Leveraging Multimodal Shapley Values to Address Multimodal Collapse and Improve Fine-grained E-Commerce Product Classification

1st Ajibola Obayemi
School of Architecture, Technology and Engineering
University of Brighton
Brighton, UK
e.a.obayemi@brighton.ac.uk

2nd Khuong Nguyen
Department of Computer Science
Royal Holloway, University of London
London, UK
khuong.nguyen@rhul.ac.uk

***Abstract*—Multimodal models can experience multimodal collapse, leading to sub-optimal performance on tasks like fine-grained e-commerce product classification. To address this, we introduce an approach that leverages multimodal Shapley values (MM-SHAP) to quantify the individual contributions of each modality to the model's predictions. By employing weighted stacked ensembles of unimodal and multimodal models, with weights derived from these Shapley values (MM-SHAP), we enhance the overall performance and mitigate the effects of multimodal collapse. Using this approach we improve previous results(f1-score) from 0.67 to 0.79.**

***Index Terms*—Multimodal learning, Weighted ensemble, MM-SHAP, Shapley values, Multimodal collapse, Fine-grained classification.**

I. Introduction

Multimodal classification is a machine learning technique which uses multiple data modalities such as image, text, audio, and spatial data to create an all-encompassing model which can be used for downstream tasks such as e-commerce product classification [1]–[3]. Multimodal classification has become a critical method used in e-commerce product classification of grocery items (eg: Food & Beverages with text labels and descriptions) and office supplies (Printer Ink Cartridges in this instance) and related industries where the representative data comes in the form of multiple modalities [5]–[7].

Multimodal classification has proved to be very effective in improving the results of fine-grained product classification tasks. Still, it can experience multimodal collapse which degrades the results of multimodal models and inadvertently affects its performance [9], [15], [16], [22].

Ensemble methods aim to improve the performance of machine learning models by combining multiple models using one of these: averaging, boosting, stacking, majority voting or Bayesian model averaging (BMA) [4], [10], [23], [25].

We focus on the ensemble approach in this paper and propose a different way of improving the performance and result of multimodal models by creating a weighted ensemble of unimodal and multimodal models which uses the calculated shapley values of each modality in the multimodal model as weights and training parameters for a meta-model.

The main contributions of this paper are:

1) We introduce a method of combining unimodal and multimodal models using multimodal Shapley values (MM-SHAP) as weights in the weighted stacking ensemble method.
2) We create a meta-model and demonstrate its performance on image and text multimodal datasets (publicly available at https://github.com/multimodal-research/TAIMD-17k).
3) We demonstrate how multimodal Shapley values (MM-SHAP) can improve multimodal models used for fine-grained product classification.

II. Related works

In the research work of [24], a Shapley value based methods called "Shap-CA" was introduced. Their work discussed an approach which enables both context-text and context-image pairs alignment. By leveraging the Shapley value concept, it quantifies the individual contribution of each element within the set of contexts, texts, and images to the overall semantic and modality overlaps. This quantitative evaluation is followed by a contrastive learning strategy that aims to enhance the interactive contribution within context-text/image pairs while minimising the influence across these pairs. To further refine the alignment process, an adaptive fusion module is designed to selectively combine information from different modalities, ensuring that only relevant cross-modal interactions are considered. Their work significantly differs from ours, as it is estimates the contributions differently and focuses on Shapley Value-based contrastive alignment.

In another related work by [26], multimodal contributions are observed and an approach to evaluate individual contributions per samples using sample-level modality valuation metric was introduced. This method was analysed to enhance the discriminative capabilities of low-contributing modalities at the sample level. Their approach essentially involved improving mulitmodal cooperation by evaluating unimodal contributions in the multimodal model.

1946-0759/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLA61862.2024.00244

In agreement to the issue of multimodal collapse we had previously discussed, the work of [8] discusses the problem which arises from imbalanced multimodal learning where multimodal models find it difficult to jointly and correctly utilise each modalities leading to over-reliance on one of the modality and degrading multimodal learning performance. They proposed a new method called "MMPareto"" which aims to reduce the imbalanced multimodal learning problem.

