subjects.

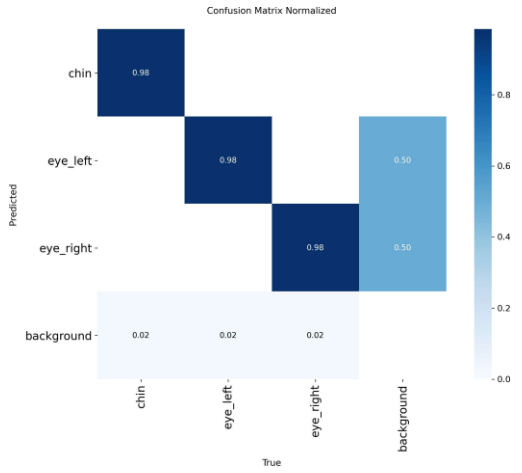

Fig. 4: Normalised confusion matrix for the ROI detector on the validation set.

ROI definition. For the bilateral eye regions we used codes covering inner-brow raising, eye-white increase, and blinking,

$$A_{\text{eyes}} = \{\text{AD1, AD133, AD160, AD1L, AU101, AU101L, AU145, AU145L, AU47, AU47L, AU5, AU5L}\}, \quad (1)$$

while for the chin/lower-lip region we defined

$$A_{\text{chin}} = \{\text{AD38, AU10, AU113, AU17, AU18, AU24}\}, \quad (2)$$

capturing upper-lip raising, sharp lip pulls, and chin-raising actions linked to lower-face pain tension [5], [7].

B. Manual ROI Annotation and Automated Detection

1) Label Studio annotation: Ground-truth boxes for chin, left eye, and right eye were created in Label Studio, an open-source annotation tool, on a subset of frames from all 12 subjects.[1] Every 10th frame plus frames around EquiFACS events were sampled, giving 50–100 frames per video and over 800 annotated frames in total, using the temporal coding of Rashid *et al.* [7]. Veterinary students drew tight boxes around the chin and peri-orbital regions, a senior veterinarian verified all labels, and on 100 double-annotated frames the mean IoU between annotators exceeded 0.85. Annotations were exported in normalised YOLO format (`class_id x_center y_center width height`) for detector training [11].

2) YOLOv8 ROI detector: A YOLOv8-nano model was trained to detect the three ROIs, using 640×640 inputs, batch size 16, up to 100 epochs with early stopping, SGD (momentum 0.937), an initial learning rate of 0.01 with cosine decay, and standard augmentations (Mosaic, MixUp, HSV jitter, horizontal flip), following the Ultralytics implementation [11]. The 800 frames were split 80 % / 20 % into train and validation sets, stratified by video. The final model achieved mAP@0.5 = 0.987 overall, with class-wise mAPs of 0.987 (chin), 0.987 (left eye), and 0.988 (right eye); the normalised confusion matrix shows over 0.98 correct assignments for all ROI classes and very few background false positives.

[1]https://labelstud.io

3) Region extraction: The trained detector was applied to every frame of all videos to generate ROI crops. Detections above 0.60 confidence (chin) and 0.65 (eyes) were kept; missing boxes (< 3 % of frames) were filled by linear interpolation or, when necessary, by the per-subject mean box. Cropped regions were stored as separate image sequences per subject and region, enabling independent processing of chin, left-eye, and right-eye streams in later stages.

C. Preprocessing Pipeline

The cropped ROI sequences undergo a three-stage preprocessing pipeline to standardise frame rate, reduce redundancy and prepare fixed-size inputs for the video backbone.

1) Temporal subsampling: The original EquiFACS clips were recorded at 7–10 fps, leading to high temporal redundancy between adjacent frames. To reduce this redundancy and the computational load, every fourth frame was retained (stride 4), yielding an effective rate of approximately 2–2.5 fps. Visual inspection confirmed that this rate still captured all salient facial changes while reducing the number of frames by about 75 %, which substantially accelerates later processing and training.

2) RIFE interpolation to 50 fps: For spatiotemporal CNNs, very low frame rates are undesirable. Therefore, each subsampled ROI sequence was temporally densified using Real-Time Intermediate Flow Estimation (RIFE), a neural frame interpolation method that combines optical flow and convolutional networks to synthesise intermediate frames between two inputs [15]. In our configuration, subsampled sequences (N frames at ≈2 fps) were interpolated to 50 fps, producing $2N - 1$ frames per sequence. Qualitative inspection showed smooth, realistic motion without visible artefacts, and the interpolated sequences preserved pain-related facial configurations observed in the original clips.

