Journal of Theoretical and Applied Information Technology
31st July 2021. Vol.99. No 14
© 2021 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DETECTING OLFACTORY IMPAIRMENT THROUGH OBJECTIVE DIAGNOSIS : CATBOOST CLASSIFIER ON EEG DATA

[1]MIN JONG CHEON, [*]OOK LEE

[1]*Department of Information Systems, Hanyang University, 222 Wangshimni-ro, Seongdong-gu, 04673 Seoul, South Korea*
[*]*Department of Information Systems, Hanyang University, 222 Wangshimni-ro, Seongdong-gu, 04673 Seoul, South Korea*

E-mail: [1]jmj2316@hanyang.ac.kr, [*]ooklee@hanyang.ac.kr

ABSTRACT

Detecting olfactory impairment using an objective diagnosis kit has been a challenge. Recently, machine learning and deep learning models have been used on EEG data with promising results. The goal of our study was to detect olfactory impairment through a machine learning classifier with EEG data. This was done by identifying the important EEG data factors affecting olfactory impairment. Finally, we compared our model to other machine learning and deep learning algorithms in order to identify possibilities for further research. Downsampling and extracting various waves from EEG data were conducted for data preprocessing. Then, an independent component analysis was performed to remove artifacts. Through this processing, a dataset in CSV format was obtained. Next, we built a CatBoost classifier model because it is recent boost model and has high performance for classification. It identified whether a subject had olfactory impairment or not. After training with the CatBoost algorithm, we compared it to different machine learning and deep learning algorithms. The CatBoost model showed 87.56 % accuracy, while other machine learning algorithms such as the random forest classifier, gradient boosting classifier, XG boosting classifier, k-nearest-neighbor classifier, decision tree classifier, Gaussian NB, and logistic regressor revealed 82.22 %, 78.89 %, 78.22 %, 75.78 %, 74 %, 69.78 %, and 41.11 % accuracy, respectively. With deep learning models, which consisted of bi-directional long short term memory, long short term memory and a deep neural network, the performance was 63.11 %, 51.33 %, and 60 %. The CatBoost model showed feature importance, which revealed that the gamma wave on the Cz channel was about 20, which was the highest among the other variables.

Keywords: *Artificial Intelligence, Machine Learning, Deep Learning, EEG, Olfactory Impairment, Diagnosis*

1. INTRODUCTION

1.1 Background

Olfactory impairment can be divided into normosmia, hyposmia, and anosmia. Normosmia is a subjectively perceived standard olfactory feature, typically defined as the ability to detect a large majority of odors tested in an olfactory test. Hyposmia results in diminished olfactory function, and anosmia is the loss of olfactory function [1]. However there is no diagnosis kit that can objectively and quantitatively determine impairment. Therefore, subjective olfactory test methods are used for diagnosis. The absence of a diagnosis kit can trigger various problems. First of all, in olfactory diagnosis, patient factors are significant, which means detecting malingering is difficult and affects the reliability of diagnosis. Secondly, a subjective diagnosis kit cannot distinguish between hyposmia and anosmia. Lastly, there is no kit for early detection of dementia, brain tumors, or Alzheimer's disease (AD). Electroencephalography (EEG) is a measurement of electrical activity in the human brain [2]. We decided to utilize EEG signals to diagnose olfactory impairment because EEG has shown promising results in detecting a number of disorders, Furthermore, in recent days, the market of medical devices has grown faster, as shown in Table1. Specifically, the technologies such as Deep Learning, Natural Language Processing have been developed rapidly[3].

