Seeing the Water: Scenery Classification using the Intel Image Classification dataset

Aurora Olaya, Zane Brown, Peter Benzoni

“There are these two young fish swimming along and they happen to meet an older fish swimming the other way, who nods at them and says "Morning, boys. How's the water?" And the two young fish swim on for a bit, and then eventually one of them looks over at the other and goes "What the hell is water?” - David Foster Wallace

Introduction

As humans, we understand the world through context—asking to go for a swim in the city means something different than in the woods, or the ocean. For machines to similarly attune their responses, they must be given, then taught the same context—they are, in our introductory metaphor, the young fish asking what water is. This project explores how machines can begin to see the world through the contextual lens humans take for granted.

This ability has practical implications: powering smarter search tools, enabling environmental study at scale, and enhancing user experiences in applications where imagery meets decision-making, such as real estate or travel platforms.

About the Dataset

For our project, we used the Intel Image Classification dataset which was originally published by Intel to host an Image Classification Challenge. This dataset comprises approximately 25,000 images, each with a resolution of 150x150 pixels, divided into six categories: Buildings, Forest, Glacier, Mountain, Sea, and Street. The data is organized into separate zip files for training, testing, and prediction, with around 14,000 images in the training set, 3,000 in the test set, and 7,000 for prediction. In the training set, each feature has roughly 2,300 examples. In the test set, each feature has 500 examples. The prediction set does not have labels as this is meant to be hidden for the challenge. We did not use the prediction set for our project.

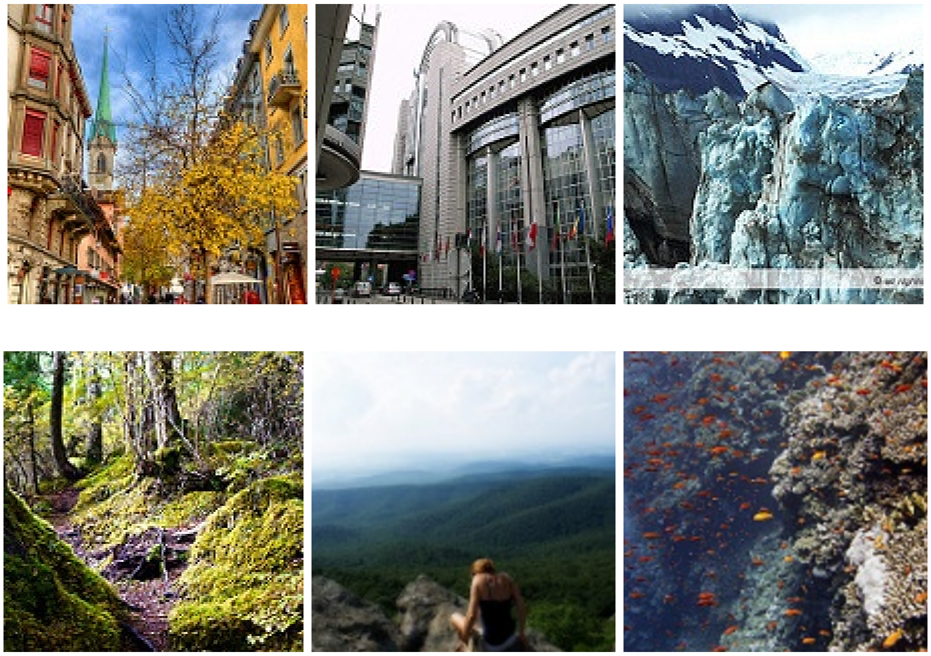

Figure 1: Sample Images

Tools and Techniques

Using these tools and techniques, our machines' previously untrained eyes can see the "water" they swim in:

1. Data Preprocessing: Images are resized to a uniform 150x150 resolution and normalized to ensure consistency in pixel values.
2. Feature Extraction: Dimensionality reduction is performed using Incremental PCA, allowing us to compress high-dimensional data into a manageable format. Additional feature vectors created were HSV and HOG.
3. Traditional Machine Learning Models: Models such as Support Vector Machines (SVM) are trained on flattened image data, with hyperparameters optimized through GridSearchCV.
4. Deep Learning Architectures: A Convolutional Neural Network (CNN) is built from scratch. Transfer learning is implemented using ResNet50, a pre-trained model fine-tuned to the dataset. These approaches leverage hierarchical feature learning to identify patterns in images.
5. Evaluation Metrics: Accuracy, time inference, confusion matrices, ROC curves, and classification reports are used to assess model performance.

Preprocessing and Challenges

The raw dataset, while comprehensive, presented some challenges, ultimately requiring the following steps:

Preprocessing Steps

To prepare the dataset for analysis, several preprocessing steps were undertaken to standardize and optimize the input data. First, all images were resized to a uniform resolution of 150x150 pixels, ensuring consistent dimensions across the dataset. Next, pixel values were normalized by scaling them to a range between 0 and 1, a technique that improves numerical stability during training and benefits gradient-based optimizers commonly used in deep learning. For traditional models such as Support Vector Machines (SVMs), images were flattened into one-dimensional arrays. While effective for certain approaches, this transformation eliminates spatial relationships between pixels—an important consideration for tasks like scene recognition. Finally, a train-test split was applied, with training and test sets loaded separately to prevent data leakage. From the training set, a validation subset was further extracted to support hyperparameter tuning.

Challenges

The project faced several challenges that required thoughtful solutions. While the dataset was relatively balanced overall, some categories exhibited significant visual variability. For example, the “forest” and “sea” categories encompassed a wide range of textures and colors, which increased the likelihood of classification errors compared to more visually distinct categories like “buildings.” Furthermore, overlapping features between categories, such as the structural lines and urban textures shared by “street” and “buildings,” complicated the ability of models to make clear distinctions. The dataset's dimensionality posed additional challenges: flattening images resulted in a feature space of 22,500 dimensions per image, which was unwieldy for traditional machine learning algorithms. Incremental PCA (IPCA) was employed to reduce the feature space while preserving critical variability, though this required careful tuning to avoid discarding essential information. Finally, noise in the dataset—stemming from variations in lighting, perspective, and image quality—added another layer of complexity, particularly for simpler models that struggled to identify distinguishing features amidst these inconsistencies. By addressing these challenges, this project takes a step toward helping machines contextualize the world in ways that parallel human perception.

Exploratory Data Analysis & Feature Extraction

Exploratory Analysis

Contrast and HSV

Analyzing pixel intensity distributions provides valuable insights into the visual characteristics of each category in the dataset. These histograms reveal the tonal range, contrast, and dominant features within images, HSV further adds both literal and figurative color to this analysis. Together, these offer a foundation for selecting appropriate preprocessing and feature extraction techniques. To illustrate these characteristics, below are sample images from each classification alongside their corresponding contrast and HSV histograms.

