Multi-modal Machine Learning for Vehicle Rating Predictions Using Image, Text, and Parametric Data

Hanqi Su[1]*, Binyang Song[1], Faez Ahmed[1]

[1]Massachusetts Institute of Technology, Department of Mechanical Engineering, Cambridge, MA

Abstract

Accurate vehicle rating prediction can facilitate designing and configuring good vehicles. This prediction allows vehicle designers and manufacturers to optimize and improve their designs in a timely manner, enhance their product performance, and effectively attract consumers. However, most of the existing data-driven methods rely on data from a single mode, e.g., text, image, or parametric data, which results in a limited and incomplete exploration of the available information. These methods lack comprehensive analyses and exploration of data from multiple modes, which probably leads to inaccurate conclusions and hinders progress in this field. To overcome this limitation, we propose a multi-modal learning model for more comprehensive and accurate vehicle rating predictions. Specifically, the model simultaneously learns features from the parametric specifications, text descriptions, and images of vehicles to predict five vehicle rating scores, including the total score, critics score, performance score, safety score, and interior score. We compare the multi-modal learning model to the corresponding unimodal models and find that the multi-modal model's explanatory power is 4% - 12% higher than that of the unimodal models. On this basis, we conduct sensitivity analyses using SHAP to interpret our model and provide design and optimization directions to designers and manufacturers. Our study underscores the importance of the data-driven multi-modal learning approach for vehicle design, evaluation, and optimization. We have made the code publicly available at http://decode.mit.edu/projects/vehicleratings/.

Keywords: Multi-modal Learning, Machine Learning, Vehicle Rating Prediction, Model Interpretability, Sensitivity Analytsis

1. INTRODUCTION

From the earliest years of their invention, vehicles have stood as a major contributing factor to both everyday consumer life and global economic development. Since the availability of the internet, most consumers research vehicle evaluation scores online and see them as important references for their vehicle purchasing decisions [1]. Vehicle evaluation is likewise at the heart of vehicle design, optimization, and improvement. Effective and efficient vehicle evaluation is essential for designers and manufacturers to enhance the appeal of their new models. Extant research has shown promise in exploiting machine learning (ML) and artificial intelligence for vehicle price prediction [2–4], vehicle sales prediction [5], vehicle purchase criteria [6], vehicle evaluation [7], and insurance services[8]. When evaluating a vehicle, consumers typically analyze multiple data types, such as images, 3D models, parametric specifications, and text reviews.

Consider a typical vehicle purchasing journey. Initially, a potential vehicle buyer determines the need to purchase a vehicle, which leads them to explore various automotive websites, such as US News, to evaluate numerous vehicle options. In order to make a well-informed choice, they might scrutinize the vehicle's exterior and interior images to assess its design and features. They might even engage with 3D models, when accessible, for a more detailed understanding of the vehicle's attributes.

Additionally, the buyer might review parametric data to measure the vehicle's specifications against others in its category, focusing on elements such as engine capacity, fuel efficiency, safety features, and cost. Reading reviews and written summaries about the vehicle's performance, reliability, and user experience also aids them in ensuring it suits their requirements. Renowned entities, like US News, often rank or rate different vehicles, which can significantly influence the buyer's decision. Through the amalgamation of this varied information, buyers are able to make knowledgeable purchase decisions. Subsequently, they might visit a vehicle dealership to inspect and test drive their chosen vehicle. Upon assessing the vehicle's performance, comfort, and additional features, buyers determine whether it's the right fit for them. If satisfied, they return to the dealership to discuss the price and finalize the purchase. It's crucial to acknowledge that individuals tend to consider multiple data modalities when interacting with designs. However, the majority of current machine learning algorithms are focused on a single modality, typically images, which limits their perspective and hence their practicality. This single-dimensional approach inevitably results in oversimplified conclusions and findings.

This paper endeavors to bridge this gap by tackling the research question: How does multi-modal information about a vehicle influence its ratings? This question is approached utilizing a multi-modal learning method and interpretability mod-

*Corresponding author: hanqisu@mit.edu

els. The application of artificial intelligence and multi-modal deep learning to evaluate and analyze vehicles is relatively unexplored, predominantly due to the substantial requirement for labeled multi-modal data to train deep neural networks. To remedy this shortfall, we also collected a novel multi-modal dataset that includes parametric specifications, images, and textual descriptions of vehicles, all labeled with various vehicle assessment scores.

On this basis, we develop and validate a multi-modal learning model to predict the rating scores of vehicles more comprehensively and accurately. We show that multi-modal learning can exploit the features learned from different types of data and capture the interactions between them to achieve better performance than unimodal learning. Our contributions include the following:

1. We propose the development of individual unimodal ML models that independently learn from parametric specifications, images, and text descriptions of vehicles. These models aim to predict five distinct vehicle rating scores, namely the total score, critics score, performance score, safety score, and interior score.

2. We introduce a multi-modal learning model capable of concurrently learning from parametric, image, and text data to predict vehicle rating scores. Our findings indicate that this multi-modal learning model markedly outperforms the unimodal models.

3. We assess the relative informativeness of different data modes. Our analysis suggests that parametric data is the most informative for predicting all rating scores, and in most instances, text descriptions offer more predictive power than images.

4. We demonstrate that the sensitivity analyses using SHAP are capable of interpreting our models and providing more detailed design, optimization, and improvement directions to designers and companies.

The rest of this article is organized as follows: Section 2 reviews the approaches to the relevant components of the proposed model. Section 3 introduces the source and composition of the data used in this paper, the data processing module, and both the unimodal and multi-modal machine learning models. Section 4 reports and discusses the performances of the unimodal and multi-modal machine learning models, interprets the models through sensitivity analyses, and summarizes the limitations of this study. Section 5 concludes this paper by highlighting its findings and contributions.

2. BACKGROUND

Good vehicle evaluation often requires the analysis of multi-modal data, often involving vehicle parametric specifications, text descriptions, and images. In this section, we first discuss why vehicle evaluation is important. We then review relevant methods for embedding parametric, text, and image data, and investigate prior research on ML techniques for multi-modal data.

2.1 Why are vehicle evaluations important?

A few websites provide vehicle reviews and ratings, such as J.D. Power[1], US News[2], Motor Trend[3], Edmunds[4], and Kelley Blue Book[5]. Among these, US News is one of the most popular websites, showing as the number one search result for the query "vehicle rating" on search engines such as Google, Bing, and Duck.

US News vehicle ratings are highly influential and are widely followed by consumers who are in the market for a new vehicle. When a vehicle receives high ratings, it can receive increasing consumer interest, ultimately resulting in more sales. US News vehicle ratings consider various factors such as safety, reliability, performance, and interior features. These ratings are based on objective data and evaluations from automotive experts, which can provide consumers with a valuable reference for making informed decisions when purchasing a vehicle. Consumers may use these ratings as a guide when comparing different models and brands and may be more likely to consider a vehicle that has received high ratings. Similarly, vehicle dealerships may use these ratings in their advertising and marketing efforts to attract customers to their inventory.

Vehicle manufacturers can use US News vehicle ratings to improve their new vehicle designs in several ways:

1. Identify Areas for Improvement: By looking at the rating scores for factors like safety, reliability, performance, and interior features, vehicle manufacturers can use these ratings to identify areas where their new vehicles are falling short and improve their designs. US News vehicle ratings take into account consumer needs and preferences. By using the ratings to inform their new vehicle designs, vehicle manufacturers can create vehicles that can better meet the needs and preferences of their target customers.

2. Benchmark Against Competitors: Vehicle manufacturers can use vehicle ratings to see how their new vehicle designs compare to those of their competitors. This can help them identify areas where they need to improve to remain competitive in the market.

3. Incorporate Best Practices: Vehicle manufacturers can analyze the highest-ranked vehicles in their category to discover and incorporate best practices into their new vehicle designs. This can help them improve their ratings in future years.

In summary, by using vehicle ratings to inform their new vehicle designs, vehicle manufacturers can create vehicles that better meet the needs of their customers, are more competitive in the market, and ultimately achieve higher ratings in future years.

What are the benefits of predicting vehicle ratings using machine learning? Predicting vehicle ratings using ML can be incredibly useful for several reasons. By analyzing a vast amount

[1] https://www.jdpower.com/cars/rankings
[2] https://cars.usnews.com/cars-trucks/rankings
[3] https://www.motortrend.com/cars/
[4] https://www.edmunds.com/new-car-ratings/
[5] https://www.kbb.com/cars/

of data, ML algorithms can identify patterns and correlations across different vehicle data, that may not be immediately apparent to humans. This can help vehicle manufacturers gain valuable insights into the features and characteristics contributing to high ratings. By predicting ratings, vehicle manufacturers can identify areas for improvement in their products and make adjustments to enhance their performance in these areas. Predicting ratings can also help vehicle manufacturers remain competitive in the market by identifying trends and preferences among consumers, allowing them to create products that better meet the needs of their target audience.

Predicting vehicle ratings can also inform marketing and advertising strategies by highlighting the features that are most important to consumers. Additionally, it can help vehicle manufacturers identify areas for improvement in their vehicle designs and assess their performance relative to their competitors. By tracking their progress over time and setting internal targets for improvement, vehicle manufacturers can use predicted vehicle ratings as a benchmark to inform their product development and competitive strategies. Ultimately, predicting vehicle ratings using ML can help vehicle manufacturers create better products, improve their marketing and advertising strategies, and gain competitive advantages in the market. Next, we discuss different modalities of data in which vehicle information is typically captured.

