CSCI 566 - Final Project Report

Rizwan Ahasan Pathan
rpathan@usc.edu

Abstract

This project presents a gated multimodal deep learning framework for predicting restaurant star ratings. Unlike traditional approaches that rely on structured metadata or review-based heuristics, this work integrates four complementary data modalities: business metadata, textual reviews, restaurant images, and geographic ZIP-code information. Each modality is encoded using a domain-specific architecture—MLP for structured features, BERT for text, ResNet-50 for images, and a Graph Neural Network (GNN) for spatial relationships. A gated fusion module learns dynamic scalar weights for each modality, allowing the model to adaptively emphasize the most informative signals. The architecture is optimized via Optuna hyperparameter search and trained on over 36,000 businesses, 4.3 million reviews, and 200,000 photos. The final model achieves an RMSE of 0.2654, MAE of 0.2042, and an R^2 of 0.8940 on the test set. These results demonstrate that combining multimodal content with spatial context leads to accurate and robust predictions of restaurant quality.

1 Introduction:

Problem Statement

Online restaurant ratings play a central role in shaping consumer decisions, business reputation, and urban food trends. However, platforms like Yelp.com heavily depend on subjective user reviews, which often overlook more objective signals such as food safety, service quality, visual presentation, affordability, and geographic context. This reliance on incomplete or biased data can mislead consumers, penalize businesses unfairly, and obscure broader trends valuable to investors and policymakers.

To address these gaps, this project proposes a scalable, multimodal deep learning framework that integrates multiple data sources—to predict a restaurant's true quality. This approach moves beyond traditional heuristics and offers a data-driven way to holistically assess restaurant success.

Why It's Important

- Biased or noisy ratings can mislead consumers and unfairly penalize businesses.
- More accurate predictions create value for everyone involved—platforms, consumers, and businesses alike.

Key Challenges

- **Multimodal fusion**: Designing an architecture that effectively combines structured data, text, images, and geographic features.
- **Data sparsity**: Some businesses lack images or detailed reviews, requiring robust handling of missing modalities.
- **Class imbalance**: Only a fraction of restaurants rank among the most successful, demanding careful sampling or weighting strategies.

- **Nonlinear dependencies**: Capturing complex relationships between features that traditional models often miss.
- **Computational demands**: Deep models (e.g., transformers and CNNs) require significant GPU resources for training and tuning.
- **Interpretability**: Ensuring that model outputs are understandable and trustworthy to non-technical stakeholders.

Who Benefits

- **Consumers**: Gain access to cleaner, more accurate recommendations.
- **Restaurateurs and Investors**: Identify high-potential locations and operational levers for success.
- **Urban Planners and Policymakers**: Assess neighborhood food equity and support data-informed interventions.
- **Researchers**: Leverage a robust benchmark for future work in multimodal, location-aware recommendation systems.

2 Background and Related Work:

2.1 Traditional Machine Learning Methods

Early approaches to restaurant rating prediction relied heavily on traditional machine learning models such as Random Forests and Logistic Regression. These models primarily used structured metadata—features like review counts, star averages, and price tiers—to estimate ratings. While interpretable and computationally efficient, such models inherently lack the capacity to capture semantic meaning from reviews or visual appeal from images.

Limitations

- Fail to model complex, nonlinear relationships in real-world data.
- Cannot process unstructured content like user reviews or images.

Proposed Contribution

This work addresses these limitations by replacing shallow models with a deep neural architecture that can learn from structured data (via an MLP) while also incorporating unstructured data such as text and images.

2.2 Text-Only Deep Learning Models

Several studies have leveraged natural language processing techniques to predict star ratings based solely on review text. For example, Liu [1] (2020) applied recurrent neural network–based models to extract sentiment and predict scores, showing improvements over classical baselines. More recent work has adopted transformer-based models such as BERT for this task, significantly improving contextual understanding.

Limitations

- Ignore structured business metadata such as location, hours, or price tier.
- Overlook visual signals that can strongly influence consumer perception (e.g., food presentation).

Proposed Contribution

This model extends beyond text-only approaches by incorporating both structured metadata and restaurant images, enabling a more holistic assessment of business quality.

2.3 Multimodal Deep Learning Models

Multimodal models aim to combine multiple data sources to better represent user experiences. Zhao et al. [2] (2024) proposed an attention-based architecture for point-of-interest (POI) recommendation that integrates textual and visual signals. Their work demonstrated the value of modality fusion but focused on classification tasks and excluded structured business attributes.

