ORIGINAL ARTICLE

# INTEGRATING MACHINE LEARNING INTO ALZHEIMER'S RESEARCH: A CLINICAL DECISION SUPPORT SYSTEM APPROACH

Turkish Journal of Geriatrics
2025; 28(2):137–149

DOI: 10.29400/tjgeri.2025.430

Ceyda ÜNAL[1]
Yılmaz GÖKŞEN[1]

[1] Dokuz Eylül University, Management Information Systems, İzmir, Turkey

*Correspondence*

Ceyda ÜNAL
Phone : +902323010785
e-mail : ceyda.unal@deu.edu.tr

Received : Mar 15, 2025
Accepted : May 25, 2025

This study is derived from Ceyda Ünal's PhD dissertation completed under the supervision of Prof. Yılmaz Gökşen at Dokuz Eylül University.

*Cite this article as:*
Ünal C, Gökşen Y. Integrating Machine Learning Into Alzheimer's Research: A Clinical Decision Support System Approach. Turkish Journal of Geriatrics 2025; 28(2):137–149. doi: 10.29400/tjgeri.2025.430

## Abstract

**Introduction:** Alzheimer's Disease is defined as a progressive brain disease that affects memory, thinking, and behavior. It accounts for 50-80% of all dementia cases. Diagnosis can be challenging, particularly in the early stages and notably in mild cognitive impairment. The growing number and diversity of health-related data have led to the widespread use of machine learning algorithms in the early detection of Alzheimer's Disease. This study focuses on developing an Artificial Intelligence-based clinical decision support system that classifies individuals as individuals with Alzheimer's Disease, Mild Cognitive Impairment, or healthy individuals.

**Materials and Method:** The dataset used in the study was obtained from the Brain Aging and Dementia Unit of the Geriatrics Department. All patients aged between 45 and 96 years and followed up in the clinic were examined. Classification was performed using the Logistic Regression, Naive Bayes, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machines, Decision Trees, and ensemble methods.

**Results:** The CatBoost algorithm outperformed the other models in terms of accuracy. Ensemble learning methods outperformed traditional methods for 176 samples in the Alzheimer class. Random Forest method exhibited the highest precision for Mild Cognitive Impairment classification.

**Conclusion:** Machine learning techniques according to the purpose of the study can serve experts as a low-cost and non-invasive diagnostic tool. The clinical decision support system developed in this study has been designed as a tool to assist the clinicians.

***Keywords:*** Machine Learning; Alzheimer Disease; Clinical Decision Support Systems.

## INTRODUCTION

Machine learning (ML) is a subset of Artificial Intelligence (AI) that integrates concepts from several disciplines, including computer science, statistics, and optimization, and fundamentally investigates tools and techniques for pattern recognition in data. Essentially, almost all ML problems can be formulated as optimization problems concerning a dataset. In such problems, the goal is to find a model that best describes the data. This explains the concept of "learning" in ML terminology (1).

Implementing developments in AI technologies at every stage of healthcare services offers convenience in early diagnosis, treatments, and follow-up of diseases. Dementia, a progressive brain disease, manifests with memory loss and diminishes cognitive abilities, disrupting daily functioning. With the growing elderly population, dementia has become more common worldwide and emerged as a global public health concern. Alzheimer's Disease (AD), which constitutes a large proportion of the recent progressive diseases, is a clinical syndrome that increases in prevalence with age, progresses with impairment in multiple cognitive domains, and eventually affects daily life.

According to statistics published by the Turkish Statistical Institute (2024), the proportion of the elderly population in the total population increased from 8.8% in 2018 to 10.2% in 2023. According to population projections, the proportion of elderly population is expected to become 12.9% in 2030, 16.3% in 2040, 22.6% in 2060 and 25.6% in 2080 (2). According to World Health Organization (2023) estimates, the number of patients with dementia, currently over 55 million, will reach 75 million by 2030 and 132 million by 2050. Studies show that "a new diagnosis of dementia is made every three seconds" 60-80% of which are Alzheimer's patients (3). In 2024, another study found that an estimated 6.9 million Americans aged 65 and older have AD and it could rise to 13.8 million by 2060 unless medical breakthroughs are developed to prevent or treat the disease (4).

