Mapping Bias: Visualizing Valence-Arousal Distributions to Reveal Affective Gaps in Face Datasets

Udayini VEDANTHAM[a*], Nithish CHOUTI[a], Nikhil Karthik C[a], Avaneesh SUNDERARAJAN[a], Ashwin TUDUR SADASHIVA[b], Manjunath K VANAHALLI[a], and, Gautam Biswas[b]

[a]*IIIT Dharwad, Karnataka, India*
[b]*Vanderbilt University, Nashville, Tennessee, USA*
*udayinivedantham@gmail.com

Abstract: Facial expression datasets play a foundational role in affective computing and emotion recognition. However, many rely on categorical emotion labels that may not reflect the true affective state captured in facial cues. The Valence–Arousal (VA) framework provides a dimensional alternative, mapping emotions along axes of polarity and activation. Despite its theoretical strengths, VA-based validation is rarely applied to large-scale facial datasets, raising questions about the fidelity of categorical annotations. This paper presents a systematic framework for auditing affective distributions in facial expression datasets through VA analysis. We evaluate two widely used datasets, DAiSEE and AffectNet, by estimating VA scores with two state-of-the-art deep learning models, HSEmotion and EmoNet. Kernel density heatmaps and statistical measures, including mean valence, mean arousal, standard deviation, and 90% coverage ellipses, are used to assess alignment with Russell's Circumplex Model of Affect. Our findings reveal significant inconsistencies: DAiSEE's engagement class and AffectNet's fear class deviate substantially from their expected VA regions, while boredom and anger demonstrate more reliable clustering. These results expose systematic biases in dataset labeling that may compromise downstream affective models. By grounding dataset validation in continuous affective space, the proposed framework enhances transparency, supports ethical dataset design, and promotes the development of more robust and interpretable emotion recognition systems.

Keywords: Affective computing, Valence-Arousal, Emotion analysis, Dataset bias, Face de-identification, Heatmap visualization, StyleGAN, HSEmotion, EmoNet.

1. Introduction

Facial expression recognition (FER) plays a central role in affective computing, supporting applications in education, mental health, human–computer interaction, and surveillance (Wang et al., 2022). Most FER systems are trained on large-scale datasets annotated with categorical emotion labels such as happy, fearful, or engaged (Ashwin & Guddeti, 2020). While these labels are convenient for classification, they fail to capture the continuous and nuanced nature of human affect, where emotions rarely fit neatly into discrete categories.

The Valence–Arousal (VA) framework offers a more expressive alternative by representing affect along two continuous dimensions: valence, which measures emotional polarity (positive to negative), and arousal, which measures activation level (calm to excited). This two-dimensional model, grounded in psychological research such as Russell's Circumplex Model of Affect (Russell, 1980, Akpanoko et al., 2024a), enables a richer understanding of emotional states, including blended or ambiguous expressions that categorical models often overlook (Akpanoko et al., 2024b; Fonteles et al., 2024).

Despite its advantages, VA-based evaluation is rarely applied in dataset curation or validation. As a result, important questions remain unanswered: Do categorical labels in widely used datasets actually align with their expected valence–arousal distributions? If not, what biases or inconsistencies might these misalignments introduce into FER systems trained on them? In this work, we define such problems as distributional misalignments between categorical labels and their expected VA coordinates. This represents a form of label bias or annotation inconsistency: distinct from demographic or societal bias, but equally important, since systematic misalignment can compromise both the reliability and fairness of affective computing systems.

In this paper, we address these questions by introducing a multi-model auditing framework that projects categorical labels into VA space for validation. We focus on two widely used datasets: DAiSEE (Gupta et al., 2015), which models student engagement in e-learning contexts, and AffectNet (Mollahosseini, et al., 2017), a large-scale benchmark for facial emotion recognition. Using two state-of-the-art models for VA estimation — HSEmotion (Kollias & Zafeiriou, 2020) and EmoNet (Toisoul et al., 2021) — we estimate valence and arousal values for selected classes and compare their distributions against the zones predicted by Russell's Circumplex Model. By generating valence–arousal heatmaps and computing distributional statistics, we reveal where dataset labels deviate from theoretical expectations. By grounding dataset validation in continuous affective space, our framework provides deeper insight into dataset quality and promotes the development of more reliable, interpretable, and ethically responsible emotion recognition systems.

