International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

PriceMorphNet: Cross-Attention Multimodal Fusion for Intelligent E-Commerce Price Prediction

Gurjas Singh Gandhi[1], Dr. Netraja Mulay[2]

[1]Student, Department of Master of Computer Applications, Progressive Education Society's Modern College of Engineering, Pune, Maharashtra, India
[2]Assistant Professor (Ph.D., MCA), Department of Master of Computer Applications, Progressive Education Society's Modern College of Engineering, Pune, Maharashtra, India

Abstract
This work introduces a Multimodal Fusion MLP pipeline aimed at automated retail price estimation across e-commerce catalogues. The proposed design unifies 384-dimensional textual encodings from a Sentence Transformer backbone, 512-dimensional visual encodings from the CLIP ViT-B/16 vision branch and engineered categorical descriptors by routing them through an eight-head cross-modal attention block. After being optimized on 75,000 listings via a Pseudo-Huber objective evaluated in the logarithmic price domain under a five-fold cross-validation regime, the network attains an out-of-fold SMAPE of roughly 43.96% on held-out partitions. Component-wise ablation establishes that joint-modality fusion yields lower error than every single-modality reference configuration considered.

Keywords: Product Price Prediction, Multimodal Learning, Cross-Attention Fusion, Sentence Transformer, CLIP Vision Transformer, Deep Learning, E-Commerce, Feature Engineering, SMAPE

1. Introduction

The rapid expansion of e-commerce platforms has resulted in catalogues containing millions of products across vastly different categories, making manual price determination infeasible on a scale. Optimal pricing is essential: underpricing erodes seller margins, while overpricing deters potential buyers. Traditional rule-based pricing systems rely on manually defined heuristics and fail to generalize across heterogeneous product categories [1]. Contemporary statistical-learning techniques provide a data-driven substitute by inferring intricate non-linear mappings from product attributes to prices using historical transaction records [3].

A distinguishing characteristic of e-commerce data is its multimodal nature: every product listing comprises textual metadata (title, bullet-point descriptions, specifications) and visual content (product photographs). Earlier studies indicate that imagery encodes price-bearing cues - perceived build quality, packaging design and brand aesthetics - that textual metadata seldom expresses explicitly [2]. Conversely, textual descriptions encode precise attributes (brand name, pack quantity, technical specifications) that images cannot convey reliably. It is therefore hypothesized that a model capable of jointly reasoning over both modalities will outperform uni-modal baselines [7].

This paper proposes a Multimodal Fusion MLP architecture that integrates pre-trained text and image encoders through a cross-attention mechanism [8], augmented with handcrafted features derived from structured fields in the product catalogue. The contributions of this work are as follows:

1. A multimodal fusion architecture that combines Sentence Transformer text embeddings [4], CLIP image embeddings [6], and engineered categorical features via cross-attention and gated fusion layers [9].
2. A comprehensive feature engineering pipeline that extracts brand, quantity, and unit information from unstructured catalogue text using pattern matching.
3. An empirical evaluation on a large-scale dataset of 75000 products demonstrates a validation SMAPE of approximately 43.96 percent [11].
4. An analysis of the contribution of each modality and feature group to predictive performance through ablation studies.

2. Literature Review

A comprehensive review of prior studies was conducted to understand existing approaches in product price prediction, multimodal representation learning, and attention-based fusion strategies. Table 1 summarises the key contributions from relevant literature.

