Machine Learning & Neural Networks

Project: Alzheimer's MRI images – Extracting numerical and statistical features for classification

Professor:

Ossnat Bar-Shira

Students:

Aracely Gutiérrez Lomelí G37064824

Tali Rozenson 208160937

Kanykei Mairambekova AC3188924

January, 11th, 2024

Aracely Gutiérrez Lomelí G37064824
Tali Rozenson 208160937
Kanykei Mairambekova AC3188924

The Leslie and Susan Gonda Multidisciplinary Brain Research Center
Bar-Ilan University

i. Summary

The present study aimed to develop a machine learning tool for the classification of Alzheimer's Disease into affected and unaffected patients based on MRI scan images. The dataset is composed of 6400 images from MRI studies. The resources used were Kaggle for dataset acquisition, Python for data preprocessing, model training, and evaluation.

To classify Alzheimer's disease, we implemented three different methods. The first method aimed at applying classification models on the statistical features extracted from the images. The second method applied 2-D Fourier transform (FFT2) and extracted additional frequency values to the statistical values in the first method. Lastly, the third method used Convolutional Neural Network (CNN) directly on the images.

After that, we experimented with a range of binary classification models, including Random Forest Classifier, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes Classifier, Decision Tree Classifier and Convolutional Neural Network (CNN). We aimed to compare their performances and identify the most suitable model for our task. Additionally, we experimented with hyperparameter tuning to optimize model performance. Figure 1 shows the general diagram of the methodologies used in the project.

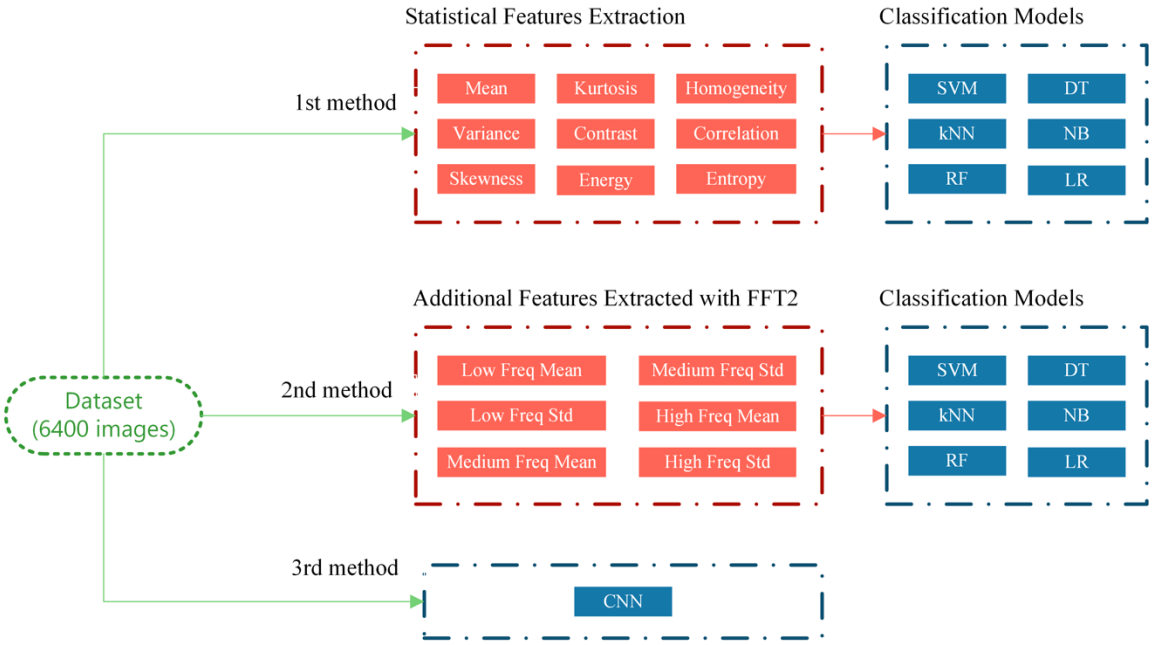


Figure 1. Methodologies used in the project.

