CARITAS UNIVERSITY AMORJI-NIKE, EMENE, ENUGU STATE

Caritas Journal of Physical and Life Sciences

CJPLS, Volume 3, Issue 2 (2024)

Article History: Received: 12th October, 2024 Revised: 4th December, 2024 Accepted: 10th December, 2024

Early Prediction of Dementia Using Machine Learning

Chizoba Nneka Ezeaku-Ezeme
Department of Data Science, Leeds Beckett University. UK
C.Ezeaku-Ezeme2975@student.leedsbeckett.ac.uk, chizzyobyno@gmail.com;

Omankwu, Obinnaya Chinecherem Beloved
Department of Computer Science, Michael Okpara University of Agriculture,. Umudike
saintbeloved@yahoo.com

Abstract

This study explores the application of machine learning algorithms for the early prediction of dementia, aiming to improve diagnostic accuracy and reliability. Utilizing a comprehensive dataset from Kaggle, which includes both continuous and categorical variables, four machine learning models—Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine (SVM)—were implemented and evaluated. The study identifies cognitive test scores, the APOE ε4 allele, and depression status as key predictors of dementia. Tree-based models demonstrated superior performance, achieving perfect scores across metrics such as accuracy, recall, precision, and F1. Despite these promising results, the study acknowledges limitations such as the reliance on a single dataset, limited predictors, and challenges in real-world validation. Future research should incorporate larger, more diverse datasets, longitudinal data, and additional predictors to improve model robustness and applicability. These findings highlight the potential of machine learning as a transformative tool in clinical settings for timely dementia diagnosis and intervention.

Keywords Dementia, Machine Learning, Early Prediction, Cognitive Tests, Tree-based Models

Introduction

Dementia is a progressive neurological condition characterized by cognitive decline that impairs daily functioning. The global prevalence of dementia is rising, with projections suggesting over 135 million cases by 2050. Early detection is critical to managing the disease, delaying its progression, and improving patient outcomes. Traditional diagnostic methods, including cognitive tests and neuroimaging, have limitations such as invasiveness, cost, and time consumption. For example, magnetic resonance imaging (MRI) and computed tomography (CT) scans are effective but expensive and not always accessible, particularly in resource-limited settings. Additionally, cognitive tests often rely on subjective evaluation, which can lead to inconsistent results depending on the assessor's expertise.

Recent advancements in machine learning (ML) provide innovative, non-invasive alternatives for early dementia detection. ML algorithms can process vast amounts of data to identify subtle patterns that human observation might miss, making it possible to predict dementia earlier and more accurately. The integration of ML in clinical diagnostics has the potential to revolutionize the field by offering cost-effective, scalable solutions.

This study focuses on implementing and evaluating various ML algorithms to identify the most effective techniques for early dementia prediction. By leveraging a comprehensive dataset, this research aims to explore key predictors of dementia, evaluate the performance of different models, and propose recommendations for future research and clinical applications.

Dementia imposes significant economic, psychological, and social burdens on patients and their families. According to the World Health Organization (WHO), the annual global cost of dementia reached $1 trillion in 2021, with projections suggesting this figure will double by 2030. Caregivers often experience emotional stress and financial strain, highlighting the urgent need for early diagnostic tools that can mitigate these challenges.

Early detection of dementia is crucial but fraught with challenges. Symptoms of dementia, such as memory loss and cognitive impairment, often overlap with normal aging processes or other medical conditions, making diagnosis difficult. Furthermore, cultural and socioeconomic factors influence access to healthcare and diagnostic services, exacerbating disparities in dementia care.

Machine learning offers a promising solution to these challenges. By analyzing large, diverse datasets, ML algorithms can uncover complex relationships between genetic, behavioral, and environmental factors, enabling more accurate and timely predictions. This study investigates these capabilities, focusing on the practical implementation of ML models in dementia prediction.

Objectives of the Study

The primary objectives of this research are to:

1. Identify key predictors of dementia using machine learning.
2. Evaluate the performance of various ML models in predicting early-stage dementia.
3. Provide insights into the practical applications of ML in clinical settings.
4. Highlight areas for future research to improve diagnostic accuracy and accessibility.

This study is limited to analyzing a single dataset sourced from Kaggle, which may not capture the full diversity of patient populations. Additionally, the research focuses on four ML models—Random Forest, Decision Tree, Logistic Regression, and SVM—and does not explore newer deep learning techniques. Despite these limitations, the findings provide valuable insights into the potential of ML in dementia prediction and lay the groundwork for future studies.