A comprehensive approach to diagnosis entails the utilization of a multifaceted array of assessment techniques, including clinical, biochemical, and cognitive evaluation and brain imaging methods, which are inherently time-consuming and demand utmost professionalism (5). Nevertheless, for the reasons previously outlined, the present clinical status of older adults may result in relatively subtle cognitive symptoms, with cognitive problems being overlooked or associated with existing noncognitive complaints. Conversely, in certain cases, findings such as simple forgetfulness may be interpreted in a manner that favors AD due to the age of the patients, which may result in overdiagnosis and overtreatment (6). These handicaps can be eliminated and the effect of personal factors on the diagnosis or exclusion of AD can be minimized with the potential of current AI technologies to provide guidance and stimulation to healthcare professionals working under heavy workload. Furthermore, AI technologies may be invaluable for physicians regarding the dementia syndromes, which are highly prevalent in older adults. In such cases, clinical evaluation remains essential, as biomarkers may be insufficient or unavailable for early diagnosis. In recent years, the increase in the number and diversity of health data has facilitated the use of several techniques for analyzing large-scale data for early AD diagnosis. Thus, cognitive status of older adults can be accurately classified as cognitively healthy, mild cognitive impairment (MCI), or AD based on a set of ML-based classification techniques (7). Most ML studies in AD diagnosis have focused on neuroimaging data, particularly MRI or PET scans. However, access to imaging tools may be limited in many clinical settings due to cost, infrastructure, or patient-specific contraindications. Therefore, there is a growing need for reliable diagnostic approaches using easily obtainable, non-imaging data. This study addresses this gap by integrating neuropsychological assessments, laboratory findings, demographic variables and comprehensive geriatric assessment to classify AD and distinguish it from MCI and healthy aging. This study aimed to classify older adults with AD, MCI, or cognitively healthy based on risk factors and clinical features determined using classification methods.

## MATERIALS AND METHOD

### Dataset

The dataset used in the study was obtained retrospectively from the Dokuz Eylül University Hospital, Department of Geriatrics, Unit for Brain Aging and Dementia (Approval of Non-Interventional Research Ethics Committee). Within the scope of the study, the file records of all patients aged between 45 and 96 years who were followed up in the Geriatrics Clinic of the Geriatrics Department at Dokuz Eylül University Hospital were examined. The demographic data, chronic diseases, dementia diagnosis, laboratory data, and detailed geriatric assessment parameters of the patients included in the study were recorded. In other words, the independent variables included age, gender, educational status, smoking status, comorbidities, detailed geriatric assessment tests (cognition, emotion, activities of daily living) and laboratory data whereas the dependent variables comprised the presence of AD, MCI and healthy controls.

Considering an average of 1000 new case records per year applying to Outpatient Clinic of the Dokuz Eylül University Faculty of Medicine, Department of Geriatrics, and given that the prevalence of cognitive impairment in patients over 60 years of age with subjective memory complaints was 26% according to the study conducted by (8), it was decided to include at least 228 patients in the study with 5% acceptable error rate and 95% confidence level. The initial dataset comprises 176 AD cases, 73 MCI cases, and 351 healthy cases, totaling 600 cases. Table 1 presents the dataset.

Validated and commonly used instruments including the MMSE and CDR mostly formed cognitive assessment. Usually applied in both clinical and research environments, the MMSE is a quick 30-point questionnaire extensively validated for identifying cognitive impairment. Conversely, well-known in staging AD, the CDR offers a thorough rating of dementia severity by means of semi-structured interviews with both patients and carers. Strong psychometric qualities, simple administration, and regular usage in geriatric and dementia-related studies led to the selection of these tools.

As illustrated in Table 1, the dataset did not include imaging data and was ethically approved for the use of only clinical, demographic, neuropsychological and comprehensive geriatric assessment features. The phases of Cross-Industry Standard Process Model for Data Mining (CRISP-DM) (9) were followed and applied to the data in Table 1.

### Data Preprocessing

In the data preprocessing phase, the "File Number" feature, which would not be used in the analyses, was first removed from the dataset. The "Label Encoding" approach was utilized in numerical representation of categorical features. The relevant library was imported as a Python module. Consequently, the categorical features in the dataset as "present" and "absent", were coded as 0 for "present" and 1 for "absent".

Concurrently, the target variable (class) "DIAGNOSIS" was recoded as 0 for AD, 1 for MCI, and 2 for healthy subjects. Regarding the "Gender" feature, no coding was made since it was already expressed with numerical representations in the dataset. Accordingly, the "Gender" feature