SHapley vAlue based PErceptual (SHAPE) was proposed by [27] to quantify the marginal contribution of individual modalities and the degree of cooperation across modalities in multimodal models. Using different multimodal datasets on various tasks, the authors find that multimodal models often ignore the cooperation across modalities, especially when one modality dominates. However, when different modalities are indispensable, the models learn to exploit cross-modal cooperation, and early fusion is beneficial.

III. Methods

Weighted ensemble methods are machine learning methods which combine multiple pre-trained models. A crucial distinction between weighted ensembles and other ensemble approaches lies in the introduction of weights [11]. In our case, we create an ensemble of fine-tuned unimodal models and multimodal models from previous experiments to compute the weight elements using the MM-SHAP to determine the modality contribution of our multimodal model.

In this section, we deconstruct the technical aspects of our weighted ensemble approach. We introduce the ensemble architecture, specifically the algorithm employed to combine predictions from the unimodal and multimodal models. We then discuss the selection and integration of pre-trained models, including their modalities (text or image) and the justification for their choice. Finally, we explain the weight computation technique that leverages multimodal Shapley (MM-SHAP) values to determine the contribution of each modality. Figure 1 visually complements our technical explanations.

A. Fine-tuned/ Pre-trained models

Building upon some of the established performance of ensemble models in prior research which capitalises on the strengths of individual models, leading to improved generalisation and robustness compared to relying on a single model [12], we propose to leverage this approach to create a better and more robust learner. We aim to combine the unimodal and multimodal models into an ensemble using the proposed weighted ensemble architecture.

For our experimental evaluation, we utilised the pre-trained checkpoints of the top-performing models from our previous studies on multimodal and image-only fine-grained product classification. These models, respectively, employed a multimodal architecture, a transformer (BERT) and a CNN (ResNet) based approach.

To generate the required inputs for the meta-model, we retrieved the logits from previously trained models. We provide further details about the steps we took below:

1) Multimodal model logits
   We load a checkpoint from our CLIP and MultiModal BiTransformers (MMBT) based multimodal model. This checkpoint essentially captures the model's state at a specific point during training. We then pass the evaluation set through this model. During the forward pass, the model makes computations on the input data and generates logits, which are the raw outputs before the final classification layer applies a function (like softmax) to convert them into probabilities.
2) Shapley values from the multimodal model
   Shapley values are a way to quantify the contribution of each feature or data modality to a model's prediction. In our case, we are interested in understanding how much the text data and the image data each contribute to the predictions made by our multimodal model. To achieve this, we calculate Shapley values using the equation 2. This formula considers all possible combinations of features (text only, image only, and both together) and analyses how each feature permutation influences the model's predictions. The resulting Shapley values represent the proportional contribution of each modality to the final prediction.
3) Stacking Unimodal and Multimodal model logits
   The logits from all models (both unimodal and multimodal) and the Shapley values from the multimodal model are then fed into the meta-model. Additionally, we incorporate an additional weight parameter within the meta-model. This weight is directly set to the computed Shapley value, allowing the meta-model to consider the relative importance of text and image information during the final decision-making process. Essentially, the meta-model can learn to pay more attention to the modality (text or image) that has a historically greater influence on the multimodal model's predictions according to the Shapley value.

B. SHAP (Shapley Values)

The concept of Shapley values was first introduced in the work of Lloyd Shapley in 1953, where a proposal was made for a novel method to quantify the individual contribution of each player in a collaborative game [13]. This background was expanded over the years and applied across several fields, including machine learning. Theoretically, shapley values can be used to assign a value to each player in a collaborative game using an estimation of their impact on the overall outcome which can be achieved by all collaborative players in a group [14].

