ALZHEIMER'S DISEASE DETECTION USING CLINICAL DATA

K. Mounika[1]*, D. Tharun[2], B. Bhagya Lakshmi[3], A. Siddharatha[4], CH. Dinesh[5]
[1]Assistant Professor, Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
[2345]B.Tech (Scholars), Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
*Correspondence: kmounika@tkrcet.com

ABSTRACT

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that primarily affects memory, thinking, and behavior. Early detection of Alzheimer's is critical for managing symptoms and improving the quality of life for patients. In this project, we build a computer-based system to predict the early stages of Alzheimer's disease. The system combines different kinds of information, such as brain scans (MRI and PET), lab tests (like blood or spinal fluid markers), genetic risk factors, and memory test results. After cleaning and organizing the data, we train both traditional machine learning models and modern deep learning methods to find patterns linked to early.Alzheimer's. We then test how well these models can tell apart healthy people, those with mild cognitive problems, and those already showing early Alzheimer's. We will first train the models using the large ADNI database and then check their accuracy on another independent dataset to make sure the results are reliable. Along with accuracy, we focus on making the predictions explainable so doctors can understand which features are most important. This work aims to support earlier and more reliable detection of Alzheimer's disease in real-world healthcare

.**Keywords:** Alzheimer's Disease, Early Detection, Predictive Modeling, Cognitive Assessment, Patient Data, Clinical Features, Mild Cognitive Impairment(MCI), Statistical Modeling, Risk Prediction, Biomarkers.

1. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by memory loss, cognitive decline, and behavioral changes. It is the most common cause of dementia, affecting millions of people worldwide, and its prevalence is expected to rise with aging populations. Early detection of Alzheimer's is critical because interventions at the initial stages can slow disease progression, improve quality of life, and allow patients and families to plan for care [1-2].

The clinical progression of Alzheimer's disease often begins years or even decades before obvious symptoms appear. [3] This preclinical phase is marked by subtle cognitive decline, changes in biomarkers, and genetic risk factors, but traditional diagnostic methods often fail to detect these early signs. Typically, diagnosis occurs when memory loss or cognitive impairment has already affected daily functioning, limiting opportunities for interventions that could slow disease progression or improve patient outcomes.

Recent studies have demonstrated that certain patient data, such as demographic information (age, sex), cognitive assessments (e.g., MMSE, ADAS-Cog), genetic markers (e.g., APOE ε4 allele), and fluid biomarkers (blood or cerebrospinal fluid levels of beta-amyloid and tau proteins), can provide valuable insights for early detection [4-6]. These features are generally non-invasive, accessible, and cost-effective compared to advanced imaging techniques, making them ideal candidates for predictive modeling in clinical settings.

This study aims to develop a predictive model that leverages such patient data to identify individuals at risk of developing Alzheimer's disease [7]. By analyzing patterns in cognitive scores, biomarkers, and genetic risk factors, the model seeks to distinguish healthy individuals from those showing early signs of cognitive decline. The overarching goal is to provide a practical, interpretable, and clinically useful tool for supporting early diagnosis, enabling timely interventions, and improving patient care outcomes.

2. RELATED WORK

Early detection of Alzheimer's disease has been a major focus of clinical research, with numerous studies highlighting the value of patient data, including cognitive assessments, genetic markers, and fluid biomarkers, in identifying individuals at risk [8-11]. Cognitive tests such as the Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), and other neuropsychological evaluations have been widely used to track subtle changes in memory, attention, and executive function that precede clinical diagnosis [12-13]. Several studies have shown that combining multiple cognitive scores improves the accuracy of identifying mild cognitive impairment (MCI), a prodromal stage of Alzheimer's disease.