2.2 Representing Engineering Data in Different Modalities

Parametric data Engineering product specifications are often provided in the form of tables in a structured way. Parametric data is one of the most commonly used forms of data, consisting of samples (rows) that share the same feature set (columns), which has been used in many applications [9]. Compared with image or text data, parametric data is mostly heterogeneous, consisting of continuous-valued and categorical-valued attributes. Parametric data features dense values but sparse classification. Although parametric data modeling has been explored intensively using traditional ML methods in the past decades, such as linear regression [10], the Gaussian process [11], and gradient-boosted decision trees (GBDT) [12], deep neural networks can learn parametric data in a gradient-based way and allow for the integration of parametric data with other data modalities for multi-modal learning. Typically, parametric data can be learned by simple neural networks, such as multi-layer perceptrons (MLPs). Prior studies have reported that regularization can improve the performance of MPLs in learning parametric data [13]. Deep learning techniques like attention mechanisms [14] and transformer [15] architectures have also been applied to parametric data learning and have shown good prospects.

Image data With the recent advances of deep learning in computer vision, convolutional neural networks (CNNs) have made breakthroughs in image recognition [16], image classification [17], image segmentation [18], image generation [19], and other applications. Therefore, we focus on CNNs for image learning. A few pre-trained image embedding modules are commonly used for image learning tasks, including AlexNet [20], VGGNet [21], ResNet [22], and Inception [23]. Although current image learning for prediction tasks mostly focuses on classification and recognition, this study particularly focuses on the prediction of vehicle rating scores, which is essentially a regression problem. Different from classification problems, the features learned by the image embedding modules are not used to predict a categorical class through the Softmax activation function but are employed to predict continuous values (i.e., vehicle rating scores) through the Rectified Linear Unit (ReLU) activation function.

Text data In addition, natural language processing (NLP) has made significant strides toward automatic comprehension of text data. A few neural network language models (NNLMs) [24] first appeared to learn massive text data. Then, deep recurrent neural networks (RNNs) [25] brought NLP to the next level with their strengths in learning sequential data. Its variants, such as long short-term memory (LSTM) [26] or gated recurrent unit (GRU) [27], were proposed to resolve the problems of gradient vanishing and the explosion of RNNs. Recently, with the advent of the transformer models [15], large Transformer-based language models have gradually gained prominence in fulfilling various NLP tasks. These models have many advantages over the previous NNLMs and RNN-based models: taking entire sequences as input, they can understand the context of each word in a sequence more comprehensively; transformers can process and train more data in less time and utilize the embedded self-attention mechanism to enhance learning. Deep learning models, such as the generative pre-trained transformers (GPT) models [28, 29] proposed and constantly updated by the OpenAI team, and the bidirectional encoder representations from transformers (BERT) model [30], and its variants [31, 32], significantly improve NLP tasks. In this study, we use a BERT model to encode text data and predict different rating scores.

2.3 Multi-modal Learning

On the vehicle rating websites, each vehicle is represented in multiple data modes. Capturing the complementarity and alignment of multi-modal data can lead to a better understanding and more accurate evaluation of a vehicle. Multi-modal learning models that can learn vehicle features simultaneously from the multi-modal information are required to predict vehicle rating scores using such information. In multi-modal learning, the unimodal models are often pre-trained to learn features from each data modality first. On this basis, the multi-modal model can be constructed by fusing the features learned by multiple unimodal models for the downstream tasks. In this paper, we focus on employing multi-modal learning to learn vehicle images, text descriptions, and parametric specifications to predict vehicle rating scores.

Obtaining multi-modal latent representations with effective information fusion lies at the heart of multi-modal learning for this prediction task. Joint representations and coordinated representations are the common options to represent multi-modal data [33]. Joint representation is better at capturing complementary information from different modalities compared to coordinated representations [33, 34], making it more suitable for prediction tasks [35]. Fusing the information from multiple modalities effectively is critical to learn informative joint representations. Operation-based methods, bilinear pooling methods, and

attention-based methods are commonly used for information fusion in multi-modal learning [33]. The operation-based methods integrate features learned from unimodal data using simple operations, such as concatenation [35–38], averaging [39], element-wise multiplication [40], (weighted) summation [37, 38], linear combination [37], and majority voting [41]. Bilinear pooling fusion integrates features learned from unimodal data by calculating their outer product or Kronecker product [42, 43]. This approach can capture the high-order multiplicative interactions among all modalities, leading to more expressive and predictive multi-modal representations for fine-grained recognition [42, 44]. In comparison, the attention-based methods can model dependencies between two data modalities dynamically and assign higher weights to the elements more relevant to the other modality [15, 45]. The integration of the features learned from different modalities can be joined at early or late stages. It is easier to learn the interactions between different data modalities when the features are joined at early stages. However, early joining results in higher-dimensional joint representations, which need more computational resources to train [33].

In recent years, multi-modal learning has been explored for a variety of tasks, such as cross-modal synthesis [46–49], multi-modal prediction [50], and cross-modal information retrieval [51, 52]. However, it is still underexplored in the engineering domain. Recently, Yuan, Mation, and Moghaddam [53] proposed a multi-modal learning model to capture features from images and text for shoe evaluation. Li et al. [54] developed a multi-modal target embedding variational autoencoder model for 2D silhouette-to-3D shape translation. Song et al. [55] developed an attention-enhanced multi-modal learning model that learns design sketches and text descriptions simultaneously for design metric prediction.

2.4 Machine Learning Model Interpretability

In the realm of engineering, there is a growing emphasis not only on the effectiveness and predictive capabilities of ML models but also on their interpretability [56–58], which indicates if the reasoning process behind the model predictions can be easily comprehended by humans. The greater the interpretability of a model, the more readily people can understand and trust its predictions.

The rapid advancement of deep learning models has facilitated diverse model-independent explanation techniques. For instance, permutation feature importance [59–61] and Shapley Additive exPlanations (SHAP) [62–64] have seen widespread applications. Fisher et al. devised the model class reliance (MCR) approach to facilitate the comprehension of Variable Importance (VI) for unknown models [59]. Within Engineering Design, Ahmed et al. [61] employed a feature permutation-based technique to interpret the predictions of a graph neural network in predicting product relationships. They found factors such as car make, body type, and segment were important for determining co-consideration relationships. Mukund Sundararajan and Amir Najmi [63] proposed Baseline Shapley (BShap), a technique to explore differences in Shapley value attribution across multiple operations. Shrikumar et al. [62] developed Deep Learning Important FeaTures (DeepLIFT), a method that analyzes the contribution of neurons to input in backpropagation networks to ascertain feature importance. And DeepExplainer, an implementation of Deep SHAP, was developed based on SHAP [64] and DeepLIFT [62]. Additionally, Integrated Gradients [65], in conjunction with SHAP [64] and SmoothGrad [66], has given rise to the GradientExplainer, which is a variant of the SHAP Explainer. This variant enables the interpretation of image or text model outputs. In our study, we deploy a SHAP-based approach [64] to interpret the outputs of the image, text, and parametric models.

3. DATA AND METHOD

This section introduces the different types of data we use and the multi-modal learning models for vehicle rating score prediction in this research. This prediction problem is viewed as a regression task. The multi-modal learning model can be divided into five modules, as shown in Figure 1. The first module is a pre-processing module to prepare the original data (parametric vehicle specifications, images, and text descriptions). The processed data are the input to the respective unimodal models. The second, third, and fourth modules are the unimodal models capturing features from the parametric, image, or text data, respectively. After the three unimodal models are pre-trained, they can be combined to construct the multi-modal learning model. The rest of this section will separately introduce the data used in this study and each module of the proposed multi-modal learning model.

3.1 Data

The data for developing the multi-modal learning model comes from U.S.News[6]. The website provides detailed information on vehicles from different categories, such as sedans, trucks, vans, and sport utility vehicles. The available information covers expert reviews, photos, prices, specifications, performances, rating scores, and so on. In this study, the rating scores, including the overall score, the performance score, the interior score, the critics score, and the safety score, are used as the labels of each vehicle described by the other information. Among them, the performance score reflects the vehicle's performance in terms of acceleration, braking, ride quality, handling, and other qualitative performance metrics. The interior score is regarding vehicle interior manufacturing quality, interior comfort, decoration and features, cargo space, and styling. The critics score represents the reviewer's degree of recommendation and their overall tone regarding the vehicle. The safety score is based on two factors: the number of advanced accident-avoidance technologies and crash test results from the National Highway Traffic Safety Administration and the Insurance Institute for Highway Safety. The overall score for each vehicle is the weighted average of the other four component scores and a few other factors, which are not available on the US News website and are not considered in our study. The rating scores range from 0 to 10, with 10 being the best.

