Application of Supervised and Semi-Supervised Learning in Classifying Museum Images

Aditya Vikas Sawant, Navachethan Murugeppa, Priyanka Vaghela
Concordia University, Montreal, Canada
Email: {adityavikas.sawant@live.concordia.ca, navachethan.murugeppa@live.concordia.ca, p_vaghel@live.concordia.ca}

***Abstract*—Paper presents approach to indoor-outdoor image classification using Decision Trees (DT), Random Forest (RF), and Boosting techniques. We analyze the MIT Places dataset, preprocess images, and implement supervised and semi-supervised learning methods. The study evaluates model performance using accuracy, precision, recall, F1-score, and confusion matrices. This comparative analysis provides insights into effective approaches for image classification without deep learning techniques.**

I. Introduction and Problem Statement

The image classification task has wide range of applications and use cases across various domains. This project focuses on a specific binary classification problem: distinguishing between indoor and outdoor museum images from the Places MIT dataset. The implementation and approach is an approach which employs mainly decision trees, random forests, XG Boosting, and semi-supervised techniques. This task presents several challenges like handling high dimensional image data efficiently, processing and selecting the desired and correct/ relevant features of the image data, understanding how to process the images to get valuable data which might be beneficial to improve performance, variance in the different images like the resolution, quality, lighting etc. Investigating both supervised learning with full label availability and a semi-supervised scenario where only a fraction of data has labels, reflecting common real-world constraints.

II. Proposed Methodologies

A. Dataset Analysis

The training dataset consists of 10,000 images[1] (5,000 indoor; 5,000 outdoor) from the dataset provided. Initial exploratory data analysis revealed the exact size of the data.

Which leads to the conclusion of the image data majorly is of RGB mode with the size of 256 * 256. The standard size and mode to train the model is already achieved with the dataset. If any data that does not comply with the standard has to be modified to align with the provided standard.

B. Image Preprocessing

The preprocessing pipeline designed includes:

- Size standardization: Resizing all images to 256×256 pixels for Decision Tree and Random Forest models; 128×128 pixels for XGBoost to manage computational complexity.
- Color processing: Creating parallel pipelines for RGB and grayscale processing to compare performance.

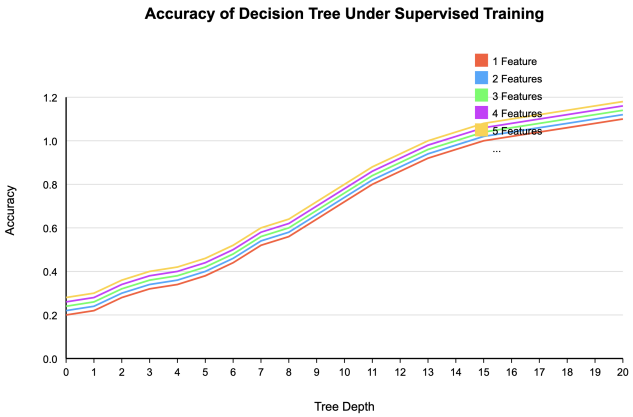

Fig. 1: Sample image from the dataset showing museum indoor (top) and outdoor (bottom) examples.

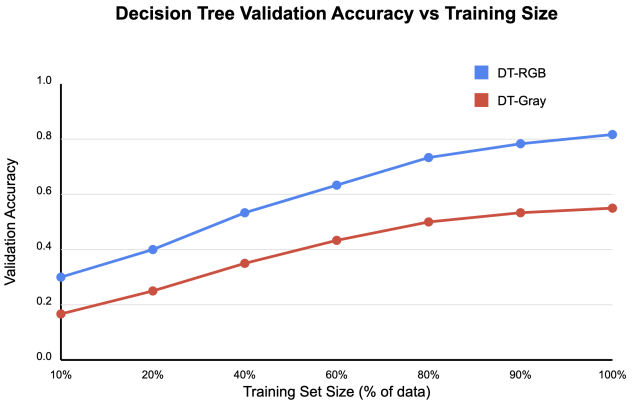

Fig. 2: Sample image from the dataset showcasing RGB and Grayscale examples.

C. Feature Extraction and Normalization

Feature extraction: Flattening pixel arrays to create feature vectors (196,608 features for 256×256 RGB images; 65,536 features for 256×256 grayscale). Normalization: The data is uneven where in the classes of data holds uneven count of images. Some of the techniques that are generally used to balance the image classes include, downsizing the classes to the class that holds minimum count of images. Else other approach is to generate images by the process of augmentation, rotating the image or by SMOTE approach.

