Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 21(4): 281-286, Dec. 2023 Regular paper

Diagnosing Reading Disorders based on Eye Movements during Natural Reading

Yongseok Yoo*

School of Computer Science and Engineering, Soongsil University, Seoul 06978, Republic of Korea

Abstract

Diagnosing reading disorders involves complex procedures to evaluate complex cognitive processes. For an accurate diagnosis, a series of tests and evaluations by human experts are required. In this study, we propose a quantitative tool to diagnose reading disorders based on natural reading behaviors using minimal human input. The eye movements of the third- and fourth-grade students were recorded while they read a text at their own pace. Seven machine learning models were used to evaluate the gaze patterns of the words in the presented text and classify the students as normal or having a reading disorder. The accuracy of the machine learning-based diagnosis was measured using the diagnosis by human experts as the ground truth. The highest accuracy of 0.8 was achieved by the support vector machine and random forest classifiers. This result demonstrated that machine learning-based automated diagnosis could substitute for the traditional diagnosis of reading disorders and enable large-scale screening for students at an early age.

Index Terms: Eye movements, Machine learning, Natural reading, Reading disorder

I. INTRODUCTION

A. Reading Disorder

Reading is one of the most important skills for acquiring knowledge. Children shift from "learning to read" to "reading to learn" in around the fourth grade [1]. Students with poor literacy skills experience difficulties in acquiring new knowledge [2]. However, a significant proportion of the third- and fourth-grade students lagged because of poor reading skills. According to the 2019 National Assessment of Educational Progress report, 34% of the fourth graders performed below the basic reading level (for example, identifying main ideas and making simple inferences) in 2019 [3].

Diagnosing reading disorders (RDs) is challenging. This is because reading involves complex interactions between multiple cognitive processes such as letter-sound correspondence, phonological memory, word recognition, sentence processing, and comprehension. A deficiency in any low-level process can result in RDs [4,5]. For instance, deficiencies in lexicography and phonics can result in difficulties in word identification and dyslexia [4,5]. Higher-level cognitive processes, such as reading comprehension, are less understood [5].

The lack of systematic screening for RDs reduces opportunities for early intervention and support. Many students with RD go unnoticed until they experience a significant academic setback. Identifying the reasons for their low academic achievement can be challenging for both teachers and caregivers. Currently, existing diagnostic methods rely on standardized tests and evaluations by experts (Fig. 1A) [5], that require expertise in reading and considerable resources.

Eye-tracking methodology has gained significant attention as a cost-effective means of quantitatively assessing reading

Received 18 May 2023, **Revised** 3 September 2023, **Accepted** 9 September 2023
***Corresponding Author** Yongseok Yoo (E-mail: yyoo@ssu.ac.kr)
School of Computer Science and Engineering, Soongsil University, Seoul 06978, Republic of Korea

Open Access https://doi.org/10.56977/jicce.2023.21.4.281 print ISSN: 2234-8255 online ISSN: 2234-8883

skills. Eye tracking is a technique that records and analyzes the movements and positions of the eyes as participants process visual stimuli such as text on a screen [6,7]. By measuring the duration and frequency of gazes on individual words, real-time monitoring of the cognitive processes of a participant becomes possible, allowing researchers to gain valuable insights into the cognitive processes involved in reading [8-14]. Compared with other modalities such as electroencephalography and functional magnetic resonance imaging, eye tracking offers an affordable option to investigate the cognitive processes involved in reading [15,16].

However, the current applications of eye tracking in studying RDs are primarily focused on detecting dyslexia and understanding low-level processes such as word recognition [17, 18]. Although eye tracking provides valuable insights into gaze behavior and visual processing during reading, its relationship with reading comprehension remains unclear. Further research is required to explore the relationship of eye movements and visual processing with the higher-level cognitive processes involved in reading comprehension.

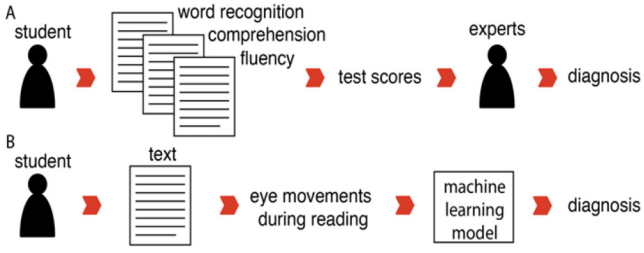

Fig. 1. Comparison of traditional diagnosis of reading disability based on standardized tests (A) and the proposed method based on natural reading using machine learning (B).