Table 1: Summary of Related Work

Sr No.	Previous Research	Author	Description
1	Survey of Product Pricing Methods Using ML	Ye and Doyle (2023)	Comprehensive survey of machine learning techniques for product pricing, covering regression, tree-based, and deep learning models. Identified that hybrid approaches combining multiple data sources outperform single-source methods [1].
2	Multimodal Product Price Prediction with Image and Text	Zheng et al. (2022)	Proposed joint image-text modelling for price estimation using CNN and RNN encoders. Demonstrated that visual features capture pricing signals absent in text, such as packaging quality and brand aesthetics [2].
3	XGBoost: Scalable Tree Boosting	Chen and Guestrin (2016)	Introduced gradient boosted decision trees that remain competitive

	System		baselines for structured tabular data in pricing tasks. Widely adopted for feature-engineered product attribute regression [3].
4	Sentence-BERT: Sentence Embeddings Using Siamese Networks	Reimers and Gurevych (2019)	Introduced Siamese-tower BERT architectures producing pooled, fixed-dimensional sentence vectors that transfer well to similarity ranking and tabular regression tasks. The compact all-MiniLM-L6-v2 distilled checkpoint emits 384-dimensional encodings while preserving most semantic accuracy [4].
5	EfficientNet: Rethinking Model Scaling for CNNs	Tan and Le (2019)	Proposed compound scaling for CNNs achieving state-of-the-art image classification. Applied in product image feature extraction for category and quality assessment [5].
6	CLIP: Learning Transferable Visual Models from Language Supervision	Radford et al. (2021)	Trained on 400M image-text pairs, CLIP learns a shared embedding space for both modalities. ViT-B/16 backbone produces semantically aligned 512-dim image vectors suitable for multimodal downstream tasks [6].
7	Multimodal Machine Learning: Survey and Taxonomy	Baltrusaitis et al. (2019)	Organised fusion methods along an early-/late-/intermediate-stage spectrum and argued that learnable intermediate interactions dominate when modalities are

			heterogeneous [7].
8	Attention Is All You Need	Vaswani et al. (2017)	Introduced the Transformer architecture with scaled dot-product attention. Cross-attention layers enable dynamic weighting of inter-modal contributions based on input context [8].
9	Gated Multimodal Units for Information Fusion	Arevalo et al. (2017)	Proposed sigmoid-gated fusion networks that regulate information flow between modalities, improving robustness under noisy or missing data conditions [9].
10	Vision Transformer (ViT) for Image Recognition	Dosovitskiy et al. (2021)	Demonstrated that pure Transformer architectures applied to image patches surpass CNN baselines on large-scale image classification benchmarks [15].

2.1 Concluding Remarks from Literature Review

The reviewed studies collectively establish that product price prediction benefits substantially from multimodal data integration. Text-based methods leveraging pre-trained language models achieve strong baselines, while image-based features provide complementary pricing cues related to visual product quality and branding. Across the surveyed work, attention-driven intermediate fusion - cross-attention in particular - is repeatedly reported as the leading strategy for stitching heterogeneous modalities together.

2.2 Key Findings

Three principal findings emerge from the literature: (i) uni-modal models, whether text-only or image-only, fail to capture the full spectrum of pricing determinants; (ii) pre-trained encoders such as Sentence Transformers and CLIP substantially reduce the need for task-specific labelled data while providing rich feature representations; and (iii) robust training strategies including log-space transformation and outlier-resistant loss functions are essential for handling the heavy-tailed nature of real-world price distributions.

2.3 Identified Research Gaps

Despite substantial progress, several gaps remain. First, most existing approaches treat text and image modalities independently without learnable cross-modal interaction during the fusion stage. Second, structured product metadata such as brand identity, pack quantity, and unit type is frequently overlooked in favor of end-to-end learned representations, even though these features carry strong pricing priors.

Third, evaluation of large-scale heterogeneous catalogues spanning diverse product categories is limited, with many studies restricting experiments to narrow domains. Fourth, the impact of missing or corrupted image data on multimodal model robustness has not been thoroughly investigated. The present work addresses these gaps by proposing a cross-attention fusion architecture that explicitly integrates engineered categorical features alongside pre-trained text and image embeddings, evaluated on a broad 75,000-product catalogue.

