Development and Optimization of Deep Learning Systems for MRI Analysis in Alzheimer's Disease Monitoring

Jolanta Podolszańska

Jan Długosz University, Częstochowa, Poland

https://doi.org/10.26636/jtit.2024.4.1815

Abstract — Alzheimer's disease is one of the leading causes of dementia worldwide, and its increasing prevalence presents significant diagnostic and therapeutic challenges, particularly in an aging population. Current diagnostic methods, including patient history reviews, neuropsychological tests, and MRI scans, often fail to achieve adequate sensitivity and specificity levels. In response to these challenges, this study introduces an advanced convolutional neural network (CNN) model that combines ResNet-50 and Inception V3 architectures to classify, with a high degree of precision, the stages of Alzheimer's disease based on MRI. The model was developed and evaluated using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and classifies MRI scans into four clinical categories representing different stages of disease severity. The evaluation results, based on the precision, sensitivity and F1 score metrics, demonstrate the effectiveness of the model. Additional augmentation techniques and differential class weighting further enhance the accuracy of the model. Visualization of results using the t-SNE method and the confusion matrix underscores the ability to distinguish between disease categories, supporting the model's potential to aid in neurological diagnosis and classification.

Keywords — Alzheimer disease, convolutional neural network, Inception V3, ResNet-50, t-SNE analysis

1. Introduction

Dementia is one of the leading causes of disability and mortality, particularly among the elderly [1]–[5]. It significantly affects cognitive functions, especially memory, making daily activities more difficult. According to the World Health Organization (WHO), approximately 55 million people worldwide suffer from dementia, with this number projected to increase to 139 million by 2050 [6]. Alzheimer's disease, the most common form of dementia, accounts for 60-70% of cases [7]. This paper focuses on developing and validating deep learning models for brain image analysis, specifically designed to assess the severity of Alzheimer's disease.

Recent advances in machine learning and deep learning have contributed significantly to the diagnosis and prognosis of neurodegenerative diseases, such as Alzheimer's. Techniques such as magnetic resonance analysis, deep neural networks, and feature extraction methods have improved the degree of precision achieved in diagnostic processes.

Key studies highlight the effectiveness of various machine learning models in diagnosing Alzheimer's disease. A classification model using ResNet-50 with self-attention layers improved the classification of the stage of Alzheimer's, demonstrating the potential that deep feature extraction techniques combined with Bayesian optimization offer in terms of hyperparameter tuning [8]. The study achieved a high accuracy rate, although concerns about overfitting remain, due to optimized hyperparameters [9].

Another approach from Batangas State University applied principal component analysis (PCA) and the synthetic minority oversampling technique (SMOTE) to address data imbalance, achieving a 99% accuracy rate using the extra tree algorithm [10]. However, the potential for overfitting requires a cautious interpretation of these results.

Other research has focused on more computationally efficient models, such as vision transformers (ViT) which demonstrated a high precision rate of 99.83% in classifying Alzheimer's stages with minimal computational resources [11]. This makes it a promising option for clinical environments with limited resources. Similarly, studies exploring random forest (RF), support vector machine (SVM), and convolutional neural networks (CNN) for dementia classification showed that SVM, especially when combined with advanced feature extraction techniques, could outperform other models [12].

In addition, combining biomarkers, such as amyloid beta and tau proteins, with machine learning classifiers has proven effective in identifying the early stages of Alzheimer's. Ensemble methods, such as boosted trees and logistic regression, have shown strong precision, highlighting the potential of integrating biomarkers into diagnostic models [13]–[18].

These findings suggest that deep learning models, particularly when optimized for data imbalance and computational efficiency, hold promise in improving early diagnosis and classification of Alzheimer's disease. However, more research is needed to validate these models in clinical settings and address potential problems such as overfitting and interpretability.