3) Spatial preprocessing and normalisation: After temporal processing, each ROI frame was resized to 112×112 pixels

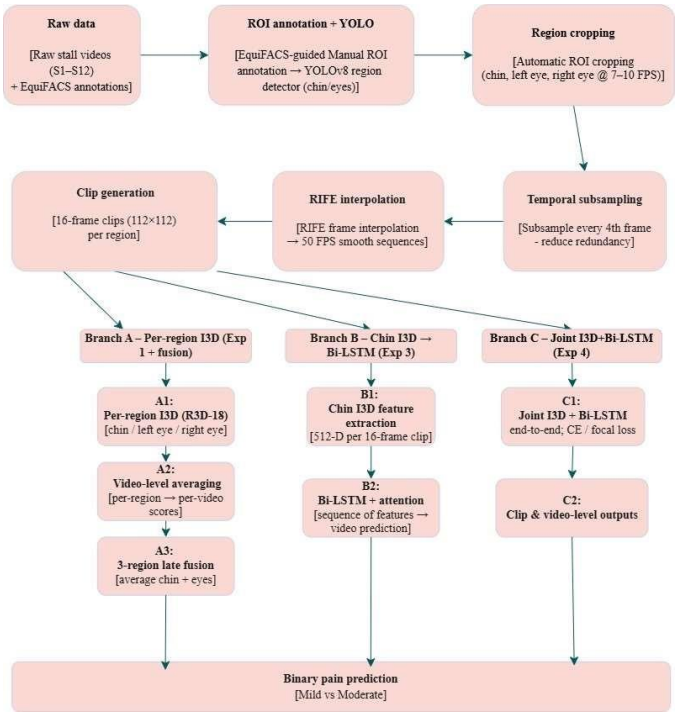

Fig. 5: Preprocessing pipeline from raw videos to standardised region clips: raw video → YOLO ROI detection and cropping → temporal subsampling → RIFE interpolation to 50 fps → spatial resizing and normalisation.

using bilinear interpolation to match the input size of the R3D-18/I3D backbone. Pixel values were scaled to $[0, 1]$ and standardised using ImageNet statistics (mean $[0.485, 0.456, 0.406]$, standard deviation $[0.229, 0.224, 0.225]$) to align with the pretraining distribution of the network. Frames were then converted to PyTorch tensors of shape $(3, 112, 112)$, ready for batching into fixed-length video clips in subsequent experiments.

D. I3D Feature Extraction (Experiments 1a–1c)

We use an Inflated 3D ConvNet based on the R3D-18 architecture as a region-specific spatiotemporal feature extractor for chin, left-eye and right-eye clips [13], [14]. The network is initialised from Kinetics-400 pretraining and fine-tuned separately for each region on the EquiFACS pain dataset.

1) Architecture: The R3D-18 backbone follows a ResNet-18-style design with 3D convolutions. Input clips have shape $(B, 3, 16, 112, 112)$ (batch, channels, frames, height, width). A $7 \times 7 \times 7$ stem convolution with batch normalisation, ReLU and 3D max pooling is followed by four residual stages with channel sizes $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$, each containing two $3 \times 3 \times 3$ residual blocks. Global average pooling over space and time produces a 512-dimensional feature vector per clip, which is passed to a new fully connected layer with two outputs for mild vs. moderate pain. The original 400-way Kinetics classifier is discarded.

2) Clip generation: For each ROI (chin, left eye, right eye), the 50 fps sequences are segmented into non-overlapping 16-frame clips using a temporal stride of 16. This avoids clip overlap and ensures that each clip is treated independently during training. All clips from a given video inherit the video-level pain label, assuming that pain intensity is stable over the 20–40 s recording window. Across the 12 subjects this yields approximately 2.6k chin clips, 2.4k left-eye clips and 2.5k right-eye clips.

TABLE III: Hyperparameters for region-specific I3D (R3D-18) models.

Setting	Value
Backbone	R3D-18, Kinetics-400 pretrained
Input clip shape	$3 \times 16 \times 112 \times 112$
Regions	Chin, left eye, right eye (separate models)
Optimizer	AdamW, lr 1×10^{-4}, wd 0.01
Loss	Class-weighted CE ($w_{\text{mild}} = 2.0$, $w_{\text{mod}} = 1.0$)
Batching	Batch size 8, grad. accumulation ×4 (eff. 32)
Augmentations	Random flip (0.5), colour jitter, random temporal crop
Scheduler	ReduceLROnPlateau (factor 0.5, patience 5, min lr 1×10^{-6})
Stopping	Early stopping after 10 epochs without macro-F1 gain

3) Training procedure: Three independent I3D models are trained, one per region (Experiments 1a–1c), with no parameter sharing between regions. Optimisation uses AdamW with learning rate 1×10^{-4}, weight decay 0.01 and default betas. To address the 4:8 mild-to-moderate imbalance, a class-weighted cross-entropy loss is used with weights $w_{\text{mild}} = 2.0$ and $w_{\text{mod}} = 1.0$, penalising misclassification of mild clips more strongly. The physical batch size is 8 clips; gradients are accumulated for 4 steps to obtain an effective batch size of 32.

During training, clips are augmented with random horizontal flips ($p = 0.5$) and colour jitter (brightness, contrast and saturation ± 0.2, hue ± 0.1); when longer sequences are available, the 16-frame window is randomly cropped in time. Mixed-precision training (FP16 with FP32 master weights) is used to reduce memory consumption and speed up training. A ReduceLROnPlateau scheduler monitors validation macro-F1 and halves the learning rate after 5 epochs without improvement (minimum 1×10^{-6}); early stopping is triggered after 10 stagnant epochs.

E. Region-wise I3D Baselines (Experiments 1 and 2)

Separate I3D (R3D-18) models were trained on chin, left-eye and right-eye clips using 5-fold subject-wise cross-validation on the 12 EquiFACS videos [7]. Each fold withheld two subjects for validation while the remaining ten were used for training, and macro-F1 across mild and moderate classes was recorded for each region. Across folds, the mean macro-F1 over regions was 0.494 ± 0.067, with chin generally outperforming the eye regions.