More recently, in the efforts to improve the interpretability of machine learning models, SHapley Additive exPlanations (SHAP) have emerged as a model-agnostic way of interpreting how a machine learning model works. This method is built upon the theoretical foundation of SHapley's work in 1953 as explained in the paragraph above. SHAP leverages core game theory concepts which determine the fair distribution of payoffs or contributions amongst players in a cooperative game.

Essentially, it quantifies the individual contribution of each feature within a model to its overall prediction or performance. Compared to other methods such as LIME: Local Interpretable Model-Agnostic Explanations, SHAP works for more complex models and provides a better-performing framework when it comes to the interpretability of machine learning models [17], [18].

As expressed in Equation (1), the Shapley value for a player represents the average of its marginal contributions to the value of all possible predecessor sets. This average is weighted based on the number of players in each predecessor set.

$$\phi_i = \sum_{S \subseteq M \setminus \{i\}} \underbrace{\rho(m, |S|)}_{\text{weight factor}} \underbrace{(v(S \cup \{i\}) - v(S))}_{\text{marginal contribution}} \tag{1}$$

SHapley Additive exPlanations (SHAP) is expressed in Equation (2) below. Where $g(z')$ denotes the explanation model and simplified features (or coalition vector) respectively. $\phi_j z_j$ denotes the feature attribution for feature j and z' describes coalitions in the coalition vector.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2}$$

Our exploration of SHapley Additive exPlanations (SHAP) in this section establishes a theoretical foundation for the subsequent exploration of MM-SHAP. It is essential to discuss this foundation as MM-SHAP directly leverages the fundamental principles of Shapley values/ SHAP and this is comprehensively discussed in the next section.

C. MM-SHAP

MM-SHAP, a performance agnostic method of estimating the multimodality score of multimodal models based on Shapley values, can be used to quantify the proportion of individual contribution of each modality in multimodal models [19]. MM-SHAP provides a way to measure each modality's influences on the model's predictions, regardless of the model's overall accuracy.

The MM-SHAP method, introduced by [19], was specifically designed for analysing vision (image) and language (text) multimodal models, which aligns well with the focus of our work. This shared foundation in the analysis of individual contribution in image-text modality models makes their research particularly relevant to our solution exploration. Additionally, they discussed 3 ways of applying MM-SHAP in their work as listed below:

1) Sample-level
2) Dataset and model level
3) Measuring fine-tuning effects

Building on the work presented in [19], we can utilise MM-SHAP to quantify the impact of each modality within a multimodal model. Equations 3 and 4 define the calculations for text and image contributions, respectively.

$$\Phi_T = \sum_{j}^{PT} |\phi_j| \tag{3}$$

$$\Phi_I = \sum_{j}^{PI} |\phi_j| \tag{4}$$

where Φ_T & Φ_I represent the percentage of the final prediction that can be attributed to the textual and image modalities, respectively. These values are calculated by summing the absolute Shapley values for each modality.

Based on the established foundation, we calculate the individual contributions of each modality in our multimodal model. These contributions are then used to assign weights to the different components within the meta-model of our weighted ensemble system, as depicted in Figure 1. The estimated multimodal contribution can be positive, indicating they enhance the model's prediction, negative, meaning they weaken it, or zero, signifying no significant impact [20].

D. Meta Model

In ensemble-based architectures, meta-models are typically trained on the raw outputs from the base learners (in our case, image unimodal, text unimodal, and image-text multimodal) along with their corresponding true targets [21]. This training process allows the meta-model to implicitly learn how to best combine these individual predictions for a more robust final output. However, unlike traditional approaches, we have opted to introduce a different approach by incorporating an explicit weight parameter into our meta-model architecture. This weight parameter goes beyond the implicit weighting learned through standard meta-model training. Instead, it allows the model to directly assign importance scores based on each modality (text or image). We leverage the power of MM-SHAP (explained in Section III-C) to estimate these weights effectively. By incorporating this explicit weighting mechanism, we aim to empower the meta-model to make a more nuanced assessment of each base learner's contribution, ultimately leading to a potentially more accurate and reliable final prediction.

