# ImagoAI Intern Task Report

Rakshatha Vasudev

March 2025

## 1 Introduction

Data Preprocessing and Feature Engineering The dataset is loaded from a CSV file and explored using basic statistical and structural analysis.The dataset shape is (500, 450).

## 2 Exploratory Data Analysis and cleaning

To ensure data quality and optimize model performance, missing values in numeric columns are imputed using their mean values, and outliers are detected using the Z-score method. Feature scaling techniques like StandardScaler and MinMaxScaler are applied to normalize data, improving stability for the models. Dimensionality reduction using PCA helps retain essential variance with fewer features, enhancing computational efficiency, while t-SNE is leveraged for clustering and visualization, revealing hidden patterns in high-dimensional data. Together, these preprocessing steps refine the dataset, making it more suitable for predictive modeling and insightful analysis.

## 3 Model-Convolution Neural Networks

The Conv1D layer learns spectral patterns that are important for predicting the target variable. MaxPooling ensures the model focuses on the most important features. The fully connected layers learn high-level relationships between extracted features. The model is trained using the Mean Squared Error (MSE) loss, which is suitable for regression tasks.

## 4 Model Hyperparameter Tuning and Results

- Lower MAE (Mean Absolute Error): Random Forest (0.0156) vs. CNN (0.0219) → Random Forest makes more precise predictions.

- Higher $R^2$ Score: Random Forest (0.8889) vs. CNN (0.8569) → Random Forest explains more variance in the data

Since CNN underperformed I considered hyper parameter tuning using Grid search.GridSearchCV was applied with 3-fold cross-validation, using Mean Absolute Error (MAE) as the scoring metric.

The hyperparameter tuning resulted in the optimal settings for the CNN model: With these optimal parameters, the Best CNN MAE (Mean Absolute Error) is 0.0156, and the Best CNN $R^2$ (R-squared) is 0.919. These results indicate a strong performance from the CNN model, with a relatively low error and a high $R^2$ score, suggesting that the model explains most of the variance in the data.

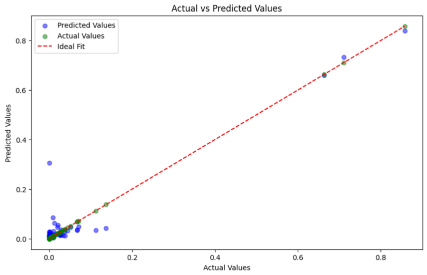

Figure 1:

- Alignment with Ideal Fit:Most points lie close to the red dashed line, indicating that the model's predictions are generally accurate.
- Outliers:A few points deviate significantly from the ideal fit, indicating potential outliers or areas where the model struggles to make accurate predictions.
- Model Performance:The clustering of points near the ideal line suggests a good overall fit.

# 5 Challenges and Tradeoffs

: Initially experimented with Feed Forward Neuaral Network(FNN) which had a very high MAE-0.91 compared to the baseline indicating poor performance.This could stem from the fact that FNNs wasn't able to handle the curse of dimensionality even after rigorous cleaning and dimensionality reduction.CNN models are better suited for Hyperspectral data as they can capture spatial and spectral patterns via convolution operations.

Choosing between PCA and t-SNE for training data:PCA was used for training because it reduces dimensionality in a way that preserves the data's variance and structure, which is crucial for training machine learning models effectively. t-SNE, on the other hand, is better suited for exploratory data analysis and visualization, but it is not appropriate for training because it distorts the data's relationships and does not preserve global information.