INTRODUCTION

The paradigm of Alzheimer’s disease (AD) treatment is at a major turning point, with positive reports of disease-modifying therapies, including anti-amyloid approaches. This gives hope that early detection and curative treatment of AD using biomarkers may be possible in the future. However, this approach is not yet suitable for application to all populations in real clinical settings due to costs and accessibility.

The population is aging and it is known that multiple factors affect the onset of AD, and no one can be completely free from the onset of AD. Therefore, cost-effective methods for the early diagnosis of AD are required.

Currently, screening tools, such as the Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment, are used to detect cognitive impairment. MMSE is the most widely used tool that provides a global measure of cognitive impairment in clinical, research, and community settings. However, a recent study found no evidence supporting the significant role of MMSE as a “stand-alone single-dose test” in identifying patients who develop AD. Bondi et al. stated that conventional criteria for cognitive decline can be problematic because they rely on impairment on a single neuropsychological test, clinical judgment, or limited neuropsychological assessment using a “one test equals one domain” methodology. Jak et al. suggested that the neuropsychological (NP) criteria using comprehensive neuropsychological tests offer the ideal balance of sensitivity and reliability by using a more liberal >1 SD cut-off.

Objective

This study aimed to use deep learning (DL) to develop a cost-effective and accessible screening tool to improve the detection of cognitive decline, a precursor of Alzheimer’s disease (AD). This study integrating a comprehensive battery of neuropsychological tests adjusted for individual demographic variables such as age, sex, and education level.

Methods

A total of 2,863 subjects with subjective cognitive complaints who underwent a comprehensive neuropsychological assessment were included. A random forest classifier was used to discern the most predictive test combinations to distinguish between dementia and nondementia cases. The model was trained and validated on this dataset, focusing on feature importance to determine the cognitive tests that were most indicative of decline.

Results

Subjects had a mean age of 72.68 years and an average education level of 7.62 years. The DL model achieved an accuracy of 82.42% and an area under the curve of 0.816, effectively classifying dementia. Feature importance analysis identified significant tests across cognitive domains: attention was gauged by the Trail Making Test Part B, language by the Boston Naming Test, memory by the Rey Complex Figure Test delayed recall, visuospatial skills by the Rey Complex Figure Test copy score, and frontal function by the Stroop Test Word reading time.

Conclusion

This study showed the potential of DL to improve AD diagnostics, suggesting that a wide range of cognitive assessments could yield a more accurate diagnosis than traditional methods. This research establishes a foundation for future broader studies, which could substantiate the approach and further refine the screening tool.

Keywords

Deep learning; Cognition; Dementia; Neuropsychology.
for impairment (rather than a 1.5−2 SD cut-off), with the essential need for at least two impaired scores within one cognitive domain. When analyzed using Alzheimer's Disease Neuroimaging Initiative (ADNI) data, approximately one-third of participants diagnosed with mild cognitive impairment (MCI) using the conventional criteria were cognitively normal (CN) diagnosed using NP criteria. Additionally, when participants re-diagnosed with CN were further explored, many characteristics, such as imaging findings, genetic biomarkers, and pathological findings, were more consistent with CN than with MCI.

The validity of the NP criteria has been tested in many studies. However, the three cognitive domains included in the NP criteria were memory, language, and executive function; the remaining visuospatial functions and attentional abilities were excluded. In addition, the ADNI data, mainly used to verify the NP criteria, lacked cognitive function test results corrected for participants’ education level. Therefore, in this study, we conducted a comprehensive neuropsychological test that included all five cognitive domains and attempted to create a cost-effective screening test combination based on the results adjusted for each subject’s age, sex, and education level. Additionally, we used deep learning (DL), an artificial intelligence technology, for this analysis. DL works by mimicking the operation of the human brain, generating automated predictions from input data. In other words, because it can extract predictive values by identifying and learning patterns from input data with complex relationships, DL is establishing itself as a new tool showing promising results in early stage AD research.

Consequently, this study aimed to use DL technology to create a cost-effective screening tool for cognitive decline.

