Classification of Patients with the Development of Alzheimer's Disease using an Ensemble of Machine Learning Models

Mariia Nykoniuk[a], Nataliia Melnykova[b], Yurii Patereha[c], Dariusz Sala[d], Dariusz Cichoń[e]

[a,b,c]*Lviv Polytechnic National University, Stepan Bandera 12, Lviv, 79013, Ukraine*
[d,e] *AGH University of Krakow, al. Adama Mickiewicza 30, 30-059 Kraków, Poland*

Abstract

Every year, the number of diagnosed cases of Alzheimer's disease (AD) continues to grow. Dementia affects memory, orientation, language, learning ability, and the ability to perform daily activities. It is very important to correctly diagnose the stage of Alzheimer's disease, as each stage requires different treatment and support strategies for the person and their caregivers. Machine learning (ML) methods have been shown to be effective in the classification of AD patients based on medical images, such as magnetic resonance imaging (MRI). However, individual ML models often have limited performance due to overfitting or the inability to capture all of the complex patterns in the data. In this study, an ensemble of ML models is proposed to improve the classification of patients with the development of AD. The ensemble model combines the predictions of multiple individual ML models, such as Random Forest, Multi-Layer Perceptron and SVM, to produce a more accurate and robust prediction. The ensemble model achieved an accuracy of 96% in classifying patients into five stages of AD: cognitively normal, early mild cognitive impairment, late mild cognitive impairment, mild cognitive impairment, and Alzheimer's dementia.

Keywords [1]

Classification, Alzheimer's disease, magnetic resonance imaging, MRI, machine learning

1. Introduction

Alzheimer's disease is a progressive neurodegenerative disorder that affects the brain, especially in areas related to memory, thinking, and behavior. According to the World Health Organization, by 2030, the number of dementia patients will increase by 40% to 78 million [1]. Currently, there is no cure for Alzheimer's disease, but modern rehabilitation and therapy methods can slow down the rate of disease development. There are often three stages of Alzheimer's disease, but experts also use a more detailed scheme that breaks down the stages into five categories: preclinical, mild cognitive impairment, mild dementia, moderate dementia, and severe dementia. Traditional methods for diagnosing AD, such as cognitive tests and neuropsychological assessments, are subjective and can be time-consuming. Medical imaging techniques, such as magnetic resonance imaging (MRI), can provide objective and quantitative data on the structural and functional changes in the brain that occur in AD. Alzheimer's disease is characterized by the loss of brain tissue and the formation of abnormal protein deposits that can be visualized using MRI. The classification of Alzheimer's disease stages using MRI is relevant for early detection, which allows for improvement of patients' condition through early intervention. In addition, the correct determination of the stage allows for the planning of high-quality treatment, as it directly depends on the stage of the disease.

The contribution of this work is highlighted below:

DDM'2023: 6th International Conference on Informatics & Data-Driven Medicine, November 17 - 19, 2023, Bratislava, Slovakia
EMAIL: mariia.nykoniuk.mknssh.2023@lpnu.ua (A.1); nataliia.i.melnykova@lpnu.ua (A.2); yurii.i.patereha@lpnu.ua (A.3); dsala@zarz.agh.edu.pl (A.4); dcichon@agh.edu.pl (A.5)
ORCID: 0000-0003-3521-1740 (A.1); 0000-0002-2114-3436 (A.2); 0009-0002-5110-008X (A.3); 0000-0003-1246-2045 (A.4); 0000-0003-4198-1530 (A.5)

- The development of an ensemble of Random Forest, Multi-Layer Perceptron and SVM for AD classification;
- ML model evaluation using multiple performance measures such as Accuracy, Recall, Precision and F1-score;
- Multiclass classification by predicting 5 classes, namely CN, EMCI, LMCI, MCI and AD.

2. Methods and tools

2.1. Dataset

The dataset used in this paper is from Kaggle [2]. This dataset contains MRI images of patients' brains divided into five stages of Alzheimer's disease. This set consists of 2953 JPEG files with a size of 208x176. The images are already divided into training and test images, but for convenient analysis and processing, it is worth combining them into one set. Examples of images of each stage are shown in Figure 1.