The goal of our multi-modal regression model is to predict the five scores for new vehicle designs, given their specifications, images, and text descriptions. The text description provides an overall review of the vehicle, including its advantages and disadvantages, and changes to the vehicle model compare to its last

[6] https://cars.usnews.com/cars-trucks

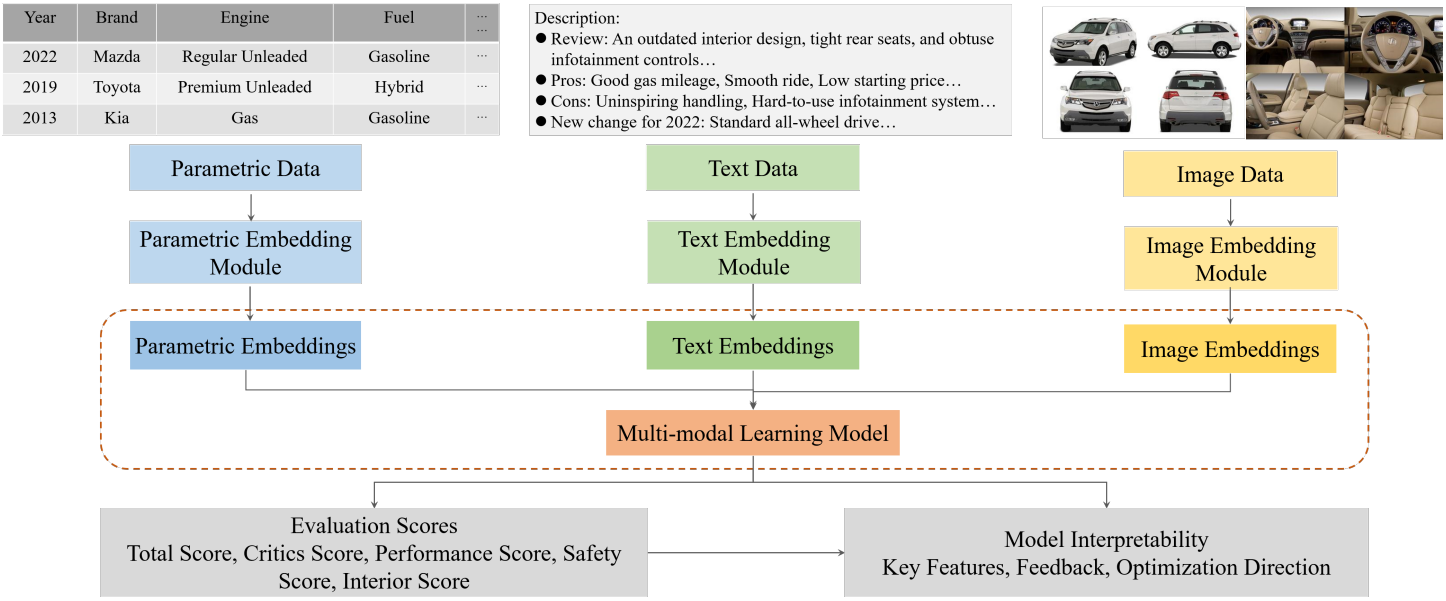

Figure 1: The outline of the proposed multi-modal learning model for predicting vehicle rating scores[7].

version. The image data are the exterior and interior photos of the vehicle. The parametric data conveys detailed specification information of the vehicle, such as body style, dimensions, and other mechanical, safety, and interior features. Our dataset covers 4,517 different vehicle models from different categories from 2007 to 2022. However, some relevant information is missing on the US News website for 1,946 vehicle models, which are excluded from this study. Accordingly, the data of 2,571 vehicles is used to develop the multi-modal learning model.

3.2 Data Processing

The raw data from the US News website contains three types of data: parametric specifications, text descriptions, and image data. The parametric specifications consist of five categories: general information, exterior information, interior information, mechanical information, and safety information. Each category covers multiple subcategories, as listed in Table 1. Notably, the subcategories further comprise varying numbers of features, resulting in a total of 302 features for each vehicle. Some of these features are numeric, while others are categorical. When preprocessing the data, we normalize all numeric features to [0,1] and use one-hot encoding to represent the categorical features.

The image data consists of exterior and interior photos. Among a large number of exterior and interior photos, we select the four most representative exterior or interior photos as the input to the image model. Specifically, the selected exterior photos include photos taken from four fixed views: angular front, front, rear, and side. The original size of these photos is 776×776. First, we resize the original images to 224×224. Second, we remove part of the white background at the periphery of the exterior photos to further reduce the size of the images. Third, we integrate the four resized exterior photos into a single exterior image with a size of 448×290, as shown in Figure 2.

[7] https://cars.usnews.com/cars-trucks/acura/mdx/2007

Category	Subcategory
General Information	Years Brand Drivetrain Manufacturer Suggested Retail Price (MSRP) Mile Per Gallon (MPG) City Mile Per Gallon (MPG) Highway
Exterior Information	Exterior Body Style Exterior Dimensions Exterior Measurements
Interior Information	Interior Convenience & Comfort Interior Dimensions Interior Entertainment Interior Heating Cooling Interior Navigation & Communication Interior Seats
Mechanical Information	Mechanical Transmission Mechanical Fuel Engine & Performance
Safety Information	Safety Airbags Safety Brakes Safety Features

Table 1: Parametric specification information.

The selected interior photos cover the major interior components, including the dashboard, front seat, rear seat, and steering wheel. While their original size is 776×517, we resize the interior photos to 224×150 proportionally and integrate them to produce a single interior photo with a size of 448×300.

For text data, the information of different features is formatted as "The name of this feature: content", and the information on different features is concatenated successively into a single

machine-readable string. The maximum, minimum, and average word lengths for the text descriptions are 224, 32, and 74, respectively.

Figure 2: Examples of vehicle exterior[8] and interior[9] photos.

3.3 Models

In this subsection, we first introduce the three unimodal models for embedding the parametric data, image data, and text data, respectively. Then, we describe how the multi-modal learning model is constructed based on these three unimodal models.

Unimodal Model Figure 3 illustrates the architectures of the unimodal models that respectively learn the parametric, image, and text data. All these unimodal models adopt the ReLU activation function for the regression task in this study. The final output for each unimodal model is the predicted value of the rating score.

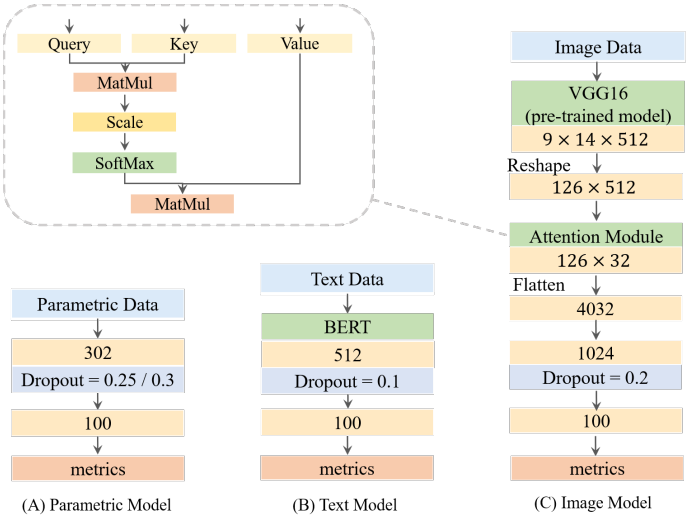

Figure 3: The architectures of three unimodal models.

(1) **Parametric model**: Firstly, we construct an MLP model to learn the parametric data. To find the optimal neural network architecture and hyper-parameters, we conduct a set of pilot experiments. This process leads us to a simple neural network architecture containing two hidden layers, as depicted in Figure 3-A. The number of neurons in the first hidden layer is equal to the dimension of the input data (302), and that of the second hidden layer is 100. We add a dropout layer after the first hidden layer with dropout rates ranging from 0.25 or 0.3 for predicting different rating scores.

[8] https://cars.usnews.com/cars-trucks/acura/mdx/2007/photos-exterior
[9] https://cars.usnews.com/cars-trucks/acura/mdx/2007/photos-interior

(2)**Text model**: Secondly, the text model adopts a pre-trained transformer-based BERT [30] text embedding module. We use the pooled output from the BERT model as the final embedding of the input text with a dimension of 512. During pre-training, the BERT embedding module is trained on large text databases (e.g., Wikipedia.) for multiple tasks. The adoption of the pre-trained BERT model allows for effective knowledge transfer from the large external text dataset to our target text descriptions when we fine-tune the model with our dataset for the regression task. Following the BERT embedding layer, a dropout layer with a dropout rate of 0.1 and a dense layer with 100 neurons are attached before the final output layer, as shown in Figure 3-B. We unfreeze all layers to train the text model.

(3) **Image model**: Thirdly, we construct a CNN-based model to learn the vehicle images. We experiment with multiple CNN models, including ResNet [22], Inception [23], and VGG16 [21], during our pilot experiments and get similar performances from them. VGG16 [21] is selected for this study because it takes less time to train. The output from the VGG16 embedding module exhibits a dimension of $9 \times 14 \times 512$. Following the image embedding module, we add a self-attention mechanism, as visualized in Figure 3-C. It reshapes the output to 126×512, which is seen as a set of 126 latent features with a dimension of 512. The self-attention mechanism employs a latent dimension of 32 in this study. Since the input to the image model integrates four exterior or interior vehicle photos, which complement or align with each other to different degrees. The self-attention mechanism is expected to facilitate capturing the interactions between different regions of the input images. The self-attention mechanism employs the dot-product attention proposed in "Attention Is All You Need" [15]. After that, a flatten layer, a dense layer with 1024 neurons, a dropout layer with a dropout rate of 0.2, another dense layer with 100 neurons, and a final output layer are attached sequentially.

For the image model, to enhance our evaluation of vehicle features, we use interior photos to evaluate the interior score and use exterior images to evaluate the other rating scores.

Multi-modal Learning Model After the three unimodal models are trained, we integrate them to construct the multi-modal learning model. To facilitate the learning of the interactions between the three data modalities, we do not directly integrate the unimodal models by concatenating their final outputs. Instead, we concatenate the final embedding of the input parametric, text, and image data from the corresponding unimodal models as the final joint representation of the multi-modal input data. The final output of the multi-modal learning model is calculated from the joint representation through a dense layer with the ReLU activation function. The architecture of the multi-modal model is shown in Figure 4. In comparison, we also construct three bi-modal models that respectively combine two of the three unimodal models. These different combinations give us four multi-modal learning models. For the sake of simplicity, we refer to the bi-modal learning model combining parametric and text data as $Par\_Text - MML$ model, the bi-modal learning model combining parametric and image data as $Par\_Img - MML$ model, the bi-modal learning model combining the image and text data as $Img\_Text - MML$ model. The

multi-modal model combining the parametric, text, and image data is called $Par\_Text\_Img - MML$ model in this paper hereafter. When training the multi-modal models, we initialize the multi-modal learning models with the pre-trained weights from the unimodal models and fine-tune the weights jointly to learn the interactions between the three data modalities for better vehicle rating score prediction.