B. Contributions of the Study

This study investigated a scalable method for diagnosing RDs. The study proposed a quantitative evaluation of reading capabilities based on everyday reading experiences. Specifically, eye movements during natural reading were used to evaluate the reading capability using machine learning algorithms (Fig. 1B). The proposed approach provided an efficient way to screen students with RDs on a wider scale such that they could receive early intervention when necessary. The key contributions of this study are summarized as follows:

First, the study focused on the relationship between eye-tracking measurements and the cognitive processes underlying reading. Whereas previous eye-tracking-based methods for diagnosing RDs primarily focused on dyslexia [17,18], this study extended the research to include students with abnormalities in higher-level processes such as comprehension. By studying eye movements during naturalistic reading, this study aimed to identify the specific patterns associated with comprehension difficulties. Therefore, this study has the potential to contribute to a more comprehensive understanding of RDs and to improve personalized interventions and support for students facing diverse reading difficulties.

Second, this study used machine learning models to establish objective criteria for distinguishing the eye movements of students with RDs from those of normal students. Unlike previous research that focused on the reader's gaze on a few selected words [8-10], the proposed approach involved analyzing eye movements on all words in a given text, thereby providing a comprehensive response to the material. This holistic analysis minimized the evaluator input and potential bias during diagnosis, leading to a more efficient screening process for larger groups of students. Consequently, a combination of eye tracking and machine learning offers a practical and effective method for early identification and targeted support for struggling readers.

Third, this study strategically targeted specific age groups for RD screening and proposed an automated screening methodology using eye tracking. The third and fourth grades play a crucial role in a student's education, representing a pivotal juncture for developing essential comprehension skills that should ideally be mastered before the fourth grade [1]. These skills include understanding not only literal meaning but also context, and making inferences. Comprehension difficulties during this phase have been linked to later academic underperformance across subjects [2]. Thus, recognizing and addressing comprehension challenges in these grades is vital to establish a strong foundation for future academic success. Therefore, this study focused on identifying and screening students with RDs at this stage, with the aim of providing timely interventions and support to ensure their educational progress.

The remainder of this paper is organized as follows: Section 2 describes the data collection, diagnosis of RDs using standardized tests, and machine learning-based diagnosis of RDs. In Section 3, the accuracies of the machine learning-based diagnoses are compared, leading to their interpretation and discussion in Section 4. Finally, we conclude this study with future research directions in Section 5.

II. METHODS

A. Participants Information

Seventy third and fourth-grade students from 14 public elementary schools located in a Metropolitan City of South Korea participated in this study. Each school had enrolled approximately 500 students. Within each grade level, classes were divided into four or five sections, with each class contained 20-24 students. None of the participants received any financial or social assistance.

Among them, 27 (39%) were diagnosed with RD based on the outcomes of the two sets of standardized tests. Specifically, students who scored below the 16th percentile on both the district-wide reading assessments and the standardized reading assessment battery tests of reading achievement and reading cognitive process ability (RA-RCP) [19] were identified as having RD.

A comprehensive written consent form was provided to all participants and their parents. These forms outlined the objectives of the research, extent of participation, potential benefits and risks, assurances of confidentiality, and the right of the participants to withdraw from the study at their discretion.

B. Data Collection

The stimulus text was adopted from a teacher's guidebook for reading courses. The text included 12 sentences comprising 58 words. The text was presented in nine lines on a 24-inch liquid crystal display (LCD) screen with a resolution of 1920×1080 pixels. The size of the characters on the screen was approximately 32×32 pixels.

The eye movements of the participant were recorded as sequences of (x, y) coordinates on the screen using a Tobii Pro Spectrum eye tracker with a sampling rate of 150 Hz and a precision of 0.01°. For each participant, calibration was first performed and a practice session was provided, where instructions were provided on the screen and the participant was requested to press any button to proceed to the main session. In the main session, the stimulus text was presented on the screen and participants read the text at their own pace.