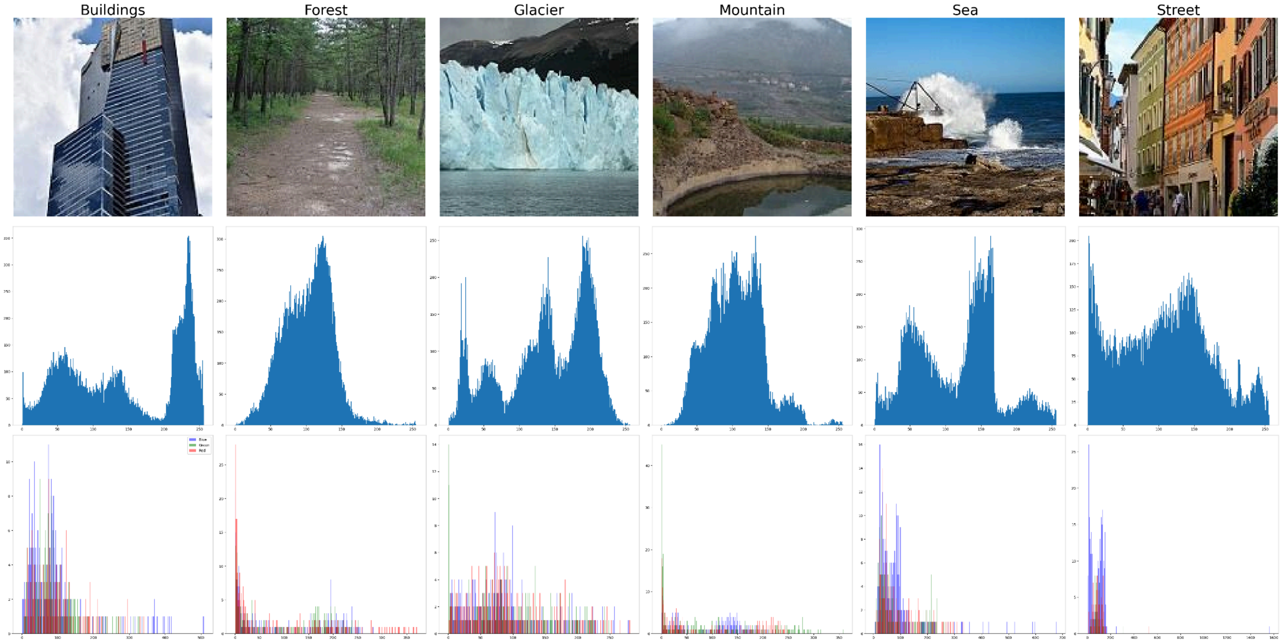

Figure 2: Sample Contrast & HSV Results

The middle row displays contrast histograms, which plot pixel intensities from dark (0) to light (255), revealing how brightness and shadows are distributed throughout each image. The bottom row shows RGB histograms, where overlapping peaks of red, green, and blue channels help identify dominant colors and their variations within each scene. This dual analysis allows us to understand both the structural and chromatic patterns unique to each landscape category.

Looking at the specific characteristics of each category, buildings' histograms have peaks at high pixel intensities, indicating bright regions such as windows or reflective surfaces, with noticeable multimodal peaks implying contrasting elements like dark building structures against bright skies and limited color diversity with low saturation and high brightness, reflecting urban tones and bright skies. Forest histograms are heavily concentrated toward mid-range intensities, reflecting the green and brown tones common in forests, with a lack of extreme peaks suggesting a uniform texture without sharp contrasts, characterized by green hues, mid-to-high saturation, and moderate brightness, reflecting vibrant vegetation and shadowed canopies. In 'Glaciers', peaks are in higher intensity ranges, which may speak to the bright whites and light blues of glaciers, with the narrow distribution indicating a more homogeneous brightness than other categories, showing blue-cyan hues, low saturation, and high brightness, indicative of reflective snow and ice surfaces. Mountains' distributions are highly diverse, with peaks in mid-to-high ranges—reflecting the varying textures of mountains, including rocky surfaces, trees, shadows, and snow, featuring earthy tones with broad hue ranges, moderate saturation, and varied brightness due to rugged terrain and sunlight contrasts. Similar to glaciers, Sea histograms have peaks in higher intensities, corresponding to bright skies or sunlight reflections on the water, though some samples show distributions skewed toward mid-tones, suggesting variability in weather or lighting conditions, featuring blue-green hues, high saturation, and high brightness, capturing the vivid and reflective nature of water bodies. The distributions of Street resemble those of buildings but tend to have more pronounced peaks at mid-intensity levels, representing asphalt roads and shaded areas, with multimodal peaks indicating a mix of bright urban elements (e.g., streetlights or signage) with darker road surfaces, displaying muted tones with narrow hue ranges, low saturation, and a wide brightness range from shadows to illuminated surfaces.

Analysis

Based on these histograms, two pairs of categories share similar intensity distributions. Buildings with Streets and Sea and Glacier. Buildings and Streets share multimodal peaks reflecting mixed urban elements. Sea and Glacier have high-intensity peaks that illuminate the bright, reflective surfaces common in both. These shared features may introduce classification challenges, as the overlap in tonal characteristics can confuse models.

Hue-based features do appear to effectively separate natural categories (e.g., forests and glaciers) from urban ones. Saturation and value are critical for distinguishing vibrant categories like forests from muted ones like streets. However, overlapping features between glacier and sea, as well as buildings and streets, highlight the need for complementary texture and spatial features to enhance classification accuracy.

Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients (HOG) analysis provides valuable structural insights into the distinct characteristics of each category in the dataset. This technique reveals edge patterns, geometric structures, and textural features within images, while effectively filtering out noise that could obscure these key elements. By reducing complex visual data to essential gradient information, HOG creates a robust foundation for efficient classification tasks. Its particular strength in identifying recurring patterns and subtle variations in texture proves especially valuable for distinguishing between visually similar categories, such as the nuanced differences between sea and glacier landscapes, or the complex boundary between glacier and mountainous scenes.

Results

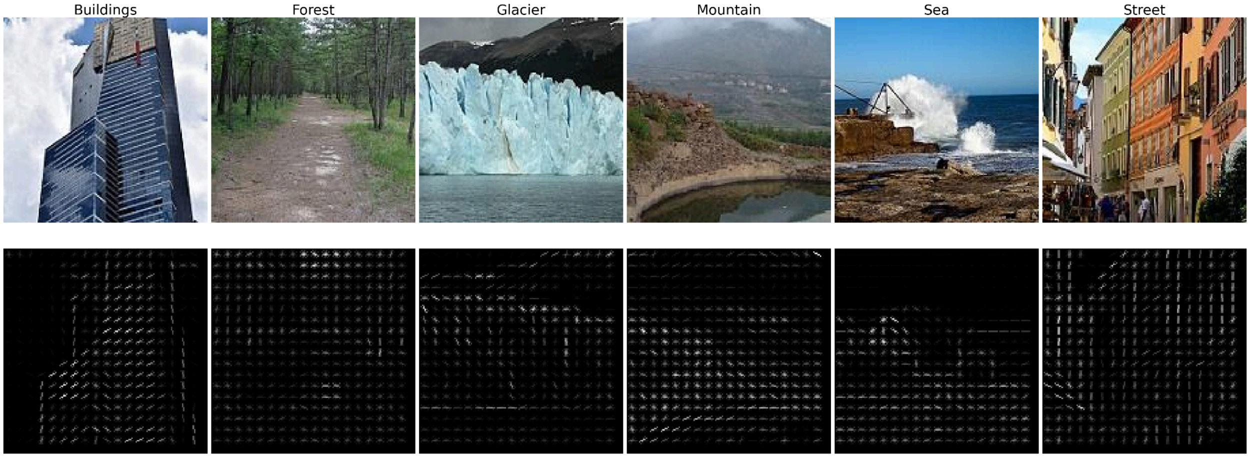

Figure 3: Sample HOG Results

To illustrate these structural characteristics, above are sample images from each classification alongside their corresponding HOG visualizations. In buildings, features clearly outline the edges of structural elements like windows, walls, and rooftops, with strong grid-like patterns reflecting urban architecture's repetitive geometry. Forest images show diffused and irregular edges indicating organic, non-linear textures of trees and foliage, with patterns lacking strong orientation despite some parallel gradients. In glaciers, gradual, smooth edges dominate, capturing ice formation contours with less complex patterns that reflect the uniform texture of snow and ice. Mountain landscapes display sharp, angular edges corresponding to rocky terrains and slopes, with highly variable patterns that reflect diverse textures and natural irregularities. Sea images are characterized by dominant horizontal edge patterns from wave crests and breaks between sea and sky, exhibiting smoother features compared to urban categories. Street scenes present strong linear features outlining roads, sidewalks, and urban elements, sharing buildings' geometric patterns but with greater variation due to additional textures like asphalt or vehicle edges.

Analysis

Both buildings and streets exhibit strong, geometric patterns with well-defined edges, reflecting man-made structures. Streets show slightly more variability due to mixed textures, while buildings retain grid-like consistency. Still, overlapping geometric features may make differentiation challenging, especially for urban environments with mixed architecture and roadways.

