J Korean Med Sci. 2025 Mar 24;40(11):e26
https://doi.org/10.3346/jkms.2025.40.e26
eISSN 1598-6357·pISSN 1011-8934

JKMS

Original Article
Medical Informatics

Explainability Enhanced Machine Learning Model for Classifying Intellectual Disability and Attention-Deficit/Hyperactivity Disorder With Psychological Test Reports

Tong Min Kim [1], **Young-Hoon Kim** [2], **Sung-Hee Song** [3], **In-Young Choi** [1], **Dai-Jin Kim** [1,4] and **Taehoon Ko** [1,5,6]

[1]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Korea
[2]Department of Pediatrics, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Uijeongbu, Korea
[3]Wellysis Corp., Seoul, Korea
[4]Department of Psychiatry, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea
[5]Department of Medical Sciences, College of Medicine, The Catholic University of Korea, Seoul, Korea
[6]CMC Institute for Basic Medical Science, The Catholic Medical Center of The Catholic University of Korea, Seoul, Korea

Received: Jul 17, 2024
Accepted: Oct 17, 2024
Published online: Nov 15, 2024

Address for Correspondence:
Taehoon Ko, PhD
Department of Medical Informatics, College of Medicine, The Catholic University of Korea, 222 Banpo-daero, Seocho-gu, Seoul 06591, Korea.
Email: thko@catholic.ac.kr

ORCID iDs
Tong Min Kim
https://orcid.org/0000-0001-7381-8303
Young-Hoon Kim
https://orcid.org/0000-0002-6151-6911
Sung-Hee Song
https://orcid.org/0000-0001-6225-2412
In-Young Choi
https://orcid.org/0000-0002-2860-9411
Dai-Jin Kim
https://orcid.org/0000-0001-9408-5639
Taehoon Ko
https://orcid.org/0000-0002-4045-0036

Funding
This work was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number

ABSTRACT

Background: Psychological test reports are essential in assessing intellectual functioning, aiding in diagnosing and treating intellectual disability (ID) and attention-deficit/hyperactivity disorder (ADHD). However, these reports can have several problems because they are diverse, unstructured, subjective, and involve human errors. Additionally, physicians often do not read the entire report, and the number of reports is lower than that of diagnoses.
Methods: We developed explainable predictive models for classifying IDs and ADHDs based on written reports to address these issues. The reports of 1,475 patients with IDs and ADHDs who underwent intelligence tests were used for the models. These models were developed by analyzing reports using natural language processing (NLP) and incorporating the physician's diagnosis for each report. We selected n-gram features from the models' results by extracting important features using SHapley Additive exPlanations and permutation importance to make the models explainable. Developing the n-gram feature-based original text search system compensated for the lack of human readability caused by NLP and enabled the reconstruction of human-readable texts from the selected n-gram features.
Results: The maximum model accuracy was 0.92, and the 80 human-readable texts were restored from four models.
Conclusion: The results showed that the models could accurately classify IDs and ADHDs, even with a few reports. The models were also able to explain their predictions. The explainability-enhanced model can help physicians understand the classification process of IDs and ADHDs and provide evidence-based insights.

Keywords: Neurodevelopmental Disorder; Intellectual Disability; Attention-Deficit Hyperactivity Disorder Psychological Test Reports; Natural Language Processing; Machine Learning; Explainable Model

RS-2022-KH124685, RS-2021-KH114279, RS-2023-KH134396), and the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (grant number 2022R1C1C2002987).

Disclosure

The authors have no potential conflicts of interest to disclose.

Author Contributions

Conceptualization: Kim TM, Kim YH, Ko T. Data curation: Kim TM. Formal analysis: Kim TM, Song SH. Funding acquisition: Kim TM, Ko T. Investigation: Kim TM, Kim YH. Methodology: Kim TM, Choi IY, Ko T. Project administration: Kim TM, Ko T. Software: Kim TM, Song SH. Supervision: Ko T. Validation: Kim TM, Song SH, Ko T. Visualization: Kim TM. Resources: Kim YH, Kim DJ. Writing - original draft: Kim TM. Writing - review & editing: Kim TM, Kim YH, Song SH, Choi IY, Kim DJ, Ko T.

INTRODUCTION

Neurodevelopmental disorders are defined as conditions that begin during the developmental period and result in deficits that cause impairment in functioning.[1] Among them, intellectual disability (ID) and attention-deficit/hyperactivity disorder (ADHD) are mostly prevalent.[2] According to the Diagnostic and Statistical Manual of Mental Disorders,[3] patients are diagnosed with an ID if they have deficiencies in both intellectual functioning and adaptive behaviors that originate during developmental stages. Individuals with IDs often experience challenges in learning, reasoning, problem-solving, and adjusting to daily life. Such disabilities can impact cognitive abilities and the capacity to learn and adapt to unfamiliar scenarios.[3] Therefore, treatment and support for IDs should be individualized and may involve educational interventions, speech therapy, occupational therapy, and social skills training. The emphasis is on helping individuals obtain adaptive skills and reach their maximum potential despite their cognitive limitations.[4,5]

On the other hand, ADHD patients are diagnosed with specific symptoms, which fall into two main categories: inattention and hyperactivity-impulsivity. The primary symptoms of ADHD are inattention (difficulty sustaining attention, careless mistakes), hyperactivity (excessive restlessness or fidgeting), and impulsivity (difficulty waiting one's turn, blurting out answers). It does not inherently affect an individual's intellectual abilities. Unlike ID patients, ADHD patients do not have deficiencies in intellectual functioning and adaptive behaviors but have average or above-average intellectual functionality.[3] Therefore, their difficulties are mainly related to execution function and attention control rather than the cognitive impairment suffered by ID patients.[3] Treatment for ADHD generally involves behavioral interventions, psychoeducation, and medication, such as stimulant or non-stimulant medications. These aims to improve attention, control impulses, and reduce hyperactivity.[6]

To summarize, ID and ADHD differ significantly regarding cognitive functioning when comorbidities are not considered. ID is characterized by below-average intellectual functioning, whereas ADHD is characterized by average or above-average intellectual functioning.[3] Therefore, the treatments and interventions for these two conditions should be prescribed differently based on intellectual functioning.[7-10]

Psychologists select and administer various tests in clinical settings, recording observations and scores in a psychological test report, which will be stored in the hospital's electronic medical record (EMR) system.[11,12] Physicians then diagnose comprehensively based on these reports and other clinical data. However, these reports have limitations, including potential human error in scoring and subjective analysis.[13,14] Variations in writing styles, language use, and potential typographical errors further complicate consistency.[15,16] Also, physicians often do not review entire reports, and with an 84% match rate between psychologists' suggestions and final diagnoses (**Table 1**), understanding the complete report is crucial. Additionally, psychological tests are conducted infrequently, leading to fewer reports relative to physician diagnoses.