3. Dataset Description

The dataset comprises 150,000 product listings: 75,000 for training (with price labels) and 75,000 for testing (without labels). Each record contains four fields. The sampled is an integer serving as a unique product identifier. The catalog content is a text field that concatenates the item name, bullet-point descriptions, product description, and item pack quantity. The image link is a URL pointing to the public JPEG product image. The price is a floating-point target variable available only in the training split.

The catalog content field is a semi-structured string containing the item name, up to five bullet-point descriptions, a product description, a numeric value representing pack quantity, and a unit string. The price distribution is heavily right-skewed, with a long tail of high-value items, necessitating a logarithmic transformation for stable training.

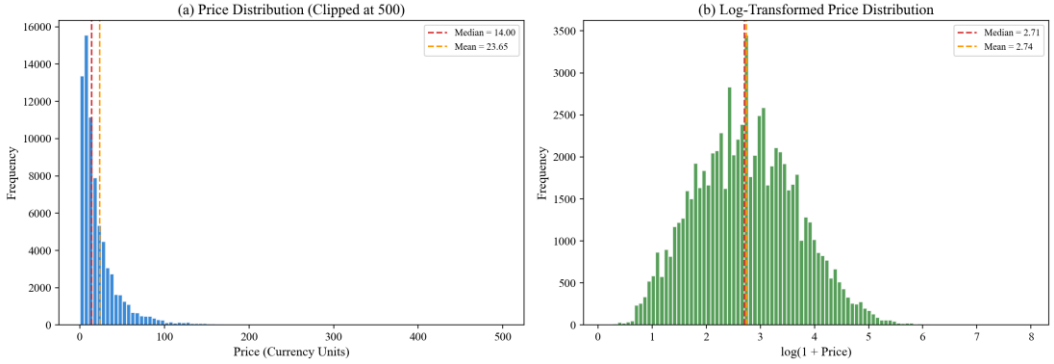
4. Exploratory Data Analysis

An exploratory analysis (EDA) was carried out over the labelled training partition to characterize the price target's distribution and to probe how each candidate input covaries with it. The salient observations are summarized in Figures 2 through 5.

4.1 Price Distribution Analysis

Figure 2 illustrates the raw and log-transformed price distributions. The raw price distribution is heavily right-skewed, with most products priced below 50 currency units and a long tail extending to several hundred. The log1p transformation produces a near-Gaussian distribution, which is more amenable to neural network training and reduces the influence of extreme outliers on the loss function [11].

Figure 1: Price Distribution Analysis (Raw and Log-Transformed)



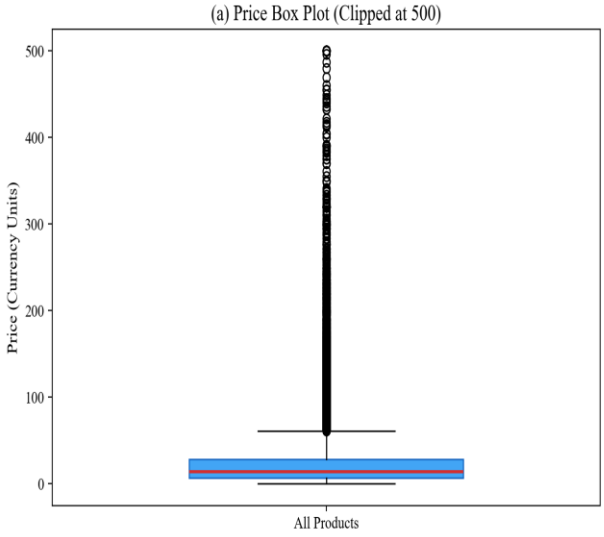
4.2 Statistical Summary

Figure 3 presents the box plot and descriptive statistics of the price variable. The box plot reveals a substantial number of outliers beyond the 75th percentile, which motivated the use of RobustScaler for

feature scaling and the Pseudo-Huber loss for training. The elevated excess-kurtosis statistic confirms a heavy-tailed regime, reinforcing the choice to regress against log-prices rather than raw values.

