# A Machine Learning Classification of Individuals with Mild Cognitive Impairment into Variants from Writing

Hana Kim, PhD[1], and Charalambos Themistocleous, PhD[2*]

[1]Department of Communication Sciences and Disorders, University of South Florida, Tampa, FL 33620, United States

[2]Department of Special Needs Education, University of Oslo, Oslo 0266, Norway

[*] Corresponding Author:

Charalambos Themistocleous, Ph.D.

Department of Special Needs Education

University of Oslo, Oslo 0266, Norway

cthemistocleous@icloud.com

**Abstract**

**Introduction**: Individuals with Mild Cognitive Impairment (MCI), a transitional stage between cognitively healthy aging and dementia, are characterized by subtle neurocognitive changes. Clinically, they can be grouped into two main variants, namely into patients with amnestic MCI (aMCI) and non-amnestic MCI (naMCI). The distinction of the two variants is known to be clinically significant as they exhibit different progression rates to dementia. However, it has been particularly challenging to classify the two variants robustly. Recent research indicates that linguistic changes may manifest as one of the early indicators of pathology. Therefore, we focused on MCI's discourse-level writing samples in this study. We hypothesized that a written picture description task can provide information that can be used as an ecological, cost-effective classification system between the two variants.

**Methods**: We included one hundred sixty-nine individuals diagnosed with either aMCI or naMCI who received neurophysiological evaluations in addition to a short-written picture description task. Natural Language Processing (NLP) and BERT pre-trained Language Models were utilized to analyze the writing samples.

**Results**: We showed that the written picture description task provided 90% overall classification accuracy for the best classification models, which performs better than cognitive measures.

**Discussion**: Written discourses analyzed the AI models can automatically assess individuals with aMCI and naMCI and facilitate diagnosis, prognosis, therapy planning, and evaluation.

## 1 Background

With the growth in the number of older adults, age-related neurodegenerative diseases such as Alzheimer's disease (AD) have dramatically increased. These neurodegenerative diseases cause a great deal of financial and emotional burdens not only for patients and their caregivers but also for society. The global cost of dementia care was estimated to exceed $500 billion in the United States [1]. It is expected to rise to $2 trillion by 2030 [2]. Research has suggested that the preclinical phase of dementia may start earlier than the diagnosis. Detecting the preclinical stage of dementia and providing an intervention will delay the onset of AD. This will significantly minimize the socio-economic burden, which is expected to reduce societal costs by 40% [3].

Mild cognitive impairment (MCI) is an intermediate stage between cognitively healthy aging and dementia [4]. It represents a critical preclinical stage of the AD [5-7]. MCI includes four different clinical subtypes. Two main subtypes are amnestic MCI (aMCI) and non-amnestic MCI (naMCI); this subtyping is determined based on the impairment in memory. Individuals with aMCI are characterized by memory loss, while individuals with naMCI demonstrate impairment in domains such as executive functions, attention, and language [8, 9]. Also, depending on the number of cognitive domains impaired, individuals can be categorized into single-domain and multi-domain MCI. Although a higher risk of developing dementia characterizes individuals with MCI, not all individuals with MCI will progress to dementia; some may remain stable, and others even regress to a condition of healthy aging [10-12]. Therefore, it is essential to discriminate against those who are more likely to progress to dementia for early intervention since most treatment strategies are more effective in the presymptomatic stage of dementia [13].

Depending on the two main subtypes of MCI, differences in the progression from MCI to dementia have been reported. In general, it has been suggested that aMCI represents the earliest symptomatic manifestation of AD pathophysiology, while naMCI is likely to progress to non-Alzheimer's dementia [14-16]. A recent 20-year retrospective study supports this and adds more information with a large dataset (N = 1188). The authors demonstrated that aMCI represents a greater risk for progressing to dementia (not only for AD) compared to naMCI. The odd ratio of the progression to dementia between aMCI and naMCI was statistically different [17]. This highlights the clinical need of a robust, reliable system for classifying aMCI and naMCI [18].

There have been several approaches for MCI diagnosis. Behaviorally, a brief cognitive screening test can assist in identifying whether an individual has an apparent cognitive impairment [9]. Neuropsychological tests can be administered depending on the need for further assessments to determine the presence or degree of impairment in cognitive functions. The tests for MCI biomarkers require magnetic resonance imaging (MRI) or lumbar puncture for cerebrospinal fluid (CSF). Increased amyloid burden was found to be specific to aMCI, while naMCI does not exhibit a specific abnormality in neuroimaging (see review for Yeung, Chau [19]). Blood biomarkers, considered a comparatively more straightforward means of testing, have also been investigated [20]. Unfortunately, such tests for MCI biomarkers are not routine care in clinical settings [21-24]. Moreover, the cost and availability of the testing technique (e.g., MRI) may limit its impact on individuals' care [25].

Linguistic changes are considered to manifest as one of the earlier indicators of pathology in cognitive impairment. It has been reported that they emerge years before deficits in other cognitive systems become apparent [26]. In particular, writing is a cognitively and linguistically complicated activity. Writing consists of distinct phases: planning, generating, and revising [27]. Writers initially set a goal for organizing their knowledge and executing the plan in response to the topic of the writing activity. Then, writers revisit and revise their output. All phases should be well orchestrated to accomplish successful writing within cognitive systems such as executive functions, attention, and working memory. A recent review article highlighted the diagnostic value of writing tests, especially at the discourse level (Kim et al., 2023). Discourse is any language beyond the sentence level [28, 29]. Kim and colleagues (2022) investigated the prognostic value of discourse-level writing tests. They conducted a chart review of individuals diagnosed with MCI and visited a neurology outpatient clinic more than once (N = 71). They classified the study participants into the stable MCI group and the converter group. The authors examined whether a written discourse task using the Cookie Theft picture [30] predicts clinical course in the MCI group. They found that the stable MCI group produced more core words than the converter group at their baseline assessment. This underscores the potential clinical utility of discourse-level writing tasks for early detection of those who are likely to progress to dementia from MCI.

In recent years, computational methods such as Natural Language Processing (NLP) have been used to analyze written language samples in individuals with neurodegenerative disease [31-

36]. Computational methods offer two advantages. First, they allow the elicitation and combination of measures from different linguistic domains. A decisive property of ML models is their ability to find patterns between features associated with a specific group of individuals, i.e., patients with aMCI and naMCI.

Earlier studies successfully distinguished healthy adults from individuals with MCI from healthy adults [32], MCI from dementia [37-40], and the subtypes of primary progressive aphasia [41, 42]. These findings highlight the role of ML as an important method that can contribute to the existing approaches [35] and to inform clinical assessment and therapy.

This is the first attempt to classify two subtypes of MCI (aMCI vs naMCI) using discourse-level writing samples in NLP. Since writing involves several cognitive functions (especially language, vision, and motor control), we hypothesized that a written picture description task could distinguish individuals with aMCI and naMCI. This work could potentially provide a quick and easy tool to facilitate diagnosis from written language tasks.