**Table 1.** Dataset with Descriptions

| Feature | Description |
|---|---|
| File Number | Archive number |
| Age | Patient age |
| Gender | Gender of the patient |
| Education (year) | Years of education |
| Smoking | Smoking status |
| Family History | Whether there is a family history of dementia |
| **COMORBIDITIES** | |
| Hypert | Hypertension status |
| DMM | Diabetes status (as of patient's visit) |
| CAD | Whether the patient has coronary artery disease |
| Depression | Whether the patient has depression |
| CHF | Heart failure (Congestive Heart Failure) |
| PAD | Peripheral arterial disease (vascular disease) status |
| **LAB TESTS** | |
| VITB12 | Vitamin B |
| VITD | Vitamin D |
| TSH | Thyroid hormone |
| FOLAT | Folic acid |
| **COMPREHENSIVE GERIATRIC ASSESSMENT** | |
| **NEUROCOGNITIVE TESTS** | |
| MMSE-MOCA | Mini Mental State Examination/Montreal Cognitive Assessment Scores |
| Attention-AttMOCA | Saying words carefully |
| Orientation (time) | Score when asked for time |
| Language | Identification |
| Memory-Recall | Remembering words after a certain period of time |
| Visuospatial | Visual skills |
| Orientation (place) | Score when asked about location (where are you now?) |
| Reg-Naming | Repeating spoken words |
| Clock Drawing Test | Clock drawing test score |
| CDR | Dementia Stages -Clinical Dementia Rating Scale |
| YGDS | The Yesavage Geriatric Depression Scale |
| CORNELL | Depression in patients with dementia |
| **ACTIVITIES OF DAILY LIVING (ADL)** | |
| BASIC (BARTEL) | Activities of Daily Living Scale |
| INSTRUMENTAL (LAWTON-BRODY) | Instrumental Activities of Daily Living Scale |
| DIAGNOSIS | Diagnosis feature-consists of three classes (0: AD, 1: MCI, 2: healthy controls) |

was included in the dataset as 0 for "female" and 1 for "male". There were outlier observations in Vitamin B, Vitamin D, TSH, and Folic Acid values. At this point, the strategy to follow regarding outliers was decided upon. Outliers were included in the dataset to minimize data loss; however, the possibility that models working with outliers would provide better performance results for this dataset was also considered. Missing values were imputed using the "missingno" library, one of the Python libraries. This library was adopted because it provided information about the columns in a dataset containing missing data and enabled visualization of the general model of missing data practically. The MinMaxScaler module from the Scikit-learn library was employed in the normalization process due to the varying ranges of the dataset's features.

**Modelling**

Some models that are suitable for the classification problem were utilized in the modeling phase. In this context, the preprocessed dataset was analyzed using ML algorithms. The 5-fold cross-validation method was administered during the analyses. In 5-fold cross-validation, the entire dataset was systematically partitioned into five subsets to be used for training and testing.

Following the preprocessing steps previously outlined and the implementation of 5-fold cross-validation, the following ML models were employed.

- Logistic Regression (LR): A regression method utilized to predict two or multiple dependent variables (10).
- Multinomial LR: LR outputs two classes, whereas multinomial LR accepts multiple data classes.
- Naïve Bayes (NB): This method can estimate the probabilities of class membership, specifically the likelihood that a specific sample is associated with a certain class. The Bayesian classifier is built upon Bayes' theorem (11).
- K-Nearest Neighbor (KNN): The KNN approach classifies new unlabeled data by identifying the classes of its neighboring data points.
- Artificial Neural Networks (NN): Computational systems designed to autonomously simulate the functions of the human brain, including the derivation, creation, and discovery of new information through self-learning. The network is trained with suitable data to enable generalization. This generalization identifies output sets that may correlate to similar events (12).
- Support Vector Machine (SVM): This method delineates margins between classes. The margins seek to optimize the separation between classes, hence reducing classification error (13). SVMs aim to identify the separating hyperplane that maximizes the margin between the classes to be categorized (14).
- Decision Trees (DTs): The primary objective is to derive significant insights from extensive datasets through applying decision rules.
- Random Forest (RF): The RF algorithm (15) is an ensemble learning method that integrates several classifiers to enhance model performance. The RF approach was developed to address the issues of prolonged model performance resulting from the complexities associated with overfitting or memorizing when the number of observations in decision trees is substantial.
- XGBoost: The XGBoost algorithm is based on the "boosting" approach of ensemble learning techniques. Boosting implies that a high-performing learner can be constructed by combining ensembles with poor-performing learners, each of which surpasses the others solely by chance (16).

- LightGBM: The primary distinction between LightGBM and conventional decision tree algorithms is that the tree is expanded leaf-wise rather than evaluating all preceding leaves for each new leaf.
- CatBoost: CatBoost (17) is a combination of the terms "Category" and "Boosting". It is a more sophisticated variant of gradient boosting-based decision trees that can effectively handle categorical variables within a dataset.
- ChefBoost: ChefBoost (18) is a basic decision tree framework for Python that supports categorical features. This method clearly illustrates the decisions taken by the tree to reach a specific prediction and supports Explainable Artificial Intelligence (XAI).

