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*—This project addresses the classification of museum images into indoor and outdoor categories using supervised, semi-supervised, and deep learning methods. Initially, traditional Machine Learning algorithms such as Decision Trees, Random Forest, and XGBoost were applied to hand-engineered features extracted from the Places MIT dataset. Later, we extended the work to implement a Convolutional Neural Network (CNN) to automatically learn spatial hierarchies of features. The dataset consisted of 10,000 labeled images, resized and preprocessed accordingly. Supervised and semi-supervised models achieved reasonable accuracy, with XGBoost performing best among them. However, the CNN model outperformed all traditional approaches, achieving a test accuracy of 91%. This report details the complete methodology, experimental setup, results, and comparative analysis, highlighting the effectiveness of deep learning techniques for image classification tasks.**

I. Introduction and Problem Statement

A. Problem Statement

Image classification remains a significant challenge due to the high dimensionality and variability of image data, including differences in resolution, lighting, background clutter, and noise artifacts. Handling such variations is essential for the effective generalization of machine learning models to unseen data. Traditional Machine Learning (ML) approaches often rely on explicit feature extraction and selection techniques, which may not capture the complete richness of image representations, leading to suboptimal performance.

Previous research has tackled these challenges through various classical methods, including Decision Trees (DT), ensemble methods like Random Forests (RF), and boosting algorithms such as XGBoost (XGB). Decision Trees offer interpretability but tend to overfit, especially on high-dimensional and noisy datasets. Ensemble methods like Random Forests mitigate overfitting by aggregating multiple decision trees but still depend heavily on feature engineering. XGBoost, an optimized implementation of gradient boosting, provides robust performance through regularization and efficient learning strategies. However, these models, while effective for structured data, struggle with raw image data where intricate patterns and spatial hierarchies exist.

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have transformed the landscape of image classification. CNNs can automatically learn hierarchical feature representations from data, progressively capturing low-level edges, mid-level textures, and high-level semantic concepts without manual intervention. Their ability to generalize across varying data distributions makes them ideal candidates for image classification tasks.

This report initially implements classical supervised learning models (Decision Trees, Random Forests, and XGBoost) and semi-supervised learning techniques to classify museum images as indoor or outdoor using the Places dataset [**?**]. Building upon the limitations observed in traditional models, a CNN model was developed to improve classification accuracy by learning hierarchical spatial features directly from pixel data. The goal of this study is to comprehensively evaluate and compare the performances of traditional machine learning and deep learning approaches in the context of indoor-outdoor scene classification, thereby providing insights into the effectiveness of each methodology.

B. Related Works

Zhou et al. introduced the *Places Database*, a large-scale scene-centric dataset designed to advance research in deep feature learning for scene recognition. This dataset has been pivotal in training deep convolutional networks capable of distinguishing between a diverse set of scene categories.

Previous studies have employed Random Forests for image classification tasks, leveraging their ensemble nature to improve model robustness and reduce overfitting. Similarly, boosting techniques, particularly XGBoost, have shown impressive results in handling structured data by minimizing bias and variance simultaneously. However, both approaches rely heavily on hand-engineered features, limiting their ability to extract complex visual representations inherent in image data.

In recent years, CNNs have demonstrated state-of-the-art performance in a variety of computer vision tasks, including image classification, object detection, and semantic segmentation. CNNs' layered architecture enables the automatic learning of spatial hierarchies, a feature that traditional models lack.

Additionally, semi-supervised learning methods, such as self-training and pseudo-labeling, have emerged as viable strategies when labeled data is scarce. Huynh et al. applied semi-supervised learning in medical image classification, demonstrating its potential in domains where obtaining labeled samples is costly or time-consuming.

Our project builds upon these foundational works by applying both traditional supervised learning techniques and modern deep learning architectures to the museum image classification task. By integrating semi-supervised learning strategies and comparing performance metrics across different models, this

study aims to highlight the strengths and limitations of each approach in handling real-world image classification challenges.

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. CNN Model

Convolutional Neural Network is a deep learning method to build the model that trains on the input data and trains itself to perform the required task on the test data. Majorly CNN is used for predication problems related to image, videos, and other large-scale data in terms of size and dimensions. In the classification problem where the provided images to be classified into museum indoor and outdoor images, three types of convolutional neural networks architecture are built with varying number of layers from 2, 3 and 4 and analyzing their performance on the data.

