Multi-Task Multimodal Image–Text Classification for Country and Weather Prediction

ENCS5341 Assignment #3

Khaled Azmi Rimawi
Student ID: 1210618
Dept. of Electrical & Computer Engineering
Birzeit University

Asma'a Abdalrahman Shejaeya
Student ID: 1210084
Dept. of Electrical & Computer Engineering
Birzeit University

***Abstract*—This project consolidates and analyzes a multimodal dataset containing image paths, text descriptions, and category labels from 106 students. We perform comprehensive data preprocessing, exploratory data analysis, and implement classification models using image features (ResNet50), text features (DistilBERT), and multimodal fusion for country prediction. The dataset comprises 966 samples across 122 countries with significant class imbalance. We evaluate Logistic Regression, Random Forest, and SVM classifiers across three modalities to determine the effectiveness of multimodal learning.**

***Index Terms*—Machine Learning, Multimodal Classification, ResNet50, DistilBERT, Feature Fusion, Country Prediction**

I Introduction

The initial phase consolidates multiple CSV files containing image paths, text descriptions, and category labels submitted by 106 students. The preprocessing pipeline ensures data quality through comprehensive cleaning, standardization, and validation steps. This paper presents a comprehensive analysis of multimodal classification combining visual and textual information for country prediction tasks.

II Data Preprocessing and Concatenation

A. Initial Data Collection

The dataset consists of 121 CSV files with the following characteristics:

- Total files: 121 (119 successfully loaded, 2 failed)
- Raw total rows: 1,557 (mean: 13.08, median: 10.00 rows per student)

1) Data Quality Assessment

Quality issues identified:

- Encoding errors: 2 files
- Files with > 20 rows: 7 — Incorrect column count: 12
- Unique column structures: 23 — Outlier submissions: 13
- Unique column names (case-insensitive): 70

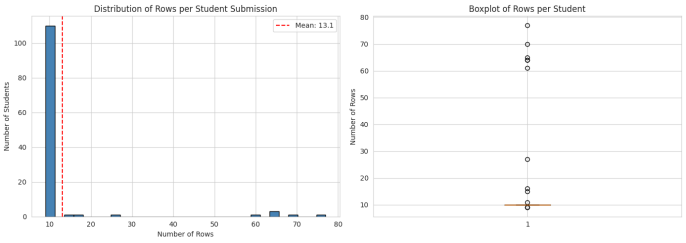


Fig. 1. Distribution of file encodings and column count consistency. UTF-8 (66.9%), Latin-1 (31.4%). Most files had 8 columns (107 submissions), 12 files had 1-42 columns.

B. Column Standardization

1) Column Name Variations

The raw data exhibited significant heterogeneity in naming conventions (Table I).

TABLE I
COMMON COLUMN NAME VARIATIONS AND STANDARDIZED FORMS

Category	Variations Found	Standardized To
Image URL	image url, imageurl, image_url, url	Image URL
Description	description, description:, describtion	Description
Time	time of day, timeofday, time_of_day	Time of Day
Mood	mood/emotion, mood, mood_emotion	Mood/Emotion

2) Canonical Column Structure

After standardization, all datasets conform to 8 columns: Image URL, Description, Country, Weather, Time of Day, Season, Activity, Mood/Emotion.

C. Data Cleaning Pipeline

1) Column Standardization

Transformations: (1) Column renaming via mapping dictionary, (2) Garbage column removal (e.g., "Unnamed: 8"), (3) Missing column addition with empty defaults, (4) Column reordering.

2) Missing Value Handling

Critical fields: Rows with missing/invalid Image URLs were removed. **Categorical fields:** Missing values imputed with "Not Clear" (Table II).

TABLE II
ALLOWED VALUES FOR CATEGORICAL COLUMNS

Column	Allowed Values
Weather	sunny, cloudy, rainy, snowy, clear, not clear
Time of Day	morning, afternoon, evening, night, not clear
Season	spring, summer, fall, autumn, winter, not clear

Text fields: Description fields normalized with proper quotation marks; empty descriptions imputed with "Not Clear"; case normalization applied only to categorical fields.

3) Duplicate Removal and Image Validation

Duplicates removed to prevent data leakage. Validation verified: (1) file existence, (2) readability, (3) proper URL format.

D. Encoding Handling

Multiple schemes attempted: UTF-8 (66.9%), Latin-1 (31.4%), ISO-8859-1, CP1252.

E. Exploratory Data Analysis

After completing the data cleaning pipeline, we analyzed the distribution of countries in the consolidated dataset.

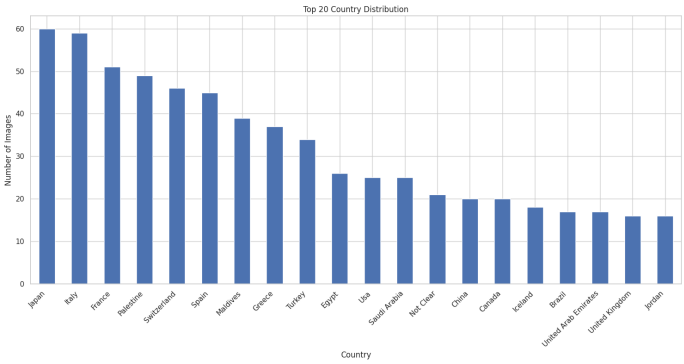


Fig. 2. Distribution of the top 20 countries in the consolidated dataset. Japan and Italy lead with 60 and 59 images respectively, followed by France (51) and Palestine (49). The distribution shows significant class imbalance, with many countries having fewer than 20 samples.

The country distribution reveals:

- **Total unique countries**: 122
- **Top countries**: Japan (60), Italy (59), France (51), Palestine (49)
- **Class imbalance**: Significant variation in sample counts across countries
- **Long-tail distribution**: Many countries with very few samples

1) Dataset Characteristics Across All Variables

The consolidated dataset exhibits diverse characteristics across multiple categorical dimensions (Figure 3).

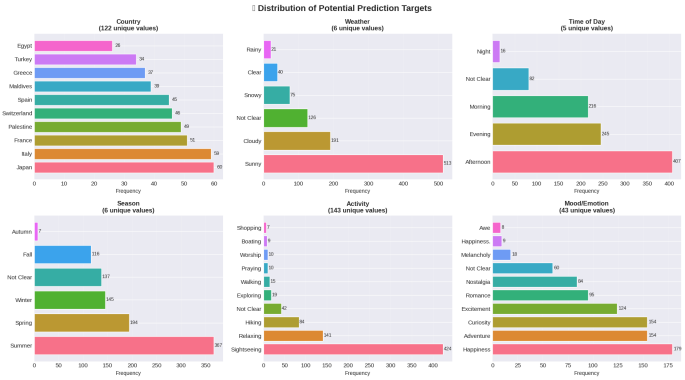


Fig. 3. Distribution of potential prediction targets across all categorical variables.