To address challenges, this study aimed to develop a system with high classification accuracy for ID and ADHD using few reports. It focused on minimizing human error by converting subjective texts into objective factors and ensuring report standardization. Sensitive data handling necessitated the use of explainable artificial intelligence[17,18] to provide evidence-based insights.[19]

Table 1. Demographic characteristics of the ID and ADHD groups

Characteristics	No. (%) of participants			*P* value[a]
	ID (n = 1,014, 68.7%)	ADHD (n = 461, 31.3%)	Total (N = 1,475, 100%)	
Sex				< 0.001
Male	641 (43.5)	350 (23.7)	484 (32.8)	
Female	373 (25.3)	111 (7.5)	991 (67.2)	
Age, yr				< 0.001
< 10	239 (16.2)	233 (15.8)	472 (32.0)	
10–19	360 (24.4)	217 (14.7)	577 (39.1)	
20–29	153 (10.4)	11 (0.7)	164 (11.1)	
30–39	73 (4.9)	0 (0.0)	73 (4.9)	
40–49	79 (5.4)	0 (0.0)	79 (5.4)	
50–59	88 (6.0)	0 (0.0)	88 (6.0)	
60–69	19 (1.3)	0 (0.0)	19 (1.3)	
70–79	3 (0.2)	0 (0.0)	3 (0.2)	
Department				< 0.001
Pediatrics	48 (3.3)	91 (6.2)	139 (9.4)	
Psychiatry	797 (54.0)	364 (24.7)	1,161 (78.7)	
Rehabilitation Medicine	168 (11.4)	6 (0.4)	174 (11.8)	
Plastic and Reconstructive Surgery	1 (0.1)	0 (0.0)	1 (0.1)	
Diagnostic concordance rate				< 0.001
Consistent	913 (61.9)	327 (22.2)	1,240 (84.1)	
Inconsistent	38 (2.6)	90 (6.1)	128 (8.7)	
Both	63 (4.3)	44 (3.0)	107 (7.3)	

ID = intellectual disability, ADHD = attention-deficit/hyperactivity disorder.
[a]The difference in the distribution of each variable between ID and ADHD was tested using the χ^2 test.

We utilized Natural Language Processing (NLP) to transform unstructured data into structured formats, facilitating text data use in EMRs.[20] Four classification models—Naïve Bayes (NB),[21] random forest (RF),[22] eXtreme Gradient Boosting (XGB),[23] and Light Gradient Boosting Machine (LGBM)[24]—were selected and optimized via Bayesian Optimization (BO).[25] Explainability was achieved by extracting features using SHapley Additive exPlanations (SHAP)[26] and permutation importance (PI).[27] An n-gram feature-based original text search (NOTS) system was developed to present NLP-derived features in a human-readable format to enhance interpretability.

METHODS

Participants and study design

The study was conducted on a cohort of patients (n = 1,816) with IDs and ADHDs between January 1, 2011, and May 31, 2022. The selection criteria were age 0–90 years, a confirmed diagnosis of ID or ADHD, and having undergone an intelligence test before the diagnosis. The final number of reports was 1,475, as 341 reports were excluded based on the following exclusion criteria: 1) identical contents (n = 202), 2) a diagnosis of both ID and ADHD (n = 136), and 3) missing contents (n = 3).

Fig. 1 shows the flowchart of developing explainable predictive models for classifying IDs and ADHDs in this study. We first went through pre-processing (**Fig. 1A**) and NLP (**Fig. 1B**) to transform unstructured data into model-optimized structured data. Then, we used the generated structured data to develop classification models (**Fig. 1C**). Finally, we provided the explainability of the models' results (**Fig. 1D**).