2. Methods

This section outlines the steps involved in preparing the preprocessed models for application. The project uses the

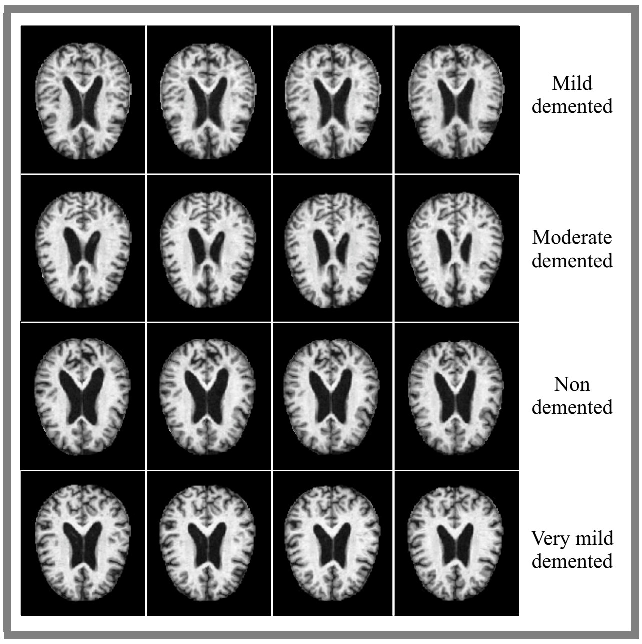

Fig. 1. Sample MRI images representing different stages of Alzheimer's disease classification from the ADNI dataset.

"Alzheimer MRI preprocessed dataset" made available by Kumar and Shastri in 2022 on the Kaggle platform. The dataset consists exclusively of magnetic resonance imaging (MRI) scans of the brains of patients diagnosed with various stages of Alzheimer's disease, as well as healthy control subjects. Data have been preprocessed by applying normalization and other techniques to improve image quality, which is crucial for the effectiveness of deep learning algorithms [19].

The dataset contains images collected from various sources, including websites, hospitals, and public repositories (Fig. 1). They were preprocessed and normalized to a uniform 128×128-pixel format to facilitate data analysis and processing. It contains a total of 6 400 images which have been classified into four classes corresponding to different stages of Alzheimer's disease:

- class 1: mild dementia (896 images),
- class 2: moderate dementia (64 images),
- class 3: non-dementia (3200 images),
- class 4: very mild dementia (2240 images).

The primary objective of using this dataset is to develop and validate advanced predictive models capable of accurately classifying and predicting different stages of Alzheimer's disease through computer analysis. By relying on machine learning and deep learning techniques, the objective is to create models that not only improve medical diagnosis but also contribute to a deeper understanding of the neurodegenerative processes associated with the disease in question.

2.1. Feature Selection and Development of Predictive Models

Two advanced neural network models, ResNet-50 and Inception V3, were used to extract features from the images. Both models were pre-trained on datasets consisting of MRI images. The images were pre-processed and upscaled to a 299×299-pixel format for processing with the use of deep learning algorithms.

The pre-trained network, trained for 50 epochs, utilizes a residual architecture to optimize the training process of the neural networks. In contrast, the model incorporating a multipath data flow concept allows for efficient processing with fewer parameters. The outputs of both models are merged to form a single high-dimensional feature vector. This combined feature vector is passed through a fully connected layer, followed by a dropout layer to regularize the model and prevent overfitting. Finally, the processed features are passed through an output layer which generates predictions for each of the four severity classes.

2.2. Model Validation

A thorough validation was conducted to evaluate the performance of the hybrid CNN model. The data were divided into a training set (approximately 80%) and a validation set (approximately 20%), ensuring that the model was tested on data unseen during training, allowing for the assessment of its generalization capability. The metrics used to evaluate quality included precision, sensitivity, F1 score, and a confusion matrix. The validation process was designed to demonstrate whether the hybrid CNN model, which combines the ResNet-50 and Inception V3 architectures, is capable of effectively classifying different stages of Alzheimer's disease based on MRI images. High precision, sensitivity, and F1 score values should confirm the model's potential to serve as a valuable tool in supporting medical diagnosis and advancing research into Alzheimer's disease.

3. Results

As mentioned earlier, the data were divided into training and validation sets. Cross-validation was used to provide a more comprehensive evaluation of the model, ensuring that the results were not random and adequately represented the entire dataset. The precision metric, which measures the precision of a classifier in predicting the positive class, was used to evaluate the models. Precision is calculated as the ratio of true positives (TP) to the sum of true positives and false positives (FP).

$$Precision = \frac{TP}{TP + FP} \, . \tag{1}$$

Sensitivity is the classifier's ability to provide information about the detection of all true positives and is defined as:

$$Recall = \frac{TP}{TP + FN} \, , \tag{2}$$

where FN stands for false negatives.