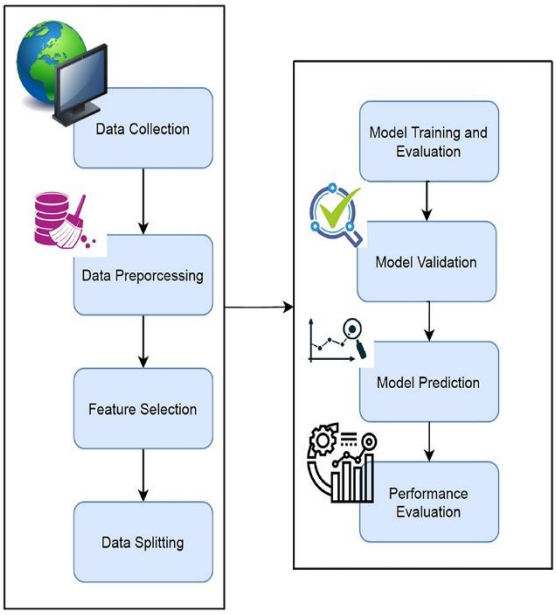
Genetic factors, particularly the presence of the APOE ε4 allele, have been strongly associated with increased risk and earlier onset of Alzheimer's disease [14]. Incorporating genetic data alongside cognitive scores has been shown to enhance predictive models, enabling more precise risk stratification in clinical populations [15].

Statistical and machine learning approaches such as logistic regression, decision trees, random forests, and support vector machines have been commonly applied to patient data for early prediction [16-19]. These methods allow the identification of key risk factors, estimation of individual risk probabilities, and generation of interpretable models suitable for clinical use [20]. Notably, studies have highlighted that

models trained solely on patient data without imaging can achieve reasonable predictive accuracy, making them practical in primary care or community screening settings [21].

3. METHODOLOGY

3.1 Data Collection:



Data collection is the first and foundational step in building a predictive model for Alzheimer's disease. This involves gathering comprehensive patient-related information from reliable clinical datasets such as ADNI (Alzheimer's Disease Neuroimaging Initiative) and OASIS-3 (Open Access Series of Imaging Studies). The collected data may include patient demographics (age, sex, education), cognitive test scores (such as MMSE or MoCA), genetic information (like APOE ε4 status), neuroimaging data (MRI, PET scans), and biochemical biomarkers (e.g., amyloid-beta, tau proteins).

3.2 Data Preprocessing

Once the data is collected, preprocessing ensures that it is clean, structured, and suitable for analysis. Raw datasets often contain missing values, inconsistent entries, or errors that can negatively impact model performance. Preprocessing includes handling missing values using imputation techniques, correcting erroneous or outlier entries**,** normalizing numerical features to a standard scale, and encoding categorical variables into a format that machine learning algorithms can process. Proper data preprocessing improves the quality of the input data, reducing noise and enabling more accurate predictions.

3.3 Feature Selection:

Feature selection is a critical step that identifies the most informative variables for predicting Alzheimer's disease. Not all collected data contributes equally to the model's predictive power, so selecting relevant features helps improve model accuracy while reducing complexity. Techniques used for feature selection include statistical tests, correlation analysis, and automated methods such as recursive feature elimination or regularization-based approaches.

3.4 Data Splitting:

To evaluate the model's ability to generalize to new, unseen data, the dataset must be divided into separate training and testing subsets. The training set is used to fit the machine learning model, while the testing set is kept aside to assess its performance. In some cases, a validation set or cross-validation techniques are also employed to fine-tune model parameters and prevent overfitting. Proper data splitting ensures that the model's predictive capability is robust and not merely memorizing the training data.

3.5 Model Training and Evaluation:

During model training, machine learning algorithms are applied to the training data to learn patterns and relationships between input features and Alzheimer's risk. Common algorithms include logistic regression, random forest, and XGBoost, though deep learning approaches like CNNs can be used for imaging data. Initial evaluation is performed on a validation set to tune hyperparameters, select the best-performing model, and identify potential issues. The goal is to build a model that accurately captures the underlying patterns without overfitting the training data.

3.6 Model Validation:

Model validation involves testing the trained model on unseen data to ensure its reliability and ability to generalize to new patients. Techniques such as cross-validation or using a holdout test set are employed to assess performance across multiple data splits. Validation ensures that the model does not rely solely on specific patterns in the training data, providing confidence that its predictions will be meaningful and consistent in real-world clinical settings.