2. Related Works

Early research in facial expression recognition has been dominated by categorical models of emotion, particularly Ekman's six basic emotions (Ekman, 1992). While these categories provide a simple framework for annotation, they often fail to capture gradual transitions, blended states, and cultural variability in how affect is expressed. In contrast, dimensional models such as Russell's Circumplex Model of Affect (Russell, 1980) represent emotions along two continuous axes—valence, describing the polarity of affect from positive to negative, and arousal, describing activation from calm to excited. This two-dimensional approach offers a richer characterization of affective states, yet relatively few facial expression datasets provide annotations in valence–arousal (VA) space.

Recent advances in deep learning have enabled direct estimation of valence and arousal from facial images. Models such as HSEmotion (Kollias & Zafeiriou, 2020) and EmoNet (Toisoul et al., 2021) predict continuous affective values rather than discrete categories and have been applied to tasks such as affect tracking, facial animation, and privacy-preserving transformations (Wang et al., 2022,). However, these approaches raise new questions about cross-model consistency and whether dataset labels align with expected VA distributions, since different models may yield divergent affective predictions for the same images.

Concerns about dataset quality and annotation reliability have been raised in several studies. Benitez-Quiroz et al. (2016) demonstrated the difficulty of capturing subtle expressions at scale, while Barsoum et al. (2016) showed that crowd-sourced labeling often introduces variability and disagreement. We follow prior work that has described these issues as annotation noise or label inconsistency (Ashwin et al., 2025; Ashwin and Biswas, 2024; Benitez-Quiroz et al., 2016; Barsoum et al., 2016). In this paper, we frame such systematic distributional misalignments as a form of label bias. These issues highlight what we define in this paper as label bias or annotation inconsistency—systematic misalignments between categorical labels and their expected VA coordinates. Unlike demographic or societal bias, this form of bias arises from inconsistencies in the affective content of labeled data itself, yet it can equally undermine the generalization and fairness of downstream models.

Visualization techniques offer an important opportunity to bridge this gap. Prior work in psychology and human–computer interaction has used heatmaps to illustrate affective distributions derived from physiological signals or self-reports (Kreibig et al., 2013). In machine learning, however, their application has been limited. By combining VA estimation with kernel density estimation (KDE) heatmaps, it becomes possible to expose distributional misalignments, reveal outliers, and highlight overlaps between emotion categories—providing a more transparent assessment of dataset fidelity.

3. Methodology

To examine affective inconsistencies and reveal potential bias in facial expression datasets, we designed an experimental pipeline that combines dataset curation, standardized preprocessing, valence–arousal (VA) estimation using multiple deep learning models, and comparative analysis against established psychological emotion theory.

We initially considered three widely used facial expression datasets—DAiSEE, AffectNet, and CK+ (Lucey et al., 2010). However, CK+ was excluded from the analysis due to its low resolution (24×24 pixels) and grayscale format, which make it unsuitable for modern VA estimation models that require high-resolution color images. The final analysis therefore focused on DAiSEE and AffectNet. From each dataset, we sampled 100 images per emotion class, yielding a total of 400 images across four categories. DAiSEE contributed the classes engagement (typically associated with positive valence and moderate arousal) and boredom (low valence, low arousal), while AffectNet provided fear (negative valence, high arousal) and anger (negative to moderately negative valence, high arousal). Ambiguous or co-occurring expressions were excluded to ensure greater label purity. This selection creates a controlled subset that spans a wide region of VA space, ranging from calm to intense emotional states.

Each image was standardized using Multi-task Cascaded Convolutional Networks (MTCNN) for face preprocessing. MTCNN detects the facial region, aligns landmarks such as the eyes and mouth, and outputs a normalized 256×256 face crop. This process reduces noise from pose, scale, and alignment variations, ensuring consistent input representation across datasets. When multiple faces were detected, the largest was retained under the assumption that it corresponded to the primary subject (Zhang et al., 2016; Ashwin & Guddeti, 2019)

Valence–arousal estimation was performed using two state-of-the-art pretrained models. The first, HSEmotion (Kollias & Zafeiriou, 2020), employs an EfficientNet-B0 backbone trained on large-scale affective datasets and outputs continuous valence and arousal scores in the range (−1, 1). The second, EmoNet (Toisoul et al., 2021), is a convolutional architecture optimized for emotion recognition in video but applicable to still images, also producing continuous VA predictions in the range (−1, 1). For each sample, both models generated independent VA values, which were then averaged to reduce model-specific variance. This dual-model approach provides more robust estimates of underlying affective states while mitigating biases introduced by individual training distributions.