1.2 Objectives

We applied machine learning and deep learning algorithms to EEG data for detecting and discriminating olfactory impairment. Therefore, the goal was to create an objective olfactory test method for diagnosis and hopefully solve the problems mentioned above. We want to come up with the novel method of detecting olfactory impairment and encourage the further research related to ours. Furthermore, as we show a possibility of EEG data for diagnosis, we believe that other data sources could also be used as input data. To the end, our paper shows a various possibility for upcoming researches. For discriminating olfactory impairment, we labelled the patient data in two categories, 1 for olfactory impairment and 0 for normosmia. Our research was structured in two stages. In the first stage, we utilized the CatBoost algorithm for discrimination and compared the result to other machine learning algorithms such as decision tree, logistic regression, k-nearest neighbors (KNN), naive Bayes, random forest, gradient boosting, XG boosting, and the light gradient boosting model (LGBM). The second stage compares deep learning models such as deep neural network(DNN), long short term memory(LSTM), and bi-directional long short term memory(Bi-LSTM) to the CatBoost model.

2. Related Works

There are various researches conducted for utilizing EEG data to discriminate diverse diseases or states such as sleep stage, odors, confused states and emotions. Those researches apply machine learning and deep learning methods in common. Jeon et al.[4] utilized multi -domain hybrid neural network(HNN-multi) consisting of a convolutional neural network(CNN) and bidirectional long short-term memory for discriminating three sleep stages. This research achieved F1 score of 92.21%. Li et al.[5] applied Support Vector Machines (SVM) and K- Nearest Neighbors (KNN) classifiers to the resting state EEG data and achieved accuracy of 98%. Zhang et al.[6] used K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Back Propagation Neural Network (BPNN) and Convolutional Neural Network (CNN) to classify five different odors and got accuracy of 82.2%. For classifying confused state, Ni et al.[7] suggested possibility of applying EEG data to confused state which achieved accuracy of 73.3%. Wang et al.[8] showed possibility of associating EEG data with emotional state by achieving average classification accuracy of 91.77%.

3. MATERIALS AND METHODS

3.1 Data Description

Mobilab+, brain wave measuring equipment, was used for the EEG data. Through this equipment, 4 types of channels were measured: Cz, Pz, P1, P2, and Fp2. The purpose was to remove artifacts by eye blink. Figure 1 provides the EEG channel locations. At this time, the air presented to the subjects was kept under the conditions of flow rate (8 L/min), temperature (38.5°C), and humidity (80 %). We used n-butanol (99.5 %) as the olfactory source. The total number of subjects involved in the experiment with the average age of the subjects being 24.4 years (19–38). There was only one person with olfactory impairment and forty seven subjects who did not have an impairment

3.2 Data Preprocessing

Firstly, down sampling to 256 Hz for each channel was conducted. We then extracted various types of brain waves, which are alpha (8–13 Hz), beta (14–30 Hz), theta (4–7 Hz), and gamma (30–47 Hz). Subsequently, we conducted an independent component analysis to remove eye blink artifacts and movements [9]. To the end, we got 4 types of brain waves from each channel through preprocessing.

Table 1. Market size of medical devices

	2016	2018	2020	2022	2023	CAGR
Hardware	191.63	441.89	999.87	2219.65	3283.41	49.7
Software	940.91	2107.34	4633.51	9999.99	14587.77	47.6
Service	308.25	696.04	1543.04	3357.77	4918.56	48.2
Total	1440.79	3245.28	7176.42	15577.40	22789.74	48.7

