Society of Medical Education & Research

2024, Volume 4, Issue 2, Page No: 69-93
ISSN: 3108-4826

Journal of Medical Sciences and Interdisciplinary Research

Integration of HOIRT-Derived Scores and Machine Learning for Accurate Detection of Cognitive Impairment

Ivana Marija Babic[1]*, Nikola Petar Jovanovic[1]

[1]Department of Management, Faculty of Economics, University of Belgrade, Belgrade, Serbia.

***E-mail** ✉ ivana.babic.bg@yahoo.com

Abstract

Early identification of cognitive impairment (CI) is essential for the health and quality of life of older adults. The MyCog assessment utilizes two established iPad-administered tests drawn from the NIH Toolbox® for Assessment of Neurological and Behavioral Function (NIH Toolbox). These tests target key cognitive areas: Picture Sequence Memory (PSM) for episodic memory and Dimensional Change Card Sort Test (DCCS) for executive function/cognitive flexibility. The research included 86 participants and examined various machine learning approaches to improve CI classification. This included conventional classifiers as well as neural network techniques. Following 100 bootstrap iterations, the Random Forest algorithm performed best, achieving strong metrics: precision of 0.803, recall of 0.758, accuracy of 0.902, F1 score of 0.742, and specificity of 0.951. Importantly, the model used a combined score obtained from a 2-parameter higher-order item response theory (HOIRT) framework that combined DCCS and PSM performance. A central conclusion of the work highlights the limitations of depending only on a single fixed threshold for the composite score. Rather, it supports the use of machine learning algorithms that integrate HOIRT-based scores along with additional variables such as age. This strategy offers greater potential for accurate CI prediction, thereby supporting earlier screening and management in older populations.

Keywords: MyCog, NIH toolbox, Machine learning, Deep learning, IRT, Higher order item response theory

Introduction

Cognitive impairment (CI) represents a major public health issue with substantial effects on older individuals. Mild CI (MCI) frequently remains undiagnosed in clinical settings, as providers often depend on patients voluntarily reporting issues, yet only about one-third typically do so [1–4].

Prompt recognition of CI provides numerous advantages for patients and caregivers. It can uncover reversible conditions (e.g., depression or vitamin B12 deficiency) (Office of Disease Prevention and Health Promotion, n.d.). When CI exists, early identification allows time for emotional adaptation, future planning, access to symptom-mitigating interventions (to preserve independence and well-being), and proactive management of safety risks (e.g., driving or home modifications) [5]. The aim was to create a short screening tool (< 10 minutes) suitable for primary care to facilitate early CI detection, including dementia (CID). The MyCog assessment incorporates modified versions of two validated iPad tests from the NIH Toolbox® for Assessment of Neurological and Behavioral Function Cognitive Battery (NIHTB-CB), targeting two domains often affected earliest: Picture Sequence Memory (PSM) for episodic memory and Dimensional Change Card Sort Test (DCCS) for cognitive flexibility [6–8]. This investigation sought to test, improve, and initially validate the MyCog assessment for identifying CID among community-dwelling older adults in primary care. A total of 86 individuals were enrolled from an ongoing cohort and underwent a short in-person evaluation. CI status was established via documented dementia/CI

Access this article online https://smerpub.com/

Received: 23 May 2025; **Accepted:** 04 August 2025

How to cite this article: Babic IM, Jovanovic NP. Integration of HOIRT-Derived Scores and Machine Learning for Accurate Detection of Cognitive Impairment. J Med Sci Interdiscip Res. 2024;4(2):69-93. https://doi.org/10.51847/y761zroHQX

diagnosis or results from a full cognitive evaluation within the prior 18 months. Besides the MyCog PSM and DCCS tasks, participants completed three remote Mobile Toolbox tests on smartphones: Arrow Matching (ARW), Sequences (MFS), and Number-Symbol Match (NSM). These five tests are described further in the following section.

The primary goal was to determine the most effective derived scores for separating those with and without CI. The research systematically compared various scoring approaches within the same dataset to find the strongest predictors for correct group assignment. In particular, it contrasted scores from two cognitive domains against those from five domains, seeking insight into subtest scoring mechanisms and evaluating the final MyCog composite score's discriminatory power. Furthermore, machine learning (ML) techniques were applied to boost classification performance. ML employs non-parametric methods to achieve better fit and flexibility in data representation [9]. The analysis also identified key predictors ("features") from the MyCog tasks associated with cognitive status.

In summary, the work offers a thorough examination of multiple tests, scoring strategies, and ML methods to reliably differentiate CI status using fewer measures. The study recruited 86 patients who completed NIH Toolbox PSM and DCCS via iPad, plus self-administered Mobile Toolbox ARW, MFS, and NSM on smartphones. Various composite scoring models and ML algorithms with differing feature sets were tested to optimize MCI prediction.

Materials and Methods

Real data

The dataset comprised results from the five tests administered to participants: MyCog DCCS and PSM, plus Mobile Toolbox ARW, MFS, and NSM. Following data cleaning and integration, complete records were available for 86 individuals, including item-level responses across the five domains and response times for three tests (ARW, PSM, and DCCS). Demographic details—age, race, education, and income—were also recorded for each participant. Impairment was binary-coded (0 = normal cognition, 1 = CI). **Table 1** presents demographic and clinical characteristics for the 86 participants, with cognitive status stratified by gender. For instance, among 60 females, 46 had normal cognition, and 14 had CI. Regarding education, the category coded as 9 (Master's degree) showed the highest proportion at 32%, followed by 23% for code 8 (Bachelor's degree). **Table 2** displays response time statistics (in seconds) for ARW, DCCS, and PSM across the 86 participants.

Table 1. Statistics of the 86 patients.

Category	Code	Description	Count	Percent	Details
Cognitive Status	0	Normal Cognition	67	78%	
	1	Cognitive Impairment	19	22%	
Gender	0	Female	60	70%	60 (46 Normal, 14 Impaired)
	1	Male	26	30%	26 (21 Normal, 5 Impaired)
Race	0	Asian	4	5%	
	1	Black	11	13%	
	2	Unreported	1	1%	
	3	White	70	81%	
Income Level	0	Level 2	2	2%	
	1	Level 3	3	3%	
	2	Level 4	3	3%	
	3	Level 5	8	9%	
	4	Level 6	10	12%	
	5	Level 7	12	14%	
	6	Level 8	21	24%	
	7	Level 9	26	30%	
	8	Declined to answer	1	1%	

Education Level	4	4 years	5	6%
	6	6 years	10	12%
	7	7 years	3	3%
	8	8 years	23	27%
	9	9 years	32	37%
	10	10 years	5	6%
	11	11 years	6	7%
	777	Not reported	2	2%

Table 2. Statistics of the item response times for ARW, DCCS, and PSM (in s).

Statistic	PSM	DCCS	ARW
Number of observations	86	86	86
Average	73	36.049	47.421
Standard deviation	34	14.909	11.619
Minimum value	30	15.301	31.906
25th percentile	52	26.823	39.207
Median (50th percentile)	63	33.014	44.966
75th percentile	84	40.248	52.492
Maximum value	233	103.959	94.313

Cognitive measures

MyCog DCCS

The MyCog DCCS evaluates executive function, with a focus on cognitive flexibility. Participants see two fixed reference images at the bottom of the screen that differ along two attributes—shape and color. Target images appear one at a time in the center, and respondents must match each target to one of the references based on either shape or color, depending on the current rule. The task includes five blocks: (1) two practice trials using the shape rule, (2) five pre-switch trials using shape, (3) two practice trials using the color rule, (4) five post-switch trials using color, and (5) 30 mixed trials where the required dimension (shape or color) changes in a pre-set random sequence. Accuracy and response times (RTs) are recorded for every trial. RT is measured in milliseconds from the appearance of the target to the participant's selection. Average completion time is approximately 3 min.

MyCog PSM

The MyCog PSM measures episodic memory. Participants are shown a sequence of 12 pictures, each depicting distinct activities, presented one by one in the center of the screen with accompanying spoken descriptions, and then placed into one of twelve designated boxes (encoding phase). After all pictures have been shown, they are rearranged in the center, and participants must drag them back into the boxes in the exact original order (recall phase). The task starts with a practice round using four pictures, followed by one main trial with twelve pictures. Average completion time is approximately 5 min.

Mobile toolbox ARW

The Mobile Toolbox ARW assesses inhibitory control and attentional processing. This 50-item task evaluates executive attention, involving voluntary, top-down control mechanisms that substantially overlap with executive function. Performance on similar dual-task formats has revealed deficits in presymptomatic familial Alzheimer's disease carriers and in cancer-related cognitive impairment. Participants must judge the left-or-right direction of a central arrow while ignoring surrounding arrows that may point incongruently. Average completion time is approximately 3 min.

Sequences

The Sequences task (MFS) evaluates working memory—the capacity to hold and manipulate unrelated items mentally. It is a recently created working memory assessment with a 30-item stimulus bank. Working memory difficulties are well-documented in various conditions involving cognitive decline, such as Parkinson's disease and HIV-related neurocognitive disorders. Participants must recall a sequence of letters and numbers and reorder them alphanumerically. Trials start with three characters and increase in difficulty up to a maximum of 10 characters. The test terminates after three consecutive errors. Scoring is based on the number of correct trials. Average completion time is approximately 5 min.

Number-symbol match

This task is presented in landscape mode. A legend at the top displays nine symbols, each linked to a digit from 1 to 9. Below, rows of symbols appear, and participants must quickly input the corresponding digit for each. The main score is the total correct responses within 90 seconds. As a 144-item measure, Number-Symbol Match engages multiple abilities, including perception, encoding, active memory transformation, retrieval, and decision-making. Its involvement of diverse cognitive processes results in high sensitivity but lower specificity across various impairments, such as age-associated changes, dementia, cancer-related cognitive issues, and multiple sclerosis. Average completion time is approximately 2 min.

Item and test level psychometrics

Internal consistency was evaluated using Cronbach's alpha for the five domains (DCCS, PSM, ARW, MFS, and NSM), yielding values of 0.866, 0.828, 0.593, 0.833, and 1, respectively. The first three tests showed complete data with no missing responses, whereas missing item responses rose progressively in MFS and NSM. This pattern in later tests likely reflects participant fatigue or the application of discontinuation criteria. Primary focus remains on the two MyCog tests (DCCS and PSM). **Table 3** displays correlations linking performance on these cognitive tasks (DCCS and PSM) with demographic variables across the 86 participants. Intercorrelations among correct-response scores for all five domains ranged between 0.2 and 0.6. **Table 4** summarizes means and standard deviations (SD) of correct responses across the five domains, stratified by demographic characteristics. Overall, no statistically meaningful differences emerged by gender, race, education, or income. Nevertheless, performance was consistently lower in the cognitively impaired group relative to the unimpaired group.

Table 3. Relationships between measures and demographics.

	Cognitive Impairment	Race	Education Level	Income Level	Gender	Age	DCCS Score	PSM Score
Cognitive Impairment	1	-0.136	-0.087	-0.335	-0.045	0.268	-0.284	-0.334
Education Level	-0.087	-0.105	1	-0.012	0.069	0.095	0.088	-0.07
Race	-0.136	1	-0.105	0.076	0.132	0.125	-0.014	0.157
Income Level	-0.335	0.076	-0.012	1	0.167	-0.132	0.109	0.097
Age	0.268	0.125	0.095	-0.132	0.067	1	-0.174	-0.27
Gender	-0.045	0.132	0.069	0.167	1	0.067	0.065	-0.331
PSM Score	-0.334	0.157	-0.07	0.097	-0.331	-0.27	0.268	1
DCCS Score	-0.284	-0.014	0.088	0.109	0.065	-0.174	1	0.268

Table 4. Measure differences (mean (SD)) among demographics.

