PHANTOMATRIX: Explainability for Detecting Gender Bias in Affective Computing

Anne Schwerk[1,2] and Armin Grasnick[1]

[1] IU International University, Department IT & Engineering, Extended Artificial Intelligence, Erfurt, Germany

[2] Berlin Institute of Health, Berlin, Germany

anne.schwerk@iu.org
armin.grasnick@iu.org

Abstract: The PHANTOMATRIX project is a research incubator running at the International University of Applied Sciences and aims to advance the field of Human-Machine Interaction by integrating machine learning (ML) techniques to predict emotional states using physiological and facial expression data within Virtual Reality environments. A major focus of the PHANTOMATRIX project is on employing trustworthy ML models by using explainable AI (XAI) methods that allow to rank features according to their predictive power, which aids in understanding the most influential factors in emotional state predictions. In addition, a comparative analysis of XAI techniques to emotion prediction models allows us to assess and correct for the effect of gender on the predictive performance. As affective computing is a highly sensitive research arena, it is of outmost importance to ensure bias free models. Key XAI methods such as Deep Taylor Decomposition (DTD), and SHapley Additive exPlanations (SHAP) are employed to clarify the contributions of features towards model predictions, providing insights into how specific signals influence emotion detection across individuals. This allows for a comprehensive comparison of different XAI approaches and their utility in gender bias detection and mitigation. To further our understanding of gender dynamics within emotional predictions, we develop intuitive visualizations that graphically represent the link between multimodal input data and the resulting emotional predictions to support the interpretation of complex model outputs and to make them more accessible not only to researchers but also to novice users of the system. Our background research demonstrates the effectiveness of XAI methods in identifying and mitigating gender bias in emotion prediction models. By applying XAI, the project reduces the influence of gender-based disparities in affective computing, leading to more equitable model performance across demographics. This research not only highlights the importance of transparent, bias-free AI-affect models but also sets a foundation for future developments in responsible affective computing. The findings contribute to advancing trust in AI-driven emotion analysis, promoting fairer and more inclusive applications of this highly relevant technology.

Keywords: Explainable AI, Human-Machine Interaction, Machine Learning, Affective Computing, Virtual Reality, XAI.

1. Introduction

Affective computing, a term created by Rosalind Picard (1995), is a multidisciplinary subfield of artificial intelligence (AI), develops systems that recognize, interpret, and simulate human emotions. This research area is essential for enhancing human-computer interactions and developing applications that respond empathetically to users. Understanding human emotions is not only vital for creating more user-friendly technologies but also plays a crucial role in various domains such as mental health, education, and customer service (Pei et al, 2024).

Wearable technology has revolutionized the way we collect and analyze emotional data (Schmidt et al, 2019). Devices such as smartwatches and specialized sensors can continuously and non-invasively monitor physiological signals that correlate with emotional states and hence provide insights into the emotional experiences of individuals in real-time. The integration of multi-modal wearable data offers researchers valuable tools for capturing the complexity of human emotions. However, incorporating multiple data sources requires a more sophisticated interpretation to ensure that end-users understand how insights are derived from their data. Especially when "black-box" machine learning (ML) techniques are applied, it becomes crucial to explain the outcomes to enhance transparency, provide meaningful insights, and control. Explainable AI (XAI) can play a key role in making ML models more interpretable and reveal potential bias (Pahde et al, 2023).

Despite the advancements in affective computing, the presence of inherent biases and ethical concerns poses significant challenges concerning privacy, consent, discrimination, and the potential for data misuse (Iren et al, 2023). Affective computing can gather sensitive biometric and behavioral data that is susceptible to bias, arising from datasets that do not represent diverse cultures or demographic groups, leading to discriminatory outcomes and social stereotypes, including gender bias (Manresa-Yee et al, 2023; Suman et al. 2022). As algorithms can be vulnerable to stereotyping and discrimination, particularly when their performance varies across gender groups as shown for affect classification, also Microsoft committed to the retirement of classifiers for attributes like

gender and emotion in its Face API[1]. This highlights the critical need for fairness and transparency in emotion recognition to prevent a perpetuation of harmful stereotypes and unjust outcomes, to ensure these systems are both fair and inclusive. By enhancing the transparency and reliability of emotion recognition technologies, we can better address these biases and reduce the risk of discrimination, ensuring a more equitable deployment of AI in emotion detection applications (Murindanyi et al, 2023).

In sum, PHANTOMATRIX focusses on AI-based multi-modal emotion classification and aims to explore whether XAI methods can be utilized to identify and mitigate potential gender biases present for classifying emotions.

2. Methods

2.1 Multimodal ML-analysis in PHANTOMATRIX

The resulting data from 100 participants of the PHANTOMATRIX study will be used, which is derived from the FDA-cleared Empatica EmbracePlus, a state-of-the-art wearable device, that collects multi-modal physiological data, e.g. electrodermal activity (EDA) and skin temperature. In addition, self-report data from participants that reflect their emotional experience, and a structured questionnaire (PANAS) will be used to ensure the mapping of emotional states to physiological signals. Also, facial expressions will be recorded and analyzed.