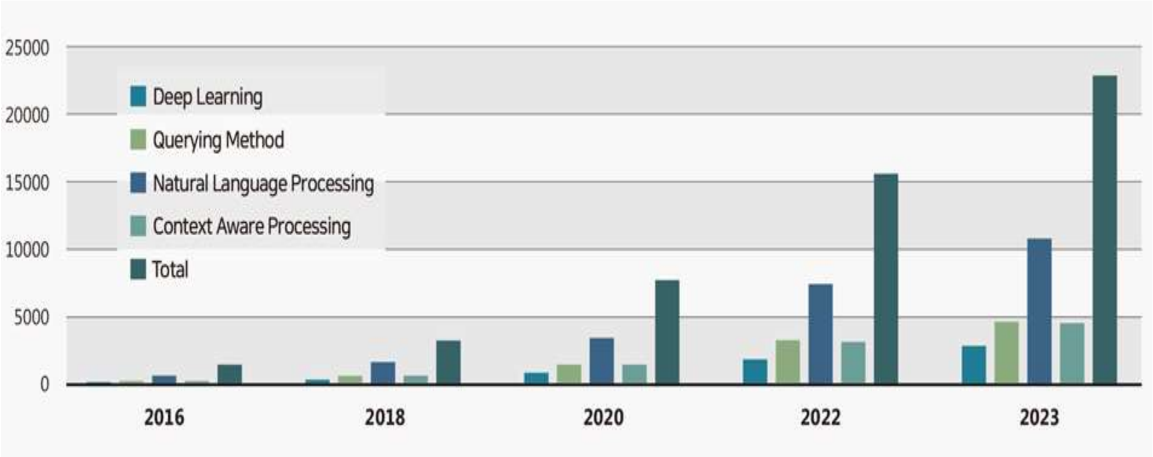


Figure 1. .Size of artificial intelligence – heath care market

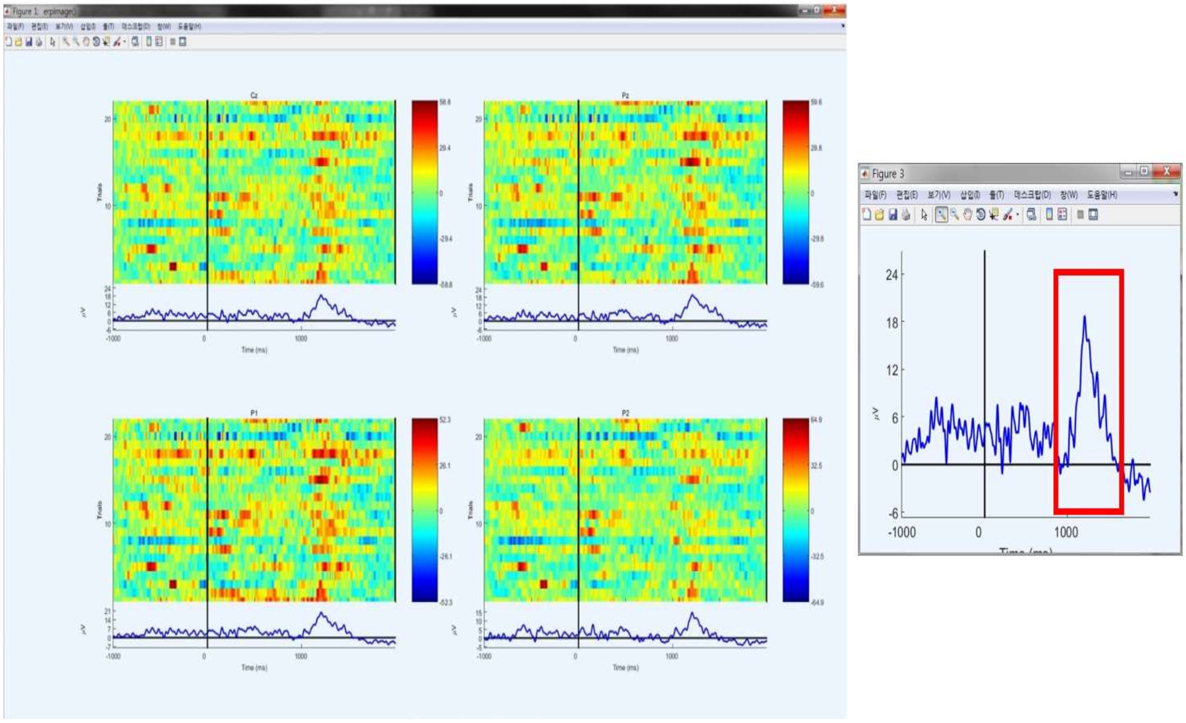


Figure 2. Result after preprocessing process

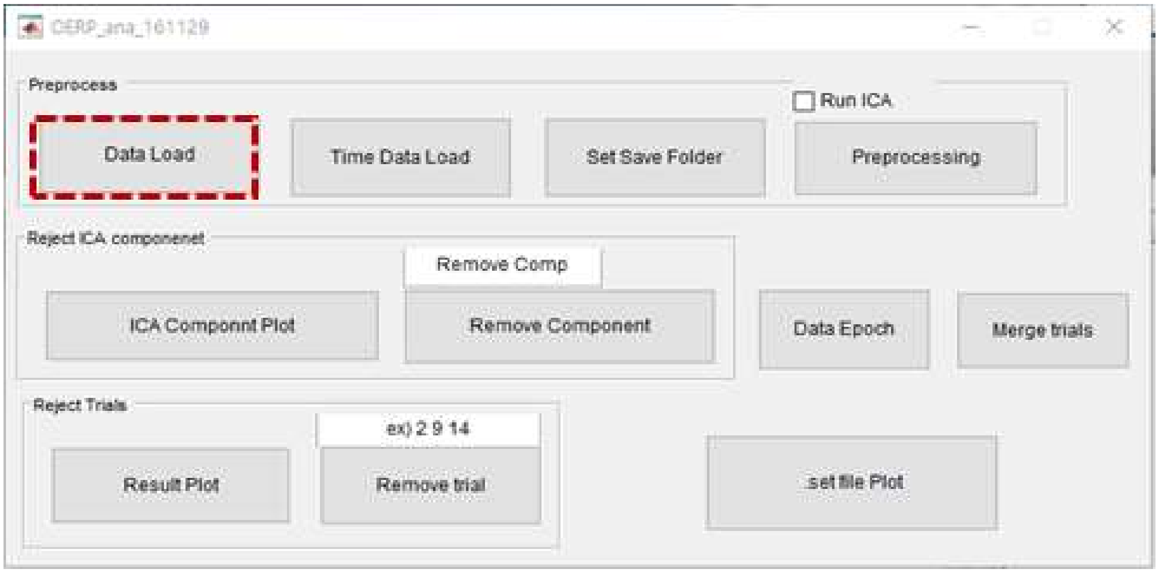


Figure 3. Preprocessing software based on matlab

Table 2. Types of brain wave

Types of brain wave	Hz
Theta	4-7 Hz
Alpha	8-15 Hz
Beta	16-31 Hz
Gamma	32-47 Hz

2.1 Boosting Algorithm

The boosting algorithm is an ensemble algorithm that improves the prediction through training of a sequence of weak models so that they can be converted to strong models. The boosting algorithm is a decision tree-based algorithm, and the decision tree cannot handle categorical variables directly. In other words, the general boosting technique requires the preprocessing of categorical variables. To this end, techniques such as one-hot encoding are used, but this is not efficient in terms of memory usage and speed. In addition, the boosting technique basically builds a model for learning residual errors, in turn learning the previous residual errors and predicting the results. As a result, this traditional boosting technique is vulnerable to overfitting [10]. Figure 2 shows the sequential process in boosting the algorithm.

2.2 Catboost Algorithm

The CatBoost algorithm is an ordered boosting algorithm that focuses on preprocessing categorical data and solving the overfitting problem. Unlike the original boosting algorithms that train every residual error in a sequence, the CatBoost algorithm only calculates the residual error that is left on certain data. Furthermore, by randomizing the data sequences through random permutation on ordered boosting, the CatBoost algorithm can prevent overfitting. For preprocessing categorical variables, the CatBoost algorithm calculates the sample mean values for variables in the same category from a dataset that has gone through random permutation.

$$\widehat{x_k^i} = \frac{\sum_{j=1}^{n}[x_j^i = x_k^j] \circ y_j + \alpha P}{\sum_{j=1}^{n}[x_j^i = x_k^i] + \alpha} \quad (1)$$

where α is the corresponding weight, P denotes the prior value, $x_k = (x_k^1, , , , , , , x_k^m)$ is the random vector of m features, and $y_k \in R$ denotes the corresponding label.