METHODS

Data preparation

The data used were the results of a comprehensive neuropsychological test conducted in outpatients or inpatients with subjective cognitive complaints who visited Yeungnam University Hospital between January 2017 and December 2023. This study used a dataset containing test results in five cognitive domains: attention, language, memory, visuospatial function, and frontal function. The data were collected by a psychiatrist and trained psychologists. Subjects completed a comprehensive neuropsychological assessment and activities of daily living to assess their cognitive and daily functions. The clinical diagnosis of each subject was based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, and it was considered that they had dementia if they met the diagnostic criteria for a major neurocognitive disorder.

The Seoul Neuropsychological Screening Battery (SNSB) was used as a comprehensive neuropsychological test. SNSB is a comprehensive neuropsychological test that evaluates five cognitive domains such as memory, language, attention, visuospatial function, and frontal (executive) function. SNSB includes the following detailed list of cognitive functions: 1) memory was evaluated by Seoul Verbal Learning Test-Elderly’s version evaluating immediate recalls, delayed recall, recognition abilities, and same abilities using Rey Complex Figure Test (RCFT); 2) language function was evaluated by spontaneous speech, comprehension, repetition, reading, writing, and Korean-Boston Naming Test (K-BNT); 3) attention ability was evaluated by Vigilance test, Digit Span Test, and Letter cancellation; 4) visuospatial function was evaluated by copy score and copy time of RCFT and Clock Drawing Test; 5) frontal functions were evaluated by Contrasting program, Go-No go test, FIST Edge-Palm, Alternating hand movement, Alternating square & triangle, Luria loop, Controlled Oral Word Association Test, Korean-Color Word Stroop Test, Digit Symbol Coding, and Korean-Trail Making Test-Elderly's version (K-TMT-E). Analysis was performed using test results that yielded z-scores adjusted for age, sex, and educational level.

The dataset was preprocessed by correcting the column names based on the first-row entries and converting all numerical data from strings to floating-point numbers. To maintain the integrity of the dataset, missing values were imputed using the mean of each test in all participants. This study was approved by the Institutional Review Board of Yeungnam University Medical Center (YUMC 2021-06-039).

Feature selection and model training

To identify the most predictive combination of cognitive tests to distinguish the two groups, a random forest classifier was used because of its ability to handle high-dimensional data and provide insights into feature importance. The classifier was trained on a subset of data, divided into training (70%) and testing (30%) sets to accurately evaluate the performance of the model.

Model evaluation

The effectiveness of the model was assessed using several statistical metrics, including accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC). These metrics were selected to provide a comprehensive understanding of the performance of the model in terms of its ability to correctly classify subjects into the appropriate dementia category.

Statistical analysis

An independent t-test was conducted for each cognitive test to determine the mean score difference between the two
groups. This analysis helped identify tests with significant differences, suggesting their potential utility in distinguishing between subjects with and without dementia.

The RandomForestClassifier also provided feature importance scores, which were used to rank cognitive tests based on their contribution to the model's prediction accuracy. This approach allowed the identification of the most relevant tests across cognitive domains for the detection of dementia.

In the present study, DL models were implemented using the TensorFlow framework, version 2.4 (Google LLC, Mountain View, CA, USA), which offers extensive support for DL algorithms. The models were built and trained using Keras (Google LLC), a high-level neural networks API running on top of TensorFlow, facilitating rapid experimentation and model iteration. The training and evaluation of the models were conducted on a system equipped with an NVIDIA Tesla V100 GPU, which provided the necessary computational power for handling large datasets and complex neural network architectures. The system operated under Ubuntu 20.04 LTS (Google LLC), utilizing Python 3.8 for scripting and data manipulation tasks. This setup ensured an efficient pipeline from data preprocessing through to model training and evaluation.