To visualize affective distributions, we projected the VA outputs into two-dimensional heatmaps. Using kernel density estimation (KDE), we constructed smooth distributions for each emotion class and overlaid them on a fixed VA coordinate system. The x-axis represents valence $v \in (-1, 1)$, while the y-axis represents arousal $a \in (-1, 1)$. These visualizations allow direct comparison with Russell's Circumplex Model of Affect, which specifies expected regions for each categorical label.

In addition to visual inspection, we computed quantitative statistics to characterize each distribution. For each emotion class and model, we report the mean valence μ_v and mean arousal μ_a, the corresponding standard deviations σ_v, σ_a, and a 90% coverage ellipse derived from KDE density contours. Formally, if (v_i,a_i) denote the valence–arousal coordinates of N images in a class, the mean values are given by:

$$u_v = \frac{1}{N}\sum_{i=1}^{N} v_i\,, \qquad u_a = \frac{1}{N}\sum_{i=1}^{N} a_i$$

and the spread of the distribution is captured by the covariance matrix (given below),

$$\sum = \begin{bmatrix} \sigma_v^2 & \sigma_{va} \\ \sigma_{va} & \sigma_a^2 \end{bmatrix}$$

From which the principal axes of the 90% confidence ellipse are derived. These metrics quantify not only whether a distribution is centered in its expected VA region but also how widely it spreads across other zones, providing a diagnostic for label noise and ambiguity.

Through this methodology, we create both a visual and statistical basis for assessing the fidelity of categorical labels against continuous affective representations. The combination of preprocessing, dual-model VA estimation, and KDE-based analysis enables a rigorous audit of dataset quality in affective computing.

4. Results

We assessed the affective fidelity of DAiSEE and AffectNet by projecting categorical labels into Valence–Arousal (VA) space using HSEmotion and EmoNet. Heatmap visualizations and statistical measures provide evidence of how closely dataset labels align with the regions defined by Russell's Circumplex Model of Affect.

For DAiSEE, two emotion classes were examined: engagement and boredom. Engagement is theoretically associated with positive valence and moderate arousal, consistent with attentiveness and interest. However, the heatmaps generated from both HSEmotion and EmoNet reveal that many engagement-labeled images drift toward neutral or slightly negative valence and frequently occupy the low-arousal region. This suggests that the engagement class contains substantial label noise, with samples that do not express the intended affective state. In contrast, boredom aligns more reliably with its expected quadrant of low valence and low arousal. Both models produced concentrated heatmaps in this region, with only minor outliers drifting toward mildly positive valence or elevated arousal, likely reflecting states such as fatigue or passive observation. Taken together, DAiSEE demonstrates that boredom is consistently captured, while engagement suffers from substantial intra-class variance, highlighting the difficulty of annotating complex cognitive-emotional states.

For AffectNet, we analyzed fear and anger, two high-arousal emotions with negative valence. Anger maps relatively well onto its expected quadrant, with both models producing clusters around the negative-valence, high-arousal region. A small subset of samples, particularly under EmoNet, drifted toward neutral valence, possibly reflecting determination or frustration, which share visual similarity with anger. Fear, however, showed far greater inconsistency. Although theoretically expected in the low-valence, high-arousal region, the heatmaps revealed significant spread across neutral and even slightly positive valence areas. These misalignments suggest annotation ambiguities and overlap with expressions of surprise or sadness, emotions that are difficult to disentangle visually.

Table 1. *Summary of Valence-Arousal Statistics Across Classes and Models*

Emotion Class	Model	Mean Valence	Mean Arousal	Std(V,A)	90% Coverage
Engagement	HSEmotion	-0.015	0.270	0.1280	82.00%
	EmoNet	-0.415	0.375	0.2145	84.00%
Boredom	HSEmotion	0.077	0.143	0.1518	84.00%
	EmoNet	0.007	0.013	0.0138	84.00%
Fear	HSEmotion	-0.249	0.452	0.2823	82.83%
	EmoNet	-0.023	0.041	0.0257	82.83%
Anger	HSEmotion	-0.249	0.452	0.2125	82.11%
	EmoNet	-0.044	0.066	0.0193	82.11%