## 2 Methods

### 2.1 Participants

Our participants were comprised of 169 individuals diagnosed with either aMCI or naMCI. All individuals were recruited through the Johns Hopkins Hospital and were diagnosed by an experienced neurologist. The diagnosis was based on history, neuroimaging, neurological examination, and neuropsychological testing, and all individuals met the current criteria for MCI. The exclusion criteria for the study included individuals 1) who were younger than 18 years old, 2) who had a lack of English competence, 3) who had significant psychiatric illness and alcohol and drug use, 4) who had significant neurological problems affecting the brain (e.g., stroke, multiple sclerosis, and Parkinson's disease), and 5) who had uncorrected visual or hearing loss. All individuals with MCI fulfilled the recent criteria of the 2018 National Institute on Aging-Alzheimer's Association (NIA-AA) research framework [43].

Demographic information for individuals with MCI can be found in Table 1.

Table 1 Participants' Age and Education across variants (Amnestic and Non-Amnestic) and gender.

| Variant | Gender | N | Mean | SD | Median | Mode |
|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Age | Amnestic | F | 71 | 67.4 | 12.99 | 70 | 53 |
| | | M | 53 | 69.7 | 15.28 | 74 | 69 |
| | Non Amnestic | F | 21 | 54.2 | 13.48 | 52 | 48 |
| | | M | 25 | 65.6 | 12.04 | 66 | 65 |
| Education | Amnestic | F | 70 | 16.1 | 3.19 | 16 | 16 |
| | | M | 52 | 17.5 | 3.42 | 18 | 16 |
| | Non Amnestic | F | 21 | 15.5 | 3.53 | 16 | 16 |
| | | M | 24 | 16 | 3.06 | 16.5 | 12 |

Specifically, participants underwent a battery of standardized neuropsychological tests to assess their cognitive and linguistic abilities. These tests comprehensively evaluated various aspects of language and cognitive functioning, offering a detailed assessment of their cognitive strengths and weaknesses. The neurocognitive tests include the Mini-Mental State Examination (MMSE, Folstein, Folstein [44]), the Orientation and Information subset from Wechsler Memory Scale-Third Edition (WMS-III; Wechsler [45]), Digit span subtests of the WMS-III [45], Rey Auditory Verbal Learning Test (RAVLT; Rey, 1941), Rey Complex Figure (RCF; Rey [46]), Boston Naming Test [30], Verbal Fluency Task (FAS), the Free narrative writing section from BDAE [30], Trail Making tests (TMT; Reitan and Wolfson [47]), and Stroop test [48]. The tests were carefully selected to provide a sensitive measure of abnormalities compared to individuals with normal cognitive functioning. Table 2 includes neurocognitive test results for all individuals with MCI. The study protocol underwent rigorous review and received approval from the Johns Hopkins Institutional Review Board (IRB00266221). The data were accessed in August 1st 2023 for the purposes of this study and the authors had no information to identify the participants.

Table 2. Performance in Neurocognitive Testing in Individuals with MCI.

| | **Variant** | **Mean** | **Median** | **Mode** | **SD** |
|---|---|---|---|---|---|
| MMSE | Amnestic | 27.5081 | 28 | 28 | 1.746 |
| | Non Amnestic | 28.0476 | 29 | 29 | 1.821 |
| WMS | Amnestic | 13.25 | 14 | 14 | 0.942 |
| | Non Amnestic | 13.6804 | 14 | 14 | 0.592 |
| Digit Forward | Amnestic | 6.7016 | 7 | 7 | 1.169 |

| | | | | | |
|---|---|---|---|---|---|
| | Non Amnestic | 6.7391 | 7 | 6 | 1.437 |
| Digit Backward | Amnestic | 4.2984 | 4 | 4 | 1.044 |
| | Non Amnestic | 4.4565 | 4 | 4 | 1.187 |
| RAVLT (total) | Amnestic | 29.2177 | 29 | 30 | 9.373 |
| | Non Amnestic | 37.8587 | 37 | 37 | 11.187 |
| RAVLT (Delayed) | Amnestic | 3.5081 | 3 | 3 | 2.95 |
| | Non Amnestic | 6.8333 | 7 | 7 | 3.151 |
| RCF (Immediate) | Amnestic | 7.8487 | 7 | 0 | 5.934 |
| | Non Amnestic | 14.5435 | 12 | 6 | 8.989 |
| RCF (delayed) | Amnestic | 6.2391 | 5 | 0 | 5.25 |
| | Non Amnestic | 13.1739 | 12.25 | 0 | 8.568 |
| BNT | Amnestic | 49.2033 | 52 | 56 | 10.265 |
| | Non Amnestic | 52.2826 | 54 | 56 | 7.12 |
| Verbal Fluency (FAS) | Amnestic | 35.5772 | 35 | 32 | 13.073 |
| | Non Amnestic | 34.3261 | 32.5 | 23 | 12.994 |
| BDAE writing | Amnestic | 4.1441 | 4 | 4 | 3.733 |
| | Non Amnestic | 3.7778 | 4 | 4 | 0.56 |
| TMT A | Amnestic | 55.2218 | 48.5 | 30 | 31.634 |
| | Non Amnestic | 45.5993 | 36.5 | 25 | 24.149 |
| TMT A error | Amnestic | 0.042 | 0 | 0 | 0.302 |
| | Non Amnestic | 0.087 | 0 | 0 | 0.354 |
| TMT B | Amnestic | 132.8319 | 113 | 110 | 99.71 |
| | Non Amnestic | 121.9254 | 96 | 57 | 75.288 |
| TMT B error | Amnestic | 0.5439 | 0 | 0 | 1.863 |
| | Non Amnestic | 0.3696 | 0 | 0 | 0.878 |
| Color | Amnestic | 111.7168 | 112 | 112 | 2.647 |
| | Non Amnestic | 110.55 | 112 | 112 | 8.852 |
| Color (Word) | Amnestic | 67.2 | 66 | 112 | 29.593 |
| | Non Amnestic | 68.8158 | 64.5 | 112 | 25.877 |

MMSE = Mini-Mental State Examination; WMS = Wechsler Memory Scale; RAVLT (total) = total score of the Rey Auditory Verbal Learning Test ; RAVLT (delayed) = score for the delayed

recall of the Rey Auditory Verbal Learning Test; (delayed recall); RCF (immediate) = score for the immediate recall of the Rey Complex Figure; RCF (delayed) = score for the delayed recall of the Rey Complex Figure; BNT = Boston Naming Test; BDAE writing = free narrative writing from the Boston Diagnostic Aphasia Examination; TMT A = Trail Making Test Part A; TMT A error = Errors made in Trail Making Test Part A; TMT B = Trail Making Test Part B; TMT B error = Errors made in Trail Making Test Part B; Color = Color Stroop test; Color-Word = Stroop Color and Word Test; SD: Single-domain MCI; MD: Multiple-domain MCI.

## 2.2 Written Picture Description Task

Writing samples were collected using the Cookie Theft picture from the Boston Diagnostic Aphasia Examination-3 (BDAE-3; Goodglass, Kaplan [30]. Participants were seated with the picture stimulus and a piece of paper. The clinicians used the prompt to encourage the participants to provide a written description, "Write as much as you can about what you see going on in this picture." Once the participants completed the task, their writing samples were transcribed into a text document by experienced researchers.

## 2.3 Machine Learning Process

The analysis involved the preprocessing of the data, the extraction of significant features from the written picture description task, and the study of those measures.