Ensemble methods were subjected to both manual and automated hyperparameter optimization strategies to guarantee best model performance. GridSearchCV and RandomizedSearchCV were used to manually tune combinations of important parameters including learning_rate, max_depth, n_estimators. Apart from manual tuning, AutoML-based systems—specifically Optuna, Hyperopt, and TPOT—were employed to quickly investigate a larger hyperparameter area. These techniques find high-performing configurations using sophisticated search techniques including genetic programming and Bayesian optimization.

**Clinical Decision Support Systems Prototype (Deployment)**

There are different and various risk assessment methods for AD. Currently, data-driven/data-informed "Clinical Decision Support Systems (CDSS)" automate these assessments and assist clinicians in their decision-making processes. In the most general sense, CDSS are computer systems to assist the clinician in patient-related decisions such as diagnosis and treatment (19). CDSS is also defined as "software designed to directly assist the clinical knowledge base with patient-specific assessments or recommendations, presented to the clinician or patient to make a decision (20). These systems, which were previously designed as rule-based, are currently handled within a ML-based approach thanks to the increase in health data. Therefore, systems trained from existing data and offer decision support accordingly can detect undiscovered patterns in the diagnosis of a disease. In the case of AD, a ML-based decision support system prototype has been presented within the scope of the study using the data described in Table 1.

The prototype for the CDSS was developed using Flask, a microframework designed in the Python programming language. Initially, the CatBoost model, which had been previously trained and achieved the highest score (refer to Results for details), was integrated into the system using the Flask structure. The model was saved as *predMed.pkl*, and the test data was evaluated on the Web utilizing Flask technology. For this purpose, an interface was developed, and the trained model was activated by clicking the "Analyze Cognitive Status" button after inputting the value of each feature in the dataset into the system.

Figure 1 presents a sample data entry. As previously stated, the system defines 0 for "present" and 1 for "absent" values.

At this stage, validation procedures were also conducted. The objective was to verify whether the outcomes from the expert physician aligned with those generated by the developed prototype. For instance, experiments were conducted on a patient diagnosed with AD by a specialist physician, and it was observed that the same result was supported by the system.

Although a formal usability study did not test the CDSS interface with end users, a small-scale consistency check was carried out to compare the system's output with the decisions made by clinical experts. In certain cases, this first step helped the

