Spatiotemporal Modeling of Bodily Emotional Expressions for Continuous Valence-Arousal-Dominance Prediction in Video

Anonymous submission

Abstract

Predicting Valence-Arousal-Dominance (VAD) dimensions from bodily-expressed emotions in videos remains a fundamentally challenging task in affective computing, requiring models that capture subtle spatiotemporal patterns while balancing computational efficiency and interpretability. We present a comprehensive investigation of VAD prediction approaches on the newly introduced Annotated Bodily Expressed Emotion (ABEE) dataset, which contains approximately 3,200 video clips spanning 8 primary emotion categories and 20 subcategories. We explore two complementary methodologies: a feature-based gradient boosting approach using XGBoost with carefully engineered spatiotemporal features and dimensionality reduction, and deep learning architectures capable of learning hierarchical representations directly from raw video data. Our feature-based approach demonstrates exceptional computational efficiency, with subsecond training times and minimal resource requirements, while our deep models reveal the fundamental difficulty of capturing continuous VAD dimensions from bodily expressions. Through systematic evaluation on the ABEE dataset, we establish baseline performance for the VAD prediction task, achieving R^2 scores of -0.090, -0.014, and -0.058 for valence, arousal, and dominance, respectively, with our gradient boosting approach. These results highlight the substantial gap between current methodologies and the inherent complexity of bodily emotion signals, providing benchmarks for future research. We further discuss critical insights regarding feature engineering, temporal dynamics, and the intrinsic challenges of continuous emotion prediction from naturalistic video data, emphasizing the need for dedicated spatiotemporal modeling strategies tailored to bodily expressions.

Keywords: Valence-Arousal-Dominance, Bodily-Expressed Emotion, Affective Computing, Video Emotion Recognition, Spatiotemporal Modeling

Introduction

Understanding and quantifying human emotions through automated systems has emerged as a critical challenge in affective computing, with applications spanning healthcare, human–computer interaction, education, and mental health assessment. While facial expressions have dominated emotion recognition research (Koolagudi and Rao 2012), recent studies highlight that bodily expressions convey equally rich emotional information, often more accessible in real-world scenarios where faces may be occluded, distant, or partially visible (Bolles and Cain 1982). This motivates a paradigm shift toward full-body affect understanding, particularly in ecological settings where facial cues are unreliable. Moreover, bodily affect interpretation is central for socially competent robots, embodied agents, and physical intelligence systems, aligning closely with the vision of the BEEU challenge.

The Valence–Arousal–Dominance (VAD) model provides a dimensional framework for representing emotions as continuous values across three fundamental axes (Barrett et al. 2001). Let $(v, a, d) \in \mathbb{R}^3$ denote continuous affect coordinates capturing pleasantness, activation, and control. Unlike categorical labels $\mathbf{y} \in \{1, \ldots, K\}$, VAD enables modeling smooth affect transitions, crucial for naturalistic emotion expression. VAD modeling from the body therefore serves as a principled bridge toward fine-grained affect dynamics, moving beyond discrete action units or emotion classes.

However, predicting VAD dimensions from bodily video data presents significant challenges due to (1) reliance on subtle and dynamic cues, (2) the need to model long-range temporal structure, (3) ambiguity in continuous affect perception, and (4) high cross-subject variability driven by cultural, biomechanical, and personality factors. These factors render bodily VAD prediction substantially more complex than facial affect, where smoother spatial priors and larger corpora exist. Addressing these gaps requires rigorous baselines and diagnostic analysis to guide future model design.

The BEEU Challenge 2025 introduces the ABEE dataset containing 3,200 clips with 8 emotion categories, 20 subcategories, and VAD scores from 1–9. To our knowledge, ABEE is the first public benchmark for bodily VAD regression, creating a unique opportunity to characterize the limits of body-only affect modeling. We position this work as an early, systematic baseline effort to ground future research on this task.