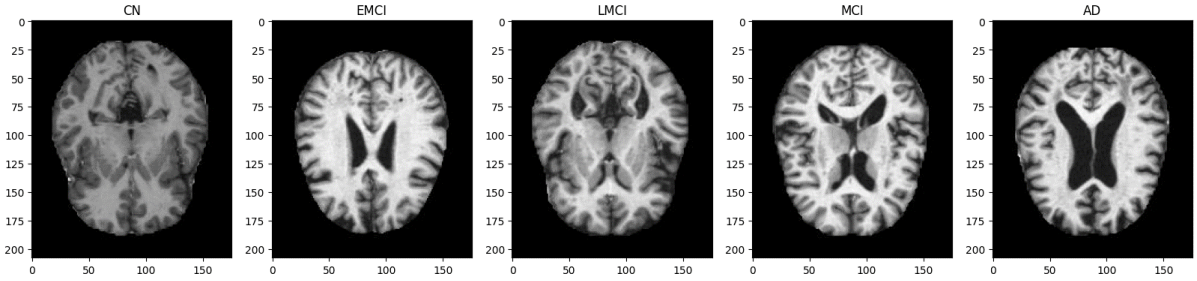


Figure 1. Example of images of each stage of the disease in the set

As you can see, the images have already removed non-brain tissue to focus only on brain structures.

2.1.1. Dataset preprocessing

To process the dataset, the steps shown in Figure 2 are performed. This consists of the steps of checking for duplicates, removing duplicates, balancing the dataset, and normalizing the images.

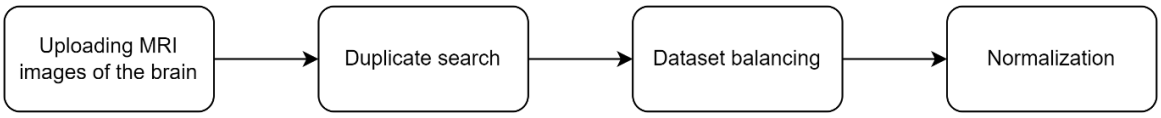


Figure 2. Stages of preprocessing an MRI image dataset

First of all, we have checked for duplicates in this set. For this purpose, we used the SHA-256 algorithm. This algorithm is one of the most popular hashing algorithms used to find duplicate images. One of the main advantages of SHA-256 is that it is more secure and reliable because it generates a longer hash code (256 bits) than MD5 (128 bits). This means that the probability of a collision when generating a hash code for different data is very low. Thus, the chances that two different photos have the same SHA-256 hash code are very small. So, all detected duplicates have been removed and are not used for research.

In addition, the study of the dataset revealed data imbalances. The problem with an unbalanced image set is that the number of images in different classes is unequal. This can lead to the learning model paying more attention to the categories with more images, thereby reducing the classification accuracy for the less represented categories. To balance the dataset, artificial images were added to the less represented classes. This is done by augmenting the data by displaying the images relative to the vertical axis and rotating them by a certain angle.

The next step in data processing is image normalization. MRI brain images of the five stages of Alzheimer's disease are normalized to ensure standardization of the brightness scale. This is done to better control the effects of different light sources and other differences between images. Normalization involves converting pixel values to a range from 0 to 1 using a standard formula:

$$normalized\_X = \frac{X - \min(X)}{\max(X) - \min(X)}, \quad (1)$$

where X is the original image, min(X) is the minimum pixel value of the image, max(X) is the maximum pixel value of the image.

For further work, the data was divided into a training, validation, and test sample, keeping the balance of classes in each set at 60:20:20.

2.2. Models used for classification

Image classification is the process of assigning each image to a specific class based on its properties and characteristics. This task is usually solved using machine learning methods that provide class identification using patterns and algorithms. The task of medical image classification is to automatically identify the characteristics and features of medical images and classify them based on these features. For the study, four classification methods were selected, including Random Forest, Decision Tree, SVM, Multi-Layer Perceptron.

2.2.1. SVM

One of the most popular machine learning methods for image classification is the support vector machine (SVM). The basic idea of the method is that the image data is first converted into vector forms and then used to train the SVM. Each image vector becomes a separate training example, and the classification is based on the position of the vector relative to the hyperplane that separates the different classes.