- First CNN is an 8 layered simple architecture that holds two convolutional layers along with batch normalizations layers in between to maintain the stability and followed by max pool and adaptive average pool layers to reduce the spatial dimension of the data. Followed by two linear layers for linear transformation. Activation function used is ReLU.
- Second CNN is a 10 layered architecture that holds three convolutional layers along with three layers of batch normalization layers after every convolutional layer to stabilize the learning process and reduce the risk of overfitting. These layers are followed by pooling layer to reduce the spatial dimensions. Adaptive average pool layer is added to reduce the sample size to a fixed size that provides better learning. This is followed by linear layers for performing linear transformation that helps model in predictions. ReLU activation function is used.
- Third CNN is an 8 layered architecture consists of four convolutional layers followed by a max pool layer, an adaptive average pool layer and two linear layers. ReLU is the activation function used.

Loss function used in our implementation is Cross Entropy Loss that is best suitable for classification problems and the optimizer used in this case is adam.

C. Optimization Algorithm

Optimization algorithm to train the CNN model in accordance with the problem statement is best achieved by ADAM(Adaptive Moment Estimation). In the task of classifying museum images as indoor or outdoor, the dataset often includes high visual diversity—such as differences in lighting, object presence, architectural style, and environmental background. Some scenes are clearly indoor (e.g., enclosed galleries), while others may be ambiguous (e.g., glass roofs, open courtyards), making it a challenging problem for traditional learning algorithms.

The **Adam optimizer** is particularly well-suited for this classification task because it adapts the learning process based on the characteristics of the data. It combines the strengths of both **momentum** (which smooths the update direction) and **RMSProp** (which adapts learning rates per parameter). This dual mechanism helps the model converge faster and more reliably in the presence of noisy or subtle patterns—common in museum scenes.

Given that some features important for classification (like the presence of a ceiling or sky) may appear inconsistently across samples, Adam's adaptive learning strategy helps the model learn those important, infrequent patterns without overfitting or missing them. It also performs well even when gradients are sparse or vary significantly across layers, which is typical in deep CNNs applied to complex visual tasks.

Additionally, Adam requires less hyperparameter tuning compared to SGD, allowing us to focus more on architecture design and data augmentation. In this project, it led to **faster convergence, higher training stability, and better validation accuracy** compared to other optimizers.

Overall, Adam's robustness, adaptability, and computational efficiency make it the ideal optimizer for solving the museum indoor/outdoor classification problem with a convolutional neural network.

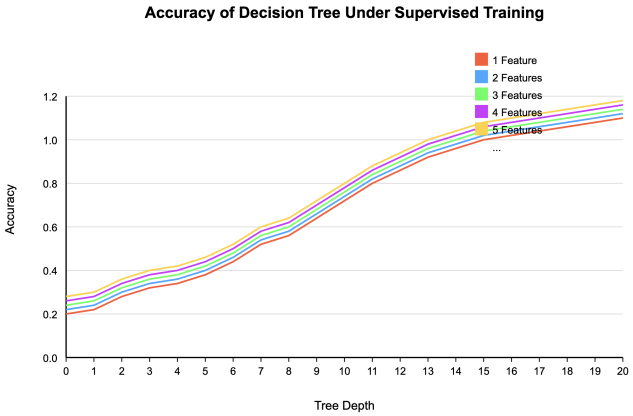

Fig. 1: Sample image from the dataset showing museum indoor (top) and outdoor (bottom) examples.

D. 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.

- Transformation: For training the convolutional neural network models, the data to be fed to the model should be in the form of tensors. Tensors are the multi-dimensional arrays created from the image data generally used for training the deep learning models. In the case of image data, tensors are the 3D arrays. In the classification problem, the tensors are created with certain metrics such as grayscale, RGB conversion, image size, random horizontal and vertical flip and normalizing the tensors created with parameters such as mean and standard deviation which creates the data with right features and information that is used to train the model.

