International Journal of Advanced Smart Convergence Vol.12 No.2 17-24 (2023)
http://dx.doi.org/10.7236/IJASC.2023.12.2.17

IJASC 23-2-2

Prediction of Cognitive Ability Utilizing a Machine Learning approach based on Digital Therapeutics Log Data

Yeojin Kim[1], Jiseon Yang[2], Dohyoung Rim[3], Uran Oh[†]

[1]*Master Student, Department of Computer Science and Engineering, Ewha Womans University, Seoul, Korea*
[2] *AI Researcher, Department of Artificial Intelligence, Rowan, Korea*
[3]*Chief Technology Officer, Department of Artificial Intelligence, Rowan, Korea*
[†]*Corresponding Author, Professor, Department of Computer Science and Engineering, Ewha Womans University, Seoul, Korea*
[1]*rowenna03@ewhain.net,* [2]*jiseon.yang@rowan.kr,* [3]*dhrim@rowan.kr,* [†]*uran.oh@ewha.ac.kr*

Abstract

Given the surge in the elderly population, and increasing in dementia cases, there is a growing interest in digital therapies that facilitate steady remote treatment. However, in the cognitive assessment of digital therapies through clinical trials, the absence of log data as an essential evaluation factor is a significant issue. To address this, we propose a solution of utilizing weighted derived variables based on high-importance variables' accuracy in log data utilization as an indirect cognitive assessment factor for digital therapies. We have validated the effectiveness of this approach using machine learning techniques such as XGBoost, LGBM, and CatBoost. Thus, we suggest the use of log data as a rapid and indirect cognitive evaluation factor for digital therapy users.

Key words: *Digital Therapeutics, Machine Learning, Mild Cognitive Impairment*

1. INTRODUCTION

Due to the rapid aging of the domestic population, experts predict that the proportion of elderly people over the age of 65 will reach 24.3% in 2030 [1], expressing concerns about the super-aging society. In this context, digital therapeutic agents for elderly diseases are receiving great attention since they can receive treatments steadily at home without a direct visit to a hospital. Digital therapeutics (DTx) refers to software programs that are appropriate for preventing, managing, or treating medical measures alone or in combination with other treatments such as drugs or devices [2]. Among them, this paper focuses on DTx for improving Mild Cognitive Impairment (MCI). SUPERBRAIN is a DTx for MCI patients released by Rowan Inc, a software program that increases accessibility by using a tablet for cognitive training based on cognitive function improvement experiments performed in the FINGER study [3, 4]. Figure 1 shows SUPERBRAIN's treatment environment. According to the prior study, MCI DTx showed significant cognitive function improvement through active

Manuscript Received: March. 5, 2023 / Revised: March. 10, 2023 / Accepted: March. 13, 2023
Corresponding Author: uran.oh@ewha.ac.kr
Tel: +82-2-3277-6896, Fax: +82-2-3277-6896
Professor, Department of Computer Science and Engineering, Ewha Womans University, Seoul, Korea

multi-domain intervention such as exercise, diet, cognitive training, and prevention and management of vascular diseases in the existing dementia high-risk group.

However, in the process of assessment of cognitive function, doctors do not utilize the outcome of digital therapeutics log data that is derived while the user uses the DTx [5, 6, 7]. Currently, when assessing whether a patient's cognitive abilities have improved during the use of digital therapeutics, doctors rely on the conventional method of referring to the diagnostic criteria of DSM-5 to make judgments. However, they are not incorporating the log data of the digital therapeutic used during the patient's cognitive improvement process into the diagnostic assessment.

Our paper suggests a key performance indicator to help the evaluation of a patient’s cognitive ability. Prior studies have addressed the creation of a score by adherence to weight according to its importance [8, 9]. In this regard, we propose a CogScore which is a derived feature based on log data with weights assigned. The level feature was selected to impose weight based on the SHAP method (SHapley Additive exPlanations). Furthermore, we confirmed CogScore through a comparison of expectation performance done by machine learning techniques (XGBoost, LGBM, CatBoost). Therefore, this paper proposes CogScore as an indicator that can be indirectly used for cognitive function judgments in the app by assigning weights to significant features among the collected features using machine learning-based data analysis techniques, in addition to the diagnosis of physicians. In the future, it is necessary to contribute to the analysis of DTx log data by using accumulated log data from developing DTx agents as an auxiliary tool that can be used effectively for users' cognitive function judgments.