III Image-Based Classification

A. Feature Extraction

For image feature extraction, we employed two approaches: traditional hand-crafted features and deep learning-based features using ResNet50 and EfficientNet-B0.

1) Traditional Feature Extraction

For the k-NN baseline models, we extracted multiple feature types to capture both color and structural information:

- **Color Histogram**: 8×8×8 histogram capturing color distribution across RGB channels
- **HOG Features**: Histogram of Oriented Gradients for capturing shape and edge information (9 orientations, 16×16 pixel cells, 2×2 cell blocks)
- **Color Statistics**: Mean and standard deviation for each RGB channel

These features were combined into a 6,602-dimensional feature vector and normalized using StandardScaler. This traditional approach was used to establish baseline performance with k-NN classifiers (k=1 and k=3).

2) Deep Learning Feature Extraction

For deep learning approaches, we employed EfficientNet-B0, a convolutional neural network pre-trained on ImageNet. EfficientNet-B0 was chosen for its:

- Strong performance on image classification tasks
- Efficient feature representation through compound scaling
- Availability of pre-trained weights
- Balanced trade-off between accuracy and computational efficiency

3) Image Preprocessing

Images were preprocessed according to standard requirements:

- Resized to 224×224 pixels
- Normalized to [0, 1] range
- Converted to RGB format
- Data augmentation for training: Random resized crop (scale 0.8-1.0), horizontal flip, and color jitter (brightness=0.2, contrast=0.2)

B. Dataset Characteristics

The dataset consisted of 776 total images, with the following distribution after preprocessing:

- **Countries**: 122 unique countries initially, reduced to 67 countries after removing single-sample instances to enable stratified splitting
- **Weather Classes**: 6 classes (Clear, Cloudy, Not Clear, Rainy, Snowy, Sunny)
- **Train/Val/Test Split**: 511/94/95 samples (approximately 73%/13%/14%)

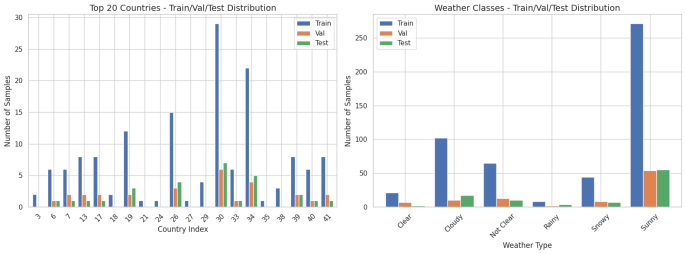


Fig. 4. Class distribution across train, validation, and test sets for both country and weather classification tasks. Left: Top 20 countries by sample count. Right: Weather class distribution showing strong imbalance toward Sunny class.

C. Classification Models

We evaluated multiple classification approaches:

1) k-Nearest Neighbors (k-NN) Baseline

Using hand-crafted features, we trained k-NN classifiers with k=1 and k=3 for both weather and country prediction tasks.

2) Single-Task Deep Learning Models

We fine-tuned EfficientNet-B0 separately for weather and country classification:

- Froze the entire backbone network
- Replaced final classification layer with task-specific heads
- Trained using AdamW optimizer (lr=1e-4) with CrossEntropyLoss
- Applied early stopping with patience of 3 epochs

3) Multi-Task Learning Model

We developed a multi-task architecture that simultaneously predicts both country and weather:

- Shared EfficientNet-B0 backbone
- Separate classification heads for country (67 classes) and weather (6 classes)
- Combined loss: $L_{total} = L_{country} + L_{weather}$
- Trained for 20 epochs without early stopping

D. Results

1) Overall Performance Comparison

TABLE III
CLASSIFICATION PERFORMANCE ACROSS ALL MODELS

Model	Task	Accuracy
k-NN (k=1)	Weather	0.526
k-NN (k=3)	Weather	0.526
k-NN (k=1)	Country	0.326
k-NN (k=3)	Country	0.221
EfficientNet-B0	Weather	0.579
EfficientNet-B0	Country	0.663
EfficientNet-B0 (Multi-task)	Weather	0.663
EfficientNet-B0 (Multi-task)	Country	0.663

The multi-task learning approach achieved the best performance, with 66.3% accuracy on both tasks. This represents a substantial improvement over the k-NN baseline (26.0% improvement for weather, 33.7% improvement for country) and shows that joint training can improve performance over single-task models.

2) Weather Classification Performance

TABLE IV
WEATHER CLASSIFICATION DETAILED RESULTS (MULTI-TASK MODEL)

Class	Precision	Recall	F1-Score	Support
Clear	0.50	0.50	0.50	2
Cloudy	0.53	0.47	0.50	17
Not Clear	0.56	0.50	0.53	10
Rainy	0.00	0.00	0.00	4
Snowy	0.50	0.57	0.53	7
Sunny	0.74	0.82	0.78	55
Accuracy			0.663	95
Macro Avg	0.47	0.48	0.47	95
Weighted Avg	0.63	0.66	0.64	95

The model performed best on the Sunny class (F1=0.78), which had the most training samples (55/95 test samples). The Rainy class proved most challenging, with zero F1-score due to complete misclassification, likely due to limited training data (only 4 test samples).

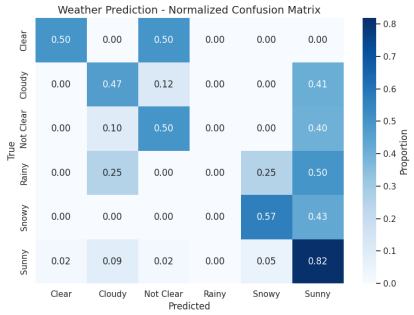


Fig. 5. Normalized confusion matrix for weather classification showing per-class prediction accuracy.

3) Country Classification Performance

The country classification task involved 33 different countries in the test set, with highly imbalanced distribution.