https://doi.org/10.3346/jkms.2025.40.e26

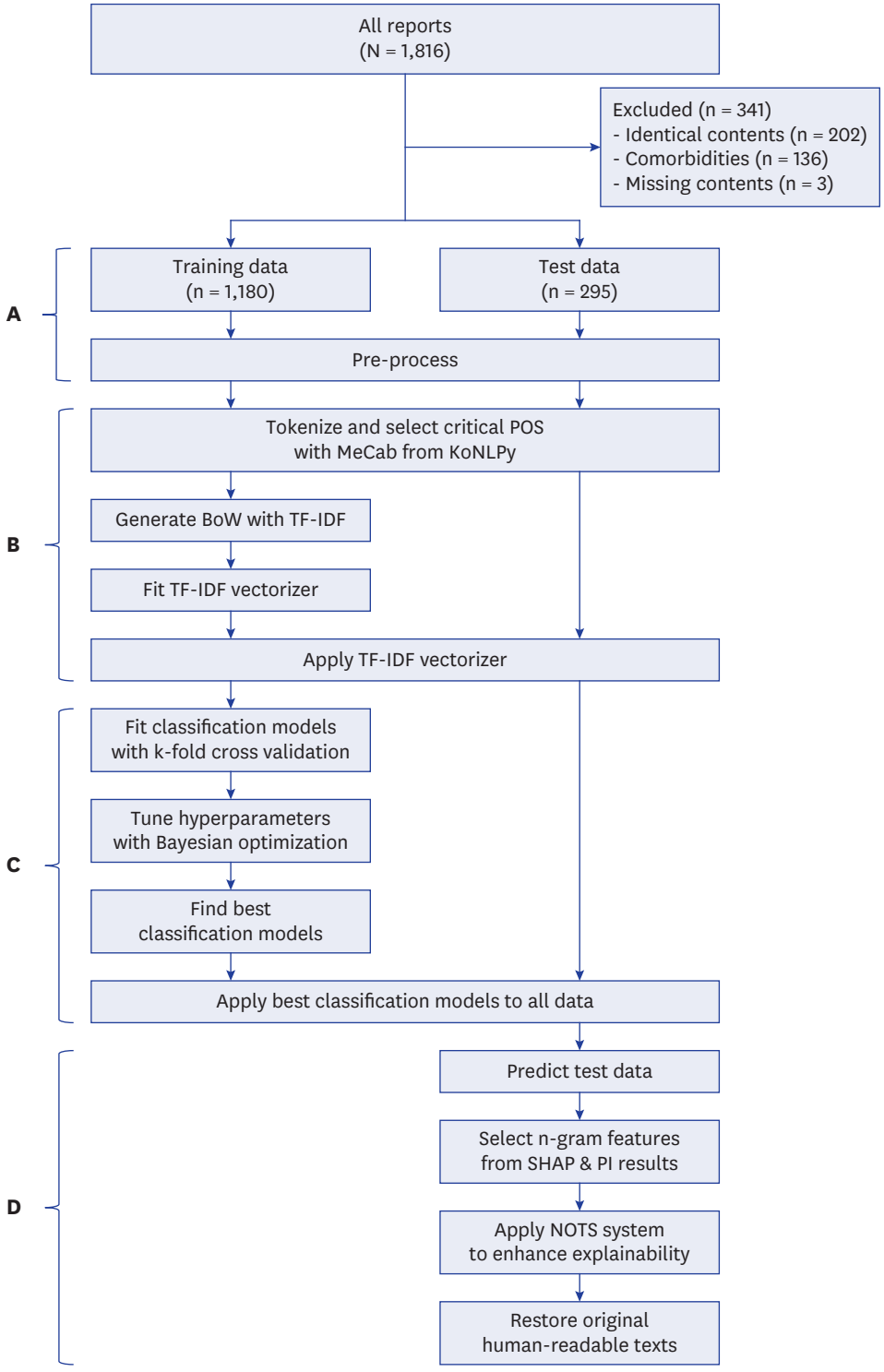


Fig. 1. Flowchart of developing explainable predictive models for classifying intellectual disabilities and attention-deficit/hyperactivity disorders. (**A**) Data pre-processing, (**B**) natural language processing, (**C**) classification model development, (**D**) explainable model development.
POS = part of speech, KoNLPy = Korean natural language process in Python, BoW = Bag-of-Words, TF = text frequency, IDF = inverse document frequency, SHAP = SHapley Additive exPlanations, PI = permutation importance, NOTS = n-gram feature-based original text search.

https://doi.org/10.3346/jkms.2025.40.e26

Data pre-processing

The input data used in this study were the contents of reports written by the psychologists who examined the patients. The output data were the diagnoses of ID or ADHD by a physician, integrating the report with other EMR data. **Fig. 1A** shows how the data of 1,475 reports were distributed. The reports were then divided in an 8:2 ratio by stratified sampling into a training and a test set (n = 1,180 and 295, respectively). All digits and special characters were deleted from all reports to measure the impact of the natural languages used by psychologists. Deleting the digits allows the models to classify IDs and ADHDs by focusing on natural languages rather than being influenced by various test scores (**Fig. 1A**).

We also compared the diagnostic concordance rates between the physicians' diagnoses and the psychologists' diagnostic suggestions to compare the accuracy of the models developed in this study. After comparing the physicians' and psychologists' diagnoses, we found three prominent categories: cases where the psychologist diagnosed both ID and ADHD as comorbid conditions, cases where the diagnoses of the physician and psychologist were consistent, and cases where their diagnoses were inconsistent ('Diagnostic concordance rate' of **Table 1**).

NLP

The psychologists' reports were mainly written in Korean. **Fig. 2** shows an example of the English translation of one such report. Because NLP performs best when applied appropriately to a language,[28] we used NLP optimized for Korean. First, we analyzed the reports based on morpheme analysis or phoneme separation[29,30] using the MeCab Korean natural language process in Python (KoNLPy) library.[31] Based on the analysis results, we tokenized all words in the reports and selected principal parts of speech (POSes): verbs, adjectives, adverbs, nouns, and foreign languages.[32]

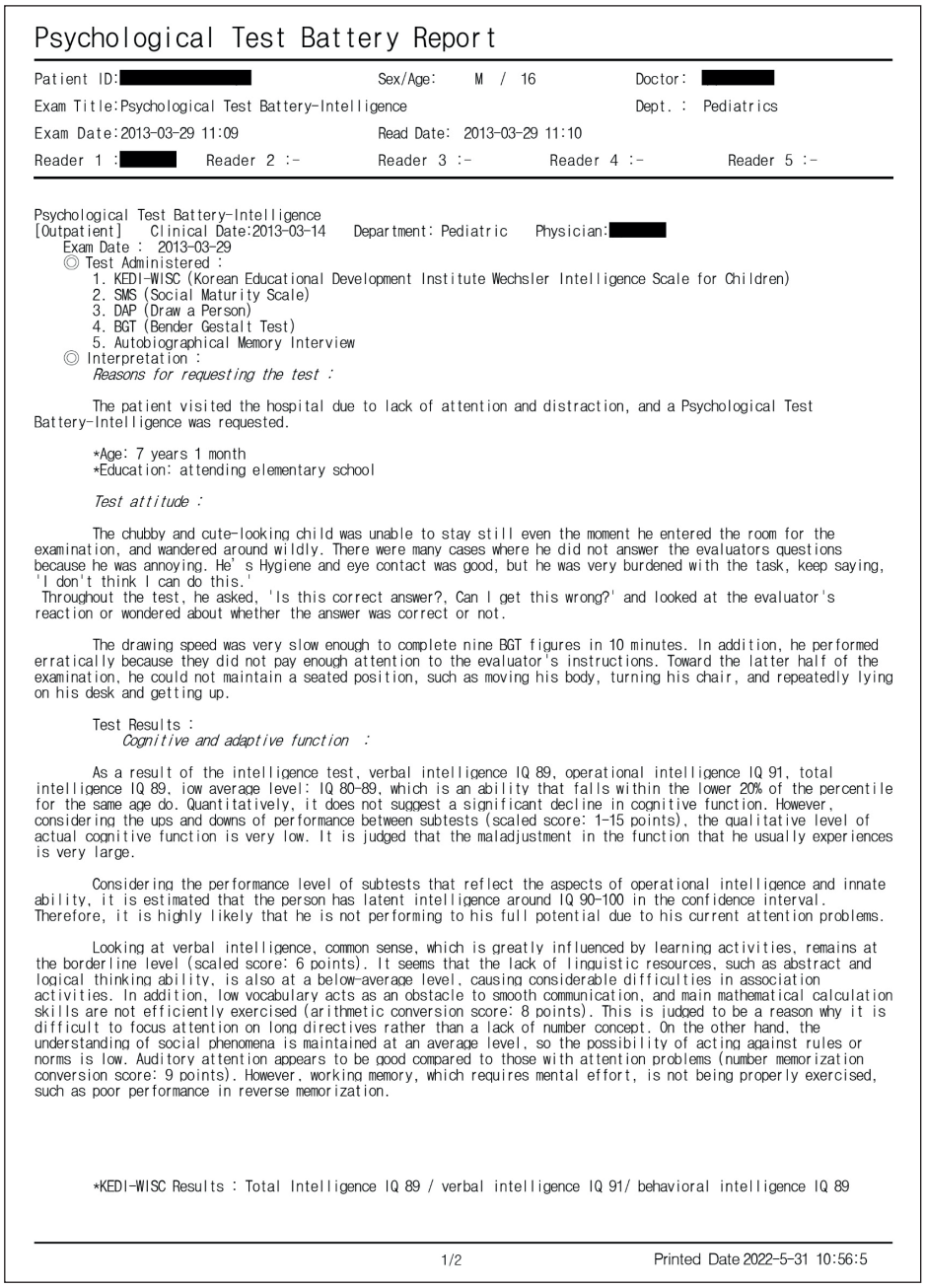
Before the classification models could be trained using the tokenized and selected words from the reports, these words needed to be feature-engineered and transformed into numeric representations. We used the text frequency-inverse document frequency (TF-IDF) method[33] for feature engineering. The TF-IDF is a frequency-based numeric representation of each word and is calculated as the product of TF and IDF. TF is the word count in each observation, and IDF is calculated as the log of document counts divided by the count of documents containing the word in question.[34] We first fit the TF-IDF vectorizer with the tokenized and selected training data. Then, we applied the fitted TF-IDF vectorizer to all data. We calculated the TF-IDF weight scores from unigram to heptagram (n-gram = 1 to 7) because a continuous word series can create new meanings (**Fig. 1B**).