The accuracies for the models assessed on the test set were for the first method: 73% in Random Forest, 75% in SVM, 71% in Logistic Regression, 69% in Naive Bayes, 69% in Decision Tree; for the second method: 74% in Random Forest, 81% in SVM, 73% in Logistic Regression, 69% in Naive Bayes, 69% in Decision Tree; and lastly 99.53% in CNN. The latter was chosen as the best method to classify the dataset.

ii. Introduction

Alzheimer's disease (AD) is a devastating condition that affects millions of people worldwide. Early and accurate diagnosis is crucial for effective management and treatment. This is a neurodegenerative disease causing around 60-80% of all dementias (Kasáč, *et al.*, 2024). Examination of AD has widely been done with MRI (Magnetic Resonance Imaging) because it can give detailed graphical information about the volume of brain structures, characterized in healthy and demented subjects. For this reason, it is a recommended technique for AD diagnosis (Myers, 2022).

Our dataset (Kumar & Shastri, 2022) consists of 6400 MRI images with a size of 128x128 pixels. These images were collected from different websites, hospitals, and public repositories consisting of four classes: mild demented (896 images), moderate demented (64 images), non-demented (3200 images), and very mild demented (2240 images). The number of samples within the classes makes the dataset imbalanced, requiring some techniques further explained to regularize the data.

iii. Methods

a. Statistical Features Extraction

In our project, the first method of identifying Alzheimer's disease using MRI scan images entailed applying different binary classification models to the statistical features that were taken out of the images. Statistical features such as contrast, correlation, energy, entropy, homogeneity, kurtosis, mean, skewness, and variance were extracted into a tabular Excel file for red color channel. We focused only on the one channel as the values for all color channels were the same. To guarantee uniformity and equal scaling, the extracted features underwent normalization. After, the classification models received the extracted features as input. We also used GridSearch for the hyperparameter tuning iterating over a range of potential values to optimize each model's performance. Lastly, measures including accuracy, precision, recall, and F1 score were used to assess each model's performance.

b. Additional Frequency Features Extraction using FFT2

Building on the initial approach, the second method extended the feature set by including frequency domain features extracted using the 2-D Fourier Transform (FFT2). The rationale behind this method was to capture not just the spatial but also the frequency characteristics of the MRI images, which may provide additional discriminative information useful for classification.

The FFT2 decomposes an image into its sine and cosine components, revealing the spectrum of frequencies present in the image. After applying the FFT2 to each image, we split the frequency components based on their distance from the center of the frequency matrix (which corresponds to the low-frequency region). The center of the frequency matrix (65, 65) was considered the point of reference, as the image size is (128, 128). A pixel distance of less than 4 units from the center was considered low frequency. A pixel distance between 4 and 9 units from the center was considered medium frequency. A pixel distance less than 35 units from the center (and thus more than 9 units) was considered high frequency.

After that, we found mean and standard deviation for each frequency group and combined these new frequency domain features with the previously extracted statistical features. This hybrid feature set aimed to leverage the strength of both spatial and frequency domain information. Using this enhanced feature set, we retrained the same set of binary classification models.

Moreover, upon analyzing the MRI images for Alzheimer's Disease classification, we made a strategic decision to focus the classification task on the most distinct classes: 'Non-Demented' versus 'Moderate Demented,' excluding the 'Mild Demented' category. This decision was informed by the premise that 'Mild Demented' cases might present subtler and more ambiguous features, potentially introducing noise and complexity into the model training process. The sample size for this approach was 128 (64 for each group).

c. Convolutional Neural Network (CNN)

The next figure summarizes the trials with the CNN model.