TABLE IV: Region-wise I3D baselines. Validation scores are macro-F1 on validation clips; test accuracy is on the three held-out test videos (S2, S4, S7).

Region	Best Val Macro-F1	Test Acc. (%)
Chin	0.801	66.7
Left eye	0.541	33.3
Right eye	0.715	33.3

After cross-validation, final region models were retrained on the combined train+validation subjects and evaluated on the three held-out test videos (S2: Mild; S4, S7: Moderate). Table IV reports the best validation macro-F1 for each region together with the corresponding video-level test accuracy. The chin model achieved the highest validation macro-F1 (0.801) and correctly classified both moderate test subjects, whereas the eye models reached lower macro-F1 and did not improve test accuracy beyond the chin-only baseline. Consequently, all subsequent temporal experiments (Bi-LSTM and joint I3D+Bi-LSTM) were conducted on chin features only.

F. Bi-LSTM Temporal Modeling (Experiment 3)

Since the chin-only I3D model achieved the highest macro-F1 and was less affected by occlusions than the eye models, all temporal experiments (Bi-LSTM and joint I3D+Bi-LSTM) were conducted on chin features only. While the chin I3D model captures spatiotemporal patterns within 16-frame clips, it treats clips independently. To model pain trajectories over full recordings, Experiment 3 adds a Bi-LSTM with attention on top of frozen chin I3D features.

1) Chin I3D feature sequences: For Experiment 3 we first freeze the best chin I3D checkpoint from Experiment 1a and use it as a feature extractor. The final classification layer is removed so that each 16-frame chin clip is mapped to a 512-dimensional vector. For each video, all clips are passed through the frozen network in evaluation mode, and their 512-D outputs are stacked into a variable-length sequence of size $(T_{\text{clips}}, 512)$, where T_{clips} depends on video duration. These sequences are saved as `.npy` files (one per subject) and form the input to the temporal model.

2) Bi-LSTM with temporal attention: The temporal backbone is a bidirectional LSTM operating on the sequence of chin features. Padded sequences of shape $(B, T_{\max}, 512)$ are fed to a 2-layer Bi-LSTM with hidden size 256 per direction (512 combined), dropout 0.3 between layers, and bidirectional outputs of shape $(B, T_{\max}, 512)$. On top of the LSTM output, a temporal attention module computes a scalar weight for each time step via a small MLP (Linear(512→256)–Tanh–Linear(256→1)) followed by a masked softmax along the time axis; padded positions are masked before normalisation. The attention weights are used to obtain a context vector as a weighted sum over time, yielding a single 512-D representation per video.

This context vector is passed through a three-layer MLP classifier: 512→256→128→2 with ReLU activations and dropout 0.3 after the first two layers. The output logits are used with the same class-weighted cross-entropy loss as for I3D (weight 2.0 for mild, 1.0 for moderate). Overall the temporal module has approximately 2.3 M parameters, substantially smaller than the I3D backbone.

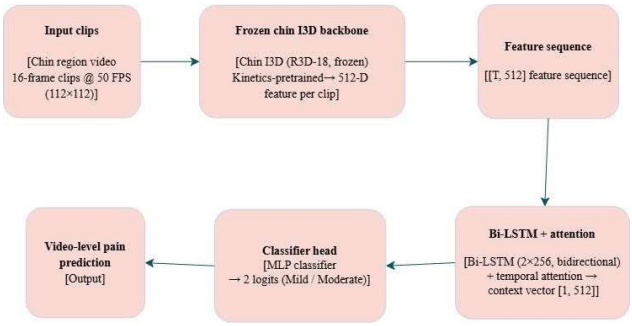

Fig. 6: Frozen chin I3D + Bi-LSTM architecture used in Experiment 3. Chin-region clips are encoded by a frozen I3D backbone into 512-D features, which are fed to a Bi-LSTM with temporal attention and an MLP classifier to produce video-level pain predictions.

3) Training setup and main result: Bi-LSTM models are trained on frozen chin features using AdamW with learning rate 1×10^{-3}, weight decay 0.01, batch size 4 (videos), and the same subject-wise train/val/test split as in earlier experiments (S1, S3, S5, S6, S8, S9, S10 for training; S11, S12 for validation; S2, S4, S7 for testing) [7]. A ReduceLROnPlateau scheduler halves the learning rate after 5 epochs without improvement in validation macro-F1, and early stopping is applied with patience 10. The best validation model (Bi-LSTM with attention, 2 layers, hidden size 256, dropout 0.3) achieves a validation macro-F1 of 0.49 and test macro-F1 of 0.36 with test accuracy 47.8 %, indicating that temporal aggregation alone does not close the generalisation gap on unseen subjects.[2]

4) Ablation study: To understand the effect of attention depth and capacity, several Bi-LSTM variants were trained on the same split. Table V summarises the main configurations

[2]Values obtained from the Experiment 3 summary logs.