Quantitative results reinforce these observations. Table 1 summarizes the mean valence, mean arousal, standard deviations, and 90% coverage of each class. Engagement exhibits relatively wide spread, with HSEmotion centering near neutral valence ($\mu_v \approx -0.015$) and moderate arousal, while EmoNet predicts more negative valence ($\mu_v \approx -0.415$) with higher variance. Boredom, by contrast, remains concentrated near low valence and arousal with limited variability across both models. Fear shows notable dispersion, with EmoNet producing near-neutral valence ($\mu_v \approx -0.023$) and low arousal compared to HSEmotion's more negative valence ($\mu_v \approx -0.029$), reflecting the instability of this class. Anger, though also high-arousal and negative, maintains stronger alignment across both models, with similar mean values and relatively compact coverage.

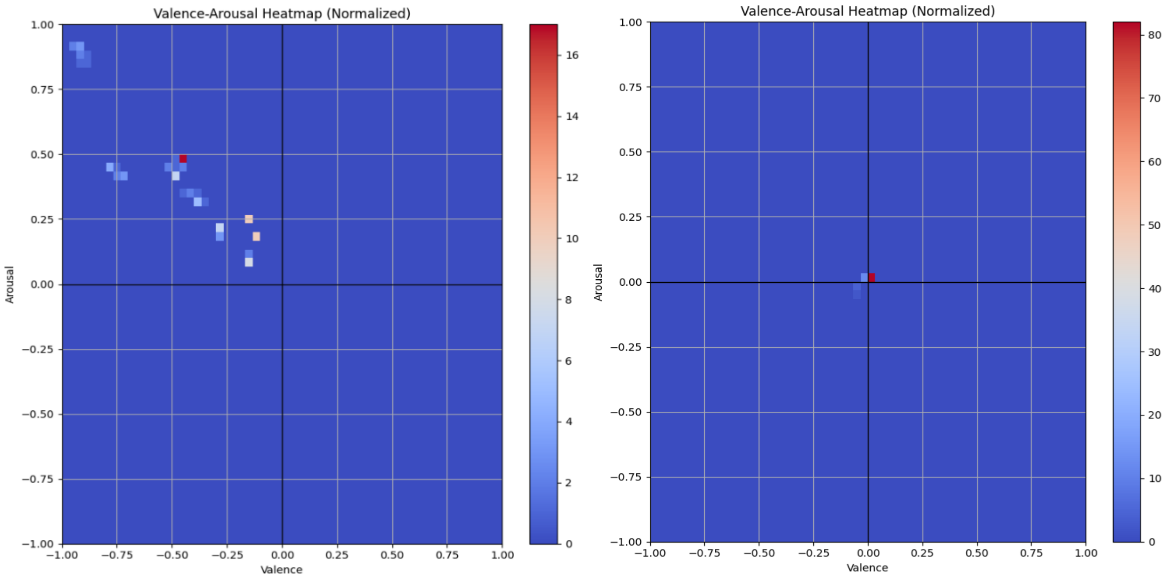

Figure 1. Valence Arousal Heatmaps of Engagement (left) and Boredom (right) Emotion from DAiSEE Dataset using EMONet