By contrast, forests and mountains display more irregular, less repetitive features, consistent with organic form. Both display irregular features, but mountains are marked by sharper, angular edges, whereas forests have diffuse textures.

Glaciers and seas share smoother patterns, with glaciers focusing on contours and seas on horizontal edges. Both categories exhibit smooth and less complex patterns, with differentiation likely relying on horizontal dominance in seas and contour-like gradients in glaciers.

Overall, HOG is effective for capturing structural and geometric features in urban environments. The similarity in structural patterns between categories like streets and buildings or seas and glaciers underscores the need for additional contextual or color-based features to improve classification accuracy, though texture-based methods may complement HOG for differentiating natural categories like forests and mountains.

Principal Components Analysis (PCA)

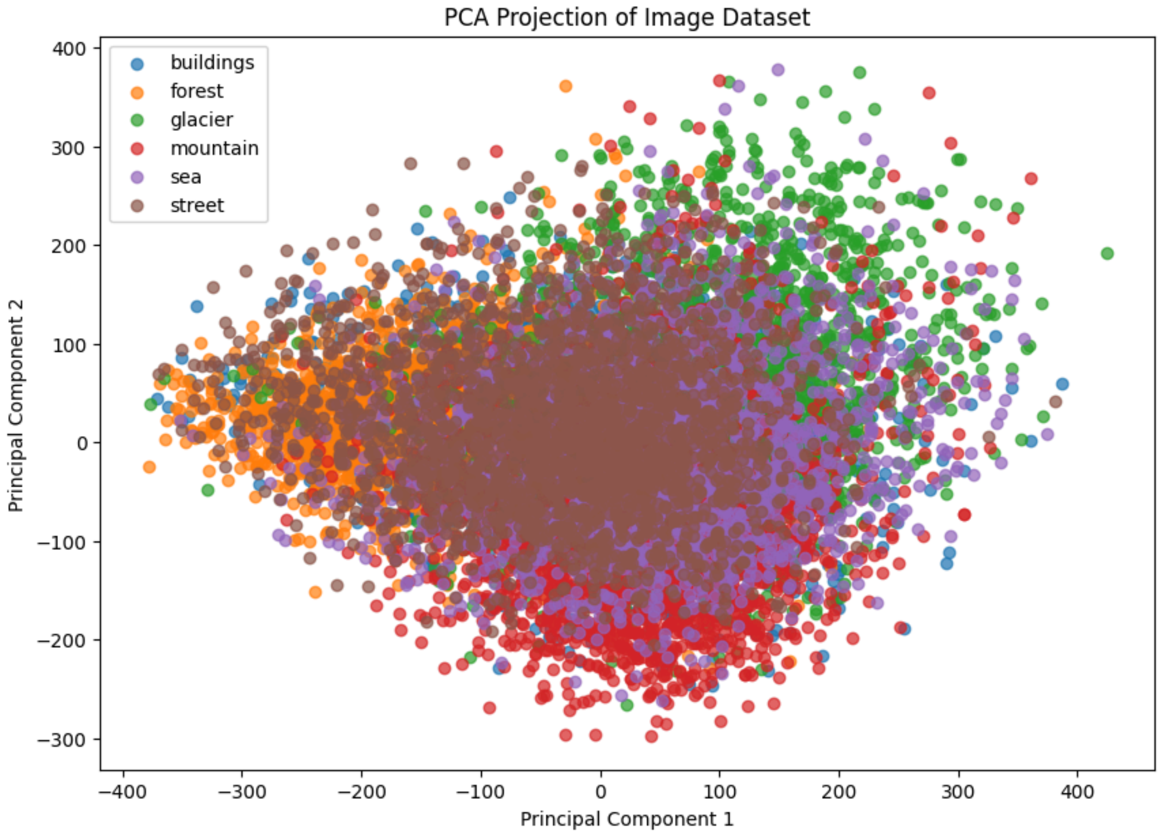

Figure 4: PCA Results

PCA appears to have limited utility for this dataset due to significant overlap between many categories. Urban categories like buildings and streets exhibit substantial overlap, reflecting their shared geometric and structural features, such as edges, lines, and muted tones. Natural categories like forests, mountains, and seas show broader dispersion, indicative of variability in natural textures and patterns. However, glaciers form a tighter cluster, suggesting that their consistent visual features—dominated by brightness and smooth textures—make them easier to separate.

Despite some broad distinctions, overlap persists between natural and urban categories, such as forests and mountains or seas and glaciers. These overlaps reflect shared characteristics like irregular textures, earthy tones, or reflective surfaces, which PCA's linear nature struggles to distinguish effectively. This emphasizes the need for non-linear dimensionality reduction techniques or advanced feature extraction to better separate nuanced features.

While PCA provides an accessible summary of feature variance and relationships between categories, it highlights the challenges of achieving clear separability in this dataset. Techniques like t-SNE and texture-sensitive feature extraction are essential for improving classification performance and tackling the complexities of scene classification.

t-distributed Stochastic Neighbor Embedding (tSNE)

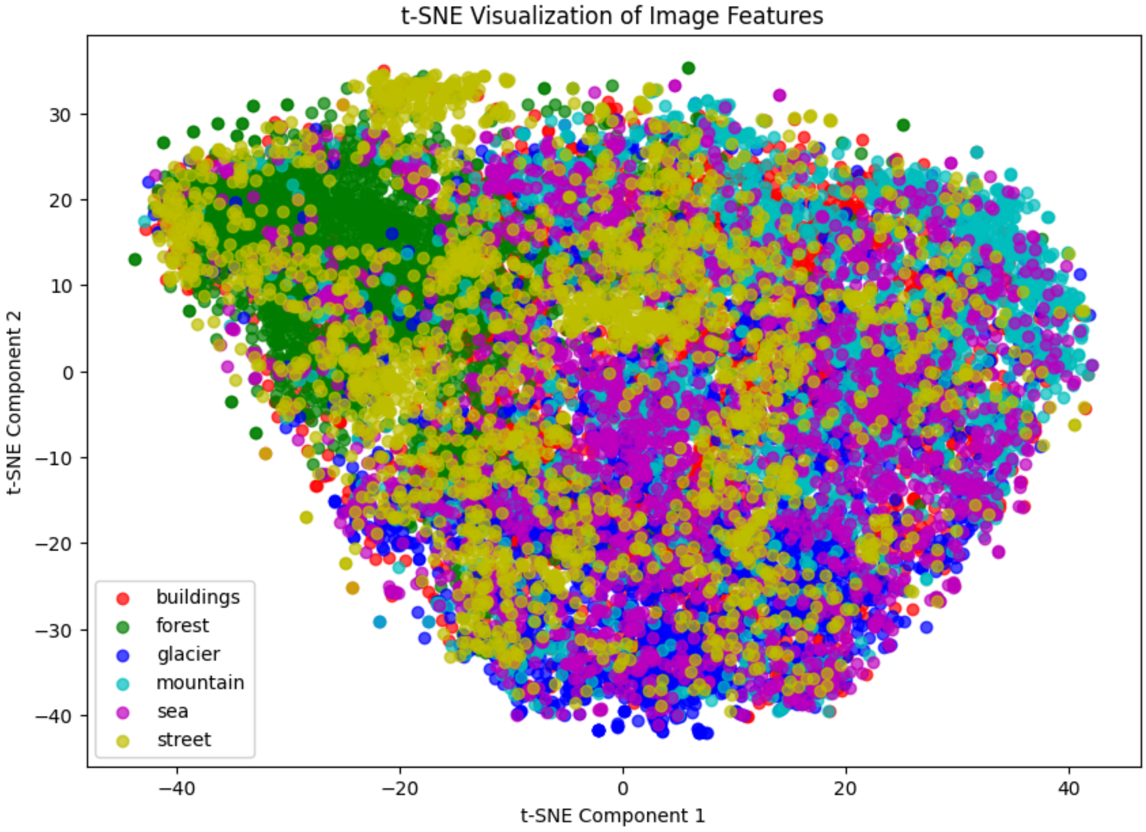

Figure 5: t-SNE Results

The t-SNE visualization offers a non-linear projection of the dataset, improving on PCA by better preserving local feature structure. While some categories, like forest, form tighter clusters due to their uniform brightness and smooth textures, others, such as buildings and streets, exhibit significant overlap. This overlap reflects shared structural features like edges and linear patterns, particularly within urban categories. Natural categories like forests, mountains, and seas display broader dispersion, with partial clustering for forests and overlap between mountains and glaciers, due to shared natural textures and color tones.

