

Multi-modal Biosignal and Media Content Fusion for Human Interest Prediction Using HGNNs

Teoman Berkay Ayaz, Akhan Akbulut

Department of Computer Engineering, Istanbul Kültür University, Istanbul, Turkey

Email: 1800004169@stu.iku.edu.tr; a.akbulut@iku.edu.tr;

Abstract—Prediction of consumer interest in advertisements is an increasingly significant challenge in the interdisciplinary field of Neuromarketing. Traditional neuromarketing studies often rely on electroencephalographic (EEG) signals or facial expression data, in addition to self-reported responses to estimate user preferences. While this information can provide rich insight into human decision-making, it often suffers from limited availability in real-world deployment scenarios, in addition to causing overfitting when employing predictive solutions for user interest. To overcome these challenges, our study presents a novel multi-modal data fusion approach working with a hierarchical deep learning pipeline to overcome overfitting and availability problems. By representing emotion signals, consumer demographic information, and advertisement semantic information on a heterogeneous graph, our study uses an emotion intensity regressor and a hybrid Heterogeneous Graph Neural Network (HGNN) model working consecutively to effectively predict user interest in advertisements. The empirical results demonstrate robust performance on limited data with an accuracy of 0.71 and an F1-score of 0.82.

Keywords—Neuromarketing, Graph Data Structure, Graph Neural Networks, Multi Modal Data Fusion, Consumer Behavior Prediction

I. INTRODUCTION

In recent years, the interdisciplinary field of neuromarketing emerged as a research domain to understand human decision making in response to stimuli. Neuromarketing typically leverages methods more commonly found in domains like biomedical engineering such as, signal acquisition, image processing, and statistical modeling to capture the link between cognitive and emotional mechanics responsible for consumer behavior [1]. NeuroBioSense dataset [2] provides a multi-modal corpus of participant demographics, wearable sensor signals, facial expression recordings, and interest ratings for a diverse array of products and brands, making a significant contribution to this field.

Despite the richness of data, predicting customer interest in advertisements remain a difficult problem. Emotional signals are strong predictors of engagement, yet oftentimes deployment scenarios will have no observed emotional signals. Even in scenarios where emotion signals are available, they can often lead to overfitting, rendering any predictive solution for this task unusable. Furthermore, neuromarketing data often comes in heterogeneous and complex forms. This creates a necessity for a solution to handle the context-dependent engagement between advertisements and customers.

In this paper, we propose a multi-modal fusion of media content with biosignals and participant demographics to create a heterogeneous customer-advertisement graph for the prediction of user interest with Heterogeneous Graph Neural Networks (HGNN). The study leverages the NeuroBioSense dataset as a testbed to construct a heterogeneous graph with participant nodes, connected to advertisement nodes to achieve a machine learnable form for the multi-modal fusion of data. We convert natural language data and video data into machine-learnable embeddings using pre-trained encoders. We then train and evaluate a hybrid HGNN model based on Graph Attention (GAT), advanced linear networks, and internal feature gating mechanisms. Additionally, we compare our model against more commonly used tabular models to provide a more comprehensive evaluation. Lastly, our study provides an emotion intensity predictor that works on customer demographics and advertisement information to predict and generate the emotion signals, which are often unavailable at inference time, overcoming the availability and overfitting problems. The four-fold contributions of our study can be summarized as:

- A multi-modal feature integration pipeline which combines demographic information, media content, and self-reported emotional signals.
- A signal predictor aiming to provide biosignal data that is often unavailable at inference time, allowing for higher-performing model generalization.
- A hybrid HGNN architecture that leverages GAT, internal feature gating mechanisms, and advanced linear networks for improved predictive performance.
- An empirical demonstration of the effectiveness of the proposed approach on the NeuroBioSense dataset, achieving considerable predictive performance.

II. RELATED WORKS

Neuromarketing as a field of study often leverages methods borrowed from the Biomedical Engineering Domain. Majority of studies done in Neuromarketing often leverage biosignals recorded from volunteers in participants. These studies use neural signals and advanced machine learning to predict user preferences and interest in specific ad campaigns and commercials.

One relevant study conducted by Alimardani et al. [3] in 2021, compared traditional machine learning (ML) algorithms

to deep learning based methods, specifically Convolutional Neural Networks (CNNs). Their approach leveraged these predictive methods to predict user preferences on the basis of electroencephalographic (EEG) signals on two independent EEG datasets. The researchers found that CNN architecture performed considerably better in predicting movie ratings based on EEG signals, showcasing the potential of neural networks in neuromarketing tasks. A similar study done in 2022 by Ullah et al. [4] leveraged EEG signal analysis to predict product preferences among users, using deep learning methods and support vector machines (SVM). Their findings demonstrated that using EEG information for consumer behavior prediction is an effective solution. Similarly, Derdiyok and Patlar Akbulut [5] proposed a biosignal-based emotion-oriented video summarization framework that utilizes physiological signals such as EDA, BVP, and accelerometer data, demonstrating the effectiveness of emotion-driven approaches in multimedia analysis. A recent survey by Patlar Akbulut [6] provides a comprehensive overview of emotion recognition methods based on physiological signals, highlighting the role of biosignals as a reliable source for affective computing applications.