Category	Subgroup	ARW	PSM	DCCS	NSM	MFS
Cognitive Status	Normal (0)	49 (3)	6 (3)	29 (2)	31 (8)	11 (3)
	Impaired (1)	45 (10)	4 (3)	27 (5)	21 (8)	6 (4)
Gender	Female	48 (4)	7 (3)	28 (3)	31 (9)	10 (4)
	Male	48 (8)	4 (2)	29 (3)	25 (9)	10 (4)
Race	Asian	49 (2)	5 (3)	27 (5)	30 (3)	10 (4)
	Black	47 (5)	5 (2)	29 (2)	29 (9)	10 (5)
	White	48 (6)	6 (3)	28 (3)	29 (9)	10 (4)
Education (years)	4	46 (3)	5 (4)	24 (6)	21 (7)	6 (4)
	6	44 (13)	5 (4)	27 (5)	2 (10)	8 (5)
	7	50 (1)	6 (4)	29 (1)	31 (9)	11 (2)
	8	48 (4)	6 (3)	29 (1)	29 (10)	10 (4)
	9	48 (4)	6 (3)	29 (3)	31 (9)	12 (3)

	10	50 (1)	6 (2)	30 (0)	31 (4)	10 (4)
	11	50 (0)	6 (4)	30 (1)	32 (5)	10 (1)
	Not reported (777)	48 (1)	4 (1)	30 (0)	28 (2)	12 (2)
Income Level	2	49 (1)	7 (1)	29 (1)	30 (8)	8 (4)
	3	35 (25)	3 (3)	27 (2)	1 (10)	6 (5)
	4	43 (9)	7 (3)	29 (1)	26 (6)	10 (1)
	5	48 (3)	6 (4)	29 (2)	27 (11)	9 (4)
	6	48 (2)	4 (3)	26 (6)	30 (8)	9 (5)
	7	49 (2)	6 (4)	28 (4)	29 (10)	11 (4)
	8	48 (3)	6 (3)	29 (2)	30 (9)	11 (3)

Study design

The research approach centers on leveraging the complete dataset to pinpoint the strongest predictors and modeling strategies for classifying impairment status. Key comparisons include predictive performance when relying on two versus five cognitive tasks. The investigation also tests whether derived composite scores alone suffice for reliable impairment classification (detailed in the following section). Finally, machine learning techniques are applied to assess whether additional variables enhance classification beyond the core cognitive measures.

An exploratory unsupervised step applied KMeans clustering to partition participants into homogeneous groups. The elbow technique guided, selection of the ideal cluster count, after which clusters were examined for associations involving composite scores, income, age, and diagnosed impairment.

Supervised modeling used established clinical impairment labels to isolate drivers of classification performance. This phase explored feature-label associations to highlight variables most predictive of impairment. Findings inform the identification of risk indicators and support the development of screening instruments for timely detection and management.

The item response theory models

The five domains featured different item counts: DCCS with 30 items, PSM with 12, ARW with 50, MFS with 30, and NSM with 144. To derive IRT-based ability estimates, a higher-order two-parameter item response theory (HOIRT) framework was implemented [10]. Multidimensional HOIRT estimation for the full dataset employed Markov chain Monte Carlo methods via the BMIRT software [11]. The five-dimensional HOIRT model (HOIRT-5D) produced an overall higher-order score (SSHO) plus dimension-specific scores labeled dccs, psm, arw, mfs, and nsm. The two-dimensional HOIRT model (HOIRT-2D) yielded an overall score (SSHO2D) together with dimension scores dccs2D and psm2D for the two targeted domains.

Composite scores

A straightforward approach to creating composite indices involves fitting linear regression models across possible combinations of k predictors.

$$Impaired = f(x_1, \dots x_k), \quad (1)$$

where the function f represents a linear regression model, and k denotes the count of predictors included. Five separate linear regression analyses were performed on the dataset, with resulting coefficients reported in **Table 5**. Probabilities were computed using the coefficients from **Table 5** in conjunction with the following formula:

$$P = \sigma(f(x_1, \dots x_k)) \quad (2)$$

$$= \frac{exp^{f(x_1,\dots x_k)}}{1 + exp^{f(x_1,\dots x_k)}} \quad (3)$$

Table 5. Coefficients associated with the five composite scores.

Predictor	Composite 5	Composite 4	Composite 3	Composite 2	Composite 1
Intercept (Classical)	-0.129	0.479	-0.0066	-0.07	1.08
SSHO			3.137	2.413	0.0607
arw				-0.0082	0.001

psm	-0.316	-1.165	-0.919	-0.0231	-0.004
nsm				-0.0146	-0.01
dccs	-0.115	-1.755	-1.381	-0.0279	-0.16
mfs				0.028	-0.03
arw-rt					0.000012
dccs-rt					0.0000009
psm-rt					-0.00149
age				0.048	0.0068

Composite1 is derived by integrating scores from the five content domains through traditional approaches. In particular, the count of correct responses was employed for the PSM, MFS, and NSM assessments. For the DCCS and ARW assessments, accuracy measures were computed by dividing correct responses by response time. These accuracy measures were subsequently incorporated into the composite calculation.

Composite2 represents a composite index that incorporates both cognitive performance scores and response time data for ARW, DCCS, and PSM. It is obtained via a linear combination of six primary scores (ssho, dccs, arw, psm, mfs, and nsm) along with their corresponding response times (arw-rt, dccs-rt, and psm-rt1). These values stem from five-dimensional parameter estimates generated by the HOIRT model, encompassing an overall estimate plus five domain-specific estimates via the BMIRT approach described by Yao [11, 12].

Composite3 consists of a linear combination involving the three parameters SSHO2D, dccs2D, and psm2D obtained from the two-dimensional HOIRT model, together with age.

Composite4 is formed as a linear combination of the three parameters SSHO2D, dccs2D, and psm2D from the two-dimensional HOIRT model.

Composite5 is constructed as a linear combination of DCCS accuracy (total correct responses divided by response time) and the number of correct responses from PSM.

Feature space

In the present investigation, multiple feature spaces were examined to determine the optimal set of variables for forecasting clinical impairment based on data from 86 participants. These spaces were constructed to evaluate the influence of various factors—including demographic variables, response times, and cognitive test results—on impairment prediction.

- Feature space F1 includes scores from all five content domains {SSHO, dccs, psm, arw, mfs, nsm}, response times for three domains {psm−rt, arw−rt, dccs−rt}2, and four demographic variables {race, gender, income, education}. Comprising 13 features, this space serves as the baseline owing to its inclusion of the broadest range of relevant variables.
- Feature space F2 removes the four demographic variables from F1, retaining six cognitive scores and three response times, for a total of nine features.
- Feature space F3 eliminates the three response times from F1, resulting in 10 features overall {race, gender, income, education, SSHO, dccs, psm, arw, mfs, nsm}.
- Feature space F4 consists of six features: the overall score plus the five domain-specific scores {SSHO, dccs, psm, arw, mfs, nsm}.
- Feature space F5 incorporates the four demographic variables {race, gender, income, education}, HOIRT-2D parameters for DCCS and PSM {SSHO2D, dccs2D, psm2D}, and response times for two domains {dccs–rt, psm–rt}, yielding nine features.
- Feature space F6 omits the two response times from F5, leaving seven features.
- Feature space F7 includes only the HOIRT-2D parameters for DCCS and PSM {SSHO2D, dccs2D, psm2D}, totaling three features.
- Feature space F8 comprises a single feature {Composite1}.
- Feature space F9 comprises a single feature {Composite2}.
- Feature space F10 comprises a single feature {Composite3}.
- Feature space F11 includes two features {age, Composite4}.
- Feature space F12 comprises a single feature {Composite5}.

- Feature space F13 utilizes three features drawn exclusively from the PSM assessment: {psm, psm−rt, age}.
- Feature space F14 utilizes three features drawn exclusively from the DCCS assessment: {dccs, dccs−rt, age}.
- Feature space F15 includes two features {age, SSH2D}.
- Feature space F16 includes five features {age, dccs2D, psm2D, dccs−rt, psm−rt}.

Machine learning models

The dataset comprising 86 cases was divided into training and testing sets using an 80/20 ratio, resulting in 68 training cases and 18 testing cases. Following this split, various supervised machine learning and deep learning algorithms were implemented and evaluated. For each model and model family, a systematic grid search was conducted on the training data via GridSearchCV from sklearn [13], employing the standard 5-fold cross-validation. To mitigate risks of overfitting and bias in model development, the training set was further partitioned into five subsets: four used for model training and one for validation and aggregation of performance metrics. With 5-fold cross-validation, each subset contained approximately 13 or 14 samples. The grid search assessed performance on the validation fold across iterations, then aggregated outcomes to identify the optimal hyperparameter configuration. Given the critical role of training data quality and volume in model performance, three categories of approaches were investigated: hyperplane-based separation, ensemble methods, and deep learning. Models implemented using the "sklearn" and "Keras" libraries were as follows: (1) hyperplane separation models (SVM, LRG, Tree, and KNN); (2) ensemble models (RF and GB); (3) deep learning models (ANN, FNN, and RNN).

Support vector machine (SVC)

SVC [14] represents a supervised machine learning technique applied to classification and regression tasks. The configuration employed the "rbf" (Radial Basis Function) kernel, along with four values for the regularization parameter C: 1, 10, 50, and 100.

Logistic regression (LRG)

LRG [15] constitutes a classification method grounded in regression that estimates the probability of a specific outcome. Hyperparameter tuning involved an L2 penalty, the standard "lbfgs" solver, and 20 levels of regularization strength ranging from 0.01 to 5.

Decision tree (Tree)

This supervised classification technique partitions data by selecting optimal thresholds across features. Seven values for max_depth were tested, spanning from 3 to 30.

K-Nearest neighbors (KNN)

KNN is a non-parametric classification approach that assigns labels according to the proximity of instances in the feature space.

Random forest (RF)

RF [16] is an ensemble supervised learning algorithm for classification that aggregates predictions from multiple bootstrapped decision trees [17]. Ten values for n_estimators were evaluated, ranging from 10 to 500.

Gradient boosting (GB)

GB [18] encompasses tree-based ensemble methods for classification. In contrast to RF, which builds trees independently, the boosting approach [18] develops trees sequentially in an additive manner. The GradientBoostingClassifier implementation from sklearn was utilized, with three n_estimators settings: 10, 30, and 50.

Artificial neural network (ANN)

ANN [19, 20] refers to computational models inspired by biological neural structures. The MLPClassifier from sklearn was applied. Grid search encompassed two hidden layer architectures ([(150, 100, 50), (120, 80, 40)]), two optimizers (["sgd", "adam"]), two activation functions (["tanh", "relu"]), three alpha values ([0.01, 0.3, 1]), and two learning rate schedules (["constant", "adaptive"]). Both tanh and relu serve as non-linear activation functions3.

Feedforward neural network (FNN)

FNN [21] designates a type of artificial neural network characterized by forward signal propagation. The model was built using Sequential from keras.models, incorporating five Dense layers each with 2000 units and 'relu' activation (rectified linear unit). The output layer employed 'sigmoid' activation. Unlike the other classifiers, this architecture required transforming target labels into one-hot encoded binary matrices for multi-class prediction. For instance, a score of 8 would set the

8th column to 1 and all others to 0. Given that the highest possible score across prompts 1–6 was 12, the encoded matrix contained 13 columns.

Recurrent neural network (RNN)

The RNN implementation [22] relied on the Keras library to construct a Sequential model [23] featuring one SimpleRNN layer. This recurrent architecture is suited for sequential data processing while retaining memory of prior inputs. The SimpleRNN layer comprised 200 units with 'tanh' (hyperbolic tangent) activation. Subsequently, three Dense layers with 'relu' activation were appended, followed by a final layer using 'sigmoid' activation. Training involved two epoch configurations, sized 64 and 32, respectively. Hyperparameter exploration included three batch sizes: 5, 12, and 32.