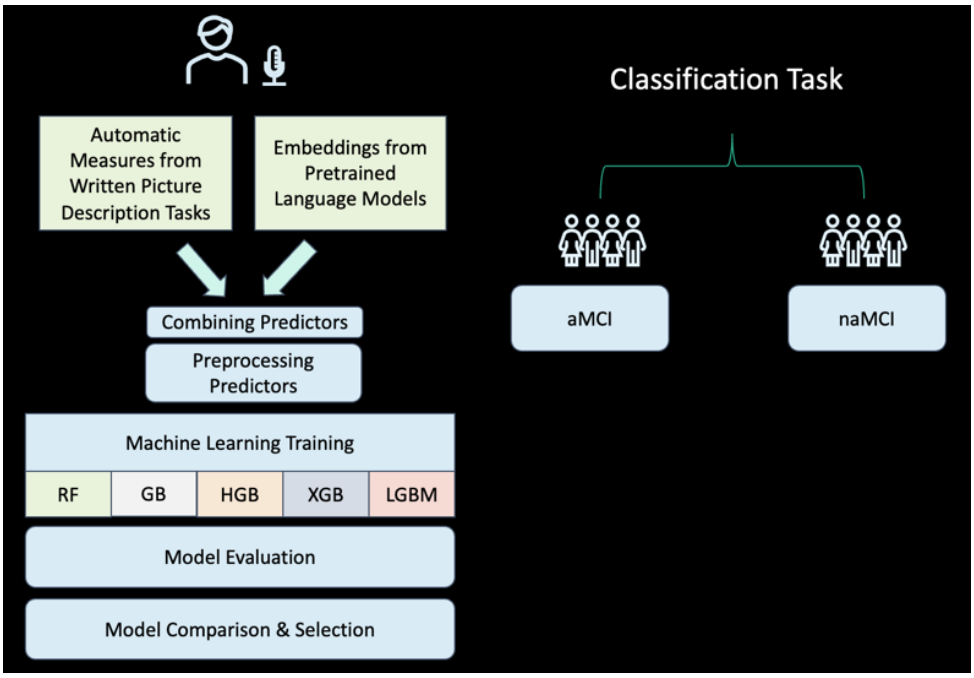


Figure 1. The machine learning classification process and classification task.

**Analysis of narrative speech**

We analyzed the written transcripts from the text documents using two natural language processing tools, including the tokenization of the text, the tagging of morphological categories, and the parsing of the syntactic constituents. Specifically, each word in the text was labelled using Open Brain AI's POS tagger and syntactic dependency parser, which uses a variety of linguistic information to determine the dependency structure of a sentence [49]. Open Brain AI provided automatic measures that included counts and the ratio of each word / total count of words that appears in the text for each participant.

Specifically, the automatically elicited morphosyntactic measures shown in Table 3 include Part of Speech (POS) categories (i.e., *adjective, adposition, adverb, auxiliary verb, coordinating conjunction, determiner, interjection, noun, numeral, particle, pronoun, proper noun, subordinating conjunction, symbol, verb*), the number of words and characters and their character/word ratio, and syntactic dependency measures indicating the grammatical relationships between words in a sentence and their count to total word ratio.

Table 3. Means and Standard Deviations of features in individuals with Non-Amnestic and Amnestic MCI.

| | Non-Amnestic | | Amnestic | |
|---|---|---|---|---|
| | Mean | SD | Mean | SD |
| Adjectival Clause | 0.021 | 0.054 | 0.014 | 0.031 |
| Adjectival Complement | 0.007 | 0.016 | 0.009 | 0.017 |
| Adjective | 0.022 | 0.032 | 0.028 | 0.033 |
| Adposition | 0.097 | 0.054 | 0.113 | 0.058 |
| Adverb | 0.013 | 0.022 | 0.015 | 0.025 |
| Adverbial Clause | 0.022 | 0.030 | 0.021 | 0.031 |
| Adverbial Modifier | 0.012 | 0.021 | 0.014 | 0.024 |
| Agent | 0.000 | 0.004 | 0.000 | 0.002 |
| Adjectival Modifier | 0.021 | 0.031 | 0.017 | 0.033 |
| Apposition | 0.004 | 0.016 | 0.003 | 0.013 |
| Attribute | 0.003 | 0.007 | 0.002 | 0.007 |

| | | | | |
|---|---|---|---|---|
| Auxiliary | 0.080 | 0.060 | 0.075 | 0.065 |
| Auxiliary (Passive) | 0.001 | 0.005 | 0.002 | 0.007 |
| Case Marking | 0.002 | 0.008 | 0.002 | 0.007 |
| Coordinating Conjunction | 0.019 | 0.026 | 0.018 | 0.026 |
| Clausal Complement | 0.020 | 0.033 | 0.021 | 0.035 |
| Coordinating Conjunction | 0.019 | 0.026 | 0.018 | 0.026 |
| Character-Word Ratio | 5.244 | 0.410 | 5.256 | 0.517 |
| Compound | 0.032 | 0.044 | 0.034 | 0.057 |
| Conjunction | 0.020 | 0.028 | 0.020 | 0.028 |
| Dative Case | 0.002 | 0.008 | 0.004 | 0.013 |
| Dependent | 0.046 | 0.095 | 0.032 | 0.056 |
| Determiner | 0.125 | 0.086 | 0.110 | 0.088 |
| Direct Object | 0.086 | 0.045 | 0.084 | 0.067 |
| Expletive | 0.003 | 0.007 | 0.001 | 0.006 |
| Interjection | 0.001 | 0.008 | 0.001 | 0.006 |
| Marker | 0.018 | 0.028 | 0.007 | 0.016 |
| Meta Data | 0.000 | 0.004 | 0.010 | 0.056 |
| Negation Modifier | 0.005 | 0.012 | 0.004 | 0.012 |
| Noun | 0.362 | 0.109 | 0.376 | 0.119 |
| Nominal Subject | 0.137 | 0.055 | 0.142 | 0.057 |
| Nominal Subject (Passive) | 0.001 | 0.005 | 0.002 | 0.007 |
| Numeral | 0.004 | 0.012 | 0.004 | 0.013 |
| Numeric Modifier | 0.003 | 0.011 | 0.004 | 0.012 |
| Object Predicate | 0 | 0 | 0.001 | 0.009 |
| Parataxis | 0 | 0 | 0.000 | 0.002 |
| Particle | 0.026 | 0.026 | 0.025 | 0.029 |
| Prepositional Complement | 0.000 | 0.002 | 0.002 | 0.008 |
| Prepositional Object | 0.078 | 0.052 | 0.096 | 0.052 |
| Possessive Modifier | 0.011 | 0.021 | 0.011 | 0.021 |
| Preposition | 0.083 | 0.057 | 0.099 | 0.058 |
| Pronoun | 0.037 | 0.041 | 0.027 | 0.033 |

| | | | | |
|---|---|---|---|---|
| Proper Noun | 0.003 | 0.014 | 0.003 | 0.012 |
| Particle | 0.012 | 0.017 | 0.011 | 0.018 |
| Punctuation | 0.111 | 0.073 | 0.104 | 0.081 |
| Relative Clause | 0.004 | 0.010 | 0.004 | 0.010 |
| Root | 0.105 | 0.056 | 0.099 | 0.061 |
| Subordinating Conjunction | 0.018 | 0.028 | 0.008 | 0.017 |
| Symbol | 0.000 | 0.003 | 0.001 | 0.007 |
| Verb | 0.196 | 0.071 | 0.192 | 0.057 |
| Other | 0.002 | 0.013 | 0.010 | 0.051 |
| Open Clausal Complement | 0.015 | 0.021 | 0.018 | 0.024 |
| Words [count][1] | 31.943 | 15.318 | 29.650 | 14.205 |
| Characters [counts][1] | 170.927 | 78.647 | 158.829 | 70.947 |

Note: All measures indicate the count / total word; features marked with the index ([1]) are counts.