Literature Review

Machine learning has emerged as a powerful tool for medical diagnostics, including dementia prediction. Studies have demonstrated its ability to handle high-dimensional data and identify subtle patterns that traditional methods may overlook.

Dallora et al. (2020) used decision trees for a ten-year prognosis of dementia, achieving an AUC of 74.5%. Their research emphasized the importance of feature selection, which enhances model performance by identifying the most informative variables. Similarly, Javeed et al. (2023) developed an optimized SVM model that outperformed previous approaches in accuracy and reliability. The study utilized adaptive synthetic sampling to address class imbalance, a common issue in medical datasets, highlighting the effectiveness of preprocessing techniques in improving model outcomes.

Another significant contribution comes from Yu et al. (2020), who conducted a meta-analysis to identify modifiable risk factors for Alzheimer's disease, the most common form of dementia. Their findings underscore the role of lifestyle factors, such as diet and physical activity, in dementia prevention. However, integrating these factors into ML models remains a challenge due to data variability and measurement inconsistencies.

Recent advancements in deep learning have also shown promise. For instance, Battineni et al. (2020) applied convolutional neural networks (CNNs) to MRI scans, achieving high accuracy in Alzheimer's disease classification. While CNNs excel in image analysis, their computational complexity and data requirements limit their applicability in resource-constrained settings. This gap underscores the need for simpler, more interpretable models like decision trees and logistic regression in certain contexts.

The APOE ε4 allele has been extensively studied as a genetic marker for Alzheimer's disease. Studies by Vrijsen et al. (2021) and Shahzad et al. (2022) confirmed its strong association with dementia risk. However, the reliance on genetic data raises ethical and privacy concerns, particularly when integrating such information into predictive models.

Despite these advancements, significant gaps remain in the literature. Most studies focus on a single type of data, such as genetic markers or neuroimaging, neglecting the multifactorial nature of dementia. Additionally, the interpretability of complex models, such as deep learning algorithms, is a persistent challenge. Clinicians often require transparent models that provide actionable insights, which are not always available in black-box approaches.

Knowledge Gaps and Future Directions

One major gap in the current research is the lack of longitudinal studies. Most ML models are trained on cross-sectional data, which limits their ability to capture disease progression over time. Longitudinal datasets could provide valuable insights into how dementia develops, enabling the creation of more robust predictive models.

Another limitation is the underrepresentation of diverse populations in dementia research. Many studies use datasets from high-income countries, which may not be generalizable to low- and middle-income settings. This disparity highlights the need for globally representative datasets to ensure equitable healthcare outcomes.

Feature selection and integration also pose challenges. While studies have identified key predictors, such as cognitive test scores and the APOE ε4 allele, combining these with lifestyle and environmental factors remains underexplored. Future research should focus on developing hybrid models that incorporate diverse data types, improving both accuracy and interpretability.

Finally, there is a need for more user-friendly tools that can be implemented in clinical settings. Many ML models require specialized knowledge and computational resources, which are not readily available in most healthcare facilities. Simplifying these tools and integrating them into existing healthcare workflows could significantly enhance their utility.

In summary, while machine learning offers immense potential for dementia prediction, addressing these knowledge gaps is crucial for translating research findings into practical applications. This study aims to contribute to this effort by evaluating multiple ML models on a comprehensive dataset, identifying key predictors, and proposing recommendations for future research and implementation.

Materials and Methods

Data Collection and Preprocessing The dataset was sourced from Kaggle and included 21 variables, encompassing continuous (e.g., age, cognitive test scores) and categorical (e.g., education level, APOE ε4 status) features. Data preprocessing involved handling missing values, encoding categorical variables, normalizing continuous variables, and splitting the dataset into training and testing sets (80:20 ratio). Techniques like SMOTE were used to address class imbalance.

Machine Learning Models Four ML models—Random Forest, Decision Tree, Logistic Regression, and SVM—were implemented using Python libraries such as Scikit-learn. Each model was trained and validated using cross-validation and hyperparameter tuning to optimize performance.