Our Approach. We study two complementary pipelines for VAD prediction from bodily movement. The first employs a lightweight XGBoost regressor using 882 engineered spatiotemporal features reduced to 194 dimensions via PCA, providing an interpretable and low-latency baseline suitable for edge cognition. The second uses a 13.9M-parameter 3D-CNN adapted from (Alakwaa, Nassef, and Badr 2017), trained in a multi-task setting to jointly predict

28 discrete emotion labels and continuous VAD values:

$$\mathcal{L} = \lambda_c \mathcal{L}_{\text{CE}} + \lambda_r \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2,$$

where $\mathbf{z} = (v, a, d)$ and $(\lambda_c, \lambda_r) = (1.0, 0.5)$. This design encourages shared affective representations while mitigating VAD sparsity. Together, these models span the classical–deep spectrum, offering insight into representational biases and data efficiency trade-offs.

Our Contributions.

- Among the first systematic baselines contrasting classical and deep architectures for bodily VAD prediction on ABEE.
- Foundational benchmarks highlighting the difficulty of body-only continuous affect regression.
- Multi-task 3D-CNN formulation for joint discrete and continuous bodily emotion prediction.
- In-depth failure and efficiency analysis providing actionable insights for future architectures, including transformer-based and pose-aware models.

Our results show that body-only VAD estimation remains an open frontier, with lightweight models offering real-time deployment but limited accuracy, and 3D-CNNs showing representational advantage yet struggling with multi-objective optimization. These insights establish core challenges and motivate future work in pose-aware transformers, temporal attention, and large-scale embodied affect pretraining. By surfacing bottlenecks and design signals rather than focusing solely on scores, we aim to catalyze progress in embodied emotion reasoning.

Related Work

Emotion Recognition from Videos. The evolution of video-based emotion recognition has progressed from geometric and appearance descriptors to deep spatiotemporal models. Early works relied on facial landmark tracking (Wu and Ji 2019) and texture operators such as LBP-TOP (Almaev and Valstar 2013), which provided interpretable cues but struggled to capture long-range emotional dynamics. With the advent of deep learning, recurrent architectures (Sherstinsky 2020) and CNN-LSTM pipelines (Ullah et al. 2017) enabled temporal modeling, though they process spatial and temporal cues separately, limiting joint feature learning.

The introduction of 3D CNNs and transformers further improved temporal representation learning. Recent works leverage large-scale video transformers and motion-aware backbones for affect modeling, yet most focus on facial or audiovisual modalities rather than isolated body signals. Despite these advances, several gaps remain: (1) bodily expressions are underexplored relative to facial cues, (2) deep models require large annotated datasets that are scarce for body affect, (3) interpretability challenges persist, and (4) computational demands hinder edge deployment (Sherstinsky 2020). To our knowledge, no prior work systematically benchmarks lightweight and deep spatiotemporal pipelines specifically for continuous VAD estimation from bodily expressions.

Bodily Expression of Emotions. Body language plays a critical role in affect communication, complementing or even overriding facial signals (De Gelder 2009). Ekman's work (Ekman 2006) emphasized facial universality, yet later studies showed rich affect conveyed through body posture and kinematics (Dael, Mortillaro, and Scherer 2012). Biomechanics-informed affect models and pose-based representations further confirm that motion patterns encode fine-grained affective cues.

However, computational research remains limited due to dataset scarcity and annotation difficulty (Kleinsmith and Bianchi-Berthouze 2012). Recent datasets (BEEU/ABEE) address this but remain relatively small, motivating efficient learning strategies and robust modeling under limited supervision. Our work uses ABEE as a benchmark to deepen understanding of bodily emotion cues, providing the first results for continuous VAD prediction under this challenge setting.

Dimensional Emotion Representations. Dimensional affect models represent emotions continuously along valence, arousal, and dominance axes (Russell 1980; Mehrabian 1996). These models capture nuanced affect better than categorical labels and align with psychophysiology. However, VAD regression from body movements remains largely unexplored, with most prior work focusing on facial or multimodal settings. This gap motivates computational pathways beyond facial affect toward full-body understanding in naturalistic settings.