Compared to PCA, t-SNE provides better separation for certain categories, such as glaciers and forests, while still struggling with complex overlaps, particularly between urban and natural scenes. Categories with high variability, such as mountains and seas, remain more dispersed, highlighting the difficulty of isolating scenes with diverse features.

While t-SNE effectively reveals local relationships and achieves partial clustering, significant category overlap underscores the need for advanced feature engineering, such as texture- or context-sensitive methods, to further enhance separability and classification accuracy. This visualization demonstrates t-SNE's strength in uncovering data structure while emphasizing its limitations for datasets with complex, overlapping features.

Classifiers & Results

Data Splitting

Before implementing our classifiers, we split 20% of the 14,034 training data (2,807 images) for validation, leaving 11,227 images for training, 2,807 for validation, and 3000 for test. This three-way split into training, validation, and test sets allowed us to:

1. Train our models on a substantial portion of the data (11,227 images)
2. Tune hyperparameters using the validation set (2,807 images) without contaminating our final test results
3. Evaluate final model performance on the original test set (3,000 images) that remained completely unseen during both training and tuning phases

This partitioning strategy helped ensure our experimental results would generalize well to new, unseen data while providing sufficient samples for both training and validation.

Perceptron

The perceptron, one of the foundational machine learning models, serves as an essential baseline for understanding classification challenges in computer vision. In our analysis, we evaluated its performance across three distinct preprocessing techniques: raw images, HSV color space transformation, and Histogram of Oriented Gradients (HOG) features, focusing on both classification accuracy and computational efficiency.

	Test Accuracy	**Training Time Inference**	**Classification Time (per Image)**
Original Images	40.0%	7 m 29 s	0.000467 s
HSV	34.0%	7 m 14.6 s	0.000167 s
HOG	43.0%	9 m 3.2 s	0.001633 s

Figure 6: Perceptron Results

From the results illustrated in the table above, our experimental results revealed that HOG features achieved the highest accuracy at 43% on both validation and test sets, though this came at the cost of longer training times (9m 3.2s). Raw images and HSV transformation showed comparatively inferior performance, with accuracies of 40% and 34% respectively. The model demonstrated significant variability across different categories, with forest classification achieving a notable 82% recall, while urban and natural scenes proved more challenging to differentiate.

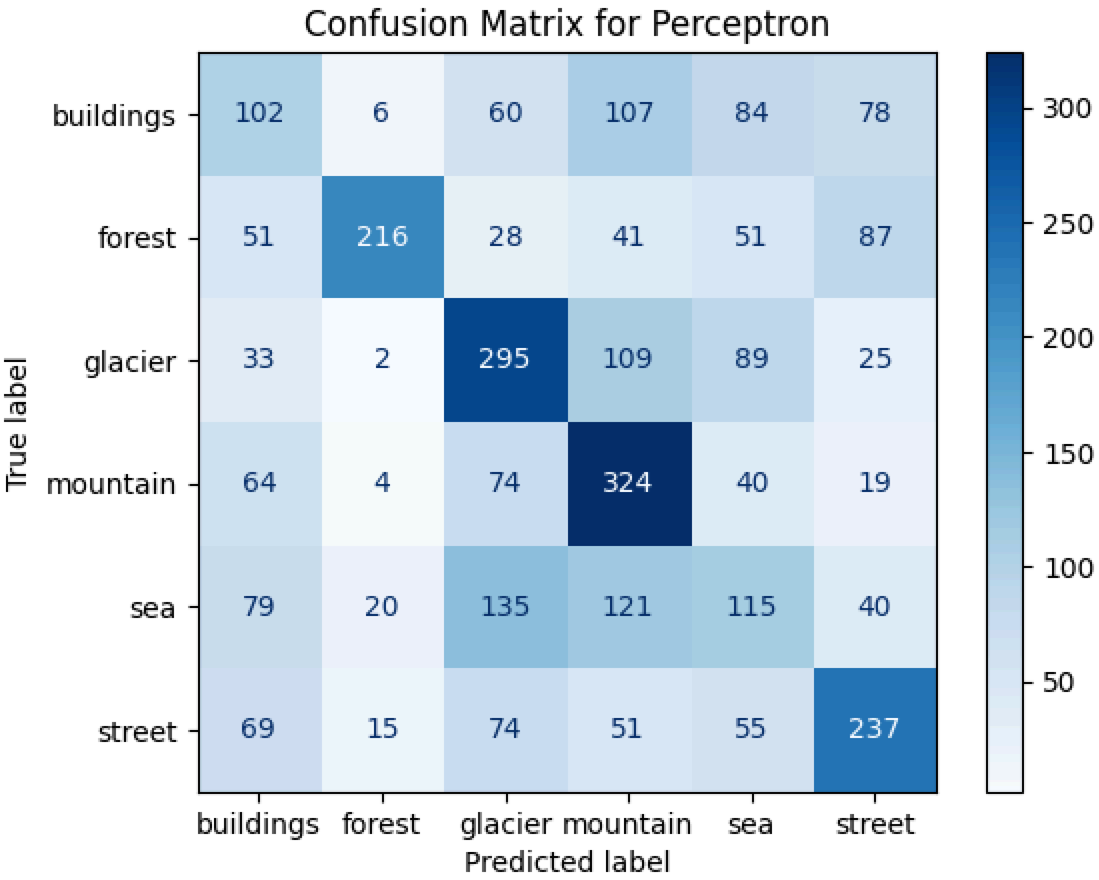

Figure 7: Confusion Matrix for Perceptron on Test Data

Error analysis through the confusion matrix unveiled systematic misclassifications. Urban categories like buildings and streets were frequently confused due to their shared geometric patterns, while natural scenes such as sea and glacier categories showed significant overlap, likely due to similar color distributions and reflective properties. The forest category stood out with 82% precision, as mentioned before, indicating reliable positive predictions despite lower recall rates.

This analysis not only demonstrates the historical significance of the perceptron but also clearly illustrates its limitations in modern computer vision tasks. The findings underscore the necessity for more advanced deep learning architectures in practical applications, particularly when dealing with complex, multi-category image classification challenges. Its linear boundary constraints severely limit its effectiveness in capturing non-linear relationships, while the high-dimensional nature of image data poses additional challenges. The significant performance disparity across categories clearly suggests the need for more sophisticated architectures. As such, we further explored logistic regression, SVM (with PCA & ResNet) & convolutional neural networks (CNNs) for better spatial feature extraction, and advanced preprocessing techniques to enhance feature discrimination.

Logistic Regression

Logistic regression extends beyond simple binary classification by providing probabilistic predictions for each class. Unlike the perceptron's hard decisions, this model assigns probability scores to indicate the likelihood of an image belonging to each category. Our implementation utilized a maximum of 2000 iterations to ensure convergence and evaluated the model's performance across three preprocessing techniques: raw images, HSV color space transformation, and Histogram of Oriented Gradients (HOG) features.

	Test Accuracy	**Training Time Inference**	**Classification Time (per Image)**
Original Images	41.0%	14 m 7.7 s	0.00073 s
HSV	42.0%	17 m 3.0 s	0.00390 s
HOG	49.0%	23 m 30.0 s	0.00617 s

Figure 8: Logistic Regression Results

The model achieved its best results using HOG feature extraction, reaching a 49% accuracy on the test set as seen in Figure *8*, marking a substantial improvement over the perceptron model. Performance varied significantly across categories, with forest scenes achieving the highest recall at 92% and street scenes following at 77%. Building categories maintained balanced precision and recall scores, demonstrating the model's ability to effectively identify consistent architectural features. However, challenges persisted in distinguishing between visually similar natural categories, with mountain, glacier, and sea showing lower performance metrics and significant confusion between them.