The CatBoost algorithm speeds up training through feature combinations that combine variables with the same information gain [11]. In addition, unlike other ensemble algorithms that use GridSearchcv or RandomizedSearchcv to find the optimal hyperparameters, the initial hyperparameter values are well optimized and do not need to go through parameter tuning procedures [12].

2.3 Train and Test Dataset

In this research, our final dataset contains columns that consist of four channels (P1, P2, Cz, Pz) and four types of brain waves for each channel (alpha, beta, theta, and gamma). Therefore, the number of total columns is 20. The number of rows in our dataset is 1500. With this dataset, 70 percent are used as training sets, and the remaining 30 percent as test sets.

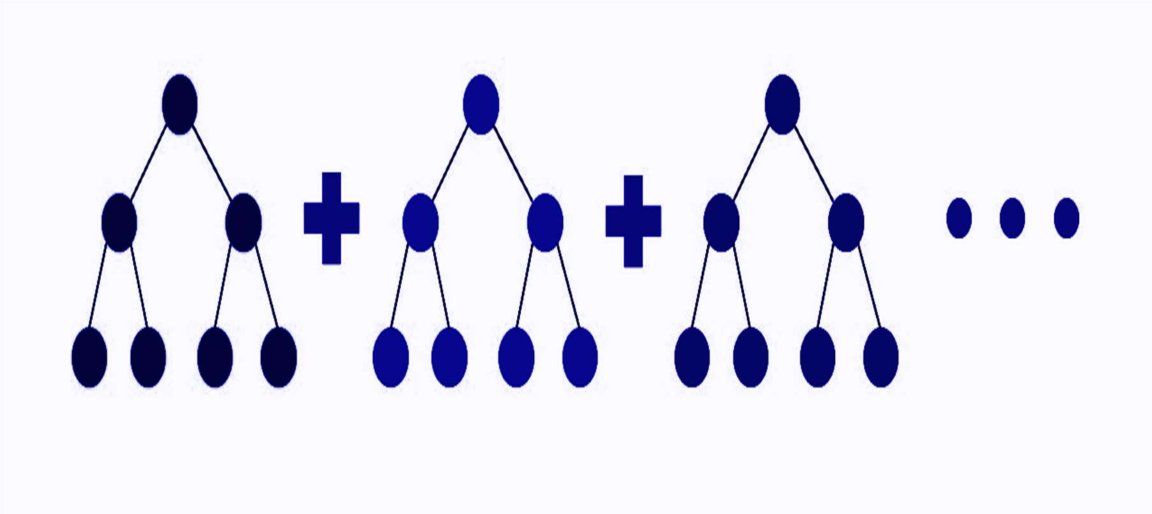
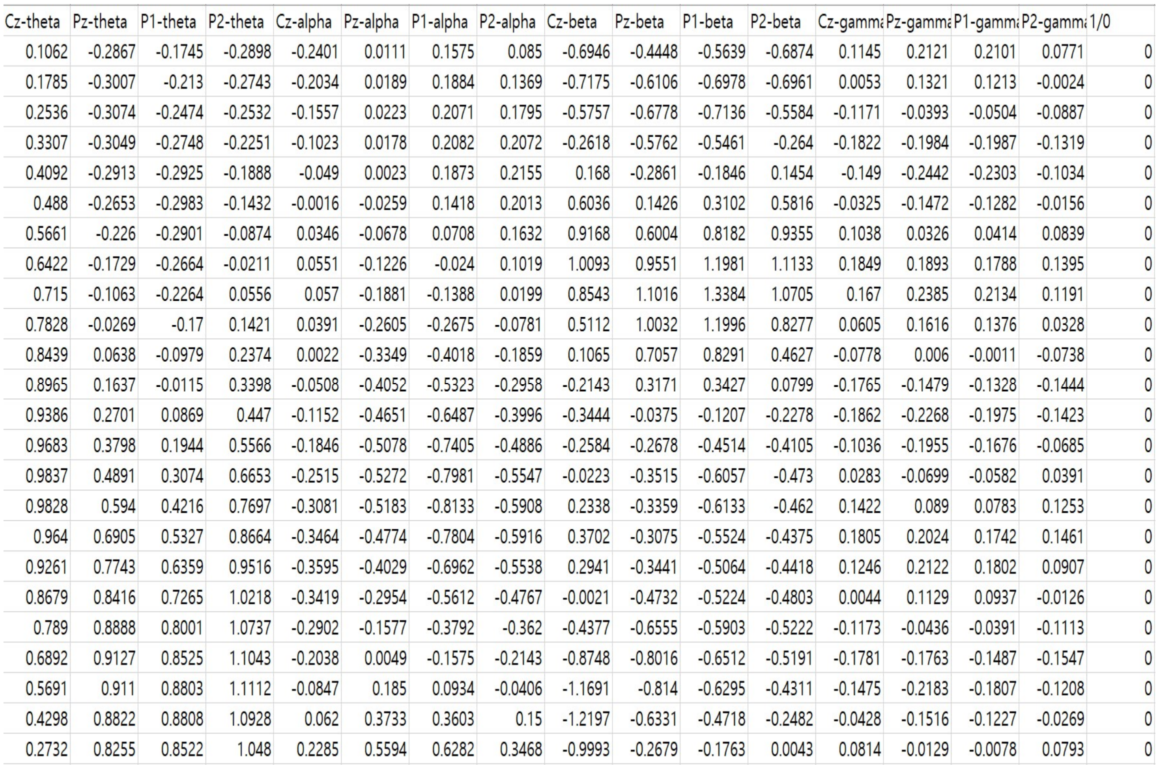