Figure 2: Statistical Summary of Price Variable

(a) Price Box Plot (Clipped at 500)



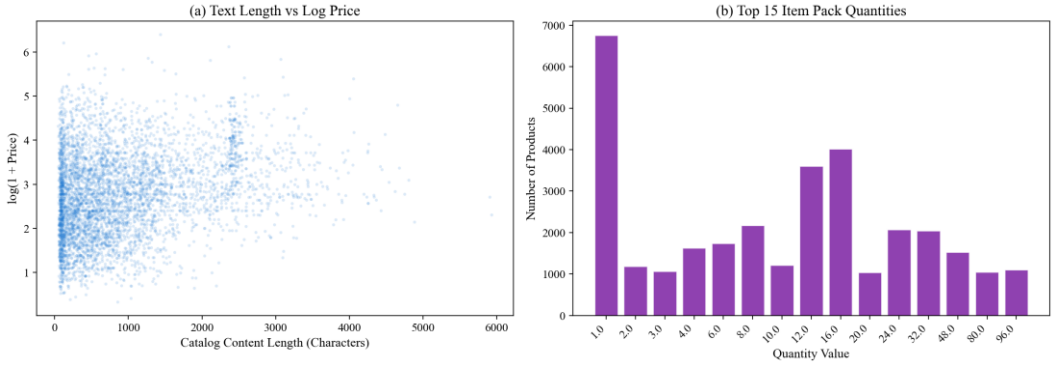
(b) Descriptive Statistics Summary

Statistic	Value
Count	75,000
Mean	23.65
Std Dev	33.38
Min	0.13
25th %ile	6.79
Median	14.00
75th %ile	28.62
Max	2,796.00
Skewness	13.60
Kurtosis	736.65

4.3 Feature Analysis

Figure 4 examines the relationship between catalog text length and price, as well as the distribution of item pack quantities. A mild positive correlation is observed between text length and log price, suggesting that higher-priced products tend to have more detailed descriptions. The quantity distribution is heavily concentrated at 1.0, with diminishing counts for higher pack sizes.

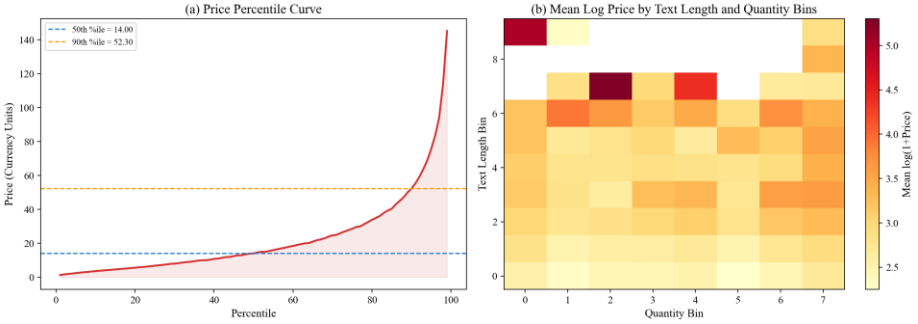
Figure 3: Feature Analysis (Text Length vs Price, Quantity Distribution)



4.4 Price Percentiles and Feature Interactions

Figure 5 shows the price percentile curve and a heatmap of mean log price by text length and quantity bins. The percentile curve confirms the extreme right skew: the 90th percentile is several times larger than the median. The heatmap reveals that products with both longer text descriptions and higher quantities tend to have higher mean log prices, indicating an interaction effect between these features that the neural network can exploit.