Evaluation criteria

Standard machine learning performance measures—precision, recall (or sensitivity), accuracy, F-score, and specificity—are derived from a confusion matrix, as presented in **Table 6**. Within medical applications, recall holds particular significance as it quantifies the proportion of correctly detected impaired patients. Specificity indicates the proportion of accurately identified individuals with normal cognition. These metrics are formally defined as follows:

$$Precision = \frac{TP}{FP + TP} \quad (4)$$

$$Recall = \frac{TP}{FN + TP} \quad (5)$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (6)$$

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

$$F\_Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

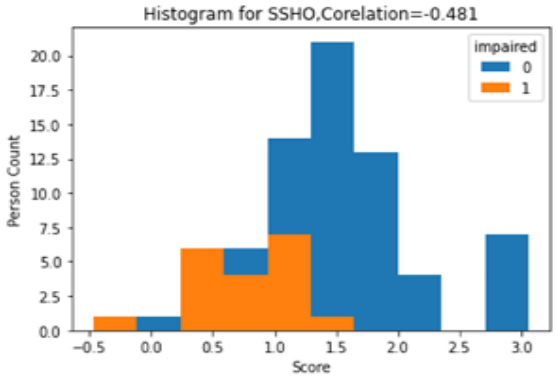
Table 6. Confusion matrix.

True	Predicted	
	0	1
0	TN	FP
1	FN	TP

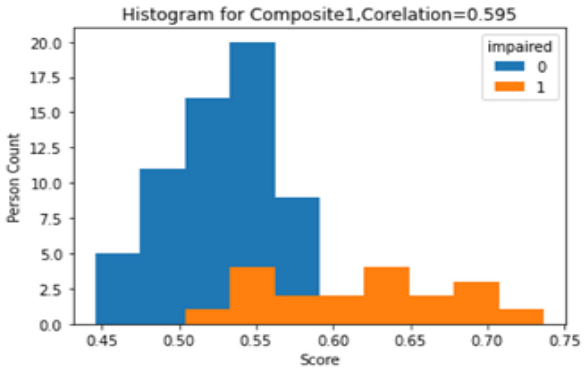
A widely used metric in machine learning is the receiver operating characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR). The area under the curve (AUC) quantifies the model's ability to separate classes. Higher AUC values indicate stronger discrimination between categories. For instance, an AUC of 0.7 implies a 70% probability that the model can correctly differentiate between impaired and non-impaired cases.

Results and Discussion

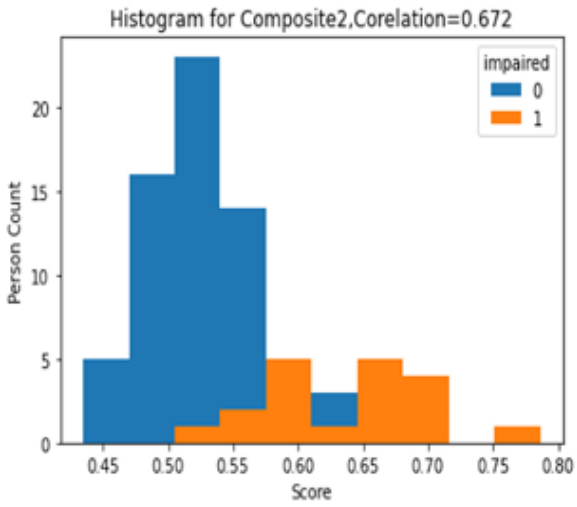
Figure 1 presents histograms with overlaid colored bars that denote impairment status across the sample. Each panel illustrates the distribution of one of six composite measures: SSHO, Composite1, Composite2, Composite3, Composite4, and Composite5. These plots display the frequency of values for each measure, including SSHO as one of the six. Such visualizations facilitate the detection of distributional patterns, score dispersion, and possible outliers. Blue bars represent cases with normal cognition, whereas orange bars denote impaired cases, enabling rapid visual comparison of score distributions between the two groups.



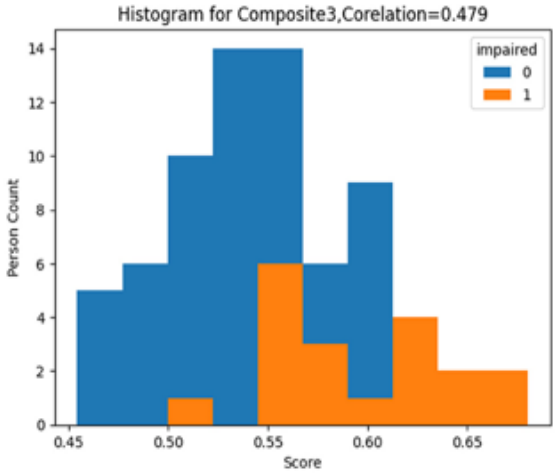
a)



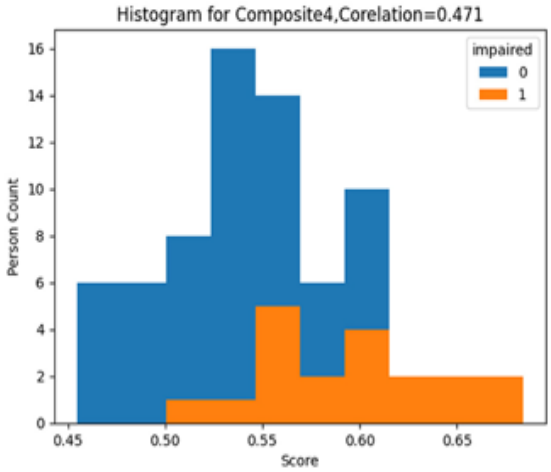
b)



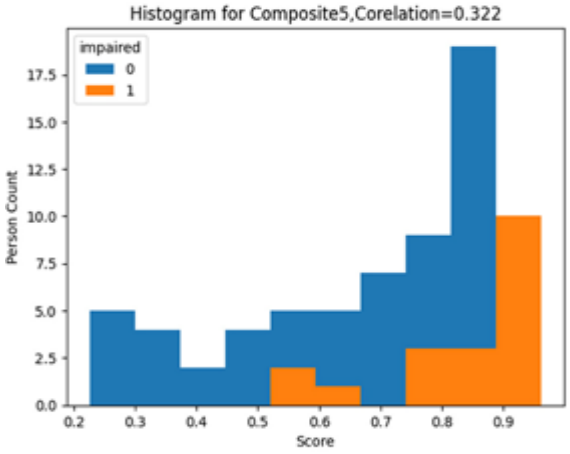
c)



d)



e)



f)

Figure 1. Histogram of six composite scores.

Unsupervised data analysis

The objective of the cognitive evaluations is to forecast impairment status from test performance without prior knowledge of actual clinical labels. Accordingly, exploratory pattern detection was pursued through unsupervised techniques. KMeans clustering from the scikit-learn library was applied to the dataset, testing cluster counts from 1 to 10 while minimizing within-cluster sum of squared distances (inertia). The resulting elbow plot **(Figure 2)** relates cluster number to inertia, revealing an inflection point at three or four clusters where further increases yield diminishing improvements.

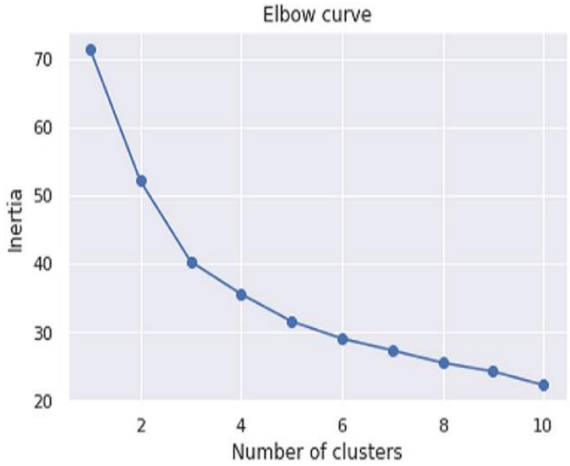
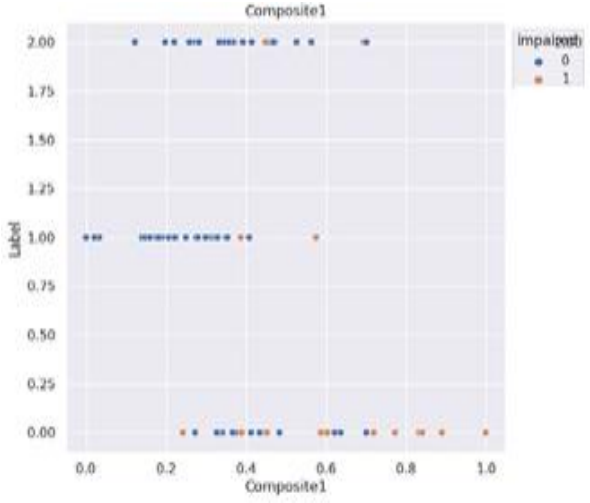


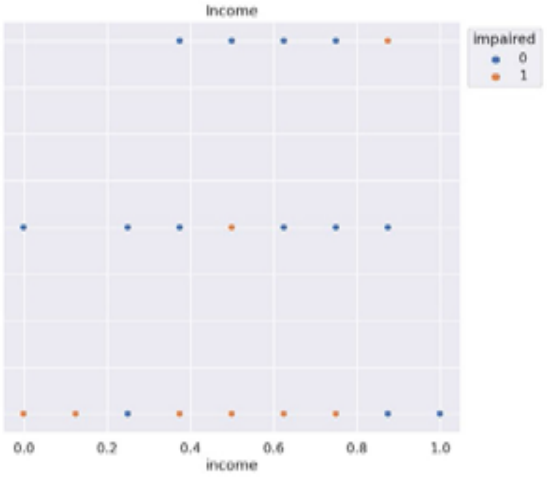
Figure 2. Elbow curve with 10 clusters.

Cluster assignment was then conducted using three clusters, labeling participants as 0, 1, or 2. **Figure 3** depicts the five composite scores (Composite1 through Composite5), SSHO, income, and age as a function of

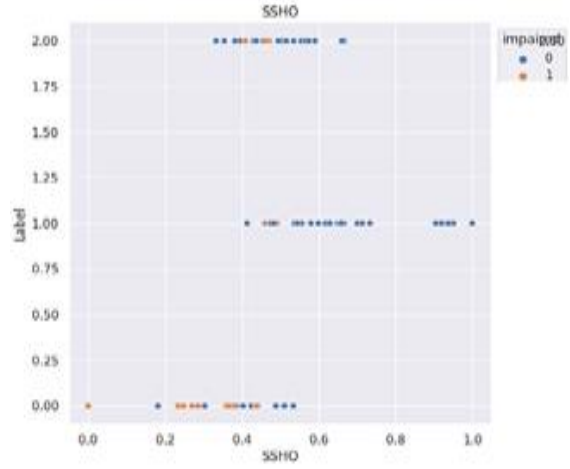
these cluster labels, with data points colored by verified clinical impairment status.