C. Preprocessing

The recorded eye movements were preprocessed as follows. First, calibration errors and drifts in gaze trajectories were reduced using the Eyekit package [20]. Next, the first fixation duration for each word was calculated. Finally, the features were normalized to zero mean and unit variance. Therefore, a 58-dimensional feature was generated for each subject.

Table 1. Comparison of chosen classifiers

Classifier	Key property and assumptions
Logistic regression	linearly separable
LDA	normal distributions with the same covariance
QDA	normal distributions
KNN	local similarity of features with the same class
Naïve Bayes	independent input features
SVM	maximum margin
Random forest	ensemble of randomly chosen decision trees

D. Classifiers to Diagnose RD

Seven machine learning models were used to classify the features (the first fixation durations on words) into normal and RD classes. The key properties and assumptions of the seven classifiers are summarized in Table 1.

The first four classifiers were chosen to investigate the effects of the assumptions regarding the distributions of the input features on the classification accuracy. The first classifier was logistic regression [21] that was simple and effective for linearly separating input features into two groups. In particular, the log odds between two classes were related to a linear function of the input features, resulting in a linear decision boundary between the two classes. The second and third classifiers were linear discriminant analysis (LDA) [22] and quadratic discriminant analysis (QDA) [23]. Both models assumed that the input features were normally distributed with different means for different classes. The covariances of the classes were assumed to be the same for LDA and allowed to differ for QDA, resulting in linear and curved decision boundaries for LDA and QDA, respectively. The fourth model was the k-nearest neighbors (KNN) classifier [24]. The KNN could capture more complex and local structures in the data because the class label of the input sample was determined by the class labels of the KNNs.

Three additional models were used to investigate the effects of potential correlations between the input features. The Naïve Bayes classifier [25] modeled the input features independently of each other, providing a baseline accuracy for ignoring the interactions between input features. Next, a support vector machine (SVM) with a linear kernel [26] was used to determine the hyperplane that separated the input features with different classes with the maximum margin. Finally, a random forest [27] was used to capture complex and potentially nonlinear patterns in the data using an ensemble of randomly chosen decision trees.

E. Cross-validation

The accuracies of the chosen classifiers were measured using leave-one-out cross-validation (LOOCV) [28]. LOOCV provided a more accurate evaluation than the other cross-validation methods because the dataset was relatively small. Specifically, the accuracy was measured as follows: Each classifier was trained using all features except for one sample and tested on the held-out sample. This was repeated for all the samples, and the average classification accuracy was measured.

F. Hyperparameter Tuning

The hyperparameters of the KNN, SVM, and random forest classifiers were varied for a wide range of values, and the highest accuracy was reported for each classifier.

In the KNN classifier, the number of neighbors (k) controlled the smoothness of the decision boundary. The value of k was increased from 1 to 15 in steps of 2.

In the SVM, the level of robustness to outliers was controlled by the weight (C) of the penalty for misclassification during training. The value of C was varied from 10^{-5} to 10^{5} by the factor of $\sqrt{10}$.

The complexity level of the random forest was controlled by the number of estimators (n). Increasing the number of estimators enabled the classifier to consider more complex data patterns. However, excessive estimators would result in overfitting and poor generalization to new inputs. The value of n was increased from 5 to 50 in steps of 5.

III. RESULTS

A. Correlation Analysis

The correlation between the input features was slightly positive, with an average correlation coefficients of 0.11. The histogram of the correlation coefficients (Fig. 2, left) showed that more features were positively correlated. In the correlation matrix (Fig. 2, right), consecutive features tended to have correlations with the same sign, indicating that the first fixation on neighboring words tended to be correlated.

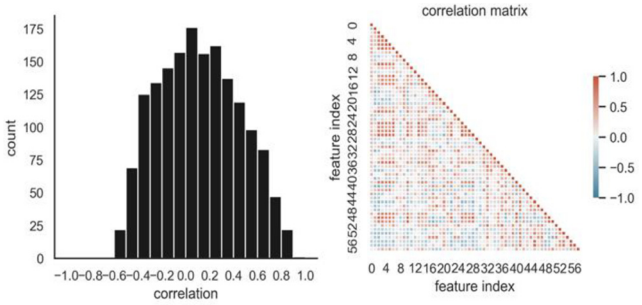

Fig. 2. Histogram (left) and correlation matrix (right) of the correlation coefficients between the features on different word pairs.