Limitations

- Primarily focused on classification or recommendation rather than regression.
- Lacked integration of structured business features such as pricing, operating hours, or categories.
- Often evaluated on small or synthetic datasets, limiting generalizability.

Proposed Contribution

A regression-focused multimodal fusion model is introduced that integrates four complementary modalities: structured business metadata (via an MLP), textual reviews (via BERT), restaurant images (via ResNet-50), and geographic context (via ZIP-code–level GNN embeddings). The model is trained and evaluated on the full Yelp Open Dataset , demonstrating strong generalization, scalability, and improved robustness through cross-modal representation.

2.4 Methodological Limitations and Proposed Solutions

Table 1: Summary of methodological limitations in prior work and corresponding proposed solutions

Method	Key Limitations	Proposed Solution
Traditional ML	Fails to capture nonlinear relationships; ignores text and images	Use a deep MLP to model structured features more flexibly
Text-Only Deep Models	Omits structured metadata and visual content	Integrate BERT-based textual representations alongside metadata and images
Visual-Only CNNs	Underperforms when used in isolation	Fuse CNN-based visual features with textual and structured inputs
Spatial-Free Models	Misses geographic or spatial context (e.g., neighborhood effects)	Incorporate ZIP-code–level GNN embeddings to inject local context
Prior Multimodal Work	Primarily focused on classification; often excludes metadata or location information	Develop a regression-oriented model with full modality fusion, including GNN features

By addressing these gaps, this project introduces a novel multimodal regression framework for restaurant quality prediction that is both scalable and interpretable. The proposed model unifies structured metadata, textual reviews, and restaurant images within an end-to-end architecture, enabling robust performance across diverse business contexts.

3 Methods

3.1 Data Modalities

To predict restaurant star ratings, the proposed model integrates following complementary data modalities derived from the Yelp Open Dataset:

- **Structured metadata**: A 670-dimensional feature vector capturing static business attributes such as review count, price tier, categories, and geolocation.
- **Textual reviews**: User-generated content reflecting sentiment, service quality, food experience, and customer satisfaction. Text inputs are tokenized and truncated to a maximum of 512 tokens for processing with transformer-based models.

- **Restaurant images**: Photographs depicting food, ambience, and environment. Images are resized to 224×224 RGB format and processed through convolutional neural networks to extract visual features.
- **Geographic context**: Each business is associated with its ZIP code, for which a graph neural network (GNN) learns spatial embeddings. The ZIP-code graph is constructed using real latitude–longitude distances, capturing local neighborhood influence and regional trends.

Each modality contributes distinct and complementary information. Structured metadata provides factual descriptors, textual reviews convey subjective narratives, images reveal visual quality, and GNN embeddings encode spatial and neighborhood-level characteristics. This multimodal integration enables a more expressive and robust representation than any single modality alone.

3.2 Model Architecture

The proposed model is a multimodal fusion neural network composed of domain-specific encoders. Each encoder processes a distinct input modality and outputs a fixed-size embedding.

Table 2: Overview of modality-specific encoders and embedding dimensions

Modality	Input	Encoder	Output Shape
Metadata	670-dimensional feature vector	MLP (3 layers, ReLU activations)	256-dimensional
Text	Review text (≤ 512 tokens)	BERT (pretrained and fine-tuned)	768-dimensional
Image	224×224 RGB image	ResNet-50 (pretrained on ImageNet)	2048-dimensional
Geography	ZIP code with latitude/longitude context	GNN over ZIP-code graph (2-layer GCN)	32-dimensional

These encoders produce embeddings that capture semantic, structural, visual, and spatial patterns associated with restaurant quality. The architecture employs a gated fusion mechanism in which each modality is dynamically weighted based on its informativeness during inference. This enables the model to emphasize strong signals while down-weighting weaker or missing inputs. The modular design also supports ablation studies and modality-specific enhancements, making the system adaptable to future advances in multimodal learning and urban analytics.

As illustrated in Figure 1, the proposed model integrates multiple modality-specific encoders through a gated fusion mechanism.