Figure 4. A sequential approach in boosting algorithm



Cz-theta	Pz-theta	P1-theta	P2-theta	Cz-alpha	Pz-alpha	P1-alpha	P2-alpha	Cz-beta	Pz-beta	P1-beta	P2-beta	Cz-gamma	Pz-gamma	P1-gamma	P2-gamma	1/0
0.1062	-0.2867	-0.1745	-0.2898	-0.2401	0.0111	0.1575	0.085	-0.6946	-0.4448	-0.5639	-0.6874	0.1145	0.2121	0.2101	0.0771	0
0.1785	-0.3007	-0.213	-0.2743	-0.2034	0.0189	0.1884	0.1369	-0.7175	-0.6106	-0.6978	-0.6961	0.0053	0.1321	0.1213	-0.0024	0
0.2536	-0.3074	-0.2474	-0.2532	-0.1557	0.0223	0.2071	0.1795	-0.5757	-0.6778	-0.7136	-0.5584	-0.1171	-0.0393	-0.0504	-0.0887	0
0.3307	-0.3049	-0.2748	-0.2251	-0.1023	0.0178	0.2082	0.2072	-0.2618	-0.5762	-0.5461	-0.264	-0.1822	-0.1984	-0.1987	-0.1319	0
0.4092	-0.2913	-0.2925	-0.1888	-0.049	0.0023	0.1873	0.2155	0.168	-0.2861	-0.1846	0.1454	-0.149	-0.2442	-0.2303	-0.1034	0
0.488	-0.2653	-0.2983	-0.1432	-0.0016	-0.0259	0.1418	0.2013	0.6036	0.1426	0.3102	0.5816	-0.0325	-0.1472	-0.1282	-0.0156	0
0.5661	-0.226	-0.2901	-0.0874	0.0346	-0.0678	0.0708	0.1632	0.9168	0.6004	0.8182	0.9355	0.1038	0.0326	0.0414	0.0839	0
0.6422	-0.1729	-0.2664	-0.0211	0.0551	-0.1226	-0.024	0.1019	1.0093	0.9551	1.1981	1.1133	0.1849	0.1893	0.1788	0.1395	0
0.715	-0.1063	-0.2264	0.0556	0.057	-0.1881	-0.1388	0.0199	0.8543	1.1016	1.3384	1.0705	0.167	0.2385	0.2134	0.1191	0
0.7828	-0.0269	-0.17	0.1421	0.0391	-0.2605	-0.2675	-0.0781	0.5112	1.0032	1.1996	0.8277	0.0605	0.1616	0.1376	0.0328	0
0.8439	0.0638	-0.0979	0.2374	0.0022	-0.3349	-0.4018	-0.1859	0.1065	0.7057	0.8291	0.4627	-0.0778	0.006	-0.0011	-0.0738	0
0.8965	0.1637	-0.0115	0.3398	-0.0508	-0.4052	-0.5323	-0.2958	-0.2143	0.3171	0.3427	0.0799	-0.1765	-0.1479	-0.1328	-0.1444	0