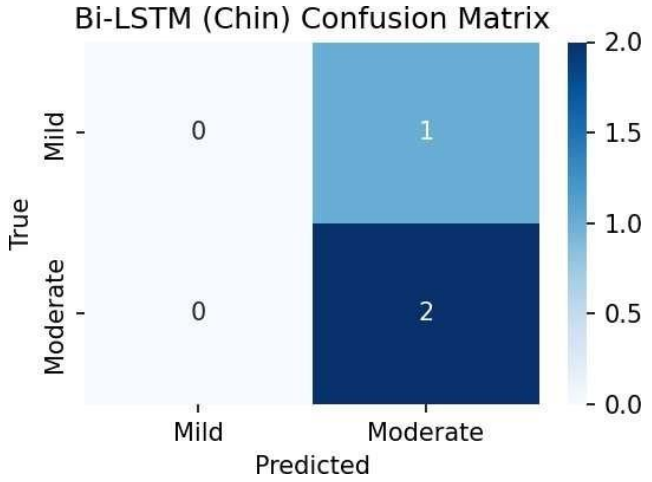

Fig. 7: Video-level confusion matrix for the chin I3D + Bi-LSTM model on the test set (S2, S4, S7). The model correctly identifies the moderate cases but remains uncertain about the single mild case.

TABLE V: Bi-LSTM chin ablation variants (Experiment 3). All models use class-weighted loss and frozen chin I3D features.

Name	Attn	Layers	Hidden	Drop.	Test Acc.
base_attn_L2_H256	Yes	2	256	0.3	66.7 %
no_attn_L2_H256	No	2	256	0.3	33.3 %
attn_L1_H256	Yes	1	256	0.3	66.7 %
attn_L2_H128	Yes	2	128	0.3	66.7 %
attn_L2_H256_do02	Yes	2	256	0.2	66.7 %

and test results. Across all ablations, validation macro-F1 remained around 0.33, while test accuracy ranged from 33.3 % to 66.7 %, reflecting high variance due to the very small test set. Notably, removing attention reduced test accuracy to 33.3 %, whereas attention-based models with either one layer (256 hidden units), two layers with smaller hidden size (128), or reduced dropout (0.2) all matched the best accuracy of 66.7 %, suggesting that attention is beneficial but model capacity beyond 128–256 hidden units offers little advantage on this dataset.

G. End-to-End Joint I3D+Bi-LSTM (Experiment 4)

Experiment 3 used frozen chin I3D features extracted offline. In Experiment 4 we instead train the entire chin I3D + Bi-LSTM pipeline end-to-end, allowing the R3D-18 backbone to adapt from generic Kinetics pretraining to the equine pain domain.

1) Joint architecture: The joint model receives multiple 16-frame chin clips per video of shape $(B, N_{\text{clips}}, 3, 16, 112, 112)$. Each clip is passed independently through the trainable R3D-18 I3D backbone, which outputs a 512-dimensional feature vector per clip. Clip descriptors are stacked into a temporal sequence of shape $(B, N_{\text{clips}}, 512)$ and fed to the same 2-layer bidirectional LSTM with hidden size 256 per direction and temporal attention as in Experiment 3. Attention pooling produces a single 512-D context vector per video, which is classified by a three-layer MLP (512→256→128→2) into mild vs. moderate logits.

TABLE VI: Joint chin I3D + Bi-LSTM performance (Experiment 4, class-weighted cross-entropy). Clip-level metrics are computed over all test clips; video-level metrics are computed on S2 (Mild), S4 and S7 (Moderate).

Level / Metric	Accuracy (%)	Macro-F1
Clip-level	95.7	0.49
Video-level	66.7	0.36

2) Training configuration: End-to-end optimisation uses AdamW with a reduced learning rate of 1×10^{-4} to avoid catastrophic forgetting of Kinetics features, weight decay 0.02, and batch size 2 videos due to the memory cost of backpropagating through the 3D backbone. Gradient clipping with max-norm 1.0 is applied to stabilise training. The same subject-wise split as in Experiment 3 is used (S1, S3, S5, S6, S8, S9, S10 for training; S11, S12 for validation; S2, S4, S7 for testing), and models are trained for up to 50–100 epochs with ReduceLROnPlateau on validation macro-F1 and early stopping after 10 stagnant epochs.

Two loss functions are considered. The first is the class-weighted cross-entropy used in earlier experiments with weights $w_{\text{mild}} = 2.0$ and $w_{\text{mod}} = 1.0$. The second replaces this with focal loss

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t),$$

using $\gamma = 2.0$, $\alpha_{\text{mild}} = 0.67$ and $\alpha_{\text{mod}} = 0.33$ to focus learning on the under-represented mild class.

3) Clip- and video-level evaluation: Because the model processes all clips in a video jointly, it naturally supports both clip- and video-level evaluation. For clip-level metrics, each 16-frame clip in the test set is treated as an individual sample and accuracy and macro-F1 are computed over all clips. For video-level metrics, logits are averaged across all clips from the same subject to obtain a single prediction per video; accuracy and class-wise F1 scores are then computed on the three held-out horses (S2: Mild; S4, S7: Moderate).

Preliminary experiments with focal loss used the same architecture and optimisation settings but replaced the weighted cross-entropy with focal loss (γ=2.0, α_{mild}=0.67, α_{mod}=0.33). However, focal loss did not improve mild-class F1 or video-level accuracy compared to class-weighted cross-entropy: S2 remained misclassified as moderate and the overall test accuracy stayed at 66.7 %. These observations suggest that, under the present data constraints, reweighting the loss cannot compensate for the lack of mild examples at the subject level.