Classification model development and evaluation

We used four classification models for explainability and enhanced the model: NB, RF, XGB, and LGBM. Deep learning models such as LSTM and Transformer can provide explicability through technologies such as SHAP and Attention Mechanisms,[35] but we excluded them because they cannot be fully converted to human-readable text using the NOTS system described later.

All models were fitted on the vectorized training split using stratified 5-fold cross-validation to mitigate potential bias in diagnosing ID and ADHD. This method ensures that each subset of the data used during the training and evaluation phases is representative of the overall distribution of key demographic and diagnostic variables. By maintaining a proportional

https://doi.org/10.3346/jkms.2025.40.e26



Psychological Test Battery Report

Patient ID: [redacted] Sex/Age: M / 16 Doctor: [redacted]

Exam Title: Psychological Test Battery-Intelligence Dept. : Pediatrics

Exam Date: 2013-03-29 11:09 Read Date: 2013-03-29 11:10

Reader 1 : [redacted] Reader 2 :- Reader 3 :- Reader 4 :- Reader 5 :-

Psychological Test Battery-Intelligence
[Outpatient] Clinical Date:2013-03-14 Department: Pediatric Physician: [redacted]
Exam Date : 2013-03-29
◎ Test Administered :
1. KEDI-WISC (Korean Educational Development Institute Wechsler Intelligence Scale for Children)
2. SMS (Social Maturity Scale)
3. DAP (Draw a Person)
4. BGT (Bender Gestalt Test)
5. Autobiographical Memory Interview
◎ Interpretation :
Reasons for requesting the test :

The patient visited the hospital due to lack of attention and distraction, and a Psychological Test Battery-Intelligence was requested.

*Age: 7 years 1 month
*Education: attending elementary school

Test attitude :

The chubby and cute-looking child was unable to stay still even the moment he entered the room for the examination, and wandered around wildly. There were many cases where he did not answer the evaluators questions because he was annoying. He' s Hygiene and eye contact was good, but he was very burdened with the task, keep saying, 'I don't think I can do this.'
Throughout the test, he asked, 'Is this correct answer?, Can I get this wrong?' and looked at the evaluator's reaction or wondered about whether the answer was correct or not.

The drawing speed was very slow enough to complete nine BGT figures in 10 minutes. In addition, he performed erratically because they did not pay enough attention to the evaluator's instructions. Toward the latter half of the examination, he could not maintain a seated position, such as moving his body, turning his chair, and repeatedly lying on his desk and getting up.

Test Results :
Cognitive and adaptive function :

As a result of the intelligence test, verbal intelligence IQ 89, operational intelligence IQ 91, total intelligence IQ 89, iow average level: IQ 80-89, which is an ability that falls within the lower 20% of the percentile for the same age do. Quantitatively, it does not suggest a significant decline in cognitive function. However, considering the ups and downs of performance between subtests (scaled score: 1-15 points), the qualitative level of actual cognitive function is very low. It is judged that the maladjustment in the function that he usually experiences is very large.

Considering the performance level of subtests that reflect the aspects of operational intelligence and innate ability, it is estimated that the person has latent intelligence around IQ 90-100 in the confidence interval. Therefore, it is highly likely that he is not performing to his full potential due to his current attention problems.

Looking at verbal intelligence, common sense, which is greatly influenced by learning activities, remains at the borderline level (scaled score: 6 points). It seems that the lack of linguistic resources, such as abstract and logical thinking ability, is also at a below-average level, causing considerable difficulties in association activities. In addition, low vocabulary acts as an obstacle to smooth communication, and main mathematical calculation skills are not efficiently exercised (arithmetic conversion score: 8 points). This is judged to be a reason why it is difficult to focus attention on long directives rather than a lack of number concept. On the other hand, the understanding of social phenomena is maintained at an average level, so the possibility of acting against rules or norms is low. Auditory attention appears to be good compared to those with attention problems (number memorization conversion score: 9 points). However, working memory, which requires mental effort, is not being properly exercised, such as poor performance in reverse memorization.

*KEDI-WISC Results : Total Intelligence IQ 89 / verbal intelligence IQ 91/ behavioral intelligence IQ 89

1/2 Printed Date 2022-5-31 10:56:5

Fig. 2. Example of an English-translated version of a report.

representation of these variables within each fold of the cross-validation process, we aim to enhance the generalizability and reliability of our diagnostic model outcomes, thereby reducing the likelihood of biased results. This stratified approach is particularly crucial in medical and psychological assessments where discrepancies in sample representation can lead to skewed interpretations and ultimately affect clinical validity. Hyperparameter tuning through trial and error can be tedious and may lead to unsatisfactory outcomes.[36]

For this reason, robust tuning methods are crucial, particularly when aiming to find the maximum value at a sampling point for an unknown function.[37] Therefore, we used BO[25] to

tune hyperparameters of NB, RF, XGB, and LGBM models using the Gaussian process as a posterior distribution. BO is a more efficient algorithm for hyperparameter optimization than commonly used Grid Search and Random Search techniques.[38] After training the models, we applied the fitted model and a combination of hyperparameters to all the data with the best accuracy (**Fig. 1C**).