1) Weights
   Our proposed weighted stacked ensemble incorporates a novel element: a weight parameter. This tailored weighting formally expressed in Algorithm 2, empowers the model to effectively leverage the strengths of individual modality contributions in the base learners.
2) Architecture
   To provide a comprehensive understanding of our proposed architecture, we can break it down into five core components followed by a dedicated prediction layer, as seen in Figure 1. We also present a detailed step-by-step breakdown of the entire process in Algorithm 2.

1: **Input:** Training data D, validation data D_v, testing data D_t
2: **Output:** Ensemble model prediction $\hat{y}$
3: Split D into training set D_{tr} and validation set D_v
4: **for** $l \in \{1, \ldots, L\}$ **do**
5: Train base learner l on D_{tr} (e.g., unimodal and multimodal models)

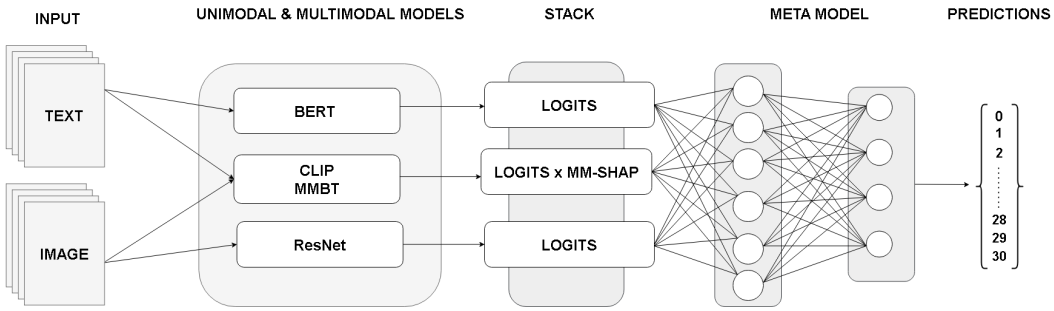

Fig. 1. An architecture depicting the weighted ensemble which combines unimodal and multimodal models, using MM-SHAP to calculate modality contributions.

end for
for $i \in D_v$ **do**
  for $l \in \{1, \ldots, L\}$ **do**
    Generate meta-feature z_{li} from base learner l prediction on i
  end for
end for
Train stacking model (meta-learner) M_s on $\{z_i, y_i\}_{i \in D_v}$ (e.g., feed-forward networks)
Determine weights w_l for base learners (using MM-SHAP to estimate weights based on multimodal contributions)
for $i \in D_t$ **do**
  $\hat{z}_i = \sum_{l=1}^{L} w_l \cdot$ predict(l, i)
  (classification) $\hat{y} = \text{argmax}_k P(y = k|\hat{z}_i)$
end for
Return: $\hat{y}$

3) Training
We trained the weighted stacked ensemble on two NVIDIA RTX 4090 GPUs, each equipped with a dedicated 48GB of memory. Details of the configuration and hyperparameters can be found in Table I.

TABLE I
CONFIGURATION AND HYPER-PARAMETER FOR THE WEIGHTED STACKED ENSEMBLE

Feature	Value
Image encoder	CLIP
Image encoder size	288
Number of image embeddings	4
Text encoder	BERT
Token Sequence Length	120
Loss function	Cross Entropy
Optimiser	SGD
Learning rate	1e-4
Input dimension	93
Hidden dimension	512
Output dimension	31
Epoch	50,000

IV. EXPERIMENTS

In this section, we discuss the experimental setup designed to evaluate the performance of the weighted ensemble methods expounded in Section III. As mentioned previously, we performed all experiments using the "TAIMD-17k" text and image multimodal dataset. This dataset is a suitable representation of our problem domain, as it is designed for fine-grained classification and contains both image and text components, making it well-suited for multimodal learning applications.