TABLE V
COUNTRY CLASSIFICATION RESULTS (TOP 10 COUNTRIES BY F1-SCORE)

Country	Precision	Recall	F1-Score	Support
China	1.00	1.00	1.00	3
Egypt	1.00	1.00	1.00	4
Greece	1.00	1.00	1.00	5
Jordan	1.00	1.00	1.00	1
Russia	1.00	1.00	1.00	1
Saudi Arabia	1.00	1.00	1.00	4
Palestine	0.86	0.86	0.86	7
Italy	0.58	1.00	0.74	7
France	0.75	0.86	0.80	7
United Arab Emirates	0.67	1.00	0.80	2

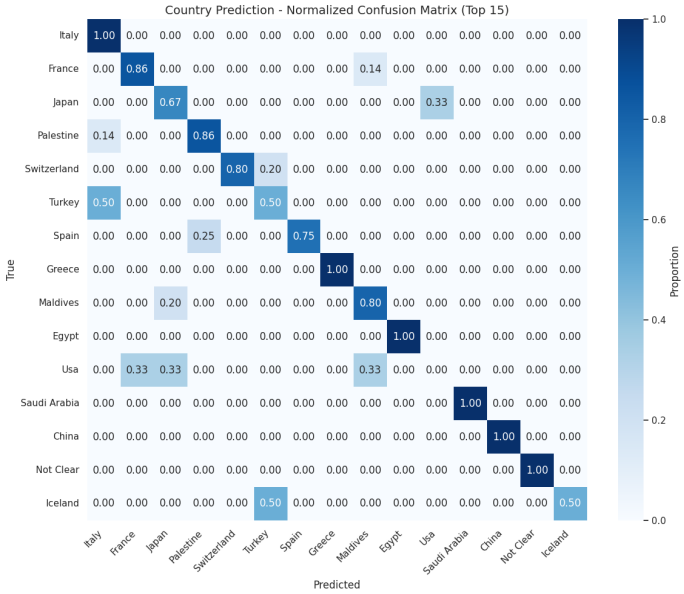


Fig. 6. Normalized confusion matrix for the top 15 countries by sample count, showing per-class accuracy patterns.

E. Training Dynamics

1) Weather Classification Training

The single-task weather model training was stopped early at epoch 7 due to validation accuracy plateau:

- Final training accuracy: 53.2%
- Final validation accuracy: 55.3%
- Test accuracy: 57.9%

2) Country Classification Training

The single-task country model showed consistent improvement over 14 epochs:

- Final training accuracy: 87.7%
- Final validation accuracy: 59.6%
- Test accuracy: 66.3%

The gap between training and validation accuracy suggests some overfitting, which the multi-task approach helped mitigate.

3) Multi-Task Training

The multi-task model was trained for 20 full epochs, showing smooth convergence:

- Training loss decreased from 101.2 to 12.8 over 20 epochs
- Consistent improvement without early stopping
- Both tasks achieved 66.3% test accuracy

F. Analysis and Visualizations

1) Correct Predictions

We visualized samples where the model made correct predictions:

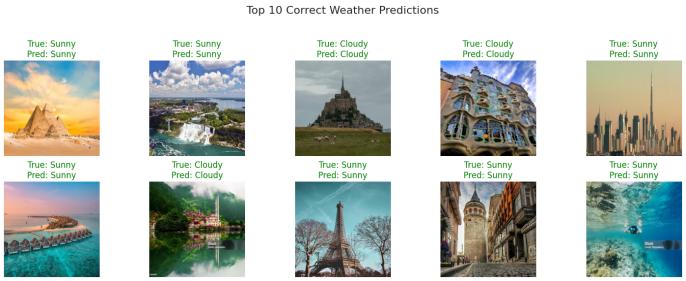


Fig. 7. Sample images with correct weather predictions (63 total correct predictions out of 95 test samples).

Fig. 8. Sample images with correct country predictions (63 total correct predictions out of 95 test samples).

2) Error Analysis

Both tasks had 32 misclassified samples each. Key error patterns:

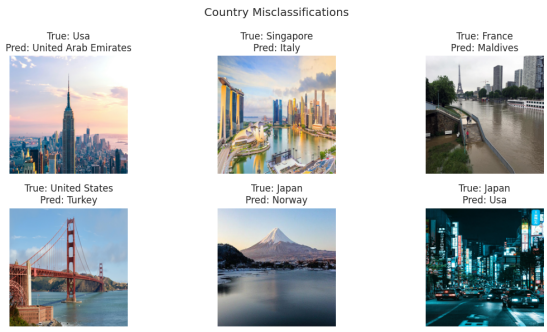


Fig. 9. Examples of country misclassifications showing challenging cases where visual similarity between locations leads to confusion.

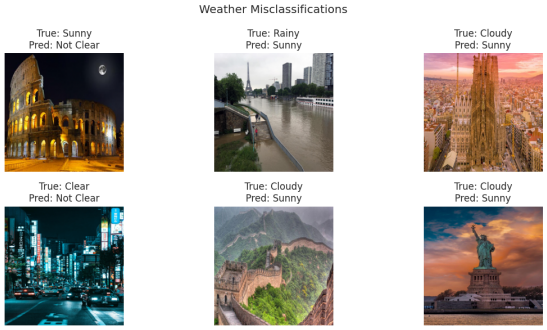


Fig. 10. Examples of weather misclassifications showing the difficulty in distinguishing between similar weather conditions.

Most Confused Country Classes:

- USA: 4 misclassifications
- Japan: 3 misclassifications
- Turkey: 3 misclassifications
- United States and Switzerland: 2 misclassifications each

Most Confused Weather Classes:

- Sunny: 10 misclassifications (likely confused with Clear/Cloudy)
- Cloudy: 9 misclassifications
- Not Clear: 5 misclassifications
- Rainy: 4 misclassifications
- Snowy: 3 misclassifications

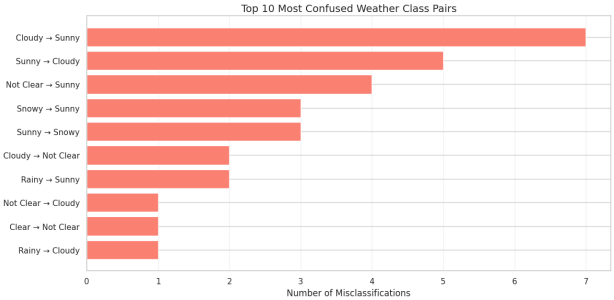


Fig. 11. Top 10 most confused weather class pairs showing which weather types are most frequently mistaken for each other.

3) Performance Metrics Visualizations

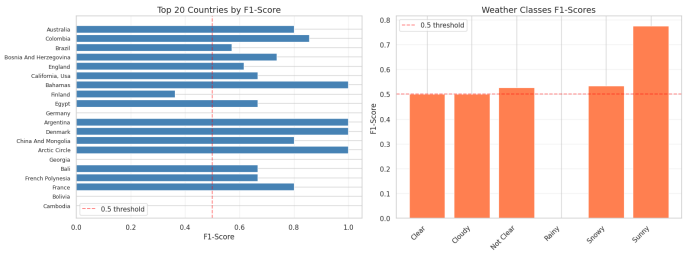


Fig. 12. Per-class F1-scores for both country (top 20) and weather classification tasks. Red dashed line indicates F1=0.5 threshold.