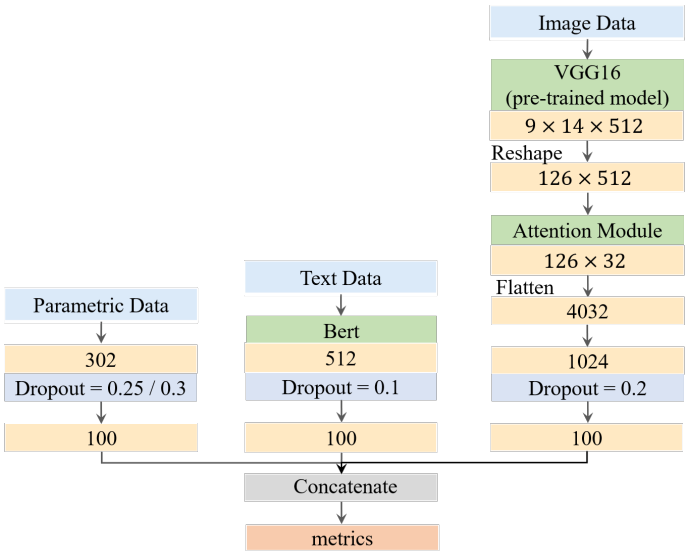

Figure 4: The architecture of the multi-modal learning model.

4. RESULTS AND DISCUSSION

In this section, we compare the performances of the three unimodal models and the four multi-modal learning models to verify the effectiveness of the proposed multi-modal learning model. Specifically, the performance of each model is assessed in terms of the explanatory power for the variances of the vehicle rating scores, which is known as the determination coefficient (i.e., R^2 value) in statistics. In regression, the degree of fit improves as the R^2 value increases. To train and test the model, the 2,571 vehicles in our dataset are divided into the training set, validation set, and test set following the ratio of 0.8:0.1:0.1. In the process of data split, we ensure that the stratified distribution of the rating scores within each set is consistent with that of the entire dataset. We observe that the distribution of different rating scores could be very different, so we generate a unique data split for each of the five rating scores. All the models use the same data split to predict the same rating score for easy comparison. During training, different unimodal and multi-modal learning models are trained with the same batch size of 32 and the initial learning rates ranging from 0.001 to 0.00005, which are selected through a series of pilot experiments. We also apply different decay rates ranging from 0.0 to $e^{-0.015}$ to schedule the learning rates during training different models. The training process is ended if the validation loss does not decrease for 20 consecutive epochs. In order to demonstrate the stability of the model and test the statistical significance of the differences between different models, we repeat each experiment 10 times.

4.1 Performance of Unimodal Models

The three unimodal models show varying performances in predicting different rating scores, as shown in Figure 5. The parametric model best predicts all rating scores. Its R^2 values are at least 0.04 higher than that of the corresponding image and text models for predicting all rating scores. Moreover, in most cases, the text model outperforms the image model. That is, the parametric data is most informative while the image data is least informative in predicting these rating scores. The information conveyed by these different types of data may explain the differences in model performance. The parametric data intuitively shows the detailed specifications of a vehicle, including general information, exterior information, interior information, mechanical information, and safety information of the vehicle, which summarizes the vehicle's major characteristics. The compact representation may make it easier for the model to capture the key features, leading to better predictions. In comparison, the text data describes the advantages and disadvantages of a vehicle compared to other vehicles or its previous version, which is also valuable for rating the vehicle. The images of a vehicle show its aesthetic features and body design, which influence customers' affection for it and its aerodynamic performance. Since exterior design is not considered by the five rating scores directly, the information conveyed by the images might be less informative for predicting these scores.

Figure 5: The performances of the unimodal models. The columns show the average R^2 values from the 10 repeated experiments with the bars indicating one standard error. We observe that parametric models have higher R^2 across different metrics.

Moreover, we find that the three unimodal models are relatively less effective in predicting the total score compared to the other scores. The total score is an overarching evaluation of a vehicle, which is the weighted average of the other four rating scores and several other indicators. The prediction of such an overarching score needs more complicated and comprehensive information from multiple perspectives, which is more challenging for the unimodal models to learn from a single data modality. Therefore, the unimodal models may struggle to learn enough features during training and thus do not predict the overall score as well as the other four rating scores. In addition, the dataset used in this study is small, which cannot provide sufficient information to train these large models. We observe that it is easy to overfit these models and the training is terminated early be-

fore better model weights can be learned, which may lead to insufficient final predictions. Among all five rating scores, the parametric model exhibits the highest R^2 value in predicting the safety score, and the R^2 values of the three unimodal models differ greatly. The R^2 value of the parametric model is higher than that of the worst model (image) by 0.34. One potential reason is that its evaluation is partly based on the advanced accident avoidance technologies implemented by a vehicle, which is described clearly in the parametric data. In comparison, although the vehicle body design reflected by the images affects the safety of the vehicle, the material of the vehicle body is unknown from the images and the importance of body design is being weakened by the incorporation of the technologies in recent years.

4.2 Effect of Multi-modal learning

Multi-modal learning models outperform the unimodal models. For predicting all rating scores, the average R^2 values of the multi-modal learning models are significantly higher than that of the corresponding unimodal models, as shown in Figure 6. The results suggest that compared to the unimodal models, the joint learning of multi-modal data enables the multi-modal learning models to leverage the complementary features learned from different modalities to better predict the rating scores. Moreover, the $Par\_Text\_Img - MML$ model also performs better than the three bi-modal learning models that integrate two types of data for predicting all rating scores except for the total score. The $Par\_Text - MML$ model slightly outperforms the $Par\_Text\_Img - MML$ model for predicting the total score. This may seem counterintuitive, as adding one more mode should logically allow the model to learn more information and, thus, likely make better predictions. However, as discussed above, the evaluation of the total score relies on more complicated, interrelated, and comprehensive information. This is more challenging for the models to learn. The features learned from the three data modalities are fused through simple concatenation in this paper, which may not be able to capture the complex interactions among the modalities for better total score prediction when the image data is involved. Another possible reason is the dataset used in this study is not large enough to support learning the complex cross-modal interactions for predicting the total score.

The effect of multi-modal learning varies in predicting different rating scores. The effect of multi-modal learning is most substantial in predicting the total score. The best multi-modal model ($Par\_Text - MML$) outperforms the best unimodal model (parametric) by 0.12. The characteristics of the total score imply that its evaluation relies more on a comprehensive understanding of a vehicle. Accordingly, the multi-modal features learned by the multi-modal models help significantly in this regard. The effect of multi-modal learning is least obvious in predicting the safety score. The best $Par\_Text\_Img - MML$ model only improves the R^2 value by 0.02 compared with the best unimodal model, which is the parametric model. As mentioned above, the parametric data clearly describe the advanced accident avoidance technologies implemented by a vehicle, which inform the evaluation of the safety score. Since the R^2 values achieved by the image model and the text model are much lower than the parametric model, the incorporation of the text and image data only slightly complements the parametric information for this evaluation. In general, the $Par\_Text\_Img - MML$ model exhibits similar or slightly better performances compared to the best bi-modal learning models for predicting all rating scores. The simple information fusion mechanism and the small size of the dataset in this study may help explain this.

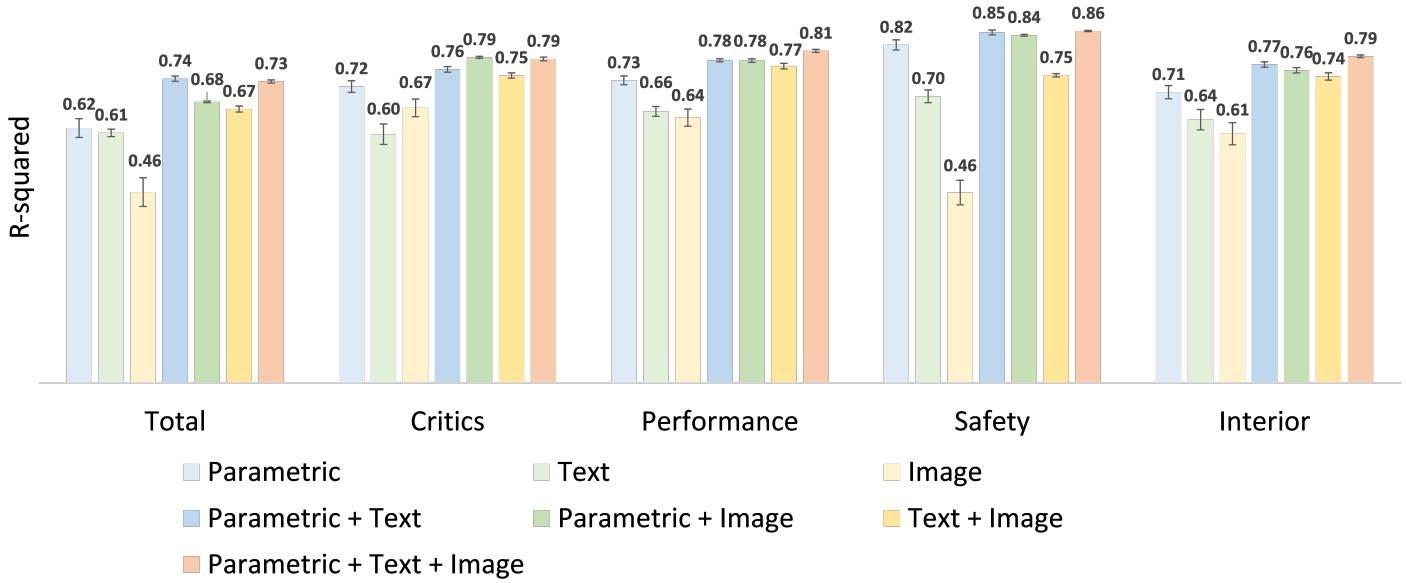

Figure 6: The comparison in performance among the unimodal and multi-modal learning models. The columns show the average R^2 values with one standard deviation bar. We observe that the multi-modal model using three modalities outperforms unimodal or bi-modal models.