**Semantic measures from BERT**

Semantic measures are crucial in individuals with aMCI and can differentiate individuals aMCI and naMCI. To depict semantic relationships, we included word and sentence embeddings from BERT-large-uncased, a BERT (Bidirectional Encoder Representations from Transformers) pretrained language model [50]. Specifically, the BERT-large-uncased is a deep neural network trained on a large dataset of text corpora and can be used for various natural language processing (NLP) tasks, such as question answering, text summarization, and sentiment analysis. The BERT-large-uncased has been shown to achieve state-of-the-art performance on various NLP tasks. It consists of 12 encoder layers, each containing a self-attention mechanism and a feed-forward network. The self-attention mechanism allows the model to learn long-range dependencies between words in a sentence, while the feed-forward network adds non-linearity.

### 2.3.1 Addressing Imbalance and Cross-validation

We employed Random Over-Sampling (ROS) to balance the class distribution and address the limitations of the relatively small dataset [51]. This technique alleviates the models' tendency to favor the majority class, a common challenge in imbalanced datasets. Additionally, we implemented group 5-fold cross-validation. This approach minimized data leakage and provided

a more reliable model performance evaluation. Furthermore, we standardized the non-BERT features to ensure uniformity in scale.

### 2.3.2 Model Evaluation and Selection

We selected ML models that do not require massive amounts of training data. To choose the best model for our data, we have trained ML models that roughly belong to four main categories of models, namely ensemble learning models (Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), and LightGBM (LGBM)). RF is a ML method combining several decision trees to enhance prediction accuracy. This approach can manage high-dimensional data and is resilient to overfitting. GB sequentially combines weak ML learners, each correcting the predecessor's errors. GB is used in classification and regression tasks for large, complex datasets. XGB and LGBM implement gradient boosting with speed and accuracy. They are employed in scenarios requiring rapid processing of large datasets. Hist Gradient Boosting (HGB), a gradient boosting variant, uses histograms for feature representation, enhancing efficiency with large-scale, high-dimensional data structures. Each ML algorithm has unique strengths, making these models suitable for specific data types and prediction tasks. Only comparing and selecting ML models provides versatility, adaptability, and improved performance in the ML process, enabling the model to tackle the various underlying characteristics of the data.

### 2.3.3 Hyperparameter Tuning and Model Comparison

A grid search with cross-validation was employed to evaluate and compare the performance of the different machine-learning models. The hyperparameter tuning involved finding the optimal hyperparameters for each model using grid search and calculating the evaluation metrics.

Grid search is a method for hyperparameter tuning evaluates different combinations of predefined hyperparameter values to determine the combination that produces the best performance for a given model. In this case, a grid search was performed for each of the machine learning models included in the study.

We evaluated each model using five-fold cross-validation, which involves evaluating the performance of a model by splitting the data into multiple folds. Each fold is used as a validation set, while the remaining folds are used as the training set. The model is trained on the training set

and evaluated on the validation set. This process is repeated for each fold, and the average performance across all folds is used as the final performance estimate.

Various evaluation metrics were used to assess the performance of the different machine learning models. These metrics included accuracy, F1 score, precision, recall, ROC AUC, and Cohen's kappa score: i. Accuracy is the proportion of correct predictions; ii. F1 score measures a model's ability to correctly classify positive and negative cases; iii. Precision is the proportion of positive predictions that are positive; iv. Recall is the proportion of positive cases correctly classified as positive, and v. ROC AUC (Receiver Operating Characteristic Area Under the Curve) measures a model's ability to distinguish between positive and negative cases.

## 3 Results

Written picture description tasks were processed using combined NLP analysis and Bidirectional Encoder Representations from Transformers (BERT) models to elicit measures representing the embeddings. We have implemented two machine learning-supervised classification tasks.

A classification model was designed to distinguish individuals with aMCI and naMCI. The model included only information from the Cookie-Theft picture description task. The model distinguished individuals with aMCI and naMCI. These results suggest that the written discourse from a picture description task provides sufficient information to identify the individuals with the two variants of MCI.

In the ML models, the ROC curves were nearly 98% for classifying individuals with aMCI and naMCI (Figure 2). This suggests that written discourse productions, as manifested in a picture description task, can distinguish the two groups of individuals from language measures. Regarding accuracy, the ensemble models with boosting had the best performance (Table 3). Gradient Boosting, Hist Gradient Boosting, XGBoost, and LightGBM. The consistency in the output of those models further demonstrates their effectiveness for real-world applications.

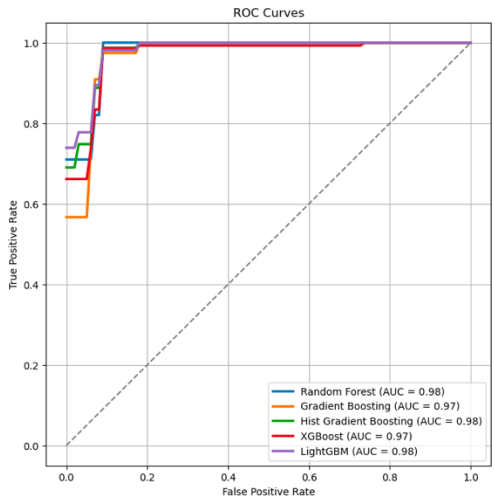



Figure 2. MLs performance on the classification task: individuals with aMCI vs. individuals with naMCI from language measures.

Table. 4. Model performance in the classification task: individuals with aMCI vs. individuals with naMCI from language measures.

| | RF | GB | HGB | XGB | LGBM |
|---|---|---|---|---|---|
| Accuracy | 0.90 | 0.90 | 0.89 | 0.89 | 0.89 |
| F1 | 0.71 | 0.72 | 0.70 | 0.71 | 0.70 |
| Precision | 0.74 | 0.75 | 0.75 | 0.75 | 0.75 |
| Recall | 0.68 | 0.70 | 0.67 | 0.68 | 0.66 |
| ROC/AUC | 0.98 | 0.97 | 0.98 | 0.97 | 0.98 |

*Note.* RF: Random Forests, GB: Gradient Boosting, HGB: Hist Gradient Boosting; XGB: XGBoost, LGBM: LightGBM.

As indicated by the outcomes, the utilization of machine learning models shows the potential of MLs in diagnosing and differentiating the two MCI subtypes. The reported standardized metrics – Accuracy, F1 Score, Precision, Recall, and ROC/AUC – indicate the effectiveness of these models, with one (1) being the best value.