The F1 measure involves calculating the harmonic mean of precision and sensitivity, which allows for a balanced model evaluation that considers both accuracy and completeness of predictions:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \, . \tag{3}$$

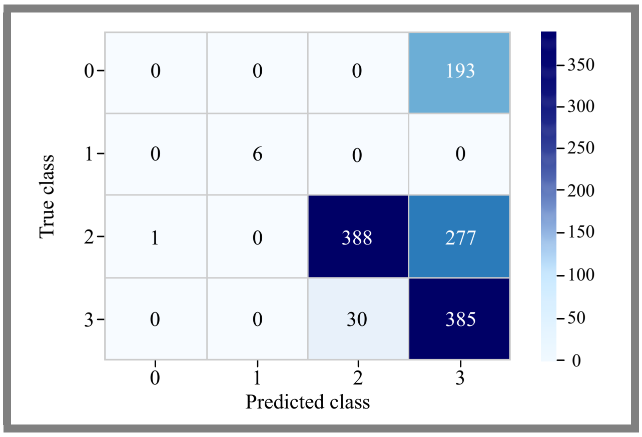

Fig. 2. Confusion matrix showing classification results for four Alzheimer's severity classes, highlighting correct and misclassified samples.

The overall performance of the model was then calculated as the ratio of the sum of true positive and accurate negative results to the total number of samples.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \, . \tag{4}$$

The confusion matrix allowed to assess the classification performance by identifying the numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) outcomes for each class (Fig. 2). This analysis helps determine which classes are most frequently misclassified, indicating potential areas where the model may require improvement. The validation results were presented using graphs and confusion matrices which provide a detailed evaluation of the effectiveness of the classifiers. Despite good overall performance, certain classes, in particular those with smaller sample sizes, may require further optimization, as highlighted in the detailed interpretation of the results.

3.1. Interpretation of Results and Evaluation of Model Effectiveness

This article focuses on developing predictive models that use advanced deep learning techniques to analyze brain magnetic resonance images to assess the severity of Alzheimer's disease. A hybrid neural network architecture, combining the ResNet-50 and Inception V3 models, was used to create a more accurate classification system. Research aimed not only to enhance the accuracy of Alzheimer's disease diagnosis, but also to contribute to a deeper understanding of the neurodegenerative processes associated with the condition. The experiments involved detailed performance analyses of the model, including validation through confusion matrices and visualization of the feature space using t-SNE.

Below is a summary of the evaluation of the convolutional neural network (CNN) model based on precision, recall, F1 score, and accuracy metrics (Tab. 1). The model was evaluated using the test dataset, ensuring its applicability at different stages of Alzheimer's disease.

Tab. 1. Results of the collective evaluation of the convolutional neural network (CNN) model.

Metric	Value
Precision	99.09%
Recall	98.84%
F1 score	98.93%
Accuracy	99.1%

Tab. 2. Results of a collective evaluation the convolutional neural networks (CNN) model's precision, recall, and F1 score, expressed as values between 0 and 1.

Class	Precision	Recall	F1 score
Non-demented	0.995	0.993	0.994
Mild demented	0.987	0.989	0.988
Moderate demented	0.985	0.984	0.985
Very mild demented	0.990	0.988	0.989

Table 2 presents detailed results for each of the four classes, providing a more in-depth analysis of the model's ability to classify different stages of Alzheimer's disease.

The t-distributed stochastic neighbor embedding (t-SNE) projection shown in Fig. 3 illustrates the two-dimensional feature space of the model's output. Each point corresponds to a sample from the dataset. The separation between classes is clear, indicating that the model has learned to effectively distinguish between different stages of Alzheimer's disease. Class 2 (green) and class 3 (red) form distinct clusters, suggesting that these classes are well recognized. Class 0 (orange), represented by a smaller group, is located toward the bottom, indicating potential difficulties in classification or fewer samples. Class 1 (blue) appears to be dispersed across several groups, hinting at challenges in distinguishing this particular category. Most points are tightly clustered, suggesting decent