4.3 Implications for Engineering Design

As demonstrated in the last subsection, our multi-modal learning models can predict vehicle rating scores accurately using vehicle parametric specifications, text descriptions, and images. However, a more detailed interpretation of the outputs from the models is needed to inform designers and companies about potential directions to optimizing the inferior design or advertising the superior design of a vehicle. For this purpose, we utilize the SHAP [64] method to interpret the outputs from the image, text, and parametric models. Through backward gradient-based sensitivity analysis, the output SHAP values indicate the influence of each element of the input data on the final prediction made by a model. A higher absolute SHAP score suggests a higher influence. Therefore, the SHAP method can help us interpret how a deep learning model makes its decision.

We first conduct SHAP analysis for the parametric model. As we mentioned before, the parametric data conveys rich information across five feature categories as listed in Table 1. The SHAP analysis can help us identify the most informative and influential vehicle feature categories from the parametric data. We run SHAP analysis for the parametric models that predict the five rating scores, respectively. Figure 7 illustrates the average absolute SHAP values of the five feature categories across all vehicles in the test set for predicting different scores. These values indicate the extents to which different feature categories affect the model's predictions. The findings indicate that the impact of each category varies, with the interior information category having the greatest influence on all score predictions and the exterior information category having the least impact on them.

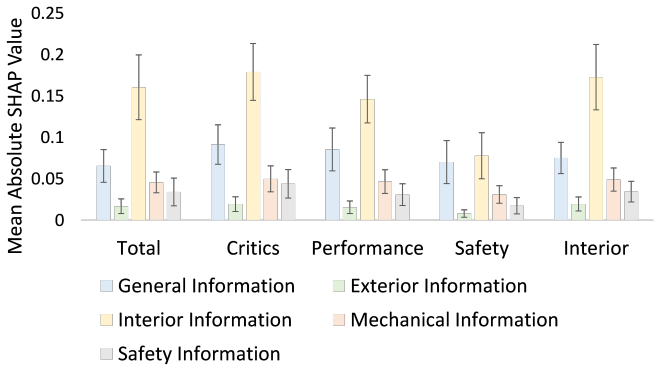

Figure 7: Mean Absolute SHAP values of the five feature categories of the parametric data with one standard error. We observe that the interior information category exhibits the largest SHAP values while the exterior information category holds minimal significance.

On this basis, we further analyze the influences of the 21 feature subcategories on the model predictions, and Table 2 shows the 21 subcategories used to predict the total score. The findings are in line with our expectations and quite interesting. When buying a vehicle, customers tend to base their decisions on the vehicle brand and the appearance of the vehicle's body. For instance, some people prefer Toyota sedans, while others may prefer Subaru SUVs. This indicates that the features like "Brands" and "Exterior Body Style" can significantly influence the prediction. Furthermore, the comfort and convenience of driving a vehicle are crucial since the most straightforward feeling that drivers and passengers may have for a vehicle is how comfortable and convenient it is to drive or ride in it. That is why most of the interior information subcategories have prominent impacts on rating predictions. Additionally, people value the safety and performance of a vehicle since if the vehicle's safety and performance are not up to par, people are less likely to trust and purchase it. Accordingly, "Safety Features", "Engine & Performance" and "Mechanical Transmission" features are also important. Designers and companies need to focus on these feature subcategories to improve or advertise their designs.

In addition to examining the average absolute values, we also analyze the variation of the average SHAP values for the 302 individual features over time. Although the SHAP values of the majority of the features do not show noticeable trends (e.g., sharp fluctuations or little changes), a subset of the features exhibit clear increasing or decreasing trends, as shown in Figure 8. As electronic and information technologies continue to advance, these technologies have been enhancing the driving experience and promoting driving safety. For example, the "Heated Steering Wheel" prevents hand stiffness during long driving hours in winter, and "Keyless Start" eliminates the need for manual key insertion by pressing a button inside the vehicle, or turning a knob, making the process more convenient. Furthermore, the "Hands-Free Liftgate" automatically opens and closes the liftgate. Other technologies, such as "Back-UP Camera," "Lane Keeping Assist," and "Lane Departure Warning" help improve driving safety on the road. These features have experienced an increase in their SHAP values over time, highlighting their growing positive influence on the model's prediction. On the other hand, the SHAP values of a few others show the opposite pattern, such as "auxiliary power outlet", "regular unleaded (fuel)", and "high-intensity discharge (HID) headlights". These features play increasingly negative roles in affecting the predictions, which means having these features may result in lower rating scores as time passes by. For example, "Auxiliary Pwr Outlet," also known as the "car cigarette lighter," is an outdated feature that has been excluded by many new vehicle models. By analyzing the original data, we observe that most vehicles manufactured before 2014 were equipped with cigarette lighters, but very few vehicles had them after that year. Similarly, "HID headlights" were once popular for their high brightness and long service life compared to traditional halogen bulbs. However, due to their expensive manufacturing costs and slow response time to peak brightness, they have been gradually replaced by LED headlights that offer lower power consumption, longer lifespan, and faster response times. Consequently, HID headlights are disappearing from the market, which aligns with the observed changes in its SHAP values.

Then, We employ the SHAP method to analyze the informativeness of different image regions for predicting different rating scores. Since we use the interior and exterior images to predict the interior score and the other four scores, respectively. The mean absolute SHAP values of interior image regions are displayed in Table 3, while Figure 9 showcases the mean absolute SHAP values of the exterior image regions for predicting the four rating scores. For predicting the interior score, the dashboard and steering wheel regions present higher mean absolute SHAP

Subcategory	Mean Absolute Shap Values	Sample Features
Interior Convenience & Comfort	$9.11 \cdot 10^{-2}$	Auxiliary Pwr Outlet, HID headlights, Back-Up Camera
Brand	$3.93 \cdot 10^{-2}$	Toyota, Honda, Ford
Interior Seats	$2.59 \cdot 10^{-2}$	Heated Rear Seat, Heated Front Seats, Driver Lumbar
Interior Entertainment	$2.58 \cdot 10^{-2}$	WiFi Hotspot, Smart Device Integration, Entertainment System
Engine & Performance	$2.17 \cdot 10^{-2}$	Engine Type, Premium Unleaded, Regular Unleaded
Safety Features	$2.06 \cdot 10^{-2}$	Lane Keeping Assist, Blind Spot Monitor
Mechanical Transmission	$2.01 \cdot 10^{-2}$	Continuously Variable Trans, Manual, 6-Speed Automatic
Years	$1.78 \cdot 10^{-2}$	2010, 2014, 2020
Exterior Body Style	$1.25 \cdot 10^{-2}$	SUV, Sedan, Hatchback
Interior Heating Cooling	$7.44 \cdot 10^{-3}$	Climate Control, Dual Zone A/C, Rear A/C
Safety Airbags	$7.21 \cdot 10^{-3}$	Rear Side Air Bag, Passenger Air Bag On/Off Switch
Interior Navigation & Communication	$7.15 \cdot 10^{-3}$	Navigation System, Onboard Communications System
Safety Brakes	$6.48 \cdot 10^{-3}$	Front Disc/Rear Drum Brakes, 4-Wheel Disc Brakes
Drivetrain	$6.21 \cdot 10^{-3}$	AWD, FWD, RWD
Mechanical Fuel	$3.99 \cdot 10^{-3}$	Hybrid Fuel, Plug-In Electric/Gas, Gasoline Fuel
Exterior Dimensions	$3.43 \cdot 10^{-3}$	Dim Width (in), Wheelbase (in), Dim Length (in)
Interior Dimensions	$3.30 \cdot 10^{-3}$	Front Head Room (in.), Front Leg Room (in.)
Manufacturer Suggested Retail Price(MSRP)	$1.46 \cdot 10^{-3}$	MSRP
Exterior Measurements	$1.08 \cdot 10^{-3}$	Base Curb Weight (lbs)
Mile Per Gallon (MPG) City	$5.25 \cdot 10^{-4}$	MPG City
Mile Per Gallon (MPG) Highway	$4.53 \cdot 10^{-4}$	MPG Highway

Table 2: The feature subcategories with the corresponding SHAP Values and a few sample features. We observe that Interior convenience and brand are found most important in ratings prediction.