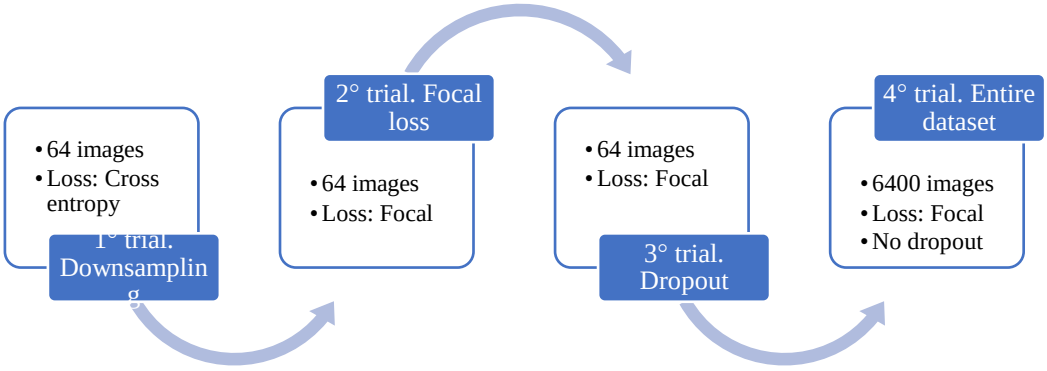


Figure 2. CNN trials

For the convolutional neural network, it was opted to use a drive folder of the dataset creating a train and test set in a ratio of 80% and 20%, respectively. Since the classes are imbalanced in the number of samples, it was decided to downsample every label in the first trial into the minimum amount of images in any of the classes, specifically 64 images for the moderate demented label. Cross entropy was applied since its recommended use for multiclass classification. This function is combined with softmax which gives a probability distribution of the classes while the cross-entropy function is the loss function measuring how well the guess resembles the true answer (Charan, 2023).

Later on, for the second trial, focal loss was used because it is said to encourage the model to learn on misclassified samples and reduce the impact of the overrepresentation of a certain class (Niyaz, 2023). For the third attempt, and with the aim of results comparison, a regularization technique was used to avoid overfitting. This technique disregards some nodes to ensure there is no codependence between units (Marimuthu, 2023). Finally, the fourth trial aimed at analyzing the possibility of classifying the entire dataset using focal loss that considers the imbalances within the classes. In this trial, the model was also assessed in the evaluation set.

iv. Results

a. Statistical Features Extraction

The results of the first method indicated that among the tested models, the Support Vector Machine (SVM) achieved the highest accuracy at 75%, with a precision of 0.71, recall of 0.82, and an F1 score of 0.76. Close contenders included the Random Forest Classifier with an accuracy of 73%, and K-Nearest Neighbors (KNN) with an accuracy of 74%. The other models demonstrated lower performance, with accuracies ranging from 69% to 71%. Table 1 shows the performance of each model on the test set along with its best hyperparameters.

Table 1. Performance and best hyperparameters of each algorithm for the first method

Model	Optimized Hyperparameters	Accuracy	Precision	Recall	F1 Score
SVC	{'C': 100, 'gamma': 'scale', 'kernel': 'rbf'}	75%	71%	82%	76%
KNN	{'leaf_size': 10, 'n_neighbors': 3}	74%	72%	75%	73%
NB	{'priors': None, 'var_smoothing': 1e-09}	69%	69%	67%	68%
LR	{'C': 0.1, 'max_iter': 1000, 'penalty': 'l2', 'tol': 0.0001}	71%	68%	76%	72%
DT	{'max_depth': 2, 'min_samples_leaf': 1, 'min_samples_split': 4, 'min_weight_fraction_leaf': 0.1}	69%	66%	74%	70%
RF	{'criterion': 'gini', 'max_depth': 8, 'max_features': 'sqrt', 'n_estimators': 200}	73%	69%	80%	74%

b. Additional Frequency Features Extraction using FFT2

The incorporation of FFT2 features had a profound impact on model performance. Remarkably, both the Support Vector Machine (SVM) and Random Forest Classifier achieved perfect scores with an accuracy, precision, recall, and F1 score of 1.0. The KNN and Logistic Regression models also exhibited excellent performance with accuracies of 0.97. Naive Bayes and Decision Tree classifiers had relatively lower metrics but showed improvements from the first method. Table 2 shows the performance of each model on the test set along with its optimized hyperparameters.