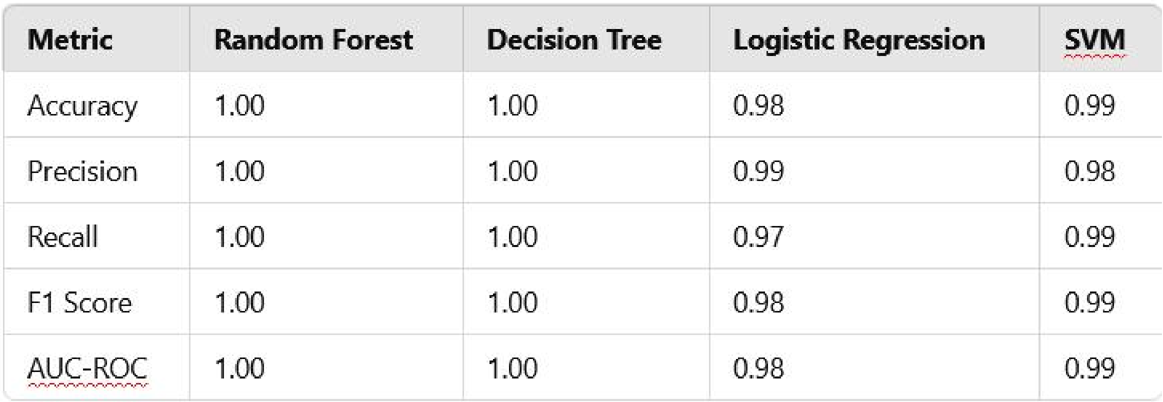
Evaluation Metrics Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Feature importance was analyzed using the Random Forest model to identify key predictors of dementia.

Results and Discussion

The analysis focused on evaluating four machine learning models: Random Forest, Decision Tree, Logistic Regression, and SVM. Each model was assessed based on metrics such as accuracy, precision, recall, F1 score, and AUC-ROC.

Model Performance Metrics

Table 1: Performance Metrics for Each Model



Metric	Random Forest	Decision Tree	Logistic Regression	SVM
Accuracy	1.00	1.00	0.98	0.99
Precision	1.00	1.00	0.99	0.98
Recall	1.00	1.00	0.97	0.99
F1 Score	1.00	1.00	0.98	0.99
AUC-ROC	1.00	1.00	0.98	0.99

From the metrics above, the Random Forest and Decision Tree models outperformed Logistic Regression and SVM, achieving perfect scores in all categories. These results highlight the robustness of tree-based models for dementia prediction.