When training the model, a kernel is first selected that determines how the classes are separated in space. Next, the SVM searches for the optimal hyperplane that maximizes the separation of points from different classes and minimizes classification errors. This process is called hyperplane optimization [3]. Once trained, the SVM can classify new images by converting them to vector form and determining their position relative to the hyperplane built during training. The distance from the hyperplane to each image vector allows the SVM to determine the probability that the vector belongs to a class. SVM usually performs well with small data and provides good resolution, so it may be suitable in this study. The disadvantage of SVM is that it can become computationally demanding when the amount of data increases [4]. Also, when creating a classifier, there are difficulties with choosing the optimal values of hyperparameters, which affect its performance [5]. In addition, this method cannot work well with data containing noise or a large number of outliers.

Therefore, SVM can be an effective classification method for MRI brain images for Alzheimer's disease diagnosis, but care must be taken when setting up the model parameters and processing the data accordingly.

2.2.2. Decision Tree

Decision Tree is a supervised learning method used for classification and regression. The goal is to create a model that predicts the value of the target variable by learning simple decision rules derived from the characteristics of the data [6]. The most important feature of a decision tree is its ability to transform complex decision-making problems into simple processes, thus finding a solution that is clear and easier to interpret [7]. Thus, it is necessary to create a decision tree model based on the features obtained from the input images.

The Decision Tree method is popular in image classification because it is easy to interpret and visualize the decisions made by the algorithm. It also works well for datasets with a small number of classes and a small number of features and does not require complex input processing, but the dataset must be balanced [6]. However, there are disadvantages, as this method may not be the best choice for datasets with many classes or high-dimensional feature spaces, where other algorithms may be more efficient.

Thus, the Decision Tree method may be suitable for classifying images of Alzheimer's disease stages, since the dataset is not large and does not contain large image dimensions. However, one should be careful when tuning the model parameters to avoid overfitting.

2.2.3. Random Forest

Random Forest is an ensemble learning method that uses a set of decision trees to make predictions. In image classification, it works by using multiple decision trees to assign an image to one of several classes.

The algorithm starts by creating a large number of decision trees, each based on a random subset of the input data. Each decision tree in the random forest makes a prediction for the class of the input image based on its own set of rules, and these predictions are combined to make the final prediction for the image. During training, each decision tree is built based on a random subset of the input data, and at each split point in the tree, a random subset of features is considered [8]. This helps to reduce overfitting and increase the model's ability to generalize. At the prediction stage, each decision tree in the forest independently predicts the class of the input image, and the final prediction is the majority vote of all decision trees.

The Random Forest method is known for its high accuracy and ability to handle large datasets with high dimensionality compared to a single decision tree model. It also has the advantage of being able to handle data with noise and outliers [9]. However, training a random forest model can be computationally expensive and time-consuming, especially when dealing with large datasets. In addition, the model may not be as interpretable as a single decision tree, as it involves combining the predictions of several trees [8].

Thus, Random Forest can be an effective tool for the task of classifying images of disease stages, provided that the dataset is properly prepared and the model parameters are tuned.

2.2.4. Multi-Layer Perceptron

Multi-Layer Perceptron Classifier is a type of neural network that can be used for image classification. A supervised learning algorithm learns from a defined dataset using backpropagation.

A multilayer perceptron works by taking an input image and reducing it to a one-dimensional array of pixel values. This array is then fed to the input layer of the neural network. The input layer is connected to one or more hidden layers, each of which consists of several neurons. These hidden layers are responsible for learning the features in the input image. The output layer then produces a probability vector, with each element of the vector representing the probability of the input image belonging to a particular class. During the training phase, the weights of the connections between neurons are adjusted iteratively using backpropagation, which is a gradient descent optimization algorithm. This adjusts the weights in such a way as to minimize the difference between the predicted output of the network and the actual output based on the labelled data set [10]. Once a neural network is trained, it can be used to classify images by inputting new images and obtaining predicted probabilities for each class. The class with the highest probability is then assigned as the predicted class for that image.

Thus, a multilayer perceptron can be used for the task of classifying disease stage images because it performs well on small datasets, but it can also be sensitive to the choice of hyperparameters such as the number of hidden layers, the number of neurons in each layer, and the activation function.