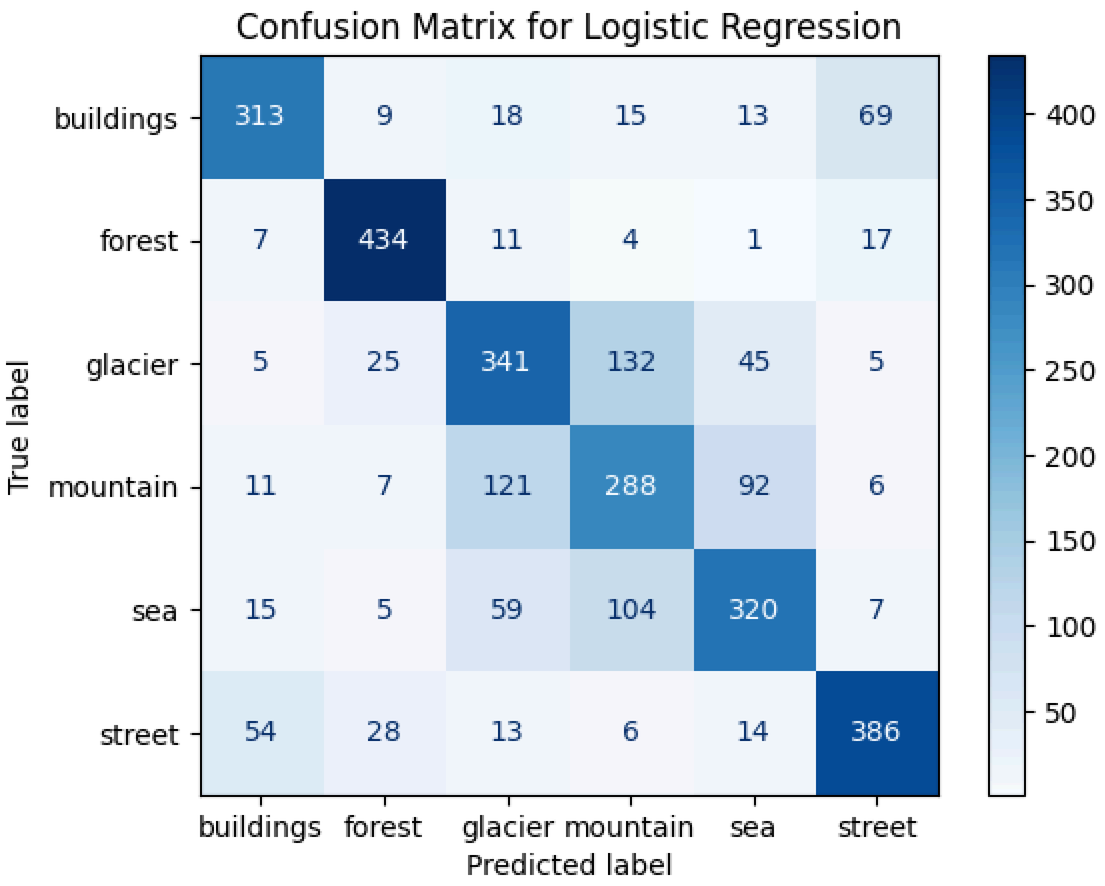

Figure 9: Confusion Matrix for Logistic Regression on Test Data

While logistic regression offers more sophisticated decision boundaries than the perceptron, its inherent linearity still constrains its ability to separate visually complex categories effectively. The model's performance suggests that more advanced approaches might be necessary, such as Support Vector Machines (SVMs) for better handling of non-linear boundaries, Neural Networks for capturing complex feature hierarchies, or ensemble methods for improving robustness across categories. This analysis demonstrates both the advantages of logistic regression over simpler models and its limitations in complex image classification tasks, pointing toward the need for more sophisticated machine learning architectures in challenging computer vision applications.

SVM with Hyper Parameter Search

C	Gamma	Mean Score
0.1	auto	0.4029

1	scale	0.5725
1	auto	0.5734
10	scale	0.5677
10	auto	0.5693
100	scale	0.5525
100	auto	0.5526

Figure 10: SVM Hyperparameter Search

In our investigation of scene classification using the Intel Image Dataset, we implemented Support Vector Machines (SVM) with various feature extraction techniques. The model's performance was optimized through a grid search over hyperparameters using a subset of validation data, with the best configuration found to be an RBF kernel with C=1 and gamma='auto'. This hyperparameter configuration balanced model complexity with generalization ability.

Feature	**Test Accuracy**	**Training Time**	**Classification Time (per Image)**
HOG (PCA = 100)	68.00%	1.84s	0.002034s
HSV (PCA = 50)	53.00%	1.53s	0.015857s
ResNet (PCA = 128)	62.50%	2.02s	0.017395s

Figure 11: SVM Results

Our analysis encompassed different approaches including HOG features (with PCA=100), HSV color space (with PCA=50), and ResNet features (with PCA=128). The HOG features demonstrated superior performance, achieving 68% test accuracy with efficient training (1.84s) and classification times (0.002034s per image). This outperformed both the HSV approach (53% accuracy) and ResNet features (62.5% accuracy).

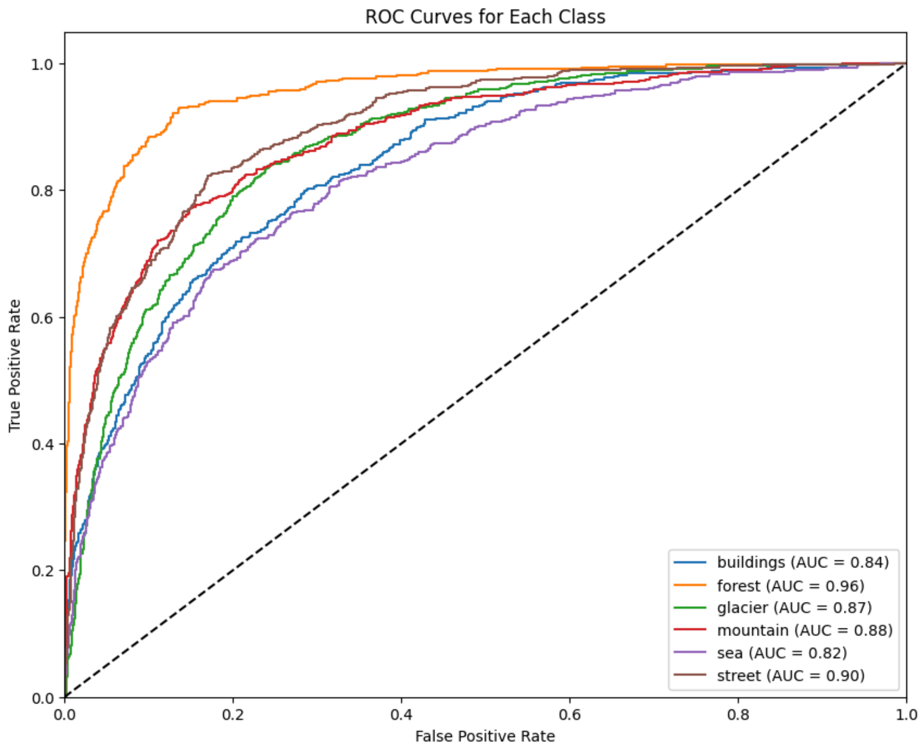

Figure 12: ROC Curves for Each Class SVM

The model's performance varied significantly across different scene categories, as evidenced by our ROC curve analysis. Forest classification showed particularly strong results with an AUC of 0.96, followed by street (AUC=0.90) and glacier (AUC=0.87) categories. The confusion matrix below revealed that while the model excelled at certain categories, it struggled with others - notably the sea category (AUC=0.82) showed significant confusion with other natural scenes. The classifier demonstrated strong diagonal elements in the confusion matrix, indicating good class separation for most categories, though some expected overlap occurred between visually similar classes such as natural scenes (forest, glacier, mountain) and urban environments (buildings, streets).