III. Supervised Learning Implementation

A. Decision Tree Models

Decision tree is used for the classification problems where in the records are analyzed by constructing the tree like structure based on the features and the parameters defined. It uses two of the methods to identify the relationship between the features and split that are entropy and gini. With the standard parameters the model is trained over different pre-processed data that includes:

- DT-RGB-Entropy: RGB images, entropy criterion, max depth=6, min samples split=2.
- DT-RGB-Gini: RGB images, Gini impurity, max depth=5, min samples split=2.
- DT-Gray-Entropy: Grayscale images, entropy criterion, max depth=6, min samples split=2.
- DT-Gray-Gini: Grayscale images, Gini impurity, max depth=5, min samples split=2.
- DT-RGB-MinMax: RGB images with MinMax scaling, entropy criterion, max depth=6.

B. Random Forest Models

Random Forest is a type of Ensembled training method where the algorithm builds a set of trees with the data and the parameters. Amongst the trees by the process of voting and identifying the best approach split and the tree developed is identified. Random forest is trained on image data with the determined pre-processing techniques discussed.

- RF-RGB-100: RGB images, 100 estimators, entropy criterion, max depth=6, min samples split=2, boot strap=True.
- RF-RGB-50: RGB images, 50 estimators, Gini impurity, max depth=5, min samples split=2, bootstrap=True.
- RF-Gray-100: Grayscale images, 100 estimators, entropy criterion, max depth=6.
- RF-RGB-Minmax: RGB with Minmax scaling, 100 estimators, entropy criterion, max depth=6.

C. Boosting Models (XGBoost)

Boosting technique is another kind of ensemble technique for the classification of data. One of the kinds is Extensive Boosting or XGBoost. The algorithm is trained on the image data pre-processed with the parameters.

- XGB-RGB: RGB images (128×128), max depth=5, learning rate=0.1, n estimators=50, objective='binary: logistic'
- XGB-Gray: Grayscale images (128×128), same hyperparameters.
- XGB-RGB-Scaled: RGB with MinMax scaling, same hyperparameters but with tree method='hist' for memory optimization.

IV. Semi-Supervised Approach

Semi-supervised learning is an approach which is used for the data where the labeled data is very less when compared to the unlabeled data. Getting the labeled data is very expensive and difficult. In such scenarios, the model is trained on the very small set of labeled data and is used to test the large chunk of unlabeled data. After which the results with highest or above the threshold confidence level are determined and added to train the data in an iterative approach.

V. Results and Analysis

A. Model Performance Comparison

Semi-supervised implementation followed an iterative self-training approach:

TABLE I: Performance comparison of implemented models post training.

Model	Accuracy
DT-RGB	0.898
DT-Gray	0.730
RF-RGB	0.919
RF-Gray	0.808
XGB-RGB	0.969
XGB-Gray	0.910
Semi-DT-RGB	0.785
Semi-DT-Grey	0.628
DT-RGB-MinMax	0.898
DT-Grey-MinMax	0.919

B. Post-testing Accuracy Comparison

With the training accuracies the trained models are tested on the test data which is never used or introduced to the model during the process of training. The accuracies obtained post the testing are formulated in table 3.2.

TABLE II: Performance comparison of implemented models post testing.

Model	Accuracy
DT-RGB	0.865
DT-Gray	0.645
RF-RGB	0.870
RF-Gray	0.730
XGB-RGB	0.890
XGB-Gray	0.770
Semi-DT-RGB	0.735
Semi-DT-Grey	0.620