RESULTS

Demographic characteristics

This study included 2,863 subjects who underwent cognitive assessments. The mean and standard deviation (SD) of the participant's age and years of education were 72.68±9.60 (dementia group 73.62±9.30, non-dementia group 72.58±9.66) years and 7.62±4.86 (dementia group 7.03±4.79, non-dementia group 8.55±4.54) years, respectively. The mean MMSE total score was 18.54±7.02 (dementia group 16.48±6.35, non-dementia group 26.33±2.51), Clinical Dementia Rating-Global Score was 1.25±0.74 (dementia group 2.31±0.87, non-dementia group 0.78±0.56), Clinical Dementia Rating-Sum of Boxes was 6.77±5.08 (dementia group 8.25±4.91, non-dementia group 1.95±0.91), Global Mental State Examination; CDR, clinical dementia rating; GS, Global Deterioration Scale; IADL, instrumental activities of daily living; B (K_TMT_E_B_time_z) emerged as the most critical factor, underscoring its significance in assessing attention and executive function; 2) language domain: The Boston Naming Test (Naming_K_BNT_z) was identified as the key test, highlighting its role in evaluating language abilities; 3) memory domain: The Rey Complex Figure Test delayed recall (RCFT_delayed_recall_z) was the predominant test, indicating its importance in memory assessment; 4) visuospatial domain: The Rey Complex Figure Test copy score (Rey_CFT_copy_score_z) was the most influential, reflecting its utility in assessing visuospatial skills; 5) frontal function domain: The Stroop

DL analysis results

Model performance

The random forest classifier demonstrated robust performance in distinguishing between the two subject groups. The model achieved an accuracy of 82.42%, with a precision of 84.74% and a recall of 94.10%, indicating a high sensitivity in the detection of subjects with clinical dementia. The F1 score, which is the harmonic mean of precision and recall, was 89.18%, reflecting the balanced effectiveness of the model in terms of precision and sensitivity. Furthermore, the AUC was 0.816, illustrating the excellent discrimination capability of the model (Table 2 and Figure 1).

Feature importance

The analysis of the feature importance provided by the RandomForestClassifier highlighted the most predictive tests within each cognitive domain. The most influential tests in each domain were: 1) attention domain: The Trail Making Test Part B (K_TMT_E_B_time_z) emerged as the most critical factor, underscoring its importance in assessing attention and executive function; 2) language domain: The Boston Naming Test (Naming_K_BNT_z) was identified as the key test, highlighting its role in evaluating language abilities; 3) memory domain: The Rey Complex Figure Test delayed recall (RCFT_delayed_recall_z) was the predominant test, indicating its importance in memory assessment; 4) visuospatial domain: The Rey Complex Figure Test copy score (Rey_CFT_copy_score_z) was the most influential, reflecting its utility in assessing visuospatial skills; 5) frontal function domain: The Stroop

Table 1. Demographic characteristics of all subjects

<table>
<thead>
<tr>
<th>Metric</th>
<th>All group</th>
<th>Dementia group</th>
<th>Non-dementia group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>72.68±9.60</td>
<td>73.62±9.30</td>
<td>72.58±9.66</td>
</tr>
<tr>
<td>Years of education</td>
<td>7.62±4.86</td>
<td>7.03±4.79</td>
<td>8.55±4.54</td>
</tr>
<tr>
<td>MMSE total score</td>
<td>18.54±7.02</td>
<td>16.48±6.35</td>
<td>26.33±2.51</td>
</tr>
<tr>
<td>CDR-GS</td>
<td>1.25±0.74</td>
<td>2.31±0.87</td>
<td>0.78±0.56</td>
</tr>
<tr>
<td>CDR-SOB</td>
<td>6.77±5.08</td>
<td>8.25±4.91</td>
<td>1.95±0.91</td>
</tr>
<tr>
<td>GDS</td>
<td>3.87±1.06</td>
<td>5.56±3.22</td>
<td>3.09±0.98</td>
</tr>
<tr>
<td>BADL</td>
<td>16.69±4.87</td>
<td>15.88±5.15</td>
<td>19.15±2.63</td>
</tr>
<tr>
<td>IADL</td>
<td>1.38±2.44</td>
<td>1.72±2.72</td>
<td>0.39±0.59</td>
</tr>
</tbody>
</table>