2.3. Ensemble Voting Classifier

Ensemble learning is a machine learning method that uses multiple models to get better prediction results.

In an ensemble method, multiple models are trained on the same training set, and their prediction results are combined to produce the final result. This approach can provide better prediction accuracy than a single model if the models are properly tuned. There are different types of ensemble learning methods, which differ in the type of models used, the data sample, and the decision-making function. One of the most widespread ensemble learning techniques, Voting, was chosen for the study.

Voting Classifier is a voting algorithm used to combine the results of different classifiers into one ensemble to achieve better results.

In this study, we will use the Voting method because it does not require complex computations. Also, Voting can be a fast method because it does not require additional training of the ensemble model compared to other ensemble learning techniques. Voting is also a flexible method because it allows you to use different combinations of base models with different algorithms, hyperparameters and settings. This allows you to experiment with different models and find the optimal combination for a particular task.

The ensemble developed in this study consists of three classification models, which are selected based on the results of image classification separately by each method discussed in Section 2.2. The diagram of the ensemble formation from the selected models using Voting Classifier is shown in Figure 3.

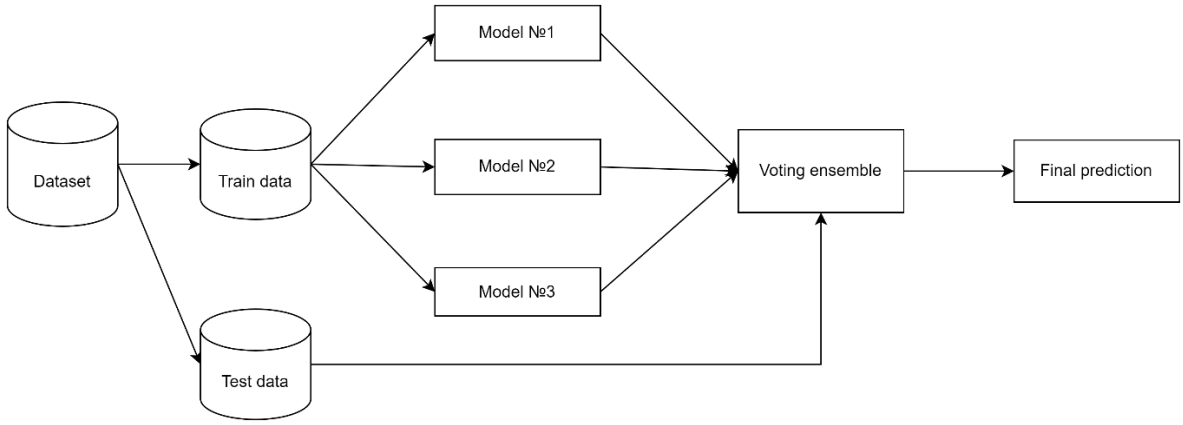


Figure 3. Diagram of creating an ensemble of classification models

The ensemble is trained on the training dataset, and the final testing is performed on the test dataset. To choose the best type of voting, we conducted a study with two types: hard voting and soft voting. Hard voting is majority voting, where each classifier gives a vote for its predicted class, and then the class with the highest number of votes is selected [11]. This can be useful for high quality data when all classifiers produce accurate results. Soft voting is a probability-based voting where each classifier provides probabilities for each class, and then the class with the highest average probability is selected [11]. This makes it possible to take into account the importance of some classifiers that may produce better results for certain classes. Soft voting usually works better when classifiers have different accuracies.

The ensemble is chosen according to the results of the study with the considered voting types.

3. Proposed solution

This section presents the results of the study obtained for AD prediction using the selected dataset. To summarize, first of all, the performance of each model for MRI image classification is studied separately. After that, the three models with the best results are selected and combined into an ensemble. To ensure the best result, two types of voting ensemble are investigated and the best one is selected.

The development of the software module is carried out on the free interactive platform Google Colab. To create a program module, the following libraries are used: Numpy, Opencv, Matplotlib, Seaborn, Scikit-learn, Hashlib, Scikit-image, Joblib.