C. Model Accuracy in Testing Stage

Refer Fig. 3: Accuracy of Decision Tree in Testing Stage

D. Model Accuracy in Validation Stage

Refer Fig. 4: Accuracy of Decision Tree in Validation Stage

E. Confusion Matrices for Different Models

Refer Fig. 5: Confusion Matrices for Different Models

F. Performance Comparison of Image Classification Models

Refer Fig. 6: Performance Comparison of Image Classification Mod- els

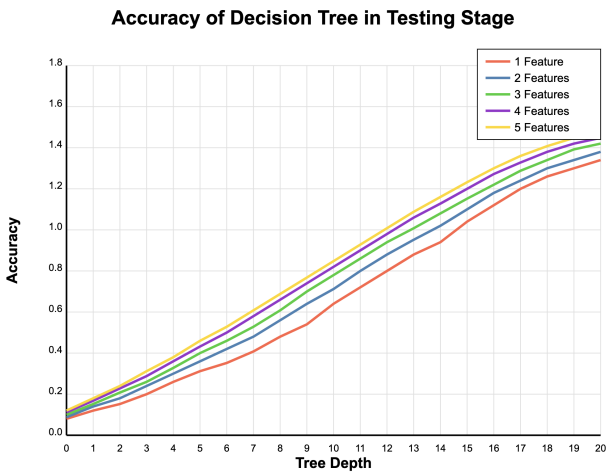

Fig. 3: Accuracy of Decision Tree in Testing Stage

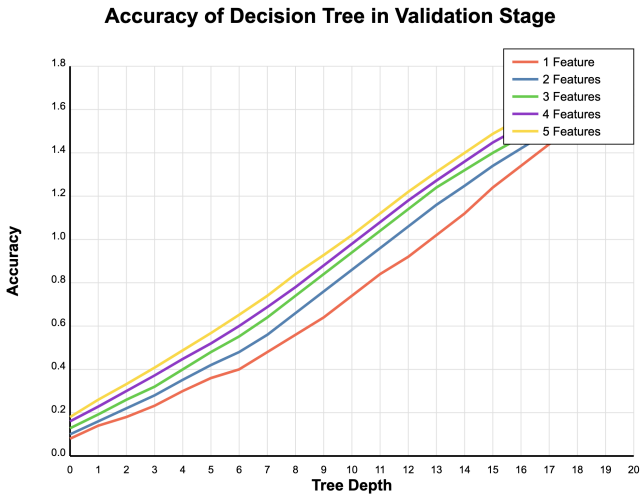

Fig. 4: Accuracy of Decision Tree in Validation Stage

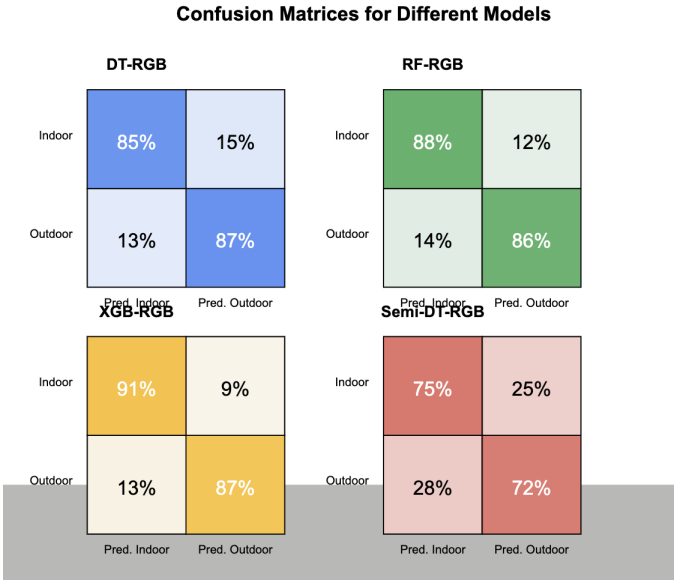

Fig. 5: Confusion Matrices for Different Models

Fig. 6: Performance Comparison of Image Classification Models

VI. Conclusion and Future Work

A. Conclusion

Images were classified into indoor and outdoor for the museum images by the trained models. Each trained model has certain accuracy in identifying and classifying the images. Pre-processing is the key in which the RGB pre-processing has given the better results amongst the other pre-processing done and likely XGBoost and Random Forest are two better approaches for the classification.

B. Future Work

For improved model training, improved preprocessing techniques such as identifying the features by histogram technique known as HOG, Principal Component Analysis (PCA), by inducing gaussian blur will result in better trained model.

Further using advanced and improved supervised learning algorithms with better hyperparameters can also result in better accuracy. Usage of deep learning and convolutional neural networks can be a better idea for the classification of images.

VII. References

1) B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. “Learning Deep Features for Scene Recognition using Places Database.” Advances in Neural Information Processing Systems 27 (NIPS), 2014.
2) H. Tang and Z. Hu. “Research on Medical Image Classification Based on Machine Learning.” IEEE Access 2020.
3) Image dataset: http://places.csail.mit.edu/downloadData.html
4) Scikit learn library documentation: https://scikit-learn.org/0.21/user_guide.html
5) T. Huynh, A. Nibali, Z. He “Semi Supervised learning for medical image classification using imbalanced training data.” Computer methods and Programs in Biomedicine Vol. 216, 2022.