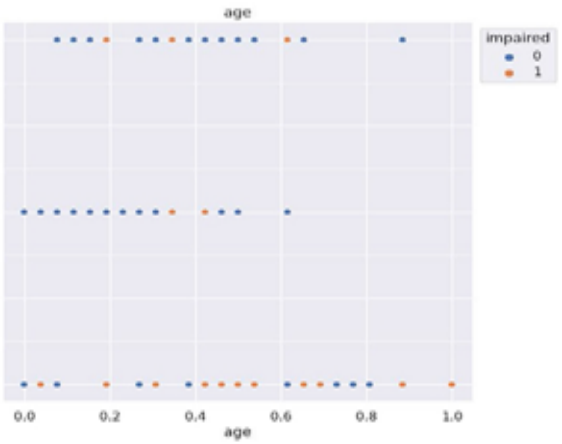
a)



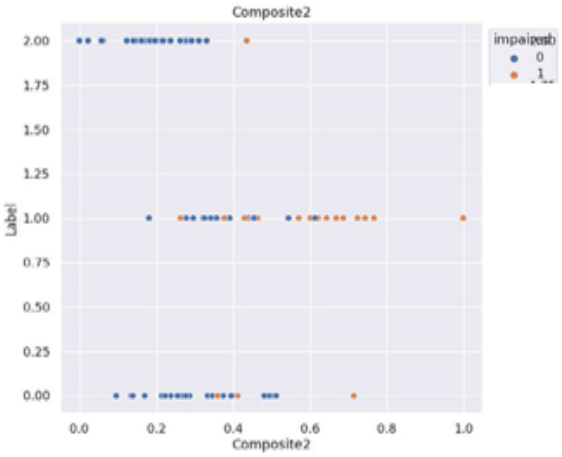
b)



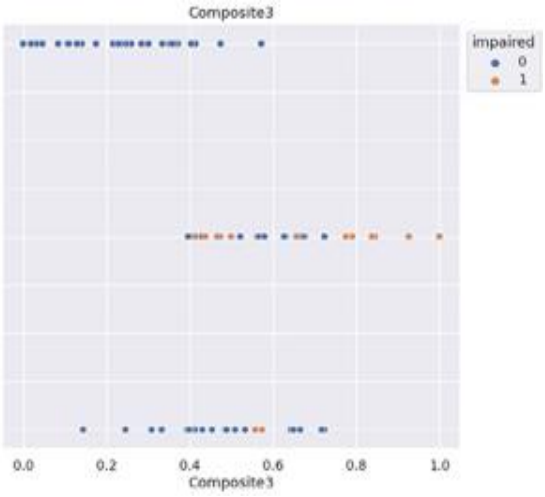
c)



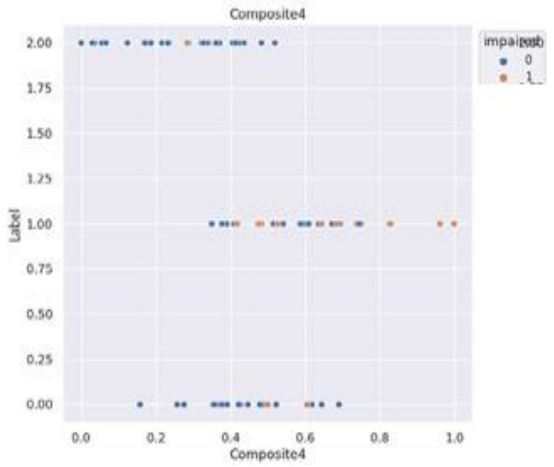
d)



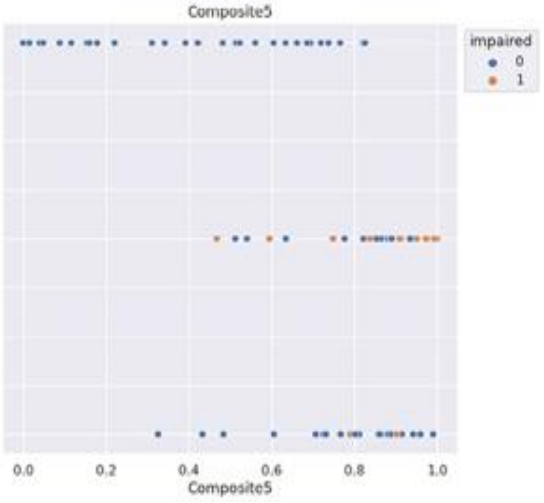
e)



f)



g)



h)

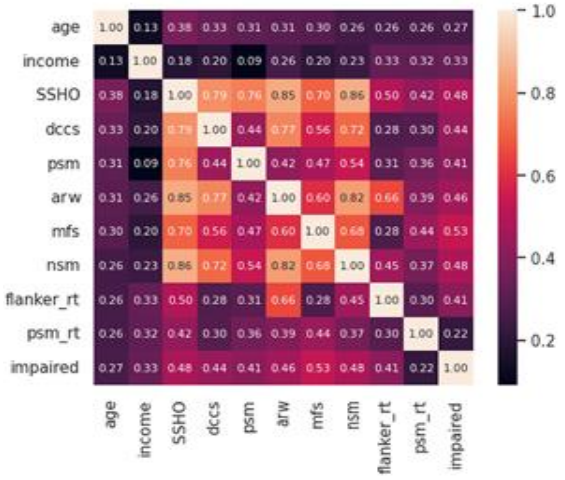
Figure 3. Five composite scores, SSHO, income, and age plotted against the three cluster labels.

The unsupervised clustering with three groups reveals certain patterns, though the ideal cluster count and threshold for label assignment remain somewhat ambiguous. Nevertheless, this approach offers valuable indications regarding patient groupings according to composite scores, SSHO, income, and age, as well as their association with verified clinical impairment status. Additional refinement of clustering techniques could enhance comprehension of the data's latent organization.

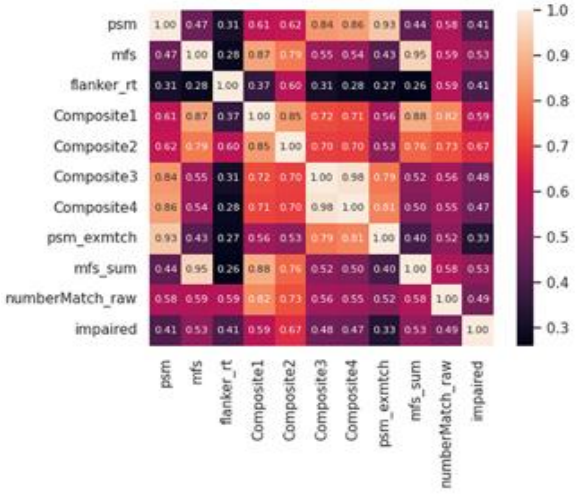
Supervised data analysis

To examine variables influencing the forecasting of clinical impairment within this dataset, supervised analyses were performed utilizing established clinical impairment labels. Exploring associations between predictors and outcomes helps identify the most predictive elements for impairment detection. Such investigations aid in pinpointing risk indicators and supporting the creation of screening instruments for timely identification and management.

Feature importance evaluation was carried out to determine the most influential variables across the included predictors. **Figure 4** presents two heatmaps highlighting the top 10 features. The initial heatmap incorporates HOIRT-5D parameters, demographic details, and response times, revealing the following ranking of influence: cognitive scores MFS, NSM, SSHO, ARW, DCCS, PSM, followed by their respective response times, and demographic factors income and age. The subsequent heatmap focuses on the five composite measures, identifying Composite2 and Composite1 as the most prominent.



a)



b)

Figure 4. Heat map of the 10 best features.

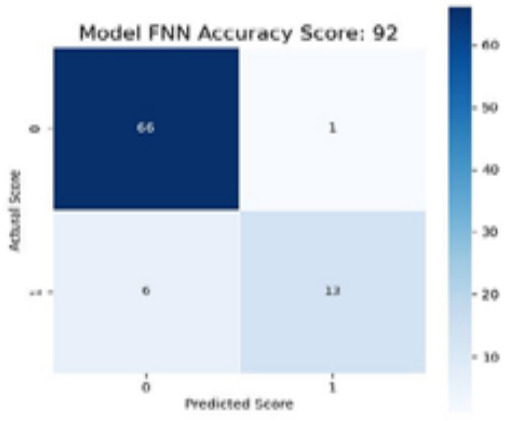
Supervised machine learning was implemented across the specified feature spaces using the described algorithms. In configurations that included not only cognitive scores but also demographic variables ("race," "gender," "income," "education") and response times—such as F1, F2, F3, F4, F5, F6, and F7—certain ensemble methods like GB and RF achieved perfect classification of all 86 cases.

Although model evaluation should primarily rely on test set performance, results across the entire dataset are reported in **Table 7** for the RF algorithm applied to all 16 feature spaces. Perfect classification of all 86 cases was attained with RF in feature spaces F1, F3, F5, and F6.

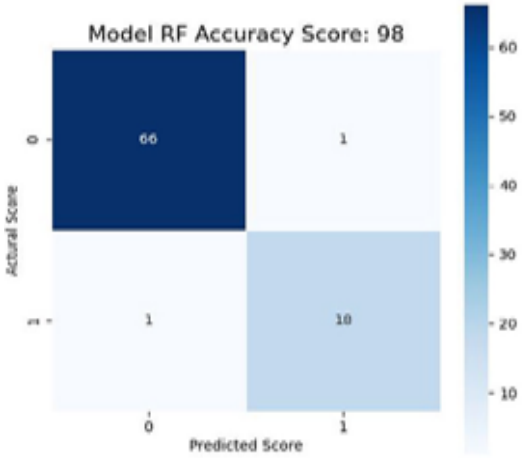
Table 7. Results for all data from RF for 16 feature spaces.

Feature	Recall	Precision	$F_{measure}$	AUC	QWK	Specificity
F1	1	1	1	1	1	1
F2	1	1	1	1	1	1
F3	0.988	0.988	0.988	0.993	0.967	0.985
F4	0.977	0.977	0.977	0.966	0.932	0.985
F5	1	1	1	1	1	1
F6	0.988	0.988	0.988	0.993	0.967	0.985
F7	0.977	0.977	0.977	0.966	0.932	0.985
F8	0.965	0.965	0.965	0.959	0.901	0.97
F9	0.93	0.93	0.93	0.918	0.805	0.94
F10	0.953	0.953	0.953	0.951	0.87	0.955
F11	0.965	0.965	0.965	0.94	0.897	0.985
F12	0.942	0.942	0.942	0.925	0.834	0.955
F13	0.988	0.988	0.988	0.993	0.967	0.985
F14	0.965	0.965	0.965	0.921	0.893	1
F15	0.988	0.988	0.988	0.974	0.966	1
F16	0.965	0.965	0.965	0.94	0.897	0.985

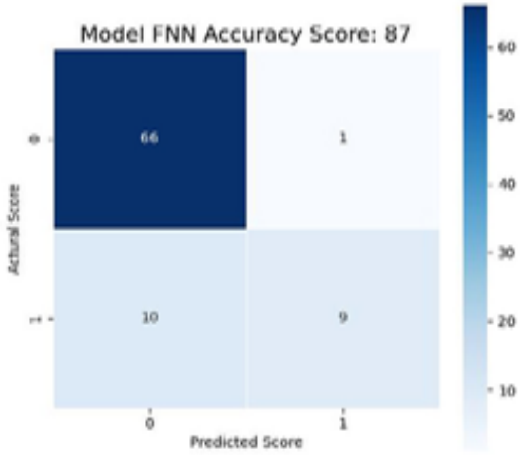
Figure 5 displays confusion matrices for the complete dataset (first row) and test set (second row) using FNN and RF models on Feature Space F4. The third and fourth rows show corresponding matrices for the full and test datasets using FNN and RF on Feature Space F7. Feature space F4 comprises six parameters (SSHO, dccs, psm, arw, mfs, nsm) obtained from the HOIRT-5D model across the five cognitive domains DCCS, PSM, ARW, MFS, and NSM. Feature space F7 includes three parameters (SSHO2D, dccs2D, psm2D) derived from the HOIRT-2D model based solely on the DCCS and PSM domains. In comparison to feature spaces F1 and F5, which incorporate demographic variables and response times and yielded perfect classification with RF, F4 exhibited one misclassification, while F7 showed two misclassifications.