Furthermore, we incorporate the previously trained unimodal models, which extract crucial features from either the textual descriptions (text-based) or the corresponding product images (image-based). Additionally, the image-text multimodal models, trained to exploit the inherent relationship between these modalities, are integrated into the ensemble construction process. To determine the relative influence of each model within the ensemble and guide the weight assignment process, the pre-computed multimodal Shapley values, which capture the marginal contribution of each model to the overall ensemble performance are utilised.

Finally, a meta-model, a higher-order model trained on the stacked logits and MM-SHAP values of the unimodal and multimodal models within the ensemble, will be introduced. This meta-model plays a pivotal role in our approach, serving as a final classification layer or a mechanism for feature extraction for the weighted ensemble. By analysing the performance of the weighted ensemble on our dataset, we aim to assess its ability to outperform the constituent unimodal and multimodal models across various performance metrics (f1-score, recall & precision). The results are presented and critically evaluated to establish the viability and potential advantages of the weighted ensemble approach for fine-grained product classification tasks.

A. Unimodal models

To build upon the previous work, we employ the text and image unimodal models as the foundation for our weighted ensemble method. These pre-trained models act as our base learners. We leverage checkpoints from the prior experiments to generate prediction scores, which are then fed as inputs into the ensemble architecture described in Algorithm 2. For the

text unimodal model, we fine-tuned BERT and for the image unimodal model, we fine-tuned ResNet-152.

B. Multimodal model

For the multimodal model, we train an architecture based on MMBT and CLIP using the image and text dataset (TAIMD-17k). We utilise the checkpoints from this multimodal model to generate predictions which we use as inputs in the weighted ensemble method. A breakdown of the process is detailed below:

1) Pre-processing
   To prepare the image and text data for the multimodal model, we applied suitable preprocessing techniques. The text was tokenized using the BERT tokenizer, and images were divided into patches using CLIP. To capture a richer representation of the input, we generated embeddings for both the processed text and image data. Additionally, we create a dictionary containing the following key-value pairs: $input\_ids$, $input\_modal$, $attention\_mask$, $start\_tokens$ and end_tokens.
2) Load checkpoint
   We load a checkpoint file containing the trained weights and biases from the multimodal model. This process essentially restores the model's learned parameters.
3) Generate Prediction
   Once the preprocessed data and checkpoint are loaded, we activate the multimodal model. The processed image and text data are fed through the model's architecture, triggering computations across its interconnected neural layers. This activation process ultimately results in the final output layer generating a set of prediction scores. These scores represent the model's confidence level for each possible category for the input. These scores then become the new input for the meta-model, which will be discussed in the following section.

C. Weighted ensemble model

Our approach utilises a meta-model architecture, detailed in Figure 1. This model was extensively trained for $50,000$ epochs using the specific configuration and hyper-parameters outlined in Table I. The meta-model plays a critical role in constructing the weighted ensemble.

V. Result

Having conducted the experiments outlined previously, this section discusses an evaluation of the derived results. We establish comparisons between our findings and those of prior experiments, utilising the same established evaluation metrics: F1-score, precision, and recall. To facilitate a comprehensive understanding of the results, visual summaries are presented in Figures 2 & 3. Additionally, we present our final results in Table II, comparing the performance of the unimodal, multimodal and weighted ensemble methods.

TABLE II
An Evaluation of the Performance of Unimodal, Multimodal, and Weighted Ensemble Approaches.

Method	Precision	Recall	F1-Score
Unimodal (ResNet-152)	0.70	0.61	0.59
Multimodal (MMBT + CLIP)	0.75	0.71	0.67
Weighted Ensemble	**0.80**	**0.82**	**0.79**