B. Accuracies of the Classifiers

Fig. 3 shows the classification accuracies of the seven classifiers. The classification accuracy of the logistic regression was 0.66 that was the baseline score achieved using a simple linear model. The accuracy of LDA (0.54) was lower than that of logistic regression, and QDA showed an even lower accuracy of 0.49. In contrast, KNN, which is the most flexible of the four classifiers, showed the highest accuracy (0.73) of the first four classifiers. The Naïve Bayes classifier exhibited a slightly higher accuracy of 0.76 than the KNN classifier. The SVM and random forest showed the highest accuracy of 0.8.

By simulating the model performance on unseen data when using the entire dataset, LOOCV provided a robust measure of the extent to which the model generalized to new instances, rendering it particularly valuable for smaller datasets in which the impact of each instance was significant.

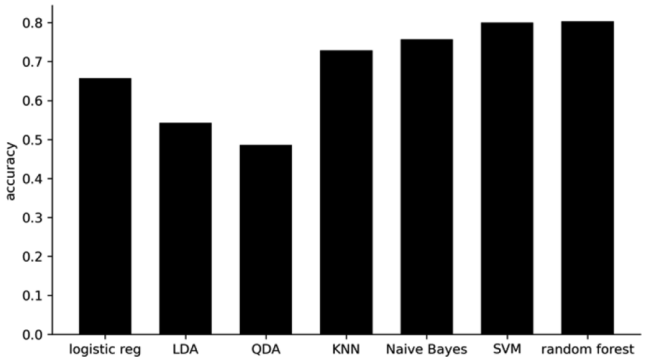

Fig. 3. Classification accuracies of RD using different classifiers.

C. Analysis of the Classifier Coefficients

Fig. 4 shows the absolute values of the coefficients of the SVM classifier that achieved the highest classification accuracy (red) and fixation durations averaged across the participants (blue). Notably, words that corresponded to large absolute values of the SVM coefficients existed (peaks of the red curve in Fig. 4), indicating that these words were more important for classifying students with RDs from normal students. These words tended to be fixated longer than neighboring words, as indicated by the co-located local peaks in the average fixation durations (peaks of the blue curve in Fig. 4).

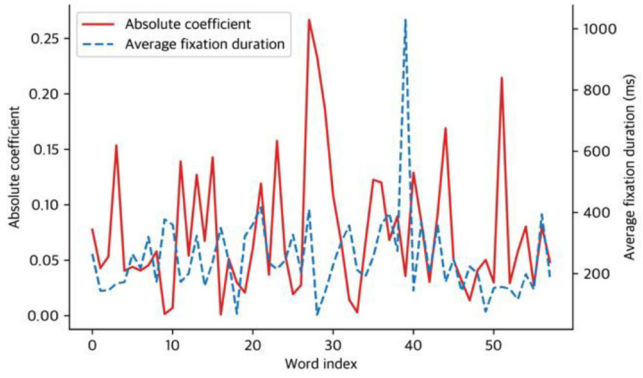

Fig. 4. Absolute values of the best SVM classifier coefficients (red) and average fixation durations (blue).

IV. DISCUSSION

A comparison of the accuracies of the first four classifiers indicated that the distribution of the first fixation durations was more complex than a normal distribution and was not easily described by simple parametric models. The logistic regression and LDA classifiers are linear models; however, the accuracy of the LDA classifiers was lower than that of the logistic regression, demonstrating that a multivariate Gaussian distribution, which is used for LDA, was not suitable to fit the first fixation durations on words. The lower accuracy of QDA compared with that of LDA corroborated this argument. The complex patterns in the first fixation were better captured by a more flexible model such as KNN.

The higher accuracies of the remaining three classifiers indicated that complex interactions between the first fixation durations were necessary to classify normal and RD students based on their eye movements. The Naïve Bayes classifier was designed to efficiently fit high-dimensional data with the assumption of independence between features. Consequently, the Naïve Bayes classifier achieved higher classification accuracy than the simpler classifiers. However, the assumption of independence contradicted the positive correlations observed in the input features. This led to a suboptimal performance, resulting in a suboptimal classification accuracy. The higher accuracies of SVM and random forest compared with those of Naïve Bayes classifiers demonstrated the importance of careful modelling of the correlated structure in the input feature.