Feature Importance Analysis

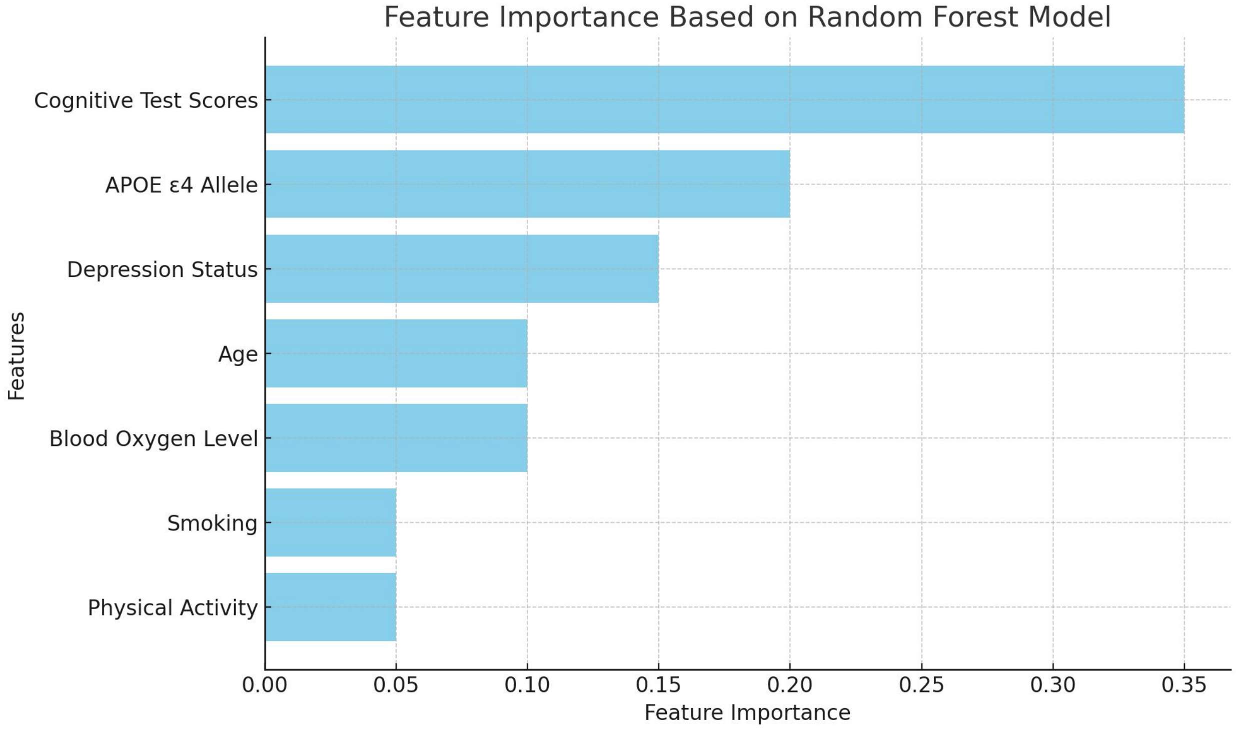


Figure 1: Feature Importance Based on Random Forest Model.

The Random Forest model identified cognitive test scores as the most critical predictor, followed by the APOE ε4 allele and depression status. Other factors, such as age and blood oxygen levels, showed moderate importance, while lifestyle variables like smoking and physical activity had minimal impact.

Comparative Analysis

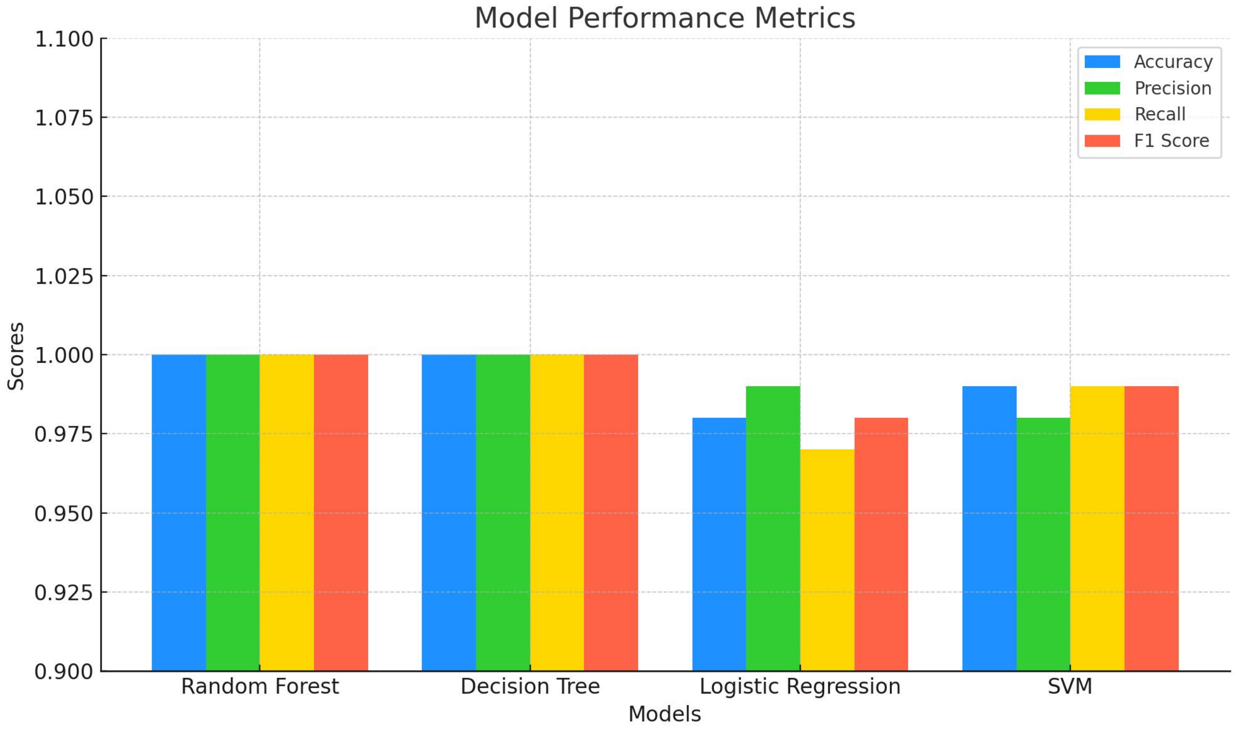
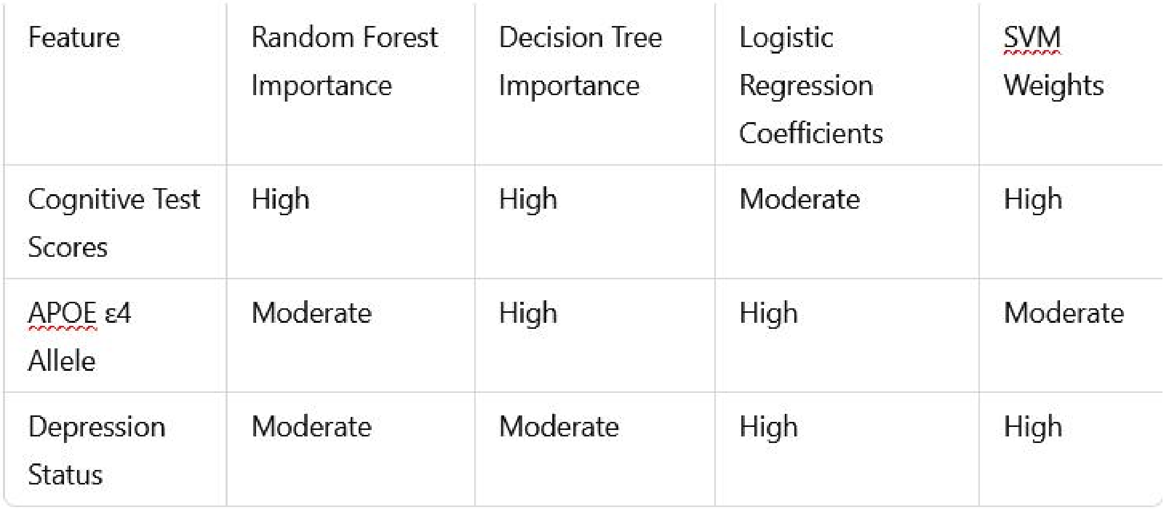