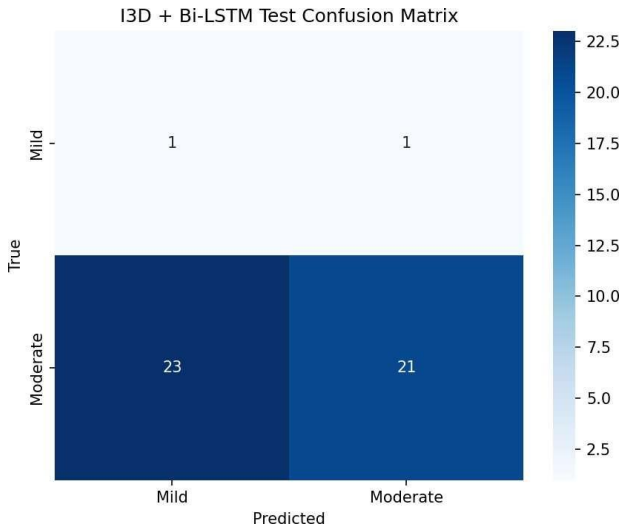

Fig. 8: Video-level confusion matrix for the joint chin I3D + Bi-LSTM model on the held-out test subjects (S2: Mild; S4, S7: Moderate). The model predicts both moderate horses correctly but fails on the mild case.

III. Experimental Setup

A. Cross-Validation Strategy

For region-wise I3D baselines (Experiments 1a–1c) we employed 5-fold subject-wise cross-validation on the 12 EquiFACS videos (S1–S12) [7]. In each fold, two subjects were held out for validation and the remaining ten were used for training, ensuring that no horse appeared in both train and validation sets within the same fold. The folds were stratified so that each contained one mild and one moderate subject in the validation partition, reflecting the global 4:8 class ratio.

For experiments involving temporal modelling (Experiments 3 and 4) and multi-region fusion (Experiment 2), a fixed subject-independent split was used to simplify comparison across architectures. Subjects S1, S3, S5, S6, S8, S9 and S10 were used for training, S11 and S12 for validation, and S2 (Mild) together with S4 and S7 (Moderate) served as the held-out test set following Rashid *et al.* [7]. Table VII summarises the fold compositions and final train/validation/test allocation.

B. Evaluation Metrics

Models are evaluated using accuracy, precision, recall and macro-F1 score over the two pain classes. Accuracy measures the fraction of correctly classified clips or videos, while macro-F1 averages per-class F1 scores and is therefore insensitive to the mild/moderate class imbalance. For each method we report macro-F1 as the primary metric, accompanied by per-class F1 scores to highlight performance on mild vs. moderate pain.

For region-wise I3D baselines, metrics are computed at the clip level and averaged across the 5 folds. For temporal and joint models, both clip-level and video-level metrics are considered. At clip level, each 16-frame segment contributes one prediction; at video level, logits are averaged across all clips from the same subject to obtain a single prediction per video. Confusion matrices are reported for ROI detection, Bi-LSTM temporal models and joint training to visualise error patterns and the difficulty of correctly identifying the single mild test subject.

TABLE VII: Data split composition for region-wise cross-validation (Experiments 1a–1c) and the fixed split used in Experiments 2–4.

Fold	Train subjects	Val subjects	Pain labels
1	S2,S3,S4,S5,S7,S8,S9,S10,S11,S12	S1,S6	1 Mild, 1 Mod.
2	S1,S4,S5,S6,S7,S8,S9,S10,S11,S12	S2,S3	1 Mild, 1 Mod.
3	S1,S2,S4,S6,S7,S8,S9,S10,S11,S12	S3,S5	1 Mild, 1 Mod.
4	S1,S2,S3,S5,S6,S8,S9,S10,S11,S12	S4,S7	1 Mild, 1 Mod.
5	S1,S2,S3,S4,S5,S6,S7,S8,S10,S11	S9,S12	1 Mild, 1 Mod.
Fixed split for Experiments 2–4			
Train	S1, S3, S5, S6, S8, S9, S10		
Val	S11, S12		
Test	S2 (Mild), S4, S7 (Moderate)		

TABLE VIII: Summary of main hyperparameters for Experiments 1–4.

Setting	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Backbone	I3D (R3D-18)	I3D	Frozen I3D	Joint I3D
Temporal head	–	Late fusion	Bi-LSTM+attn	Bi-LSTM+attn
Optimiser	AdamW	AdamW	AdamW	AdamW
LR	1×10^{-4}	1×10^{-4}	1×10^{-3}	1×10^{-4}
Weight decay	0.01	0.01	0.01	0.02
Batch size	8 clips	8 clips	4 videos	2 videos
Loss	WCE	WCE	WCE	WCE / Focal
Scheduler	RLROP	RLROP	RLROP	RLROP
Early stopping	Yes	Yes	Yes	Yes

C. Hyperparameter Selection and Tuning

Hyperparameters were selected by grid or coarse random search on the validation subjects, using macro-F1 as the selection criterion. For region-wise I3D models we tuned learning rate, weight decay, batch size and class weights, finding that AdamW with learning rate 1×10^{-4}, weight decay 0.01, batch size 8 (effective 32 with gradient accumulation) and loss weights $(w_{\text{mild}}, w_{\text{mod}}) = (2.0, 1.0)$ provided a good trade-off between stability and performance.

