Museum Image Classification
Indoor vs. Outdoor Using Supervised and Semi-Supervised Learning

[#]*COMP 6721 Applied Artificial Intelligence*

Devshree Shah
40269569
shahdevshree7@gmail.com

Sara Ezzati
40295160
sara84ez@gmail.com

Yug Kotak
40264255
yugkotak9745@gmail.com

***Abstract*— This project classifies museum images taken from Places MIT Dataset as indoor or outdoor using machine learning models. Key features such as color histograms, Local Binary Patterns (LBP), HOG, brightness, and edge density are extracted from the images. Supervised models like Decision Trees, Random Forest, and XGBoost are trained on these features, while a semi-supervised approach expands the labeled dataset using pseudo-labels from the Decision Tree. Results show XGBoost achieves the highest accuracy of 92%. Challenges like class imbalance and corrupt images are addressed, with suggestions for future work including deep learning and feature selection.**

***Keywords*— Museum Image Classification, Supervised Learning, Semi-Supervised Learning, Decision Tree, Random Forest, XGBoost, Feature Extraction, Pseudo-Labels, Image Processing.**

I. Introduction and problem statement

Image classification is a critical task in computer vision that involves categorizing images into predefined classes based on their content. With the rapid advancement of machine learning and deep learning techniques, image classification has found numerous applications across various industries, such as healthcare, autonomous driving, and entertainment.

In the context of museums, digitalization has been a pivotal part of efforts to preserve and share collections. Museums often capture images of exhibits, artifacts, and events for various purposes, such as archival records, promotional materials, and virtual exhibitions. However, classifying and organizing these images automatically is an ongoing challenge, particularly when distinguishing between indoor and outdoor settings. Given that museums house diverse exhibits in both indoor and outdoor environments, the ability to accurately categorize images based on their location can vastly improve the efficiency of curating and organizing these images for different uses.

This study aims to address the challenge of automatically classifying museum images as either indoor or outdoor. The primary obstacles include dealing with noisy or inconsistent data, the presence of visually similar images across categories, and the limited availability of labeled data for training machine learning models. In particular, distinguishing between indoor and outdoor images is complicated by various factors such as lighting conditions, similar architectural elements, and environmental features. Additionally, the lack of sufficient labeled data for model training adds to the difficulty. Therefore, the goal is to develop a robust and efficient classification model that can accurately differentiate between indoor and outdoor museum images, even with noisy, limited, and visually challenging data.

II. Proposed methodologies

This study focuses on proposing a methodology for classifying museum images as either indoor or outdoor. Specifically, we aim to explore various machine learning approaches for classifying images based on their contextual characteristics. The basic purpose of a classification study can be either to produce an accurate classifier or to uncover the predictive structure of the problem.[2] By tackling this problem, we intend to enhance museum operations by enabling more efficient management of image collections and streamlining the categorization process. Furthermore, the approach is valuable for museums looking to automate the curation of virtual galleries and enhance the visitor experience with online platforms.

The first objective is to create a machine learning model that can accurately classify museum images as either indoor or outdoor. The model should take into account various features of the image, such as lighting conditions, background textures, and overall visual context, to determine the correct category. Given the challenge of limited labeled data, this study aims to compare the performance of supervised and semi-supervised learning techniques.The dataset includes a variety of museum environments, with outdoor images featuring architecture, landscapes, and statues, while indoor images showcase exhibitions, lighting variations, and confined spaces. Exploratory Data Analysis (EDA) revealed no corrupt images, ensuring data quality. However, duplicate images were detected and skipped during the preprocessing phase. The dataset consists of indoor and outdoor museum images, with color distribution and pixel intensity differences observed between categories. Images were resized to 128x128 pixels, converted to grayscale, and histogram equalization was applied for better contrast. Canny edge detection was used for structural details. For feature extraction, Histogram of Oriented Gradients (HOG), color histograms, and Local Binary Pattern (LBP) were employed. The features were normalized using z-score normalization. The data was then used to train decision tree-based models (Decision Tree, Random Forest, XGBoost).