Multi-modal data fusion is another field of study investigated by researchers. Studies investigating this curious subject will often leverage a heterogeneous graph data structure to represent the various types of data in a machine-learnable form. One such study conducted by Silva et al. in 2022 [7], leveraged an unsupervised HGNN model for hit song prediction by fusing together a multi-modal dataset. The researchers included musical data like lyrics and artist information. Following this, the researchers formulated a one-class prediction task and found promising results as opposed to other strategies when it comes to representing similar data.

This study proposes a novel approach, which works by fusing multi-modal data to create a machine-learnable heterogeneous graph. By combining participants' recorded emotions as biosignals, demographic information, and advertisement information from both videos and textual data, our study builds an informative heterogeneous graph. Building on this information, the method trains two consecutive back-to-back models to form a production-ready pipeline for predicting user interest.

III. METHODOLOGY

As mentioned our methodology combines multi-modal data by the usage of heterogeneous graphs and pre-trained encoders to process multimedia data in a machine-learnable format. The initial stage is the processing of the NeuroBioSense Dataset [2] to extract relevant features and combine the data into a tabular dataframe. Following this, an emotion signal regressor is trained to make the biosignals available for the interest prediction stage without overfitting. Following this, the graph construction is done, and an HGNN model is trained to predict participant interest.

A. NeuroBioSense Dataset

The NeuroBioSense dataset [2] provides multidimensional data designed to study emotional and cognitive responses of consumers to advertisements. The dataset contains physiological signals, facial expressions, and demographic information of participants, presenting a rich source for emotion detection, consumer behavior, and signal processing studies, making it a suitable dataset for this study. The dataset contains 58 unique participants, each participant watching 10 or 15 ads based on the advertisement category. Each participant's age, sex, and emotional responses are stored in the dataset, allowing for demographic and biosignal information to be included in the study. Outside of the presented dataset, the advertisement videos were collected to process alongside the related participant information. One detail regarding the collection of videos is that only 31 of the 35 presented ads were collected, as the 4 remaining videos were removed from their sources. Hence, the instances related to those 4 ads were omitted from this study.

1) *Multi-Modal Data Fusion & Preprocessing*: The preprocessing for this study was done in three main stages: participant demographic info processing, advertisement info processing, and emotion signal processing. The first and last of these stages leverage standard preprocessing applications such as one-hot encoded binning for participant age, one-hot encoding for gender, and one-hot encoded emotion intensities for 7 recorded participant emotions.

In the dataset and collected videos, there are three types of available information: numeric advertisement information (duration, fps, etc.), advertisement videos, and textual information (advertisement title and category descriptions). The numeric information was preprocessed using min max scaling as a standard approach. Aside from this, to convert the natural language data belonging to these ads into machine-learnable, unidimensional representations, the approach leverages the all-MiniLM-L6-v2 pre-trained sentence transformer [8], [9]. For both types of information, this pre-trained encoder was used to generate a 384-dimensional representation. Finally, for the extraction of video-level embeddings, the approach employs the Contrastive Language-Image Pre-training (CLIP) model [10]. Each video was uniformly sampled into 12 frames across the full duration, enabling even coverage for all available clips. The frames were then encoded using the CLIP image encoder (ViT-B/32). The resulting embeddings were then normalized and aggregated by mean pooling to acquire a 512-dimensional representation for all available video clips. After conversion, all acquired representations are concatenated to form advertisement representations.

Following the preprocessing, all of the available data was fused into a heterogeneous graph with two node types (participant and advertisement), one edge type (participant, watches, ad) with emotions as edge attributes, and user interest as the edge label and prediction target. One noticeable detail in the pipeline is that the insertion of edge attributes and graph creation is done using the predictions made by the emotion