For Bi-LSTM temporal models we explored the number of layers (1–2), hidden size (128, 256), dropout (0.2, 0.3) and the presence or absence of temporal attention. The base configuration (2 layers, hidden size 256, dropout 0.3, attention enabled, learning rate 1×10^{-3}) was selected based on validation macro-F1 and is used as the reference in Experiment 3. End-to-end joint training required a lower learning rate (1×10^{-4}), stronger weight decay (0.02), smaller video batch size (2) and gradient clipping to ensure convergence with the larger parameter count. Table VIII summarises the final hyperparameters for all experiments.

TABLE IX: 5-fold subject-wise cross-validation results for region-wise I3D models (Experiments 1a–1c). Values are mean ± standard deviation across folds.

Region	Acc. (%)	Macro-F1	F1 (Mild / Mod.)
Chin	73.4 ± 5.1	0.57 ± 0.06	0.46 / 0.68
Left eye	68.2 ± 9.4	0.48 ± 0.09	0.35 / 0.61
Right eye	70.1 ± 7.3	0.50 ± 0.07	0.38 / 0.62

D. Implementation Details

All models were implemented in Python using PyTorch 2.x for deep learning and the Ultralytics YOLOv8 framework for ROI detection. Training and inference were performed on a workstation with an NVIDIA RTX 3090 GPU (24 GB VRAM), Intel i9-class CPU and 64 GB RAM. ROI detection and preprocessing for the 12 subjects required approximately 1–2 hours, including RIFE interpolation to 50 fps. Region-wise I3D models trained in 2–3 hours per region, Bi-LSTM models trained in under 30 minutes, and joint I3D+Bi-LSTM models required 3–4 hours due to full backpropagation through the 3D backbone. All experiments were run with fixed random seeds, subject-wise splits and config files stored in version control to facilitate reproducibility.

IV. Results

A. Region-Wise I3D Performance (Experiments 1a–1c)

Table IX reports the 5-fold cross-validation performance of chin, left-eye and right-eye I3D models, averaged over clip-level predictions. The chin model obtains the highest mean macro-F1 and accuracy, while left eye exhibits the largest standard deviation, reflecting its greater susceptibility to occlusion and motion blur. Across all regions, F1 for the moderate class is consistently higher than for mild pain, indicating that the class imbalance and more subtle facial changes in mild cases remain challenging.

Figure 9 illustrates these results as a bar chart with error bars for macro-F1. The chin region clearly outperforms both eye regions and shows more stable performance across folds, supporting its use as the primary ROI for subsequent temporal modelling.

B. Multi-Region Late Fusion (Experiment 2)

Multi-region late fusion averages the predicted probabilities from chin and eye models at the video level. Table X lists video-level metrics on the fixed test split. Although fusion slightly improves robustness in certain occluded frames, it does not outperform the chin-only baseline at video level: test accuracy remains 66.7 % with two correctly classified moderate subjects and the mild subject (S2) still misclassified as moderate.

C. Bi-LSTM Temporal Modeling (Experiment 3)

Table XI summarises the performance of the base Bi-LSTM model and the main ablation variants using frozen chin I3D features. The attention-based Bi-LSTM achieves a

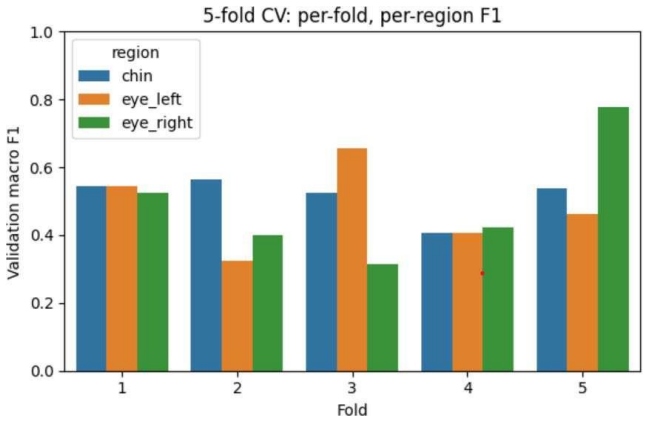

Fig. 9: Mean macro-F1 over 5 folds for chin, left-eye and right-eye I3D models with standard-deviation error bars. Chin provides the most discriminative and stable signal for pain classification.

TABLE X: Video-level performance of region-wise and fused I3D models on the test subjects (S2, S4, S7).

Method	Acc. (%)	F1 Mild	F1 Mod.
I3D (Eyes only)	33.3	0.00	0.50
I3D (Chin only)	66.7	0.00	0.80
Late Fusion (Eyes+Chin)	66.7	0.00	0.80

test accuracy of 66.7 % and macro-F1 of 0.36 at video level, matching the chin-only I3D baseline but not improving the mild-class F1, which remains close to zero.

Attention visualisations show that the network focuses on segments with pronounced lower-face motion, such as chewing or lip pressing, but the scarcity of mild training examples limits its ability to learn distinct temporal signatures for that class.