Feature-Based Machine Learning for Temporal Data. Traditional models such as SVMs (Hearst et al. 1998) and Random Forests (Breiman 2001) remain attractive for explainability and efficiency. Gradient boosting, particularly XGBoost (Bentéjac, Csörgő, and Martínez-Muñoz 2021), excels in regression via sequential tree optimization with regularization. Yet, handcrafted features struggle to represent subtle temporal affect cues, and dimensionality reduction (e.g., PCA) is often required to mitigate overfitting. Our feature-based XGBoost model establishes a transparent and low-latency baseline for bodily VAD, supporting deployment on resource-limited embodied systems.

Multi-Task Learning in Affective Computing. Joint modeling of emotion categories and VAD dimensions improves representation sharing (Zhang et al. 2018). Recent multi-task affect systems report gains through shared encoders and attention fusion, but balancing heterogeneous losses remains challenging, often requiring careful weighting and curriculum strategies. Improper balancing can cause task interference and degraded performance. We employ a balanced multi-task objective tailored to limited data in bodily affect settings, demonstrating its impact under ABEE challenge constraints.

Challenges in Bodily Emotion Recognition. Body-based affect modeling is challenging due to complex articulated motion, occlusions, clothing variation, and cultural differences (Pantic and Rothkrantz 2002). Temporal dynamics also unfold over longer horizons than facial micro-expressions, demanding richer motion encodings. Dataset size remains a barrier: while facial corpora like CK+ (Lucey et al. 2010) are large, body emotion datasets are limited in

scale. Our results underscore these bottlenecks and highlight the necessity for scalable pose-aware temporal architectures and future multimodal augmentation.

Computational Efficiency and Deployment. Real-world affect systems often run on mobile and edge devices with strict latency constraints. Lightweight architectures (MobileNets (Howard et al. 2017), EfficientNets (Tan and Le 2019)) make progress, yet 3D CNNs and video transformers remain expensive for sustained real-time inference. Feature-based models offer microsecond-level inference but depend on effective feature extraction. By contrasting efficient handcrafted representations with deep spatiotemporal learning, we provide actionable insights for embodied AI systems requiring socially aware perception and real-time reasoning.

Our work addresses these gaps by systematically comparing a lightweight feature-based model and a 3D CNN for VAD prediction from bodily expressions, highlighting trade-offs between efficiency, representation capacity, and continuous affect estimation in this underexplored modality. We position this evaluation as a foundation for future large-scale bodily affect benchmarks and socially responsive embodied intelligence.

Method

We address the task of predicting continuous affective dimensions, Valence (v), Arousal (a), and Dominance (d), collectively denoted as the VAD triplet $\mathbf{y} = (v, a, d) \in \mathbb{R}^3$, from videos containing bodily expressions of emotions. Unlike facial-expression–centric affect research, this work focuses exclusively on body cues. The VAD framework provides a fine-grained continuous representation of affective states, where valence quantifies emotional pleasantness, arousal measures physiological activation, and dominance reflects sense of control.

Let $X \in \mathbb{R}^{T\times H\times W\times C}$ represent a video tensor comprising T frames, each of spatial dimensions $H \times W$ with C color channels. For our data, $T = 10$, $H = W = 112$, and $C = 3$ (RGB). The goal is to learn a mapping

$$f_\theta : X \mapsto \hat{\mathbf{y}}, \quad \hat{\mathbf{y}} \in [1, 9]^3, \tag{1}$$

where θ denotes model parameters and the range $[1, 9]$ corresponds to human-annotated VAD scales. We pursue two complementary modeling paradigms: (i) feature–engineered gradient boosting and (ii) end-to-end spatiotemporal deep learning.

This is the first study to systematically benchmark continuous bodily affect prediction on the ABEE dataset using both classical machine-learning and neural video encoders. We highlight (i) a rigorous multi-branch feature pipeline capturing pose-independent temporal energy, (ii) a PCA–regularized XGBoost baseline providing interpretable benchmarks, and (iii) an adapted 3D-CNN multi-task architecture jointly predicting categorical emotions and continuous VAD targets, with principled multi-objective loss coupling.