A. Supervised Decision-Tree

The Decision Tree classifier is a supervised learning model that splits the dataset into subsets based on feature values, creating a tree-like structure of decisions. Each node in the tree represents a feature, and branches represent decision rules. By utilizing decision trees, researchers aim to enhance the accuracy and efficiency of detecting physical activity types in various real-life settings, contributing to

the advancement of personalized healthcare solutions.[7] The leaves correspond to the predicted class. In this project, the Decision Tree model was trained on the preprocessed features extracted from the museum indoor and outdoor images. The Gini impurity was used to measure node purity, ensuring efficient splits during the training process.

The max_depth=5 parameter limits the depth of the tree to prevent overfitting and reduce the model complexity. The min_samples_split=10 ensured that nodes would only be split if they contained at least 10 samples, helping improve the model's ability to generalize.

B. Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their results to improve accuracy and reduce overfitting. The simplest random forest with random features is formed by selecting at random, at each node, a small group of input variables to split on.[3] Each tree is trained on a random subset of the data, and the final prediction is based on the majority vote from all trees. In this project, the Random Forest classifier was trained on the same set of preprocessed features used for the Decision Tree model. This model benefits from combining the predictions of many trees, leading to better generalization.

The max_depth=None allowed trees to grow to their full depth, enabling the model to capture more intricate relationships in the data. The min_samples_split=2 allowed splits even when a node had only two samples, promoting deeper tree growth. With n_estimators=200, the model built 200 trees in the forest to improve its overall performance. This resulted in a significant improvement in classification accuracy compared to the Decision Tree model, showcasing the power of ensemble learning for complex datasets like image classification.

C. XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting algorithm known for its high performance in classification tasks. It builds multiple weak models (decision trees) sequentially, each trying to correct the errors of the previous one. The scalability of XGBoost is due to several important systems and algorithmic optimizations. These innovations include: a novel tree learning algorithm is for handling sparse data; a theoretically justified weighted quantile sketch procedure enables handling instance weights in approximate tree learning.[6] For this project, XGBoost was trained on the preprocessed image features. The dataset was first converted into DMatrix, a memory-efficient format used by XGBoost to handle large datasets.

The subsample=0.8 parameter ensured that only 80% of the data was used for training each tree, helping reduce overfitting. The n_estimators=150 allowed for 150 boosting rounds, which provided a good balance between model complexity and training time. The max_depth=3 limited the depth of each individual tree to prevent overfitting, while the learning_rate=0.15 controlled the step size at each iteration. The gamma=0.1 helped control the complexity of the model by specifying a minimum loss reduction required to make a further split. Finally, colsample_bytree=0.7 randomly selected 70% of features for each boosting round, reducing the likelihood of overfitting.

D. Semi-Supervised Decision Tree

Semi-supervised learning leverages both labeled and unlabeled data to enhance model performance while minimizing labeling efforts. Here, we start by splitting the dataset into 20% labeled and 80% unlabeled data. Labeled and unlabeled data are represented as vertices in a weighted graph, with edge weights encoding the similarity between instances.[1] A Decision Tree classifier (max depth = 10) is trained on the labeled subset and then used to predict pseudo-labels for the unlabeled data. To ensure reliability, only predictions with at least 85% confidence are selected, reducing the risk of introducing noisy labels. The high-confidence pseudo-labeled data is then combined with the original labeled dataset, effectively expanding the training set. The classifier is retrained using this augmented dataset, improving its ability to generalize. This iterative process refines the model's decision boundaries while gradually incorporating more data.

This approach demonstrates the effectiveness of semi-supervised learning, especially when confident pseudo-labels can enhance model accuracy.