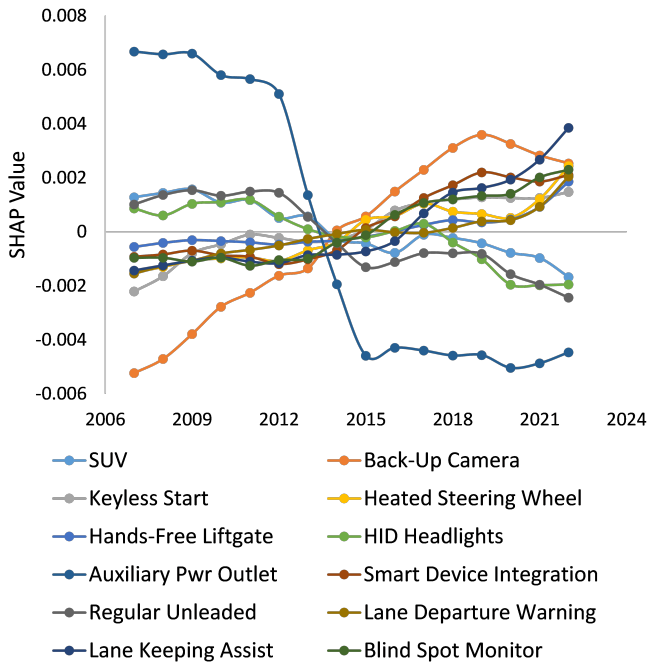

Figure 8: The SHAP values of 12 example features for predicting the total score over time. We observe that the importance of some features such as back-up cameras has increased in the last decade.

values than the front and rear seat regions, suggesting that the model may primarily rely on features in these regions to predict the interior score. Notably, the SHAP values of the front and rear views in the exterior images are higher than that of the other two views in predicting most of the rating scores. The results indicate the critical role of the front and rear views in predicting the other scores. This indicates that when purchasing a vehicle, people prioritize the front and back views, as they are the most visible when driving. Well-designed front and rear portions of a vehicle can also potentially reduce safety risks.

Score	Dashboard	Steering Wheel	Front Seat	Rear Seat
Interior	1.159	1.011	0.448	0.389

Table 3: Mean absolute SHAP values of the interior image regions for predicting the interior score. The SHAP values of the dashboard and steering wheel highlight their high importance in predicting the interior score

To gain further insights into the performance of different features in each region, we used the SHAP method to analyze the image data of individual samples. We exemplify the SHAP values of two representative images for predicting the total and interior scores in Figure 10 and Figure 11, respectively. The red regions positively influence the predictions, while the blue regions negatively influence the predictions. The color intensity indicates the influence extent.

Figure 10 showcases the SHAP values of the exterior image regions of the 2020 GMC Terrain for predicting the total score. We find that in the front view and angular front view, the regions on the front wheels, the vehicle brand logo, the fog lamps, the front bumper, and the front fenders have positive influences on the total score prediction of this vehicle. Similarly, in the rearview, the regions on the rear wheels, taillights, and bumpers also play a bigger role in predicting the total score. This is reasonable. For example, during night driving, turning on the taillight can alert the following vehicle to maintain a safe distance, and a well-designed

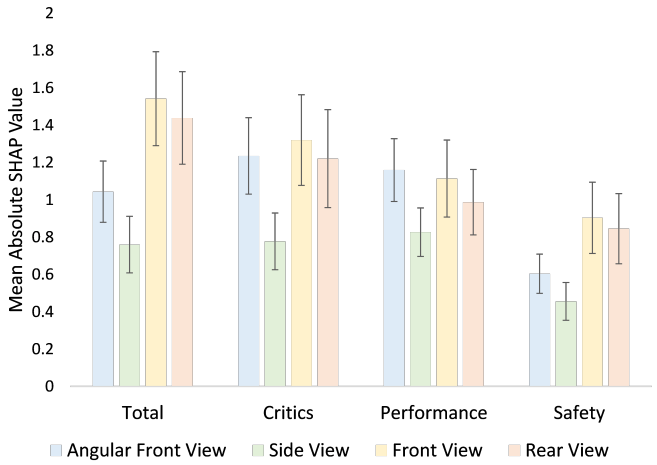

Figure 9: Mean absolute SHAP values of the exterior image regions for predicting the total, critics, performance, and safety scores with one standard deviation bar. The SHAP values of the front view and rear view play a bigger role in predicting these scores.

bumper can offer better protection in case of accidents. This will undoubtedly have a positive impact on the overall rating of the vehicle. Additionally, our SHAP analyses for predicting the other scores show similar trends. However, it is important to note that different vehicles may have distinct components that contribute to different score predictions, thus necessitating a case-by-case analysis by designers and engineers.

Figure 10: SHAP values of the exterior image regions of the 2020 GMC Terrain[10] for total score prediction: the SHAP values on the right with the corresponding exterior image on the left.

Figure 11 displays the SHAP values of the interior image regions of the 2020 Acura RLX for predicting the interior score. The instrument panel, the gearshift, the dashboard, the steering wheel, the steering wheel controls, and the front and rear seats are likely to have positive impacts on the interior score prediction of this vehicle. Among these, the steering wheel and dashboard have the most substantial impacts, as they are among the most used interior components. We believe that people can only truly experience the comfort and convenience of the front and rear seats when they are in the vehicle, rather than from a picture, so their effect may be weaker than that of the steering wheel. Overall, considering these details and features can lead to a better interior score and increase the vehicle's appeal to potential consumers.

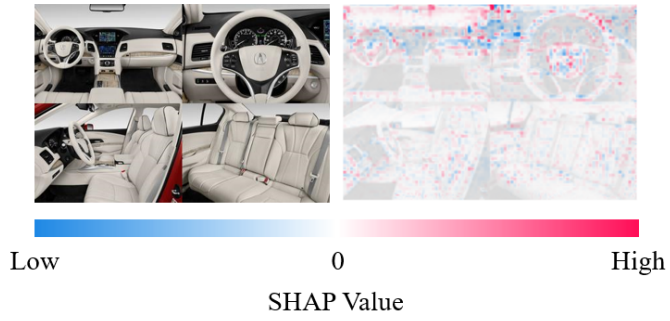

Figure 11: SHAP values of the interior image regions of the 2020 Acura RLX[11] for the interior score prediction: the SHAP values on the right with the corresponding interior image on the left. One can observe that most points are clustered near the dashboard, indicating its importance for ratings.

Lastly, We use the SHAP method to analyze the informativeness of different text segments for predicting different rating scores. The results are displayed in Figure 12. Our findings reveal that the review of a vehicle and its advantages and disadvantages significantly influence the model predictions. In general, the review segment of the text has the largest influence on the rating score predictions. If this segment provides a negative evaluation (e.g.,) of the vehicle, often with words like "bottom", "however", or "but", the corresponding SHAP value is mostly negative, indicating a negative impact on the rating prediction. Otherwise, words like "top" indicate a positive evaluation and tend to have positive impacts on the predictions. Moreover, the brand and year of the vehicle mentioned in the text also have relatively important impacts on the model predictions. Additionally, the pros segment usually has a positive SHAP value, while the cons segment has a negative SHAP value, as expected. The "New Change" segment of the text indicates if there are any new changes in the vehicle compared to the previous year. If there are positive changes, the corresponding SHAP value is usually positive.

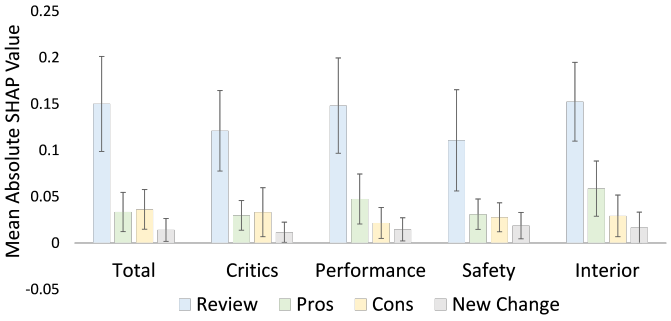

Figure 12: Mean Absolute SHAP values of different segments of the text data for predicting different scores with one standard error. The review segment is most important for all predictions.

Figure 13 is an instance of SHAP analysis applied to an

[10] https://cars.usnews.com/cars-trucks/gmc/terrain/2020/photos-exterior
[11] https://cars.usnews.com/cars-trucks/acura/rlx/photos-interior

individual vehicle - the 2020 Mazda - for predicting its total score. The light blue colors assigned to "2020" and "Mazda" indicate slightly negative effects on the prediction. The positive SHAP value of the word "top" confirms the esteemed position held by the vehicle in the US News evaluation system, positively influencing the prediction. This positive reputation may sway potential buyers towards considering this vehicle over others within this system. In addition, the vehicle has several advantages such as a "premium cabin," "pleasant ride," and "thrilling handling," all of which are positive aspects of the vehicle and present positive SHAP values. These aspects probably lead to a positive perception of the vehicle and attract potential buyers. However, the vehicle also has some drawbacks, such as "subpar cargo space" and a "cramped third row," indicating that it may not be the best option for those looking for more space. These downsides have negative impacts on the total score forecast, as indicated by their negative SHAP values. New changes such as "standard heated front seats" and "Mazda i-active sense suite of safety features made standard," are positive changes and exhibit positive SHAP values. Engineers and designers need to analyze their own designs case by case.

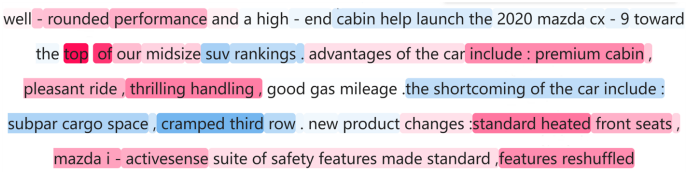

Figure 13: SHAP values of the words in the 2020 Mazda CX-9's text data[12] for its total score prediction. The red and blue colors imply positive and negative impacts, respectively. Words such as premium cabin and thrilling handling are found to have high importance.