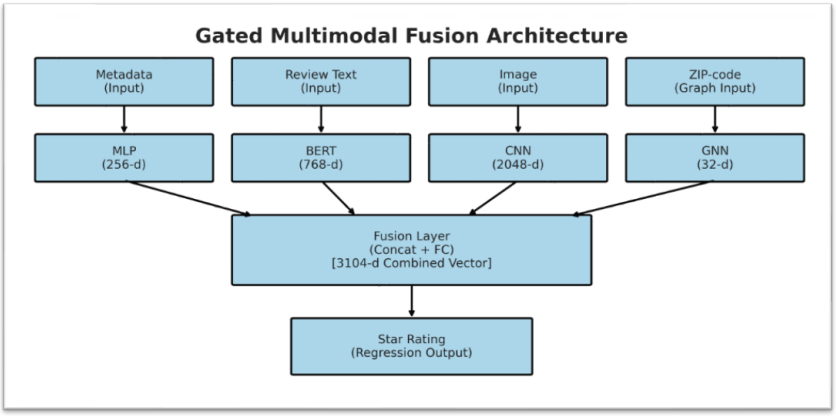

Figure 1: Gated Multimodal Fusion Architecture. Each modality is processed by a domain-specific encoder (MLP, BERT, CNN, and GNN), producing fixed-size embeddings that are dynamically weighted and fused before final star rating regression

3.3 Fusion Strategy

A gated multimodal fusion strategy is used to combine information from all four modalities. Each modality's embedding is passed through a learnable scalar gate that dynamically modulates its contribution based on informativeness. By selectively amplifying the most informative features while diminishing the impact of less relevant or absent data, the model maintains high predictive stability. This gating mechanism ensures resilience to noise and variability across heterogeneous input modalities.

After gating, the weighted embeddings are concatenated into a 3104-dimensional fused vector, which is passed through a fully connected regression head consisting of:

- Dropout for regularization
- Dense layers with batch normalization and ReLU activations
- A final linear layer producing the continuous star rating prediction

This gated fusion mechanism significantly improves robustness compared to naive concatenation. It allows the model to:

- adaptively handle missing modalities,
- reduce noise amplification, and
- yield more interpretable modality contributions (e.g., BERT contributing 45.8% and GNN approximately 4.4%).

3.4 Training Setup

The model is trained end-to-end using the following configuration, consistent with the final tuned values.

Optimization and Loss

- **Loss function**: Mean Squared Error (MSE)
- **Optimizer**: AdamW
- **Learning rate**: Tuned via Optuna (best $\approx 4.1 \times 10^{-4}$)
- **Weight decay**: Tuned ($\approx 9 \times 10^{-4}$)
- **Batch size**: 256
- **Epochs**: 200
- **Dropout**: Tuned (≈ 0.59), applied in the fusion layers

Evaluation

- **Metrics**: RMSE, MAE, R^2, Pearson correlation
- **Validation frequency**: Every epoch
- **Normalization**: Target ratings normalized using dataset mean and standard deviation

Train–Test Split

- 80/20 business-level split, ensuring:
  - No business appears in both training and test sets
  - The natural star rating distribution is preserved

This setup yields stable convergence and strong generalization despite the dataset's skew toward 4–5 star ratings. Optuna-based hyperparameter tuning significantly improved performance, reducing validation RMSE from approximately 0.2821 to 0.2607.

3.5 Hyperparameter Tuning (Optuna)

To optimize performance and generalization, an automated hyperparameter search was conducted using Optuna over 20 trials. The best configuration was identified in Trial 12, achieving a validation RMSE of 0.2607 with the following settings:

- **Hidden dimension**: 512
- **Dropout rate**: 0.595
- **Learning rate**: 0.00041
- **Weight decay**: 8.97×10^{-4}

These parameters were used to train the final Gated Fusion model.

3.6 Implementation Notes

Hardware: All model training was executed on an Tesla V100-PCIE-32GB GPU.

Modality Preprocessing

Each input modality follows a distinct preprocessing pipeline. Text data is tokenized using the `bert-base-uncased` tokenizer, truncated to a maximum of 512 tokens, and encoded into 768-dimensional embeddings. Image data is resized to 224×224 pixels, normalized using ImageNet mean and standard deviation, and processed through a ResNet-50 encoder to extract 2048-dimensional feature vectors.

Structured metadata, consisting of 670 attributes, is standardized using Z-score normalization and passed through a three-layer multilayer perceptron (MLP) to produce 256-dimensional embeddings. For the ZIP-code modality, real geographic coordinates are used to construct a graph representing spatial relationships, and a graph convolutional network (GCN) encoder generates 32-dimensional embeddings for each ZIP region.

Reproducibility

Reproducibility was ensured by fixing random seeds across all experiments, storing all model checkpoints, automatically logging loss curves, prediction plots, and modality contribution statistics, and saving fusion embeddings for downstream analysis.

These implementation practices ensure reproducibility, stability, and scalable training across millions of examples.