Data are presented as mean±standard deviation. MMSE, Mini-Mental State Examination; CDR, clinical dementia rating; GS, global score; SOB, sum of boxes; B (K_TMT_E_B_time_z) emerged as the most critical factor, underscoring its importance in assessing attention and executive function; 2) language domain: The Boston Naming Test (Naming_K_BNT_z) was identified as the key test, highlighting its role in evaluating language abilities; 3) memory domain: The Rey Complex Figure Test delayed recall (RCFT_delayed_recall_z) was the predominant test, indicating its importance in memory assessment; 4) visuospatial domain: The Rey Complex Figure Test copy score (Rey_CFT_copy_score_z) was the most influential, reflecting its utility in assessing visuospatial skills; 5) frontal function domain: The Stroop

Table 2. Model performance

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.8242</td>
</tr>
<tr>
<td>Precision</td>
<td>0.8474</td>
</tr>
<tr>
<td>Recall</td>
<td>0.9410</td>
</tr>
<tr>
<td>F1 score</td>
<td>0.8918</td>
</tr>
<tr>
<td>AUC score</td>
<td>0.8156</td>
</tr>
</tbody>
</table>

AUC, area under the curve
Cognitive Decline Criteria Using DL Technology

Test Word reading time (StroopTest_Wordreading_Time_per_item_z) was the most significant, emphasizing its value in evaluating frontal lobe functions.

When analyzed based on all the tests in the entire cognitive domain, the most important features identified by the RandomForestClassifier and their respective importance scores were as follows: 1) K_TMT_E_B_time_z (Trail Making Test Part B, time): The importance score was notably high, reflecting its significant role in the model's classification ability. This test is the most critical feature to distinguish between the two groups; 2) RCFT_delayed_recall_z (Rey Complex Figure Test delayed recall (RCFT_z): This memory-related test also showed a high importance score, indicating its strong influence on the predictive performance of the model; 3) Naming_K_BNT_z (Boston Naming Test): As a key test in the language domain, it demonstrates substantial predictive value with a significant importance score; 4) Rey_CFT_copy_score_z (Rey Complex Figure Test copy score): This test in the visuospatial domain was identified as highly influential in the analysis, as indicated by the importance score; 5) Stroop Test_Word reading_Time_per_item_z (Stroop Test Word reading time): This test was of notable importance, highlighting its relevance in the assessment of frontal lobe function and contributing to the discrimination capacity of the model (Table 3).

DISCUSSION

This study sought to take advantage of DL methodologies to improve the diagnostic precision of AD in a clinical setting. The use of a random forest classifier model, trained in comprehensive neuropsychological test results, is a testament to the potential of artificial intelligence to enhance early detection and intervention strategies for AD. The results of this study underscore the importance of employing a multifaceted approach that encompasses a wide range of cognitive domains, moving beyond traditional screening methods such as MMSE, which may not capture the complexity of cognitive impairment associated with AD.

The high accuracy and recall of the model suggest that DL can identify nuanced patterns in cognitive tests that could escape traditional analysis. The sensitivity of the model, as evidenced by the high F1 score and AUC, indicates its potential to reduce false negatives, thus ensuring that fewer cases of dementia remain unrecognized. Moreover, the feature importance analysis highlighted tests in various cognitive domains, indicating that the cognitive impact of AD is multifactorial and supports the need for a comprehensive testing strategy, as posited by Jak et al.

Importantly, the prominent role of tests that assess attention and executive function in the performance of the model aligns with emerging research suggesting that deficits in these areas may precede memory impairments in AD. It is now recognized that AD produces marked impairment in attentional and executive functions, which are related to frontal lobe function before deficits in language and visuospatial function occur. A systematic review of executive functions in AD discusses how, in the last 20 years, views have changed about the preservation of executive functions in the early stages of AD. Recent studies have confirmed the presence of early impairment in various tasks that investigate executive function due to degeneration of the prefrontal cortex. This includes specifically compromised inhibitory abilities, attentional, and visuospatial functions. The paper "Working Memory and Executive Function Decline across Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease" discusses how MCI is often characterized by slight but noticeable deficits in attention, learning and memory, executive function, processing speed, and semantic language. Early impairments in visual episodic memory, executive function, semantic language/memory, attention, and working memory are also strong predictors of progression from MCI to AD.