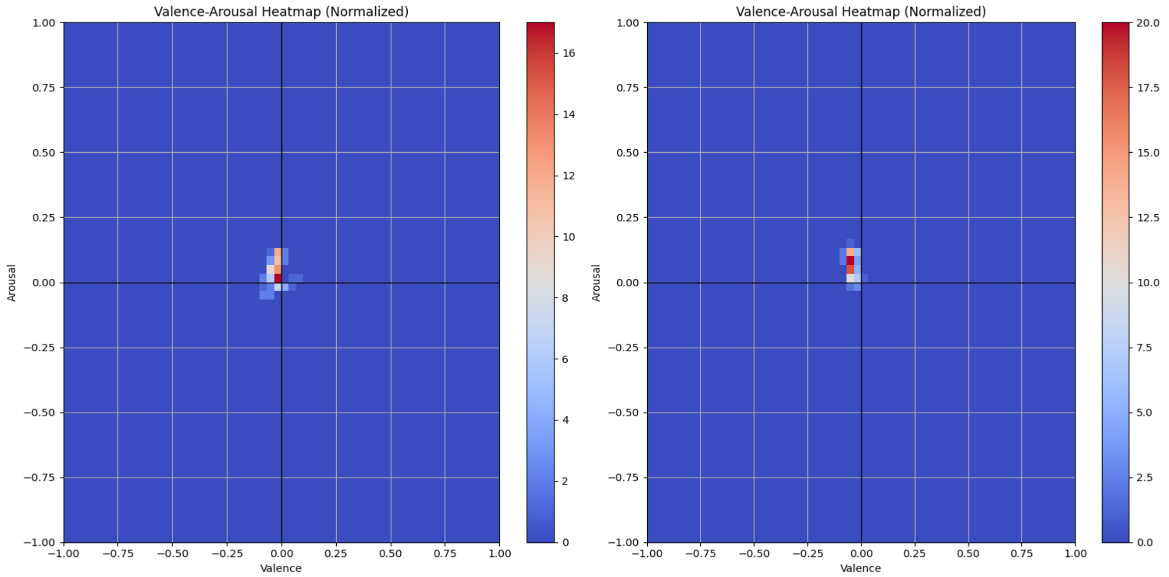

Figure 2. Valence Arousal Heatmaps of Fear (left) and Anger (right) Emotion from AffectNet Dataset using HSEMotion

Figures 1–4 illustrate the VA heatmaps for representative classes. Figure 1 and Figure 3 shows engagement and boredom from DAiSEE using EmoNet and HSEMotion respectively, highlighting the dispersion of engagement samples versus the concentration of boredom. Similarly, Figure 2 and Figure 4 shows fear and anger from AffectNet using EmoNet and HSEMotion respectively. Anger from AffectNet, demonstrates alignment with its expected region while fear reveals broad distributional drift beyond the target quadrant. Together, these heatmaps provide visual confirmation of the statistical trends.

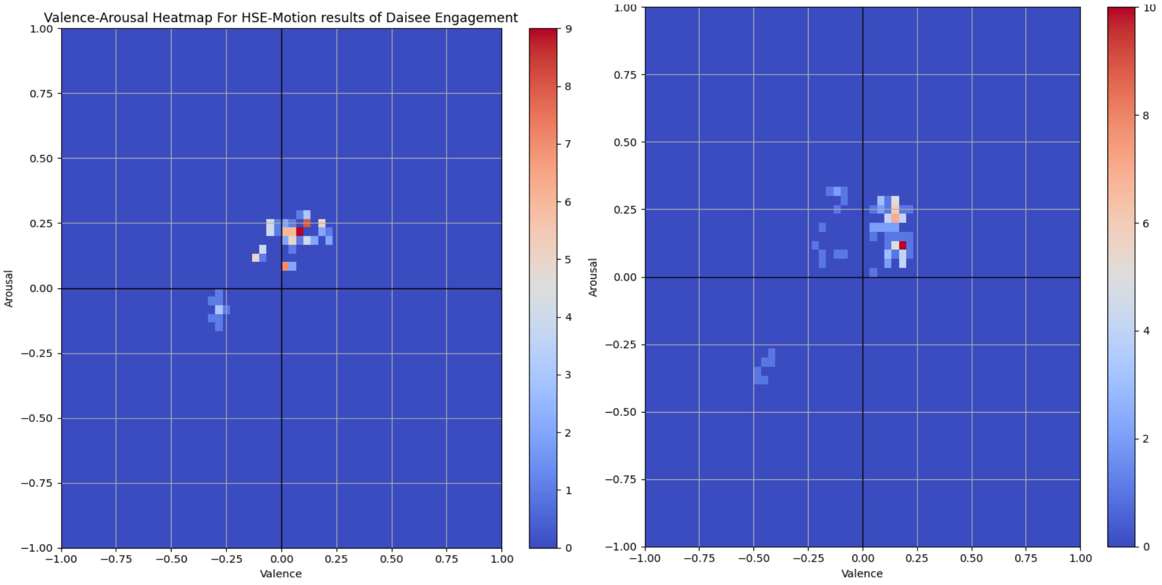

Figure 3. Valence Arousal Heatmaps of Engagement (left) and Boredom (right) Emotion from DAiSEE Dataset using HSEMotion