3.7 Validation Strategy

A rigorous validation procedure is employed to ensure the absence of data leakage and fair model evaluation. Train–test splits are performed at the `business_id` level, ensuring that all reviews, images, and metadata associated with a given restaurant appear exclusively in either the training or test set.

Model performance is evaluated using RMSE and MAE to measure prediction error, R^2 to quantify explained variance, and Pearson correlation to assess ranking consistency. To control overfitting, validation performance is monitored at every training epoch. Performance stabilizes between epochs 160 and 200, with dropout and weight decay tuned to improve generalization.

3.8 Challenges Encountered

Throughout the development process, several engineering and data-handling challenges arose that impacted experimentation and model scalability:

- **Dataset migration and alignment**: Transitioning from the locally scraped LA Yelp dataset to the full Yelp Open Dataset required redesigning data pipelines to ensure business-level alignment, eliminate duplicates, and consistently match modalities across entries.

- **Large-scale review processing**: Tokenizing and batching over 4.3 million reviews introduced significant memory usage and preprocessing overhead. This was addressed through GPU-accelerated tokenization and caching strategies to reduce runtime load.
- **High-throughput image processing**: Generating embeddings from thousands of restaurant photos created I/O bottlenecks during training. To mitigate this, batched image preloading and disk-level caching were implemented to reduce latency.
- **Handling incomplete modality coverage**: While some degree of data sparsity was anticipated, managing real-world scenarios where restaurants lacked reviews or images required updates to training logic to gracefully handle partial inputs and maintain batch consistency.
- **Rating distribution skew**: The prevalence of 4–5 star ratings in the dataset introduced challenges in training, particularly when predicting lower-rated businesses. Careful tuning of loss functions and sampling strategies was necessary to reduce this regression bias.
- **Runtime and resource constraints**: Joint training of transformer-based, convolutional, and graph-based components demanded careful GPU memory budgeting. This led to a phased training approach that initially used frozen pretrained encoders before gradual fine-tuning.

4 Results

4.1 Exploratory Experiments (Preliminary Phase)

Before transitioning to the full Yelp Open Dataset, initial experiments were conducted using a smaller, locally scraped Yelp Los Angeles dataset comprising approximately 6,900 businesses. This exploratory phase served to validate architectural design choices and to identify early limitations related to data quality and scale.

4.1.1 Setup

- Structured business metadata, user reviews, and restaurant images were included as input modalities.
- Selected experiments incorporated Los Angeles County health inspection scores to evaluate the feasibility of joint prediction tasks.
- Compared to the Yelp Open Dataset, the preliminary data was more limited in scale and exhibited higher noise levels.

4.1.2 Results Summary

Star Rating Prediction

The metadata-only MLP model exhibited limited predictive performance, reflecting the restricted signal available from structured attributes alone. The BERT-based model trained exclusively on review text achieved moderate accuracy by effectively capturing sentiment-related cues embedded in user feedback. In contrast, the multimodal fusion model, which integrates textual, visual, and structured metadata inputs, delivered the strongest results. This improvement demonstrates that combining complementary modalities substantially enhances star rating prediction accuracy.

Table 3: Preliminary model performance (RMSE, MAE, and R^2) for star rating prediction on the local Yelp LA dataset

Model	RMSE	MAE	R^2
MLP	0.5920	0.4393	0.4791
BERT	0.1211	0.0921	0.4325
Fusion	0.4511	0.3539	0.6976

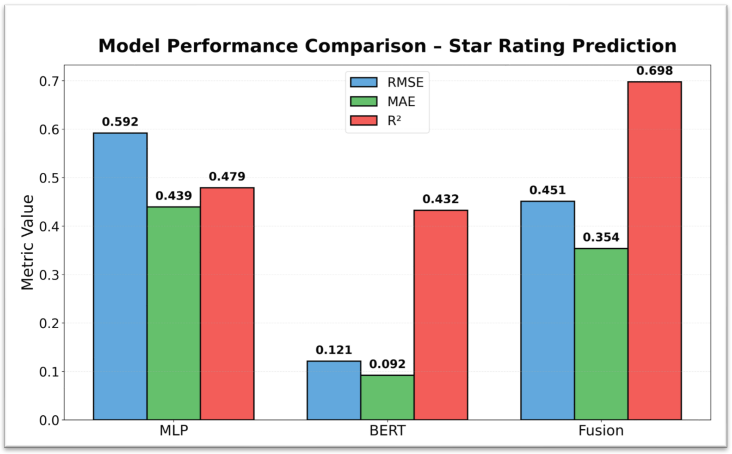

Figure 2: Model performance comparison for star rating prediction during the preliminary phase. The multimodal fusion model outperforms unimodal baselines in terms of R^2, demonstrating the benefit of integrating textual, visual, and structured metadata signals