The evaluation metrics used in this study were accuracy, Area Under the Receiver Operating Characteristic Curve (ROC_AUC), positive predictive value (PPV), sensitivity, and F1 score. A comprehensive list of metrics was used because a single metric cannot judge models' performance. For example, accuracy and ROC_AUC are sensitive to imbalanced data, such as the data used in this study, and using these measures alone would give clinicians inaccurate results. Therefore, we also measured the PPV, sensitivity, and F1 scores to compensate for the problem caused by the data imbalance. PPV, sensitivity, and F1 scores had different values depending on whether ID or ADHD was labeled positive or negative. Therefore, the PPV, sensitivity, and F1 scores were calculated considering cases where the two diagnoses were labeled as positive and negative or negative and positive, respectively.[39,40]

Explainable model development

After predicting the test data using the four models developed, we made the prediction results explainable by the NOTS system (**Fig. 1D**). **Fig. 3** shows the detailed explainable model development flowchart. The explainable model consisted of two main processes: selecting n-gram features (**Fig. 3A-C**) and applying the NOTS system (**Fig. 3C-G**).

For selecting n-gram features, we first applied the SHAP[26] and PI[27] methods to the four models to extract 240 important features, the top 30 for each method (**Fig. 3A**, **Supplementary Tables 1** and **2**). The important features were printed as a list of words up to seven principal POSes because we created the training data only with principal POS

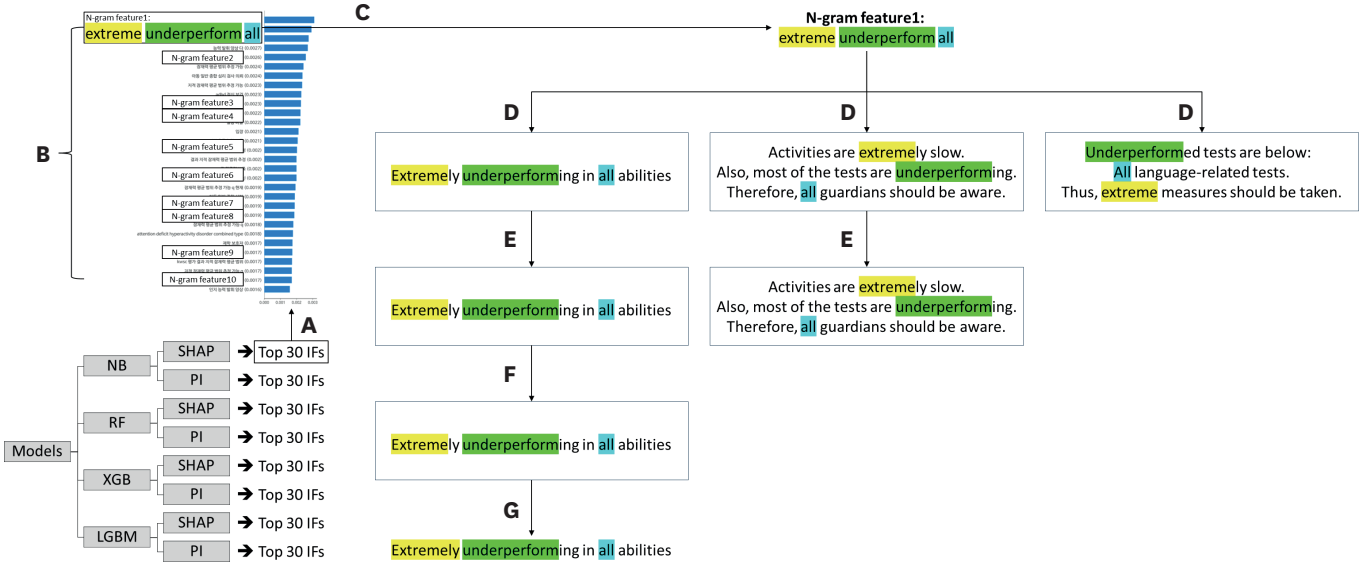


Fig. 3. Flowchart of the explainable model development. (**A**) Extract the top 30 important features from each method. (**B**) Select ten n-gram features from the top 30 important features. (**C**) Insert one of the ten n-gram features in the n-gram feature-based original text search system. (**D**) Check if the words of the n-gram feature are in the report. (**E**) Check if the words in the n-gram feature are in order. (**F**) Search texts with no other principal POS between the words of the n-gram feature. (**G**) Restore the original human-readable text.
NB = Naïve Bayes, RF = random forest, XGB = eXtreme Gradient Boosting, LGBM = Light Gradient Boosting Machine, SHAP = SHapley Additive exPlanations, PI = permutation importance, IF = important features, POS = part of speech.

https://doi.org/10.3346/jkms.2025.40.e26

(verbs, adjectives, adverbs, nouns, and foreign languages) and set the TF-IDF weight score from unigram to heptagram (n-gram = 1 to 7). However, there are cases where duplicated features appear consecutively due to the use of n-grams, and we have defined these features as 'duplicated features.' In addition, we have defined some features as 'irrelevant features' because they have very different meanings than other features. As a result, we removed 160 duplicate and irrelevant features out of 240 important features, leaving a total of 80 features (10 for each method), which we defined as 'n-gram features' (**Fig. 3B**). Finally, we inserted n-gram features into the NOTS system to restore them as human-readable texts, because these n-gram features were not in sentence form, they had to be restored to human-readable sentences or phrases (**Fig. 3C**). In summary, the process of selecting n-gram features resulted in the selection of 80 n-gram features out of 240 important features.

We developed a NOTS system that restores human-readable texts from the n-gram features (**Fig. 3C-G**). The system receives the inserted n-gram features (**Fig. 3C**). Then, the system checks if the words of the n-gram feature are in the report (**Fig. 3D**). Next, the system checks if the words in the n-gram feature are present and if the words are in the same order as the inserted n-gram feature (**Fig. 3E**). The system then searches texts with no other principal POS between the words of the n-gram feature (**Fig. 3F**). Because the n-gram features are already made up of only principal POS, there should be no principal POS other than n-gram features. Finally, the system restores the original human-readable text (**Fig. 3G**). Based on the results of the NOTS system, we restored 80 human-readable texts from 80 n-gram features (**Supplementary Tables 3** and **4**).

Consequently, the NOTS system enhances the transparency and comprehensibility of machine learning model decisions for clinicians by converting important features that significantly influence the model's outcomes into a more human-readable format.