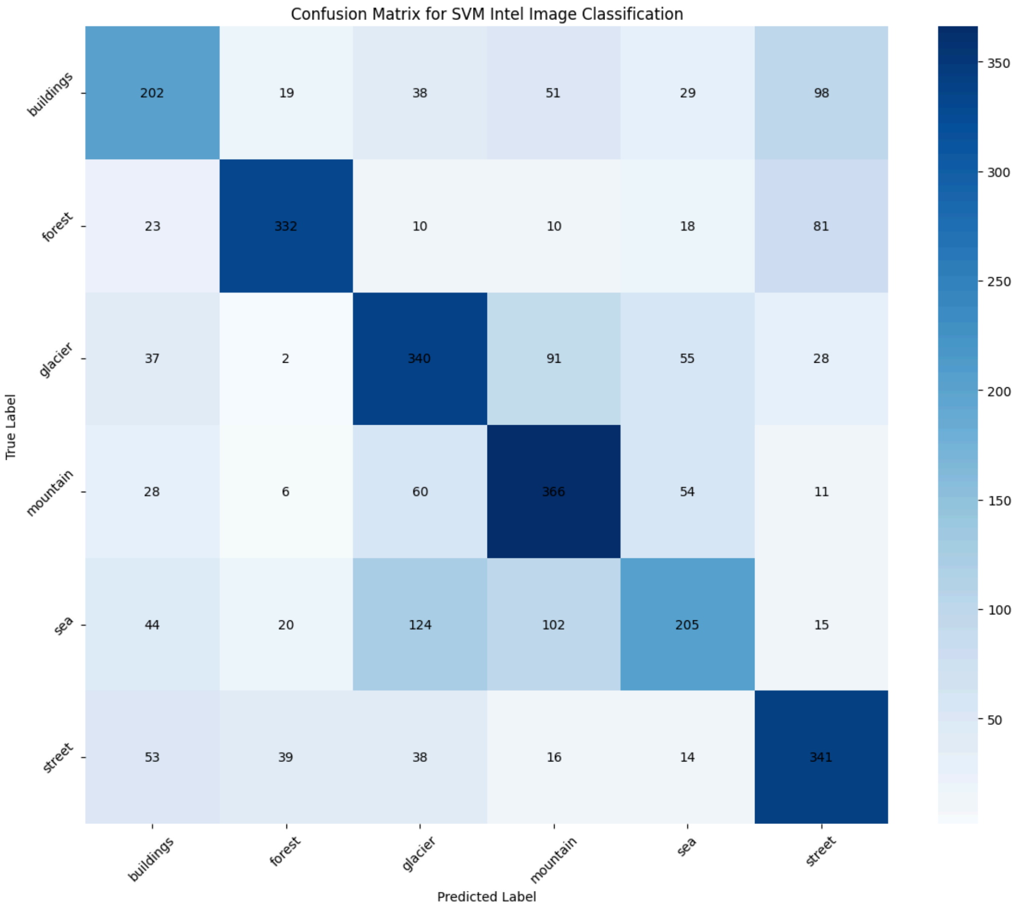

Figure 13: Confusion Matrix for SVM on Test Data

These results suggest that while SVM with HOG features provides a solid baseline for scene classification, there's room for improvement in handling visually similar categories. The model's strength in distinguishing forest and street scenes, coupled with its challenges in separating similar natural landscapes, points to the importance of feature engineering in capturing subtle visual distinctions. Our findings indicate that while SVM offers reasonable performance for this multi-class scene classification task, more sophisticated approaches like deep learning architectures are likely necessary to achieve higher accuracy across all categories.

CNN

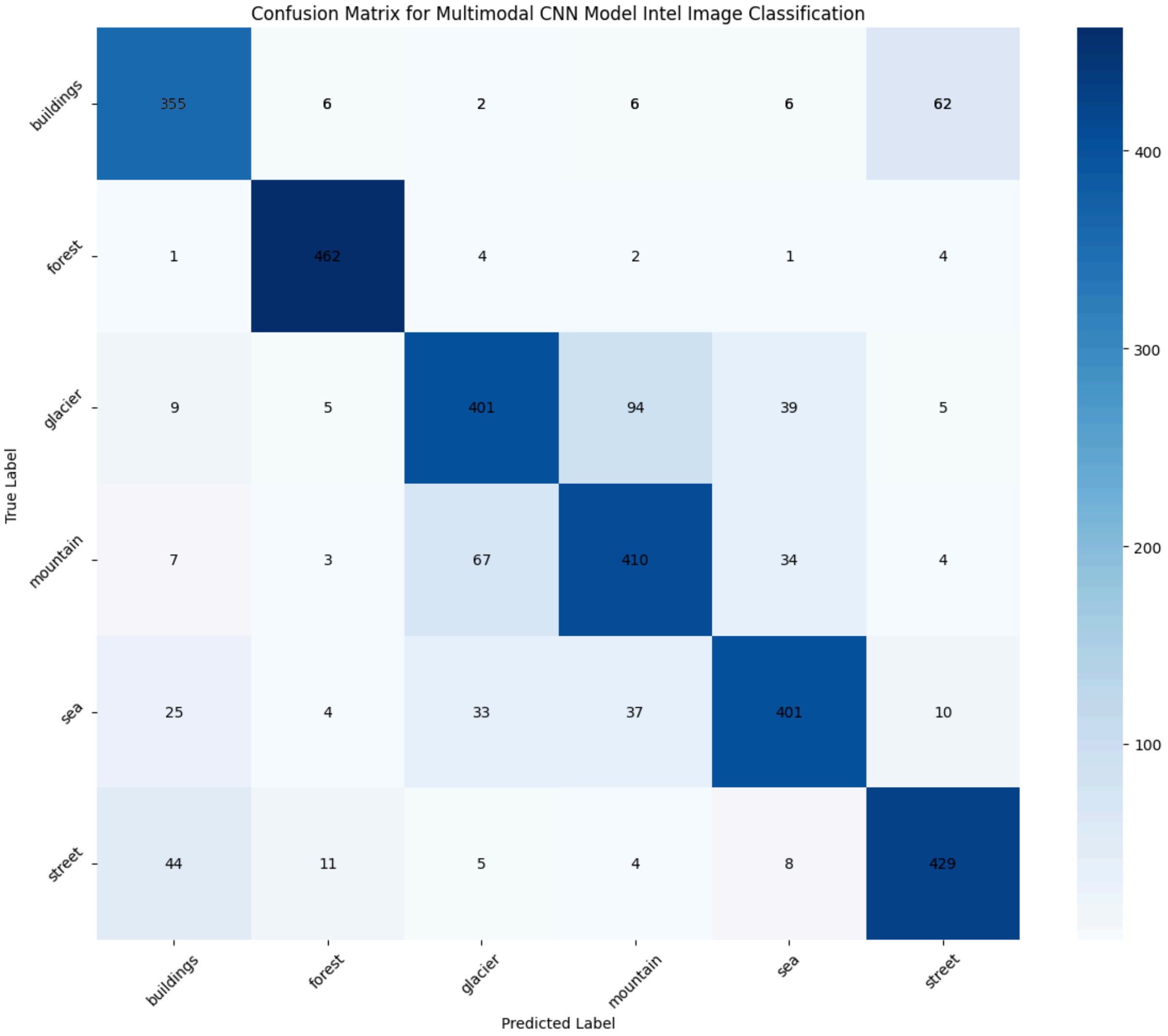

Figure 14: Confusion Matrix for CNN on Test Data

Naive

The CNN model, trained solely on raw image data, achieves an overall accuracy of approximately 79.7%. While this represents a substantial improvement over the previously reported performance, the confusion matrix reveals a heavy bias toward predicting the “buildings” category. The model consistently identifies buildings correctly but frequently misclassifies images from other classes—such as forest, glacier, mountain, sea, and street—as buildings or occasionally as glacier. This imbalance suggests that the model may be over-relying on certain visual cues strongly associated with the buildings category, potentially due to class distribution in the dataset or insufficient feature diversity extracted by the network. Although the numerical accuracy appears relatively high, it is likely influenced by underlying class imbalances and does not reflect uniformly strong performance across categories. Further refinement, including more balanced training data, enhanced augmentation strategies, or architectural adjustments, could help the model distinguish the other categories more effectively.