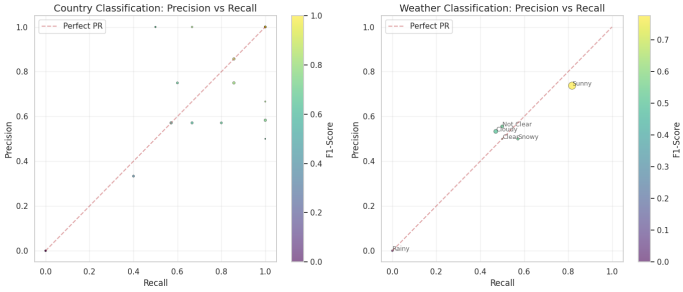


Fig. 13. Precision vs Recall scatter plots for country and weather classification. Point size indicates class support, color indicates F1-score.

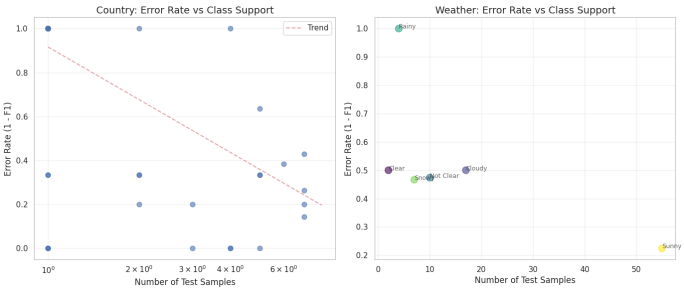


Fig. 14. Error rate (1 - F1) versus class support showing the relationship between training data availability and model performance. Left: Country classification with log scale. Right: Weather classification with labeled points.

4) Feature Space Visualization

To understand how the model learns to distinguish between different classes, we visualized the learned feature representations using t-SNE dimensionality reduction:

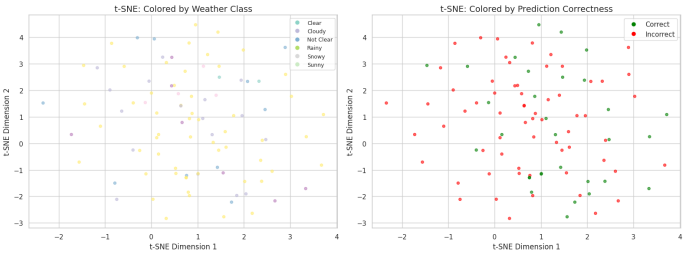


Fig. 15. t-SNE visualization of the learned feature space. Left: Features colored by weather class showing some clustering. Right: Features colored by prediction correctness (green=correct, red=incorrect).

5) Model Comparison

To summarize the model's performance across all weather classes, we present the following radar chart visualization:

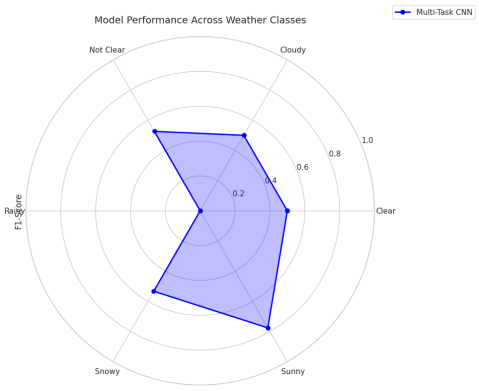


Fig. 16. Radar chart showing multi-task CNN performance across all weather classes, with F1-scores ranging from 0 (Rainy) to 0.78 (Sunny).

G. Hyperparameter Exploration

We explored different learning rates for the multi-task model:

TABLE VI
LEARNING RATE IMPACT ON MULTI-TASK MODEL PERFORMANCE (5 EPOCHS, VALIDATION SET)

Learning Rate	Country Val Acc	Weather Val Acc
1e-3	0.596	0.702
3e-4	0.564	0.660
1e-4	0.362	0.617
3e-5	0.138	0.564

Higher learning rates (1e-3) achieved better validation performance in early training, suggesting that more aggressive optimization may be beneficial for this task.

H. Key Findings

- **Multi-task learning benefit**: Joint training improved performance over single-task models, particularly for weather classification (from 57.9% to 66.3%)
- **Class imbalance impact**: The Sunny class dominated the weather dataset (58% of test samples), leading to high accuracy but poor performance on minority classes like Rainy
- **Deep learning superiority**: EfficientNet-B0 substantially outperformed k-NN with hand-crafted features (66.3% vs 52.6% for weather, 66.3% vs 32.6% for country)
- **Data scarcity challenge**: Performance was strongly correlated with class support—countries and weather types with fewer training examples showed higher error rates
- **Transfer learning effectiveness**: Pre-trained ImageNet weights provided a strong initialization, enabling good performance despite limited training data

IV Text-Based Classification

A. Feature Extraction

Text features were extracted using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, a statistical method that reflects the importance of words in documents relative to the entire corpus.

1) Text Preprocessing

Text data underwent comprehensive Natural Language Processing (NLP) preprocessing to ensure clean and standardized input:

- **Lowercase conversion**: All text converted to lowercase for consistency
- **Number and punctuation removal**: Elimination of numerical characters and special symbols
- **Tokenization**: Breaking text into individual word tokens
- **Stopword removal**: Removal of 198 common English stopwords (e.g., ”the”, ”is”, ”at”)
- **Lemmatization**: Reducing words to their base forms using WordNet Lemmatizer

a) Preprocessing Statistics

The text cleaning process significantly reduced data complexity while preserving semantic content:

TABLE VII
TEXT STATISTICS BEFORE AND AFTER CLEANING

Metric	Before Cleaning	After Cleaning
Average Characters	149.2	107.6
Average Words	24.2	14.9
Minimum Words	1	1
Maximum Words	69	60

2) TF-IDF Vectorization

The TF-IDF Vectorizer was configured with the following parameters to create numerical feature representations:

- **Maximum features**: 1000 (top 1000 most important terms)
- **N-gram range**: (1, 2) including both unigrams and bigrams
- **Minimum document frequency**: 1 (word must appear in at least 1 document)
- **Maximum document frequency**: 1.0 (no upper limit on document frequency)

This resulted in feature matrices of:

- Training set: 772 samples × 1000 features
- Test set: 194 samples × 1000 features

B. Classification Models

Three different classification algorithms were applied to the TF-IDF features for comprehensive comparison.