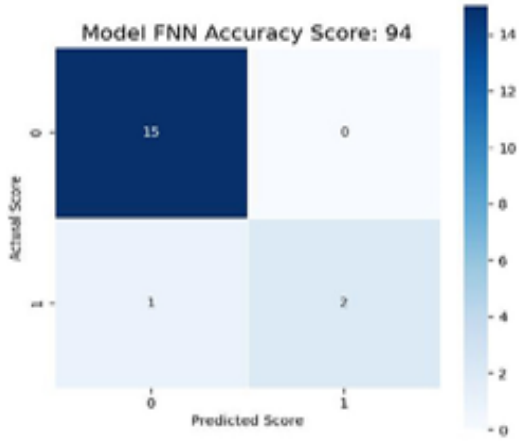
a)


Model RF Accuracy Score: 98
Actual Score
Predicted Score
66
1
1
10

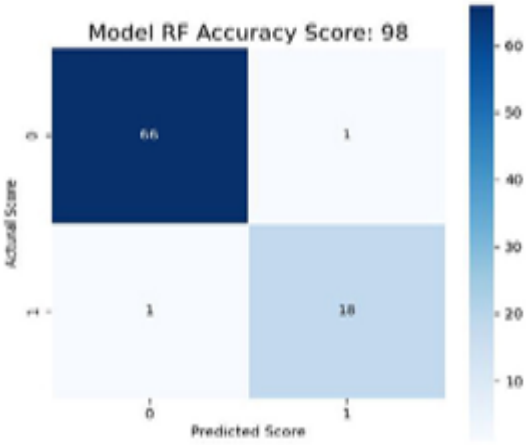
b)


Model FNN Accuracy Score: 87
Actual Score
Predicted Score
66
1
10
9

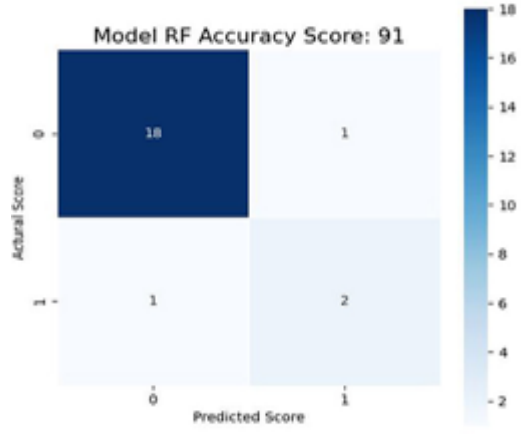
e)


Model FNN Accuracy Score: 94
Actual Score
Predicted Score
15
0
1
2

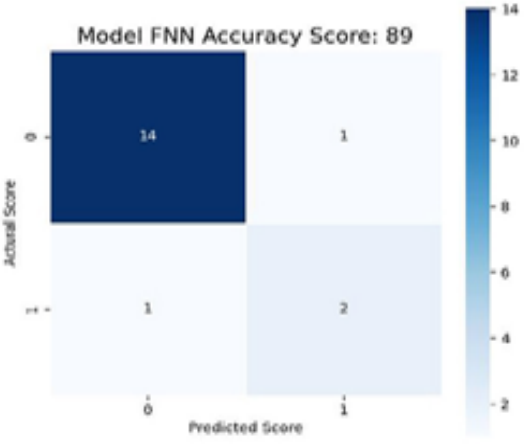
c)


Model RF Accuracy Score: 98
Actual Score
Predicted Score
66
1
1
18

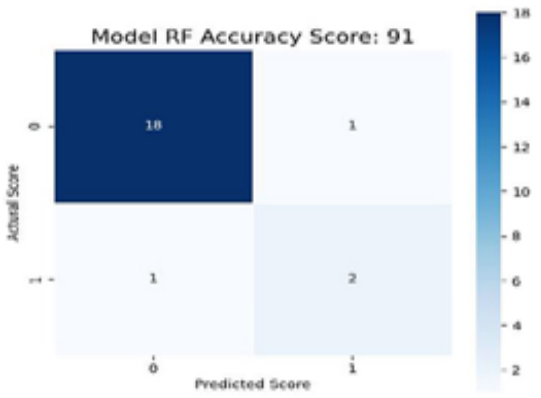
f)


Model RF Accuracy Score: 91
Actual Score
Predicted Score
18
1
1
2

d)


Model FNN Accuracy Score: 89
Actual Score
Predicted Score
14
1
1
2

g)



h)

Figure 5. Confusion matrix for the full dataset and test set using FNN and RF on feature space F4, and GB and RF on feature space F7.

The evaluation of feature spaces F7 through F16 focused primarily on configurations relying on only two cognitive tasks (DCCS and PSM) together with derived composite measures. **Figure 6** presents specificity, recall, AUC, F-measure, precision, and QWK values for every model across the complete dataset (top three rows) and the test subset (bottom three rows) for these feature spaces. The visualization highlights comparative model effectiveness.

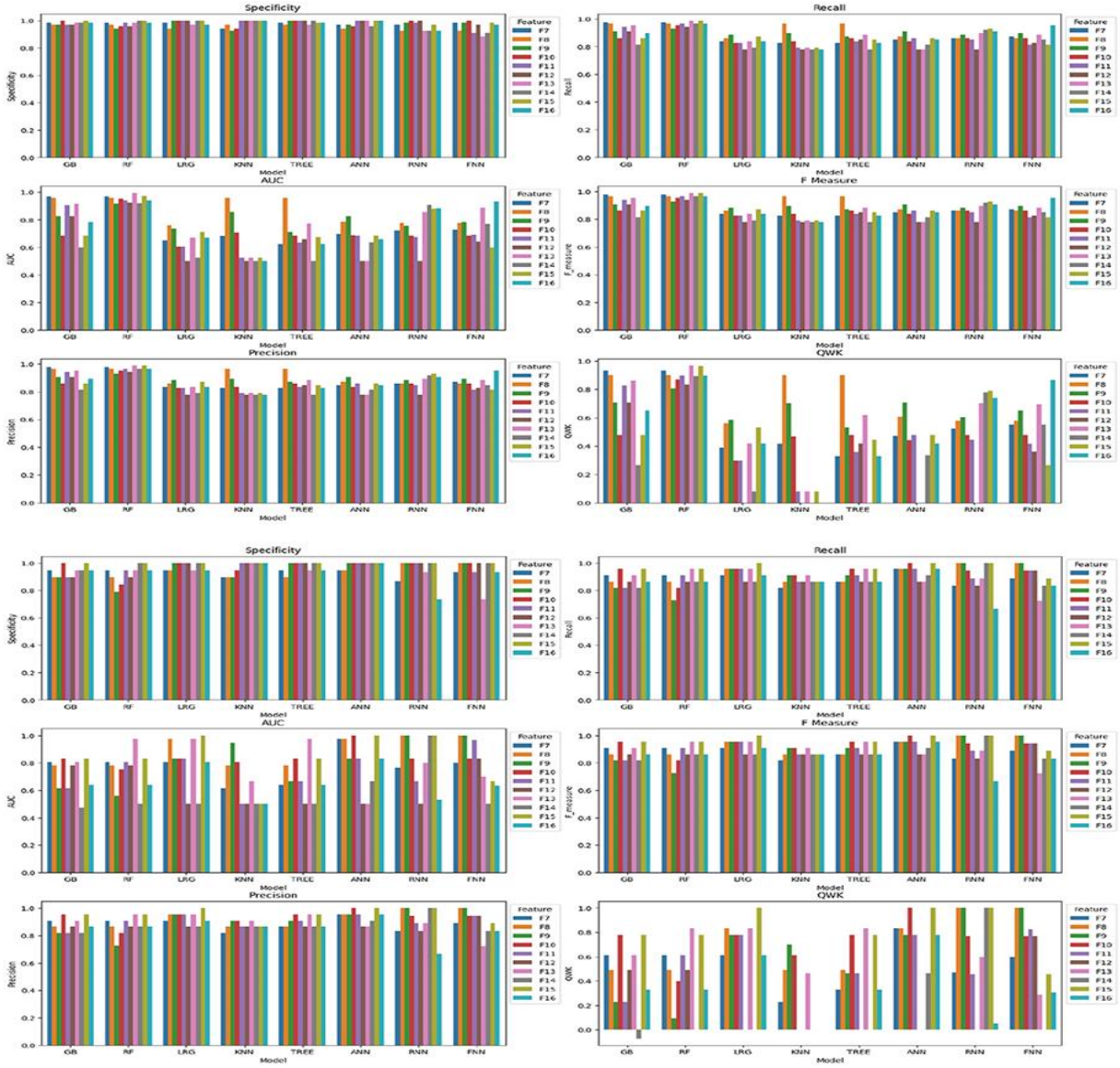


Figure 6. Specificity, recall, AUC, F-measure, precision, and QWK for all models on the full dataset (first three rows) and test dataset (last three rows) for feature spaces F7–F16.

Regarding the F-measure, which balances precision and recall, the ANN model achieved its strongest test-set results in feature spaces F3, F6, F10, and F15. Feature space F3 combined HOIRT-5D scores with demographic variables, F6 included HOIRT-2D scores with demographics, F10 used Composite3 (HOIRT-2D scores plus age), and F15 incorporated SSHO-2D and age.

When evaluated using QWK, which reflects agreement between predicted and true labels, the ANN model again showed superior test-set performance in the same four feature spaces: F3, F6, F10, and F15. Notably, age was a common element across all four.

For the abbreviated assessments centered on DCCS and PSM (feature spaces F15 and F16), the ANN model delivered the highest F-measure among all tested algorithms.

When restricting analysis to a single task (either DCCS or PSM), the ANN model's recall dropped markedly in feature spaces F13 and F14 compared with spaces that included age.

Although the dataset was limited to 86 participants and the results may not generalize to other samples, the analysis consistently indicated that incorporating age improved predictive accuracy.

Performance of composite scores

The next section evaluates the effectiveness of the composite scores using both threshold-based cutoffs and machine learning classifiers. Overall, the ANN model demonstrated the strongest results across most feature spaces, particularly when composite scores served as the sole predictors. **Table 8** reports test-set performance of the ANN model across all 16 feature spaces. Perfect classification on the test set was achieved in feature spaces F2, F6, F10, and F15.

Performance metrics were also computed for the entire set of 86 cases using the ANN model. Optimal cut points were determined for each of the five composite scores by minimizing total misclassifications. **Figure 7** contains six panels, each plotting score values against ANN-predicted impairment status for all 86 participants for Composite1 (F8), Composite2 (F9), Composite3 (F10), Composite4 (F11), Composite5 (F12), and SSHO-2D, respectively. The computed optimal cut point for each measure is indicated by a red vertical line. **Table 9** lists the optimal cut scores and corresponding number of misclassified cases for all 86 participants using linear thresholds for the five composite scores and SSHO-2D. The table also includes misclassification counts from the ANN model on both test and full datasets, as well as corresponding results from the RF model across all 86 cases.

Table 8. Results for the test data from ANN for 16 feature spaces.

Feature	Recall	Precision	$F_{measure}$	AUC	QWK	Specificity
F1	0.955	0.955	0.955	0.833	0.776	1
F2	0.955	0.955	0.955	0.833	0.776	1
F3	1	1	1	1	1	1
F4	0.955	0.955	0.955	0.974	0.831	0.947
F5	0.955	0.955	0.955	0.833	0.776	1
F6	1	1	1	1	1	1
F7	0.955	0.955	0.955	0.974	0.831	0.947
F8	0.955	0.955	0.955	0.974	0.831	0.947
F9	0.955	0.955	0.955	0.833	0.776	1
F10	1	1	1	1	1	1
F11	0.955	0.955	0.955	0.833	0.776	1
F12	0.864	0.864	0.864	0.5	0	1
F13	0.864	0.864	0.864	0.5	0	1
F14	0.909	0.909	0.909	0.667	0.463	1
F15	1	1	1	1	1	1
F16	0.955	0.955	0.955	0.833	0.776	1