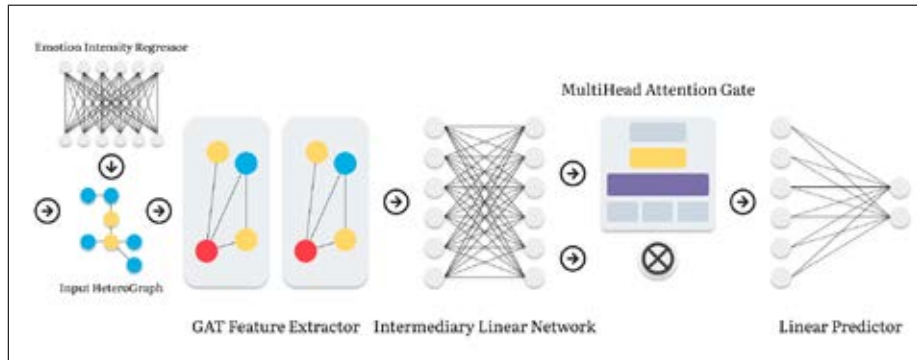


Fig. 1: Architecture of the proposed hierarchical model, combining the Emotion Intensity Regressor and the HGNN predictor

intensity predictor. Given that emotion signals can often act as standalone predictors of user interest, they can often lead to overfitting in addition to being unavailable in most production environments. The emotion intensity predictor closes this prevalent gap and makes it possible to incorporate these signals into a production-ready pipeline for further predictive tasks.

B. Hierarchical Model Architecture

To address the gap caused by missing or unreliable biosignals, the study proposes a hierarchical architecture composed of two consecutive models. The initial approximates an emotion signal based on demographic and advertisement information, generating a continuous feature stream when genuine biosignals are unavailable. The latter uses these approximated signals on a heterogeneous graph, and makes supervised user interest predictions based on this data.

1) *Emotion Intensity Predictor:* As mentioned, the users emotional response, such as shock or joy, is often unavailable in production environments. Given this, any approach aiming real-world deployment needs to overcome and fill this gap to make these biosignals available at inference time. To close this gap, this problem can be approached as a data imputation problem. Researchers working in the Artificial Intelligence and Data Science domains will often use ML models to impute the missing fields in their datasets by treating the column with missing values as a prediction target. The NeuroBioSense dataset and collected videos provide key features that can be used for this downstream prediction task. By using a concatenated set of participant demographic representations and advertisement representations, we can formulate an input feature set of X . By treating the one-hot encoded signal intensities as our prediction target y , we can formulate this problem as a multidimensional regression problem and train a model to approximate these missing signals.

For this multidimensional regression problem, our study employs a fully connected XNet model to approximate these signals. As opposed to the traditional Multi-Layer Perceptron (MLP) architecture with Rectified Linear Unit (ReLU) activation terms, the XNet architecture uses Cauchy Activation terms, which are based on the Cauchy Integration Theorem [11]. Recent studies have demonstrated that XNet architecture provides higher performance in both predictive

and regressive tasks [12]. In addition to its reported predictive advantages, XNet was chosen as it provides smoother nonlinear transformations and stronger resistance to overfitting compared to conventional MLPs, making it more suitable for the limited-sample, high-dimensional setting of this study. This study employs this architecture to provide a better proof of concept for the proposed methodology.

2) *Hybrid HGNN Predictor:* As mentioned, the HGNN model does not use the original recorded emotion signals but instead relies on the approximations generated by the emotion signal predictor. As shown in Figure 1, the HGNN model is fed through two channels: the initial being the demographic and advertisement info, which is also fed into the emotion intensity regressor, and the latter being the outputs of the emotion intensity regressor. Once the emotion intensity regressor predicts the signal, the HGNN pipeline starts to make its predictions.

The HGNN model comprises 4 key components as shown in Figure 1. The initial component is the GAT-based feature extractor to process the graph-based features. The implemented version leverages the dynamic attention mechanism proposed by Brody et al. [13] to provide further components with richer latent representations. These representations are then processed using an XNet-based intermediary network to transform them for the Multi-Head Attention-based internal feature gating mechanism. The internal gating mechanism is integrated to learn and only pass through informative signals, making it easier for the predictor head to make more accurate predictions. The last key component in the HGNN architecture is an XNet-based predictor to make edge label predictions.

To ensure reliable prediction boundaries for the HGNN model, the study employs a threshold tuning process, as opposed to relying on a fixed 0.5 threshold. To prevent collapse onto a single class of predictions to minimize loss, an F1-score-based threshold was tuned, preventing an inflated trivial all-positive solution.

C. Model Training & Experimental Setup

The preprocessing of the dataset yields a 592-row dataframe where each row represents a distinct edge on the graph structure. Given the relative low count of instances, the study leverages a 3-fold cross validation approach for the emotion

intensity regressor, and a 2:1 train-test split for the entirety of the study, with cross-validation done solely on the training dataset and further testing on the held-out test set.

IV. EXPERIMENTAL RESULTS

A. Emotion Intensity Regressor

The emotion intensity regressor yields a mean R^2 score of 0.2159 with a standard deviation of 0.0384. The low variance across folds showcase that learning on different folds is relatively consistent. The 0.2923 R^2 score on the held out test set, shows that, despite the relatively low size of the dataset, it can still learn to approximate about 30% of the important signals contained within the dataset, especially when considering only demographic information available, such as the gender and age of a given participant.