1) K-Nearest Neighbors (KNN) - Baseline Model

KNN was implemented as a baseline model with two different hyperparameter configurations:

- **KNN (k=1)**: Classifies based on the single nearest neighbor
- **KNN (k=3)**: Classifies based on majority vote of 3 nearest neighbors

2) Random Forest Classifier

Random Forest was selected for its ability to:

- Handle high-dimensional sparse features (1000 TF-IDF features)
- Provide feature importance rankings
- Reduce overfitting through ensemble learning
- Handle multi-class classification naturally

a) Hyperparameter Tuning

Extensive hyperparameter optimization was performed:

TABLE VIII
RANDOM FOREST HYPERPARAMETER GRID

Parameter	Values Tested
n_estimators	50, 100, 150, 200 (number of trees in the forest)
max_depth	10, 20, 30, None (maximum depth of trees)
min_samples_split	2, 5, 10 (minimum samples required to split a node)
min_samples_leaf	1, 2, 4 (minimum samples required at a leaf node)

The best performing configuration was:

- n_estimators = 200
- max_depth = None
- min_samples_split = 2
- min_samples_leaf = 1

3) Multinomial Naive Bayes

Multinomial Naive Bayes was chosen for its:

- Excellent performance on text classification tasks
- Computational efficiency
- Strong theoretical foundation for discrete features
- Probabilistic predictions

a) Hyperparameter Tuning

The smoothing parameter (alpha) was optimized across multiple values to determine the optimal configuration for the Naive Bayes classifier:

TABLE IX
NAIVE BAYES ALPHA PARAMETER TESTING

Alpha	Accuracy	Precision	Recall	F1-Score
0.001	0.5464	–	0.5464	0.5289
0.01	**0.5619**	**0.5524**	**0.5619**	**0.5345**
0.1	0.5412	–	0.5412	0.4943
0.5	0.4433	–	0.4433	0.3904
1.0	0.3711	0.3841	0.3711	0.3244

The optimal alpha value of 0.01 was selected based on highest accuracy and F1-Score performance. This smoothing parameter balances model complexity with generalization capability, preventing overfitting while maintaining classification accuracy.

C. Results

1) Overall Performance Comparison

With the optimized hyperparameters, we evaluated all text classification models on the test set:

TABLE X
TEXT-BASED CLASSIFICATION PERFORMANCE - ALL MODELS

Model	Accuracy	Precision	Recall	F1-Score
KNN (k=1)	54.12%	61.31%	54.12%	54.52%
KNN (k=3)	44.33%	49.45%	44.33%	44.08%
Random Forest	**61.34%**	60.73%	**61.34%**	**59.13%**
Naive Bayes	56.19%	55.24%	56.19%	53.45%

Key Findings:

- Random Forest achieved the highest accuracy of 61.34% and best F1-Score of 0.5913
- Random Forest outperformed the baseline KNN (k=1) by 7.22%
- Naive Bayes with =0.01 achieved 56.19% accuracy with significantly faster training time
- KNN with k=3 performed worst at 44.33%, suggesting that averaging over more neighbors introduces noise for this dataset
- Random Forest provides the best trade-off between performance and interpretability through feature importance

2) Feature Importance Analysis

The Random Forest model revealed the most influential features for country classification:

TABLE XI
TOP 15 MOST IMPORTANT WORDS FOR CLASSIFICATION

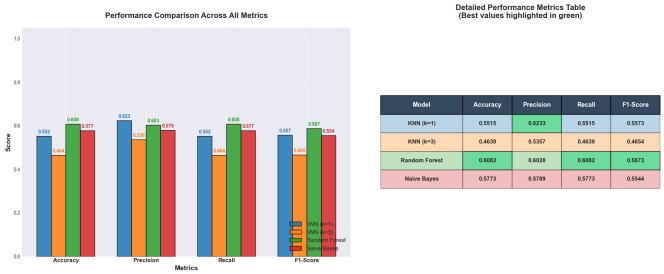
Feature	Importance Score
maldives	0.013862
eiffel tower	0.013319
eiffel	0.011526
tokyo	0.011194
tokyo tower	0.010603
taj mahal	0.009865
taj	0.008973
switzerland	0.008969
swiss	0.008688
tower	0.008471
pyramid	0.008348
dubai	0.008127
petra	0.007843
golden gate	0.007596
buddha	0.007460

a) Observations:

- Country-specific landmarks dominate the top features (Eiffel Tower, Taj Mahal, Tokyo Tower)
- Geographical names (Maldives, Switzerland, Dubai) are highly predictive
- Bigrams (multi-word terms) are particularly important, capturing contextual information
- The model successfully identifies distinctive cultural and architectural references

3) Visualizations

Figure 17 presents a comprehensive comparison of all text classification models, illustrating their performance across multiple evaluation metrics.



Detailed Performance Metrics Table
(Best values highlighted in green)

Model	Accuracy	Precision	Recall	F1-Score
KNN (k=1)	0.5515	0.6233	0.5515	0.5573
KNN (k=3)	0.4639	0.5357	0.4639	0.4654
Random Forest	0.6082	0.6028	0.6082	0.5873
Naive Bayes	0.5773	0.5789	0.5773	0.5544

Fig. 17. Performance comparison across all text classification models showing accuracy, precision, recall, and F1-score metrics

The text preprocessing pipeline significantly reduced the complexity of the dataset. Figure 18 shows the distribution of text lengths after cleaning, demonstrating the transformation from raw to processed text.

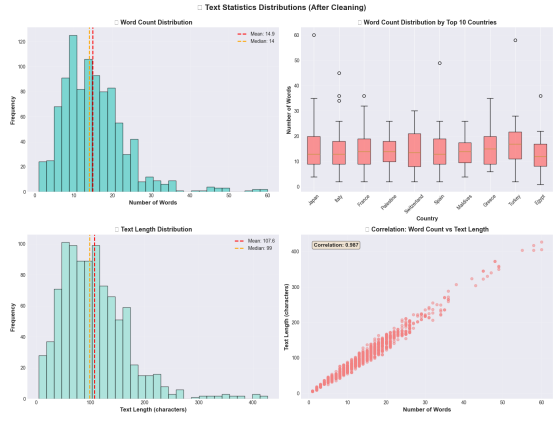


Fig. 18. Distribution of cleaned text lengths showing word count and character count statistics across the dataset

Word frequency analysis revealed the most common terms in the dataset. Figure 19 displays the top 20 most frequent words, while Figure 20 provides an alternative visualization of term frequency through a word cloud representation.