Ethics statement

This study was approved by the Institutional Review Board of The Catholic University of Korea, Catholic Medical Center (XC22WIDI0062). The ethics panel determined that the study does not involve Human Subjects. Informed consent was waived because of the study's retrospective nature, and the analysis used anonymous clinical data.

RESULTS

Data characteristics

Table 1 summarizes the report characteristics of the ID and ADHD groups. There were more ID diagnoses (n = 1,014, 68.7%) than ADHD diagnoses (n = 461, 31.3%) across all sexes, and there were more diagnoses in males (n = 991, 67.2%) than in females (n = 484, 32.8%). The < 10 years (n = 472, 32.0%) and 10–19 years (n = 577, 39.1%) age groups accounted for the most significant proportion of the study subjects (n = 1,049, 71.1%). Among them, ID (n = 239, 16.2%) and ADHD (n = 233, 15.8%) were similar in the < 10 years age group. There was no ADHD in the patients aged 30–79 (n = 0, 0.0%). Most departments diagnosed ID more commonly (n = 966, 65.5%) than ADHD (n = 370, 25.1%); however, only the pediatric department diagnosed ADHD more commonly (n = 91, 6.2%) than ID (n = 48, 3.3%). The diagnostic concordance between physicians' diagnoses and psychologists' diagnostic suggestions was only 84.1%.

https://doi.org/10.3346/jkms.2025.40.e26

Table 2. Top 20 word counts for each ID and ADHD

ID	Count	ADHD	Count
Ability	11,026	Ability	4,557
Level	8,812	Level	4,524
Society	5,428	Performance	2,638
Test	5,126	Test	2,415
Language	4,204	Task	2,315
Intellectual	3,776	Assessment	2,259
Intelligence	3,732	According	2,192
Assessment	3,649	Language	2,148
Age	3,638	Range	1,976
Performance	3,526	Understanding	1,918
According	3,428	Indicator	1,848
Disorder	3,380	Average	1,749
Development	3,099	Ordinary	1,722
Possibility	3,004	Society	1,711
Case	2,971	Processing	1,533
Indicator	2,863	Reasoning	1,524
Status	2,861	Results	1,452
Very	2,827	Demonstrate	1,396
Retardation	2,571	Intellectual	1,381
Reasoning	2,502	Behavior	1,354

ID = intellectual disability, ADHD = attention-deficit/hyperactivity disorder.

Word counts from NLP

Table 2 lists the top 20 words and their counts for each ID and ADHD when a common noun, proper noun, general adverb, verb, adjective, and foreign language were extracted using POS analysis. Eleven of the 20 words appeared equally frequently in both diagnoses, among which the words 'ability,' 'level,' and 'test' had the same frequency ranking. The number of words with the same frequency ranking was twice as high for ID as for ADHD. ID diagnoses had several words related to intelligence (e.g., intellectual, disorder, development, retardation), whereas ADHD diagnoses had many words related to behavior (e.g., task, understanding, processing, demonstrate).

Classification models

Table 3 compares the performances of the four models. The first five columns show each model's accuracy, ROC_AUC, PPV, sensitivity, and F1 score values with the test data. The accuracies of the NB, RF, XGB, and LGBM models were 0.92, 0.91, 0.87, and 0.89, respectively. The NB model exhibited the highest accuracy (0.92) and ROC_AUC (0.91). All scores were higher when the ID was set to positive than when ADHD was set to positive.

Explaining reports

Table 4 shows the counts of duplicated, irrelevant, n-gram, and total features from the SHAP and PI results. The XGB and LGBM models did not have duplicated feature counts (0, 0%), whereas the RF (31, 12.9%) and NB (29, 12,1%) models had many duplicated features.

Table 3. Comparison of the performance of the four models

Models	Accuracy	ROC_AUC	PPV (ID, ADHD)	Sensitivity (ID, ADHD)	F1 score (ID, ADHD)
NB	0.92	0.91	0.95, 0.86	0.94, 0.89	0.94, 0.88
RF	0.91	0.87	0.91, 0.90	0.96, 0.78	0.93, 0.84
XGB	0.87	0.83	0.88, 0.85	0.94, 0.73	0.91, 0.78
LGBM	0.89	0.86	0.90, 0.89	0.96, 0.75	0.93, 0.82

ROC_AUC = Area Under the Receiver Operating Characteristic Curve, PPV = positive predictive value, ID = intellectual disability, ADHD = attention-deficit/hyperactivity disorder, NB = Naïve Bayes, RF = random forest, XGB = eXtreme Gradient Boosting, LGBM = Light Gradient Boosting Machine.

https://doi.org/10.3346/jkms.2025.40.e26

Table 4. Counts of duplicated, irrelevant, n-gram, and total features from the SHAP and PI results

No. (%) of features	NB		RF		XGB		LGBM		Total count
	SHAP	PI	SHAP	PI	SHAP	PI	SHAP	PI	
Duplicated	12 (5.0)	17 (7.1)	16 (6.7)	15 (6.3)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	60 (25.0)
Irrelevant	8 (3.3)	3 (1.3)	4 (1.7)	5 (2.1)	20 (8,3)	20 (8,3)	20 (8,3)	20 (8,3)	100 (41.7)
N-gram	10 (4.2)	10 (4.2)	10 (4.2)	10 (4.2)	10 (4.2)	10 (4.2)	10 (4.2)	10 (4.2)	80 (33.3)
Total	30 (12.5)	30 (12.5)	30 (12.5)	30 (12.5)	30 (12.5)	30 (12.5)	30 (12.5)	30 (12.5)	240 (100)

NB = Naïve Bayes, RF = random forest, XGB = eXtreme Gradient Boosting, LGBM = Light Gradient Boosting Machine, SHAP = SHapley Additive exPlanations, PI = permutation importance.

Supplementary Tables 1 and **2** show the English version of the 240 important features extracted, the top 30 for each method. These important features were originally extracted in Korean since they were extracted using the PI and SHAP methods applied in Korean reports. All four models found common occurrences of the words 'borderline intellectual,' 'comprehensive ability,' 'intellectual,' etc. **Supplementary Tables 3** and **4** list the selected 80 human-readable texts. In all four models, ID- and ADHD-related words were evenly selected. However, in the case of the XGB and LGBM models, words related to ID (BGT, KFD, TMT) tended to appear more frequently, whereas terms related to ADHD (adjustment disorder, borderline intellectual functioning) appeared more often in the RF and NB models.