Table 2. Performance and best hyperparameters of each algorithm for the second method

Model	Optimized Hyperparameters	Accuracy	Precision	Recall	F1 Score
SVC	{'C': 100, 'gamma': 'auto', 'kernel': 'rbf'}	81%	78%	86%	82%
KNN	{'leaf_size': 10, 'n_neighbors': 3}	83%	82%	83%	83%
NB	{'priors': None, 'var_smoothing': 1e-09}	69%	68%	67%	68%
LR	{'C': 1000.0, 'max_iter': 1000, 'penalty': 'l2', 'tol': 0.1}	73%	69%	77%	73%
DT	{'max_depth': 2, 'min_samples_leaf': 1, 'min_samples_split': 4, 'min_weight_fraction_leaf': 0.1}	69%	66%	74%	70%
RF	{'criterion': 'gini', 'max_depth': 8, 'max_features': 'sqrt', 'n_estimators': 500}	74%	70%	81%	75%

The binary classification approach, focusing on the two opposite conditions, led to a significant improvement in model performance. Notably, the Support Vector Machine (SVM) and Random Forest Classifier achieved perfect scores across all metrics with an accuracy, precision, recall, and F1 score of 1.0. The KNN and Logistic Regression models also performed exceptionally well, with accuracies of 0.97. The Gaussian Naive Bayes and Decision Tree classifiers showed improvements, with accuracies of 0.88. Table 3 shows the performance of each model on the test set along with its optimized hyperparameters.

Table 3. Performance and best hyperparameters of each algorithm for the second method (non demented vs moderate demented)

Model	Optimized Hyperparameters	Accuracy	Precision	Recall	F1 Score
SVC	{'C': 50, 'gamma': 'scale', 'kernel': 'rbf'}	100%	100%	100%	100%
KNN	{'leaf_size': 10, 'n_neighbors': 5}	97%	94%	100%	97%
NB	{'priors': None, 'var_smoothing': 1e-09}	88%	87%	87%	87%
LR	{'C': 1000.0, 'max_iter': 1000, 'penalty': 'l2', 'tol': 0.1}	97%	100%	93%	97%
DT	{'max_depth': 2, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.1}	88%	82%	93%	88%
RF	{'criterion': 'gini', 'max_depth': 5, 'max_features': 'sqrt', 'n_estimators': 500}	100%	100%	100%	100%

c. Convolutional Neural Network

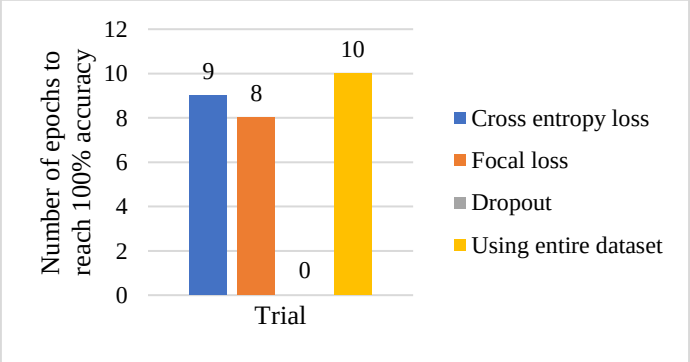
By testing and analyzing different learning rates and their respective accuracies, it was found that 0.001 gave the best result. This learning rate was set within an ADAM optimizer —the latter because of its consideration of current and previous epochs for calculation on the moving average of gradients (Singh, 2021)—.