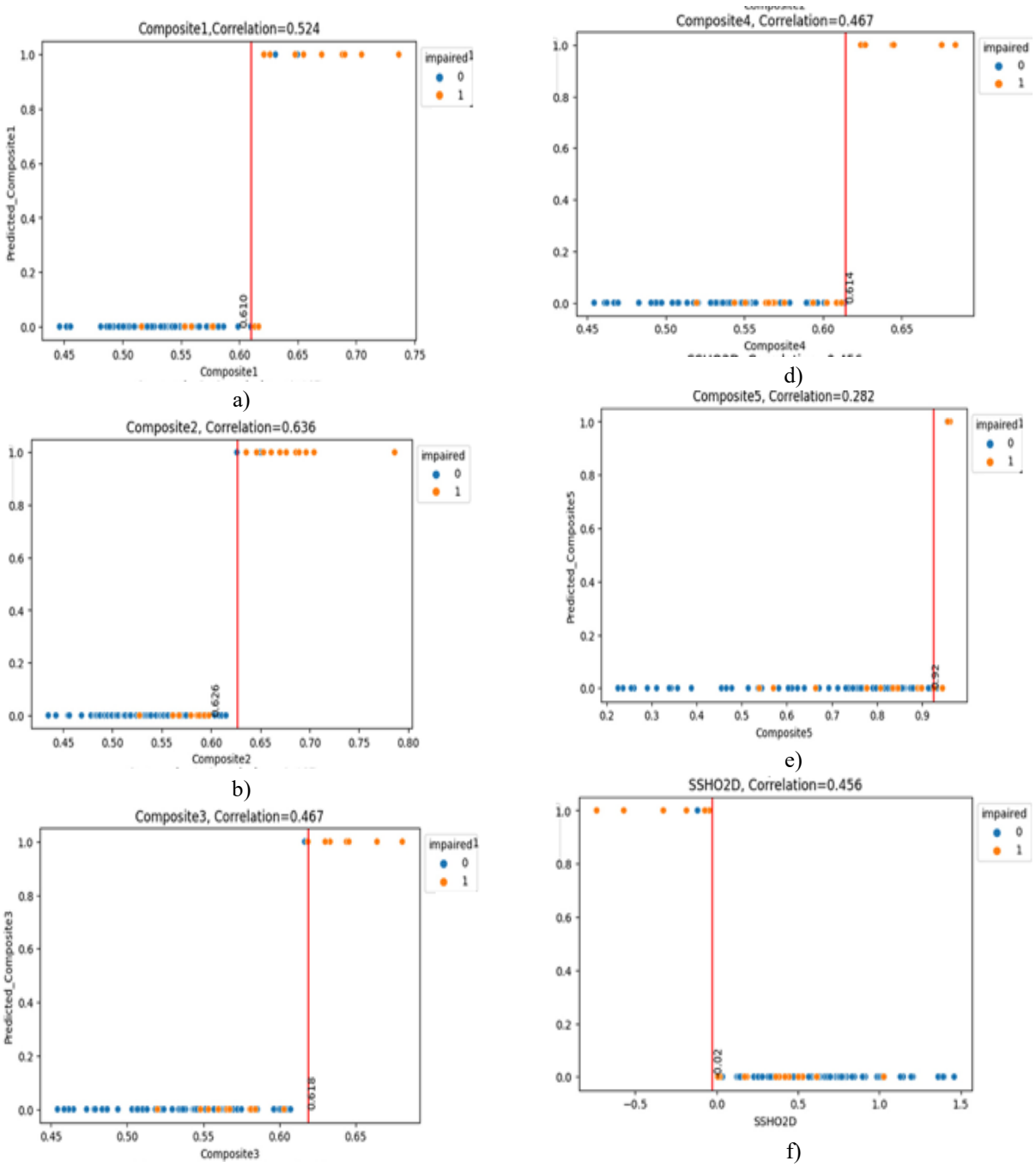


Figure 7. Scores versus ANN-predicted impairment with optimal cut points for the six composite measures across all 86 participants.

Table 9. Number of misclassified patients using both machine learning models and linear cut points for the six composite measures.

Model/Feature	Correlation Coefficient (Liner Cut)	Points Awarded	Mismatches on Test Dataset	Mismatches on Full Dataset	Rank Among ML Models (Best Performer)
F8	0.61	1	11	11	3rd (Random Forest)

F9	0.626	1	8	9	6th (Random Forest)
F10	0.618	0	14	12	4th (Random Forest)
F11	0.614	1	12	14	3rd (Random Forest)
F12	0.925	1	17	14	5th (Random Forest)
SSHO2D	-0.026	0	12	14	1st (Random Forest)

Tables 10 and 11 provide detailed information for all 86 participants, including their cognitive assessment scores, ethnicity, clinical diagnosis, and predicted impairment labels obtained from: RF model on feature space F10 (Composite3), linear cut point applied to SSHO-2D, linear cut point applied to Composite5, and RF model on feature space F16 (which includes dccs2D, psm2D, age, and corresponding response times).

Table 10. Number of misclassified patients using both machine learning models and linear cut points for the six composite measures.

Impairment Status	SSHO2D Prediction Score (Random Forest)	Composite5 Prediction Score (Linear)	Composite5 Prediction Flag	F16 Prediction Score (Linear)	F16 Prediction Flag	Unknown Value	Unknown Flag	Gender	Age	Education Level	Income Level
0	0.597	0.229	0	0.92	0	0	0	F	77	9	8
0	0.513	0.899	0	0.639	0	0	0	F	78	11	9
0	0.564	0.473	0	0.791	0	0	0	M	80	8	7
0	0.562	0.51	0	0.611	0	0	0	F	75	8	9
0	0.546	0.703	0	0.798	0	0	0	F	75	6	4
0	0.507	0.897	0	0.358	0	0	0	F	67	8	5
0	0.618	-0.117	0	0.862	1	0	0	F	75	9	6
0	0.561	0.776	0	0.882	0	0	0	F	88	*	7
0	0.519	0.567	0	0.834	0	0	0	F	67	6	8
0	0.557	0.523	0	0.582	0	0	0	M	79	11	9
0	0.56	0.769	0	0.831	0	0	0	F	86	8	8
0	0.53	0.698	0	0.765	0	0	0	M	71	9	9
0	0.606	0.027	0	0.898	0	0	0	F	77	8	6
0	0.487	1.017	0	0.466	0	0	0	M	70	8	8
0	0.597	0.14	0	0.875	0	0	0	F	79	9	8
0	0.585	0.157	0	0.624	0	0	0	F	74	4	8
0	0.596	0.036	0	0.855	0	0	0	F	69	9	9
0	0.605	0.125	0	0.86	0	1	0	M	84	9	9
0	0.618	0.125	0	0.92	0	0	0	M	83	8	9
0	0.551	0.347	0	0.843	0	0	0	F	68	7	6
0	0.572	0.377	0	0.603	0	0	0	F	74	9	9
0	0.604	0.256	0	0.88	0	0	0	F	87	9	D

0	0.6	0.281	0	0.933	0	0	1	M	81	9	8
0	0.575	0.331	0	0.859	0	0	0	M	77	9	8
0	0.617	0.135	0	0.956	0	0	1	M	90	11	8
0	0.601	0.144	0	0.9	0	0	0	M	78	6	7
0	0.607	0.017	0	0.914	0	0	0	F	75	9	9
0	0.583	0.415	0	0.856	0	0	0	F	83	8	4
1	0.56	0.386	1	0.843	0	1	0	F	68	8	8
1	0.581	0.365	1	0.891	0	1	0	M	83	10	9
1	0.553	0.612	1	0.57	0	1	0	F	78	4	6
1	0.566	0.53	1	0.808	0	1	0	M	76	9	7
1	0.581	0.446	1	0.835	0	1	0	F	86	8	2
1	0.52	1.028	1	0.539	0	1	0	F	76	10	9
1	0.56	0.617	1	0.776	0	1	0	F	85	9	8
1	0.548	0.421	1	0.664	0	1	0	F	67	7	3
1	0.681	-0.737	1	0.958	1	1	1	F	80	6	6
1	0.644	0.022	1	0.963	0	0	1	F	93	4	5
1	0.619	-0.045	0	0.957	1	0	1	F	84	8	6
1	0.567	0.503	1	0.896	0	1	0	F	78	6	6
1	0.63	0.007	1	0.943	0	1	1	F	90	8	3
1	0.603	0.187	1	0.778	0	1	0	F	79	8	8
1	0.584	0.173	1	0.9	0	1	0	M	72	8	9
1	0.646	-0.327	1	0.847	1	1	0	F	81	4	5
1	0.633	-0.072	1	0.944	1	1	1	F	84	8	5
1	0.664	-0.57	1	0.946	1	1	1	M	75	4	7
1	0.63	-0.185	1	0.927	1	1	1	M	72	6	3

*Indicates missing value.

Table 11. Results for all patients.

Impairment Status	SSHO2D Prediction Score (Random Forest)	Composite5 Prediction Score (Alternative Model)	Composite5 Prediction Flag	F16 Prediction Score (Linear)	F16 Prediction Flag	Unknown Flag 1	Unknown Flag 2	Gender	Age	Education Level	Income Level
0	0.522	0.734	1	0.756	0	0	0	F	70	9	8
0	0.478	1.382	0	0.344	0	0	0	F	79	9	9
0	0.539	0.799	0	0.693	0	0	0	F	70	9	8
0	0.534	0.667	0	0.737	0	0	0	F	68	9	8
0	0.543	0.751	0	0.761	0	0	0	M	77	10	6
0	0.546	0.679	0	0.731	0	0	0	F	80	8	7

0	0.57	0.632	1	0.874	0	0	0	M	80	9	9
0	0.51	0.907	0	0.545	0	0	0	M	70	9	9
0	0.546	0.659	0	0.79	0	0	0	F	74	8	8
0	0.53	0.847	0	0.713	0	0	0	F	72	8	9
0	0.454	1.399	0	0.262	0	0	0	F	71	9	7
0	0.51	0.914	0	0.605	0	0	0	F	75	9	6
0	0.548	0.668	1	0.831	0	0	0	F	73	11	7
0	0.552	0.527	0	0.882	0	0	0	M	75	9	9
0	0.473	1.362	0	0.311	0	0	0	F	76	9	8
0	0.544	1.017	0	0.693	0	0	0	F	83	6	2
0	0.508	1.029	0	0.515	0	0	0	F	74	9	9
0	0.465	1.393	0	0.291	0	0	0	F	74	8	5
0	0.584	0.317	0	0.581	0	0	0	F	83	9	9
0	0.503	1.03	0	0.339	0	0	0	F	78	8	9
0	0.538	0.696	0	0.455	0	0	0	F	75	6	4
0	0.462	1.207	0	0.255	0	0	0	F	70	6	7
0	0.53	0.617	0	0.747	0	0	0	M	69	6	5
0	0.535	0.756	0	0.769	0	0	0	F	78	9	9
0	0.486	1.195	0	0.389	0	0	0	F	70	9	8
0	0.537	0.833	0	0.769	0	0	0	F	78	10	8
0	0.51	0.784	0	0.672	0	0	0	F	68	9	7
0	0.523	0.682	0	0.536	0	0	0	F	69	8	8
0	0.544	0.563	0	0.817	0	0	0	M	71	9	9
0	0.484	1.147	0	0.237	0	0	0	F	68	8	7
0	0.494	1.015	0	0.478	0	0	0	F	71	9	9
0	0.546	0.564	0	0.824	0	0	0	M	72	*	7
0	0.565	0.616	0	0.863	0	0	0	M	78	10	9
0	0.544	0.477	0	0.876	0	0	0	M	74	8	6
0	0.459	1.46	0	0.226	0	0	0	F	67	11	9
0	0.506	0.986	0	0.581	0	0	0	F	73	9	5
0	0.524	1.002	0	0.672	0	0	0	M	81	9	5
0	0.479	1.13	0	0.389	0	0	0	F	70	7	8
0	0.548	0.584	0	0.879	0	0	0	M	81	11	9

*Indicates missing value.

Linear threshold approaches for impairment detection clearly underperform compared to machine learning classifiers. Misclassification rates using simple cut points on composite scores consistently exceed those achieved by machine learning algorithms. In particular, the ANN model applied to feature space F15 (incorporating SSHO2D and age) yielded flawless predictions.

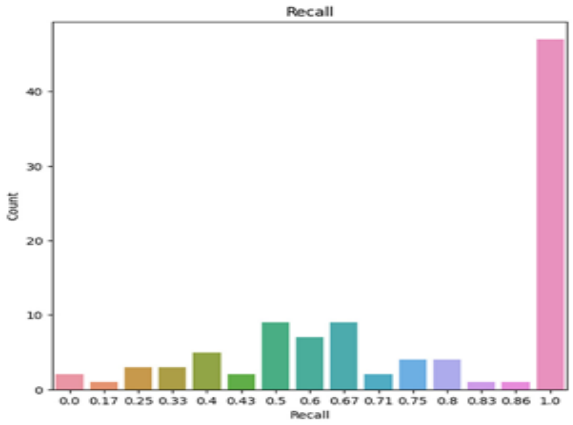
Bootstrap and replications

To address potential variability arising from the limited sample of 86 participants, 100 bootstrap resamples were generated from the original dataset, preserving sample size in each iteration while retraining models. Average performance metrics and their standard deviations across these replications are reported in **Table 12** for selected feature spaces using RF and ANN models. **Figure 8** illustrates the distribution of precision, recall, accuracy, F-measure, and specificity across the 100 bootstraps for the RF model on features {age, SSHO2D}, with mean values of 0.803, 0.758, 0.902, 0.742, and 0.951, respectively. **Figure 9** presents analogous distributions for the ANN model on features {age, dccs2D, psm2D,

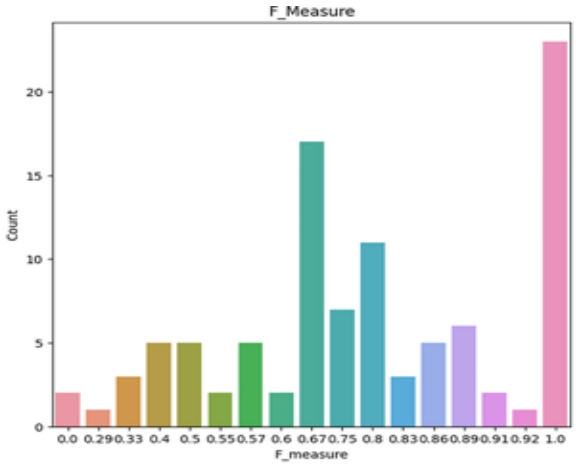
dccs−rt, psm−rt}, achieving means of 0.824, 0.706, 0.899, 0.733, and 0.955, respectively.