DISCUSSION

In this study, we developed explainable predictive models for classifying IDs and ADHDs from free text in reports. We transformed unstructured text into structured data optimized for classification models using appropriate pre-processing and NLP techniques. Four models were developed and tailored to the report data, with optimization achieved through BO and k-fold cross-validation. We selected 80 n-gram features from the classification results to enhance model explainability using SHAP and PI, highlighting important features. The NOTS system was developed to address the readability issues introduced by NLP, restoring 80 human-readable texts from the selected n-gram features. Thus, the NOTS system improves the transparency and understandability of decisions made by machine learning models for clinicians by translating important features, which substantially impact the model's outcomes, into a format that is more accessible to human interpretation.

Analysis of the diagnostic concordance rates between the physician's diagnoses and the psychologist's diagnostic suggestions showed that the four models developed in this study were more accurate than the psychologist's diagnostic suggestions. The maximum accuracy of our models was 92%, and the diagnostic consistency between the psychologists and doctors was 84%; therefore, the model's accuracy was approximately 8% higher (**Tables 1** and **3**).

The words listed in **Table 2** have the same rank, but their counts are more than twice as many in the ID group as those in the ADHD group because the amount of ID data was twice as much as the ADHD data. **Table 2** also shows that more than half of the word counts in the ID and ADHD groups were similar. Therefore, it is impossible to classify ID and ADHD based on counts, and high accuracy cannot be achieved. These results suggest that AI methodologies are needed to classify ID and ADHD with high accuracy.

Table 3 shows that the accuracy and ROC_AUC of the NB model were higher than those of the XGB, LGBM, and RF models. Since the number of ID data was more significant than that

https://doi.org/10.3346/jkms.2025.40.e26

of the ADHD data, PPV, sensitivity, and F1 scores increased when the ID was set to positive. When considering accuracy, ROC_AUC, PPV, sensitivity, and F1 score, it can be concluded that the NB model is better than the XGB, LGBM, and RF models. The NB model is more accurate than others because it tends to show high accuracy in sparse matrix data, like those in the document-term matrix we created.[41]

Duplicated and irrelevant feature results are shown in **Table 4**. The NB and RF models had many duplicate features, whereas the XGB and LGBM models did not have duplicate features. This is because duplicated features can be extracted using NB and RF, as NB assigns high weights to features with specific n-gram features[42] and RF samples data points or features by restoration extraction.[43]

The four models and two feature extraction methods have different calculation formulas, resulting in slightly different important feature extractions, as demonstrated in **Supplementary Tables 1-4**. Nevertheless, the four models have commonalities as words related to ID or ADHD, such as 'intellectual,' 'ability,' 'cognitive,' 'perceptual,' etc., frequently appear. The findings also reveal that the four models rarely chose words associated with the physical appearance of patients, such as 'clean,' 'unsanitary,' 'chubby,' etc. Instead, all four models generated inventories of subtests, subtest outcomes, and behavioral descriptions during the examination.

This study has several limitations. It is necessary to verify the model's robustness by obtaining data from other institutions and conducting external validation. With an increased data volume from additional institutions, a multi-class classification model can be developed by further subdividing labels according to severity. In the pre-processing stage, a more diverse application of NLP methods, including stop words, is required. Additionally, a comparison of the performance of each morpheme analyzer should be conducted. Finally, we employed conventional machine learning models as classification tools to utilize techniques for reconstructing human-readable sentences based on SHAP and PI results. In future work, we plan to apply advanced models, such as large language models, to compare their performance with conventional machine learning models.

This study offers several key contributions. Firstly, we developed four explainable predictive models that accurately classify IDs and ADHDs using solely physician reports, eliminating numerical scores and focusing on textual information. Despite the limited quantity of reports, the models demonstrated high accuracy, ROC_AUC, PPV, sensitivity, and F1 scores, underscoring their efficacy even with a small dataset. This was achieved through meticulous pre-processing and applying NLP techniques to convert unstructured texts into structured data. Additionally, incorporating the physician's diagnosis into each report further refined the models' PPV. To enhance the explainability of the models, we employed SHAP and PI methods to identify relevant n-gram features, providing insights into the models' decision-making processes. To mitigate any loss of readability due to pre-processing, we developed a NOTS system, which clarifies the extracted n-gram features, thereby improving human understanding. Consequently, these models not only facilitate accurate ID and ADHD classification but also aid physicians in comprehending the classification process, offering evidence-based insights.

https://doi.org/10.3346/jkms.2025.40.e26

SUPPLEMENTARY MATERIALS

Supplementary Table 1

Top 30 SHAP-based important features extracted from each model

Supplementary Table 2

Top 30 PI-based important features extracted from each model

Supplementary Table 3

Selected 10 SHAP-based human-readable texts extracted from each model

Supplementary Table 4

Selected 10 PI-based human-readable texts extracted from each model

REFERENCES

1. Morris-Rosendahl DJ, Crocq MA. Neurodevelopmental disorders-the history and future of a diagnostic concept. *Dialogues Clin Neurosci* 2020;22(1):65-72. PUBMED | CROSSREF
2. Vissers LELM, Gilissen C, Veltman JA. Genetic studies in intellectual disability and related disorders. *Nat Rev Genet* 2016;17(1):9-18. PUBMED | CROSSREF
3. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR)*. Washington, D.C., USA: American Psychiatric Association Publishing; 2022.
4. Shea SE. Mental retardation in children ages 6 to 16. *Semin Pediatr Neurol* 2006;13(4):262-70. PUBMED | CROSSREF
5. Aman MG, Buican B, Arnold LE. Methylphenidate treatment in children with borderline IQ and mental retardation: analysis of three aggregated studies. *J Child Adolesc Psychopharmacol* 2003;13(1):29-40. PUBMED | CROSSREF
6. Brandon CL, Marinelli M, White FJ. Adolescent exposure to methylphenidate alters the activity of rat midbrain dopamine neurons. *Biol Psychiatry* 2003;54(12):1338-44. PUBMED | CROSSREF
7. Hässler F, Thome J. Mental retardation and ADHD. *Z Kinder Jugendpsychiatr Psychother* 2012;40(2):83-93. PUBMED | CROSSREF
8. Handen BL, McAuliffe S, Janosky J, Feldman H, Breaux AM. A playroom observation procedure to assess children with mental retardation and ADHD. *J Abnorm Child Psychol* 1988;26(4):269-77. PUBMED | CROSSREF
9. Aman MG, Kern RA, McGhee DE, Arnold LE. Fenfluramine and methylphenidate in children with mental retardation and ADHD: clinical and side effects. *J Am Acad Child Adolesc Psychiatry* 1993;32(4):851-9. PUBMED | CROSSREF
10. Handen BL, Breaux AM, Janosky J, McAuliffe S, Feldman H, Gosling A. Effects and noneffects of methylphenidate in children with mental retardation and ADHD. *J Am Acad Child Adolesc Psychiatry* 1992;31(3):455-61. PUBMED | CROSSREF
11. Vittengl JR, Jarrett RB, Ro E, Clark LA. Evaluating a comprehensive model of euthymia. *Psychother Psychosom* 2023;92(2):133-8. PUBMED | CROSSREF
12. Kaufman AS, Raiford SE, Coalson DL. *Intelligent Testing With the WISC-V*. Hoboken, NJ, USA: Wiley; 2016.
13. Styck KM, Walsh SM. Evaluating the prevalence and impact of examiner errors on the Wechsler scales of intelligence: a meta-analysis. *Psychol Assess* 2016;28(1):3-17. PUBMED | CROSSREF
14. Institute of Medicine. *Psychological Testing in the Service of Disability Determination*. Washington, D.C., USA: National Academies Press; 2015.
15. Belk MS, LoBello SG, Ray GE, Zachar P. WISC-III administration, clerical, and scoring errors made by student examiners. *J Psychoeduc Assess* 2002;20(3):290-300. CROSSREF
16. Slate JR, Hunnicutt LC. Examiner errors on the Wechsler scales. *J Psychoeduc Assess* 1988;6(3):280-8. CROSSREF
17. Angelov PP, Soares EA, Jiang R, Arnold NI, Atkinson PM. Explainable artificial intelligence: an analytical review. *Wiley Interdiscip Rev Data Min Knowl Discov* 2021;11(5):e1424. CROSSREF

18. Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al.. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. *ACM Computing Surveys* 2023;55(13s):1-42. CROSSREF

19. Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022). *Comput Methods Programs Biomed* 2022;226:107161. PUBMED | CROSSREF

20. Mermin-Bunnell K, Zhu Y, Hornback A, Damhorst G, Walker T, Robichaux C, et al. Use of natural language processing of patient-initiated electronic health record messages to identify patients with COVID-19 infection. *JAMA Netw Open* 2023;6(7):e2322299. PUBMED | CROSSREF

21. Jiang L, Zhang H, Cai Z. A novel bayes model: hidden naive bayes. *IEEE Trans Knowl Data Eng* 2009;21(10):1361-71. CROSSREF

22. Breiman L. Random forests. *Mach Learn* 2001;45(1):5-32. CROSSREF

23. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme gradient boosting as a method for quantitative structure-activity relationships. *J Chem Inf Model* 2016;56(12):2353-60. PUBMED | CROSSREF

24. Fan J, Ma X, Wu L, Zhang F, Yu X, Zeng W. Light Gradient Boosting Machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agric Water Manage* 2019;225:105758. CROSSREF

25. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: a review of Bayesian optimization. *Proc IEEE Inst Electr Electron Eng* 2016;104(1):148-75. CROSSREF

26. Vega García M, Aznarte JL. Shapley additive explanations for NO_2 forecasting. *Ecol Inform* 2020;56:101039. CROSSREF

27. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics* 2010;26(10):1340-7. PUBMED | CROSSREF

28. Joshi P, Santy S, Budhiraja A, Bali K, Choudhury M. The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 July 5–10. Kerrville, TX, USA: Association for Computational Linguistics; 2020, 6282-93.

29. Lee S, Jang H, Baik Y, Park S, Shin H. A small-scale Korean-specific BERT language model. *J KIISE* 2020;47(7):682-92. CROSSREF

30. Park K, Lee J, Jang S, Jung D. An empirical study of tokenization strategies for various Korean NLP tasks. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing; 2020 December 4–7; Suzhou, China. Kerrville, TX, USA: Association for Computational Linguistics; 2020, 133-42.

31. Park OE, Cho S. KoNLPy: Korean natural language processing in Python. Proceedings of the 26th Korean Language and Korean Language Information Processing Conference of the Annual Conference on Human and Language Technology in 2014; 2014 October 10–11; Chuncheon, Korea. Seoul, Korea: Human and Language Technology; 2014, 133-6.

32. Matteson A, Lee C, Kim Y, Lim H. Rich character-level information for Korean morphological analysis and part-of-speech tagging. Proceedings of the 27th International Conference on Computational Linguistics; 2018 August 20–26; Santa Fe, NM, USA. Kerrville, TX, USA: Association for Computational Linguistics; 2018, 2482-92.

33. Aizawa A. An information-theoretic perspective of tf–idf measures. *Inf Process Manage* 2003;39(1):45-65. CROSSREF

34. Goodman-Meza D, Shover CL, Medina JA, Tang AB, Shoptaw S, Bui AA. Development and validation of machine models using natural language processing to classify substances involved in overdose deaths. *JAMA Netw Open* 2022;5(8):e2225593. PUBMED | CROSSREF

35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 December 4–9; Long Beach, CA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2017, 6000-10.

36. Massaoudi M, Refaat SS, Chihi I, Trabelsi M, Oueslati FS, Abu-Rub H. A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting. *Energy* 2021;214:118874. CROSSREF

37. Shi R, Xu X, Li J, Li Y. Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimization. *Appl Soft Comput* 2021;109:107538. CROSSREF

38. Močkus J. On Bayesian methods for seeking the extremum. Proceedings of Optimization Techniques IFIP Technical Conference Novosibirsk; 1974 July 1–7. Berlin, Germany: Springer; 1975, 400-4.

39. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol* 2008;56(1):45-50. PUBMED | CROSSREF

40. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Front Public Health* 2017;5:307. PUBMED | CROSSREF

41. Ramage D, Manning CD, Dumais S. Partially labeled topic models for interpretable text mining. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2011 August 21–24; San Diego, CA, USA. New York, NY, USA: Association for Computing Machinery; 2011, 457-65.

42. Lewis DD. Naive (Bayes) at forty: the independence assumption in information retrieval. Proceedings of Machine Learning: ECML-98 (10th European Conference on Machine Learning); 1998 April 21–23; Chemnitz, Germany. Berlin, Germany: Springer; 1998, 4-15.

43. Kulkarni VY, Sinha PK. Pruning of random forest classifiers: a survey and future directions. Proceedings of 2012 International Conference on Data Science & Engineering (ICDSE); 2012 July 18–20; Cochin, India. New York, NY, USA: Institute of Electrical and Electronics Engineers (IEEE); 2012, 64-8.