The batch size was calculated based on the next equation, and since the first trials only included 13 images for the test set, it was only able to do 25 gradient steps for the train while in the entire dataset, 80 gradient steps were used, and this number suited well because the model converged on the 10th epoch.

Equation 1. Batch size defined by gradient (update) steps

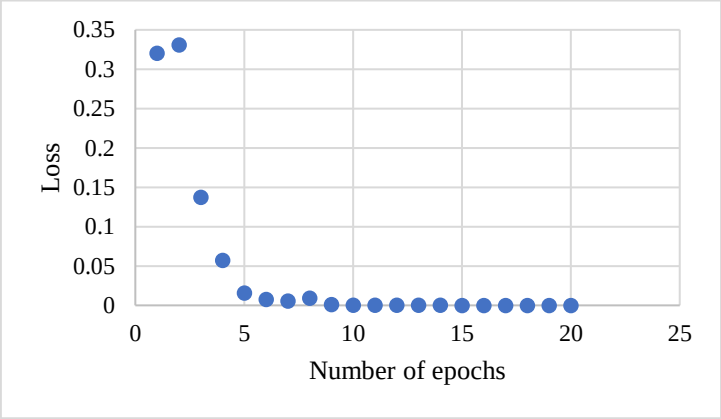
$$Effective\ batch\ size = \frac{Total\ samples}{Update\ steps}$$

The first 3 trials involved using 64 images to select the best method with the highest accuracy and interpolate it into the 6,400 images of the entire dataset. In the next graph for the training sets, we see that the dropout method never obtained 100% accuracy. In fact, the dropout method had the lowest accuracy, just getting 29.9%. In the same graph, we see that focal loss had the lowest number of epochs to reach 100% accuracy compared to cross-entropy, the reason why the use of the entire dataset considered focal loss.



Graph 1. Number of epochs to reach 100% accuracy in the training sets

In the fourth trial, where we interpolated the focal loss into the entire dataset (check Graph 2 for loss tracking, where the loss reaches 0 in the 15th epoch for the training set), we obtained an accuracy of 99.53% for the test set.



Graph 2. Track of loss per epoch in training set

v. Discussion

a. Statistical Feature Extraction

SVM was the best model, according to the study of the first method's results, for identifying Alzheimer's disease based on statistical features from MRI images. Its higher recall suggests a good potential to identify impacted individuals properly, which is crucial in a medical diagnostic environment. Though SVM led in performance, the accuracy difference between the models was negligible. This implies that, given the derived statistical data, there can be a ceiling effect on how well these models can perform. The relatively moderate performance could also indicate that the statistical features extracted might not capture all the necessary information to distinguish between affected and unaffected patients with higher accuracy. The subsequent methods in this study, involving Fourier transforms and convolutional neural networks (CNN), led to improved classification performance.

b. Additional Frequency Features Extraction using FFT2

The results from the second method suggest that the inclusion of frequency domain features via FFT2 provided the models with a more comprehensive understanding of the images, resulting in significantly improved classification accuracy. The models' ability to near-perfectly distinguish between 'Non-Demented' and 'Moderate Demented' cases suggests that the features, particularly those derived from the frequency domain, were highly effective in capturing the most salient characteristics necessary for accurate classification. The exclusion of 'Mild Demented' and 'Very Mild Demented' cases likely reduced the overlap between classes and allowed the models to learn more definitive patterns associated with each condition. These findings underscore the potential utility of targeted feature extraction and selective classification in medical diagnostics, where the ability to distinguish between more clearly defined conditions can lead to more reliable and interpretable models. The approach of excluding less distinct classes could be beneficial in various applications where the distinction between conditions is crucial for diagnosis or treatment planning.