D. End-to-End Joint Training (Experiment 4)

The joint chin I3D + Bi-LSTM model (Experiment 4) is evaluated with class-weighted cross-entropy and focal loss. As summarised in Table VI, joint training achieves high clip-level performance (95.7 % accuracy, macro-F1 0.49) but only 66.7 % video-level accuracy and macro-F1 0.36 after aggregation, with S2 (Mild) still predicted as moderate. Figure 8 provides the corresponding confusion matrix. Replacing the loss with focal loss (γ=2.0, α_{mild}=0.67, α_{mod}=0.33) did not improve mild-class F1 or video-level accuracy compared to class-weighted cross-entropy, indicating that reweighting the loss cannot compensate for the lack of mild examples at the subject level.

E. Ablation Study Summary

Table XII consolidates the main ablation results across Experiments 1–3, covering ROI selection, temporal attention, Bi-LSTM depth, hidden size and dropout. Figure 10 visualises the relative impact of each component on video-level accuracy. The most influential design choices are focusing on the chin

TABLE XI: Video-level performance of Bi-LSTM variants on chin I3D features (Experiment 3).

Variant	Acc. (%)	Macro-F1	F1 Mild / Mod.
base_attn_L2_H256	66.7	0.36	0.00 / 0.72
no_attn_L2_H256	33.3	0.25	0.00 / 0.50
attn_L1_H256	66.7	0.36	0.00 / 0.72
attn_L2_H128	66.7	0.36	0.00 / 0.72
attn_L2_H256_do02	66.7	0.36	0.00 / 0.72

TABLE XII: Summary of key ablations across experiments. All metrics are video-level accuracy on the fixed test split.

Configuration	Acc. (%)
Eyes only (I3D)	33.3
Chin only (I3D)	66.7
Chin + Eyes late fusion	66.7
Bi-LSTM (no attention)	33.3
Bi-LSTM + attention (base)	66.7
Bi-LSTM + attention (alt. configs)	66.7
Joint I3D + Bi-LSTM (CE)	66.7
Joint I3D + Bi-LSTM (Focal)	66.7

region instead of eyes or multi-region fusion, and using temporal attention instead of simple pooling. Variations in Bi-LSTM capacity, augmentation settings and RIFE interpolation parameters have comparatively smaller effects on the final metrics.

F. Computational Performance

Table XIII compares computational cost across architectures in terms of training time per run, approximate number of parameters and relative inference time per video. While joint training roughly doubles training time compared to frozen-feature Bi-LSTM, inference remains dominated by the I3D backbone, and all variants are feasible for near real-time processing at the small scale of the current dataset.

V. Discussion

A. Key Insights

The experiments confirm that the chin region carries the most reliable pain information among the considered ROIs, both in terms of macro-F1 and stability across folds. This finding aligns with HGS and EquiFACS literature, which highlights lower-face tension, lip raising and chin tightening as robust indicators of equine pain [4]–[6]. Temporal modelling with Bi-LSTM and attention improves the interpretability of the system by highlighting high-motion segments, but does not yet translate into higher video-level accuracy under the current data constraints. Overall, the results suggest that spatially precise chin features are necessary but not sufficient to resolve the subtlety of mild pain expressions.

B. Comparison with Prior Work

Table XIV summarises our results alongside representative prior work on equine pain recognition. Compared to image-based CNN approaches that operate on still frames, the proposed I3D + Bi-LSTM pipeline explicitly models temporal dynamics and uses automated ROI detection instead of manual cropping [3], [8], [9]. However, absolute performance remains limited by dataset size: while our 66.7 % video-level accuracy on unseen subjects is competitive with earlier small-cohort studies, it falls short of the higher accuracies reported on larger, more homogeneous image datasets.

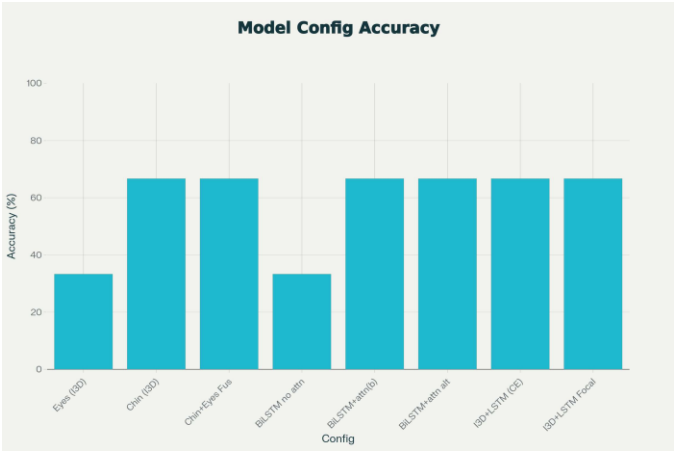

Fig. 10: Effect of key architectural and training choices on video-level test accuracy. Chin ROI and temporal attention have the largest positive impact.

TABLE XIII: Approximate computational cost for the main architectures. Training times are measured on an RTX 3090 GPU.