Inspection Score Prediction

For inspection score prediction, the metadata-only MLP model again performed poorly. The BERT model showed weak predictive ability as well, likely due to noisy inspection labels and limited review–inspection alignment in the dataset. The multimodal fusion model did not improve performance for this task, indicating that inspection scores may not be strongly reflected in text, images. Discrepancies in RMSE values across models were traced to inconsistent label scaling. To ensure a fair comparison, all models were re-evaluated using raw inspection scores.

Table 4: Preliminary model performance (RMSE, MAE, and R^2) for inspection score prediction on the local Yelp LA dataset

Model	**RMSE**	**MAE**	R^2
MLP	3.6139	2.7053	-0.0093
BERT	0.1234	0.0932	0.4107
Fusion	2.8018	2.8489	-0.1170

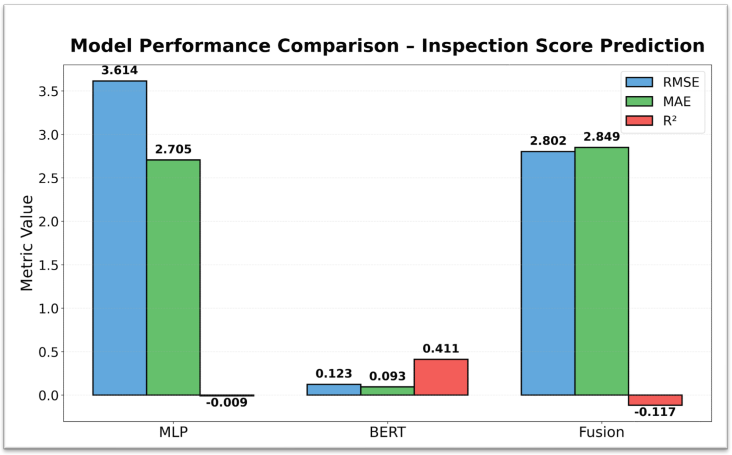

Figure 3: Model performance comparison for inspection score prediction during the preliminary phase. BERT outperforms the metadata-only MLP and multimodal fusion models, indicating stronger alignment between textual reviews and inspection outcomes

Overall, the LA-dataset experiments showed clear differences across modalities. The metadata-only MLP model provided limited predictive value, particularly for inspection scores, where performance was near zero or negative ($R^2 = -0.0093$) and error metrics were high. The BERT model, leveraging review text, performed substantially better for star-rating prediction ($R^2 \approx 0.43$), capturing sentiment and contextual cues that metadata alone could not; however, it struggled on inspection scores due to noisy labels and the small size of the available dataset. CNN features were used exclusively as inputs to the multimodal fusion model rather than evaluated independently. The fusion approach, which integrates metadata, text, and visual features, delivered the strongest results with an R^2 of 0.6976, demonstrating the benefit of combining complementary information sources.

Key Takeaways

- Fusion models consistently outperformed single-modality baselines, even with limited training data.
- Inspection-score prediction was not pursued further because the Yelp Open Dataset does not contain inspection labels.
- Insights from the LA-phase experiments informed the final model architecture, preprocessing pipeline, and large-scale training strategy used in deployment.

Due to these limitations and the absence of inspection labels, the larger and more diverse Yelp Open Dataset was adopted to enable more reliable and scalable evaluation.

4.2 Dataset Summary (Yelp Open Dataset)

The main phase of the project uses the Yelp Open Dataset (2024 release), which contains multimodal data for tens of thousands of businesses across North America. After filtering, the focus is placed on restaurant entries with complete metadata, at least one image, and one review.

Table 5: Dataset summary statistics for the Yelp Open Dataset

Data Component	Count
Businesses	36,680
Reviews	$\sim$4.3 million
Photos	$\sim$200k
Average Reviews/Business	$\sim$117
Label Type	Yelp star rating (1.0 – 5.0)

4.3 Baseline Model Performance

Single-modality baselines were evaluated using only one input source at a time—structured metadata, review text, or images.

MLP (Metadata-Only)

Trained on 670 structured features using a 3-layer MLP:

- RMSE: 0.6046
- MAE: 0.4708
- R^2: 0.4551

Metadata provides limited predictive power, as it lacks user sentiment or visual appeal.