In this paper, we examine existing studies that researched predictions of diseases by using healthcare data. After that, we explain the whole process of our proposed machine learning approach, from data collection to the creation of a new variable. To this end, we compare the performance results by using evaluation metrics. The contributions of our work are (1) the first empirical analysis of the digital therapeutics user log data, and (2) the generation of CogScore that contributes to indirect cognitive ability judgments in medical assessment.

2. RELATED WORK

DTx log data is included in the domain of healthcare data. A large body of prior work has investigated effective data analysis on healthcare data especially that explores dementia prediction or detection. Mathotaarachchi et. al. [10] conducted research with ADNI-GO/2 datasets to predict the progression from MCI to Alzheimer’s disease. Herzog et. al. [11] also worked to diagnose Mild Cognitive Impairment and Alzheimer’s disease with machine-learning classification algorithms. In addition, Deepika et. al. [12] performed a dementia detection study using four machine learning algorithms (J48, Naive Bayes, Random Forest, Multilayer Perceptron) with brain MRI datasets from OASIS_Brain.org. Furthermore, Mathkunti et.al. [13] studied the diagnosis of dementia with Parkinson’s disease data using a support vector machine (SVM), K-nearest neighbor (KNN), and linear discriminant analysis (LDA). In addition, Zhu et.al [14] evolved a dementia diagnosis design, which classifies normal people, MCI, VMD (Very Mild Dementia), and dementia based on an ML classification mechanism utilizing 37 features questionnaires completed by 5,272 participants. These earlier works primarily focused on dementia datasets analyzed by machine learning algorithms. Besides the research with the dementia datasets, various studies have been performed with a focus on lifelogging datasets. Kim et. al. [15] used to sleep and walking data from wearable devices in middle-aged men for machine learning regression to analyze the relationship between heights and BMI (Body Mass Index). Also, Kim et.al. [16] inspected a sleep habit prediction method based on sleep data from wearable devices applying ensemble machine learning algorithms. Furthermore, Kwon et.al. [17] studied the prediction of healthcare app customer attrition using Recurrent Neural Network with lifelog data and text message data from the digital healthcare app, Noom. Next, Palbar et.al [18] worked to predict blood glucose levels using a multimodal lifelog dataset, NTCIR-14, collected from wearable cameras. This work implemented diverse prediction algorithms (Random Forest, SVM, XGBoost, and Elastic Net Regression). All of these works show that healthcare datasets are well investigated through a lot of machine learning analysis.

3. METHOD

3.1 Data Collection

This paper is based on the SUPERBRAIN, DTx for MCI patients, app log data collected from the study of cognitive impairment improvement through lifestyle intervention for the elderly from September 27, 2021, to February 8, 2023 [4]. Figure 1 shows SUPERBRAIN's treatment environment. In the previous study, data were assembled from nineteen different hospitals with multi-domain intervention programs for at-risk elderly people. The collected data consisted of 211 users aged 45 to 85 and 67,493 app log data. The data include sex, education, height, age, level of the game, brain area based on hand shape, brain area based on hand movement, brain area based on the game, big cognitive area, small cognitive area, main attribute, and score accuracy. Score accuracy is the percentage of questions answered correctly by users when playing the game, calculated by dividing the number of correctly answered questions by the total number of questions. The number of questions varies depending on the levels and the type of the game. Once a level is completed, the score accuracy of the user is stacked in the log. For instance, if a user plays a Sudoku game up to level four, then the score accuracy for each of the four levels will be stored in the log data.