Preprocessing and Data Structure

Given each raw video V, we uniformly extract $T = 10$ frames, indexable as $\{F_t\}_{t=1}^{10}$. Each frame is resized to 112×112 and intensity normalized to $[0, 1]$. Missing frames (e.g., shorter clips) are padded by repeating the last observed frame to preserve tensor shape. The resulting input tensor is arranged as:

$$X = \text{reshape}(F_1, \ldots, F_{10}) \in \mathbb{R}^{B\times C\times T\times H\times W}, \tag{2}$$

where B denotes batch size. Histogram equalization via CLAHE improves local contrast to enhance motion salient regions and bodily posture contours. A train–validation split of 85-15% is formed via a deterministic random seed (42), ensuring reproducibility across experiments.

Feature–Engineered VAD Regression via XGBoost

We engineer a total of 882 handcrafted spatiotemporal features to represent bodily affect dynamics, designed to capture multi-scale statistics of color, motion, and spatial structure. As summarized in Table 1, these features include per-frame statistical moments, temporal motion descriptors, color histogram encodings in both RGB and HSV domains, spatially pooled statistics over a 3×3 grid, and global structural descriptors. This diverse feature set enables lightweight yet expressive modeling of bodily emotion cues while supporting interpretability and efficient deployment.

Let $\mathbf{z} \in \mathbb{R}^{882}$ denote the feature vector for one sample. To avoid information leakage, principal components are learned only on training data, yielding a reduced vector $\mathbf{z}' \in \mathbb{R}^{194}$ that preserves 95% variance. A three-head XGBoost model learns functions $\{g_v, g_a, g_d\}$, optimized with squared-loss objective

$$\mathcal{L}_{\text{XGB}} = \sum_{k\in\{v,a,d\}} \sum_{i=1}^{n} \left(y_{k,i} - g_k(\mathbf{z}'_i)\right)^2, \tag{3}$$

with $n\_\text{estimators} = 1200$, depth $= 6$, learning rate $= 0.02$, and early stopping patience $= 40$. This path yields a lightweight, interpretable, deployment-friendly baseline.

End-to-End 3D-CNN for Joint VAD and Emotion Learning

To directly exploit motion and posture cues, we adapt a spatiotemporal 3D convolutional encoder. Given input X, hierarchical filters learn volumetric kernels over space–time. Four sequential blocks apply 3D convolutions, batch normalization, ReLU, and average pooling. Let $\phi(X) \in \mathbb{R}^{512}$ denote the encoder embedding after spatiotemporal pooling. Two feedforward layers refine representations, followed by task–specific heads:

$$\hat{\mathbf{y}} = W_{\text{VAD}}\psi(\phi(X)) + \mathbf{b}_{\text{VAD}}, \tag{4}$$

$$\hat{\mathbf{c}} = \sigma(W_{\text{CLS}}\psi(\phi(X)) + \mathbf{b}_{\text{CLS}}), \tag{5}$$

where $\hat{\mathbf{c}} \in [0, 1]^{28}$ are multi-label emotion logits and $\psi(\cdot)$ is the MLP transformation. VAD predictions are restricted to

Table 1: Handcrafted feature taxonomy for bodily emotion representation.