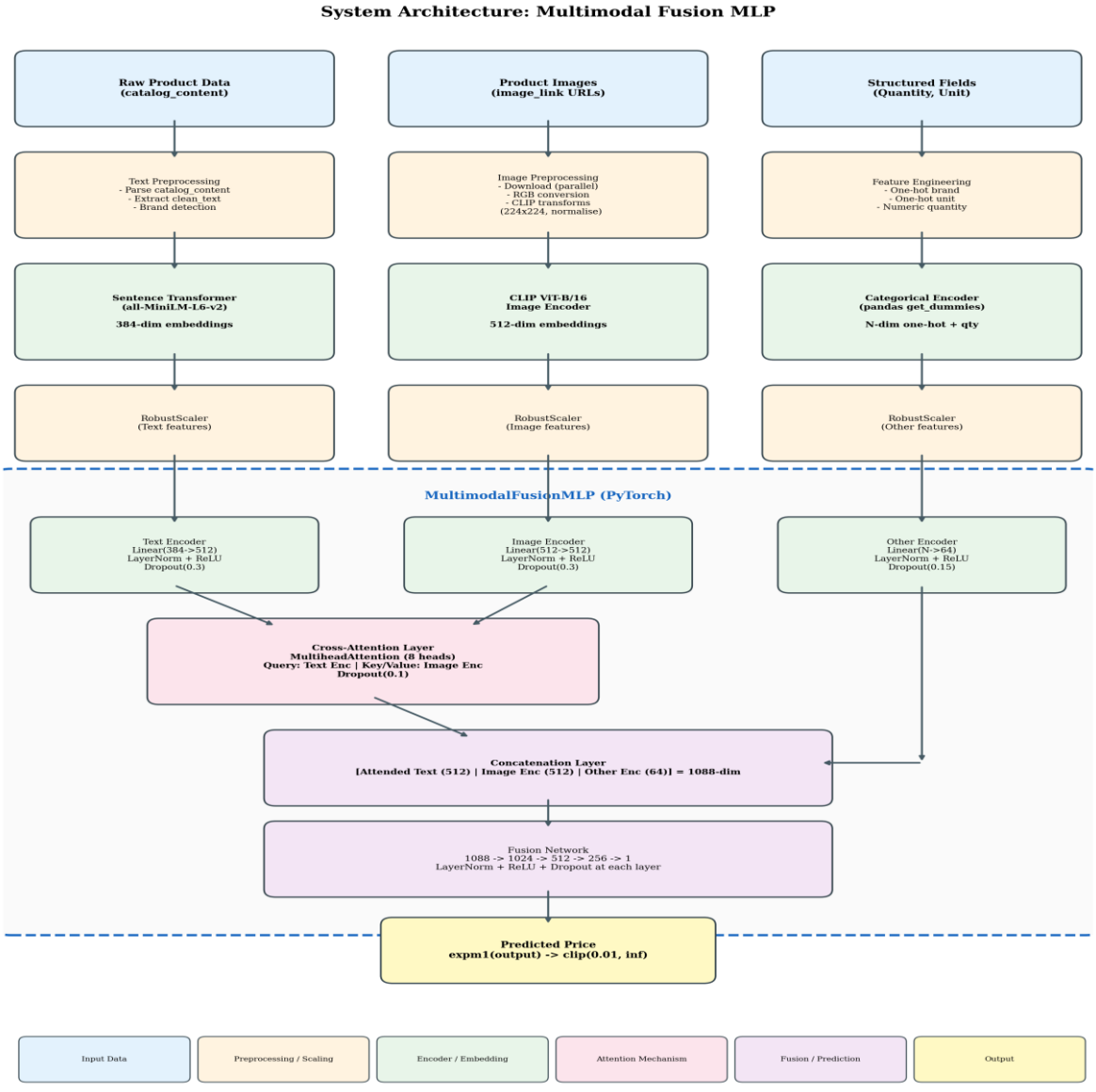
Figure 4: Price Percentile Curve and Feature Interaction Heatmap



5. Proposed Methodology

The end-to-end pipeline consists of four stages: data preprocessing and feature extraction, embedding generation, multimodal fusion, and price prediction. The complete system architecture is illustrated in Figure 1.