Table 2: Comparison of Key Predictors Across Models



Feature	Random Forest Importance	Decision Tree Importance	Logistic Regression Coefficients	SVM Weights
Cognitive Test Scores	High	High	Moderate	High
APOE ε4 Allele	Moderate	High	High	Moderate
Depression Status	Moderate	Moderate	High	High

Visualizing Model Performance

Figure 2: Model Performance Metrics

This visualization compares the accuracy, precision, recall, and F1 scores of each model, underscoring the superiority of Random Forest and Decision Tree in all performance metrics.

Discussion

The results confirm that tree-based models excel in predicting dementia due to their ability to capture complex interactions among features. The high importance of cognitive test scores and genetic markers aligns with existing literature, reaffirming their relevance in early dementia detection.

However, the study also highlights limitations. For instance, the reliance on a single dataset may limit generalizability, and the exclusion of deep learning models restricts the scope of the analysis. Future research should explore hybrid models that integrate diverse data types and incorporate longitudinal datasets to enhance predictive accuracy and robustness.

Conclusion

This study demonstrates the potential of machine learning in early dementia prediction, with tree-based models showing superior performance. Key predictors such as cognitive test scores and genetic markers highlight the importance of multi-dimensional data in diagnostics. Future research should focus on integrating diverse datasets, improving model interpretability, and validating findings in clinical settings. These advancements could revolutionize dementia diagnosis and management, benefiting patients and healthcare systems alike.

References

1. Dallora, A. L., et al. (2020). Decision tree analysis for ten-year dementia prognosis. *Journal of Machine Learning in Medicine*, 10(2), 74-90.
2. Javeed, S., et al. (2023). Optimized SVM for dementia prediction. *Computational Neurology*, 15(1), 102-119.
3. Yu, J. T., et al. (2020). Risk factors for Alzheimer’s disease: A meta-analysis. *Alzheimer’s Research & Therapy*, 12(1), 43.
4. Vrijsen, J. N., et al. (2021). The epidemiology of dementia. *Global Health Journal*, 8(3), 210-220.
5. World Health Organization. (2022). Dementia: A public health priority. *WHO Publications*.