3.1. Evaluation of the performance of individual classification models

Each model was trained on a training dataset and forecasting is performed on validation data. The Grid Search method was used to select model hyperparameters. Grid Search takes as input a list of parameters that we want to test and their possible values. For each combination of parameters, Grid Search performs cross-validation and calculates the average accuracy on the training set. The best parameter set was defined as the one with the highest average accuracy on the training set. The best parameters were used to build the final classifier on the entire dataset.

For the Decision Tree method, the hyperparameters that were selected are criterion and max_depth. Using Grid Search, we found that the best combination of parameters was criterion='entropy' and max_depth=6, which gave the best results on the sample. These parameters were used for further classification using Decision Tree. For the Random Forest method, the hyperparameters n_estimators and max_depth were selected and the best result was obtained with the parameters max_depth=None, n_estimators=300. For the Multi-layer Perceptron method, Grid Search was used to find the best result with the parameters activation='tanh', hidden_layer_sizes=(50, 50, 50), solver='sgd'. For the SVM method, the hyperparameters C=1, gamma=0.001, kernel='rbf' were chosen.

These classifications are conducted using different performance metrics in terms of Accuracy, Precision, Recall and F1 score. The results of all four models are summarized in Table 1.

Table 1
Comparison of the classification of AD using different models

Model	Metrics	Disease stages				
		AD	CN	EMCI	LMCI	MCI
Decision Tree	Precision	0.87	0.85	0.90	0.79	0.95
	Recall	0.91	0.84	0.87	0.84	0.89
	F1-score	0.89	0.84	0.88	0.81	0.92
	Accuracy			0.87		
Random Forest	Precision	0.96	0.93	0.94	0.88	0.98
	Recall	0.94	0.90	0.97	0.93	0.95
	F1-score	0.95	0.92	0.96	0.91	0.96
	Accuracy			0.94		
SVM	Precision	0.95	0.92	0.93	0.92	0.96
	Recall	0.93	0.94	0.97	0.90	0.95
	F1-score	0.94	0.93	0.95	0.91	0.95
	Accuracy			0.94		
Multi-Layer Perceptron	Precision	0.95	0.94	0.95	0.91	0.97
	Recall	0.96	0.93	0.97	0.92	0.95
	F1-score	0.95	0.94	0.96	0.92	0.96
	Accuracy			0.95		

In summary, the Random Forest, SVM, and Multi-Layer Perceptron models outperform the Decision Tree model in terms of overall accuracy. Among these three, the Multi-Layer Perceptron has the highest accuracy. All models have relatively high precision, recall, and F1-scores for most disease stages, indicating their effectiveness in classifying these stages. The best three models were selected for further integration into an ensemble: Random Forest, SVM, and Multi-Layer Perceptron.

3.2. Evaluation of the performance of different ensemble classification methods

Table 2 displays the results of classifying MRI images into five stages of Alzheimer's disease using ensembles of two types of voting.

In terms of precision, both Hard Voting and Soft Voting are strong, with high precision for the AD and CN stages, indicating low false-positive rates.

Table 2
Comparison of AD classification with different ensemble classification methods

Ensemble	Metrics	Disease stages				
		AD	CN	EMCI	LMCI	MCI
Hard Voting	Precision	0.98	0.95	0.93	0.91	0.97
	Recall	0.96	0.95	0.99	0.92	0.93
	F1-score	0.97	0.95	0.96	0.92	0.95
	Accuracy			0.95		
Soft Voting	Precision	0.98	0.95	0.94	0.94	0.97
	Recall	0.97	0.97	0.99	0.91	0.94
	F1-score	0.97	0.96	0.96	0.92	0.95
	Accuracy			0.96		