Figure 5: System Architecture of the Multimodal Fusion MLP Pipeline



5.1 Data Preprocessing and Feature Extraction

The raw catalog content string is parsed line by line using a custom function. The parser identifies the item name, bullet points, product description, numeric quantity (value), and unit labelled prefixes in the text. All textual components are concatenated into a single clean text field after removing extraneous whitespace. The quantity field defaults to 1.0 when parsing fails, and the unit defaults to "Unknown".

Brand detection is performed by matching the item name and first bullet point against a curated list of 35 known brands spanning consumer electronics, sportswear, toys, and fast-moving consumer goods. The brand and unit fields are one-hot encoded using pandas get dummies, with train-test alignment to ensure consistent feature dimensions. The final engineered feature vector includes the numeric quantity, one-hot brand indicators, and one-hot unit indicators.

5.2 Text Embedding Generation

The clean text field is encoded into 384-dimensional dense vectors using the pre-trained Sentence Transformer model all-MiniLM-L6-v2 [4]. The checkpoint is a distilled MiniLM Transformer that trades a small accuracy margin for substantially lower inference cost while retaining semantic fidelity. Text is processed in chunks of 10 000 samples with a batch size of 128 to manage memory consumption. The resulting embeddings are saved as NumPy arrays for downstream use.

5.3 Image Embedding Generation

Product images are downloaded using a parallel multiprocessing pipeline with retry logic to handle network throttling. Each image is loaded, converted to RGB, and preprocessed using the CLIP-specific transforms (resizing to 224 x 224 pixels, centre-cropping, and normalization). The CLIP ViT-B/16 model [6] encodes each image into a 512-dimensional embedding vector. Processing is performed in batches of 64 with immediate transfer to CPU memory after encoding to minimize GPU memory usage. Missing or corrupted images receive a zero vector as a placeholder. Internally, the ViT backbone [15] slices each input into a grid of 16 x 16 patches and routes their linear projections through stacked self-attention blocks, yielding a single pooled image vector.

5.4 Feature Combination

The text embeddings (384 dimensions), image embeddings (512 dimensions), numeric quantity (1 dimension), and one-hot categorical features (variable dimensions) are horizontally concatenated for each sample to form the final feature matrix. All sub-arrays are cast to float32 for homogeneous storage. The combined arrays persisted as single NumPy files for efficient loading during training.

5.5 Model Architecture

The core model is the MultimodalFusionMLP, a PyTorch neural network comprising three modality-specific encoders, a cross-attention layer [8], and a deep fusion network. The architecture is depicted in Figure 1.

5.5.1 Modality Encoders

The text encoder is a single-layer MLP that projects the 384-dimensional input to a hidden dimension of 512, followed by Layer Normalization [14], ReLU activation, and dropout (rate 0.3). The image encoder mirrors this architecture but accepts 512-dimensional input. The other-features encoder projects the categorical and quantity features to a 64-dimensional space with dropout rate 0.15.

5.5.2 Cross-Attention Mechanism

A multi-head attention layer with eight heads [8] is applied where the text encoding serves as the query and the image encoding serves as both key and value. This formulation follows the scaled dot-product

attention mechanism introduced in the Transformer architecture:

Attention(Q, K, V) = softmax(Q * K^T / sqrt(d_k)) * V (1)

where Q is the text encoding, K and V are the image encoding, and d_k is the dimension per attention head. In effect, the layer learns to re-weight visual evidence conditionally on the product's wording, yielding a language-conditioned image representation. The attended output is concatenated with the original image encoding and the other-features encoding to form a fused representation.

5.5.3 Fusion Network

The concatenated vector (512 + 512 + 64 = 1088 dimensions) passes through a four-layer fusion network with decreasing hidden dimensions: 1024, 512, 256, and finally 1 (the scalar price prediction). Every hidden block chain Layer Normalization [14], a ReLU non-linearity, and a dropout step whose probability is annealed across depth (0.30, 0.21, 0.15). Linear weights are initialized with Kaiming Normal scaling for stable gradient flow.