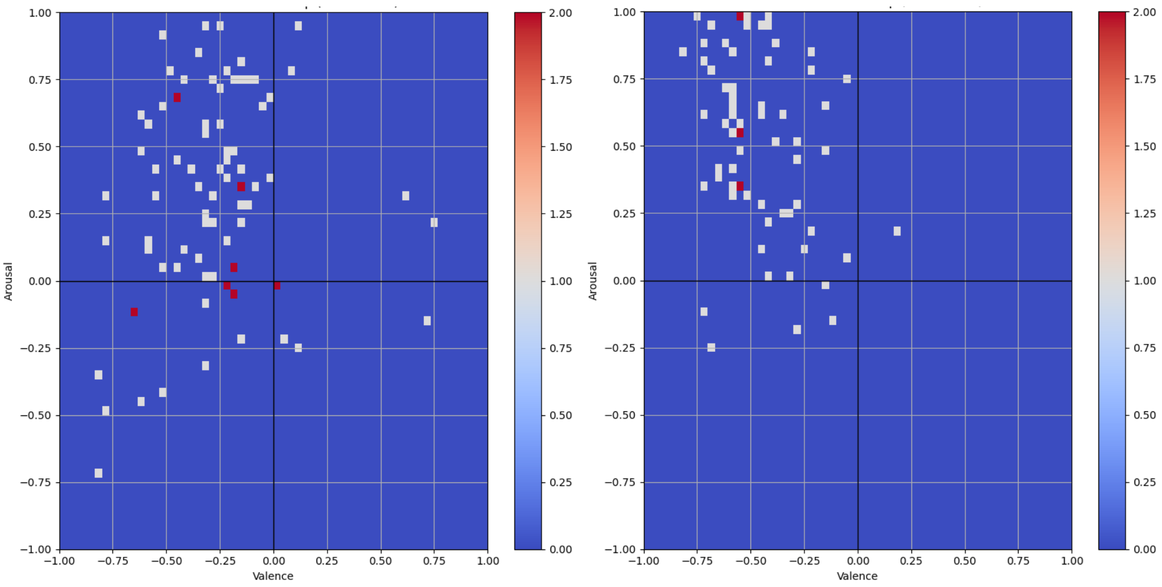

Figure 4. Valence Arousal Heatmaps of Fear (left) and Anger (right) Emotion from AffectNet Dataset using HSEMotion

Overall, the results show that boredom and anger are more reliably represented in VA space, while engagement and fear are prone to misalignment and drift. These findings expose systematic biases in categorical labeling and emphasize the importance of VA-based auditing for dataset validation in affective computing.

5. Discussion and Implications

The analysis highlights notable inconsistencies between categorical labels in widely used datasets and their actual Valence–Arousal (VA) distributions. While boredom in DAiSEE and anger in AffectNet clustered reliably in their expected quadrants, engagement and fear displayed substantial drift. Engagement often shifted toward neutral or negative valence with reduced arousal, and fear dispersed across regions associated with surprise or sadness.

These findings indicate that categorical labels do not always map cleanly onto their presumed affective coordinates, raising questions about annotation fidelity.

From an educational perspective, such inconsistencies are particularly consequential. DAiSEE, for instance, was designed to model student engagement in e-learning environments. If its engagement labels are systematically misaligned in VA space, models trained on this data may struggle to detect true engagement or confuse it with unrelated affective states. This risks producing unreliable systems in educational contexts, where accurate modeling of learner behavior is critical. Incorporating VA-based dataset auditing into curricula can help students and practitioners recognize these pitfalls and develop a more critical understanding of dataset quality.

The ethical implications extend beyond education. In applications such as healthcare, security, or human–computer interaction, reliance on datasets with misaligned labels can propagate biases and reduce fairness. A model trained on noisy engagement or fear labels may overfit to ambiguous cues, leading to flawed inferences about a person's emotional state. By contrast, VA auditing provides a systematic way to detect such biases before training. Comparing distributions against established psychological models, such as Russell's Circumplex, acts as an early warning mechanism for misrepresentation and imbalance.

Taken together, these results demonstrate the importance of moving beyond categorical validation toward dimensional analysis. VA-based auditing not only improves interpretability but also enhances accountability in affective computing systems. As affect recognition technologies become increasingly integrated into sensitive domains, ensuring that datasets are affectively consistent is both a technical and ethical imperative.