Feature Category	Count	Description
Per-frame statistics	150	Mean, variance, median, extrema across spatial axes and channels
Temporal descriptors	180	Frame-to-frame gradients, temporal derivative energy
Color histograms	96	RGB and HSV histograms with 16 bins per channel
Spatial grid pooling	324	Localized summary statistics over 3×3 spatial grid
Global descriptors	132	Kurtosis, skewness, total variation, entropy-like aggregates
Total	**882**	Comprehensive multi-scale spatiotemporal representation

$[1, 9]$ via affine sigmoid. The total loss couples binary cross-entropy and mean squared error:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{BCE}}(\hat{\mathbf{c}}, \mathbf{c}) + \lambda_2 \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2, \tag{6}$$

with $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$, where $\mathbf{c}$ denotes ground-truth multi-label vector. Adam optimizer uses learning rate $\alpha = 10^{-4}$, batch size $B = 30$, maximum epochs 12, early stopping patience 3. Hyperparameters derive from prior optimized facial-expression training but are adapted for bodily motion, aligning temporal pooling and spatial kernel scale with full-body kinematic dynamics rather than facial micro-motion.

Experimental Setup

Dataset

The experiments utilize the Annotated Bodily Expressed Emotion (ABEE) dataset from the BEEU Challenge 2025, which contains approximately 3,200 short video clips of spontaneous bodily expressions. Each clip is annotated with multi-label categorical emotions (eight coarse and twenty fine-grained classes) and continuous Valence, Arousal, and Dominance (VAD) ratings on a $[1, 9]$ scale. Neutral labels denote baseline postures without affective display. We follow the official split: 2,462 clips for training and 1,076 clips for testing. To enable validation, 15% of the training set (370 clips) is held out using a fixed seed (42). This benchmark constitutes one of the first systematic studies jointly modeling continuous bodily affect and multi-label emotion categories on ABEE, ensuring full alignment with the BEEU challenge goals.

Video Representation and Preprocessing

Ten uniformly spaced frames are extracted per clip to capture body motion dynamics. Each frame is resized to 112×112 (RGB), normalized to $[0, 1]$, and Contrast Limited Adaptive Histogram Equalization (CLAHE; clip limit 2.0, 8×8 grid) is applied to emphasize subtle posture changes. Short clips are padded by repeating the final frame. The resulting tensor is (B, C, T, H, W) with $C = 3$, $T = 10$, $H = W = 112$. This lightweight sampling strategy respects the compute constraints of the challenge while preserving salient kinematic cues.

Hardware and Software Environment

All experiments run on NVIDIA Tesla P100 GPUs (16GB) in a Kaggle Python 3.11 + CUDA 12.4 environment. We use PyTorch, NumPy, Pandas, OpenCV, scikit-learn, XGBoost, and Optuna. A unified random seed (42) ensures reproducibility. We release preprocessing and training scripts upon publication to promote transparency and open research.

Modeling Approaches

Below, we summarize modeling choices complementary to Method Section.

Feature-driven XGBoost. From each 10-frame clip, 882 handcrafted spatiotemporal features are extracted (statistics, color histograms, temporal energy, spatial grids). PCA reduces features to 194 components (95% variance), followed by three independent XGBoost regressors (V/A/D). Key settings: 1,200 trees, depth 6, learning rate 0.02, subsampling 0.85, column sampling 0.9, L2=1.0, early stopping (40 rounds), 5-fold cross-validation. This interpretable baseline supports practical deployment and controlled benchmarking.

3D-CNN. A spatiotemporal convolutional encoder (Conv–BN–ReLU–Pooling blocks) followed by adaptive average pooling and two MLP layers jointly predicts emotion labels and VAD scores. Sigmoid scaling maps VAD outputs to $[1, 9]$. Training uses Adam (10^{-4}), batch size 30, 12 epochs, early stopping (3). Loss combines BCE and MSE with weights $\lambda_1 = 1.0$, $\lambda_2 = 0.5$. The architecture emphasizes full-body kinematics over facial micro-expressions, directly addressing BEEU's core theme.

Evaluation Metrics

VAD performance is evaluated with R^2, RMSE, and MAE. For multi-label emotion recognition, we report IoU, precision, recall, and F1. We additionally report inference latency and model size to support real-world embodied-AI deployment considerations.