The findings also highlight the potential shortcomings of relying solely on memory-focused assessments, as visuospatial abilities and language function have emerged as significant predictors of cognitive decline. A study highlights that verbal

Table 3. Feature importance

<table>
<thead>
<tr>
<th>Feature</th>
<th>Importance score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trail Making Test B</td>
<td>0.1118</td>
</tr>
<tr>
<td>Rey Complex Figure Test Delayed Recall</td>
<td>0.0646</td>
</tr>
<tr>
<td>Boston Naming Test</td>
<td>0.0424</td>
</tr>
<tr>
<td>Rey Complex Figure Test Copy Score</td>
<td>0.0594</td>
</tr>
<tr>
<td>Stroop Test Word Reading Time per_item_z</td>
<td>0.0508</td>
</tr>
</tbody>
</table>

Figure 1. Receiver operating characteristic curve and area under the curve (AUC) values of the trained random forest classifier.
fluency can be more severely compromised than memory in the early stage of AD, with verbal fluency tests potentially serving as important tools for early screening of cognitive decline. Additionally, visuospatial skills, such as the ability to copy diagrams and draw a clock face, were found to show a strong negative correlation with disease duration, indicating their potential use to assess the progression of cognitive decline.\textsuperscript{16} Visuospatial processing speed is suggested to be functionally relevant throughout life, with the development of tasks such as the Visuospatial Processing Speed task aimed at capturing changes in ability that could be robust to display variations and suitable for web-based testing.\textsuperscript{17}

In other words, comprehensive neuropsychological tests are meaningful in that they aim to compensate for the vulnerability of “one test equals one domain.”\textsuperscript{5,6} Using actuarial-based neuropsychological methods, we avoid the difficulty of interpreting individual impaired scores on a single cognitive test and apply a cut-off score for impairment that optimizes classification rates; this can reduce overinterpretation and minimize the chance of a false-positive diagnosis of MCI or dementia.\textsuperscript{18-20}

This study has some limitations. First, the use of data from a single center can limit the generalizability of the findings. Furthermore, the exclusion of other factors, such as lifestyle, medical history, and socioeconomic status, which could influence cognitive function, may have affected the predictive capabilities of the model. Future studies could benefit from a more heterogeneous dataset and the inclusion of these variables to enhance the robustness and applicability of the findings.

In conclusion, this study demonstrated the viability of DL as a tool for early AD diagnosis, highlighting its ability to integrate complex datasets to produce highly sensitive diagnostic predictions. The DL model presented here could serve as a foundation for developing a cost-effective and accessible screening tool for cognitive decline, ultimately contributing to the broader goal of timely and accurate AD diagnosis across various populations. The implications of these findings are substantial, offering a potential pathway to revolutionize the standard of care for cognitive disorders and prompting a re-evaluation of current diagnostic criteria and practices.

**Availability of Data and Material**

Data supporting the findings of this study are available upon request from the corresponding author. Data are not publicly available because they contain information that can compromise the privacy of research participants.

**Conflicts of Interest**

The authors have no potential conflicts of interest to disclose.

**Author Contributions**


**ORCID iDs**

Hye-Geum Kim https://orcid.org/0000-0002-9677-7011
Wan-Seok Seo https://orcid.org/0000-0002-5122-5360
Bon-Hoon Koo https://orcid.org/0000-0001-9633-3835
Eun-Jin Cheon https://orcid.org/0000-0002-6986-0768
Seokho Yun https://orcid.org/0000-0002-2200-3494
Sohye Jo https://orcid.org/0000-0001-9471-0713
Byoungyoung Gu https://orcid.org/0009-0004-5212-0570

**Funding Statement**

This work was supported by the 2023 Yeungnam University Research Grant.

**Acknowledgments**

None

**REFERENCES**