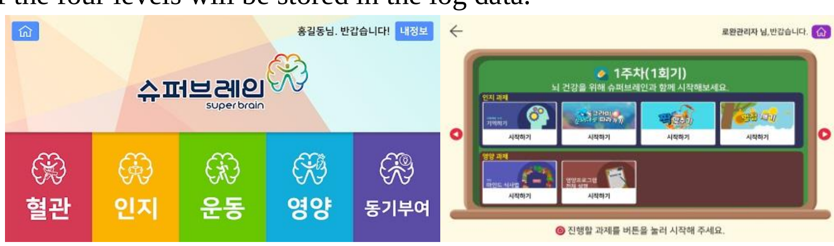

Figure 1. SuperBrain, MCI DTx by Rowan Inc

Table 1. Demographic characteristics of all participants

Characteristics		*All Participants (n=211)*
Sex	Male	76 (36%)
	Female	135 (64%)
Education		10.50 ± 4.41
Age		73.66 ± 5.91

3.2 Data Preprocessing

In the process of selecting variables to address the issue that the results of regression analysis may not be significant when there is a high correlation among independent variables, Variance Inflation Factor (VIF) stepwise feature selection method was chosen to calculate multicollinearity. Multicollinearity refers to cases where some independent variables can be expressed as a linear combination of other independent variables. The calculated VIF values are shown in Table 2. Based on the conventional analysis method that considers VIF values above 10 as indicating a significant multicollinearity problem, the big motor cortex, age, and height variables were removed. After that, the education, and score accuracy variables were normalized using min-max scaling, and then the sex, manipulation motor cortex, brain area, big cognitive domain, small cognitive domain, and main attribute were label encoded.

Table 2. Collected Variables and Descriptions

Variables	*Description*
Sex	User's sex
Education	User's duration of education
Height	User's height
Age	User's age
Level	The level of difficulty in DTx games
Big_motor_cortex	The shape of hand
Manipulation_motor_cortex	The manner of hand movement
Brain_area	The stimulation area of the brain
Big_cognitive_domain	Main cognitive domain of games
Small_cognitive_domain	Subsidiary cognitive domain of games
Main_attribute	Main type of games
Score_accuracy	Accuracy rate of game presented as score

Table 3. Variable Inflation Factor results of Independent variables

Variables	*Variance Inflation Factor*
Manipulation_motor_cortex	1.809228
Sex	1.852742
Brain_area	2.134802
Main_attribute	2.532377
Level	3.372074
Big_cognitive_domain	4.445988
Small_cognitive_domain	5.555214
Education	8.453451
Big_motor_cortex	**19.374587**
Age	**19.745817**
Height	**27.682645**

3.3 Modeling

The variables obtained through data preprocessing were trained using three machine-learning regression models (XGBoost Regression, CatBoost Regression, and LightGBM Regression). The obtained data were applied 5-fold cross-validation to find an optimal condition of the model. The train and test datasets were divided into an 8:2 ratio. XGBoost is a gradient-boosting machine that successfully handles instance weights in tree learning, applying a weighted quantile sketch procedure [19]. LightGBM is a subset of the Gradient Boosting Machine and it improves the learning process speed by using Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [20]. CatBoost (Categorical Boosting) is an ensemble algorithm that applies gradient boosting to efficiently deal with categorical features using categorical feature combinations [21]

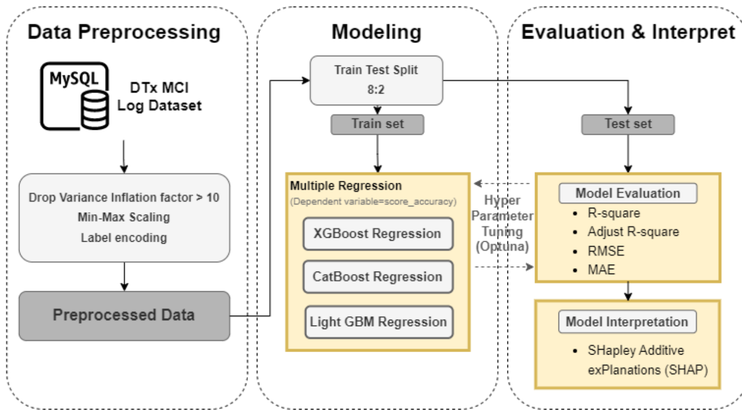

Figure 2. Flowchart for Mild Cognitive Data modeling

Figure 2 shows the overall flow of data preprocessing, modeling, and evaluation [22]. We used Optuna which is an open source python library introduced by Akiba et al. in 2019 to find the optimal hyperparameter values [23]. The hyperparameter values for each model obtained with the Optuna software framework are shown in Table 3. The machine learning algorithms were trained using these hyperparameter values.