Table 12. Means and standard deviations across 100 replications on test data for RF and ANN models in feature spaces F15 and F16.

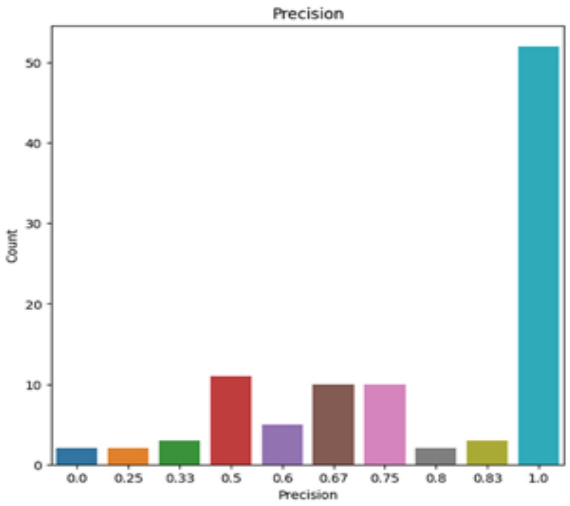
Model	**Artificial Neural Network**	**Random Forest**
Feature Set	**{dccs2D, psm2D} ∪ {age, rt}**	**{SSHO2D, age} ∪ {age, rt}**
Performance Metric	**SD**	**Mean**
----------	----------	----------
Precision	0.246	0.803
Recall	0.272	0.758
Accuracy	0.079	0.902
F-measure	0.223	0.742
Specificity	0.067	0.951
Model	Artificial Neural Network	Random Forest
Feature Set	{dccs2D, psm2D}	{SSHO2D, age}
Performance Metric	SD	Mean
----------	----------	----------
Precision	0.226	0.83
Recall	0.262	0.718
Accuracy	0.077	0.9
F-measure	0.212	0.738
Specificity	0.056	0.956



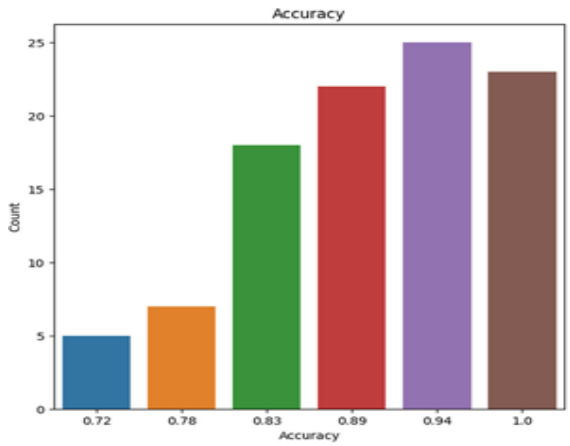
a)



b)



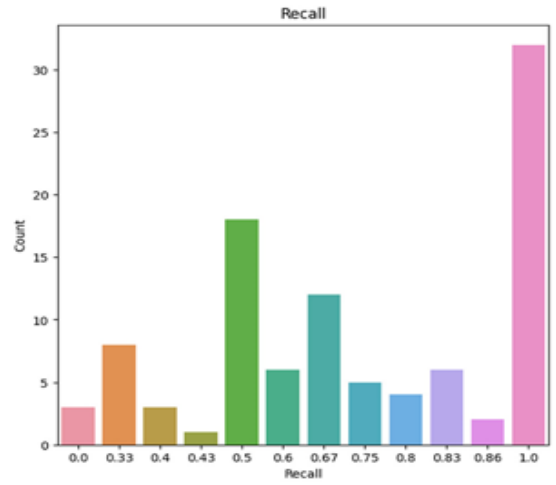
c)



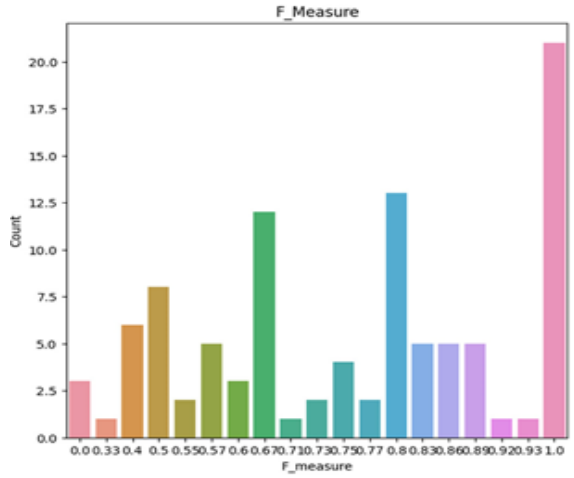
d)

Figure 8. Distribution of counts for recall, F-measure, precision, and accuracy across 100

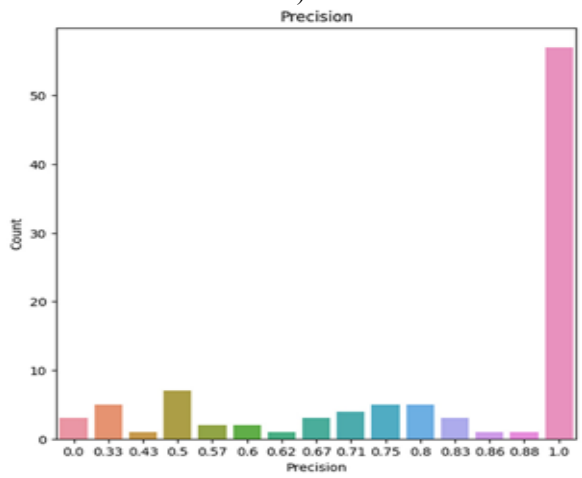
bootstrap samples using the RF model on features including SSHO2D and age.



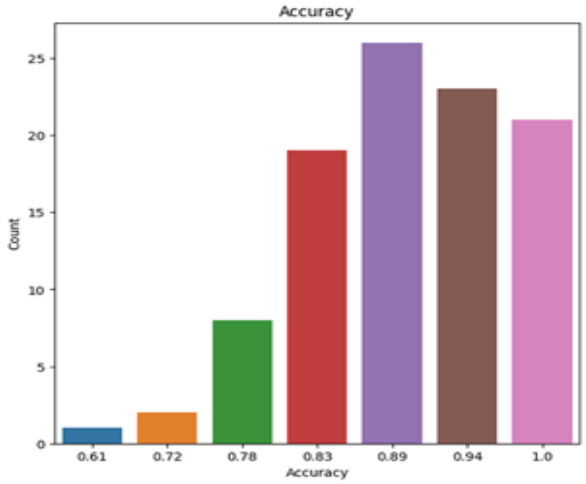
a)



b)



c)



d)

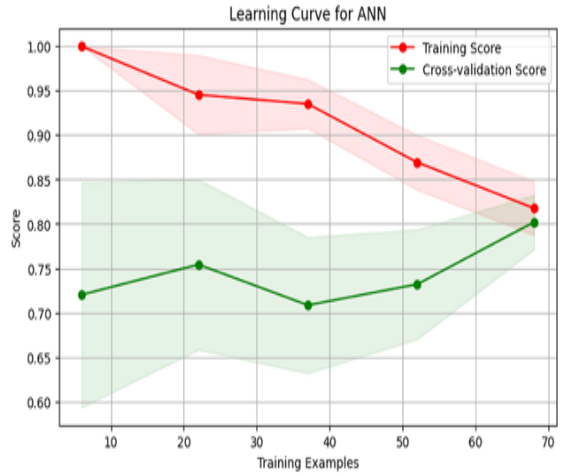
Figure 9. Distribution of counts for recall, F-measure, precision, and accuracy across 100 bootstrap samples using the ANN model on features including dccs2D, psm2D, dccs−rt, psm−rt, and age.

An additional 20 bootstrap iterations were conducted for an ANN model trained on features comprising SSHO2D, age, income, race, 30-DCCSrt, and psm-rt. The resulting metrics were precision 0.833, recall 0.748, accuracy 0.944, F1 0.777, and specificity 0.983. These findings suggest that incorporating item-level response times alongside demographic variables enhances predictive capability, especially for recall.

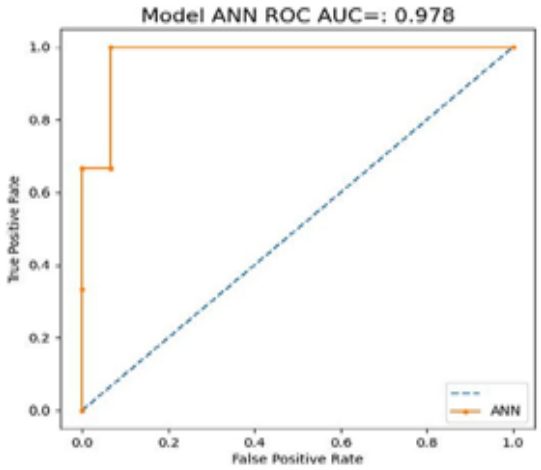
The investigation further examined the effect of reducing assessed cognitive domains. When features included demographics and total response times, recall declined from 0.772 (five domains) to 0.728 (two domains). A similar drop—from 0.762 to 0.674—was observed when using demographics with the overall HOIRT score.

To promote straightforward integration of the scoring system and avoid difficulties in obtaining sensitive demographic details (gender, education, income, race), the final models prioritized readily available predictors: age, correct counts for PSM and DCCS, and their total response times ("age," "psm," "DCCS," "dccs_rt," and "psm_rt"). This design emphasizes ease of deployment while bypassing demographic collection challenges. Performance of ANN and GB models, including AUC and confusion matrix details, is depicted in **Figure 10**. The upper panel shows the learning curve for the optimized ANN configuration: solver='sgd', activation='tanh', alpha=0.01, hidden_layer_sizes=(150, 100, 50), and max_iter=1000. Close alignment between

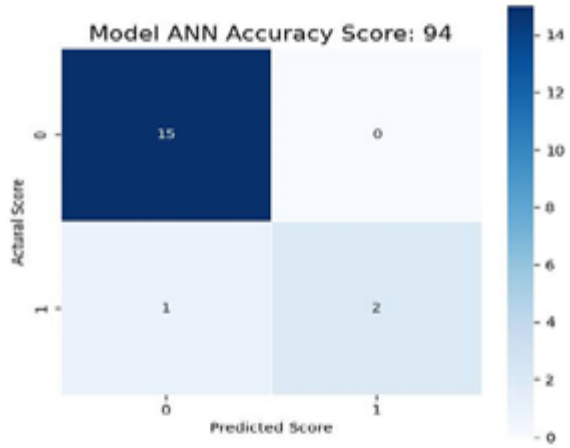
training and validation curves indicates robust generalization [24].



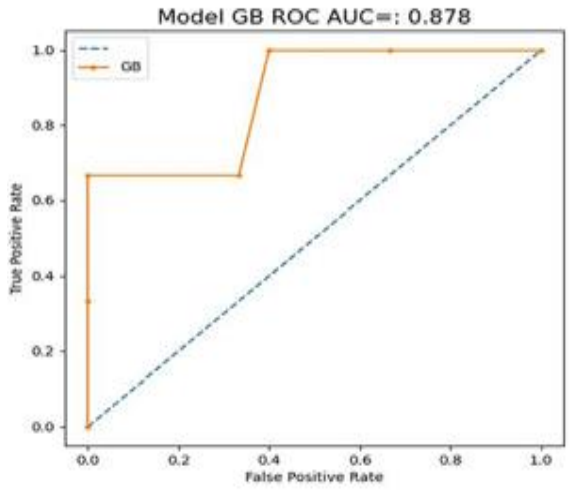
a)



b)



c)



d)

Figure 10. Learning curves, ROC curves, and confusion matrices for artificial neural network (ANN) and gradient boosting (GB) models, based on features including age, number of correct responses, and total response times from the DCCS and PSM tasks.