III. Solving the problem

The dataset, consisting of 5000 indoor and 5000 outdoor images, was first loaded, cleaned, and checked for duplicates to ensure quality. Decreasing the size of a group is always done by removing the data objects closest to the boundary.[5] Images were resized to 128x128 pixels and converted to grayscale to simplify the task, followed by histogram equalization to improve contrast. Canny edge detection was applied to enhance structural details for classification. For feature extraction, Histogram of Oriented Gradients (HOG) was used to capture texture and shape, while color histograms and Local Binary Patterns (LBP) were calculated to represent color distribution and fine textures. The extracted features were combined into a feature vector, which was then normalized using z-score normalization. Three models were trained and evaluated: a Decision Tree Classifier with hyperparameters (max_depth=5, min_samples_split=10, criterion='gini'), a Random Forest with 200 estimators, 10 iterations and max_depth=None, and an XGBoost model with n_estimators=150, max_depth=3, learning_rate=0.15, and gamma=0.1. Among these, the XGBoost model showed the highest accuracy, leveraging boosted trees and regularization to effectively capture complex patterns in the data.

A. Failed Attempts

a. Feature Reduction with PCA: To simplify the dataset and potentially enhance performance, we experimented with Principal Component Analysis (PCA) to reduce the feature dimensionality. We analyzed variance retention across different numbers of components and trained a Decision Tree on the transformed data. However, this approach led to a drop in accuracy (~0.645-0.67) compared to using the full feature

set (~0.77). The loss of critical image details seemed to hinder classification performance. This experiment highlighted the importance of preserving all extracted features, leading us to focus on optimizing models through hyperparameter tuning instead, which proved to be a more effective strategy.

b. XGBoost & Gradient Boosting Crashes: During our model training, we initially faced severe issues with XGBoost, as the training process kept crashing. Given the complexity of the image dataset and the computational demands of XGBoost, we suspected memory overload as a potential cause. To mitigate this, we switched to Gradient Boosting, hoping for a more stable training process. However, the issue persisted, preventing successful model execution. In response, we revisited XGBoost and carefully adjusted hyperparameters, such as reducing the number of estimators and tweaking learning rate and tree depth. These modifications successfully stabilized the training process, ultimately allowing XGBoost to run smoothly and deliver the highest accuracy among all models.

B. Results

The performance of the models was evaluated using accuracy, precision, recall, and F1-score to assess their classification effectiveness. Among the supervised models, XGBoost achieved the highest accuracy of 0.92, along with the best precision (0.92), recall (0.94), and F1-score (0.93). This highlights its ability to effectively capture complex patterns in the dataset and make robust predictions. Random Forest followed closely with an accuracy of 0.89, demonstrating strong generalization due to its ensemble learning approach. It maintained a high recall (0.90) and precision (0.89), making it a reliable choice for classification tasks.

On the other hand, the Decision Tree classifier, while computationally efficient, lagged in performance with an accuracy of 0.80 and slightly lower precision and recall (0.81 and 0.80, respectively). This indicates that the single-tree structure struggled with the complexity of the dataset, making it more prone to overfitting.

Overall, the results confirm that ensemble methods like Random Forest and XGBoost significantly outperform a standalone Decision Tree. The higher recall scores of XGBoost and Random Forest suggest better handling of class imbalances and misclassification reduction, making them more effective for the museum image classification task. The higher the True Positive Rate and the lower the False Positive Rate for each threshold, the better.[4]

TABLE I
EVALUATION METRICS FOR DECISION TREE, RANDOM FOREST, AND XGBOOST (HIGHEST ACCURACY)

Metric	**Decision Tree**	**Random Forest**	**XGBoosting**
Accuracy	0.80	0.89	0.92
Precision	0.81	0.89	0.92
Recall	0.80	0.90	0.94
F1-Score	0.80	0.89	0.93

Looking at Fig.1, Confusion matrices and classification reports provided deeper insights into model behavior, revealing that outdoor museum images were classified more accurately than indoor images. This discrepancy suggests that indoor images exhibit greater visual diversity, making them harder to categorize. Overall, the results demonstrated that advanced ensemble methods and semi-supervised learning approaches significantly enhanced performance, making them more suitable for this classification task.