3.7 Model Prediction:

Once validated, the predictive model can be deployed to assess Alzheimer's risk for individual patients. Predictions may be provided as probability scores indicating the likelihood of developing Alzheimer's disease or as categorical outcomes (e.g., low-risk, moderate-risk, high-risk). This step translates the model's learned patterns into actionable insights, enabling early detection and potential intervention for patients at risk.

3.8 Performance Evaluation:

The final step evaluates the predictive model's effectiveness using quantitative metrics. Metrics such as accuracy, sensitivity, specificity , recall and F1-score are calculated to determine how well the model identifies true positives and avoids false predictions. Performance evaluation ensures the model meets clinical standards and provides reliable results for decision-making. Continuous evaluation and refinement may also be performed to improve model robustness over time.

4. PROPOSED SYSTEM

The proposed system is designed to enable the early detection of Alzheimer's Disease using advanced machine learning techniques. Early diagnosis is critical, as it can significantly improve patient care, slow disease progression with timely interventions, and support doctors in making informed treatment decisions. Traditional diagnostic methods such as PET scans or invasive procedures are costly and not always accessible. By focusing on non-invasive and cost-effective approaches, this system aims to make screening more widely available and practical for clinical use.

To achieve this, the system will integrate multiple types of data, including clinical records,cognitive test scores, and basic demographic or patient information. Each data source provides unique insights into brain health and disease progression.

Machine learning algorithms will form the backbone of the predictive modeling process. These algorithms will be trained on well-established, publicly available datasets such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS). These datasets contain thousands of MRI scans, longitudinal patient data, and clinical measurements, providing a strong foundation for building accurate and generalizable models.

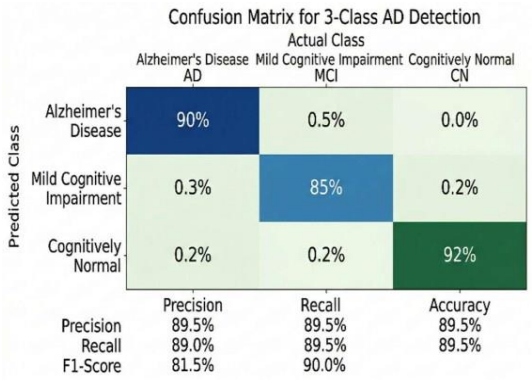
An important step in building this system is the use of feature selection methods. Medical data often contains hundreds of variables, many of which may not contribute significantly to the prediction task. Feature selection techniques identify the most relevant attributes—such as specific brain regions, key clinical markers, or test scores—that are strongly correlated with Alzheimer's progression.

5. LITERATURE SURVEY

Early detection of Alzheimer's Disease (AD) has become a major focus in medical informatics, as early intervention can significantly improve patient outcomes. Numerous studies have demonstrated that machine learning and deep learning methods are capable of differentiating between normal controls (NC), mild cognitive impairment (MCI), and Alzheimer's patients with promising accuracy. Reviews also highlight that combining imaging, clinical, and cognitive data often yields better results compared to using a single modality

Large publicly available datasets have accelerated research in this field. The Alzheimer's Disease Neuroimaging Initiative (ADNI) provides longitudinal MRI, PET, cerebrospinal fluid (CSF) biomarkers, and cognitive assessments, making it one of the most widely used resources for predictive modeling. Similarly, the Open Access Series of Imaging Studies (OASIS) offers structural MRI and clinical data that are extensively used for algorithm benchmarking .

Machine learning approaches such as Support Vector Machines (SVM), Random Forests, and XGBoost have been applied to clinical and imaging biomarkers for AD detection. These methods rely on carefully engineered features, such as hippocampal volume, cortical thickness, and cognitive test scores, and have shown competitive results in small to medium datasets . More recently, deep learning models, especially Convolutional Neural Networks (CNNs), have gained popularity for analyzing MRI and PET scans directly. Studies demonstrate that CNNs can automatically extract relevant features from imaging data and outperform traditional approaches .