Multimodal

The multimodal CNN model incorporates both raw image inputs and engineered feature sets (such as HOG, HSV, and ResNet-based embeddings) to provide a richer representation of each scene. By combining these complementary feature spaces, the model significantly enhances its classification capabilities and reaches an accuracy of approximately 81.9%. This improvement over the naive CNN underscores the value of integrating multiple information sources.

Analysis of the confusion matrix reveals a more balanced performance across categories compared to the image-only CNN. Categories that previously posed challenges—such as urban or sea-oriented scenes—benefit from the additional features. For example, what once appeared to be uniform structures in “buildings” or subtle color gradients in “sea” now stand out due to the added feature sets. HOG descriptors may help distinguish line patterns common in cityscapes, HSV histograms can emphasize unique color distributions, and ResNet-based embeddings provide higher-level abstract representations that complement the CNN’s original pixel-based filters.

Despite the clear improvements, some confusion still persists between classes with shared visual characteristics, such as “glacier” and “mountain” or “forest” and “street.” While the added features reduce uncertainty, overlapping textures, lighting conditions, and seasonal variations present persistent classification hurdles. Thus, while the multimodal CNN approach proves more robust than single-stream methods, there is still room for refinement. More targeted feature engineering, advanced data augmentation (e.g., simulating different weather conditions or viewpoints), or attention-based mechanisms could help the model focus on the most discriminative elements of each scene, pushing performance even closer to ideal.

ResNet

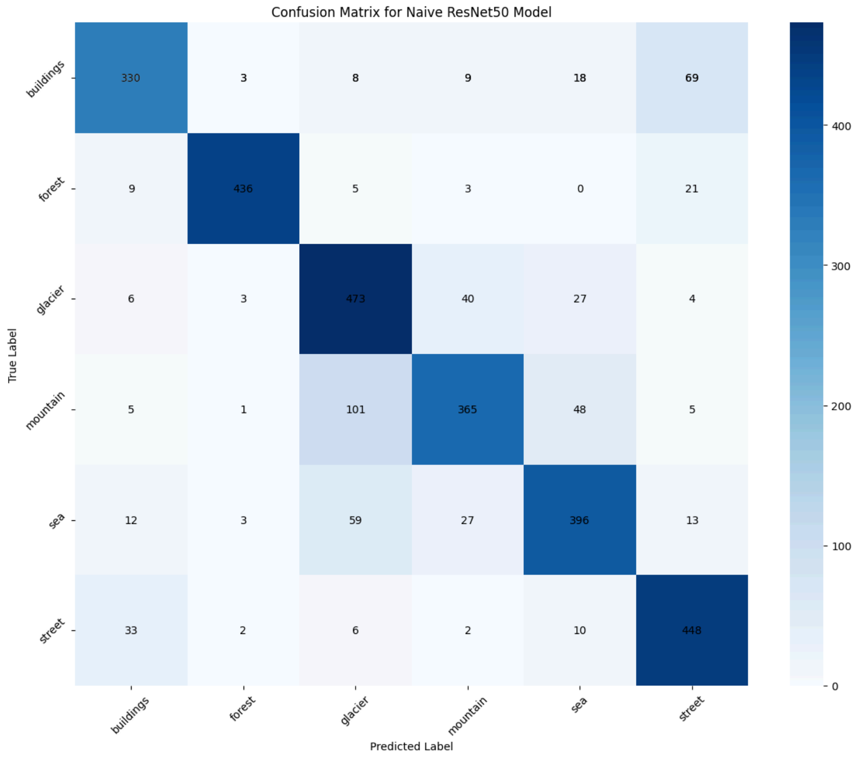

Figure 15: Confusion Matrix for ResNet on Test Data

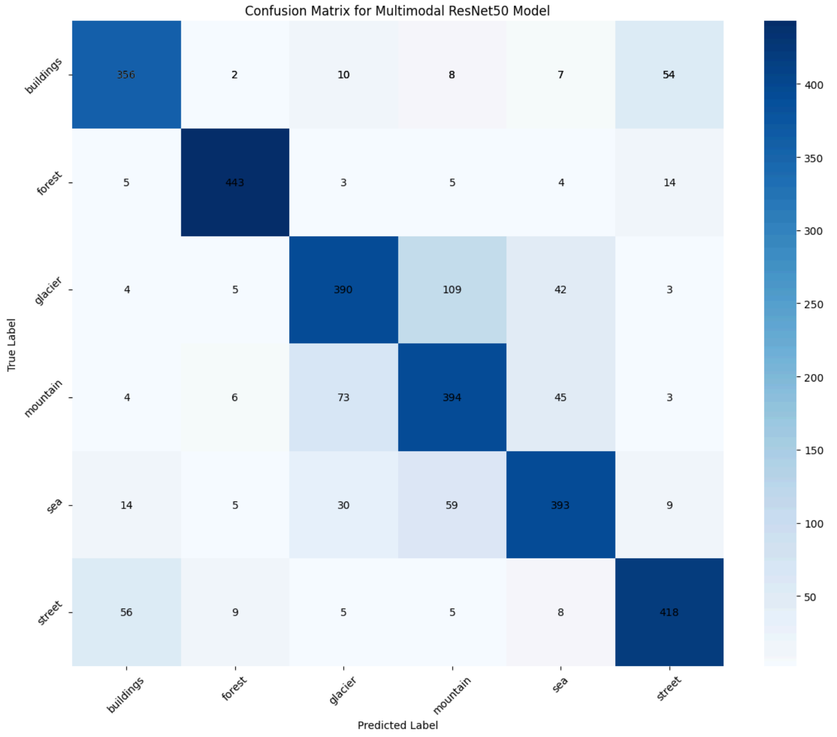

Figure 16: Confusion Matrix for ResNet on Test Data

Naive

The ResNet50-based model, despite leveraging a more advanced architecture and transfer learning from ImageNet, attains about 69.6% accuracy, which is lower than the naive CNN in this scenario. Contrary to initial expectations that a deeper, pre-trained network would significantly outperform simpler models, the ResNet50 model here does not achieve the previously stated high accuracy nor the robust class-by-class performance. It still may capture more nuanced features than simpler architectures, but the results indicate that either the model has not been fine-tuned sufficiently for the given dataset or that the particular training strategy and hyperparameters were suboptimal. Modifications such as unfreezing more layers, applying more targeted data augmentation, or tuning hyperparameters (learning rates, batch sizes, or number of epochs) could improve its discriminative power and close the gap with the CNN or surpass it.

Multimodal

The multimodal ResNet50 model leverages a pre-trained feature extractor integrated with engineered vectors to create a hierarchical representation of scene elements. Although it achieves about 79.8% accuracy—slightly below that of the multimodal CNN—it still represents a meaningful improvement over the naive ResNet50 baseline. By incorporating multiple feature types, the model more effectively captures subtle differences that raw pixel data alone might miss. Glacier images, for example, benefit from HOG features that highlight distinct ice formations, and HSV histograms can bring out nuanced color palettes in natural scenes. Urban categories also gain from integrating handcrafted features that emphasize geometric patterns and color transitions not immediately evident in raw images.