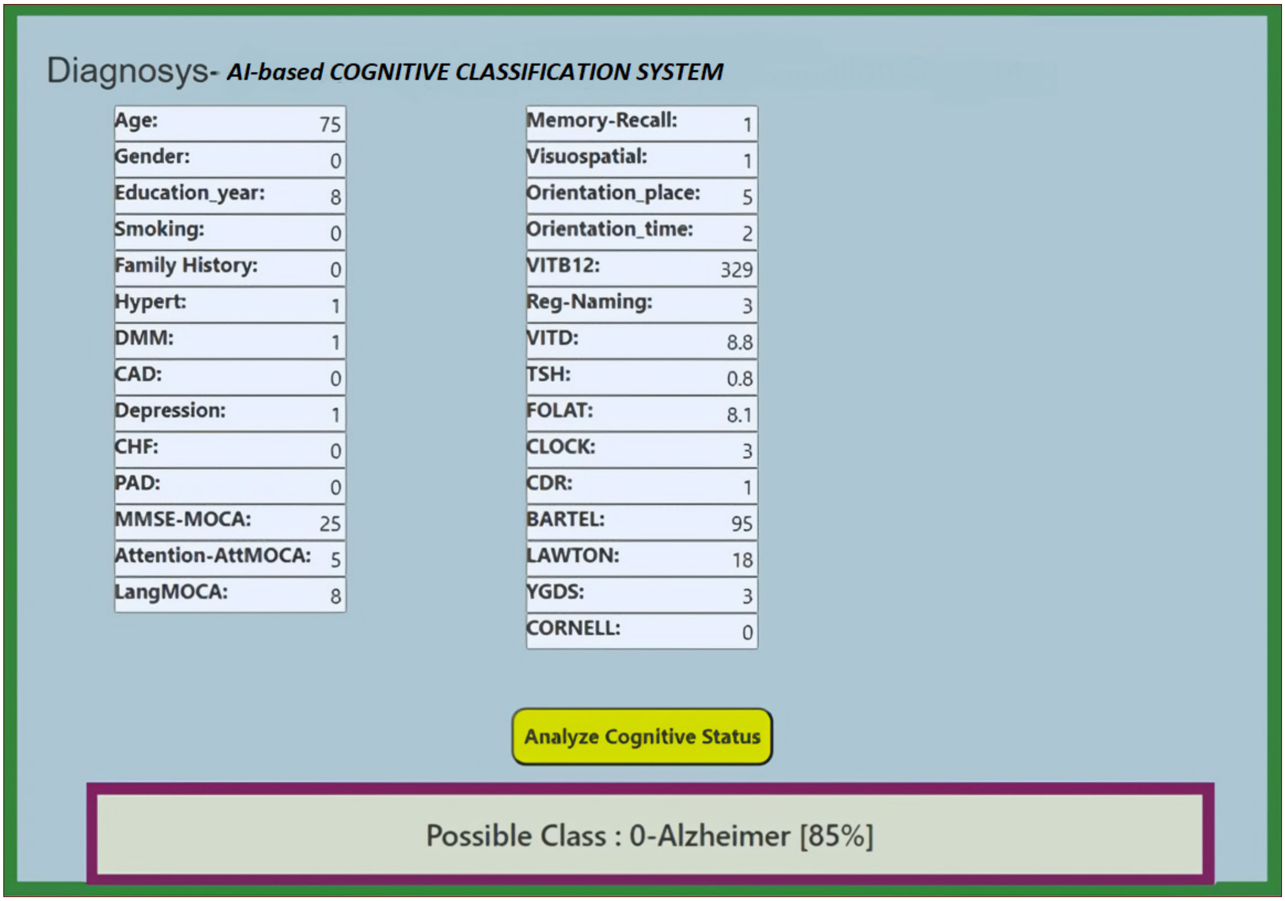


**Figure 1.** User Interface of the Prototype Clinical Decision Support System

model match expert judgment. As shown in Figure 1, the user interface allows input of patient data and displays the predicted cognitive class. In order to improve the system's clinical integration, future research might concentrate on thorough usability assessments conducted with medical experts.

## RESULTS

This part presents the findings and interpretations derived from the application of several ML methods outlined in the "Modeling" section to the dataset. In overall accuracy, CatBoost algorithm (with hyperparameter optimization) demonstrated the highest performance (87.5%).

This was particularly advantageous in health datasets where domain knowledge might be scarce and hyperparameter optimization may provide challenges. In addition, precision, recall (TPR/sensitivity), and F1-Score were presented for each class specifically.

Table 2 illustrates the performance metrics for the AD classification. It can be asserted that ensemble learning methods outperform traditional methods for 176 samples in the AD class. Upon detailed analysis of Table 2, the XGBoost (fine-tuned) and CatBoost approaches were observed to exhibit the maximum recall. In the context of AD classification, recall quantifies a classification model's or diagnostic test's capacity to correctly

**Table 2.** Precision, Recall, and F1-Score for AD Class

| Algorithm / Model | Precision (PPV) | Recall (TP rate) | F1-Score |
|---|---|---|---|
| LR | 0.83 | 0.86 | 0.84 |
| NB | 0.86 | 0.72 | 0.79 |
| KNN | 0.81 | 0.78 | 0.79 |
| MLP | 0.78 | 0.82 | 0.8 |
| SVM | 0.79 | 0.83 | 0.81 |
| DTs | 0.73 | 0.72 | 0.73 |
| RF | 0.78 | 0.84 | 0.81 |
| RF (fine-tuned) | 0.78 | 0.93 | 0.85 |
| XGBoost (fine-tuned) | 0.71 | 1.00 | 0.83 |
| LightGBM | 0.7 | 0.97 | 0.81 |
| CatBoost | 0.74 | 1.00 | 0.85 |
| ChefBoost | 0.81 | 0.69 | 0.63 |

**Table 3.** Precision, Recall, and F1-Score for MCI Class

| Algorithm / Model | Precision (PPV) | Recall (TP rate) | F1-Score |
|---|---|---|---|
| LR | 0.65 | 0.55 | 0.60 |
| NB | 0.33 | 0.56 | 0.41 |
| KNN | 0.20 | 0.04 | 0.07 |
| MLP | 0.70 | 0.63 | 0.67 |
| SVM | 0.74 | 0.49 | 0.59 |
| DTs | 0.30 | 0.30 | 0.30 |
| RF | 0.75 | 0.18 | 0.29 |
| RF (fine-tuned) | 1.00 | 0.12 | 0.21 |
| XGBoost (fine-tuned) | 0.40 | 0.18 | 0.25 |
| LightGBM | 0.40 | 0.18 | 0.25 |
| CatBoost | 0.50 | 0.18 | 0.27 |
| ChefBoost | 0.14 | 0.20 | 0.17 |

identify all actual cases of AD, emphasizing the reduction of false negatives, wherein individuals with Alzheimer's are erroneously classified as non-AD.