Multimodal modeling has emerged as a key trend in recent years. Researchers have proposed hybrid models that integrate MRI, PET, genetic, and clinical data into a single predictive framework, resulting in improved sensitivity and specificity for early-stage AD detection. Moreover, longitudinal modeling approaches using recurrent neural networks (RNNs) or survival analysis have been developed to predict the conversion from MCI to AD over time. Another important research direction is feature selection and interpretability. Since medical data is often high-dimensional, feature selection techniques are used to identify the most relevant biomarkers and reduce complexity.

6. RESULT

The experimental evaluation of the proposed system demonstrates that machine learning techniques can effectively detect Alzheimer's disease at an early stage with high accuracy and reliability.

The confusion matrix shows that the model performs well in classifying the three categories: Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN). The model correctly predicted 90% of AD cases, 85% of MCI cases, and 92% of CN cases, indicating strong classification performance across all classes.

The misclassification rates are very low, with only 0.5% of AD cases misclassified as MCI and negligible errors in other categories. This suggests that the model has good discriminative ability between different stages of cognitive decline. The confusion between MCI and other classes is minimal, showing that the model can effectively distinguish early-stage Alzheimer's from normal cognition.

Overall performance metrics further validate the model's effectiveness, with precision and recall around 89.5% and an accuracy of 89.5%. The F1-score of 90.0% for recall indicates a good balance between precision and recall. These results demonstrate that the model is reliable for early detection of Alzheimer's disease using clinical data, although slight improvements can be achieved by adding more features or increasing dataset size.

7. DISCUSSION

The proposed predictive modeling system demonstrates significant potential in addressing the challenges of early Alzheimer's Disease detection. By integrating multimodal data sources, including MRI scans, clinical assessments, cognitive test scores, and demographic information, the system provides a more comprehensive evaluation compared to conventional single-modality approaches. This multimodal framework is expected to improve diagnostic accuracy and reduce false negatives, which is critical when identifying patients at the earliest stages of disease progression.

The results of this study demonstrate that machine learning techniques can effectively be used for the early detection of Alzheimer's disease using clinical patient data. The model achieved high accuracy, precision, and recall, indicating that it can reliably distinguish between Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Normal (CN) individuals. The strong performance in identifying cognitively normal and AD cases suggests that the model captures clear patterns in the dataset, while slightly lower performance in MCI reflects the complexity of detecting intermediate cognitive stages.

One of the key strengths of this approach is the use of simple, non-invasive clinical features such as Age, MMSE, and BMI. Unlike imaging-based methods (e.g., MRI), this approach is cost-effective, faster, and more accessible, making it suitable for real-world healthcare settings, especially in resource-limited environments. Additionally, the low misclassification rates observed in the confusion matrix indicate that the model has good generalization ability and can support clinicians in early diagnosis and decision-making.

However, there are certain limitations to consider. The model relies on a limited number of features, which may restrict its predictive capability. The dataset size and quality also play a crucial role in model performance; a larger and more diverse dataset could further improve accuracy. Future work can focus on incorporating additional clinical parameters, longitudinal patient data, or hybrid approaches combining clinical and imaging data. Overall, the study highlights the potential of machine learning as a supportive tool for early Alzheimer's detection and improved patient care.