Results

VAD Prediction Performance

Gradient Boosting Results. Table 2 presents the VAD regression performance of our XGBoost-based gradient boosting model on the validation set. The results demonstrate the intrinsic difficulty of predicting continuous emotional dimensions from bodily expressions. Consistent with the goals of the BEEU challenge, these benchmarks establish the first empirical baseline for continuous bodily VAD prediction

Table 2: XGBoost VAD Prediction Results on Validation Set (370 samples)

Dimension	R^2	RMSE	MAE
Valence	-0.0904	0.4463	0.3521
Arousal	-0.0145	0.3879	0.3102
Dominance	-0.0576	0.3841	0.3087
Average	**-0.0542**	**0.4061**	**0.3237**

Table 3: Classification Threshold Analysis on Validation Set

Threshold	Mean IoU	Avg Labels	Zero Predictions
0.50	0.0000	0.00	369/370
0.40	0.0000	0.00	370/370
0.30	0.0050	0.03	359/370
0.25	0.1764	1.83	98/370
0.20	**0.2301**	3.86	0/370
0.15	0.2135	5.95	0/370
0.10	0.1855	7.62	0/370

and highlight the non-trivial nature of learning affect from body-only signals.

All three VAD dimensions yield negative R^2 values, indicating worse-than-mean prediction. Cross-validation produced similar behavior (R^2: 0.0077 valence, 0.0086 arousal, 0.0019 dominance), suggesting that the model is signal-limited rather than overfitting. RMSE values (0.38–0.45 on a 1–9 scale) outperform random-guess RMSE (≈ 2.3), but remain insufficient for meaningful affect estimation. Thus, this model provides a computationally efficient lower-bound reference for future approaches.

3D-CNN Results. The 3D convolutional neural network exhibits similar regression difficulty, emphasizing the complexity of bodily emotion learning even with deep spatiotemporal features. These results validate that direct transfer of video emotion architectures designed for faces does not trivially extend to full-body affect.

The classification head initially produced extremely low-confidence predictions (IoU=0.0000 at threshold 0.5). Threshold sweep revealed optimal IoU=0.2301 at 0.2 (Table 3), suggesting underfitting and loss imbalance from $\lambda_1 : \lambda_2 = 1.0 : 0.5$. This informs future avenues such as uncertainty-based loss weighting and curriculum design.

Computational Analysis

Table 4 compares resource requirements. XGBoost trains $\sim$32,400$\times$ faster and infers 50$\times$ faster than the 3D-CNN with negligible GPU cost. This contrast highlights that early-stage bodily affect models can benefit from scalable classical baselines before moving to large models. However, neither approach achieves practical VAD performance, indicating that compute alone does not resolve the representational challenge.

Both approaches show prediction range collapse, implying missing pose dynamics and insufficient temporal abstraction. Thus, body emotion learning demands architectures that explicitly model expressive movement primitives, temporal rhythm, and joint-level kinematics.

Table 4: Computational Efficiency Comparison

Metric	XGBoost	3D-CNN
Training Time	$<$1 seconds	540 minutes
Inference (per sample)	$<$1 ms	50 ms
Model Size	3.2 MB	53.1 MB
GPU Memory Required	-	12.5 GB
Parameters	194k features	13.9M params

Failure Analysis

Handcrafted statistics fail to capture subtle shifts in posture/-motor dynamics; 10-frame clips are insufficient for longer gestures; and multi-task gradients interfere. This section provides actionable guidance for the field by pinpointing failure modes specific to bodily emotion tasks.

In contrast to facial models achieving 97.56% accuracy on CK+, our results ($R^2 \approx -0.05$, IoU=0.23) show a two-orders-of-magnitude gap, validating BEEU's focus on body cues and underscoring need for pose-aware attention, larger sequences, and pretraining beyond VAD labels.

Limitations. Limitations include: 10-frame window, small validation set, handcrafted features, and basic multi-task balancing. Future work will explore pose-conditioned temporal transformers, contrastive pretraining, and wider contextual cues (scene, interaction) to advance body-only affect prediction. Despite these constraints, our study offers the first comprehensive baseline and diagnostic report for bodily VAD regression, serving as a benchmark for subsequent BEEU entries.