The MCI class had the smallest sample size within the dataset; however, it could be regarded as an intermediate class. Diagnosing MCI is a challenging task, even for specialist clinicians, and marks the onset of cognitive decline, as well as the increased risk of developing AD. There are 73 instances of MCI in the sample. Analysis of Table 3 revealed that the MCI classification exhibits lower performance compared to the AD classification. This was mainly due to the insufficient quantity of data. As illustrated in Table 3, the RF method exhibited the highest precision value for MCI classification.

**Table 4.** Precision, Recall, and F1-Score for Healthy Subjects Class

| Algorithm / Model | Precision | Recall (TP rate) | F1-Score |
|---|---|---|---|
| LR | 0.88 | 0.95 | 0.91 |
| NB | 0.91 | 0.85 | 0.88 |
| KNN | 0.80 | 0.95 | 0.87 |
| MLP | 0.86 | 0.93 | 0.89 |
| SVM | 0.92 | 0.95 | 0.93 |
| DTs | 0.84 | 0.85 | 0.85 |
| RF | 0.84 | 0.98 | 0.91 |
| RF (fine-tuned) | 0.87 | 0.97 | 0.91 |
| XGBoost (fine-tuned) | 0.96 | 0.89 | 0.92 |
| LightGBM | 0.96 | 0.90 | 0.93 |
| CatBoost | 0.96 | 0.93 | 0.94 |
| ChefBoost | 0.85 | 0.81 | 0.83 |

Table 4 illustrates the classification outcomes for the "Healthy" class. Healthy cases comprised 351 instances within the sample. In this setting, analysis revealed that the healthy class attained better results in both statistical methods and more sophisticated ensemble learning models. The substantial volume of data was the primary explanation behind these outcomes.

## DISCUSSION

The first notable result of the study was that the performance of the CatBoost model after hyperparameter optimization was observed as the highest compared to the other models, considering the accuracy rate. This finding is consistent with existing literature, which shows that RF or XGBoost performed better in clinical assessment (21,22).

The combined use of manual and AutoML-based hyperparameter tuning (including Bayesian and evolutionary strategies) contributed to the model's robustness and adaptability. This multi-pronged optimization approach may be especially valuable for clinical datasets where fine-tuning directly impacts diagnostic reliability.

In particular, as in the problem at hand, overfitting becomes a major concern due to the often small, imbalanced, and noisy nature of health datasets. The regularization mechanisms of CatBoost, including depth control and learning rate optimization, mitigate this issue. Another factor contributing to CatBoost's efficacy in the examined problem was the presence of missing values within the dataset, a common characteristic of health data. The CatBoost algorithm employs distinctive mechanisms to address missing values without necessitating imputation. It offers competitive performance relative to other algorithms with minimum hyperparameter tuning. In this sense, CatBoost's ability to perform high-scoring classification can also be attributed to its robustness to outliers. Health-related datasets can contain outliers that can affect model performance. Therefore, the robustness of CatBoost to outliers can be advantageous in these scenarios. CatBoost can effectively leverage multi-core CPUs and GPUs for accelerated training, which is essential for managing extensive health datasets.

The dataset exhibited class imbalance (AD: 176, MCI:73 HC: 351) among groups. To address this, class weights were adjusted during model training to penalize misclassification of the minority class. In addition, model performance was evaluated using F1-Score which are more reliable in imbalanced classification problems. As a result, the CatBoost model was able to distinguish between healthy and AD class more efficiently.

RF method exhibited the highest precision value for MCI classification. The emphasis is on accurately identifying actual positive MCI cases. In other words, it assesses the model's efficiency in differentiating between cases that genuinely have MCI and those that do not. A high precision score indicates that the model effectively identifies individuals with MCI while maintaining a low false positive rate. In medical practice, it is crucial to reduce the probability of misdiagnosing an individual with MCI when they are not affected, as this may result in unnecessary stress and redundant testing.

Accurate differentiation of the healthy cohort is critical for research and therapeutic trials. This enables researchers to enhance their comprehension of AD progression and the efficacy of prospective treatments. Moreover, the recognition of healthy people can aid in pinpointing potential risk factors for AD by concentrating on their cognitive health. This may facilitate a deeper comprehension of the disease's etiology and possible prevention strategies. Consequently, the primary contribution of identifying healthy individuals would be to understand and, if necessary, redefine "healthy cognition" to uncover the deviations associated with AD.

While the developed model was able to differentiate the healthy and AD groups with high performance, it showed poorer performance in differentiating the MCI class from the healthy class. This result was also supported by another study (23). It was suggested that the reason for the relatively lower performance for the MCI class may be the smaller amount of data and, more importantly, the fact that the features defined for MCI but not for the other two classes (AD/HC) were not included in the dataset. At this point, the experiments conducted by increasing the number of data and the performance for the MCI class did not increase support this view.