This study analyzed clinical data from 86 patients who completed five cognitive tasks. Unsupervised clustering was applied to detect underlying patterns and group patients according to selected features. Composite scores were calculated using both higher-order two-parameter item response theory (HOIRT) models and traditional correct-score approaches. Five different linear combinations of these scores were explored to assess their potential for detecting cognitive impairment.

Supervised machine learning techniques were used to classify impairment in the 86 patients across multiple algorithms and feature sets. Random forest (RF) and gradient boosting (GB) models generally achieved the highest performance on feature spaces F1–F4, which incorporated a larger set of variables, including the six task scores, race, gender, income, education, and response times. With these feature spaces, the models accurately classified all 86 patients.

Optimal thresholds for each of the six composite scores were established by minimizing classification errors among the 86 patients. The resulting misclassification counts were 11, 9, 12, 14, 14, and 14 for Composite1, Composite2, Composite3, Composite4, Composite5, and SSHO2D, respectively. Composite2—constructed from six HOIRT-5D scores, response times, and age—produced the fewest errors. Using only SSHO2D (cut-off

-0.026) or Composite4 (cut-off 0.925) as a single predictor yielded 14 misclassifications. In contrast, applying an RF model with these same scores as the sole input reduced errors to 4. These results indicate that employing composite scores within a machine learning framework can enhance prediction accuracy over traditional cut-point methods or individual domain scores.

When composite scores alone served as features, the RF model generated substantially fewer misclassifications than threshold-based approaches. This supports the value of composite scores as predictors of impairment and highlights the utility of algorithms like RF for clinical impairment detection.

The study also compared predictive accuracy between patients completing all five assessments (DCCS, PSM, ARW, MFS, and NSM) and those completing only two (PSM and DCCS). Performance improved with additional assessments overall. Nevertheless, for some models, using just the two tasks (DCCS and PSM) combined with age and task scores produced similar outcomes. When composite scores from five versus two assessments were contrasted, incorporating only the two tasks plus demographic variables with the three corresponding scores (SSHO2D, dccs2D, and psm2D) enabled the RF model to correctly classify all 86 patients. In comparison, feature space F7—which included only the three scores without demographics—misclassified one normal patient as cognitively impaired and one impaired patient as normal. Thus, adding demographic factors such as age substantially improves model precision in cognitive impairment prediction.

Given the limited sample size for training, result stability raises concerns. To address this, bootstrap methods were applied to selected feature spaces. Performance metrics were derived from 100 bootstrap samples, enabling computation of means and standard deviations. The random forest (RF) model, using features comprising a composite score and age, achieved precision, recall, accuracy, F1-score, and specificity of 0.803, 0.758, 0.902, 0.742, and 0.951, respectively. This composite score was obtained from a two-parameter higher-order item response theory (HOIRT) model based on the two assessments, DCCS and PSM.

The chosen final model acknowledges the practical difficulties in obtaining certain demographic variables, such as gender, education, income, and race. It was therefore designed to rely on a minimal, accessible feature set—namely age, raw response accuracy, and response times—centered primarily on the DCCS and PSM tasks. Visual displays of AUC values and confusion matrices for the artificial neural network (ANN) and gradient boosting (GB) models demonstrate strong overall performance.

Overall, the investigation highlights the value of including demographic variables and item-level response times in machine learning feature sets to enhance predictive capability [25, 26]. It indicates that depending solely on a fixed threshold for any composite score, no matter how carefully constructed, is likely suboptimal. Rather, applying machine learning algorithms to HOIRT-derived scores together with features like age can produce highly effective classification systems.

Larger training datasets are typically required to boost both performance and generalizability of machine learning approaches. For real-world deployment, assembling a diverse and representative dataset is essential to train models that can robustly handle new cases.

Future work aims to incorporate temporal elements from patients' PSM tasks—such as sequences of actions during testing and survey answers—into machine learning frameworks to detect atypical patterns and identify individuals with Alzheimer's disease or cognitive impairment (AD/CI). Recent literature shows notable progress in developing advanced machine learning techniques.

Jiao and colleagues [27] provided a comprehensive review focused on measurement and assessment in this domain. More recently, Ke and colleagues [28] investigated deep convolutional neural networks (CNN) and a frequency-channel-based CNN (FCCNN) for rapid and accurate depression detection. The FCCNN processes data in the frequency domain, revealing patterns less visible in spatial representations. This method resembles term-document weighting (TDW) in natural language processing [29], which emphasizes term frequencies to identify key features. Their results showed that FCCNN delivered high accuracy on a public EEG dataset for major depressive disorder. Such innovations reflect continued advancement and offer direction for subsequent studies.

Functional magnetic resonance imaging (fMRI) offers detailed views of brain function and has become central to diagnosing attention-deficit/hyperactivity disorder (ADHD), a common childhood neurodevelopmental condition. Studies have combined fMRI with other neural signals, including electroencephalography (EEG), to

examine diverse cognitive functions such as perception, memory, decision-making, and emotion processing [30]. Similarly, integrating clinical data with perfusion-CT imaging shows substantial potential, as reported by Strambo and colleagues [31]. Combining fMRI results with our cognitive task scores from DCCS and PSM, along with demographic details, could greatly advance AD/CI detection and understanding. Pursuing this integration remains a priority to deepen insight into cognitive variability across populations.

Conclusion

The long-term goal is to build the MyCog platform capable of independently delivering reliable and rapid AD/CI predictions for patients. Ongoing hyperparameter refinement using new data is critical to this objective. In related work, Ke and colleagues [30] proposed a dual-CNN architecture to support brain e-health platforms. Their system employs automated machine learning to create a dual-CNN that balances accuracy and speed, while enabling a deep neural network (DNN) model to progressively improve through hyperparameter tuning and adaptation to fresh data. These findings illustrate the power of machine learning and optimization strategies in creating adaptive systems, aligning closely with the vision for MyCog to provide precise and efficient AD/CI predictions.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Centers for Disease Control and Prevention (CDC) . Comprehensive School Physical Activity Programs: a Guide (2013). Available online at: https://www.cdc.gov/healthyschools/professional_development/e-learning/CSPAP/_assets/FullCourseContent-CSPAP.pdf
2. Petersen RC, Lopez O, Armstrong MJ, Getchius T, Ganguli M, Gloss D, et al. Practice guideline update summary: mild cognitive impairment: report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. Neurology. (2018) 90:126–35. 10.1212/WNL.0000000000004826
3. Ward A, Arrighi HM, Michels S, Cedarbaum JM. Mild cognitive impairment: disparity of incidence and prevalence estimates. Alzheimers Dement. (2012) 8:14–21. 10.1016/j.jalz.2011.01.002
4. Rajan KB, Weuve J, Barnes LL, Wilson RS, Evans DA. Prevalence incidence of clinically diagnosed Alzheimer's disease dementia from 1994 to 2012 in a population study. Alzheimers Dement. (2019) 15:1–7. 10.1016/j.jalz.2018.07.216
5. Alzheimer's Association. 2018 Alzheimer's disease facts and figures. Alzheimers Dement. (2018) 14:367–429. 10.1016/j.jalz.2018.02.001
6. Dikmen S, Bauer P, Weintraub S, Mungas D, Slotkin J, Beaumont J, et al. Measuring episodic memory across the lifespan: NIH toolbox picture sequence memory test. J. Int. Neuropsychol. Soc. (2014) 20:611–9. 10.1017/S1355617714000460
7. Curtis L, Batio S, Benavente J, Shono Y, Nowinski C, Lovett RM, et al. Pilot testing of the MyCog assessment: rapid detection of cognitive impairment in everyday clinical settings. Gerontol Geriatr Med. (2023) 9:23337214231179895. 10.1177/23337214231179895
8. Lovett R, Bonham M, Benaventee JY, Hosseinian Z, Byrne GJ, Diaz MV, et al. Primary care detection of cognitive impairment leveraging health and consumer technologies in underserved US communities: protocol for a pragmatic randomized controlled trial of the MyCog paradigm. BMJ Open. (2023) 13:e080101. 10.1136/bmjopen-2023-080101
9. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning With Applications in R. New York: Springer. (2013).
10. Yao L. BMIRT, Bayesian Multivariate Item Response Theory. (2003).
11. Yao L. Reporting valid and reliable overall scores and domain scores. J Educ Measurem. (2010) 47:117. 10.1111/j.1745-3984.2010.00117.x
12. Yao L, Schwarz RD. A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests, Appl Psychol Measurem. (2006) 30:469–92. 10.1177/0146621605284537
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine

learning in Python. J Mach Learn Res. (2011) 12:2825–30.
14. 14.Cortes C, Vapnik V. Support-vector networks. Mach Learn. (1995) 20:273–97. 10.1007/BF00994018
15. Cox DR. The regression analysis of binary sequences. J R Stat Soc Series B. (1958) 20:215–42. 10.1111/j.2517-6161.1958.tb00292.x
16. Breiman L. Random forests. Mach Learn. (2001) 45:5–32. 10.1023/A:1010933404324
17. Breiman L. Bagging Predictors. Mach Learn. (1996) 24:123–40. 10.1007/BF00058655
18. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann. Statist. (2001) 29:1189–232. 10.1214/aos/1013203450
19. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bullet Mathem Biophys. (1943) 5:115–33. 10.1007/BF02478259
20. Zou J, Han Y, So SS. Overview of artificial neural networks. In:Living-stone DJ, editor. Artificial Neural Networks. Totowa, NJ: Humana Press. (2008) p. 14–22.
21. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. (1986) 323:533–6. 10.1038/323533a0
22. Medsker LR, Jain LC. Recurrent neural networks. Design Applicat. (2001) 5:64–67.
23. Elman JL. Finding structure in time. Cognit. Sci. (1990) 14:179–211. 10.1207/s15516709cog1402_1
24. Ke H, Wang F, Ma H, He Z. ADHD identification and its interpretation of functional connectivity using deep self-attention factorization. Knowl-Based Syst. (2022) 250:109082. 10.1016/j.knosys.2022.109082
25. Tamnes CK, Fjell AM, Westlye LT, Østby Y, Walhovd KB. Becoming consistent: Developmental reductions in intraindividual variability in reaction time are related to white matter integrity. J Neurosci. (2012) 32:972–82. 10.1523/JNEUROSCI.4779-11.2012
26. Dykiert D, Der G, Starr JM, Deary IJ. Age differences in intra-individual variability in simple and choice reaction time: Systematic review and meta-analysis. PLoS ONE. (2012) 7:e45759. 10.1371/journal.pone.0045759
27. Jiao H, He Q, Yao L. Machine learning and deep learning in assessment. Psychol Test Assessm. (2023) 65:179–90.
28. Ke H, Cai C, Wang F, Hu F, Tang J, Shi Y. Interpretation of Frequency Channel-Based CNN on Depression Identification. Front Comput Neurosci. (2021) 15:773147. 10.3389/fncom.2021.773147
29. Yao L., Jiao H. Comparing performance of feature extraction methods and machine learning models in essay scoring. Chin/Engl J Educ Measurem Evaluat. (2023) 4:1. 10.59863/DQIZ8440
30. Ke H, Chen D, Shi B, Zhang J, Liu X, Zhang X, Li X. Improving Brain E-Health Services via High-Performance EEG Classification With Grouping Bayesian Optimization. IEEE Trans Serv Comp. (2020) 13:696–708. 10.1109/TSC.2019.2962673
31. Strambo D, Rey V, Rossetti AO, Maeder P, Dunet V, Browaeys P, et al. Perfusion-CT imaging in epileptic seizures. J Neurol. (2018) 265:2972–79. 10.1007/s00415-018-9095-1