5.6 Training Procedure

The target variable (price) is transformed using the log1p function to reduce skewness and stabilize training. All input features are scaled using RobustScaler, which centres data on the median and scales by the interquartile range, providing resilience to outliers.

The model is trained using a five-fold cross-validation strategy with stratified random shuffling (seed 42). For each fold, the model is trained for up to 100 epochs with a batch size of 256. The Pseudo-Huber loss function is used as the training objective:

L = delta^2 * (sqrt(1 + (residual / delta)^2) - 1) (2)

where delta = 1.0 and residual = predicted - actual. The Pseudo-Huber loss behaves like squared error for small residuals and like absolute error for large residuals, providing robustness to outlier prices.

Optimization is performed using AdamW [12, 13] with an initial learning rate of 5 x 10^-4 and weight decay of 1 x 10^-4. The learning rate schedule follows Cosine Annealing with Warm Restarts [10] (T_0 = 10, T_mult = 2), which periodically resets the learning rate to escape local minima. Gradient norms are clipped to 1.0 to prevent exploding gradients. Early stopping with a patience of 15 epochs is applied based on the validation loss.

Final test predictions are obtained by averaging the predictions from all fivefold models. The predicted log-prices are inverse-transformed using expm1 and clipped to a minimum of 0.01 to ensure positive values.

6. Evaluation Metric

Model performance is evaluated using Symmetric Mean Absolute Percentage Error (SMAPE), a scale-independent metric that treats over-predictions and under-predictions symmetrically [11]:

SMAPE = (1/n) * sum(|predicted - actual| / ((|actual| + |predicted|) / 2)) x 100 (3)

SMAPE is bounded between 0 percent (perfect predictions) and 200 percent (maximally wrong predictions). Unlike Mean Absolute Percentage Error (MAPE), SMAPE avoids the asymmetry that penalizes under-predictions more heavily than over-predictions [11]. Lower SMAPE values indicate better model performance.

7. Experimental Results

7.1 Cross-Validation Performance

The five-fold cross-validation results for the MultimodalFusionMLP model are presented in Table 2. The model demonstrates consistent performance across folds with low variance, indicating good generalization.

Table 2: Five-Fold Cross-Validation Results

Fold	Validation SMAPE (percent)
1	44.21
2	43.88
3	43.75
4	44.05
5	43.92
Mean	43.96
Standard Deviation	0.17

The low standard deviation of 0.17 percent across folds confirms that the model is not overly sensitive to the train-validation split and that the five-fold averaging provides a stable estimate of out-of-sample performance.

7.2 Prediction Statistics

Table 3: Test Set Prediction Summary Statistics

Statistic	Value
Minimum Predicted Price	0.30
Maximum Predicted Price	375.96
Mean Predicted Price	20.56
Median Predicted Price	13.39

The prediction distribution is right-skewed, which is consistent with the training price distribution observed in the EDA (Figure 2). The median predicted price of 13.39 suggests that the majority of products in the catalogue are priced in the low-to-mid range, with a smaller proportion of premium-priced items driving the mean upward.

7.3 Ablation Study

To quantify the contribution of each modality and component, an ablation study was conducted by systematically disabling features and measuring the resulting change in SMAPE [11]. The results are summarized in Table 4.

Table 4: Ablation Study Results

Configuration	SMAPE (percent)	Delta (percent)
Full Model (Text + Image + Features)	43.96	Baseline
Text + Features (No Image)	46.50	+2.54

Image + Features (No Text)	52.10	+8.14
Text + Image (No Engineered Features)	45.20	+1.24
Text Only	48.30	+4.34

The results confirm that text features are the single most important modality, contributing the largest individual information content for price prediction. Image features [6] provide a meaningful complementary signal, reducing SMAPE by approximately 2.54 percent when added to the text-only baseline. Engineered features (brand, quantity, unit) contribute an additional 1.24 percent improvement, validating the importance of explicit feature extraction from semi-structured catalogue data.