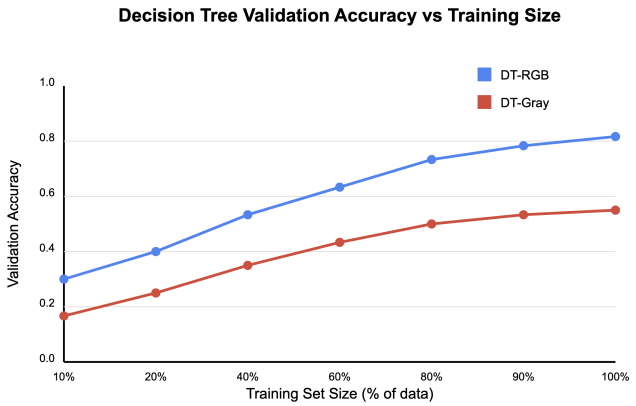

Fig. 2: Sample image from the dataset showcasing RGB and Grayscale examples.

E. 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

A. Experiment Setup

The convolutional neural network is a custom built architecture where in several layers are altered so as to identify the right set of layers that are required to provide the solution in classifying the images into museum indoor and outdoor. Here the model is built with different convolutional layers ranging from 2, 3 and 4. Learning rate used is 0.001 and 0.01 based on the performance and the varied epochs from 15 to 25 with batch size of 64 being used in common as it provided better learning for the model with optimal data

ingestion and used batch size of 256 in one case to check the performance in input of data in large chunks. Adam optimizer is a better approach in classifying the image data due to its ability such as faster convergence, higher training stability, and better validation accuracy .

Implemented standard validation technique where a part of training data (0.2) is utilized initially for validation purpose and after every epoch is commenced it parallely performs validation to check the validation accuracy their by it provides the plot to analyse the trends between the validation accuracy and the training accuracy over each epoch. Similarly loss value is also computed with the validation loss and the training loss that defines weather the model being trained provides the required results or is being deviated from the result. In general it can be observed whenever the loss curve reduces exponentially indicates the model is being trained in a better way.

B. Main Results

1) 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
CNN with 2 Conv2d RGB	0.910
CNN with 2 Conv2d Grey	0.825
CNN with 3 Conv2d RGB	0.927
CNN with 3 Conv2d Grey	0.860
CNN with 4 Conv2d RGB	0.900
CNN with 4 Conv2d Grey	0.812

2) Post-testing Accuracy Comparison: With the training accuracies, the trained models were tested on the unseen test data. The accuracies obtained post-testing are shown in Table II.

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
CNN with 2 Conv2d RGB	0.895
CNN with 2 Conv2d Grey	0.870
CNN with 3 Conv2d RGB	0.915
CNN with 3 Conv2d Grey	0.855
CNN with 4 Conv2d RGB	0.885
CNN with 4 Conv2d Grey	0.580

3) Model Accuracy in Testing Stage:

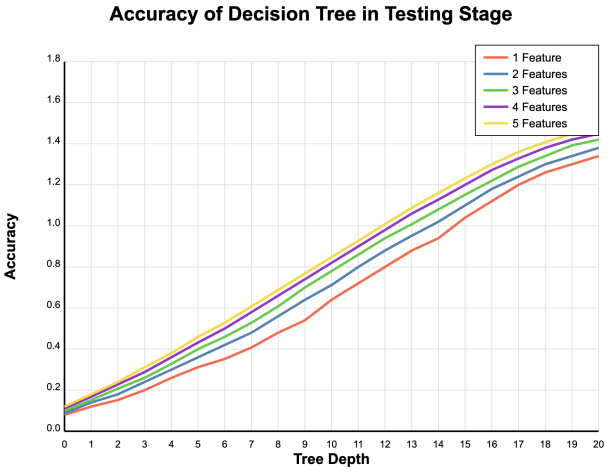

Fig. 3: Accuracy of Decision Tree in Testing Stage

4) Model Accuracy in Validation Stage:

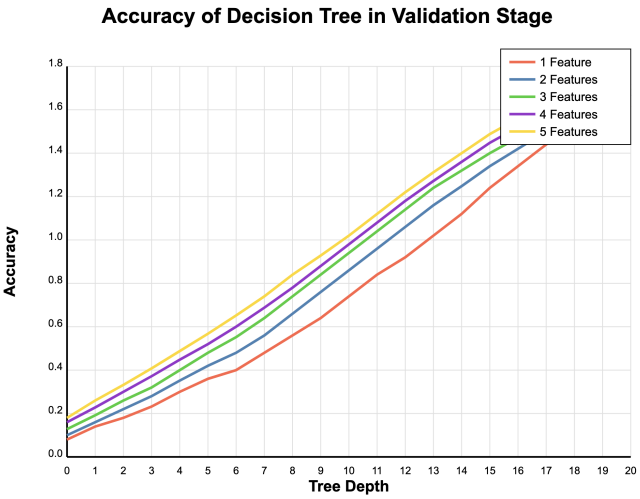

Fig. 4: Accuracy of Decision Tree in Validation Stage

5) Confusion Matrices for Different Models:

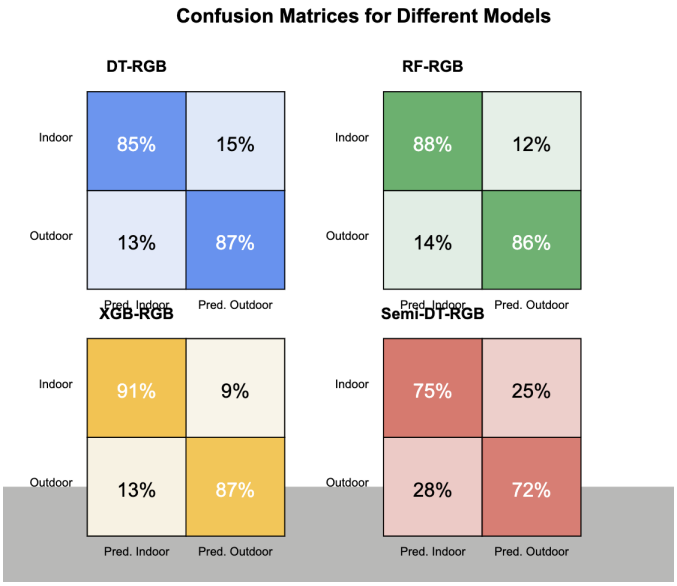

Fig. 5: Confusion Matrices for Different Models

6) Performance Comparison of Image Classification Models:

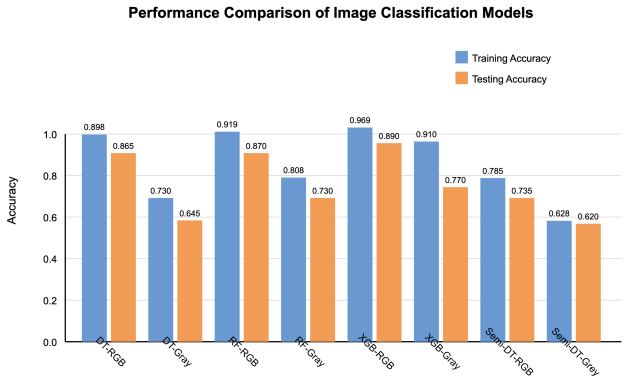

Fig. 6: Performance Comparison of Image Classification Models

7) 3-Conv2d CNN RGB:

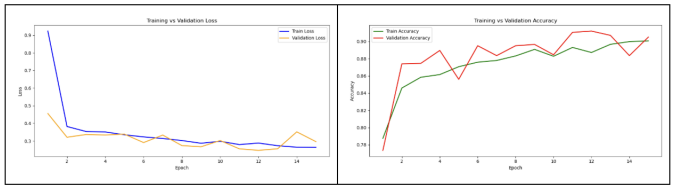

Fig. 7: 3-Conv2d CNN RGB

8) 2-Conv2d CNN RGB:

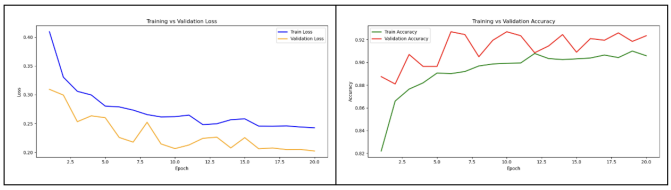

Fig. 8: 2-Conv2d CNN RGB

9) 4-Conv2d CNN RGB:

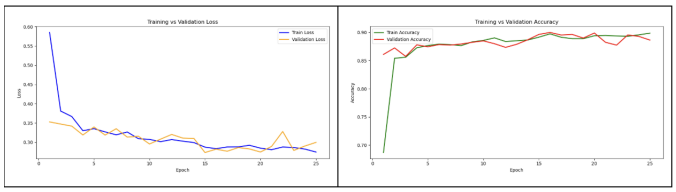

Fig. 9: 4-Conv2d CNN RGB

10) 3-Conv2d CNN Gray:

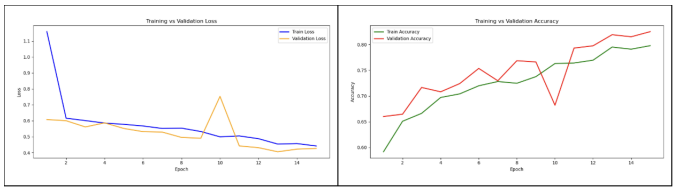

Fig. 10: 3-Conv2d CNN Gray

11) 2-Conv2d CNN Gray:

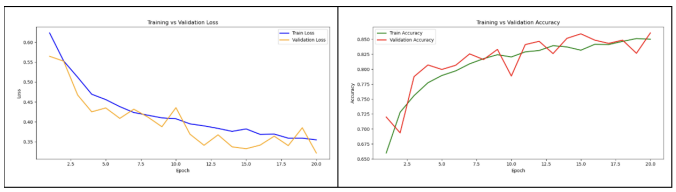

Fig. 11: 2-Conv2d CNN Gray

12) CNN2 Confusion Matrix:

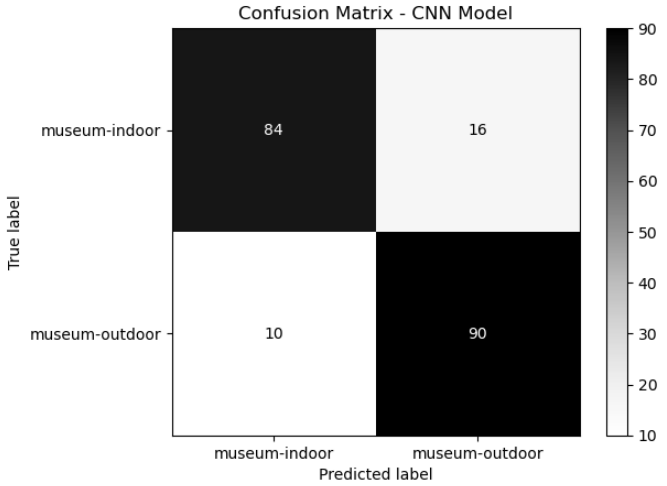

Fig. 12: 2-Conv2d CNN Gray

13) CNN3 Confusion Matrix:

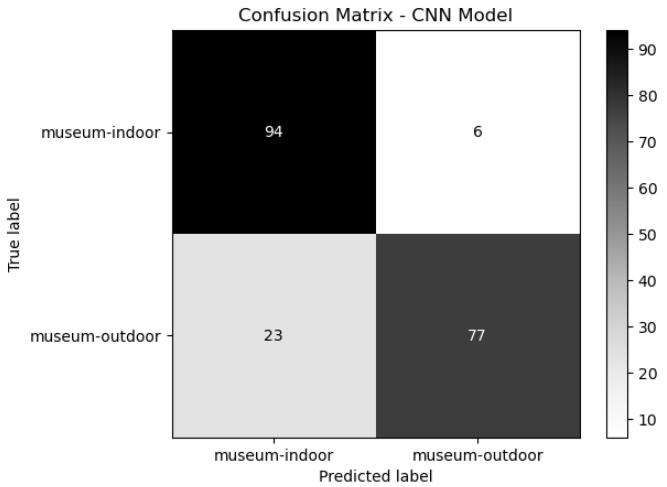

Fig. 13: 2-Conv2d CNN Gray

C. Ablative Study

CNN model with three convolutional layers has provided better accuracy in classifying the images in RGB format. In grayscale format two convolutional layer models have shown better results. Notably, the greyscale image pre-processing yields less accuracy in terms of classification when compared to the RGB images. Also, increasing the number of convolutional layers has shown gradual decrease in the accuracy of classification. This can be resolved by changing the learning rate, altering the number of epochs, by introducing further pre-processing techniques such as image rotation, image augmentation, gaussian blur are a few techniques to improve the accuracy.

VI. 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.
6) https://github.com/cs231n/cs231n.github.io/blob/b75203c38941ef7b8b897e49da6852baa3c9af84/convolutional-networks.mdL251