International Journal of Academic Research and Development
www.allstudiesjournal.com
ISSN: 2455-4197
Received: 04-01-2025, Accepted: 03-02-2025, Published: 18-02-2025
Volume 10, Issue 1, 2025, Page No. 25-27

Comparative analysis of tree-based machine learning models for early dementia detection

Dr. Chizoba Ezeaku-Ezeme
Department of Data Science, Leeds Beckett University, Leeds, England

Abstract
Early dementia detection is critical for timely intervention and improving patients' quality of life. Machine learning (ML) has emerged as a promising approach to enhance diagnostic accuracy. This study compares the performance of two prominent tree-based ML models—Random Forest (RF) and Decision Tree (DT)—in detecting early-stage dementia. Utilizing a publicly available dataset featuring demographic, genetic, and cognitive variables, the models were evaluated based on accuracy, recall, precision, and F1 scores. The results demonstrate that both models perform exceptionally well, with RF achieving marginally higher metrics due to its ensemble nature. The study underscores the potential of tree-based models as robust tools for early dementia prediction.

Keywords: Dementia detection, machine learning, random forest, decision tree, early diagnosis, tree-based models, cognitive health

Introduction

Dementia, a progressive neurological disorder, is characterized by a decline in cognitive functions such as memory, reasoning, and decision-making. This condition not only affects individuals' ability to perform daily activities but also imposes a significant emotional and financial burden on families and healthcare systems. According to the World Health Organization (WHO), by 2050, the global prevalence of dementia is projected to reach 135 million, highlighting the urgent need for effective diagnostic and management strategies.

Early diagnosis is crucial for mitigating the progression of dementia and improving the quality of life for affected individuals. Interventions initiated during the early stages can delay the onset of severe symptoms, enhance patient care, and reduce healthcare costs. However, conventional diagnostic methods, such as neuroimaging and cognitive assessments, have several limitations. These techniques are often expensive, invasive, and inaccessible to many populations, particularly in resource-constrained settings. Moreover, these methods rely heavily on clinical expertise, which can lead to variability in diagnosis accuracy.

In recent years, advances in machine learning (ML) have provided new opportunities to address the challenges associated with dementia diagnosis. ML algorithms can analyze large, complex datasets to identify subtle patterns and correlations that may be indicative of early-stage dementia. Among these techniques, tree-based models, such as Random Forest (RF) and Decision Tree (DT), have gained prominence due to their ability to handle diverse data types, interpretability, and high classification accuracy. These models are particularly suitable for healthcare applications, where interpretability and reliability are critical.

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve classification accuracy and reduce overfitting. This approach leverages the strengths of individual decision trees while mitigating their weaknesses, making it robust for handling noisy and imbalanced datasets. Decision Tree, on the other hand, is a simpler model that uses a tree-like structure to make decisions based on feature splits. While DT is highly interpretable and easy to implement, it is prone to overfitting and may not perform well on complex datasets without careful tuning.

This paper focuses on comparing the performance of RF and DT models in early dementia detection. By analyzing their effectiveness across various evaluation metrics, this study aims to provide insights into their applicability in clinical and research settings. The findings will help healthcare professionals and researchers select the most appropriate model for their specific needs, ultimately contributing to the development of more accurate and accessible diagnostic tools for dementia.

The Growing Burden of Dementia

Dementia is a major public health concern, with significant implications for individuals, families, and society. The aging population is one of the primary drivers of the increasing prevalence of dementia. According to demographic projections, the number of people aged 60 and above will double by 2050, leading to a corresponding rise in dementia cases. In addition to aging, factors such as genetics, lifestyle, and comorbidities contribute to the risk of developing dementia. Alzheimer's disease, the most common form of dementia, accounts for approximately 60-70% of cases, followed by vascular dementia, Lewy body dementia, and frontotemporal dementia.

The economic impact of dementia is staggering. In 2019, the global cost of dementia care was estimated at $1 trillion, and this figure is expected to double by 2030. These costs include direct medical expenses, informal caregiving, and lost productivity. Moreover, the emotional toll on caregivers is immense, with many experiencing stress, depression, and burnout. Given the scale of the problem, there is an urgent need for innovative solutions to improve dementia diagnosis and management.

Challenges in Early Diagnosis

The early stages of dementia are often subtle and easily overlooked. Symptoms such as mild memory loss, difficulty concentrating, and changes in mood or behavior may be attributed to normal aging or other conditions. This diagnostic ambiguity delays intervention and limits the

effectiveness of treatments. Conventional diagnostic methods, while useful, are not without limitations:

1. **Neuroimaging Techniques:** Methods such as magnetic resonance imaging (MRI) and positron emission tomography (PET) are used to detect structural and functional brain changes associated with dementia. However, these techniques are expensive, require specialized equipment, and are not widely available.

2. **Cognitive Assessments:** Tools like the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) are commonly used to evaluate cognitive function. While these tests are non-invasive and relatively inexpensive, their accuracy depends on the skill of the administrator and the patient's cooperation.

3. **Biomarker Analysis:** The detection of biomarkers such as amyloid-beta and tau proteins in cerebrospinal fluid (CSF) is another approach to diagnosing dementia. This method, however, involves invasive procedures like lumbar punctures, which may not be feasible or acceptable to all patients.

4. **Clinical Expertise:** The reliance on clinicians for interpreting diagnostic results introduces subjectivity and variability, leading to potential misdiagnosis or delayed diagnosis.