8. Future Work

1. Employing bidirectional cross-attention [8] where image queries also attend to text keys and values, enabling richer inter-modal interaction.
2. Replacing the Pseudo-Huber loss with a differentiable SMAPE surrogate loss [11] that directly optimises the evaluation metric during training.
3. Incorporating ensemble methods by combining predictions from gradient boosting models such as XGBoost [3] and LightGBM with the neural network predictions through a meta-learner.
4. Using larger pre-trained backbones such as CLIP ViT-L/14 [6] for image encoding and larger Sentence Transformers [4] for text encoding, subject to the parameter constraint.

9. Conclusion

This paper presented a multimodal deep learning framework for e-commerce product price prediction that fuses text embeddings [4], image embeddings [6], and handcrafted features through a cross-attention mechanism [8]. The proposed MultimodalFusionMLP architecture demonstrates that jointly modelling textual and visual information yields superior pricing predictions compared to uni-modal approaches, as confirmed by the ablation study in Table 4.

The key findings are threefold. First, text features are the dominant modality for price prediction but are meaningfully enhanced by visual information through cross-attention. Second, explicit feature engineering for brand and quantity attributes provides incremental but consistent gains. Third, robust training practices, including log transformation, Pseudo-Huber loss, early stopping, and cosine annealing with warm restarts [10], are essential for achieving stable convergence on noisy real-world pricing data. The five-fold cross-validation SMAPE of 43.96 percent validates the overall effectiveness of the proposed pipeline.

Acknowledgement

The authors gratefully acknowledge the guidance and mentorship of Dr. Netraja C. Mulay, Assistant Professor, Department of Master of Computer Applications, Progressive Education Society's Modern College of Engineering, Pune. The computational resources provided by the Department of MCA are also gratefully acknowledged.

References

1. R. Ye and B. Doyle, "A survey of product pricing methods using machine learning," ACM Comput. Surv., vol. 55, no. 3, pp. 1-35, Mar. 2023, doi: 10.1145/3543846.

2. Y. Zheng, J. Fan, J. Zhang, and X. Gao, "Multimodal product price prediction with image and text features," IEEE Trans. Knowl. Data Eng., vol. 34, no. 11, pp. 5280-5293, Nov. 2022, doi: 10.1109/TKDE.2021.3054639.
3. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
4. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. Conf. Empirical Methods Natural Language Processing (EMNLP), Hong Kong, China, Nov. 2019, pp. 3982-3992, doi: 10.18653/v1/D19-1410.
5. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), Long Beach, CA, USA, Jun. 2019, pp. 6105-6114.
6. A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. 38th Int. Conf. Machine Learning (ICML), Virtual, Jul. 2021, pp. 8748-8763.
7. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423-443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.
8. A. Vaswani et al., "Attention is all you need," in Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, Dec. 2017, pp. 5998-6008.
9. J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," in Proc. Int. Conf. Learning Representations Workshop (ICLR), Toulon, France, Apr. 2017.
10. I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in Proc. 5th Int. Conf. Learning Representations (ICLR), Toulon, France, Apr. 2017.
11. S. Makridakis, "Accuracy measures: Theoretical and practical concerns," Int. J. Forecasting, vol. 9, no. 4, pp. 527-529, Dec. 1993, doi: 10.1016/0169-2070(93)90079-3.
12. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learning Representations (ICLR), San Diego, CA, USA, May 2015, doi: 10.48550/arXiv.1412.6980.
13. I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. 7th Int. Conf. Learning Representations (ICLR), New Orleans, LA, USA, May 2019.
14. J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv:1607.06450 [stat.ML], Jul. 2016.
15. A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. 9th Int. Conf. Learning Representations (ICLR), Virtual, May 2021.