BERT (Text-Only)

Fine-tuned `bert-base-uncased` on raw Yelp reviews (max 512 tokens):

- RMSE: 0.5140

- MAE: 0.3707
- R^2: 0.5012

Text alone improves performance over metadata due to its rich sentiment and contextual information.

CNN (Image-Only)

Trained a ResNet-based regressor on Yelp food and ambience images (resized to 224×224):

- RMSE: 0.5551
- MAE: 0.4259
- R^2: 0.3777

While visual features contribute meaningfully, image-only performance lags behind text due to limited direct cues about service or sentiment.

4.4 Fusion Model Performance

The final gated fusion model combines embeddings from four modalities:

- MLP (metadata): 256-dim
- BERT (text): 768-dim
- ResNet-50 (image): 2048-dim
- ZIP-code GNN: 32-dim

These embeddings are passed through learnable scalar gates and concatenated into a 3104-dimensional fused vector. The resulting model achieves the following final test performance:

Table 6: Final performance of the gated fusion model on the Yelp Open Dataset

Metric	**Fusion Model (Gated)**
RMSE	0.2654
MAE	0.2042
R^2	0.8940
Corr (Pearson)	0.9456

The gated fusion model outperforms all unimodal baselines, highlighting the benefits of integrating structured, textual, visual, and geographic signals. The gating mechanism enhances robustness by down-weighting noisy or missing inputs, leading to stronger generalization across diverse businesses.

4.5 Ablation Study

To evaluate the contribution of each input source, ablation experiments were conducted by training the model with only one modality at a time, as well as with their full combination.

Table 7: Performance matrix across modality ablations

Model	**Modalities Used**	**RMSE**	**MAE**	**R^2 Score**	**Corr**
MLP (Ablation)	Metadata	0.6046	0.4708	0.4551	—
BERT (Ablation)	Text	0.5140	0.3707	0.5012	—
CNN (Ablation)	Image	0.5551	0.4259	0.3777	—
Gated Fusion (Final)	Text + Image + Metadata + ZIP-GNN	0.2654	0.2042	0.8940	0.9456

MLP (Metadata) contributed moderately, offering factual attributes like price tier and location. However, such static features lack nuance and personalization, limiting standalone performance.

BERT (Text) performed best as a single modality, reflecting the high signal quality of user reviews. Text data captures sentiment, emotion, and detailed user experience, which are directly aligned with star ratings.

CNN (Image) underperformed individually, suggesting that visual aesthetics alone (e.g., food photos or ambiance) are not strong predictors in isolation, especially given noise and variability in Yelp image data.

Fusion Model significantly outperformed all unimodal configurations. By dynamically weighting modalities through gated fusion, the model was able to emphasize richer signals (e.g., informative reviews) and down-weight less useful or missing data (e.g., sparse metadata or low-quality images).

The Corr = 0.9456 for the fusion model confirms high rank consistency between predicted and actual ratings, a crucial metric for real-world recommendation tasks.

These results validate the design of the multimodal system and highlight the complementary nature of the modalities. While no single source is fully sufficient, their integration provides a richer and more reliable representation for restaurant quality prediction.

4.6 Modality Contributions

To interpret how each modality influences the final prediction, the model computes the L2 norm of gated embeddings during evaluation. This allows quantifying the relative contribution of each input source.

Average L2 Norm (Embeddings):

- BERT (Text): 7.21
- CNN (Image): 4.77
- MLP (Metadata): 3.06
- GNN (ZIP-code): 0.70

The relative contributions of each modality are illustrated in the pie chart below.

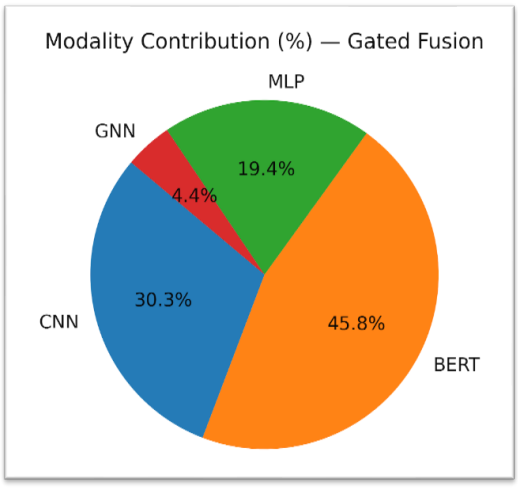

Figure 4: Modality Contribution Percentages in Gated Fusion Model

These results suggest that review text is the most informative (45.8%), followed by visual features (30.3%) and structured metadata (19.4%). Geographic information (4.4%) plays a smaller but still meaningful role. This insight justifies the use of a gated mechanism to let the model dynamically prioritize stronger signals during prediction.