Increasing the number of data, repeating the training process by feeding data to the model at certain intervals and examining the model performance, adding image data obtained as a result of brain imaging methods to the dataset, and combining different data types (such as using multimodal data) could increase the accuracy of the model.

Correlation analysis was performed to examine the relationships between key demographic and neuropsychological variables such as age, education, and test scores. Although these insights informed the understanding of feature interactions, explicit modeling of interaction effects was not implemented in the final CDSS model due to concerns related to model complexity and overfitting. Future studies could benefit from integrating feature interactions to enhance predictive accuracy and system interpretability.

When the study is considered within the scope of original value, the use of ML techniques (particularly, deep learning, ensemble learning, etc.) based on the purpose of the study can be made available to experts as a low-cost and non-invasive diagnostic tool. As far as examined, the literature contains no active decision support system.

In other words, it was observed that limited studies, especially studies using data based on neuropsychological and geriatric assessment, lacked the decision support part of deployment, especially when considered from the perspective of management information systems. In this sense, as the study suggested, an approach for automatic detection of Alzheimer's type dementia with ML-based methods can contribute to literature considering the diagnosis, treatment, and cost of AD.

The developed CDSS was designed as a tool to assist the clinicians. In other words, AI technologies are not expected to replace the physicians, but to improve the decision process within the framework of human-AI collaboration.

## LIMITATIONS

The most significant limitation of this study is the absence of neuroimaging data, which may enhance diagnostic performance when combined with clinical and cognitive features. However, the dataset used in this research was ethically approved for the use of only structured non-imaging data, including clinical, demographic, neuropsychological and comprehensive geriatric assessment features. Future studies may benefit from incorporating multimodal data, including neuroimaging, to further improve the predictive power and generalizability of ML models in AD classification.

While the interaction between age, education, and MMSE scores was not explicitly modeled, a partial dependence analysis conducted during the development phase showed that age had a relatively limited marginal effect on model predictions. Nevertheless, it is widely acknowledged in the literature that age and education level significantly influence MMSE scores.

Although interaction-based interpretability methods such as SHAP (SHapley Additive exPlanations) and PDP (Partial Dependence Plots) were not elaborated in this paper, they were explored in detail in a complementary study focused on the same dataset (24). Different XAI methods for comprehensive interaction analysis between these factors should be considered in future studies.

The application developed within the scope of the study can be characterized as context-specific research; therefore, its generalizability is limited. The most important reason is modeling of the decision-making process of two specialist physicians as decision makers. It is believed that the more expert opinions are included in the decision-making process, the higher the generalizability.

## CONCLUSION

In this age with paramount data, the healthcare sector experiences a significant data revolution. In recent years, advancements in AI within healthcare have been accelerated by hardware and software solutions that enhance data collection. The healthcare ecosystem comprises several distinct stakeholders, including patients and their families, clinical care teams, public health program managers, hospital administrators, and researchers. Patients constitute a fundamental component of this ecosystem. They consistently produce data and transmit it to various applications.

The use of ML methods in the early diagnosis of neurodegenerative diseases like AD has attracted the attention of both researchers and clinicians. As a subfield of AI, ML methods can detect AD in advance. These approaches possess the capability to forecast MCI during the progression to AD.

The study aimed to classify patients as healthy, having MCI, or diagnosed with AD through classification techniques. Given the group's considerable heterogeneity and sensitivity to external influences, it is anticipated that employing ML to derive uniform outcomes from this data, particularly in forecasting the impact of these external factors on the results, will enhance the originality of the study.

The CDSS is positioned as an aid to expert clinicians. (25) supports this view by stating that AI can contribute to a more efficient and modern healthcare system by redesigning roles in the healthcare ecosystem, and human-in-the-loop highlights a collaborative partnership between AI and human expertise to optimize outcomes.

Given the proposed system will form an infrastructure for further research, the potential to facilitate the follow-up of individuals with AD is another advantage of the study. It is thought that the system proposed within the study can help detect the disease before it progresses, slow down the course of the disease by focusing on modifiable risk factors in the early period, or apply existing treatment approaches.

In this context, the suggested and planned studies can be summarized as follows:

- Especially in the context of AD, the datasets available in Türkiye are quite limited. Considering the cultural effects of the disease, studies should be carried out to create a dataset.
- Results with data from different hospitals should be compared.
- Since the health status is dynamic and changing over time, especially when neurodegenerative diseases are considered, it is planned to obtain the data of the same patient at different time intervals to monitor the course of the disease more effectively.