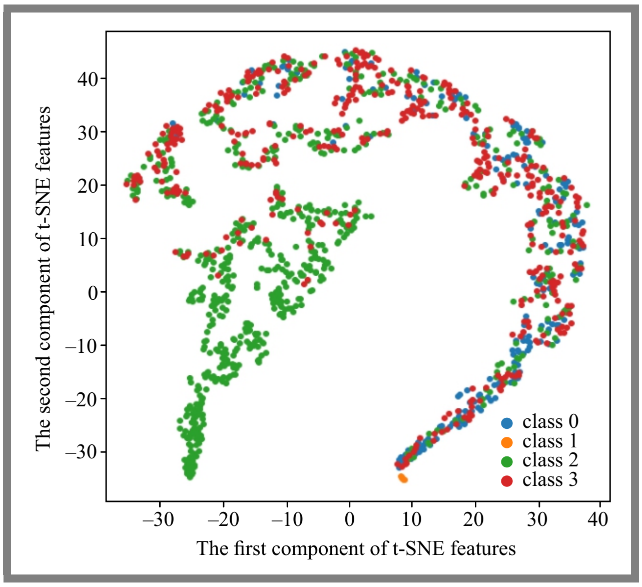

Fig. 3. Projection of t-SNE showing the clustering of Alzheimer's disease severity classes.

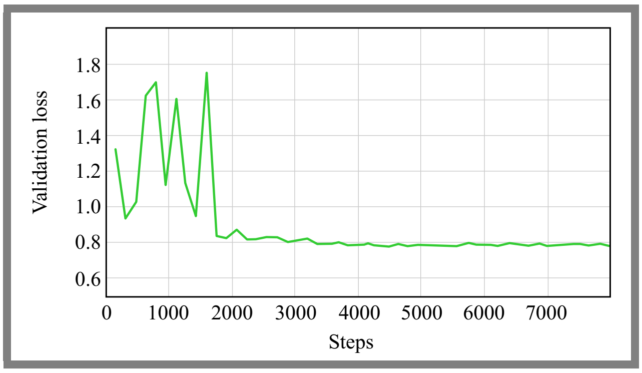

Fig. 4. Validation loss vs. steps, showing stabilization after 2 000 steps and indicating model convergence.

model performance in separating the data, although some dispersion in class 1 and partially in class 0 indicates structural complexity the model struggled with.

The plot shown in Fig. 3 presents the projection of t-SNE onto the two-dimensional feature space of the model's output. Each point on the graph represents one sample of the data set. One may notice a clear separation of classes, suggesting that the model has successfully learned to distinguish between different classes at the level of output characteristics. Class 2 (green) and class 3 (red) form very distinct groups, proving that the model is more efficient at recognizing these particular classes. Class 0 (orange) appears to have the form of a small group and is located at the bottom of the graph, which may suggest that this class has fewer samples or is more difficult for the model to recognize. Class 1 (blue) is present in several groups, which may indicate some problems in distinguishing this class or the complexity of its features. Most of the points in each class are well clustered, which is a positive result.

In the plot, one may notice that the initial changes are large, especially before approximately 2 000 steps, where the value of the loss decreases rapidly (Fig. 4). Once this limit has been exceeded, the graph clearly begins to flatten out, suggesting that further training has a diminishing effect on reducing the loss. This is the point that we call the "elbow", and it indicates the optimal number of epochs after which further training of the model no longer provides significant benefits. The network was trained in 50 epochs.

When analyzing the model training results, one should pay attention to the validation loss function depending on the number of epochs, which can be represented as a function of $L(epoch)$. In the elbow method, we look for a point k for which the change in the value of the loss between epochs k and $k+1$ ceases to be significant and its gradient flattens out. Mathematically, this can be put in the following form:

$$\Delta L = L(k+1) - L(k). \tag{5}$$

In Eq. (5), if

$$|\Delta L| =<< \epsilon \tag{6}$$

for a certain tolerance $\varepsilon > 0$, then k can be considered the optimal number of epochs from the validation loss graphs shown. One may observe that after reaching approximately 30 epochs, ΔL decreases to values close to zero, suggesting that further training of the model does not produce significant improvements.