- Accuracy (0.90 for most models) reflects the ML model's overall correctness in classifying the MCI type.
- F1 Score balances precision and recall, with values around 0.70-0.72, indicating a good balance between false positives and false negatives.

- Precision (0.74-0.75) measures the proportion of correctly identified positive cases among all positive calls made by the model.
- Recall (ranging from 0.66 to 0.70) indicates the model's ability to identify all actual positive cases.
- ROC/AUC (between 0.97 and 0.98) reflects the model's ability to distinguish between the two classes across various thresholds, with values close to 1 indicating excellent performance.

We have evaluated feature importance. BERT features dominate the rankings of the 15 contributing factors for the RF classification. The following features contribute to RF classification, from more important to less important: prepositional object, adposition, dependent, particle, auxiliary, root (verb), adjective, and subordinating conjunction.

These results suggest a reliable performance in distinguishing patients with naMCI vs. aMCI highlight the potential of advanced machine-learning techniques in medical diagnostics, especially for complex conditions like MCI. The high performance of these models suggests that they could be valuable tools in clinical practice for early and accurate identification of MCI types, thereby enabling more tailored and effective treatment strategies.

## 4 Discussion

MCI is an early stage of cognitive decline due to pathology reasons [4]. Individuals with aMCI are characterized primarily by memory deficits, while individuals with naMCI are impaired in other cognitive functions, such as language, attention, and executive functions. Identifying the type of MCI is important for predicting the progression of the condition, as individuals with aMCI are more prone to progress into Alzheimer's disease [52, 53] or all types of dementia (Glynn et al., 2021). This study aimed to determine the potential diagnostic utility of computational methods in classifying two subtypes of MCI. We found that a written picture description task can distinguish individuals with aMCI and naMCI at approximately 90% accuracy. This finding confirms that written discourse analysis, which is infrequently done in clinical settings, provides clinically essential information (Kim et al., 2023) and can be a powerful approach for better characterizing the subtypes of MCI.

Importantly, our study shows that a single behavioral task (i.e., a picture description task) can provide substantial information about domains that require multiple separate tasks. As mentioned earlier, either multiple pen-and-pencil tasks or neuroimaging techniques need to be

conducted clinically to classify MCI. Previous studies using machine learning algorithms and neuroimaging data demonstrated an accurate classification of MCI subtypes [54, 55]. However, data can be obtained only with advanced techniques. They are not often feasible for individual patients [56]. Behaviorally, multiple tasks that evaluate different cognitive components, such as memory and executive functions, need to be administered, which is considered a time-intensive process. From a clinical perspective, computational assessment of language with machine learning and natural language processing opens the door for exciting opportunities to expand the analysis to both longer and more complex tests productions.

Besides the cost-effective assessment, it is also significant to note that the current study used written discourse samples, which received little attention in research (Kim et al., 2023) and are not often collected and evaluated in clinical settings [57]. He, Chapin [58] used a spoken discourse task to investigate the classification among healthy adults, subtypes of MCI, and dementia. In the study, the researchers used both linguistic and acoustic features, but the classification accuracy (aMCI vs naMCI) was 88%. Our findings shed light on the clinical value of written discourse as the linguistic features in writing lead to higher classification accuracy. This also indicates that linguistic features in writing can be potential markers of memory deficits and may provide enough information for the classification.

Written discourse offers a plethora of information about individuals' linguistic functioning, including textual macrostructure and microstructure. However, it is not clear which components of written discourse in this population are more influenced by cognitive impairment in MCI. This is evidenced by 102 different measures used to quantify writing behaviors in research with little repetition of the same measure (Kim et al., 2023). In the current study, using written discourse samples, we calculated the POS of each word and syntactic relationships [59] that appears in the written picture description task [60]. Together, this can be an optimal approach for analyzing such language samples in that it adds to the efficiency of written picture description analysis. This also provides a comprehensive and detailed grammar analysis in a standardized and less subjective manner.

Moreover, we found that the BERT semantic features dominated the hierarchy of analytical constructs that we have used. This finding is consistent with the consensus that impairments in semantic domains of language are a key manifestation of disease progress in neurodegenerative disorders [61-63]. These features can be seen in the literature to be associated

with one or more elements of the writing skills of individuals with MCI, as they interface linguistic and semantic memory domains.

Specifically, context-sensitive embeddings from BERT [50] played a critical role in the high accuracy of the classification. These result from averaging the token-level embeddings from the last layer of a BERT model for each input text, which creates a single, comprehensive vector representation for the entire text, capturing its overall contextual meaning. Traditional word embedding techniques, such as Word2Vec [64] and GloVe [65], generate a single word embedding for each word in the vocabulary. The embeddings are decontextualized, which fails to capture the meanings of polysemous words. For instance, the word *bank* can mean a financial institution that accepts deposits and makes loans or the sloping edge of a river or other body of water. On the other hand, BERT uses a technique known as contextual embedding. This means that the representation of a word is based on sentence context. So, the word *bank* would have different representations in the sentences "*I went to the bank to retrieve money"* and "*the little house next to the river bank,"* which offers a better representation of ambiguous meanings, improving the accuracy of text classification. Again, these contextual embeddings utilized in this study demonstrate a better understanding of the syntactic and semantic relationships between words in a sentence. This is crucial for quantifying the overall thematic content of the written picture descriptions. Additionally, since individuals with amnestic and non-amnestic MCI differ in their semantic memory [66, 67], the contextual sensitivity of BERT's embeddings helps the model to adapt to differences in vocabulary and jargon.

Although it is well-known that picture description tasks are valuable for eliciting connected language samples in individuals with MCI [68], the Cookie-Theft picture offers a less ecological way of personal expression through writing. Such productions are constrained substantially in their context and effectively identifying differences in pragmatic language usage and speech and voice parameters. Also, the task does not allow the assessment of non-epistemic domains, such as deontic modality expressions of wish and hope and non-present tense verb-tense semantics, as it does not provide opportunities to discuss past or future events. Additionally, picture description tasks do not offer opportunities for expressing emotional and other affective content, which might be necessary for assessing the interface of language, emotion, and pragmatics. An open-ended essay writing could have offered the potential to assess more stylistic, linguistic, and communicative speech characteristics. Nevertheless, written picture

descriptions demonstrate the potential to detect speech and language characteristics in neurodegenerative diseases such as MCI and dementia, as suggested by a recent review (Kim et al., 2023). Considering the brief time to elicit writing samples, NLP combined with discourse-level writing samples will enable more efficient methods for analyzing these linguistic and communicative features, further enhancing the diagnostic accuracy and the clinical utility of written discourse analysis.

The results of the current study suggest that written discourse samples can offer a quick and efficient means of gaining valuable insights into linguistic abilities while minimizing the burden placed on individuals with MCI. Future research is necessary to verify this finding with a balanced sample size between aMCI and naMCI. For a better diagnostic tool, future studies, including MCI-dementia conversion, are needed to test the predictive value of the automatic classification of MCI.