3.4 Creation of CogScore, A New Derived Variable for Cognitive Ability

SHAP values are based on a unification of ideas from game theory and local explanations [22]. After training the data, the feature importance of the machine learning algorithm was evaluated using the SHAP method. Feature importance results of three machine learning models show that level, and brain area features are more important than other features. Considering that the data has many categorical variables, we focus on the feature importance results of CatBoost. Figure 2 shows the feature importance result of CatBoost using the SHAP method.

Table 4. Hyperparameters for XGBoost, Light GBM, CatBoost

XGBoost		*Light GBM*	
num_leaves	3	num_leaves	43
colsample_bytree	0.9486043979	colsample_bytree	0.8951191546
reg_alpha	0.04689631939	reg_alpha	0.6010389534
reg_lambda	6.262871483	reg_lambda	8.052231968
max_depth	10	max_depth	9
learning_rate	0.1877465453	learning_rate	0.00283069635
n_estimators	677	n_estimators	1026
min_child_weight	13	min_child_samples	13
subsample	0.5520731753	subsamples	0.5268907503
CatBoost			
max_depth	10		
min_child_samples	6		
n_estimators	354		
learning_rate	0.1276656705		

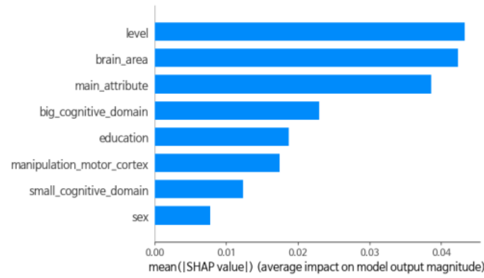

Figure 3. Feature Importance of Shapley Additive exPlanations Method

This study created a newly derived variable called CogScore by applying level-specific weights to the score accuracy. SUPERNBRAIN DEX (Dementia eXit), the latest version of dementia DTx by Rowan Inc, distributed game levels to three groups (easy, basic, advanced). In the DEX application, easy groups contain levels 1 and 2, basic groups contain levels 3 and 4, and advanced groups contain other levels. According to this level distribution of the DEX app, we divided the level into three groups (easy, basic, and advanced) and assigned different weights, called LevelWeight. The reason we divided into three groups is that the distribution of the level was clustered into each group which led to efficiency in the grouping. The easy group's weights are assigned 0.15, the basic group's weights are 0.35, and the advanced group's weights are assigned 0.5. We made a new variable CogScore applying LevelWeight to score accuracy variables.

$$CogScore = ScoreAccuracy \times LevelWeight$$
$$LevelWeight = \{0.15,\ if\ Level = 1, 2\ 0.35,\ if\ Level = 3, 4\ 0.5,\ Otherwise$$

4. RESULTS AND DISCUSSION

After preprocessing the data, regression analysis was performed on the score accuracy, which is a dependent variable using XGBoost, Light GBM, and CatBoost models. To evaluate the performance, the metrics R-square, Adjusted R-square, MAE, and RMSE were used, and the results are shown in Table 5 in the section "Without CogScore". After introducing CogScore, regression analysis was performed using the same models as before, and improved results were observed in all four metrics, which are shown in the "With CogScore" section of Table 5. The performance of predicting CogScore improved by about two times compared to the previous performance in all metrics, demonstrating the potential utility of CogScore. CogScore is an indirect cognitive ability indicator, based on the Feature Importance results of the SHAP method to analyze the impact of level on score accuracy. Prior studies have addressed the creation of scores that predict the disease well. Clark et al. [8] adhered to the weight according to sex and created a chronic disease score. Boldrini et al. [9] also introduced echocardiography scores for the efficient diagnosis of cardiac amyloidosis. All these works have focused on introducing a new score system according to the weight and features based on the data as well as our research.

Furthermore, as an aspect of computational efficiency, Light GBM is the fastest model, and XGBoost is the slowest model. Since Light GBM is a machine learning algorithm that enhances learning speed, it is obvious that Light GBM's computational efficiency is better than XGBoost.

Additionally, when considering the overall results, CatBoost showed the highest performance among the three models, which can be attributed to its high performance with categorical data, as seven out of the eight columns used in this study were categorical. According to prior studies, CatBoost is empirically known as the algorithm that has high performance in categorical datasets [25, 26, 27]. The findings suggest that our work is significantly related to the prior work.