Model	Params (M)	Train time	Inf. time / video
I3D (single region)	33	2–3 h	≈0.2 s
Chin I3D + Bi-LSTM	35	≈0.5 h	≈0.25 s
Joint I3D + Bi-LSTM	35	3–4 h	≈0.25 s

C. Limitations and Challenges

The main limitation of this study is the small number of subjects and the extreme imbalance between mild and moderate pain at the subject level: only one mild horse is available for testing. This makes it difficult to draw strong statistical conclusions and encourages overfitting to individual appearance cues rather than general pain patterns. The reliance on RIFE interpolation and relatively low original frame rates may also distort very fast facial actions, and class-weighted or focal losses cannot fully compensate for the scarcity of mild examples. Finally, all experiments focus on acute, procedure-related pain; transfer to chronic or low-grade discomfort remains untested.

D. Clinical Implications

Despite these limitations, the results demonstrate that automatic analysis of chin-region facial dynamics can detect moderate equine pain in unseen subjects with reasonable accuracy. In a clinical workflow, such a system could provide

TABLE XIV: Comparison with representative automatic equine pain recognition methods.

Method	Modality	Subjects	Reported Acc.
Lencioni *et al.* [3]	Images (ROIs)	7	73–83 % (frame)
Pessanha *et al.* [8]	Images (face)	39	70–80 % (frame)
Ruhof *et al.* [9]	Images (YOLO ROIs)	39	74 % (frame)
Ours (I3D chin only)	Video (chin)	12	66.7 % (video)
Ours (I3D+Bi-LSTM)	Video (chin)	12	66.7 % (video)

continuous, unobtrusive monitoring in stalls or recovery boxes, flagging intervals of suspected pain for human review rather than replacing expert judgement. The use of interpretable ROIs and attention maps also facilitates qualitative inspection by veterinarians, potentially increasing trust and aiding training of less experienced staff.

E. Future Work

Future work will focus on expanding the dataset with more horses, especially in the mild and no-pain ranges, and on exploring ordinal or regression formulations that better match the graded nature of pain. Self-supervised pretraining on large collections of unlabeled equine video may help reduce reliance on Kinetics and improve feature transferability. Additional directions include multi-view fusion, incorporation of body posture and locomotion cues, real-time optimisation for deployment in clinics, and the development of explainable interfaces that link model decisions to specific EquiFACS action units.

VI. Conclusion

This paper presented a region-aware spatiotemporal pipeline for automated equine pain recognition from facial videos, combining YOLOv8 ROI detection, I3D-based chin feature extraction and Bi-LSTM temporal aggregation. Region-wise experiments showed that the chin region provides the most discriminative signal, and temporal models with attention offered interpretable focus on high-motion segments, although overall video-level accuracy remained 66.7 % on unseen subjects. End-to-end joint training achieved strong clip-level performance but did not overcome the lack of mild examples, highlighting dataset size and intensity imbalance as the primary bottlenecks. Nevertheless, the study demonstrates the feasibility of clinically grounded, ROI-based deep learning for equine pain assessment and outlines concrete steps towards more robust, deployable systems.

References

[1] K. B. Gleerup and C. Lindegaard, "Recognition and quantification of pain in horses: A tutorial review," *Equine Vet. Educ.*, vol. 28, no. 1, pp. 47–57, 2016.

[2] P. H. Andersen *et al.*, "Towards machine recognition of facial expressions of pain in horses," *Animals*, vol. 11, no. 6, p. 1643, 2021.

[3] G. C. Lencioni *et al.*, "Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling," *PLOS ONE*, vol. 16, no. 10, p. e0258672, 2021.

[4] E. Dalla Costa *et al.*, "Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration," *PLOS ONE*, vol. 9, no. 3, p. e92281, 2014.

[5] J. Wathan, A. M. Burrows, B. M. Waller, and K. McComb, "EquiFACS: The equine facial action coding system," *PLOS ONE*, vol. 10, no. 8, p. e0131738, 2015.

[6] K. B. Gleerup, B. Forkman, C. Lindegaard, and P. H. Andersen, "An equine pain face," *Vet. Anaesth. Analg.*, vol. 42, no. 1, pp. 103–114, 2015.

[7] M. Rashid, A. Silventoinen, K. B. Gleerup, and P. H. Andersen, "Equine Facial Action Coding System for determination of pain-related facial responses in videos of horses," *PLOS ONE*, vol. 15, no. 11, p. e0231608, 2020.

[8] F. Pessanha, A. A. Salah, T. J. van Loon, and R. Veltkamp, "Facial image-based automatic assessment of equine pain," *IEEE Trans. Affective Comput.*, vol. 14, no. 3, pp. 2064–2076, 2023.

[9] J. J. Ruhof, A. A. Salah, and T. J. van Loon, "Automatic pain estimation in equine faces: More effective uses for regions of interest," in *Proc. 12th IEEE Int. Conf. Affective Computing and Intelligent Interaction Workshops*, 2024.

[10] S. Broome *et al.*, "Going deeper than tracking: A survey of computer-vision-based recognition of animal pain and emotions," *Int. J. Comput. Vis.*, vol. 131, no. 2, pp. 572–590, 2023.

[11] G. Jocher *et al.*, "Ultralytics YOLOv8: Real-time object detection," GitHub repository, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[12] Heartex, "Label Studio: Open source data labeling tool," 2020. [Online]. Available: https://labelstud.io

[13] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *Proc. CVPR*, 2017.

[14] D. Tran *et al.*, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*, 2018.

[15] Z. Huang *et al.*, "RIFE: Real-time intermediate flow estimation for video frame interpolation," in *Proc. ECCV*, 2020.