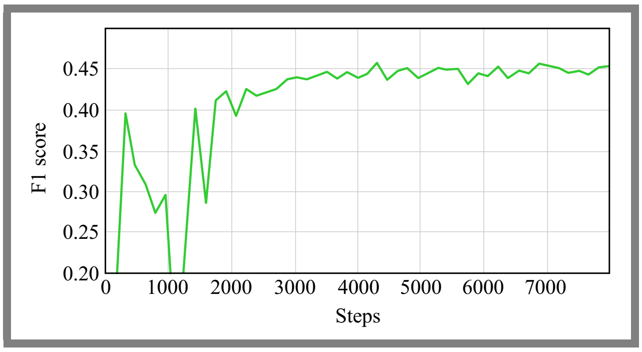

Fig. 5. Graph showing validation of F1 score vs. steps demonstrates a consistent improvement in model's performance after 2 000 steps.

Formally, we can assume that there is a certain point where $k \geqslant 30$:

$$\frac{\delta L}{\delta\, epoch} \approx 0\,, \tag{7}$$

which indicates that further increases in the number of epochs result in minimal changes in the value of the loss function, and thus in the quality of the model. Calculating the differences between successive epochs using data from the validation loss chart. For epochs in the 1–30 range, the differences in the loss values are significant. From epoch 30 onwards, the loss values begin to stabilize, and the differences between successive epochs are marginal. The values of these differences can be calculated in the form of finite differences in the following way:

$$\Delta L_30 = L(31) - L(30)\,. \tag{8}$$

For ΔL_30 close to zero, we reach the point where continuing training does not significantly improve performance. The result of the analysis suggests that the number of epochs equaling 30 is the optimal choice for this model, as once this number is reached, the changes in the validation loss value are marginal.

In terms of the F1 validation score vs. steps plot, as shown in Fig. 5, we observe a change in the value of the F1 score for validation data as the model's training process progresses, i.e., the number of steps corresponding to training epochs increases. The F1 score is a measure of the harmonic mean between precision and sensitivity, making it one of the key metrics for classification models, especially when the data are unbalanced.

At first (up to approx. 2 000 steps), the F1 score fluctuates and is subject to irregular changes. This could mean that the model needs more epochs to learn effective data representations and adjust to the classification task. Once this number of steps is exceeded, the F1 score values begin to rise more steadily until approx. 5 000 steps, after which they become flatter, indicating stabilization. The elbow method may be used in this case to identify the point at which the F1 score value stops increasing significantly.

We are looking for the step after which the change in the F1 score value becomes marginal, and further training does not result in any significant improvement. Change $\Delta F1$ between

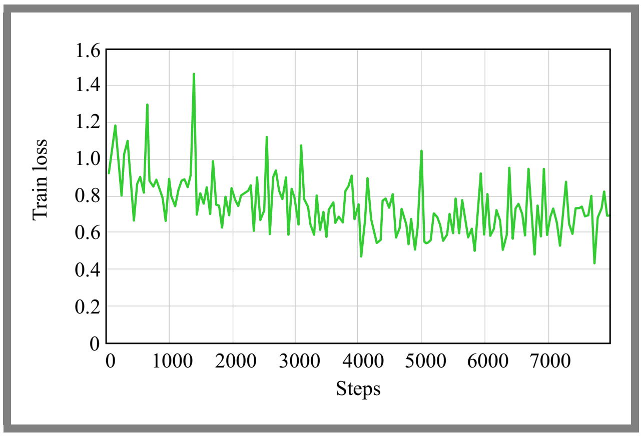

Fig. 6. Changes in the cost function as a function of the number of steps.

steps K and $k+1$ can be described as:

$$\Delta F_1(k) = F_1(k+1) - F_1(k)\ . \tag{9}$$

If $|\Delta F_1|$ becomes close to zero after a certain step k, the model can be considered to have reached optimal performance. In this case, the graph suggests that after approx. 5 000 steps ΔF_1 is declining, suggesting that it is less and less beneficial to continue to train the model. Assuming that we have the corresponding F1 score values for steps 5 000 and 6 000, one may calculate the difference as follows:

$$\Delta F_1(5000) = F_1(6000) - F_1(5000). \tag{10}$$

If $|\Delta F_1(5000)| \approx 0$, it means that further training steps do not bring significant improvement, which would confirm the elbow point at approx. 5 000 steps. According to the elbow method, the number of steps equaling approx. 5 000 may be considered optimal, since after this step changes in the F1 score values are marginal, which means that further training of the model no longer brings significant benefits in improving the quality of classification.