Although this study added weight to score accuracy only for the level with the highest feature importance, future work may add various features such as response time and age from user log data as weights. If cognitive ability is measured not only by the scores assessed by physicians but also by scoring user log data from digital therapeutic use, it can be used as a complementary, indirect, and fast indicator for assessing cognitive ability.

Table 5. Performance Comparison of XGBoost, Light GBM, CatBoost

	Model	*R-square*	*Adj. R-square*	*MAE*	*RMSE*	*Time (s)*
Without CogScore	XGBoost	0.4762	0.4758	0.1258	0.1839	117.6
	Light GBM	0.4101	0.4097	0.1389	0.1951	18.6
	CatBoost	0.4596	0.4593	0.1293	0.1868	80.09
With CogScore	XGBoost	**0.864**	**0.864**	**0.0673**	**0.1041**	**115.4**
	Light GBM	**0.8266**	**0.8265**	**0.0788**	**0.1175**	**18.6**
	CatBoost	**0.8581**	**0.858**	**0.0697**	**0.1063**	**71.8**

5. CONCLUSION

This paper contributes to the very timely discussion of DTx log data and is the first to analyze user log data from the Korean MCI DTx. We propose a derived variable called CogScore that can be used as a supplementary tool for assessing cognitive ability when using digital therapeutics for MCI based on score accuracy and level-specific weights. Notably, the effectiveness of the proposed variable was confirmed through machine learning algorithms. In our study, variables with high multicollinearity were removed using VIF, and categorical variables were encoded using label encoding while numerical variables were normalized using min-max scaling for preprocessing. The performance of the machine learning models XGBoost, Light GBM, and CatBoost was compared based on the application of CogScore, which was created by assigning weights to level-specific score accuracy. The high performance of CogScore was demonstrated, and based on these results, it is suggested that not only physician diagnoses but also the analysis of DTx log data can be used as an efficient complementary factor in assessing cognitive ability in MCI users of DTx.

ACKNOWLEDGEMENT

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2023-00155966, Artificial Intelligence Convergence Innovation Human Resources Development (Ewha Womans University)), and received research support from Rowan in 2023.

REFERENCES

[1] Korea Institute for Health and Social Affairs, Health and Welfare Policy Issues and Policy Tasks, pp.97,2014

[2] Hong, J. S., Wasden, C., & Han, D. H. "Introduction of digital therapeutics," *Computer Methods and Programs in Biomedicine*, 209, 106319, 2021.

[3] Park, H. K., Jeong, J. H., Moon, S. Y., Park, Y. K., Hong, C. H., Na, H. R., ... & Choi, S. H. "South Korean study to prevent cognitive impairment and protect brain health through lifestyle intervention in at-risk elderly people: Protocol of a multicenter, randomized controlled feasibility trial," *Journal of Clinical Neurology (Seoul, Korea)*, 16(2), 292, 2020.

[4] Moon, S. Y., Hong, C. H., Jeong, J. H., Park, Y. K., Na, H. R., Song, H. S., ... & Choi, S. H., "Facility-based and home-based multidomain interventions including cognitive training, exercise, diet, vascular risk management, and motivation for older adults: a randomized controlled feasibility trial." *Aging (Albany NY),* 13(12), 15898, 2021.

[5] Park, S. Y., Kim, D., Kim, B. J., Ju, K. W., Kim, Y. K., Park, J. Y., & Lee, J. H, "Digital Therapeutics (DTx) for Temporomandibular Disorder (TMD)", *Archives of Clinical and Medical Case Reports*, 6(6), 725-729, 2022.

[6] de Oliveira, C. M., Bolognese, L. B., Balcells, M., Aragon, D. C., Zagury, R. L., Nobrega, C., & Liu, C., "A data-

driven approach to manage type 2 diabetes mellitus through digital health: The Klivo Intervention Program protocol (KIPDM)," *Plos one*, 18(2), e0281844, 2023.

[7] Moon, S. Y., Shin, S. A., Jeong, J. H., Hong, C. H., Park, Y. K., Na, H. R., ... & Choi, S. H., "Impact of a multidomain lifestyle intervention on regional spontaneous brain activity," *Frontiers in Aging Neuroscience*, 841, 2022.