## 5 References

[1] Alzheimer's A. 2019 Alzheimer's disease facts and figures. Alzheimer's & dementia. 2019;15:321-87.
[2] Tay LX, Ong SC, Tay LJ, Ng T, Parumasivam T. Economic Burden of Alzheimer's Disease: A Systematic Review. Value in Health Regional Issues. 2024;40:1-12.
[3] Zissimopoulos J, Crimmins E, St. Clair P. The value of delaying Alzheimer's disease onset. 1 ed: De Gruyter. p. 25-39.
[4] Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E. Mild cognitive impairment: clinical characterization and outcome. Archives of neurology. 1999;56:303-8.
[5] López-Sanz D, Bruña R, Garcés P, Martín-Buro MC, Walter S, Delgado ML, et al. Functional connectivity disruption in subjective cognitive decline and mild cognitive impairment: a common pattern of alterations. Frontiers in aging neuroscience. 2017;9:109.
[6] Villemagne VL, Chételat G. Neuroimaging biomarkers in Alzheimer's disease and other dementias. Ageing Research Reviews. 2016;30:4-16.
[7] Tabatabaei-Jafari H, Shaw ME, Cherbuin N. Cerebral atrophy in mild cognitive impairment: A systematic review with meta-analysis. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2015;1:487-504.
[8] Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, et al. Mild cognitive impairment. The Lancet. 2006;367:1262-70.
[9] Winblad B, Palmer K, Kivipelto M, Jelic V, Fratiglioni L, Wahlund LO, et al. Mild cognitive impairment – beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment. J Intern Med. 2004;256:240-6.
[10] Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia–meta-analysis of 41 robust inception cohort studies. Acta psychiatrica scandinavica. 2009;119:252-65.

[11] Ansart M, Epelbaum S, Bassignana G, Bône A, Bottani S, Cattai T, et al. Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. Medical Image Analysis. 2021;67:101848.
[12] Iraniparast M, Shi Y, Wu Y, Zeng L, Maxwell CJ, Kryscio RJ, et al. Cognitive reserve and mild cognitive impairment: predictors and rates of reversion to intact cognition vs progression to dementia. Neurology. 2022;98:e1114-e23.
[13] Giau VV, Bagyinszky E, An SSA. Potential fluid biomarkers for the diagnosis of mild cognitive impairment. International journal of molecular sciences. 2019;20:4149.
[14] Busse A, Angermeyer MC, Riedel-Heller SG. Progression of mild cognitive impairment to dementia: a challenge to current thinking. The British Journal of Psychiatry. 2006;189:399-404.
[15] Guo T, Korman D, Baker SL, Landau SM, Jagust WJ. Longitudinal Cognitive and Biomarker Measurements Support a Unidirectional Pathway in Alzheimer's Disease Pathophysiology. Biological Psychiatry. 2021;89:786-94.
[16] Roberts RO, Knopman DS, Mielke MM, Cha RH, Pankratz VS, Christianson TJH, et al. Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal. Neurology. 2014;82:317-25.
[17] Glynn K, O'Callaghan M, Hannigan O, Bruce I, Gibb M, Coen R, et al. Clinical utility of mild cognitive impairment subtypes and number of impaired cognitive domains at predicting progression to dementia: A 20-year retrospective study. International Journal of Geriatric Psychiatry. 2021;36:31-7.
[18] Pelka O, Friedrich CM, Nensa F, Mönninghoff C, Bloch L, Jöckel K-H, et al. Sociodemographic data and APOE-ε4 augmentation for MRI-based detection of amnestic mild cognitive impairment using deep learning systems. Plos one. 2020;15:e0236868.
[19] Yeung MK, Chau AK-y, Chiu JY-c, Shek JT-l, Leung JP-y, Wong TC-h. Differential and subtype-specific neuroimaging abnormalities in amnestic and nonamnestic mild cognitive impairment: A systematic review and meta-analysis. Ageing Research Reviews. 2022;80:101675.
[20] Qu Y, Ma Y-H, Huang Y-Y, Ou Y-N, Shen X-N, Chen S-D, et al. Blood biomarkers for the diagnosis of amnestic mild cognitive impairment and Alzheimer's disease: A systematic review and meta-analysis. Neuroscience & Biobehavioral Reviews. 2021;128:479-86.
[21] van Maurik IS, Zwan MD, Tijms BM, Bouwman FH, Teunissen CE, Scheltens P, et al. Interpreting biomarker results in individual patients with mild cognitive impairment in the Alzheimer's biomarkers in daily practice (ABIDE) project. JAMA neurology. 2017;74:1481-91.
[22] Handels RLH, Vos SJB, Kramberger MG, Jelic V, Blennow K, van Buchem M, et al. Predicting progression to dementia in persons with mild cognitive impairment using cerebrospinal fluid markers. Alzheimer's & Dementia. 2017;13:903-12.
[23] van Harten AC, Visser PJ, Pijnenburg YAL, Teunissen CE, Blankenstein MA, Scheltens P, et al. Cerebrospinal fluid Aβ42 is the best predictor of clinical progression in patients with subjective complaints. Alzheimer's & dementia. 2013;9:481-7.
[24] Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiology of aging. 2011;32:2322-e19.
[25] Snyder PJ, Lim YY, Schindler R, Ott BR, Salloway S, Daiello L, et al. Microdosing of scopolamine as a "cognitive stress test": Rationale and test of a very low dose in an at-risk cohort of older adults. Alzheimer's and Dementia. 2014;10:262--7.
[26] Fleming VB, Harris JL. Complex discourse production in mild cognitive impairment: detecting subtle changes. Aphasiology. 2008;22:729-40.

[27] Hayes J, Flower L. Identifying the organization of writing processes. 1980. p. 3.
[28] Ulatowska HK, Olness GS, Williams LJ. Coherence of narratives in aphasia. Brain and Language. 2004;91:42-3.
[29] Ulatowska HK, Streit Olness G, Samson AM, Keebler MW, Goins KE. On the nature of personal narratives of high quality. Advances in Speech Language Pathology. 2004;6:3-14.
[30] Goodglass H, Kaplan E, Barresi B. BDAE-3: Boston Diagnostic Aphasia Examination–Third Edition: Lippincott Williams & Wilkins Philadelphia, PA; 2001.
[31] Themistocleous C, Eckerström M, Kokkinakis D. Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. PLoS One. 2020;15:e0236009.
[32] Themistocleous C, Eckerström M, Kokkinakis D. Identification of Mild Cognitive Impairment From Speech in Swedish Using Deep Sequential Neural Networks. Frontiers in Neurology. 2018;9:975.
[33] Fraser KC, Lundholm Fors K, Kokkinakis D. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. Computer Speech & Language. 2019;53:121-39.
[34] Hernández-Domínguez L, Ratté S, Sierra-Martínez G, Roche-Bergua A. Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring. 2018;10:260-8.
[35] Fraser KC, Lundholm Fors K, Eckerström M, Themistocleous C, Kokkinakis D. Improving the Sensitivity and Specificity of MCI Screening with Linguistic Information. Proceedings of the LREC 2018 Workshop "Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)". 2018:19-26.
[36] König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring. 2015;1:112-24.
[37] Calzà L, Gagliardi G, Rossini Favretti R, Tamburini F. Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. Computer Speech & Language. 2021;65:101113.
[38] López-de-Ipiña K, Solé-Casals J, Eguiraun H, Alonso JB, Travieso CM, Ezeiza A, et al. Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. Computer Speech & Language. 2015;30:43-60.
[39] Toledo CM, Aluísio SM, Dos Santos LB, Brucki SMD, Trés ES, de Oliveira MO, et al. Analysis of macrolinguistic aspects of narratives from individuals with Alzheimer's disease, mild cognitive impairment, and no cognitive impairment. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2018;10:31-40.
[40] Clarke N, Barrick TR, Garrard P. A comparison of connected speech tasks for detecting early Alzheimer's disease and mild cognitive impairment using natural language processing and machine learning. Frontiers in Computer Science. 2021;3:634360.
[41] Fraser KC, Meltzer JA, Graham NL, Leonard C, Hirst G, Black SE, et al. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex. 2014;55:43-60.
[42] Themistocleous C, Ficek B, Webster K, den Ouden D-B, Hillis AE, Tsapkini K. Automatic subtyping of individuals with Primary Progressive Aphasia. bioRxiv. 2020:2020.04.04.025593.

