CS610 - APPLIED MACHINE LEARNING - G2, GROUP 6 - PROJECT REPORT

PROJECT TITLE:

ANOMALY ALZHEIMER'S DETECTION IN MRI

TEAM MEMBERS:

Wang Ruipeng - ruipengwang.2021

Wang Shengming - smwang.2021

Huang Jing - huangjing.2021

Yueyang Jiang - yyjiang.2021

CONTENT PAGE

Section 1: Background & Motivation

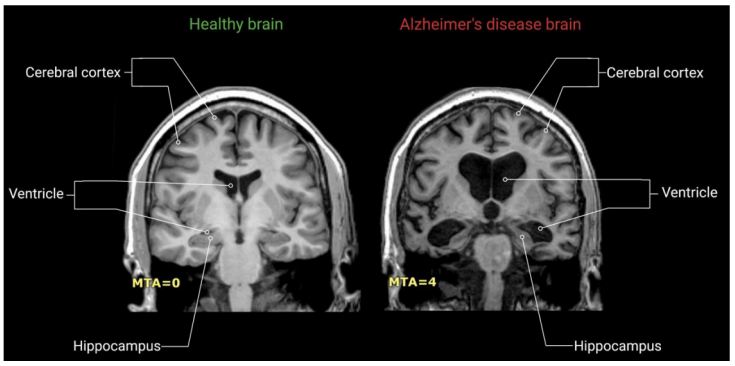
Today, AD (Alzheimer's Disease) is no longer a purely neurological disease, its social problems and potential impacts have attracted worldwide attention and raised concern about the detection and treatment for the disease.

Current diagnosis of Alzheimer's disease heavily relies on documenting mental decline, at which point, a patient already reaches moderate AD stage and the brain is severely damaged.Utilising biological markers in lumbar puncture might be recommended in current diagnosis, which indicates the presence of disease accurately and reliably to some extent. However, biomarkers are very expensive and controversial. In fact, one-time lumbar puncture can cost up to $4000.

Considering the potential high cost of the diagnosis process in the early stage, most equal or below average families will have difficulty affording it. But with the advent of the revolutionary wave of medical imaging, its accurate diagnosis and early detection of AD has brought new opportunities, making early intervention and treatment possible. Advanced imaging methods certainly broaden the AD research scope.

Doctors now can use the MRI (Magnetic Resonance Imaging) to help make diagnoses. As is shown in the graph below, there are some differences between the healthy person's and Alzheimer's patient's brain MRI image. However, sometimes the discrimination is not obvious enough to be correctly detected. In practice, the diagnostic accuracy is only about 77% for a clinical diagnosis of AD, even among experts. For less-experienced doctors, it is quite helpful to use supporting tools to improve the diagnosis accuracy of AD.

Based on the mass application of machine learning models, machine learning classification methods could be implemented to support detecting the diagnosis of AD. Creating machine learning models could help improve diagnostic efficiency and accuracy (especially for mass-population screening), collect the results for further analysis, decrease medical operation cost, and reduce the burden of insurance for both government and individual.



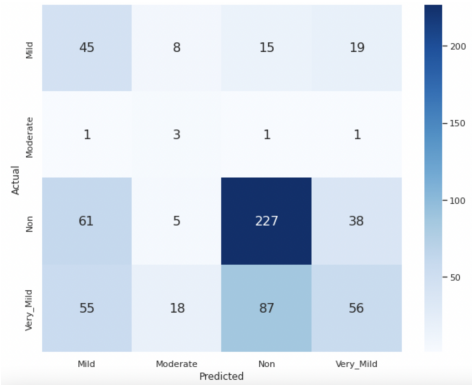
Healthy brain (left) versus AD brain (right)

Dataset

To achieve the project objective and extract the desired and useful insights, we use the datasets from kaggle. The dataset[1] has a total of 6400 images, consisting of preprocessed MRI Images. All the images are resized into 128 x 128 pixels.The Dataset has four classes (Non Demented, Very Mild Demented, Mild Demented and Moderate Demented) of images with sample size of 3200, 2240, 896, 64 respectively.

[1] https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset

Section 2: Model Evaluation Methodology



For model evaluation, confusion matrix was used to show the relationship between the predicted results and the actual classification. As the confusion matrix shown on the left, the x-axis is the predicted results of the four classes including 'Mild Demented', 'Moderate Demented', 'Non-Demented' and 'Very Mild Demented', while the y-axis is the actual case. Numbers inside such as 15 at the top of the matrix indicates 15 patients predicted as 'Non-Demented' are in fact 'Mild Demented'. Moreover, the values in the third column are emphasised as it is better to misdiagnose a healthy person than failed to diagnose a patient who has the Alzhemer's disease.