[8] Clark, D. O., Von Korff, M., Saunders, K., Baluch, W. M., & Simon, G. E., "A chronic disease score with empirically derived weights," *Medical care*, 33(8), 783-795, 1995.

[9] Boldrini, M., Cappelli, F., Chacko, L., Restrepo-Cordoba, M. A., Lopez-Sainz, A., Giannoni, A., ... & Fontana, M., "Multiparametric echocardiography scores for the diagnosis of cardiac amyloidosis," *Cardiovascular Imaging*, 13(4), 909-920, 2020.

[10] Mathotaarachchi, S., Pascoal, T. A., Shin, M., Benedet, A. L., Kang, M. S., Beaudry, T., ... & Alzheimer's Disease Neuroimaging Initiative, "Identifying incipient dementia individuals using machine learning and amyloid imaging," *Neurobiology of aging*, 59, 80-90, 2017.

[11] Herzog, N. J., & Magoulas, G. D., "Brain asymmetry detection and machine learning classification for diagnosis of early dementia", *Sensors*, 21(3), 778, 2021.

[12] Bansal, D., Chhikara, R., Khanna, K., & Gupta, P., "Comparative analysis of various machine learning algorithms for detecting dementia," *Procedia computer science*, 132, 1497-1502, 2018.

[13] Mathkunti, N. M., & Rangaswamy, S., "Machine learning techniques to identify dementia," *SN Computer Science*, 1(3), 118, 2020.

[14] Zhu, F., Li, X., Tang, H., He, Z., Zhang, C., Hung, G. U., ... & Zhou, W., "Machine learning for the preliminary diagnosis of dementia". *Scientific Programming*, 1-10., 2020.

[15] Jiyong Kim, Jisoo Lee, Minseo Park, "Analysis of Lifelong for Health of Middle-Aged Men by Using Machine Learning Algorithm", *Journal of the Korean Institute of Industrial Engineers*,47(6),504-513, 2021.

[16] Kim, J., & Park, M., "A Study on ML-Based Sleep Score Model Using Lifelog Data", *Applied Sciences*, 13(2), 1043, 2023.

[17] Kwon, H., Kim, H. H., An, J., Lee, J. H., & Park, Y. R., "Lifelog data-based prediction model of digital health care app customer churn: retrospective observational study," *Journal of Medical Internet Research*, 23(1), e22184, 2021.

[18] Palbar, T., Kesavulu, M., Gurrin, C., & Verbruggen, R., "Prediction of Blood Glucose Using Contextual LifeLog Data," *In Proceedings of MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam*, June 6–10, 2022, Proceedings, Part I (pp. 404-415). Cham: Springer International Publishing, 2022.

[19] Chen, T., & Guestrin, C., "Xgboost: A scalable tree boosting system," *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794), 2016.

[20] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y., "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, 30, 2017.

[21] Dorogush, A. V., Ershov, V., & Gulin, A., "CatBoost: gradient boosting with categorical features support," *arXiv* preprint arXiv:1810.11363, 2018.

[22] Oh, H. R., Son, A. L., & Lee, Z., "Occupational accident prediction modeling and analysis using SHAP," *Journal of Digital Contents Society*, 22(7), 1115-1123, 2021.

[23] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M., "Optuna: A next-generation hyperparameter optimization framework," *In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631), 2019.

[24] Lundberg, S. M., & Lee, S. I., "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, 30, 2017

[25] Jhaveri, S., Khedkar, I., Kantharia, Y., & Jaswal, S., "Success prediction using random forest, catboost, xgboost and adaboost for kickstarter campaigns," *In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1170-1173). IEEE, 2019.

[26] Hussain, S., Mustafa, M. W., Jumani, T. A., Baloch, S. K., Alotaibi, H., Khan, I., & Khan, A., "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," *Energy Reports*, 7, 4425-4436, 2021.

[27] Zhang, Y., Ma, J., Liang, S., Li, X., & Li, M., "An evaluation of eight machine learning regression algorithms for forest aboveground biomass estimation from multiple satellite data products," *Remote Sensing*, 12(24), 4015, 2020.