[43] Jack Jr CR, Bennett DA, Blennow K, Carrillo MC, Dunn B, Haeberlein SB, et al. Research Framework: toward a biological definition of Alzheimer's disease. Alzheimers Dement. 2018 Apr; 14 (4): 535–62. doi: 10.1016/j. jalz. 2018.02. 018.
[44] Folstein MF, Folstein SE, McHugh PR. “Mini-mental state”: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research. 1975;12:189-98.
[45] Wechsler D. WAIS-III : Wechsler adult intelligence scale. 3d ed. San Antonio: The Psychological Corporation : Harcourt Brace & Company San Antonio; 1997.
[46] Rey A. L'Examen psychologique dans les cas d'encéphalopathie traumatique ... Avec 4 figures1941.
[47] Reitan RM, Wolfson D. A selective and critical review of neuropsychological deficits and the frontal lobes. Neuropsychology review. 1994;4:161-98.
[48] Trenerry MR, Crosson BA, DeBoe J, Leber WR. Stroop neuropsychological screening test: Psychological Assessment Resources; 1989.
[49] Themistocleous C. Computational Language Assessment: Open Brain AI. arXiv. 2023;2306.06693:1-17.
[50] Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.
[51] Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data mining and knowledge discovery. 2014;28:92-122.
[52] Torres VL, Rosselli M, Loewenstein DA, Curiel RE, Uribe IV, Lang M, et al. Types of Errors on a Semantic Interference Task in Mild Cognitive Impairment and Dementia. Neuropsychology. 2019;33:670-84.
[53] Buschke H, Mowrey WB, Ramratan WS, Zimmerman ME, Loewenstein DA, Katz MJ, et al. Memory Binding Test Distinguishes Amnestic Mild Cognitive Impairment and Dementia from Cognitively Normal Elderly. Archives Of Clinical Neuropsychology: The Official Journal Of The National Academy Of Neuropsychologists. 2017;32:29-39.
[54] Jester DJ, Andel R, Cechov. Cognitive phenotypes of older adults with subjective cognitive decline and amnestic mild cognitive impairment: The Czech Brain Aging Study. Journal of the International Neuropsychological Society. 2021;27:329--42.
[55] Kwak K, Giovanello KS, Bozoki A, Styner M, Dayan E. Subtyping of mild cognitive impairment using a deep learning model based on brain atrophy patterns. Cell Reports Medicine. 2021;2.
[56] McLane HC, Berkowitz AL, Patenaude BN, McKenzie ED, Wolper E, Wahlster S, et al. Availability, accessibility, and affordability of neurodiagnostic tests in 37 countries. Neurology. 2015;85:1614--22.
[57] Beeson P, Rapcsak S. Clinical diagnosis and treatment of spelling disorders. In: Hillis A, editor. The handbook of adult language disorders2015. p. 145--70.
[58] He R, Chapin K, Al-Tamimi J, Bel N, Marqui. Automated classification of cognitive decline and probable Alzheimer's dementia across multiple speech and language domains. American Journal of Speech-Language Pathology. 2023;32:2075--86.
[59] Love T, Oster E. On the categorization of aphasic typologies: the SOAP (a test of syntactic complexity). J Psycholinguist Res. 2002;31:503-29.
[60] Themistocleous C, Webster K, Afthinos A, Tsapkini K. Part of Speech Production in Patients With Primary Progressive Aphasia: An Analysis Based on Natural Language Processing. American Journal of Speech-Language Pathology. 2020:1-15.

[61] Pekkala S, Wiener D, Himali J, Beiser A, Obler L, Liu Y. Lexical retrieval in discourse: An early indicator of Alzheimer's dementia. Clin Linguist Phonetics. 2013;27:905--21.
[62] Forbes-McKay K, Shanks M, Venneri A. Charting the decline in spontaneous writing in Alzheimer's disease: a longitudinal study. Acta Neuropsychiatr. 2014;26:246--52.
[63] Kim H, Walker A, Shea J, Hillis AE. Written Discourse Task Helps to Identify Progression from Mild Cognitive Impairment to Dementia. Dement Geriatr Cogn Disord. 2021;50:446-53.
[64] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. CoRR. 2013;abs/1301.3781.
[65] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. p. 1532-43.
[66] Chang HT, Chiu MJ, Chen TF, Liu MY, Fan WC, Cheng TW, et al. Deterioration and predictive values of semantic networks in mild cognitive impairment. Journal of Neurolinguistics. 2022;61:101025.
[67] Juncos-Rabadn O, Facal D, Lojo-Seoane C, Pereiro AX. Tip-of-the-tongue for proper names in non-amnestic mild cognitive impairment. Journal of Neurolinguistics. 2013;26:409--20.
[68] Mueller KD, Hermann B, Mecollari J, Turkstra LS. Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. J Clin Exp Neuropsychol. 2018;40:917-39.

**Conflicts**

HK and CT do not have any conflicts of interest.