Features indicative of specific emotions (e.g. anxiety) are extracted from the raw physiological data using dimensionality reduction techniques, such as Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE). Relevant time intervals are then determined and mapped to each modality. The ML model design from Vu et al (2023) will be adapted, which uses an early fusion of multi-modal emotional features based on Gaussian Transformation of physiological signals followed by a Transformer Encoder processing. The design will be extended for PHANTOMATRIX with facial expressions (see Figure 1). Model performance is evaluated using a subset of the data reserved for testing (five-fold cross validation) and based on the self-report outcomes that serve as ground truth.

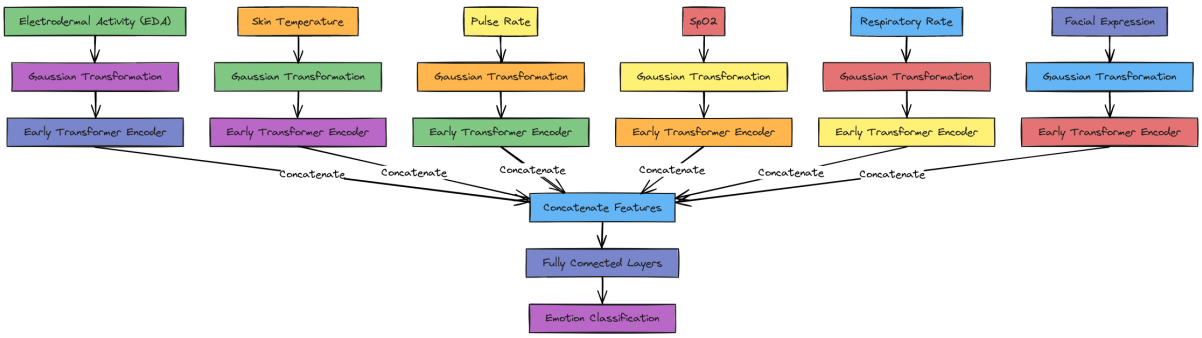

Figure 1: ML model design showing the integration of the different features

2.2 Bias Detection and Mitigation

Post-hoc XAI techniques such as Deep Taylor Decomposition (DTD) and SHapley Additive exPlanations (SHAP) will be applied to analyze and visualize the importance of sensitive features across modalities. After assessing their impact on the prediction outcome, the effects will be mitigated from the overall predictive model to obtain a bias-free general affect model.

2.2.1 Deep Taylor Decomposition

DTD explains neural network predictions by breaking down the output into contributions from each input feature by propagating relevance scores backward through the network (Lauritsen et al, 2020). Using a first-order Taylor series expansion around a chosen reference point for each neuron, DTD estimates how changes in input features affect the overall prediction. Using DTD, the relevance of each input feature (e.g., physiological signals) to the model's emotional state predictions can be traced back through the Transformer layers. By calculating the relevance scores separately for male and female data samples, DTD can identify if certain physiological features disproportionately influence the predictions based on gender.

2.2.2 Gender Bias Detection Based on SHAP

SHAP have been repeatedly used to quantify bias, like gender effects, in ML models (Kristjanpoller et al, 2023). To enable time-series analysis with SHAP, VARSHAP, a method that uses vector autoregressive models to capture temporal dependencies, is used (Villani et al, 2022). Additionally, an event detection method aggregates feature importance over time, enabling identification of influential time steps.

SHAP interaction values will be used to assess not only the individual feature contributions to emotion predictions but also the interaction effects between different features, specifically focusing on how these interactions may differ between genders. This method is adapted from Lundberg et al (2020) and allows for a nuanced understanding of how combinations of physiological signals and facial expressions interact to affect emotional state predictions, with respect to gender differences. Lundberg's SHAP interaction plots are adapted to show interactions between gender and time-dependent features to uncover whether certain feature combinations disproportionately affect predictions for specific groups. By using VARSHAP to compute feature importances and interactions for different demographic subgroups, it's possible to create interaction plots for each subgroup, indicating potential biases. Combining time series SHAP with interaction plots will visualize how interactions vary over time for different groups, based on heatmaps and line plots. In addition, event detection to interaction values (rather than just individual SHAP values) will allow to identify moments when demographic-feature interactions spike unusually. This indicates that a model's response to certain interactions is heightened for specific groups at certain times.

2.2.3 Bias Mitigation Strategies

To mitigate bias detected in the ML models, two approaches will be applied:

- An adversarial network is applied that learns to predict gender based on DTD/SHAP outputs. The main model is trained to minimize this predictability, reducing demographic-based dependencies. Then attention weights are adjusted or regularized to ensure that the focus on key features is consistent across gender. This minimizes differential impacts on predictions based on gender characteristics.
- The calculated feature importances from gender are used for each prediction and SHAP values corresponding to these sensitive features are summed to understand how much they contribute to the overall prediction. Then the summed SHAP contributions of the sensitive features are subtracted from the model's raw prediction with the SHAP Explainer[2]. This adjustment effectively "removes" the influence of demographic information, leaving only the contributions from non-sensitive features.