Figure 6 shows the value of the cost function (train loss) for the training data as training progresses, measured in steps. The cost function (loss function) measures how well the model predicts target values (labels) compared to actual performance, and the goal of the optimization process is to minimize this value.

This process can be described using the following iterative equation for gradient descent:

$$\theta_t + 1 = \theta_t - \eta \nabla_\theta L(\theta_t)\ , \tag{11}$$

where: θ_t is the vector of model parameters at step t, η is the learning rate, $\nabla_\theta L(\theta_t)$ is the loss function at point θ_t.

Initially, we observe sharp changes in the loss function, as the gradients $\nabla_\theta L(\theta)$ have high values at the beginning of the training process, leading to large parameter updates $\Delta\theta$ in each step. The early training phase is often unstable, which is reflected in Fig. 6 by large fluctuations, especially in the range of 0–1 000 steps. This can be explained by the fact that at the beginning of the training process, the model is far from the optimal local minimum, so the gradients are steep, causing jumps in the loss function.

As the model approaches the optimal minimum, the gradients decrease, resulting in smaller parameter updates θ. Therefore, we observe a flatter curve for steps $t > 2\,000$, which indicates that the optimization process is slowing down and the model is gradually approaching the minimum of the loss function.

For steps $t > 2\,000$, the curve shows fewer fluctuations, which is typical for the convergence of a gradient algorithm. In particular, the AdamW optimizer further accelerates convergence by using mechanisms that adjust the adaptive step size and regularization weights, thus helping avoid overfitting. This mechanism can be mathematically described by adding regularization to the objective function in the following way:

$$L(\theta) = L_{data}(\theta) + \lambda\|\theta\|^2, \tag{12}$$

where $L_{data}(\theta)$ is the loss function of the data and $\lambda\|\theta\|^2$ is the $L2$ applied by AdamW. In summary, the graph shows the effectiveness of minimizing the loss function as the steps progress. The sharp initial decrease and subsequent stabilization of the loss function indicate that the model is approaching an optimal solution, with a further loss reduction occurring in smaller steps, which is typical for convergence in gradient-based algorithms.

4. Conclusions

The research implemented advanced deep learning techniques using a modified CNN architecture to assess and classify various stages of Alzheimer's disease based on MRI images. The results demonstrate that the model achieves high degrees of precision, sensitivity, and effectiveness in identifying multiple stages of disease progression, highlighting its potential application as a tool supporting medical diagnostic procedures.

Systematic improvements in accuracy showcase its ability to learn complex data patterns, which may contribute to a better understanding of the neurodegenerative processes associated with Alzheimer's disease.

Based on the findings of the study, it is recommended that performance be further evaluated by incorporating more MRI images from diverse populations, which could improve generalizability and contribute to understanding variations between patient groups. This also involves obtaining the appropriate consent to prevent the misuse of patient data. Future research could explore other advanced neural network architectures, such as generative adversarial networks (GANs) and capsule networks, which may offer new perspectives and improved diagnosis accuracy.

Future work should focus on increasing the interpretability of deep learning models, which is critical in a medical context to help physicians better understand the decision-making processes relied upon by the models. Extensive clinical validations are also necessary to confirm the effectiveness in real-world medical environments and constitute a key step in progressing towards practical application.

References

[1] J.H. Shin, “Dementia Epidemiology Fact Sheet 2022”, *Annals of Rehabilitation Medicine*, vol. 46, no. 2, pp. 53–59, 2022 (https://doi.org/10.5535/arm.22027).

[2] World Health Organization, *Dementia*, Geneva, Switzerland: World Health Organization, 2024 [Online] (https://www.who.int/news-room/fact-sheets/detail/dementia).

[3] V.A. Ciurea *et al.*, “Alzheimer’s Disease: 120 years of Research and Progress”, *Journal of Medicine and Life*, vol. 16, no. 2, pp. 173–177, 2023 (https://doi.org/10.25122/jml-2022-0111).

[4] J.A. Soria Lopez, H.M. Gonzalez, and G.C. Leger, “Alzheimer’s Disease”, in: *Handbook of Clinical Neurology*, pp. 231–255, 2019 (https://doi.org/10.1016/B978-0-12-804766-8.00013-3).