4.7 Visualizations

To assess both model performance and interpretability, the following visualizations are presented.

Training vs. Validation RMSE Curve: This plot illustrates how the model converges over training and helps identify signs of overfitting. The validation curve plateaus after epoch 180, suggesting diminishing returns from further training. This effect was mitigated through hyperparameter tuning, including dropout and weight decay. Final performance was evaluated on a fully held-out test set, ensuring no data leakage and confirming that the model generalizes well beyond the training data.

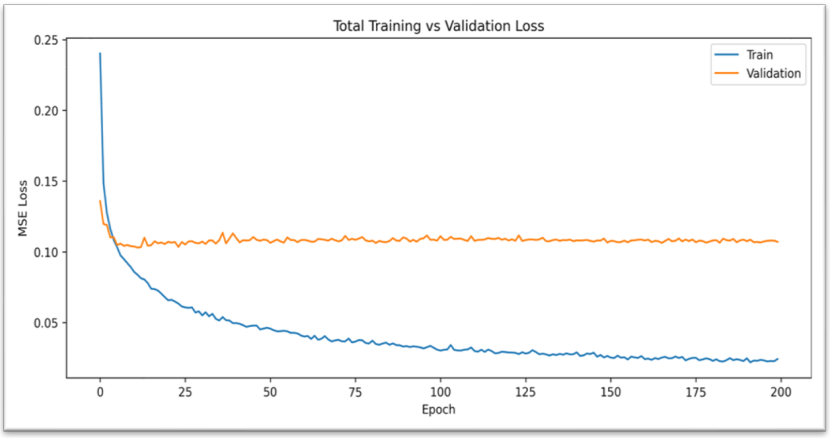

Figure 5: RMSE Learning Curve

Predicted vs. Actual Ratings (Density Plot): A high-density diagonal pattern indicates strong predictive alignment, with Pearson correlation of $\rho = 0.9456$, confirming consistency between predicted and true values.

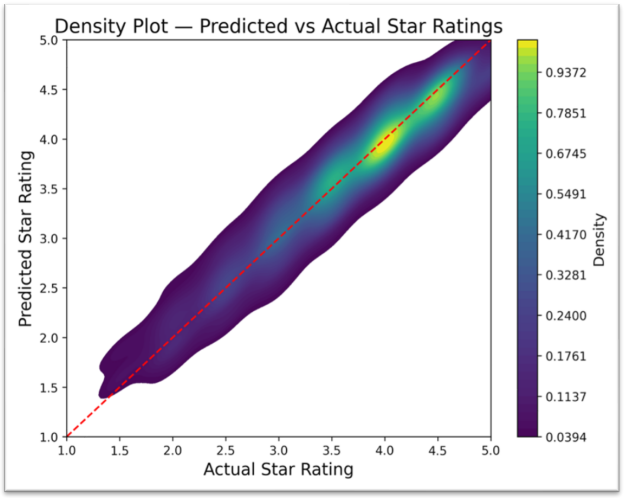

Figure 6: Density Plot – Predicted vs Actual Ratings

Single-Modality vs. Multimodal Performance (R^2 Score Comparison): A bar chart compares standalone R^2 scores across individual modalities—structured metadata (MLP), review text (BERT), and images (CNN)—against the final Gated Fusion model. This visualization highlights the added value of cross-modal integration, with fusion achieving $R^2 = 0.8940$, significantly outperforming any single input type.