Insights derived from the XAI methods, and the mitigation strategies will be used to refine the emotion detection models iteratively and the model will be adapted and retrained to minimize these biases. See Figure 2 for an overview of the process pipeline.

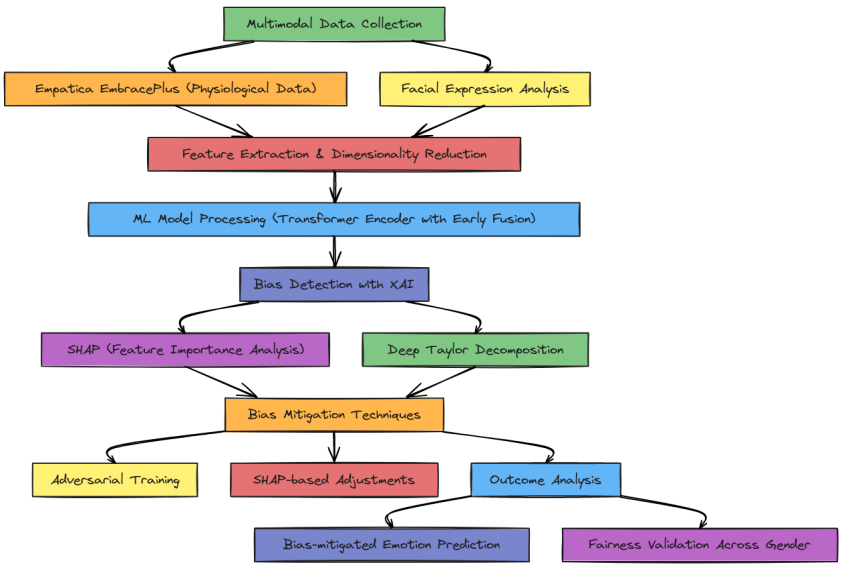

Figure 2: Overview of process for detecting bias in ML emotion classification with XAI

References

Iren, D., Yildirim, E. and Shingjergji, K., 2023, September. Ethical risks, concerns, and practices of affective computing: a thematic analysis. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 1-4). IEEE.

Kristjanpoller, W., Michell, K. and Olson, J.E., 2023. Determining the gender wage gap through causal inference and machine learning models: evidence from Chile. *Neural Computing and Applications*, *35*(13), pp.9841-9863.

Lauritsen, S.M., Kristensen, M., Olsen, M.V., Larsen, M.S., Lauritsen, K.M., Jørgensen, M.J., Lange, J. and Thiesson, B., 2020. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature communications*, *11*(1), p.3852.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, *2*(1), pp.56-67.

Manresa-Yee, C., Ramis, S. and Buades, J.M., 2023. Analysis of Gender Differences in Facial Expression Recognition Based on Deep Learning Using Explainable Artificial Intelligence.

Murindanyi, S., Kirabo, C., Kirabo, N.P., Hellen, N. and Marvin, G., 2023, November. Trustworthy Machine Emotion Intelligence Using Facial Micro-expressions. In *International Conference on Intelligent Vision and Computing* (pp. 46-62). Cham: Springer Nature Switzerland.

Pahde, F., Dreyer, M., Samek, W. and Lapuschkin, S., 2023, October. Reveal to revise: An explainable ai life cycle for iterative bias correction of deep models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 596-606). Cham: Springer Nature Switzerland.

Pei, G., Li, H., Lu, Y., Wang, Y., Hua, S. and Li, T., 2024. Affective Computing: Recent Advances, Challenges, and Future Trends. *Intelligent Computing*, *3*, p.0076.

Picard, R.W., 1995. Affective computing (No. 321).

Schmidt, P., Reiss, A., Dürichen, R. and Van Laerhoven, K., 2019. Wearable-based affect recognition—A review. *Sensors*, *19*(19), p.4079.

Suman, C., Chaudhari, R., Saha, S., Kumar, S. and Bhattacharyya, P., 2022. Investigations in emotion aware multimodal gender prediction systems from social media data. *IEEE Transactions on Computational Social Systems*, *10*(2), pp.470-479.

Villani, M., Lockhart, J. and Magazzeni, D., 2022. Feature importance for time series data: Improving kernelshap. *arXiv preprint arXiv:2210.02176*.

Vu, N.T., Huynh, V.T., Yang, H.J. and Kim, S.H., 2023, November. Multiscale Transformer-Based for Multimodal Affective States Estimation from Physiological Signals. In *Asian Conference on Pattern Recognition* (pp. 113-122). Cham: Springer Nature Switzerland.