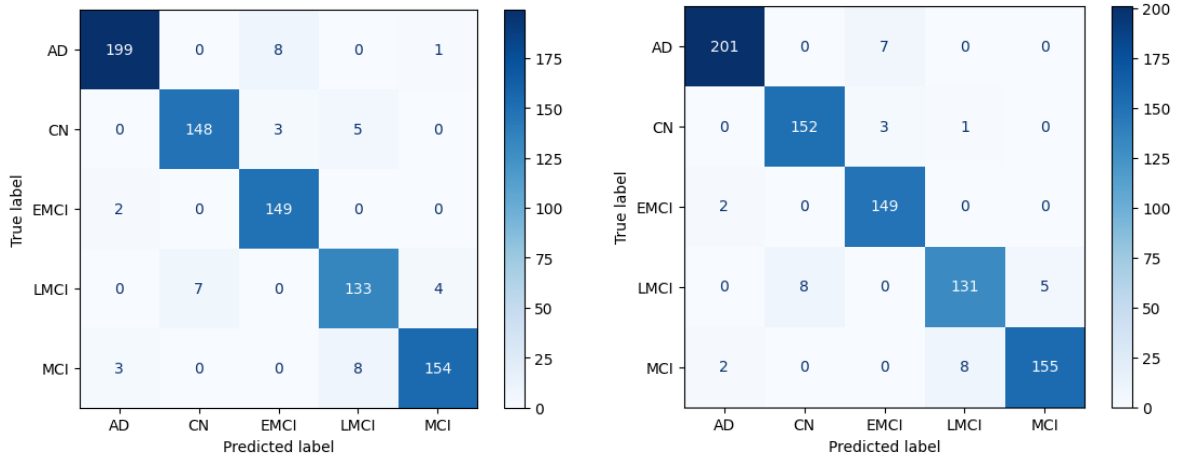
Soft Voting generally outperforms Hard Voting in terms of recall, achieving higher recall values for CN, EMCI, and MCI. This suggests that Soft Voting is better at identifying true positive cases in these stages.

F1-scores are relatively similar between the two methods, with both ensembles maintaining a good balance between precision and recall for most disease stages.

Soft Voting has a slightly higher overall accuracy (0.96) compared to Hard Voting (0.95).

In summary, Soft Voting Classifier appears to be a slightly better ensemble method in this context, as it achieves higher recall and accuracy while maintaining strong precision and F1-scores for most stages of the disease.

The Figure 4 shows the error matrices for both ensembles. From the obtained confusion matrix, we can see that in general, the results are better compared to the confusion matrix of the ensemble with hard voting type. The number of correctly classified images of the AD class is 2 samples higher than when using the ensemble with hard voting type. For the CN class, this number is 4 samples more, and for the MCI class, it is 1 sample more. For the EMCI class, the number of correctly classified images has not changed. However, for the LMCI class, the number of correctly classified data decreased by 2 samples. The problem of classifying the LMCI class may be related to the specificity of the data of this class in the dataset used in the study.



(a) (b)

Feature 4. The confusion matrix of the models. (a) Hard Voting Classifier; (b) Soft Voting Classifier

The Soft Voting Classifier with the highest efficiency and accuracy of 0.96 was selected. The results show good performance of the classification model with high precision and recall for most classes, which indicates its ability to recognize different stages of Alzheimer's disease in the studied dataset.

4. Discussion

To evaluate the effectiveness of the developed model, we will compare the results of the soft voting ensemble with the results of individual models used in this ensemble. Comparative results are shown in the Table 3.

Table 3
Combined results of the ensemble and individual models tested

Model	Metrics	Disease stages				
		AD	CN	EMCI	LMCI	MCI
Ансамбль soft voting	Precision	0.98	0.95	0.94	0.94	0.97
	Recall	0.97	0.97	0.99	0.91	0.94
	F1-score	0.97	0.96	0.96	0.92	0.95
	Accuracy			0.96		
Random Forest	Precision	0.96	0.93	0.94	0.88	0.98
	Recall	0.94	0.90	0.97	0.93	0.95
	F1-score	0.95	0.92	0.96	0.91	0.96
	Accuracy			0.94		
SVM	Precision	0.95	0.92	0.93	0.92	0.96
	Recall	0.93	0.94	0.97	0.90	0.95
	F1-score	0.94	0.93	0.95	0.91	0.95
	Accuracy			0.94		
Multi-Layer Perceptron	Precision	0.95	0.94	0.95	0.91	0.97
	Recall	0.96	0.93	0.97	0.92	0.95
	F1-score	0.95	0.94	0.96	0.92	0.96
	Accuracy			0.95		

Analyzing this table with metrics for each model, the following conclusions can be drawn for each stage of the disease.