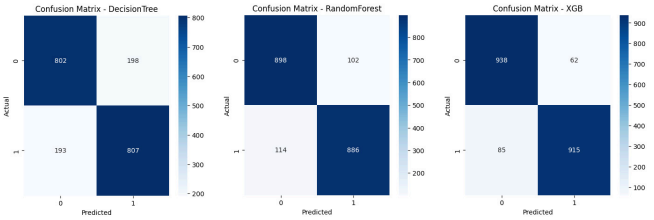

FIG 1: CONFUSION MATRICES FOR DECISION TREE, RANDOM FOREST, AND XGBOOST (HIGHEST ACCURACY)

IV. Future Improvements

For future improvements, several strategies can be explored to further enhance model performance and robustness. First, incorporating deep learning models such as Convolutional Neural Networks (CNNs) could significantly improve feature extraction by automatically learning hierarchical representations from image data. The accuracy of tied/untied CNNs is evaluated with various width, depth, and numbers of parameters.[8] This could eliminate the need for manual feature extraction methods like HOG and LBP. Additionally, leveraging transfer learning with pre-trained models like ResNet or VGG could allow for better generalization while reducing the need for extensive labeled data.

Another avenue for improvement is optimizing the semi-supervised learning approach. Instead of relying solely on Decision Trees for pseudo-labeling, more advanced techniques such as Self-Training with ensemble models or Graph-Based Label Propagation could be explored to improve label assignment accuracy. Furthermore, adjusting the confidence threshold for pseudo-label selection dynamically rather than using a fixed value might result in better model performance.

From a data augmentation perspective, introducing techniques such as random cropping, rotation, flipping, and brightness adjustment could help enhance model robustness and reduce overfitting. Additionally, expanding the dataset by including images from diverse museum environments and lighting conditions could improve the model's ability to generalize across unseen scenarios.

V. References

[1] Ghahramani, Z., Zhu, X., & Lafferty, J. (n.d.). Semi-supervised learning using Gaussian fields and harmonic functions. https://dl.acm.org/doi/10.5555/3041838.3041953

[2] Breiman, L., Friedman, J., Olshen, R.A., & Stone, C. J. (2017). Classification and Regression Trees. CRC Press LLC. https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-olshen-charles-stone

[3] Breiman, L. (January 2001). In RANDOM FORESTS. University of California Berkeley, CA 94720.

[4] Vujović, Ž. Đ. (2021). Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications. 10.14569/IJACSA.2021.0120670

[5] Kamiran, F., & Calders, T. (October 2011). Data Pre-Processing Techniques for Classification without Discrimination. 10.1007/s10115-011-0463-8

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. University of Washington. https://www.semanticscholar.org/reader/26bc9195c6343e4d7f434dd65b4ad67efe2be27a

[7] Abdulqader, H. A., & Abdulazeez, A. M. (2024). A Review on Decision Tree Algorithm in Healthcare Applications. Indonesian Journal of Computer Science. 10.33022/ijcs.v13i3.4026

[8] He, K., & Sun, J. (2015). Convolutional Neural Networks at Constrained Time Cost. (Microsoft Research). chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://openaccess.thecvf.com/content_cvpr_2015/papers/He_Convolutional_Neural_Networks_2015_CVPR_paper.pdf

[9] (n.d.). http://places.csail.mit.edu/browser.html

VI. Supplementary Material

During the evaluation of the test dataset, we assessed multiple models, including Decision Tree, Random Forest, and XGBoost. Among them, XGBoost achieved the highest accuracy, demonstrating superior performance in classifying the images. The comparison of model accuracies is visualized in the bar graph below, highlighting XGBoost as the best-performing model.

FIG 2: BAR GRAPH DEPICTING THE EVALUATION OF MODELS BASED ON TEST DATA