The significance of achieving a classification accuracy of 0.8 was twofold. First, the accuracy was determined by comparison with judgments made by human experts. Thus, an accuracy of 0.8 implied that the automated diagnosis facilitated by the SVM or random forest classifiers was in agreement with that of the human expert with an 80% probability. Second, a cross-validation technique of LOOCV was used to measure the classification accuracy. This method ensured that the calculated classification accuracy served as a reliable estimate of the classification accuracy of data that had not been encountered before. Thus, the high accuracy of 0.8 was expected to be particularly useful for screening new students.

V. CONCLUSIONS

In this study, students with RDs were diagnosed based on their eye movements during natural reading using machine learning algorithms. The third- and fourth-grade students were labeled as normal or RD students using the standardized tests and expert evaluations. The eye movements of the participants were then measured while reading and used to classify the participants into normal and RD classes.

A comparison of the accuracies of the classifiers with different properties led to two major findings. First, eye movements during natural reading contained complex patterns that were unsuitable for simple parametric models with few parameters. Second, this complex pattern was well fitted by flexible classifiers designed for high-dimensional problems, such as SVM and random forest.

The future research will focus on the scalability of the machine learning-based diagnosis of RD. This study automated the diagnosis of RD from eye movements during natural reading. Even though the sample size was relatively small, SVM and random forest were able to reliably identify RD, and the accuracy of these classifiers was close to that evaluated by human experts. This finding demonstrated the feasibility of RD screening on a larger scale. The automated RD diagnosis would be most beneficial for the third- and fourth-graders, who are in the critical period.

This study aimed to collect more eye movement data from a wider range of age groups for texts of different genres. Based on the understanding from existing studies, fixation duration is a reliable feature that is stable across different languages and reading materials. Therefore, the proposed method is expected to achieve similar accuracy in classifying RDs that should be validated using different text types in future studies. More data will provide insight into the effects of developmental and individual differences on reading. A wide range of individual differences and personal characteristics can exist even within a particular age group. These differences may influence the way in which students experience RDs and respond to interventions.

Another important direction is to explore other useful features for analyzing eye movements during reading. In addition to fixation duration, other eye movements such as saccades (rapid, involuntary eye movements) and regressions (backward eye movements) have the potential to provide a further understanding of the reading process. Combining these features to improve the classification accuracy is certainly an interesting direction for future work.

REFERENCES

[1] M. J. Snowling and C. Hulme, *The Science of Reading: A Handbook*, Oxford, UK: Blackwell Publishing, 2005.

[2] T. Shanahan and C. Shanahan, "Teaching disciplinary literacy to adolescents: Rethinking content-area literacy," *Harvard Educational Review*, vol. 78, no. 1, pp. 40-59, Apr. 2008. DOI: 10.17763/haer.78.1.v62444321p602101.

[3] National Center for Education Statistics, National assessment of educational progress (NAEP) 2019 reading assessments, 2019, [online] Available: https://nces.ed.gov/nationsreportcard/reading/.

[4] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, 5th ed., Washington, DC: American Psychiatric Publishing, 2013.