To inform the design and optimization directions of individual vehicles, brands, or other aspects, designers and engineers need to extract a suitable sub-dataset from the entire dataset, retrain the prediction model, and carry out SHAP analyses and discussions accordingly. This approach can improve the rating score prediction for a particular type of vehicle and improve the interpretability of the model.

4.4 Limitations and Future Work

This section summarizes the limitations of this study and future research directions of using multi-modal learning models to promote engineering design. First, a major limitation of this study is our dataset is much smaller than the other datasets for training large deep learning models, which may not provide sufficient information for the multi-modal learning models to learn the complex interactions between different data modalities. It is hard to harness the full potential of multi-modal learning with small datasets. We observe that the US News website only provides information regarding the vehicles on the US market, leading to the exclusion of some vehicle brands from China, India, and other countries in this study. We will work to expand this dataset by including more vehicle brands and completing the information on the vehicles with missing data items in the future.

[12] https://cars.usnews.com/cars-trucks/mazda/cx-9/2020

Second, we use the simple concatenation mechanism to fuse information from different data modalities in this paper, which may lead to less effective information fusion compared to more advanced information fusion mechanisms, such as attention-based or transformer-based information fusion. In future work, we will explore advanced techniques to fuse features learned from the parametric, text, and image data. Additionally, not all information available from the US News website is leveraged in this study. For example, we only select four exterior photos and four interior photos from a much larger photo collection and use a small part of each text description from the website for the rating score prediction. A more comprehensive understanding of a vehicle and a better rating score prediction may be achieved by incorporating all available information into ML. In the future, we need to explore more effective and efficient deep-learning models to manage richer data.

5. Conclusion

In this research, we have developed and validated a multi-modal learning model aimed at predicting five different vehicle rating scores—total score, critics score, performance score, safety score, and interior score. These predictions are facilitated using the parametric specifications, text descriptions, and images of vehicles. As the foundation of the multi-modal learning model, we developed three unimodal models to independently extract features from parametric, text, and image data. Based on this, we compared the efficacy of the multi-modal learning model against its unimodal equivalents. Our research has led to three significant discoveries: 1. Parametric data proves to be the most informative in predicting all the scores, with the text model surpassing the image model in most instances for predicting the rating scores. 2. The multi-modal learning model, which concurrently learns from parametric, text, and image data, outperforms all the unimodal models. This suggests that multi-modal data learning captures a richer array of information than learning from a single data mode for the task of prediction. 3. The sensitivity analyses conducted via SHAP can offer invaluable insights for interpreting predictions and provide crucial design, optimization, and improvement guidance to designers and engineers. Furthermore, the proposed multi-modal learning methodology can be extrapolated to a broader range of application scenarios, potentially providing fresh insights and inspiration for designers.

References

[1] Simbolon, Freddy Pandapotan, Handayani, Elvira and Nugraedy, Menik. "The Influence of Product Quality, Price Fairness, Brand Image, and Customer Value on Purchase Decision of Toyota Agya Consumers: A Study of Low Cost Green Car." *Binus Business Review* Vol. 11 (2020): pp. 187–196.

[2] Ponmalar, Punitha, Associate, P, Angelin, M Rs and Assistant, Christinal C. "Review on the Pre-owned Car Price Determination using Machine Learning Approaches; Review on the Pre-owned Car Price Determination using Machine Learning Approaches." (2022)DOI 10.1109/ICAISS55157.2022.10010958.

[3] Jin, Chuyang. "Price Prediction of Used Cars Using Machine Learning." *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*: pp. 223–230. 2021. DOI 10.1109/ICESIT53460.2021.9696839.

[4] Tsagris, Michail and Fafalios, Stefanos. "Advanced Car Price Modelling and Prediction." *Advances in Econometrics, Operational Research, Data Science and Actuarial Studies: Techniques and Theories*. Springer (2022): pp. 479–494.

[5] Xia, Zhenchang, Xue, Shan, Wu, · Libing, Sun, Jiaxin, Chen, Yanjiao and Zhang, Rui. "ForeXGBoost: passenger car sales prediction based on XGBoost." *Distributed and Parallel Databases* Vol. 38 (123). DOI 10.1007/s10619-020-07294-y. URL https://doi.org/10.1007/s10619-020-07294-y.

[6] Kumar Panda, Sandeep, Mohapatra, Ramesh Kumar, Panda, Subhrakanta, Balamurugan, S and Kumar Panda, Samdeep. "Car Buying Criteria Evaluation Using Machine Learning Approach." (2022)DOI 10.1002/9781119884392.ch10. URL https://onlinelibrary.wiley.com/doi/10.1002/9781119884392.ch10.

[7] Li, Deming, Li, Menggang, Han, Gang and Li, Ting. "A combined deep learning method for internet car evaluation." *Neural Computing and Applications* Vol. 33 . DOI 10.1007/s00521-020-05291-x. URL https://doi.org/10.1007/s00521-020-05291-x.

[8] Wang, Hui Dong. "ScienceDirect Research on the Features of Car Insurance Data Based on Machine Learning." *Procedia Computer Science* Vol. 166 (2020): pp. 582–587. DOI 10.1016/j.procs.2020.02.016. URL www.sciencedirect.comwww.sciencedirect.com.

[9] Borisov, Vadim, Leemann, Tobias, Seßler, Kathrin, Haug, Johannes, Pawelczyk, Martin and Kasneci, Gjergji. "Deep neural networks and tabular data: A survey." *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[10] Su, Xiaogang, Yan, Xin and Tsai, Chih Ling. "Linear regression." *Wiley Interdisciplinary Reviews: Computational Statistics* Vol. 4 No. 3 (2012): pp. 275–294. DOI 10.1002/WICS.1198. URL https://onlinelibrary.wiley.com/doi/full/10.1002/wics.1198https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1198https://wires.onlinelibrary.wiley.com/doi/10.1002/wics.1198.

[11] Wang, Jack, Hertzmann, Aaron and Fleet, David J. "Gaussian process dynamical models." *Advances in neural information processing systems* Vol. 18 (2005).

[12] Chen, Tianqi and Guestrin, Carlos. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*: pp. 785–794. 2016.

[13] Kadra, Arlind, Lindauer, Marius, Hutter, Frank and Grabocka, Josif. "Regularization is all you need: Simple neural nets can excel on tabular data." *arXiv preprint arXiv:2106.11189* Vol. 536 (2021).

[14] Arik, Sercan Ö and Pfister, Tomas. "Tabnet: Attentive interpretable tabular learning." *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8: pp. 6679–6687. 2021.

[15] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz and Polosukhin, Illia. "Attention is all you need." *Advances in neural information processing systems* Vol. 30 (2017).

[16] Pak, Myeongsuk and Kim, Sanghoon. "A review of deep learning in image recognition." *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*: pp. 1–3. 2017. DOI 10.1109/CAIPT.2017.8320684.

[17] Rawat, Waseem and Wang, Zenghui. "Deep convolutional neural networks for image classification: A comprehensive review." *Neural computation* Vol. 29 No. 9 (2017): pp. 2352–2449.

[18] Minaee, Shervin, Boykov, Yuri, Porikli, Fatih, Plaza, Antonio, Kehtarnavaz, Nasser and Terzopoulos, Demetri. "Image segmentation using deep learning: A survey." *IEEE transactions on pattern analysis and machine intelligence* Vol. 44 No. 7 (2021): pp. 3523–3542.

[19] Elasri, Mohamed, Elharrouss, Omar, Al-Maadeed, Somaya and Tairi, Hamid. "Image Generation: A Review." *Neural Processing Letters* Vol. 54 No. 5 (2022): pp. 4609–4646.

[20] Krizhevsky, Alex, Sutskever, Ilya and Hinton, Geoffrey E. "Imagenet classification with deep convolutional neural networks." *Communications of the ACM* Vol. 60 No. 6 (2017): pp. 84–90.

[21] Simonyan, Karen and Zisserman, Andrew. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[22] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing and Sun, Jian. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*: pp. 770–778. 2016.

[23] Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent and Rabinovich, Andrew. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*: pp. 1–9. 2015.

[24] Bengio, Yoshua, Ducharme, Réjean and Vincent, Pascal. "A neural probabilistic language model." *Advances in neural information processing systems* Vol. 13 (2000).

[25] Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural networks* Vol. 61 (2015): pp. 85–117.

[26] Hochreiter, Sepp and Schmidhuber, Jürgen. "Long short-term memory." *Neural computation* Vol. 9 No. 8 (1997): pp. 1735–1780.

[27] Chung, Junyoung, Gulcehre, Caglar, Cho, KyungHyun and Bengio, Yoshua. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).

[28] Radford, Alec, Narasimhan, Karthik, Salimans, Tim, Sutskever, Ilya et al. "Improving language understanding by generative pre-training." (2018).

[29] Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, Sutskever, Ilya et al. "Language models are

unsupervised multitask learners." *OpenAI blog* Vol. 1 No. 8 (2019): p. 9.

[30] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[31] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke and Stoyanov, Veselin. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

[32] Yang, Zhilin, Dai, Zihang, Yang, Yiming, Carbonell, Jaime, Salakhutdinov, Russ R and Le, Quoc V. "Xlnet: Generalized autoregressive pretraining for language understanding." *Advances in neural information processing systems* Vol. 32 (2019).

[33] Song, Binyang, Zhou, Rui and Ahmed, Faez. "Multi-modal Machine Learning in Engineering Design: A Review and Future Directions." (2023)DOI 10.48550/arxiv.2302.10909. URL https://arxiv.org/abs/2302.10909v1.