[5] J.S. Snowden, “Changing Perspectives on Frontotemporal Dementia: A Review”, *Journal of Neuropsychology*, vol. 17, no. 2, pp. 211–234, 2023 (https://doi.org/10.1111/jnp.12297).

[6] A. Alzheimer, “On Certain Peculiar Diseases of Old Age”, in: *History of Psychiatry*, H. Forstl and R. Levy, vol. 2, no. 5, pp. 71–101, 1991 (https://doi.org/10.1177/0957154X9100200505).

[7] A.J. Intorcia *et al.*, “A Modification of the Bielschowsky Silver Stain for Alzheimer Neuritic Plaques: Suppression of Artifactual Staining by Pretreatment with Oxidizing Agents”, *BioRxiv*, 2019 (https://doi.org/10.1101/570093).

[8] N. Yaqoob *et al.*, Prediction of Alzheimer’s Disease Stages Based on ResNet-Self-attention Architecture with Bayesian Optimization and Best Features Selection”, *Frontiers in Computational Neuroscience*, 2024 (https://doi.org/10.3389/fncom.2024.1393849).

[9] H. Habehh and S. Gohel, “Machine Learning in Healthcare”, *Current Genomics*, vol. 22, no. 4, pp. 291–300, 2021 (https://doi.org/10.2174/1389202922666210705124359).

[10] R.M. Hernandez *et al.*, “Application of Machine Learning on MRI Scans for Alzheimer’s Disease Early Detection”, *Proceedings of the 8th International Conference on Sustainable Information Engineering and Technology*, pp. 143–149, 2023 (https://doi.org/10.1145/3626641.3627609).

[11] M.H. Alshayeji, “Alzheimer’s Disease Detection and Stage Identification from Magnetic Resonance Brain Images Using Vision Transformer”, *Machine Learning: Science and Technology*, vol. 5, no. 3, art. no. 035011, 2024 (https://doi.org/10.1088/2632-2153/ad5fdc).

[12] M.G. Hussain and Y. Shiren, “Identifying Alzheimer Disease Dementia Levels Using Machine Learning Methods”, *Medical Research Archives*, vol. 11, no. 7.1, 2023 (https://doi.org/10.18103/mra.v11i7.1.4039).

[13] V.K. Tiwari, P. Indic, and S. Tabassum, “Machine Learning Classification of Alzheimer’s Disease Stages Using Cerebrospinal Fluid Biomarkers Alone”, *arXiv*, 2024 (https://doi.org/10.48550/arXiv.2401.00981).

[14] S.E. Sorour *et al.*, “Classification of Alzheimer’s Disease Using MRI Data Based on Deep Learning Techniques”, *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 2, art. no. 101940, 2024 (https://doi.org/10.1016/j.jksuci.2024.101940).

[15] D.A. Arafa *et al.*, “A Deep Learning Framework for Early Diagnosis of Alzheimer’s Disease on MRI Images”, *Multimedia Tools and Applications*, vol. 83, pp. 3767–3799, 2024 (https://doi.org/10.1007/s11042-023-15738-7).

[16] A.M. El-Assy, H.M. Amer, H.M. Ibrahim, and M.A. Mohamed, “A Novel CN Architecture for Accurate Early Detection and Classification of Alzheimer’s Disease Using MRI Data”, *Scientific Reports*, vol. 14, art. no. 3463, 2024 (https://doi.org/10.1038/s41598-024-53733-6).

[17] Y. Liu *et al.*, “MPC-STANet: Alzheimer’s Disease Recognition Method Based on Multiple Phantom Convolution and Spatial Transformation Attention Mechanism”, *Frontiers in Aging Neuroscience*, vol. 14, art. no. 918462, 2022 (https://doi.org/10.3389/fnagi.2022.918462).

[18] S. Ha, Y. Yoon, and J. Lee, “Meta-ensemble Learning with a Multi-headed Model for Few-shot Problems”, *ICT Express*, vol. 9, no. 5, pp. 909–914, 2023 (https://doi.org/10.1016/j.icte.2022.09.001).

[19] https://www.kaggle.com/datasets

Jolanta Podolszańska, M.Sc.
Department of Mathematics and Computer Science
https://orcid.org/0000-0002-6032-5654
E-mail: j.podolszanska@ujd.edu.pl
Jan Długosz University, Częstochowa, Poland
https://www.ujd.edu.pl