Conclusion

In this work, we present the first systematic investigation of continuous Valence–Arousal–Dominance (VAD) prediction from bodily-expressed emotions using the newly introduced ABEE dataset. We evaluate two complementary paradigms: a feature-based gradient boosting approach using 882 engineered spatiotemporal descriptors (PCA-reduced to 194), and a deep 3D-CNN model with 13.9M parameters adapted from state-of-the-art facial affect architectures. Our analysis reveals fundamental modeling challenges tied to bodily affect cues, including weak feature discriminability, limited short-range temporal context, and multi-task interference when jointly learning discrete and continuous affect signals.

The XGBoost model yields negative R^2 scores on all VAD dimensions (valence: 0.090, arousal: -0.014, dominance: -0.058), underperforming mean-value prediction despite 5-fold cross-validation and extensive tuning. Similarly, the 3D-CNN fails to learn effective VAD regressors while exhibiting modest multi-label performance (IoU = 0.23). These findings, contrasted against strong facial affect benchmarks, demonstrate that bodily emotion recognition cannot be directly inherited from facial pipelines and instead requires domain-specific modeling strategies.

Despite modest predictive performance (XGBoost $R^2 \approx -0.05$, 3D-CNN IoU = 0.23), our study establishes reproducible baselines, detailed failure analyses, and resource profiles for future work. These results demonstrate that bodily affect cannot be directly inherited from facial pipelines and motivate specialized modeling strategies.

Future work includes: (i) explicit skeletal pose representations for viewpoint-robust bodily cues, (ii) spatial–temporal attention to emphasize emotionally salient motion, (iii) longer temporal context via transformer or hierarchical architectures, (iv) self-supervised or auxiliary tasks to strengthen representations, and (v) multimodal fusion with audio, context, or partial facial input. Together, these directions aim to advance reliable, scalable, and cognitively aligned bodily emotion understanding, bridging the $\sim 200\times$ gap relative to facial affect recognition.

References

Alakwaa, W.; Nassef, M.; and Badr, A. 2017. Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). *International Journal of Advanced Computer Science and Applications*, 8(8).

Almaev, T. R.; and Valstar, M. F. 2013. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *2013 Humaine association conference on affective computing and intelligent interaction*, 356–361. IEEE.

Barrett, L. F.; Gross, J.; Christensen, T. C.; and Benvenuto, M. 2001. Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition & Emotion*, 15(6): 713–724.

Bentéjac, C.; Csörgő, A.; and Martínez-Muñoz, G. 2021. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3): 1937–1967.

Bolles, R. C.; and Cain, R. A. 1982. Recognizing and locating partially visible objects: The local-feature-focus method. *The international journal of robotics research*, 1(3): 57–82.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1): 5–32.

Dael, N.; Mortillaro, M.; and Scherer, K. R. 2012. Emotion expression in body action and posture. *Emotion*, 12(5): 1085.

De Gelder, B. 2009. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535): 3475–3484.

Ekman, P. 2006. *Darwin and facial expression: A century of research in review*. Ishk.

Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; and Scholkopf, B. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4): 18–28.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Kleinsmith, A.; and Bianchi-Berthouze, N. 2012. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1): 15–33.

Koolagudi, S. G.; and Rao, K. S. 2012. Emotion recognition from speech: a review. *International journal of speech technology*, 15(2): 99–117.

Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; and Matthews, I. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, 94–101. IEEE.

Mehrabian, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current psychology*, 14(4): 261–292.

Pantic, M.; and Rothkrantz, L. J. M. 2002. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on pattern analysis and machine intelligence*, 22(12): 1424–1445.

Russell, J. A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6): 1161.

Sherstinsky, A. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404: 132306.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; and Baik, S. W. 2017. Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE access*, 6: 1155–1166.

Wu, Y.; and Ji, Q. 2019. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127(2): 115–142.

Zhang, Z.; Luo, P.; Loy, C. C.; and Tang, X. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5): 550–569.