Besides, accuracy, precision and recall are calculated and compared among the models to find out a better performance. And instead of F1 score, β=2 is chosen and F2 score is calculated as recall is considered to be more important and should be given more weight in disease diagnosis. Formulas are shown below.

Precision

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ positive}$$

Recall

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Accuracy

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

F2 score

$$\text{F2 Score} = \frac{(1+2^2) * Precision * Recall}{2^2 * Precision + Recall}$$

Section 3: Solution Details & Experiments

Model 1: Naive Bayes

Naive Bayes classifier is based on Bayes' Theorem with an assumption of independence among predictors. In our project, the Naive Bayes classifier assumes that each pixel is an independent feature in a class and it is unrelated to the presence of any other pixels.

Overall, the idea of Bayes theorem is calculating posterior probability $P(C|X_1 \ldots X_{16,384})$ from $P(C)$ and $P(X_1 \ldots X_{16,384}|C)$. Each X represents an individual pixel, and the value range of X is from 0 to 1, C represents the class, which are non-demented, very mild demented, mild demented and moderate demented.

With regression models, the preprocessing is different from the deep learning models, we need to flatten the image from three dimensional to one dimensional, then we divide all the values by 255 to make the value within the range from 0 to 1. Then we manually choose 5760 samples (*90%*) as the training data and 640 samples (*10%*) as testing data.

To improve the performance of our model, we tried all the three types of Naive Bayes classifiers:

- Multinomial
- Gaussian
- Bernoulli

With all these done, we built these 3 models by our image dataset to predict the classes.

Model 2: Logistic Regression

We used multinomial logistic regression as our second model. Like the binary logistic regression, it uses maximum likelihood estimation to evaluate the probability of categorical membership.

The data preprocessing step is the same in the previous model training, and we also use the same proportion to split the train and test data.

To improve the model performance, we utilised grid search to find out the optimal combination of the hyperparameters:

- Penalty = none. (No penalty is added)
- Tol = 0.1. (Tolerance for stopping criteria)
- Solver = saga (The *SAGA* solver is a variant of SAG that also supports the non-smooth penalty L1. This is therefore the solver of choice for *sparse multinomial logistic regression* and it's also suitable for very *Large* datasets).
- Muti_class = multinomial.

After hyperparameter tuning, we fit our data into the model.

Model 3: Random Forest

For random forest, we applied the same process used in the previous two models. The grid search returns the following optimal hyperparameters:

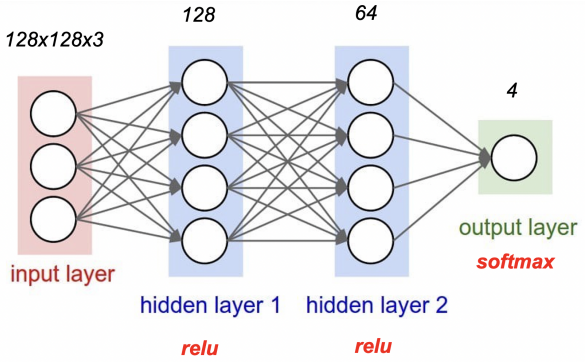
- MIn_sample_split = 3 (The minimum number of samples required to split an internal node)
- Max_depth = 9 (The maximum depth of the tree)
- Max_feature = 'sqrt' (The number of features to consider when looking for the best split)

After hyperparameter tuning, we fit our data into the model.

Model 4: Artificial Neural Network (ANN)

Before fitting into Deep Learning models, we took advantage of the existing data splitting method called split-folders. It will split image data into train, validation and test folders with the input split ratios. We splitted 80% of our data to be train set, 10% of our data to be validation set and 10% of our data to be test set.