**Funding Sources**

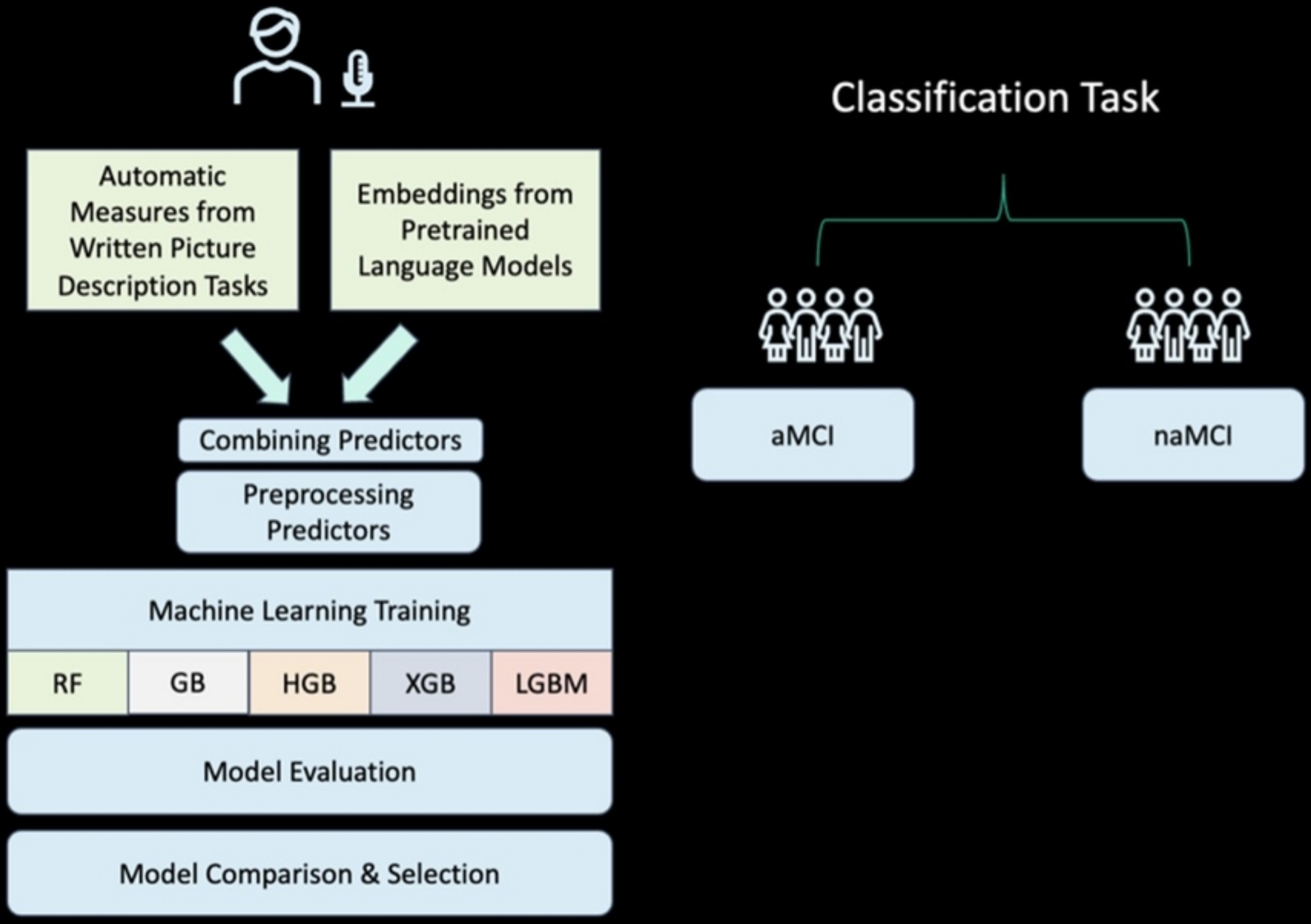
No funding source available.

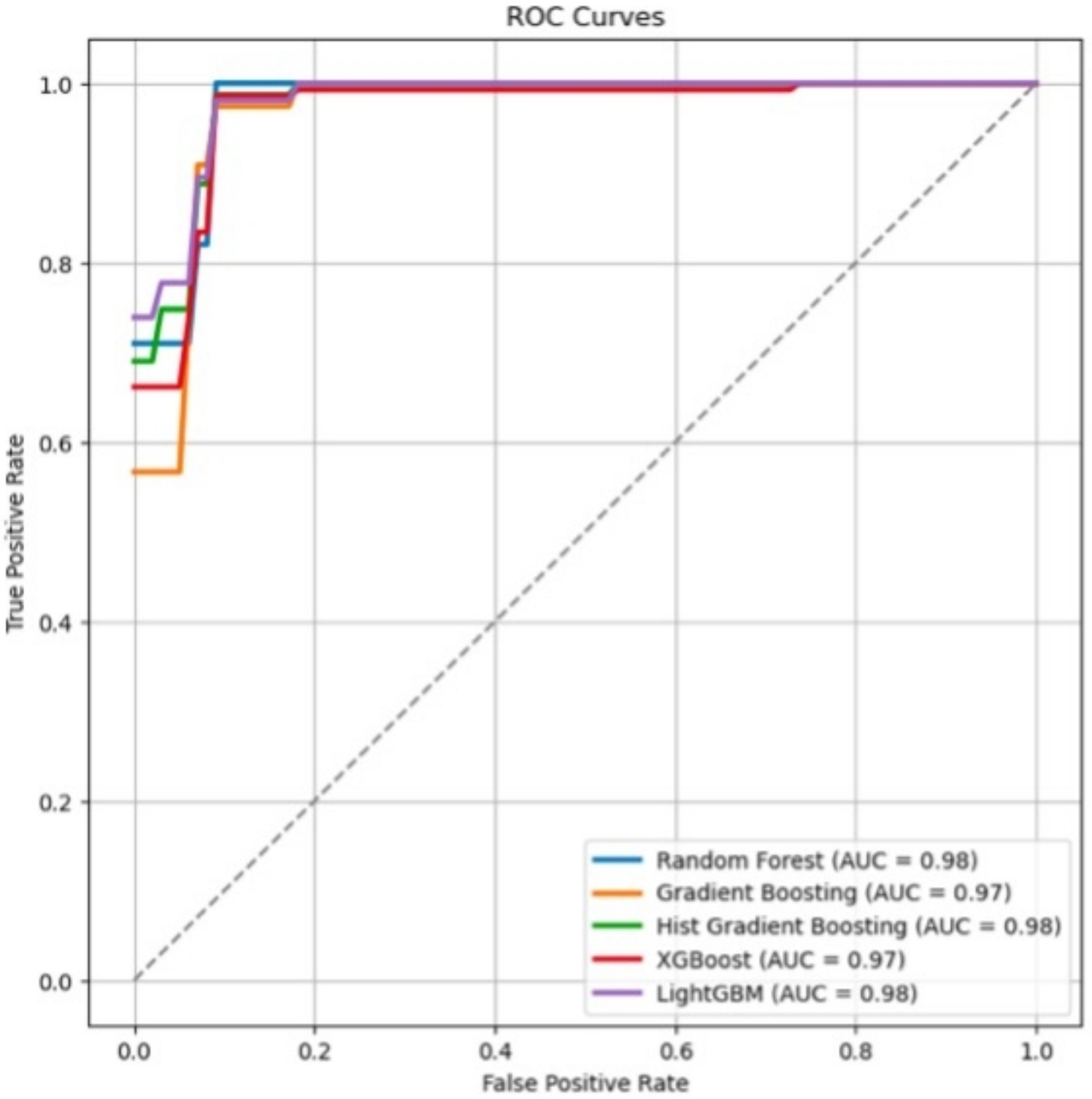
**Consent Statement**

All human subjects provided informed consent.

**Keywords**

mild cognitive impairment, writing, written discourse, machine learning, natural language processing


Automatic Measures from Written Picture Description Tasks
Embeddings from Pretrained Language Models
Combining Predictors
Preprocessing Predictors
Machine Learning Training
RF
GB
HGB
XGB
LGBM
Model Evaluation
Model Comparison & Selection
Classification Task
aMCI
naMCI


ROC Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC = 0.98)
Gradient Boosting (AUC = 0.97)
Hist Gradient Boosting (AUC = 0.98)
XGBoost (AUC = 0.97)
LightGBM (AUC = 0.98)