TABLE I: Performance of the emotion intensity regressor

Testing Phase	MSE	MAE	R^2 Score
Fold 1	0.0315	0.0907	0.1623
Fold 2	0.0302	0.0867	0.2504
Fold 3	0.0343	0.0911	0.2350
Test	0.0289	0.0829	0.2923

B. Interest Predictors

When we inspect Figure 2, we can see that the HGNN model outperforms all of its tabular counterparts in both accuracy and recall, with scores of 0.7010 and 0.9412 for each metric. The high recall showcases that it performs significantly better when it comes to predicting an interested user, which is reflected onto the F1-score 0.8153, the highest observed among the model selection. Given this, in comparison to the XGBoost, LightGBM, Random Forest, and MLP-based models, a noticeable tradeoff in precision. Whereas majority of the model selection yields a precision of approximately 0.76, the HGNN model yields 0.7191. Although this tradeoff is reflected as a higher F1-score and accuracy, it still showcases that the HGNN model falls short when it comes to predicting negative samples, indicating a biased learning process.

V. CONCLUSION

The study proposed a novel multi-modal data fusion approach and a hierarchical deep learning pipeline that integrates participant demographics, advertisement media content, and biosignals into a heterogeneous graph representation for predicting user interest. The proposed approach aims to close the gap between academic neuromarketing studies and production-ready environments, paving the way for deployment-ready predictive neuromarketing solutions. The experimental results on the NeuroBioSense dataset showcases that the emotion intensity predictor can learn to approximate meaningful, emotional responses from demographic and advertisement data, enabling the use of these signals in downstream tasks. Furthermore, the proposed hybrid HGNN architecture showcases considerable predictive performance given the constraints of the dataset size. Overall, this study provides a valuable proof of concept

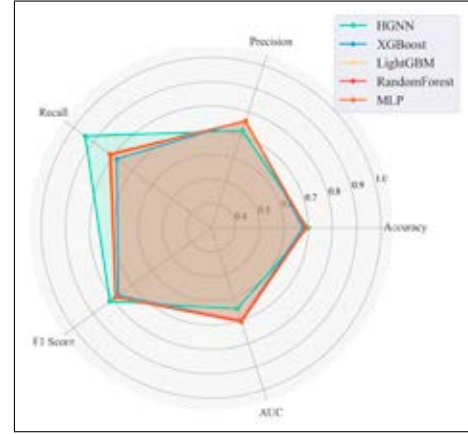


Fig. 2: Comparison of model performance across HGNN and baseline methods

for the use of heterogeneous graph learning and multi-modal data fusion to advance neuromarketing and consumer behavior prediction.

VI. ACKNOWLEDGEMENTS

This research is supported by Istanbul Kültür University under ULEP-2025-2026.

REFERENCES

- [1] S. Saban and E. Dağdevir, "Biomedical signal processing methods for neuromarketing: A comparative study," in *5th International Conference on Applied Engineering and Natural Sciences*, 2023.
- [2] B. Kocaçınar, P. Inan, E. N. Zamur, B. Çalışmşek, F. P. Akbulut, and C. Catal, "Neurobiosense: A multidimensional dataset for neuromarketing analysis," *Data in Brief*, vol. 53, p. 110235, 2024.
- [3] M. Alimardani and M. Kaba, "Deep learning for neuromarketing: classification of user preference using eeg signals," in *12th Augmented human international conference*, 2021, pp. 1–7.
- [4] A. Ullah, G. Baloch, A. Ahmed, A. B. Buriro, A. Junaid, B. Ahmed, and S. Akhtar, "Neuromarketing solutions based on eeg signal analysis using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.
- [5] S. Derdiyok and F. P. Akbulut, "Biosignal based emotion-oriented video summarization," *Multimedia Systems*, vol. 29, no. 3, pp. 1513–1526, 2023.
- [6] F. P. Akbulut, "Evaluating the effects of the autonomic nervous system and sympathetic activity on emotional states," *Istanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, vol. 21, no. 41, pp. 156–169, 2022.
- [7] A. C. M. d. Silva, M. P. S. Gôlo, and R. M. Marcacini, "Unsupervised heterogeneous graph neural network for hit song prediction through one class learning," *Anais*, 2022.
- [8] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.
- [9] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in neural information processing systems*, vol. 33, pp. 5776–5788, 2020.
- [10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
- [11] X. Li, Z. Xia, and H. Zhang, "Cauchy activation function and xnet," *Neural Networks*, vol. 188, p. 107375, 2025.
- [12] X. Li, Z. J. Xia, and X. Zheng, "Model comparisons: Xnet outperforms kan," *arXiv preprint arXiv:2410.02033*, 2024.
- [13] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" *arXiv preprint arXiv:2105.14491*, 2021.