AD:

- The Soft Voting Ensemble has the highest Precision, Recall, and F1-score.
- Random Forest and SVM also perform well, but the Soft Voting Ensemble is slightly better.

CN:

- The Soft Voting Ensemble and SVM have the highest Precision, Recall, and F1-score.
- Random Forest and MLP perform slightly worse.

EMCI:

- The Soft Voting Ensemble has the highest Precision and F1-score.
- Random Forest and SVM perform well, with slightly lower Precision.
- MLP also performs well but has slightly lower Recall and F1-score.

LMCI:

- The Soft Voting Ensemble has the highest Recall, while Random Forest has the highest Precision.
- The Soft Voting Ensemble has a better F1-score.

MCI:

- The Soft Voting Ensemble has the highest Precision, Recall, and F1-score.

A common metric for evaluating models is accuracy, which reflects the overall ability of the model to correctly classify instances across all classes. In this case, the soft voting ensemble has the highest accuracy (0.96), which means that it correctly classifies 96% of instances from all classes at different stages of the disease.

Thus, the developed soft voting ensemble proved to be the best option among the considered models, with high precision, recall, and F1-score for most classes and the highest overall accuracy for the studied dataset.

5. Conclusion

The ensemble of Random Forest, Multi-Layer Perceptron, and Support Vector Machine models proposed in this study has achieved the accuracy of 96% in classifying patients into five stages of AD. The multiclass classification of BA into five stages is of great importance in the field of diagnosis of AD.

The developed soft voting ensemble proved to be the best option among the considered models, with high precision, recall and F1-score for most classes and the highest overall accuracy for the studied dataset. This is a significant improvement over the performance of individual ML models, which typically achieve accuracy in the range of 90%. The proposed ensemble model is therefore a valuable tool for the early and accurate diagnosis of AD, which can lead to better patient outcomes.

Future work could focus on investigating the performance of the ensemble model on other AD datasets, developing and evaluating new ensemble learning algorithms for AD classification, integrating the ensemble model with other clinical data, such as imaging and neuropsychological data, to further improve its performance.

6. References

[1] S. Nebehay. Number of people with dementia set to jump 40% to 78 mln by 2030, 2021. URL: https://www.reuters.com/business/healthcare-pharmaceuticals/number-people-with-dementia-set-jump-40-78-mln-by-2030-who-2021-09-02/.

[2] Alzheimer datasets 5 classes | kaggle. URL: https://www.kaggle.com/datasets/phamnguyenduytien/alzheimer-datasets-5-classes.

[3] H. Bhavsar and M. H. Panchal, "A Review on Support Vector Machine for Data Classification", International Journal of Advanced Research in Computer Engineering & Technology, Volume 1, No. 10, pp. 185-189, 2012.

[4] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, A comprehensive survey onsupport vector machine classification: Applica-tions, challenges and trends, Neurocomputing 408(2020) 189–215. doi: https://doi.org/10.1016/j.neucom.2019.10.118.

[5] D. J. Kalita, V. P. Singh and V. Kumar, "A survey on svm hyper-parameters optimization techniques," inSocial Networking and Computational Intelligence, Berlin, Germany, Springer, 2020, pp. 243–256. doi: https://doi.org/10.1007/978-981-15-2071-6_20.

[6] 1.10. decision trees — scikit-learn 1.2.2 documentation URL: https://scikit-learn.org/stable/modules/tree.html.

[7] Priyanka, D. Kumar, "Decision tree classifier: a detailed survey", International Journal of Information and Decision Sciences, vol. 12, no. 3, pp. 246269, 2020.

[8] Ibrahim, I. Abdulazeez, A. The role of machine learning algorithms for diagnosing diseases. J. Appl. Sci. Technol. Trends 2 (2021). doi: http://dx.doi.org/10.38094/jastt20179.

[9] Speiser JL, Miller ME, Tooze J, Ip E. A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. Expert Syst Appl. 2019 Nov 15;134:93-101. doi: https://doi.org/10.1016/j.eswa.2019.05.028.

[10] Multilayer perceptron / 2023. URL: https://en.wikipedia.org/w/index.php?title=Multilayer_perceptron&oldid=1147594876.

[11] Sklearn.ensemble.votingclassifier — scikit-learn 1.2.2 documentation. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html.