**Acknowledgments:** The authors extend their sincere gratitude to Prof. Ahmet Turan Işık, Prof. Derya Kaya, and Dr. Kübra Altunkalem Seydi, a specialist in geriatrics, for generously sharing their invaluable expertise in the field of geriatrics, which has greatly enriched this study.

## REFERENCES

1. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2017 Aug; 66(1):149-153. doi: 10.1093/cid/cix731.
2. Turkish Statistical Institute. Elderly Statistics; 2024. [Internet]. Available from: https://data.tuik.gov.tr/Bulten/Index?p=Istatistiklerle-Yaslilar-2023-53710. Accessed: 05.03.2025.
3. World Health Organization Dementia Key Facts; 2023. [Internet]. Available from: https://www.who.int/news-room/fact-sheets/detail/dementia.Accessed: 05.03.2025.
4. Alzheimer's Association. 2024 Alzheimer's disease facts and figures. Alzheimers Dement. 2024; 20(5):3708-3821. doi: 10.1002/alz.13809.
5. Isik AT. Late onset Alzheimer's disease in older people. Clin Interv Aging. 2010 Oct; 5:307–311. doi: 10.2147/CIA.S11718
6. Idil E, Aydin AE, Ates Bulut E, Isik AT. Rationally decreasing the number of drugs seems to be a useful therapeutic approach in older adults: 6-month follow-up study. Arch Gerontol Geriatr. 2021 Sep; 96. doi: 10.1016/j.archger.2021.104472.
7. Weakley A, Williams JA, Schmitter-Edgecombe M, Cook DJ. Neuropsychological test selection for cognitive impairment classification: a machine learning approach. J Clin Exp Neuropsychol. 2015 Mar;37(9):899-916. doi: 10.1080/13803395.2015.1067290.
8. Işık AT, Soysal P, Kaya D, Usarel C. Triple test, a diagnostic observation, can detect cognitive impairment in older adults. Psychogeriatrics. 2018 Mar; 18(2):98-105. doi: 10.1111/psyg.12289.
9. Shearer C. The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing. 2000 ;5(4):13-22.
10. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 2nd edition. Hoboken: John Wiley & Sons; 2013.
11. Web, GI, Keogh E, Miikkulainen R. Naïve bayes. Encyclopedia of Machine Learning, 2016 Jan; 15(1), 713-714. doi: : 10.1007/978-1-4899-7502-7_581-1
12. Öztemel, E. (2016). Artificial Neural Networks (in Turkish). İstanbul: Papatya Publishing.
13. Mahesh B. Machine learning algorithms-a review. International Journal of Science and Research. 2019 Jan; 9(1):381-386. doi: 10.21275/ART20203995.
14. Metlek S, Kayaalp K. Detection of autistic spectrum disorder with machine learning algorithms. Journal of Intelligent Systems: Theory and Applications. 2020 Sep; 3(2):60-68. doi: 0.38016/jista.755481.
15. Breiman L. Random forests. Machine Learning. 2001 Oct; 45:5-32. doi: 10.1023/A:1010933404324.
16. Marsland S. Machine learning: an algorithmic perspective. Boca Raton: Chapman and Hall/CRC; 2011.
17. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds). Advances in Neural Information Processing Systems 31 (NeurIPS 2018). 2018; 31. pp 6638-6648.

18. Serengil SI. Chefboost: a lightweight boosted decision tree framework. Computer Software. 2021 Oct. 1-10. doi.org/10.5281.

19. Berner ES, La Lande TJ. Overview of Clinical Decision Support Systems., In: Eta S. Berner (Eds) Vol. 233. Clinical Decision Support Systems. 2nd edition, Springer Science+Business Media, New York; 2007, pp: 3-23.

20. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, Tang PC. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc. 2001 Nov; 8(6):527-534. doi: 10.1136/jamia.2001.0080527.

21. Gupta A, Kahali B. Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests. Alzheimers Dement (NY). 2020 Jul ;6(1): 1-10. doi: 10.1002/trc2.12049.

22. Buyrukoğlu S. Early detection of Alzheimer's disease using data mining: comparison of ensemble feature selection approaches. Konya Journal of Engineering Sciences. 2021 Sep ;9(1):50-61. doi: 10.36306/konjes.731624.

23. Pellegrini E, Ballerini L, Hernandez MDCV et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. Alzheimers Dement (Amst). 2018 Aug; 10:519-535. doi: 10.1016/j.dadm.2018.07.004.

24. Ünal, C. (2025). Explainable Artificial Intelligence Approaches: Applications for Alzheimer's Disease Detection (in Turkish). Ankara: Gazi Publishing.

25. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digit Health. 2023 Jul; 9: 1-5. doi: 10.1177/2055207623118.