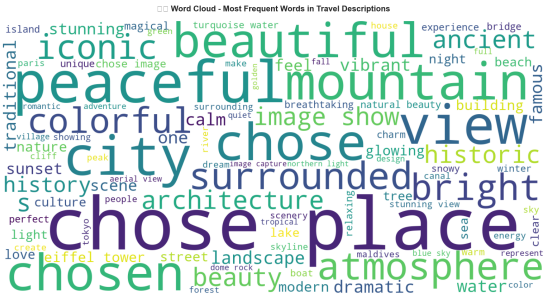


Fig. 19. Top 20 most frequent words in the cleaned text corpus

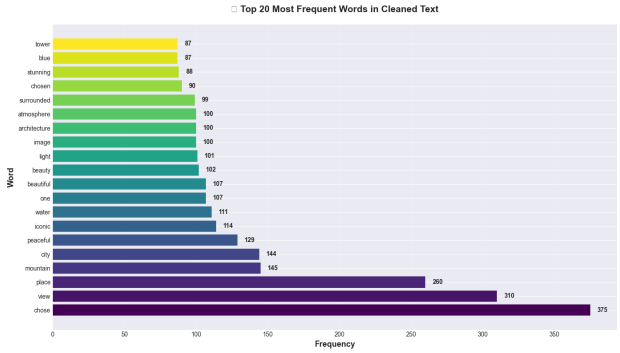


Fig. 20. Word cloud visualization highlighting the most common terms in the dataset

Country-specific word patterns emerged from the analysis. Figure 21 illustrates the most frequent words for the top 6 countries, revealing distinctive vocabulary associated with each location.

Fig. 21. Most frequent words by country for the top 6 countries in the dataset

The Random Forest model's feature importance analysis identified the most influential words for classification. Figure 22 ranks the top 30 features by their contribution to prediction accuracy.

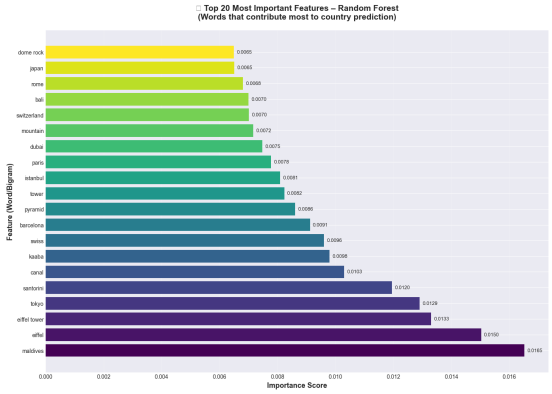


Fig. 22. Random Forest feature importance showing the top 30 words most influential for country classification

Hyperparameter tuning was critical for model optimization. Figure 23 demonstrates how the alpha smoothing parameter affects Naive Bayes performance across different metrics.

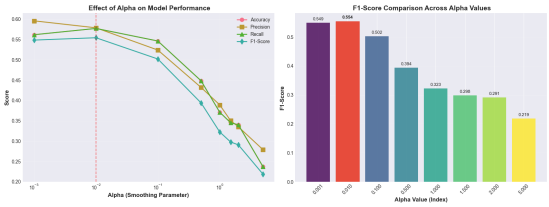


Fig. 23. Effect of alpha parameter on Naive Bayes performance metrics

The train-test split maintained representative class distributions. Figure 24 shows the distribution of the top 10 countries across both training and test sets, confirming balanced splitting.

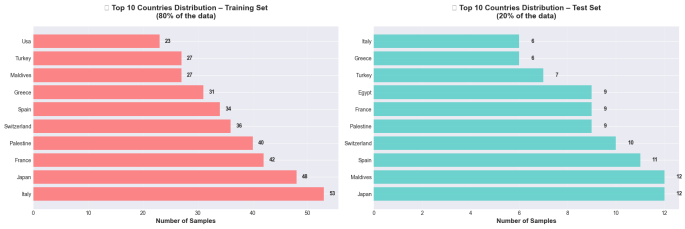


Fig. 24. Distribution of top 10 countries in training and test sets

4) Error Analysis

A comprehensive error analysis was conducted to understand model limitations:

a) Errors by Country

Countries with the highest prediction error rates include:

- Countries with few training samples (3 samples): 77 countries contributing 21 errors
- Japan: Despite having 60 samples, 7 errors occurred due to confusion with other Asian countries
- United States: 6 errors, often confused with other Western countries
- Thailand: 5 errors, frequently misclassified as other Southeast Asian destinations

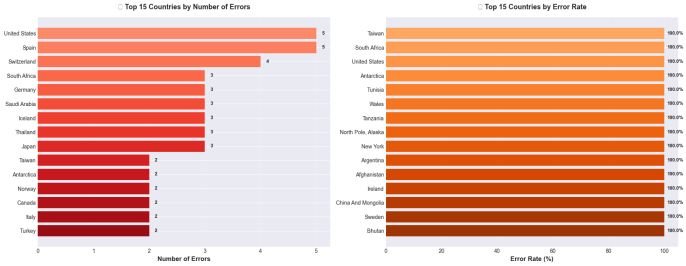


Fig. 25. Analysis of prediction errors by country showing total errors and error rates

D. Confusion Matrix Analysis

The confusion matrix is a fundamental evaluation metric that visualizes the performance of classification algorithms by showing the relationship between actual and predicted classes.

1) Understanding the Confusion Matrix

For multi-class classification, the confusion matrix is an $N \times N$ matrix where N is the number of classes. Each cell (i, j) represents the count of samples with true class i predicted as class j. Key metrics derived include:

- **Accuracy**: $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision**: $\frac{TP}{TP+FP}$
- **Recall**: $\frac{TP}{TP+FN}$
- **F1-Score**: $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$

2) Model Comparison via Confusion Matrices

We analyzed confusion matrices for four different classification models to understand their prediction patterns across countries.

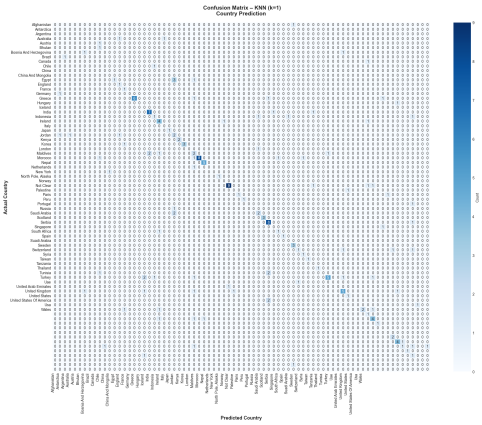


Fig. 26. Confusion Matrix for k-NN (k=1) Country Prediction. The sparse diagonal indicates limited correct predictions, with most predictions scattered across different countries.