Overall, the strategic focus on classifying only the most contrasting conditions in the dataset yielded a highly effective model, demonstrating the importance of thoughtful class selection in machine learning for medical imaging. The perfect scores achieved by SVM and Random Forest indicate that the models

could effectively capture the patterns associated with Alzheimer's Disease with the enhanced feature set.

c. Convolutional Neural Network

In the previous section, we saw that focal loss converged to 100% accuracy sooner than any other model. We can attribute this result to the fact that this method is commonly preferable for imbalanced datasets, such as the one used in this study. Nonetheless, cross-entropy still gave satisfactory results with 100% accuracy on the training set for 64 images.

On the other side, the fact that the dropout method had overly poor results on the 64 images set might be due to the small samples introduced to the model that can cause difficulties in learning. This might not seem a problem for the model without dropout because all the neurons of the CNN are used to learn forcing the model to converge. However, in the dropout method, the fact of removing some neurons with few samples on the set leads to poor performance of the model because the tools the model needs to converge are being removed, and instead of preventing overfitting in such a small set, is just finding it extremely difficult to converge.

vi. Conclusions

In conclusion, this study successfully demonstrated the application of machine learning techniques to classify Alzheimer's Disease from MRI scan images. The project explored three distinct methods, with the first two relying on statistical and frequency domain features and the third employing a Convolutional Neural Network. Briefly, each method outperformed the previous one. It was found that focusing the classification on the most distinct classes, 'Non-Demented' vs. 'Moderate Demented', significantly improved model performance, indicating the potential of targeted feature extraction and selective classification in medical diagnostics. Moreover, the CNN model showed exceptional promise with near-perfect accuracy, emphasizing the importance of leveraging advanced neural networks for complex classification tasks in medical imaging.

The limitations of this study primarily revolve around the dataset and model generalizability. The models were trained and validated on a specific set of MRI images, and their performance on external datasets remains untested. This raises questions about how well the models would generalize to data from different demographics, imaging protocols, and equipment. Another limitation is the binary classification approach that excluded 'Mild Demented' cases. While this strategy improved model clarity and performance, it also reduced the clinical applicability of the models since 'Mild Demented' cases are a significant part of the disease spectrum.

For further work, it would be recommended to validate the models on a more diverse and extensive dataset that includes varying stages of Alzheimer's Disease. This could help in understanding the models' robustness and reliability. Additionally, incorporating multimodal data such as patient demographics, genetic information, and cognitive test results could improve the models' diagnostic capabilities. Exploring transfer learning and domain adaptation techniques could also aid in overcoming the generalization challenges and enable the models to adjust to new datasets with minimal retraining.

vii.References

Charan, H. (2023). *Softmax and cross-entropy for multi-class classification*. Medium. https://charanhu.medium.com/softmax-and-cross-entropy-for-multi-class-classification-c9847690f71b

Kasáč, G., Bell, D. *et al.*, (2024). *Alzheimer disease*. https://radiopaedia.org/articles/alzheimer-disease-1

Kumar, S. & Shastri, S. (2022). *Alzheimer MRI preprocessed dataset*. Kaggle. https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset/data

Marimuthu, P. (2023). *Dropout regularization in deep learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/08/dropout-regularization-in-deep-learning/#:~:text=Conclusion-,What%27s%20Dropout%3F,are%20codependent%20with%20one%20another.

Myers, M. (2022). *Single brain scan can diagnose Alzheimer's disease*. https://www.imperial.ac.uk/news/237494/single-brain-scan-diagnose-alzheimers-disease/

Niyaz, U. (2023). *Focal loss for handling the issue of class imbalance*. Medium. https://medium.com/data-science-ecom-express/focal-loss-for-handling-the-issue-of-class-imbalance-be7addebd856#:~:text=By%20assigning%20higher%20weights%20to%20underrepresented%20classes%2C%20focal%20loss%20encourages,in%20the%20case%20of%20logistics.

Singh, M. (2021). *Multiclass classification neural network using Adam optimizer*. Medium. https://towardsdatascience.com/multiclass-classification-neural-network-using-adam-optimizer-fb9a4d2f73f4