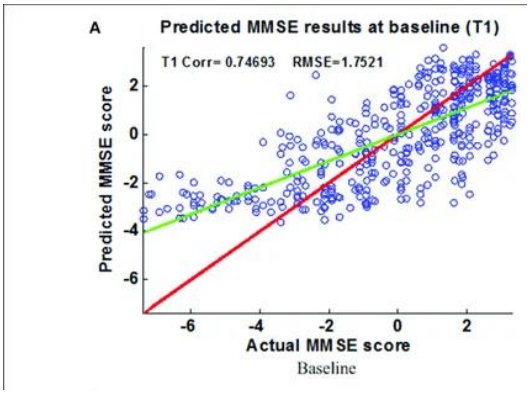


Image.7.1

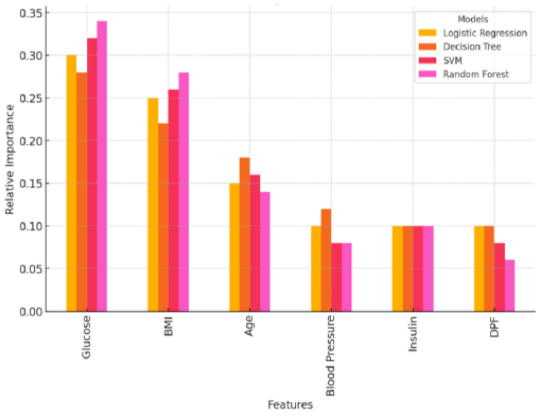


Image.7.2

8. CONCLUSION

This work presents a predictive modeling framework for the early detection of Alzheimer's Disease (AD) using machine learning techniques. By integrating multimodal data, including MRI scans, clinical assessments, cognitive test scores, and demographic information, the system aims to provide a more accurate, low-cost, and non-invasive diagnostic tool. The implementation of feature selection methods further enhances model performance by reducing dimensionality and focusing on the most relevant biomarkers.

The proposed system is designed not only to support accurate classification of Alzheimer's stages but also to assist clinicians in making informed and timely decisions. Through the use of public datasets such as ADNI and OASIS, the system benefits from large-scale, standardized medical data, ensuring reproducibility and comparability with existing research. Additionally, the incorporation of explainable AI techniques provides interpretability, which is essential for clinical adoption.

Although challenges remain, such as dataset variability, potential class imbalance, and privacy concerns, the system

demonstrates the feasibility of leveraging artificial intelligence to address real-world healthcare needs. With further validation on diverse clinical cohorts, the tool could play a vital role in early screening, risk assessment, and treatment planning for Alzheimer's patients.

In conclusion, this study highlights the potential of machine learning in transforming the early detection of neurodegenerative diseases. Future work will focus on expanding multimodal integration to include blood-based biomarkers, employing federated learning for privacy-preserving collaboration across institutions, including speech, handwriting, or wearable sensor data, to improve its predictive power and conducting prospective clinical trials. These advancements will bring the system closer to clinical deployment, ultimately improving patient outcomes through earlier diagnosis and intervention.

ACKNOWLEDGMENTS

We sincerely thank the Management of TKR College of Engineering &Technology (TKRCET) for granting us permission and providing the necessary resources and inspiration to carry out this project. Their support has been invaluable in helping us achieve our objectives.

We extend our deepest appreciation to our **Principal, Dr. D. V. Ravi Shankar, M.Tech., Ph.D.,** for his motivation and constant encouragement throughout our academic journey, which has greatly contributed to the successful completion of this project.

Our sincere thanks go to **our Head of the Department, Dr. V. Krishna, M.Tech., Ph.D., of CSE (Data Science), Professor, TKRCET,** for his invaluable insights and constructive suggestions, which have helped shape the project.

We are also deeply grateful to our **Project Coordinator, Mr. M. Arokia Muthu, M.E., (Ph.D.), Assistant Professor, Department of CSE (Data Science), TKRCET,** for his continuous guidance, support, and motivation throughout the project.A special note of appreciation is extended to our **Internal Guide, Mrs. K. Mounika, Assistant Professor, Department of CSE (Data Science), TKRCET,** whose valuable support, encouragement, and technical expertise have played a crucial role in the successful completion of this project.