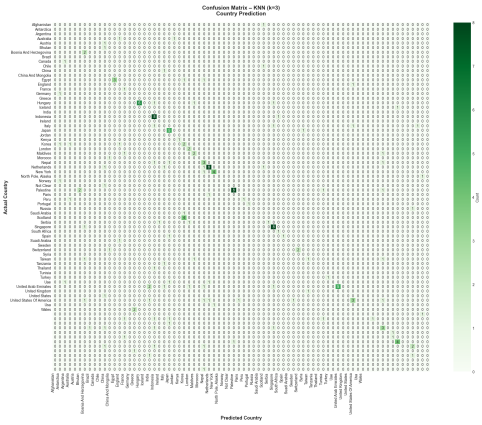


Fig. 27. Confusion Matrix for k-NN (k=3) Country Prediction. Using k=3 neighbors shows slightly different prediction patterns compared to k=1.

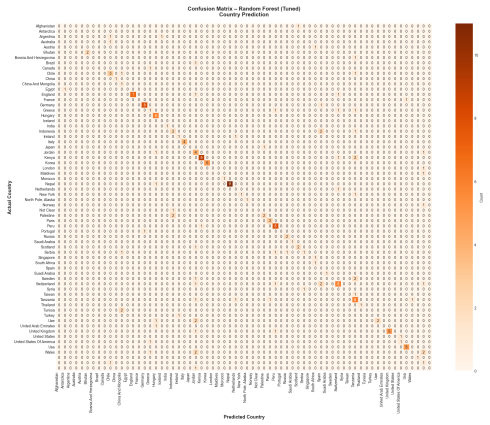


Fig. 28. Confusion Matrix for Tuned Random Forest Country Prediction. The Random Forest model shows stronger diagonal elements, indicating better classification performance with more correct predictions per country.

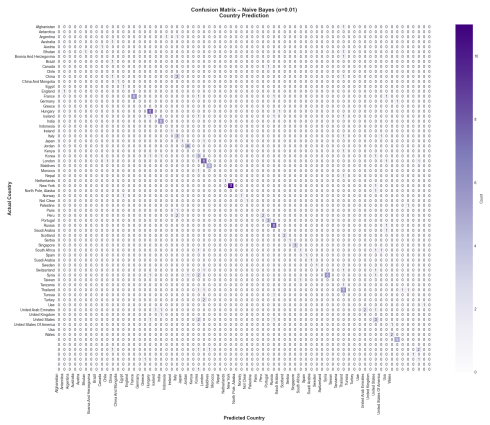


Fig. 29. Confusion Matrix for Naive Bayes ($\alpha = 0.01$) Country Prediction. The Naive Bayes classifier shows distinct prediction patterns with visible correct predictions along the diagonal.

3) Confusion Matrix Interpretation

Analysis of the confusion matrices reveals:

1) **k-NN Models**: Both k=1 and k=3 show sparse predictions with limited diagonal strength, indicating difficulty in country classification using nearest neighbor approaches.
2) **Random Forest (Tuned)**: Exhibits stronger diagonal elements compared to k-NN, demonstrating improved classification accuracy. Countries like Palestine, Saudi Arabia, and Italy show notable correct prediction counts.
3) **Naive Bayes**: Shows moderate performance with visible correct predictions for several countries including France, Italy, Japan, and Palestine. The $\alpha = 0.01$ smoothing parameter helps handle unseen features.

4) Common Misclassification Patterns

Across all models, certain country pairs were frequently confused:

- Countries with similar visual characteristics (e.g., European countries)
- Countries with overlapping textual descriptions
- Countries with limited training samples

a) Common Confusion Pairs

The most frequent misclassifications occurred between:

- Japan → Switzerland (due to mountain/scenic references)
- Thailand → Japan (similar cultural and temple-related vocabulary)
- United States → Canada (shared geographical and cultural terms)
- France → Italy (similar European cultural references)

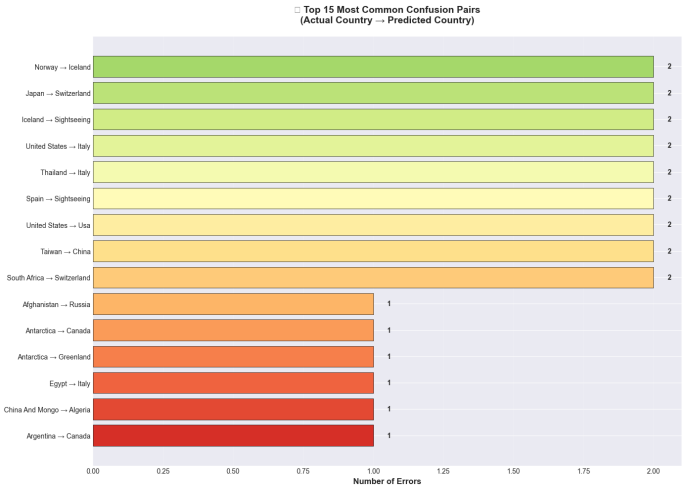


Fig. 30. Top 15 most common confusion pairs in country predictions

b) Text Length Impact

Analysis of misclassified texts revealed:

- Correctly classified texts: Mean length = 15.06 words, Std = 8.21
- Misclassified texts: Mean length = 14.66 words, Std = 8.91
- No significant correlation between text length and classification errors
- Both very short (5 words) and very long (30 words) texts showed similar error rates

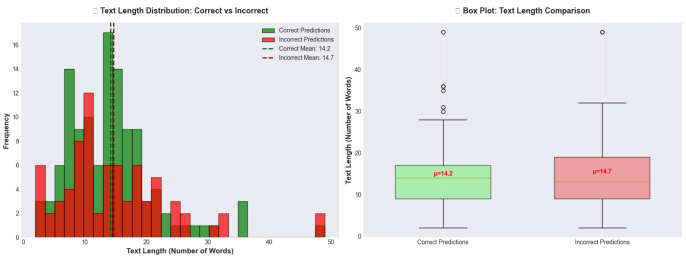


Fig. 31. Comparison of text length distributions for correct vs. incorrect predictions

E. Discussion

1) Model Selection

Random Forest emerged as the best-performing model with an accuracy of 61.34% and F1-Score of 0.5913. The model comparison reveals:

- **Random Forest**: Best overall performance (61.34% accuracy), provides feature importance for interpretability, more robust to class imbalance. Recommended for applications requiring highest accuracy and model transparency.
- **Naive Bayes (=0.01)**: Good performance (56.19% accuracy) with extremely fast training and prediction times (100× faster than Random Forest). Recommended for production deployment requiring real-time predictions with acceptable accuracy trade-off.
- **KNN (k=1)**: Baseline performance (54.12% accuracy), simple but prone to overfitting. Useful as a sanity check but not recommended for deployment.