6. Conclusion and Future Work

This study introduced a framework for auditing facial expression datasets by mapping categorical labels into Valence–Arousal (VA) space using multiple deep learning models. Applied to DAiSEE and AffectNet, the approach revealed that some emotions, such as boredom and anger, align closely with their theoretical VA regions, while others, particularly engagement and fear, exhibit significant drift. These findings underscore the risk of relying solely on categorical labels in affective computing and demonstrate the value of dimensional validation for improving dataset quality. By combining standardized preprocessing, dual-model VA estimation with HSEmotion and EmoNet, and kernel density–based heatmap analysis, the framework provides both statistical and visual diagnostics of dataset fidelity. This approach advances the field toward more transparent and interpretable evaluation of affective resources, offering a foundation for fairer and more reliable emotion recognition systems.

Future work will focus on automating the VA auditing pipeline to scale analysis across larger datasets, extending the method to temporal sequences for video-based emotion tracking, and examining how demographic attributes influence VA distributions. We also plan to explore VA-aware loss functions that can preserve affective fidelity during face de-identification or privacy-preserving transformations. By addressing these directions, this line of research aims to support the development of affective computing systems that are not only technically robust but also ethically responsible.

Acknowledgements

This work was supported by the National Science Foundation AI Institute Grant #DRL-2112635, Digital Lab Seed Grant and the TIH-IoT TDP-3 grant (No. TIH-IoT/2023-3/TDP-3-SE/Acad/SL-006). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or TIH-IoT.

References

Akpanoko, C. E., Cordell, G., & Biswas, G. (2024). Investigating the relations between students' affective states and the coherence in their activities in open-ended learning environments. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 511-517).

Akpanoko, C. E., & Biswas, G. (2024). The interplay of affective states and cognitive processes in an open-ended learning environment: A case study. In *Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 873-880*. International Society of the Learning Sciences.

Ashwin, T. S., Sanda, N., Timalsina, U., & Biswas, G. (2025, July). Challenges of Applying Computer Vision for Emotion Detection in Educational Settings: A Study on Bias. In *International Conference on Artificial Intelligence in Education* (pp. 388-395). Cham: Springer Nature Switzerland.

Ashwin, T. S., & Biswas, G. (2024, July). Identifying and mitigating algorithmic bias in student emotional analysis. In *International Conference on Artificial Intelligence in Education* (pp. 89-103). Cham: Springer Nature Switzerland.

Ashwin, T. S., & Guddeti, R. M. R. (2020). Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. *Future Generation Computer Systems*, *108*, 334-348.

Ashwin, T. S., & Guddeti, R. M. R. (2019). Unobtrusive behavioral analysis of students in classroom environment using non-verbal cues. *IEEE Access*, *7*, 150693-150709.

Benitez-Quiroz, C. R., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. CVPR.

Barsoum, E., Zhang, C., Canton Ferrer, C., & Zhang, Z. (2016). Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. ICMI.

Dalgleish, T., & Power, M. J. (2004). Handbook of Emotion Regulation. Guilford Press.

Ekman, P. (1992). An Argument for Basic Emotions. Cognition & Emotion, 6(3-4), 169–200.

Fonteles, J., Davalos, E., Ashwin, T. S., Zhang, Y., Zhou, M., Ayalon, E., ... & Biswas, G. (2024, July). A first step in using machine learning methods to enhance interaction analysis for embodied learning environments. In *International Conference on Artificial Intelligence in Education* (pp. 3-16). Cham: Springer Nature Switzerland.

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.

Koelstra, S., et al. (2012). DEAP: A Database for Emotion Analysis Using Physiological Signals. IEEE Transactions on Affective Computing, 3(1), 18–31.

Kollias, D., & Zafeiriou, S. (2020). Analyzing Affective Behavior in the First ABAW 2020 Competition. arXiv preprint arXiv:2001.11409.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010, June). 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* (pp. 94-101). IEEE.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18-31.

Russell, J. A. (1980). A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39(6), 1161–1178.

Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. Psychological Review, 110(1), 145–172.

Schuller, B., et al. (2011). Recognising Realistic Emotions and Affect in Speech: Lessons from the First Challenge. Speech Communication, 53(9–10), 1062–1087.

Toisoul, A., et al. (2021). Estimation of Continuous Valence and Arousal Values Using Deep Regression Neural Networks. IEEE Transactions on Affective Computing, 12(1), 190–203.

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., ... & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, *83*, 19-52.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, *23*(10), 1499-1503.