[5] C. Hulme and M. J. Snowling, *Developmental disorders of language*

learning and cognition. John Wiley & Sons, 2013.
[6] M. Traxler and M. A. Gernsbacher, *Handbook of psycholinguistics,* Burlington, MA: Elsevier, 2011.
[7] B. T. Carter and S. G. Luke, "Best practices in eye tracking research," *International Journal of Psychophysiology*, vol. 155, pp. 49-62, Sep. 2020. DOI: 10.1016/j.ijpsycho.2020.05.010.
[8] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin*, vol. 124, no. 3, pp. 372-422, 1998. DOI: 10.1037/0033-2909.124.3.372.
[9] K. Rayner, K. H. Chace, T. J. Slattery, and J. Ashby, "Eye movements as reflections of comprehension processes in reading," *Scientific Studies of Reading*, vol. 10 no. 3, pp. 241-255, Jul. 2006. DOI: 10.1207/s1532799xssr1003_3.
[10] B. J. Juhasz and K. Rayner, "Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 29, no. 6, pp. 1312-1318, 2003. DOI: 10.1037/0278-7393.29.6.1312.
[11] S. Nahatame, "Text readability and processing effort in second language reading: A computational and eye-tracking investigation," *Language Learning*, vol. 71, no. 4, pp. 1004-1043, Jul. 2021. DOI: 10.1111/lang.12455.
[12] D. Torres, W. R. Sena, H. A. Carmona, A. A. Moreira, H. A. Makse, and J. S. Andrade Jr., "Eye-tracking as a proxy for coherence and complexity of texts," *PLOS One*, vol. 16, no. 12, p. e0260236, Dec. 2021. DOI: 10.1371/journal.pone.0260236.
[13] M. Mak, M. Faber, and R. M. Willems, "Different kinds of simulation during literary reading: Insights from a combined fMRI and eye-tracking study," *Cortex,* vol. 162, pp. 115-135, May 2023. DOI: 10.1016/j.cortex.2023.01.014.
[14] E. De Simone, K. Moll, L. Feldmann, X. Schmalz, and E. Beyersmann, "The role of syllables and morphemes in silent reading: An eye-tracking study," *Quarterly Journal of Experimental Psychology*, vol. 76, no. 11, pp. 2493-2513, Mar. 2023. DOI: 10.1177/17470218231160638.
[15] N. Valliappan, N. Dai, E. Steinberg, J. He, K. Rogers, V. Ramachandran, and V. Navalpakkam, "Accelerating eye movement research via accurate and affordable smartphone eye tracking," *Nature Communications*, vol. 11, no. 4553, pp. 1-12, Sep. 2020. DOI: 10.1038/s41467-020-18360-5.
[16] C. Mills, J. Gregg, R. Bixler, and S. K. D'Mello, "Eye-mind reader: An intelligent reading interface that promotes long-term comprehension by detecting and responding to mind wandering," *Human–Computer Interaction*, vol. 36, no. 4, pp. 306-332, Jan. 2021. DOI: 10.1080/07370024.2020.1716762.
[17] M. N. Benfatto, G. Ö. Seimyr, J. Ygge, T. Pansell, A. Rydberg, and C. Jacobson, "Screening for dyslexia using eye tracking during reading," *PLOS One*, vol. 11 no. 12, p. e0165508, Dec. 2016. DOI: 10.1371/journal.pone.0165508.
[18] B. Nerušil, J. Polec, J. Škunda, and J. Kačur, "Eye tracking based dyslexia detection using a holistic approach," *Scientific Reports*, vol. 11, no. 1, pp. 1-10, Aug. 2021. DOI: 10.1038/s41598-021-95275-1.
[19] A. Kim, U. Kim, M. Hwang, and H. Yoo, *Test of reading achievement and reading cognitive processes ability (RA-RCP)*, Seoul: Hakjisa, 2014.
[20] Package eyekit, [Online] Available: https://jwcarr.github.io/eyekit/.
[21] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215-232, Jul. 1958. DOI: 10.1111/j.2517-6161.1958.tb00292.x.
[22] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179-188, Sep. 1936. DOI: 10.1111/j.1469-1809.1936.tb02137.x.
[23] T. W. Anderson, *An introduction to multivariate statistical analysis*, 3rd ed., Hoboken, New Jersey: John Wiley & Sons, 2003.
[24] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory,* vol. 13, no. 1, pp. 21-27, Jan. 1967. DOI: 10.1109/TIT.1967.1053964.
[25] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, pp. 103-130, Nov. 1997, [Online], Available: https://link.springer.com/article/10.1023/A:1007413511361.
[26] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, Sep. 1995. DOI: 10.1007/BF00994018.
[27] L. Breiman, "Random forest," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.
[28] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning*, 2nd ed., New York: Springer, 2009.

Yongseok Yoo

He received the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2014 (U.S.A.). He received M.S. and B.S. degrees in electrical engineering from Seoul National University (Korea) in 2005 and 2002, respectively. Currently, he is an Associate Professor with the School of Computer Science and Engineering, Soongsil University (Korea). He has worked for Electronics and Telecommunications Research Institute (2014–2015) and Samsung Advanced Institute of Technology (2005–2008).