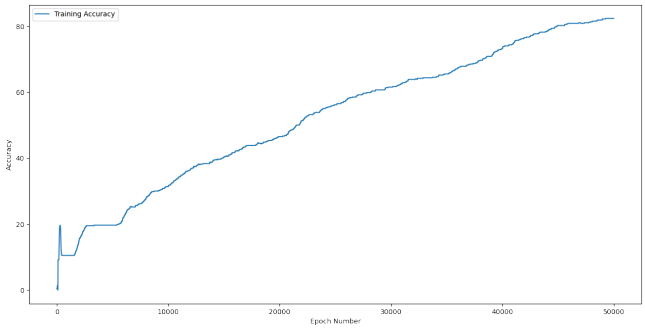

Fig. 2. This graph illustrates the model's learning progress over time. The x-axis represents the number of training epochs, while the y-axis indicates the percentage of correctly classified training samples. The curve exhibits an upward trend, suggesting that the model is learning effectively and becoming more accurate.

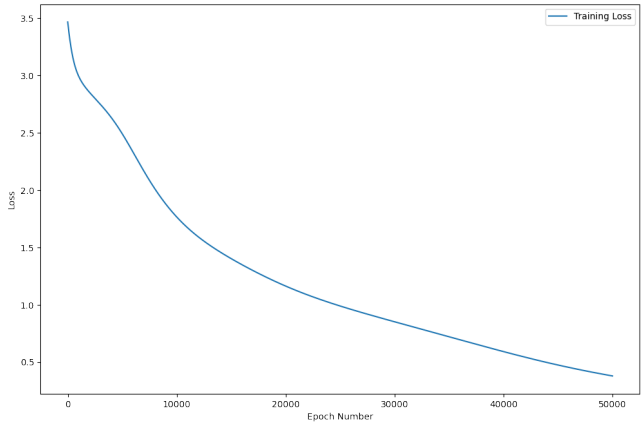

Fig. 3. The x-axis typically represents the number of training epochs, while the y-axis indicates the calculated loss value. The curve exhibits a downward trend which suggests that the model is learning effectively and minimising its errors.

VI. Conclusion & Future Work

This research introduces a new approach which addresses multimodal collapse and improves multimodal product classification through the use of a weighted ensemble model using multimodal shapley values. This model excels by combining the strengths of individual unimodal (image-only or text-only) models and multimodal models that consider both data types (image and text). The key lies in MM-SHAP values, which quantify the contribution of each data modality (image and text) to the classification process. By leveraging these values to calculate weights for each model, we create a more informed ensemble. Furthermore, the MM-SHAP values themselves

are incorporated as additional features during the training of a meta-model, further enhancing its ability to distinguish between products.

Our experiments on the TAIMD-17k multimodal product datasets containing both images and text demonstrate significant improvements in F1-score, recall & precision compared to other methods (text based unimodal, image based unimodal and text and image multimodal models).

It is crucial to acknowledge the computational cost associated with this weighted ensemble approach. Calculating MM-SHAP (Multimodal SHapley Additive exPlanations) values and training multimodal models can be computationally intensive, requiring substantial processing power and memory resources. To address this challenge, future research should focus on efficiently optimising the method. Here are some potential avenues to explore:

1) Approximation Algorithms
   We can investigate incorporating approximation algorithms to achieve a balance between accuracy and computational efficiency. These algorithms might provide an acceptable level of accuracy while significantly reducing processing time.
2) Parallelisation
   Leveraging parallel computing techniques can significantly accelerate the computation of MM-SHAP values. By distributing the workload across multiple processors or GPUs, we can achieve faster processing times without compromising accuracy. This would require restructuring the code to be compatible with parallel execution environments.

By focusing on these optimisation techniques, we can mitigate the computational bottleneck associated with this weighted ensemble approach while maintaining the gains in classification accuracy. This will ultimately lead to a more practical and scalable solution for real-world applications and adoption.

REFERENCES

[1] Bi, Ye, Shuo Wang, and Zhongrui Fan. "A multimodal late fusion model for e-commerce product classification." arXiv preprint arXiv:2008.06179 (2020).