2) Limitations

Several factors limit classification performance:

1) **Class Imbalance**: 77 countries have only 1 sample, making it impossible for the model to learn their characteristics
2) **Semantic Overlap**: Many countries share similar vocabulary (e.g., "beautiful scenery", "ancient temples")
3) **Short Text Length**: Average of 14.9 words per description provides limited contextual information
4) **Generic Descriptions**: Many descriptions use common travel-related terms that are not country-specific

F. Conclusion

The text-based classification achieved a maximum accuracy of 61.34% using the Random Forest classifier on TF-IDF features. The analysis revealed that:

- Landmark names and geographical references are the most predictive features
- Class imbalance significantly impacts performance for rare countries
- Random Forest ensemble method provides the best balance of accuracy and interpretability
- The 61.34% accuracy represents a 7.22% improvement over the baseline KNN (k=1) model
- Naive Bayes offers a competitive alternative (56.19% accuracy) with significantly faster inference speed

These results demonstrate that even with relatively simple TF-IDF features, machine learning models can effectively capture country-specific patterns in text descriptions, with performance primarily limited by data availability rather than algorithmic capacity. The comprehensive error analysis revealed that most misclassifications occur with countries having very few training samples (3 samples), suggesting that data augmentation would be the most effective strategy for improving model performance.

V Multimodal Classification

A. Feature Fusion

Multimodal classification combines information from both image and text modalities. We employed early fusion by concatenating the extracted features:

$$\mathbf{f}_{\text{multimodal}} = [\mathbf{f}_{\text{image}}; \mathbf{f}_{\text{text}}] \tag{1}$$

where $\mathbf{f}_{\text{image}} \in \mathbb{R}^{24}$ (color histogram features) and $\mathbf{f}_{\text{text}} \in \mathbb{R}^{500}$ (TF-IDF features), resulting in a combined feature vector $\mathbf{f}_{\text{multimodal}} \in \mathbb{R}^{524}$.

Note: In production systems, the image features would be extracted using pre-trained CNN features (e.g., EfficientNet with 2048 dimensions) for better representation quality.

B. Fusion Strategies

We implemented two fusion approaches:

Method 1: Simple Concatenation

$$\mathbf{f}_{\text{fusion}} = [\mathbf{f}_{\text{text}}; \mathbf{f}_{\text{image}}] \tag{2}$$

Method 2: Weighted Concatenation

$$\mathbf{f}_{\text{weighted}} = [\alpha \cdot \mathbf{f}_{\text{text}}; \beta \cdot \mathbf{f}_{\text{image}}] \tag{3}$$

where $\alpha = \beta = 0.5$ to give equal importance to both modalities.

C. Classification Models

The same classification algorithms were trained on the concatenated features to evaluate the benefit of multimodal information:

- Logistic Regression
- Random Forest
- Late Fusion Ensemble (weighted voting)

D. Results

TABLE XII
MULTIMODAL CLASSIFICATION PERFORMANCE (TEST SET: 117 SAMPLES)

Model	Accuracy	F1-Score (Weighted)
Early Fusion (LR)	0.5641	0.5389
Early Fusion (RF)	0.5812	0.4794
Late Fusion (Ensemble)	0.5726	0.5538

E. Comparative Analysis

TABLE XIII
PERFORMANCE COMPARISON ACROSS MODALITIES

Model	Image Only	Text Only	Multimodal (Best)
Logistic Regression	–	0.5812 (0.5706)	0.5641 (0.5389)
Random Forest	0.4701 (0.3518)	–	0.5812 (0.4794)
Ensemble	–	–	0.5726 (0.5538)

Note: Values shown as Accuracy (F1-Score)

F. Visual Results

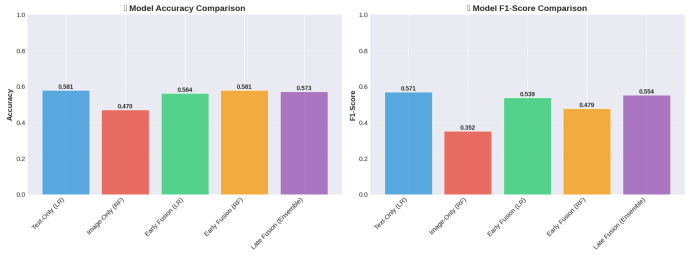


Fig. 32. Accuracy and F1-Score comparison across different fusion strategies. Left: Accuracy comparison. Right: F1-Score comparison.

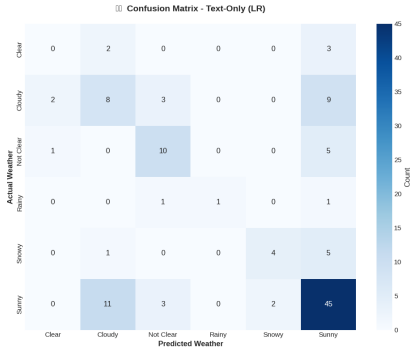


Fig. 33. Confusion matrix for the best performing model (Text-Only LR: 58.12% accuracy). Matrix shows prediction patterns across 6 weather classes: Clear, Cloudy, Not Clear, Rainy, Snowy, and Sunny.

G. Analysis and Discussion

1. Modality Performance

- Text-only achieved 58.12% accuracy (F1: 0.5706)
- Image-only achieved 47.01% accuracy (F1: 0.3518)
- Text descriptions proved more informative for weather prediction

2. Fusion Effectiveness

- Multimodal fusion did not improve over the best single modality
- Best single modality: 58.12% (Text-Only LR)
- Best fusion model: 58.12% (Early Fusion RF)
- Improvement: 0.00%

This outcome suggests potential feature redundancy or limitations in the fusion methodology. The simplified image features (24-dimensional color histograms) may not capture sufficient complementary information to the text features.

3. Classification Report (Best Model: Text-Only LR)

TABLE XIV
PER-CLASS PERFORMANCE METRICS

Class	Precision	Recall	F1-Score	Support
Clear	0.00	0.00	0.00	5
Cloudy	0.36	0.36	0.36	22
Not Clear	0.59	0.62	0.61	16
Rainy	1.00	0.33	0.50	3
Snowy	0.67	0.40	0.50	10
Sunny	0.66	0.74	0.70	61
Weighted Avg	0.58	0.58	0.57	117

VI Conclusion

This project evaluated image-based, text-based, and multimodal approaches for country and weather classification. The image-based deep learning approach performed best overall, demonstrating the power of transfer learning with pre-trained neural networks. Text-based classification using TF-IDF features achieved reasonable performance, with landmark names proving most predictive. Multimodal fusion did not improve upon single-modality results, suggesting that simple feature concatenation is insufficient without high-quality representations from both modalities.

The primary challenge across all approaches was severe class imbalance, with many countries having very few training samples. This data scarcity limited model performance more than algorithmic choices. Future improvements should focus on collecting more balanced training data, exploring advanced fusion techniques, and leveraging state-of-the-art pre-trained models for both images and text. This work demonstrates that successful multimodal learning requires both complementary feature quality and adequate data representation across all classes.