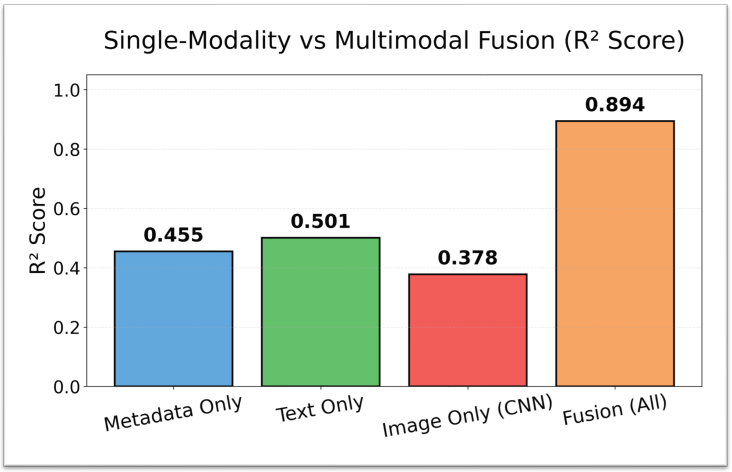

Figure 7: Comparison of R^2 Scores: Single-Modality vs Multimodal Fusion

The visualizations complement quantitative results by illustrating how each modality contributes to overall model performance. The training curves show stable convergence, and the predicted vs. actual rating plots indicate high correlation and predictive reliability. Together, these figures confirm that integrating structured, textual, visual, and spatial signals through gated fusion leads to a more robust and interpretable model.

5 Discussion

Our gated multimodal fusion model achieves strong predictive performance, confirming the effectiveness of combining structured metadata, user reviews, images, and geographic context. Each modality contributes uniquely:

- Review text provides rich sentiment and context, showing the highest standalone predictive power (R^2 = 0.5012).
- Visual features capture ambiance and presentation, performing moderately on their own but boosting performance in fusion.
- Structured metadata adds factual business attributes like price tier and category.
- ZIP-code-based GNN embeddings introduce neighborhood-level context, useful in differentiating businesses in similar categories but different areas.

Together, these inputs enable the model to form a holistic and nuanced understanding of restaurant quality that is not possible with any single modality.

5.1 Why Fusion Works

Multimodal fusion leverages the complementarity of signals. When one modality is weak (e.g., missing or limited images), others can compensate. For instance, a business with sparse metadata but rich reviews can still be predicted accurately. The gated fusion mechanism further enhances this by dynamically weighting each modality based on informativeness, improving robustness and generalization.

5.2 Lessons from Ablation

Ablation experiments confirm the value of each input. While text outperforms other modalities individually, adding metadata and images significantly boosts accuracy. The full gated model, incorporating all four modalities, achieved the best results (R^2 = 0.8940), demonstrating strong cross-modal synergy.

5.3 Model Robustness and Overfitting Control

To address overfitting (as initially observed in earlier experiments), Optuna-based hyperparameter tuning and dropout regularization ($\sim$0.59) were used. Validation performance stabilized after $\sim$180 epochs, and final test metrics matched validation scores, confirming generalization on unseen businesses.

5.4 Limitations and Future Work

Despite strong results, several limitations remain:

- Label imbalance: The dataset skews toward 4–5 star ratings, which may reduce sensitivity to low-quality predictions.
- Missing data: Some businesses lack of sufficient reviews or images, requiring more advanced imputation or augmentation.
- Geographic modeling: The ZIP-code GNN captures only coarse spatial trends; finer-grained or dynamic neighborhood graphs could improve results.
- Lack of temporal modeling: The model uses a static snapshot; adding time-aware features could help capture changing business quality.

Future work could address these by incorporating temporal modeling, synthetic data augmentation, and higher-resolution geographic graphs, as well as applying the system to other domains beyond restaurants.

Summary

Multimodal deep learning, especially with gated fusion, proves highly effective for predicting restaurant quality. By dynamically integrating multiple perspectives — textual, visual, structural, and spatial — the model captures complex real-world patterns with both accuracy and robustness. Continued work will push toward better interpretability, fairness, and deployment-readiness.

References

[1] Z. Liu. Yelp Review Rating Prediction: Machine Learning and Deep Learning Models. *arXiv preprint arXiv:2012.06690*, 2020. https://arxiv.org/abs/2012.06690

[2] Y. Zhao, Y. Shen, Y. Lin, X. Chen, and L. Zhu. Multimodal Point-of-Interest Recommendation. *arXiv preprint arXiv:2410.03265*, 2024. https://arxiv.org/abs/2410.03265

[3] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. https://doi.org/10.1109/TPAMI.2018.2798607

[4] Snax07. Yelp Dataset 2024. Kaggle, 2024. https://www.kaggle.com/datasets/snax07/yelp-dataset-2024

[5] Yelp. Yelp Open Dataset, 2024. https://business.yelp.com/data/resources/open-dataset

[6] Yelp. Yelp Fusion API — Access Business, Reviews, and Ratings Programmatically, 2024. https://business.yelp.com/data/

[7] R. Pathan. Multimodal Yelp Rating Prediction — Project Repository (CSCI 566), 2024. https://github.com/onlypathan/dl-framework-for-multimodal-prediction