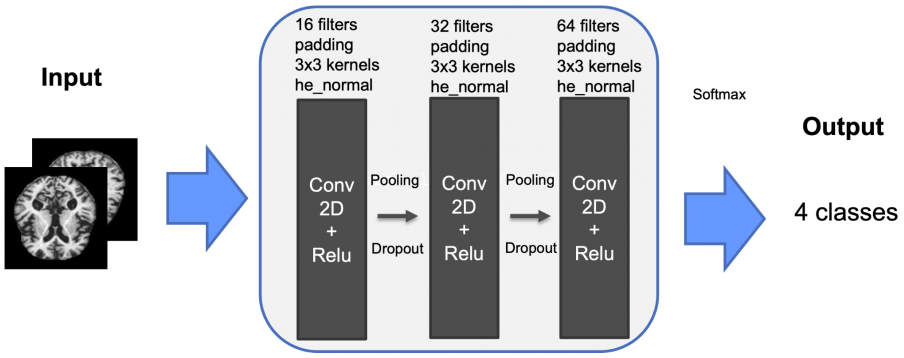
An ANN model is based on a collection of connected units or nodes called artificial neurons. The input size is (128, 128, 3), followed by two hidden layers with 128 neurons and 64 neurons respectively and Relu as their activation functions. The output layer is softmax with 4 neurons, since the output consists of 4 classes. Below is the ANN architecture:



ANN architecture

Model 5: Convolutional Neural Network (CNN)

CNN is a type of neural network model which allows us to extract high representation for the image content. We designed CNN model to have three hidden layers with Relu as the activation functions and he_normal as the weight initializer. The first layer has 16 filters, the second layer has 32 filters and the third layer has 64 filters. Each layer adds padding to the surrounding of image vectors. And there is pooling and dropout between layers. Below is the CNN architecture:

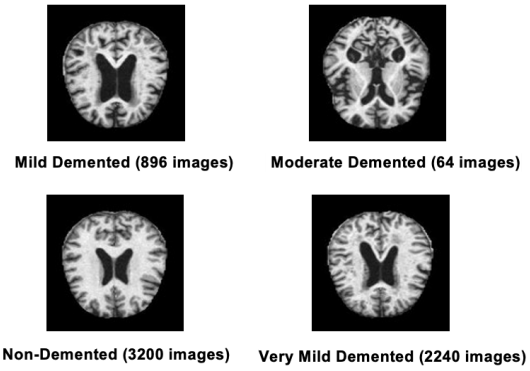


CNN architecture

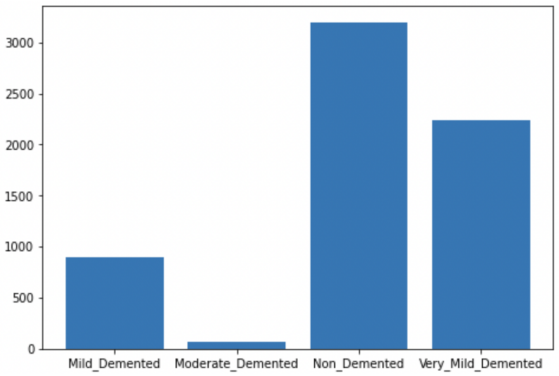
Section 4: Results & Analysis

Exploratory Data Analysis

1. Sample image for each class
2. Distribution by classes



Sample Image for Each Class



Bar chart: Class Sample Distribution

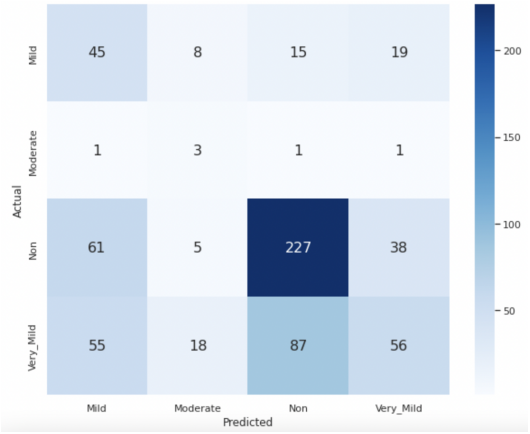
Model 1: Naive Bayes Results & Analysis

Our prediction results are shown in the table blow

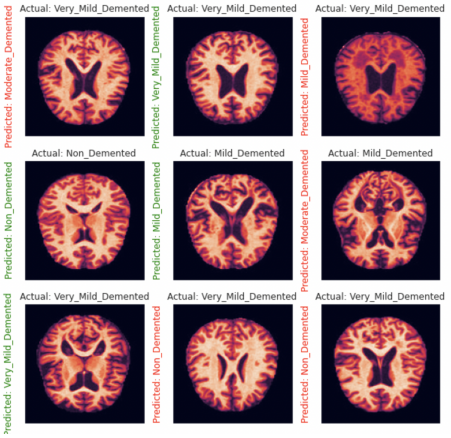
	Accuracy	Precision	Recall	F2 Score
Multinomial	0.4939	0.5223	0.4939	0.4992
Gaussian	0.5172	0.5601	0.5172	0.5252
Bernoulli	0.5328	0.4470	0.5328	0.5131

Classifier Evaluation Table

F2 score has the highest weight in our evaluation. We can see from the table that Bernoulli has both the highest accuracy and recall, but Gaussian has the highest F2 score. Hence, Gaussian has the best performance among all the classifiers in the Naive Bayes.