The Role of Machine Learning in Dementia Diagnosis

Machine learning offers a transformative approach to dementia diagnosis by leveraging data-driven algorithms to identify patterns and make predictions. Unlike traditional methods, ML models can analyze heterogeneous data, including demographic information, genetic markers, cognitive scores, and lifestyle factors. This ability to integrate and process diverse data sources enhances the accuracy and reliability of predictions.

Tree-based ML models, in particular, are well-suited for healthcare applications. Their hierarchical structure allows them to capture non-linear relationships between features, making them effective for complex classification tasks. Additionally, the interpretability of these models enables clinicians to understand the rationale behind predictions, fostering trust and acceptance in clinical practice.

The advantages of tree-based models include:

- **Robustness:** RF and DT models can handle missing data and noisy features without significant loss of accuracy.

- **Scalability:** These models are computationally efficient and can be scaled to analyze large datasets.

- **Feature Importance:** Both RF and DT provide insights into feature importance, helping researchers identify key predictors of dementia.

However, challenges remain. RF models, while accurate, can be less interpretable due to their ensemble nature. DT models, though interpretable, may require extensive tuning to prevent overfitting and ensure generalizability. This study aims to address these challenges by systematically comparing the performance of RF and DT models, providing practical recommendations for their use in dementia diagnosis.

Literature Review

Tree-based ML models have been extensively utilized in healthcare research for their ability to handle complex datasets and provide interpretable outputs. Studies have demonstrated the effectiveness of DTs in early dementia prediction by analyzing cognitive scores, genetic markers, and lifestyle factors. However, DTs are prone to overfitting, especially when datasets are small or noisy.

RF, as an ensemble learning technique, addresses these limitations by aggregating the predictions of multiple decision trees. This approach reduces variance and improves generalization. Previous research has highlighted RF's superior performance in dementia detection, with high accuracy and robustness against outliers. Despite these advantages, RF's interpretability is comparatively lower than that of single DT models, which may limit its adoption in clinical settings.

Recent advancements in feature selection and preprocessing techniques have further enhanced the capabilities of tree-based models. Studies have identified cognitive test scores, genetic markers (e.g., APOE ε4 allele), and lifestyle variables as significant predictors of dementia, providing a foundation for building reliable predictive models.

Materials and Methods

Dataset

The dataset used for this study was obtained from Kaggle and included demographic, cognitive, genetic, and lifestyle variables. Key features were:

- **Continuous variables:** Cognitive test scores, age, alcohol level, blood oxygen level, and body temperature.

- **Categorical variables:** Gender, family history, education level, APOE ε4 status, and smoking status.

The target variable was binary, indicating the presence or absence of early-stage dementia.

Data Preprocessing

1. **Handling Missing Values:** Missing data were imputed using mean or mode values based on feature type.

2. **Normalization:** Continuous variables were normalized to ensure consistency across features.

3. **Feature Selection:** Correlation analysis and feature importance metrics were used to identify the most predictive variables.

4. **Train-Test Split:** The dataset was split into training (80%) and testing (20%) subsets using stratified sampling to maintain class balance.

Model Implementation

Two tree-based models were implemented using Python's Scikit-learn library:

- **Decision Tree (DT):** A single-tree model optimized using grid search for parameters such as maximum depth and minimum samples per split.

- **Random Forest (RF):** An ensemble of 100 decision trees, tuned for the number of estimators and maximum features.

Evaluation Metrics

The models were evaluated using:

- **Accuracy:** Proportion of correctly predicted instances.
- **Precision:** Proportion of true positives among predicted positives.
- **Recall:** Proportion of true positives among actual positives.
- **F1 Score:** Harmonic mean of precision and recall.
- **ROC-AUC:** Area under the receiver operating characteristic curve.

Results and Discussion

Model Performance

Random Forest (RF):

- Accuracy: 98%
- Precision: 0.99
- Recall: 0.98
- F1 Score: 0.99
- ROC-AUC: 0.99

Decision Tree (DT):

- Accuracy: 95%
- Precision: 0.96
- Recall: 0.94
- F1 Score: 0.95
- ROC-AUC: 0.96

RF outperformed DT across all metrics, demonstrating better generalization and reduced overfitting. The ensemble nature of RF allowed it to effectively capture complex patterns in the dataset, making it more reliable for clinical applications.

Feature Importance

Cognitive test scores emerged as the most significant predictor, followed by APOE ε4 status and depression indicators. Lifestyle variables like smoking and physical activity had moderate importance, while demographic factors such as gender and education level showed minimal impact.

Interpretability vs. Performance

While RF provided higher accuracy and robustness, its complexity limited interpretability compared to DT. This trade-off highlights the need to balance performance and usability when selecting models for clinical use.

Conclusion

This study demonstrates the efficacy of tree-based ML models in early dementia detection, with RF outperforming DT in terms of accuracy and robustness. However, DT's simplicity and interpretability make it a viable option for scenarios requiring straightforward decision-making. Future research should focus on integrating larger, more diverse datasets and exploring hybrid approaches to enhance model performance and generalization.

References

1. Dallora AL, *et al*. Decision tree analysis for dementia prediction, 2020.
2. Javeed A, *et al*. Optimized SVM for dementia detection, 2023.
3. Prince M, *et al*. Global prevalence of dementia, 2015.
4. Vrijsen J, *et al*. Predictors of dementia: A systematic review, 2021.
5. Yu JT, *et al*. Risk factors for Alzheimer's disease, 2020.
6. Scikit-learn documentation: https://scikit-learn.org/