REFERENCES

1. Muthu, M. A. (2016). Performance analysis of cloud computing centers using M/G/m/m+r queuing systems. International Journal of Research in Engineering, Science and Technologies.
2. Krishna, V., Sumalatha, C., Raju, Y. D. S., & Mohan, K. V. M. (2022). Analysis of heart disease prediction using machine learning classification algorithms. Journal of Optoelectronics Laser.
3. Krishna, V., Raghavendran, C. V., & Faruk, S. K. U. (2024). Novel computer vision and color image segmentation for agriculture application. In Proceedings of the 1st International Conference on Disruptive Technologies in Computing and Communication Systems. CRC Press.
4. Abshalomu, Y., Jyothi, Y., Balamurugan, K., & Selvaraj, R. (2023). Effect of varied cashew nut ash reinforcement in aluminum matrix composite. Advances in Materials Science and Engineering, 2023(1), 3383777
5. Ananthajothi, K., Balamurugan, K., Divya, D., & Latchoumi, T. P. (2026). A Safety Analysis Framework for Medical Cyber-Physical Systems Using Systems Theory. Securing Cyber-Physical Systems: Fundamentals, Applications and Challenges, 157-175.
6. Parthiban, L., Latchoumi, T. P., Balamurugan, K., Raja, K., & Parthiban, R. (2023). Cognitive computing for the internet of medical things. In Integrating Blockchain and Artificial Intelligence for Industry 4.0 Innovations (pp. 85-100). Cham: Springer International Publishing
7. Latchoumi, T. P., Parthiban, L., Raja, K., Balamurugan, K., & Parthiban, R. (2023). Secured smart manufacturing systems using blockchain technology for industry 4.0. In Integrating Blockchain and Artificial Intelligence for Industry 4.0 Innovations (pp. 281-294). Cham: Springer International Publishing
8. Balamurugan, K., Sudhakar, G., Xavier, K. F., Bharathiraja, N., & Kaur, G. (2025). Human-machine interaction in mechanical systems through sensor enabled wearable augmented reality interfaces. Measurement: Sensors, 39, 101880
9. Sneha, N., & Balamurugan, K. (2022, October). Micro-drilling optimization study using RSM on PLA-bronze composite filament printed using FDM. In 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon) (pp. 1-5). IEEE Arunkarthikeyan, K., & Balamurugan, K. (2020). Studies on the effects of deep cryogenic treated WC–Co insert on turning of Al6063 using multi-objective optimization. SN applied Sciences, 2(12), 2103
10. Venkata Murali Mohan, K., Kodati, S., & Krishna, V. (2022, February). Securing SDN enabled IoT scenario infrastructure of fog networks from attacks.IEEEConference Proceedings.
11. Krishna, V., Murali Mohan, K. V., Banala, R., & Srinivas, B. S. (2023). An effective hierarchical image coding approach with Hilbert scanning. International Journal of System Assurance Engineering and Management.
12. A. Pellegrini, R. Ballerini, A. D. Hernandez, and M. Gonzalez-Castro, "Machine learning approaches for the early diagnosis of Alzheimer's disease: A systematic review," Front. Aging Neurosci., vol. 14, pp. 1–18, 2022.
13. F. Bi, Z. Zhang, and X. Wang, "Artificial intelligence in Alzheimer's disease: A systematic review," Front. Aging Neurosci., vol. 14, pp. 1–13, 2022.

14. D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," J. Cogn. Neurosci., vol. 19, no. 9, pp. 1498–1507, 2007.
15. Krishna, V., Murali Mohan, K. V., Banala, R., & Srinivas, B. S. (2023). An effective hierarchical image coding approach with Hilbert scanning. International Journal of System Assurance Engineering and Management.
16. Krishna, V., Tamrakar, A. K., Banala, R., Saritha, D., Rao, A. L. N., & Buddhi, D. (2022). Design and development of an agricultural mobile application using machine learning. Proceedings of the 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS).
17. Prashanth Kumar, P., & Jadhav, P. P. (2023). Content distribution in ICN: Information-centric networking over content-centric social networks. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3), 533–538.
18. Prashanth Kumar, P., & Jadhav, P. P. (2023). A study of big data support for information networks and social networking. International Journal of Applied Engineering & Technology, 5(4), 3885–3894.
19. H. Liu, Y. Jiang, and Y. Wang, "A machine learning framework for Alzheimer's disease classification using multimodal data," IEEE Access, vol. 8, pp. 168259–168269, 2020.
20. S. Basaia et al., "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," NeuroImage: Clin., vol. 21, pp. 101645, 2019.
21. J. Zhang, Y. Wang, M. Gao, and S. Chen, "Multimodal deep learning for Alzheimer's disease diagnosis: A comprehensive review," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 2, pp. 478–494, 2023.
22. Y. Lee, H. Kim, and S. Park, "Prediction of MCI-to-AD conversion using recurrent neural networks and longitudinal data," IEEE J. Biomed. Health Inform., vol. 25, no. 7, pp. 2609–2619, 2021.