Confusion Matrix of Gaussian Classifier



Actual Prediction for 9 images

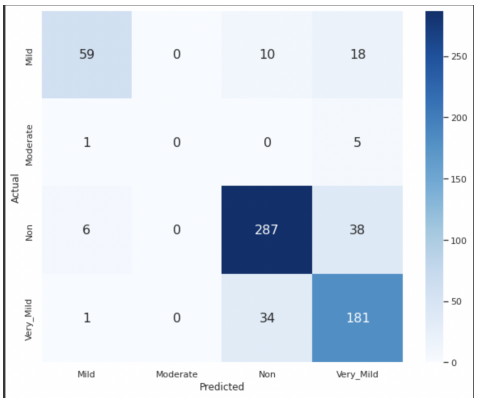
As we can see from the confusion matrix, moderate cases' recall is awful. Over 100 demented samples were marked as healthy. For the random nine-image prediction, the model only correctly predicted four images, and the accuracy was also low.

Model 2: Logistic Regression Results & Analysis

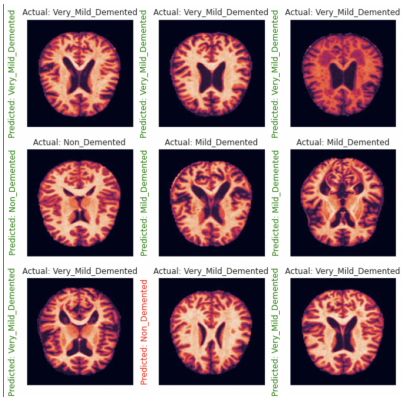
The classification report as the table shown below:

	Accuracy	Precision	Recall	F2 Score
Logistic	0.8234	0.8206	0.8234	0.8229

Classifier Evaluation Table



Confusion Matrix Heatmap of logistic regression



Actual Prediction for 9 Samples

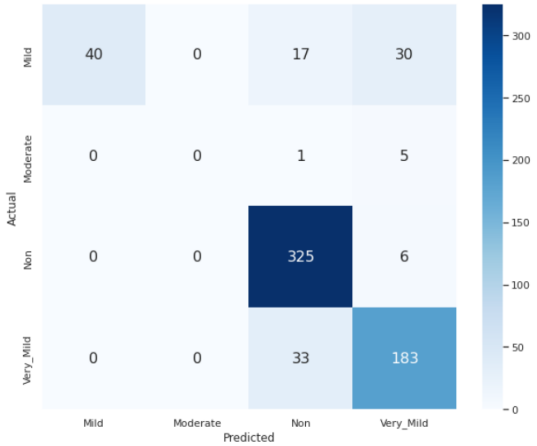
The F2 logistic regression score is *0.8229*, higher than the naive Bayes model *of 0.3*. The accuracy also reaches *0.82*. The performance of logistic regression was out of our expectation, it only predicted *44* cases of the demented patient to non-demented, but there still got room for improvement.

Model 3: Random Forest Results & Analysis

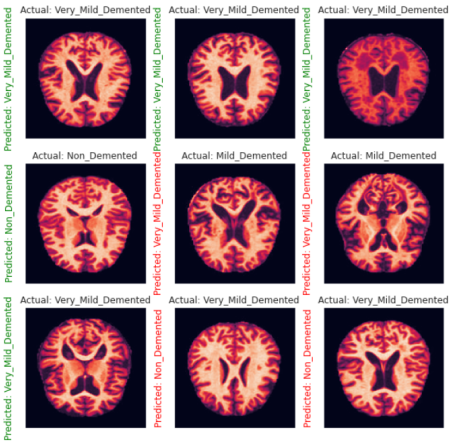
The result of random forest is shown below:

	Accuracy	Precision	Recall	F2 Score
Random Forest	0.8562	0.8587	0.8562	0.8567

Classifier Evaluation Table



Confusion Matrix Heatmap of RF

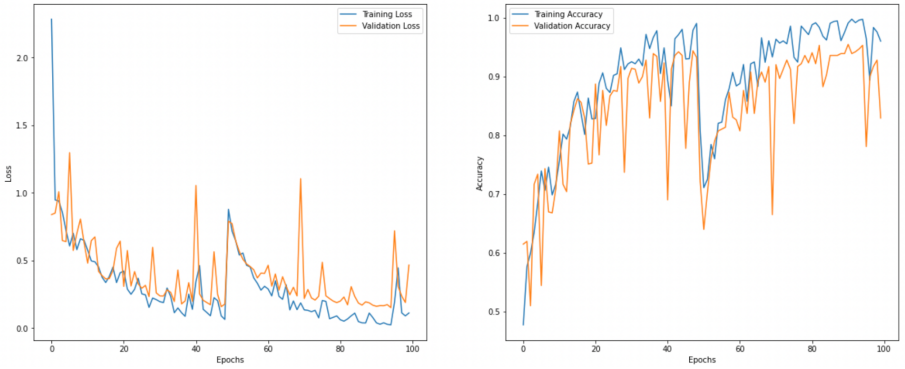


Actual Prediction for 9 Samples

The accuracy of random forest is 0.86, which is higher than Naive Bayes model and logistic regression model, and F2 score also reaches 0.86. It seems that the performance of random forest seems to be satisfying, but it would be better to improve the accuracy and F2 score above 90% so as to provide more precise diagnosis.

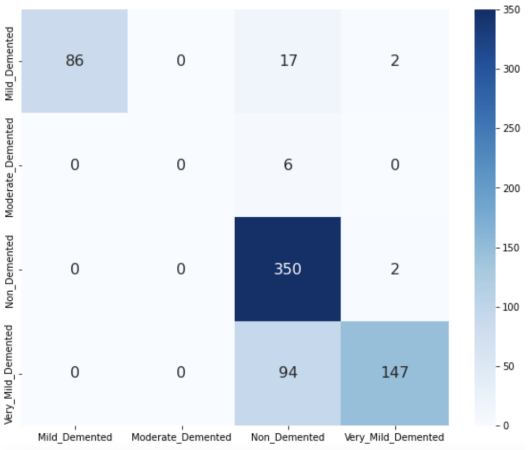
Model 4: Artificial Neural Network Results & Analysis

We Compiled ANN model with Adam optimizer, sparse categorical cross entropy as the loss function and Accuracy as the evaluation metric. We trained with 100 epochs and the accuracy on test sets is 81.78%. The F2 score is 61.06%. Below are the training/testing accuracy and loss curves:

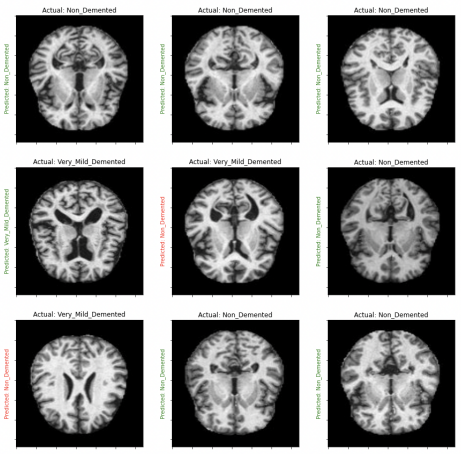


ANN loss and accuracy curve

The overall trend is decreasing and increasing for loss on training set and accuracy on test set, but there are major fluctuations when reaching around 50 epochs.
The corresponding heat map is below:



ANN heatmap

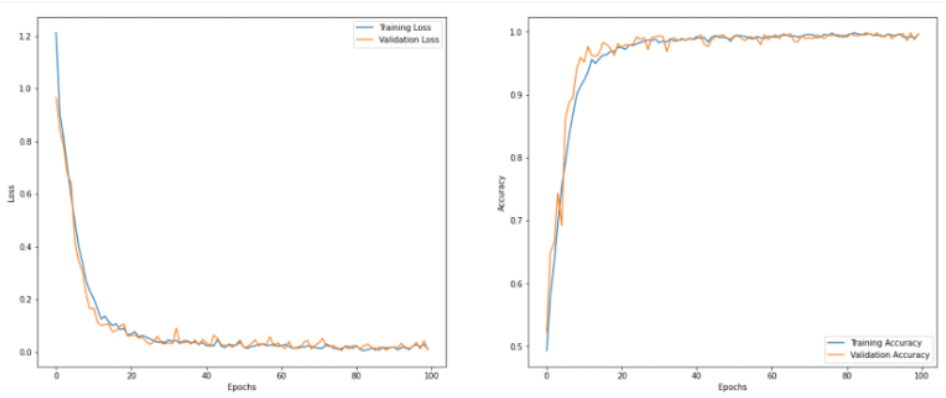


Actual Prediction for 9 Samples

This heatmap indicates that ANN model cannot recognize demented-moderate case. And there are 94 very-mild demented cases being misclassified as non-demented. Then we compared 9 random samples with model prediction and found 2 mismatched images out of 9 images.

Model 5: Convolutional Neural Network(CNN) Results & Analysis

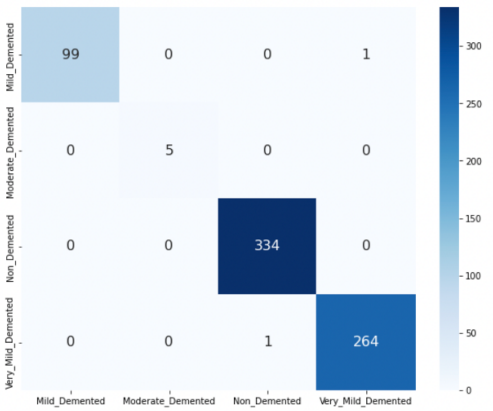
We Compiled CNN model with Adam optimizer, sparse categorical cross entropy as the loss function and Accuracy as the evaluation metric. We trained with 100 epochs and the accuracy on test sets is 99.53%. The F2 score is 99.69%. Below are the training/testing accuracy and loss curves:



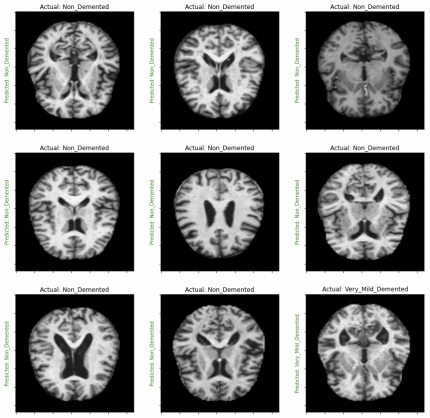
ANN loss and accuracy curve

The overall trend is stably decreasing and increasing for loss on training set and accuracy on test set.

The corresponding heat map is below:



CNN heatmap



Actual Prediction for 9 Samples

This heatmap indicates that there is only one case of very-mild demented being misclassified as non-demented. Then we compared 9 random samples with model prediction and found all images match correctly.

Conclusion

According to the model results above, Deep Learning models have generally better performance, especially CNN. CNN has nearly perfect accuracy scores and recall is close to 100%. Therefore, compared to traditional ways, which means letting experts diagnose, it is time for Machine Learning Models to shine here.

Section 5: Retrospectives - Gap Analysis, Future Work

A. Gap Analysis

Our dataset is imbalanced. When we train the models and do testing, there are too few moderate demented samples, far less than non-demented samples. Therefore, the accuracy of our prediction will be affected.

B. Future Work

In the future we need to collect more case images as new input to validate the reliability of our models. And we could try more image classification models such as ImageNet etc.

Section 6: Appendix

A. Reference

1. Xia, F., Chen, J., Fung, W. K., & Li, H. (2013). A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. *Biometrics*, *69*(4), 1053–1063. https://doi.org/10.1111/biom.12079

2. *sklearn.ensemble.RandomForestClassifier*. (2022, January 1). Scikit-Learn. Retrieved July 4, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

3. *Logistic regression python solvers' definitions*. (2016, July 28). Stack Overflow. Retrieved July 4, 2022, from https://stackoverflow.com/questions/38640109/logistic-regression-python-solvers-definitions

4. *A Review Paper on Artificial Neural Network: A prediction technique - IJSER*. (n.d.). Retrieved July 5, 2022, from https://www.ijser.org/researchpaper/A-Review-paper-on-Artificial-Neural-Network--A-Prediction-Technique.pdf

5. Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021, March 31). *Review of Deep Learning: Concepts, CNN Architectures, challenges, applications, future directions - Journal of Big Data*. SpringerOpen. Retrieved July 5, 2022, from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8