Nevertheless, the multimodal ResNet50 model does not reach the multimodal CNN's level of accuracy. This may be due to relying too heavily on pre-trained features geared toward general object recognition. Additional fine-tuning—unfreezing more layers, experimenting with attention-based feature fusion, or employing aggressive data augmentation—could help bridge the gap. Although not as successful as the multimodal CNN with current settings, these findings demonstrate that carefully chosen supplemental features can enhance performance, and that deeper optimization efforts may yield even more robust scene classification.

Model	**Test Accuracy**	**Training Time**	**Classification Time (per Image)**
CNN Naïve	79.70%	13m 33.6s	0.002034s
CNN Multimodal	81.93%	13m 37.5s	0.001954s

ResNet50 Naïve	69.60%	54m 47.4s	0.015857s
ResNet50 Multimodal	79.80%	58m 24.6s	0.017395s

Figure 17: Summary CNN & ResNet Results

Discussion

Our experiments highlight the complexities and subtleties involved in scene classification tasks, particularly when dealing with visually similar categories such as glaciers and seas or forests and mountains. The progression from linear models such as perceptron and logistic regression to SVMs, CNNs, and transfer learning models like ResNet50 underscores the importance of selecting architectures and features suited to the inherent complexity and variability of environmental images.

Early linear models struggled to capture nuanced visual cues. Although logistic regression offered some improvement over the perceptron by assigning probabilistic class scores, the underlying linear boundaries remained insufficient for robust scene differentiation. SVMs demonstrated notable gains in performance, especially when paired with HOG features. This combination offered finer-grained capture of structural details and provided a solid baseline for scene classification. Nevertheless, confusion persisted in visually similar categories where subtle differences in texture, lighting, or color distributions eluded these traditional approaches. **Still, SVM demonstrates perhaps the best tradeoff between training time/effort and performance.**

Our deep-learning experiments revealed further improvements and nuances. A naive CNN significantly outperformed traditional machine-learning models, confirming that hierarchical feature extraction is better suited to the complexity of natural and urban scenes. Multimodal CNNs performed even better by incorporating complementary handcrafted feature sets—such as HSV histograms, HOG descriptors, and feature embeddings from ResNet—into the learning process. This fusion of data-driven filters and engineered features mitigated some of the persistent classification challenges, enabling the model to better differentiate subtle texture gradients or lighting conditions that might blur category boundaries.

Interestingly, the naive ResNet50 model did not yield the dramatic boost initially anticipated. This outcome suggests that off-the-shelf transfer learning, while powerful, requires careful fine-tuning and consideration of domain-specific features. Conversely, combining ResNet features with other modalities dramatically improved accuracy, though not to the extent achieved by the multimodal CNN. This indicates that even advanced pre-trained architectures benefit from tailored adjustments and supplementary feature engineering that aligns with the dataset's particular characteristics. Given the substantial improvements gleaned from a **multimodal ResNet50, this model would likely benefit most from additional fine tuning and feature vectors.**

In examining these results, two primary themes emerge. First, richer feature representations improve classification across a range of categories, from highly structured urban environments to more amorphous natural scenes. Second, while deep learning approaches offer robust solutions, they are not universally superior without careful curation and integration of diverse features. The underlying complexity of environmental imagery—variations in brightness, color, texture, and scale—requires multifaceted strategies that combine powerful model architectures with context-sensitive preprocessing and feature selection.

Generalizability

A key aspect in computer vision and machine learning is ensuring models can generalize effectively to new, unseen data rather than simply memorizing the training set. This becomes particularly important as our model techniques increase in complexity. In our project, we implemented several strategies to promote good generalization, including rigorous dataset partitioning (training, validation, and test sets), data preprocessing to minimize noise and outliers, and maintaining balanced class distributions across all splits.

While linear models like Perceptron and Logistic Regression typically offer better generalization due to their inherent architectural constraints, our experiments revealed interesting patterns in their behavior with different feature representations.

The Perceptron model exhibited significant overfitting when trained on raw images and HSV features, with the gap between training and validation accuracy reaching approximately 50%. Similarly, Logistic Regression showed perfect training accuracy (100%) on both raw images and HSV features but performed poorly on validation sets, indicating severe overfitting. However, a notable improvement emerged when using HOG (Histogram of Oriented Gradients) features. Despite achieving a lower training accuracy of 84%, the HOG-based model demonstrated superior generalization with a validation accuracy of 69% - a much smaller gap that suggests more robust learning of underlying patterns rather than memorization.

To mitigate generalization for SVM, we applied hyperparameter fine tuning by applying GridSearchCV with 3-fold cross-validation on the training data. Ultimately, the grid search identified optimal parameters (C=1, gamma='auto'), achieving a cross-validation score of 0.57. When evaluated on the held-out validation set, the model achieved 60.49% accuracy, with test set performance at 59.53%. This small gap between validation and test performance (less than 1%) suggests good generalization, as the model performed consistently on unseen data. This suggests the model may have overfit to specific visual patterns common in building images rather than learning truly generalizable features across all categories.

For our CNN implementations, we observed distinct generalization patterns between naive and multimodal approaches. The naive CNN achieved 79.7% accuracy on the test set but showed signs of overfitting through its heavy bias toward the "buildings" category and frequent misclassification of other classes. The multimodal CNN demonstrated better generalization characteristics, reaching 81.93% test accuracy with more balanced performance across categories. The integration of engineered features (HOG, HSV, and ResNet embeddings) appeared to provide regularizing effects, helping the model learn more robust and transferable representations. However, persistent confusion between visually similar classes (like glacier/mountain and forest/street) indicates room for improved generalization through more sophisticated data augmentation strategies or architectural modifications.

Our ResNet50 experiments revealed unexpected generalization challenges. Despite leveraging transfer learning from ImageNet, the naive ResNet50 achieved only 69.6% test accuracy, significantly underperforming compared to the simpler CNN architecture. This suggests that while the pre-trained features were comprehensive, they may not have been optimally adapted to our specific scene classification task. The multimodal ResNet50 showed improved generalization (79.8% test accuracy) but still fell short of the multimodal CNN's performance. This indicates that the deeper architecture and pre-trained weights, without proper fine-tuning strategies, don't necessarily guarantee better generalization.

Overall, our experiments across different architectures revealed important insights about generalization in scene classification. While simpler models like HOG-based Logistic Regression showed reasonable generalization (69% validation accuracy), more complex architectures demonstrated higher absolute performance but varying generalization capabilities. The multimodal CNN emerged as the best performing model (81.93% test accuracy) with balanced class performance, suggesting that combining engineered features with deep learning can enhance generalization. However, the unexpected underperformance of ResNet50, despite its sophisticated architecture, highlights that model complexity and transfer learning alone don't guarantee better generalization.

These findings emphasize the importance of thoughtful feature engineering, appropriate model selection, and careful training strategies in achieving robust generalization.

Conclusion

Our findings emphasize that enabling machines to “see the water” they swim in is neither trivial nor guaranteed by simply adopting state-of-the-art models. Although deep-learning approaches outperformed simpler linear methods, no single technique fully resolved the classification challenges posed by our complex, multi-category dataset. Instead, success emerged from blending multiple strategies, whether through feature engineering or by integrating multiple modalities, to address the overlapping visual cues that confound straightforward classification.

The multimodal ResNet50 approaches offer a promising path forward. By combining raw pixel data with engineered feature sets, we achieved substantially higher accuracy (+10%) and more balanced performance across categories than the CNN equivalent . This approach represents a step toward more context-aware, interpretive modeling, as the integration of diverse feature representations begins to approximate the nuanced human perception that allows us to effortlessly recognize and categorize scenes.

Future work will benefit from exploring more sophisticated data augmentation methods, attention mechanisms that emphasize the most informative image regions, and further fine-tuning of transfer learning architectures. Integrating high-level contextual cues, such as weather conditions or geospatial data, may also improve classification outcomes. Ultimately, continued refinement of these techniques will help machines move beyond rote pattern matching toward truly contextual understanding, improving their ability to navigate and interpret the diverse visual “waters” of the real world.