[34] Baltrusaitis, Tadas, Ahuja, Chaitanya and Morency, Louis Philippe. "Multimodal Machine Learning: A Survey and Taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 41 No. 2 (2019): pp. 423–443. DOI 10.1109/TPAMI.2018.2798607.

[35] Song, Binyang, Miller, Scarlett and Ahmed, Faez. "Hey, AI! Can You See What I See? Multimodal Transfer Learning-Based Design Metrics Prediction for Sketches With Text Descriptions." *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 86267: p. V006T06A017. 2022. American Society of Mechanical Engineers.

[36] Nojavanasghari, Behnaz, Gopinath, Deepak, Koushik, Jayanth, Baltrušaitis, Tadas and Morency, Louis Philippe. "Deep multimodal fusion for persuasiveness prediction." *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction* (2016): pp. 284–288DOI 10.1145/2993148.2993176.

[37] Anastasopoulos, Antonios, Kumar, Shankar and Liao, Hank. "Neural Language Modeling with Visual Features." *undefined* (2019)URL http://arxiv.org/abs/1903.02930https://arxiv.org/abs/1903.02930v1.

[38] Vielzeuf, Valentin, Lechervy, Alexis, Pateux, Stéphane and Jurie, Frédéric. "CentralNet: A multilayer approach for multimodal fusion." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* Vol. 11134 LNCS (2019): pp. 575–589. DOI 10.1007/978-3-030-11024-6_44.

[39] Shutova, Ekaterina, Kiela, Douwe and Maillard, Jean. "Black Holes and White Rabbits: Metaphor Identification with Visual Features." *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference* (2016): pp. 160–170DOI 10.18653/V1/N16-1020. URL https://aclanthology.org/N16-1020.

[40] Cao, Yue, Long, Mingsheng, Wang, Jianmin, Yang, Qiang and Yuy, Philip S. "Deep visual-semantic hashing for cross-modal retrieval." *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Vol. 13-17-Augu (2016): pp. 1445–1454. DOI 10.1145/2939672.2939812.

[41] Morvant, Emilie, Habrard, Amaury and Ayache, Stéphane. "Majority Vote of Diverse Classifiers for Late Fusion." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* Vol. 8621 LNCS (2014): pp. 153–162. DOI 10.48550/arxiv.1404.7796. URL https://arxiv.org/abs/1404.7796v2.

[42] Zadeh, Amir, Chen, Minghai, Cambria, Erik, Poria, Soujanya and Morency, Louis Philippe. "Tensor Fusion Network for Multimodal Sentiment Analysis." *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (2017): pp. 1103–1114DOI 10.48550/arxiv.1707.07250. URL https://arxiv.org/abs/1707.07250v1.

[43] Chen, Richard J., Lu, Ming Y., Wang, Jingwen, Williamson, Drew F.K., Rodig, Scott J., Lindeman, Neal I. and Mahmood, Faisal. "Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis." *IEEE Transactions on Medical Imaging* Vol. 41 No. 4 (2019): pp. 757–770. DOI 10.48550/arxiv.1912.08937. URL https://arxiv.org/abs/1912.08937v3.

[44] Tenenbaum, Joshua B. and Freeman, William T. "Separating style and content with bilinear models." *Neural Computation* Vol. 12 No. 6 (2000): pp. 1247–1283. DOI 10.1162/089976600300015349.

[45] Graves, Alex, Wayne, Greg and Danihelka, Ivo. "Neural Turing Machines." *arXiv preprint arXiv:1410.5401*. (2014)URL https://arxiv.org/abs/1410.5401v2http://arxiv.org/abs/1410.5401.

[46] Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick and Ommer, Björn. "High-Resolution Image Synthesis with Latent Diffusion Models." (2021): pp. 10674–10685DOI 10.48550/arxiv.2112.10752. URL https://arxiv.org/abs/2112.10752v2.

[47] Liu, Zhijian, Tang, Haotian, Lin, Yujun and Han, Song. "Point-Voxel CNN for Efficient 3D Deep Learning." *Advances in Neural Information Processing Systems* Vol. 32 (2019). DOI 10.48550/arxiv.1907.03739. URL https://arxiv.org/abs/1907.03739v2.

[48] Nichol, Alex, Dhariwal, Prafulla, Ramesh, Aditya, Shyam, Pranav, Mishkin, Pamela, McGrew, Bob, Sutskever, Ilya and Chen, Mark. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models." (2021)DOI 10.48550/arxiv.2112.10741. URL https://arxiv.org/abs/2112.10741v3.

[49] Kim, Gwanghyun, Kwon, Taesung and Ye, Jong Chul. "DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation." (2021)DOI 10.48550/arxiv.2110.02711. URL https://arxiv.org/abs/2110.02711v6.

[50] Duc Tuan, Nguyen Manh and Quang Nhat Minh, Pham. "Multimodal Fusion with BERT and Attention Mechanism for Fake News Detection." *Proceedings - 2021 RIVF International Conference on Computing and Communication Technologies, RIVF 2021* (2021)DOI 10.48550/arxiv.2104.11476. URL https://arxiv.org/abs/2104.11476v2.

[51] Devlin, Jacob, Cheng, Hao, Fang, Hao, Gupta, Saurabh, Deng, Li, He, Xiaodong, Zweig, Geoffrey and Mitchell, Margaret. "Language Models for Image Captioning: The Quirks and What Works." *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference* Vol. 2 (2015): pp. 100–105. DOI 10.48550/arxiv.1505.01809. URL https://arxiv.org/abs/1505.01809v3.

[52] Kwon, Elisa, Huang, Forrest and Goucher-Lambert, Kosa. "Enabling multi-modal search for inspirational design stimuli using deep learning." *AI EDAM* Vol. 36 (2022): p. e22. DOI 10.1017/S0890060422000130. URL https://www.cambridge.org/core/journals/ai-edam/article/enabling-multimodal-search-for-inspirational-design-stimuli-using-deep-learning/2F5EA4243AD422EA74EA2B9FDAF8FF05.

[53] Yuan, Chenxi, Marion, Tucker and Moghaddam, Mohsen. "Leveraging End-User Data for Enhanced Design Concept Evaluation: A Multimodal Deep Regression Model." *Journal of Mechanical Design* Vol. 144 No. 2 (2022): pp. 1–20. DOI 10.1115/1.4052366. URL https://asmedigitalcollection.asme.org/mechanicaldesign/article/144/2/021403/1119449/Leveraging-End-User-Data-for-Enhanced-Design.

[54] Li, Xingang, Xie, Charles and Sha, Zhenghui. "A Predictive and Generative Design Approach for Three-Dimensional Mesh Shapes Using Target-Embedding Variational Autoencoder." *Journal of Mechanical Design* Vol. 144 No. 11 (2022). DOI 10.1115/1.4054906. URL https://asmedigitalcollection.asme.org/mechanicaldesign/article/144/11/114501/1141958/A-Predictive-and-Generative-Design-Approach-for.

[55] Song, Binyang, Miller, Scarlett and Ahmed, Faez. "Attention-enhanced Multimodal Learning For Conceptual Design Evaluation." *Journal of Mechanical Design* (2023): pp. 1–38DOI 10.1115/1.4056669. URL https://asmedigitalcollection.asme.org/mechanicaldesign/article/doi/10.1115/1.4056669/1156042/ATTENTION-ENHANCED-MULTIMODAL-LEARNING-FOR.

[56] Murdoch, W James, Singh, Chandan, Kumbier, Karl, Abbasi-Asl, Reza and Yu, Bin. "Definitions, methods, and applications in interpretable machine learning." *Proceedings of the National Academy of Sciences* Vol. 116 No. 44 (2019): pp. 22071–22080.

[57] Molnar, Christoph. *Interpretable machine learning*. Lulu. com (2020).

[58] Molnar, Christoph, Casalicchio, Giuseppe and Bischl, Bernd. "Interpretable machine learning–a brief history, state-of-the-art and challenges." *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings*: pp. 417–431. 2021. Springer.

[59] Fisher, Aaron, Rudin, Cynthia and Dominici, Francesca. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *J. Mach. Learn. Res.* Vol. 20 No. 177 (2019): pp. 1–81.

[60] Rosé, Carolyn P, McLaughlin, Elizabeth A, Liu, Ran and Koedinger, Kenneth R. "Explanatory learner models: Why machine learning (alone) is not the answer." *British Journal of Educational Technology* Vol. 50 No. 6 (2019): pp. 2943–2958.

[61] Ahmed, Faez, Cui, Yaxin, Fu, Yan and Chen, Wei. "Product Competition Prediction in Engineering Design Using Graph Neural Networks." *ASME Open Journal of Engineering* Vol. 1 (2022).

[62] Shrikumar, Avanti, Greenside, Peyton and Kundaje, Anshul. "Learning important features through propagating activation differences." *International conference on machine learning*: pp. 3145–3153. 2017. PMLR.

[63] Sundararajan, Mukund and Najmi, Amir. "The many Shapley values for model explanation." *International conference on machine learning*: pp. 9269–9278. 2020. PMLR.

[64] Lundberg, Scott M and Lee, Su-In. "A Unified Approach to Interpreting Model Predictions." Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (eds.). *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. (2017): pp. 4765–4774. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[65] Sundararajan, Mukund, Taly, Ankur and Yan, Qiqi. "Axiomatic attribution for deep networks." *International conference on machine learning*: pp. 3319–3328. 2017. PMLR.

[66] Smilkov, Daniel, Thorat, Nikhil, Kim, Been, Viégas, Fernanda and Wattenberg, Martin. "Smoothgrad: removing noise by adding noise." *arXiv preprint arXiv:1706.03825* (2017).

,