[2] Zhang, Heng, Vishal M. Patel, and Rama Chellappa. "Hierarchical multimodal metric learning for multimodal classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[3] Morency, Louis-Philippe, and Tadas Baltrušaitis. "Multimodal machine learning: integrating language, vision and speech." Proceedings of the 55th annual meeting of the association for computational linguistics: Tutorial abstracts. 2017.

[4] Mohammed, Ammar, and Rania Kora. "A comprehensive review on ensemble deep learning: Opportunities and challenges." Journal of King Saud University-Computer and Information Sciences 35.2 (2023): 757-774.

[5] Mehta, Karan, et al. "Multimodal Classification in E-Commerce: A Systematic." (2022).

[6] Ramachandram, Dhanesh, and Graham W. Taylor. "Deep multimodal learning: A survey on recent advances and trends." IEEE signal processing magazine 34.6 (2017): 96-108.

[7] Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." IEEE transactions on pattern analysis and machine intelligence 41.2 (2018): 423-443.

[8] Wei, Yake, and Di Hu. "MMPareto: Boosting Multimodal Learning with Innocent Unimodal Assistance." arXiv preprint arXiv:2405.17730 (2024).

[9] Han, Zongbo, et al. "Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[10] Shah, Shariq, et al. "An ensemble-learning-based technique for bimodal sentiment analysis." Big Data and Cognitive Computing 7.2 (2023): 85.

[11] Zhao, Jiakun, et al. "A weighted hybrid ensemble method for classifying imbalanced data." Knowledge-based systems 203 (2020): 106087.

[12] Xu, Qin, et al. "Fine-grained visual classification via internal ensemble learning transformer." IEEE Transactions on Multimedia 25 (2023): 9015-9028.

[13] Shapley, Lloyd S. "A value for n-person games." Contributions to the Theory of Games 2 (1953).

[14] Li, Meng, et al. "Explaining a machine-learning lane change model with maximum entropy Shapley values." IEEE Transactions on Intelligent Vehicles 8.6 (2023): 3620-3628.

[15] Javaloy, Adrián, Maryam Meghdadi, and Isabel Valera. "Mitigating modality collapse in multimodal VAEs via impartial optimization." International Conference on Machine Learning. PMLR, 2022.

[16] Hemker, Konstantin, Nikola Simidjievski, and Mateja Jamnik. "MM-Lego: Modular Biomedical Multimodal Models with Minimal Fine-Tuning." arXiv preprint arXiv:2405.19950 (2024).

[17] Lundberg, Scott. "A unified approach to interpreting model predictions." arXiv preprint arXiv:1705.07874 (2017).

[18] Molnar, Christoph. Interpretable machine learning. Lulu. com, 2020.

[19] Parcalabescu, Letitia, and Anette Frank. "Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks." arXiv preprint arXiv:2212.08158 (2022).

[20] Aggarwal, Piush, et al. "Text or Image? What is More Important in Cross-Domain Generalization Capabilities of Hate Meme Detection Models?." arXiv preprint arXiv:2402.04967 (2024).

[21] Awang, Mohd Khalid, et al. "Improving customer churn classification with ensemble stacking method." International Journal of Advanced Computer Science and Applications 12.11 (2021).

[22] Zheng, Xiao, et al. "Multi-level confidence learning for trustworthy multimodal classification." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 9. 2023.

[23] Azri, Abderrazek, et al. "Rumor classification through a multimodal fusion framework and ensemble learning." Information Systems Frontiers 25.5 (2023): 1795-1810.

[24] Luo, Wen, et al. "Shapley Value-based Contrastive Alignment for Multimodal Information Extraction." arXiv preprint arXiv:2407.17854 (2024).

[25] Liu, Zhicheng, et al. "Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion." Companion Proceedings of the ACM on Web Conference 2024. 2024.

[26] Wei, Yake, et al. "Enhancing multimodal cooperation via sample-level modality valuation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[27] Hu, Pengbo, Xingyu Li, and Yi Zhou. "Shape: An unified approach to evaluate the contribution and cooperation of individual modalities." arXiv preprint arXiv:2205.00302 (2022).