0.9386	0.2701	0.0869	0.447	-0.1152	-0.4651	-0.6487	-0.3996	-0.3444	-0.0375	-0.1207	-0.2278	-0.1862	-0.2268	-0.1975	-0.1423	0
0.9683	0.3798	0.1944	0.5566	-0.1846	-0.5078	-0.7405	-0.4886	-0.2584	-0.2678	-0.4514	-0.4105	-0.1036	-0.1955	-0.1676	-0.0685	0
0.9837	0.4891	0.3074	0.6653	-0.2515	-0.5272	-0.7981	-0.5547	-0.0223	-0.3515	-0.6057	-0.473	0.0283	-0.0699	-0.0582	0.0391	0
0.9828	0.594	0.4216	0.7697	-0.3081	-0.5183	-0.8133	-0.5908	0.2338	-0.3359	-0.6133	-0.462	0.1422	0.089	0.0783	0.1253	0
0.964	0.6905	0.5327	0.8664	-0.3464	-0.4774	-0.7804	-0.5916	0.3702	-0.3075	-0.5524	-0.4375	0.1805	0.2024	0.1742	0.1461	0
0.9261	0.7743	0.6359	0.9516	-0.3595	-0.4029	-0.6962	-0.5538	0.2941	-0.3441	-0.5064	-0.4418	0.1246	0.2122	0.1802	0.0907	0
0.8679	0.8416	0.7265	1.0218	-0.3419	-0.2954	-0.5612	-0.4767	-0.0021	-0.4732	-0.5224	-0.4803	0.0044	0.1129	0.0937	-0.0126	0
0.789	0.8888	0.8001	1.0737	-0.2902	-0.1577	-0.3792	-0.362	-0.4377	-0.6555	-0.5903	-0.5222	-0.1173	-0.0436	-0.0391	-0.1113	0
0.6892	0.9127	0.8525	1.1043	-0.2038	0.0049	-0.1575	-0.2143	-0.8748	-0.8016	-0.6512	-0.5191	-0.1781	-0.1763	-0.1487	-0.1547	0
0.5691	0.911	0.8803	1.1112	-0.0847	0.185	0.0934	-0.0406	-1.1691	-0.814	-0.6295	-0.4311	-0.1475	-0.2183	-0.1807	-0.1208	0
0.4298	0.8822	0.8808	1.0928	0.062	0.3733	0.3603	0.15	-1.2197	-0.6331	-0.4718	-0.2482	-0.0428	-0.1516	-0.1227	-0.0269	0
0.2732	0.8255	0.8522	1.048	0.2285	0.5594	0.6282	0.3468	-0.9993	-0.2679	-0.1763	0.0043	0.0814	-0.0129	-0.0078	0.0793	0

Figure 5. Final dataset with 4 types of channels and 4 types of brain waves

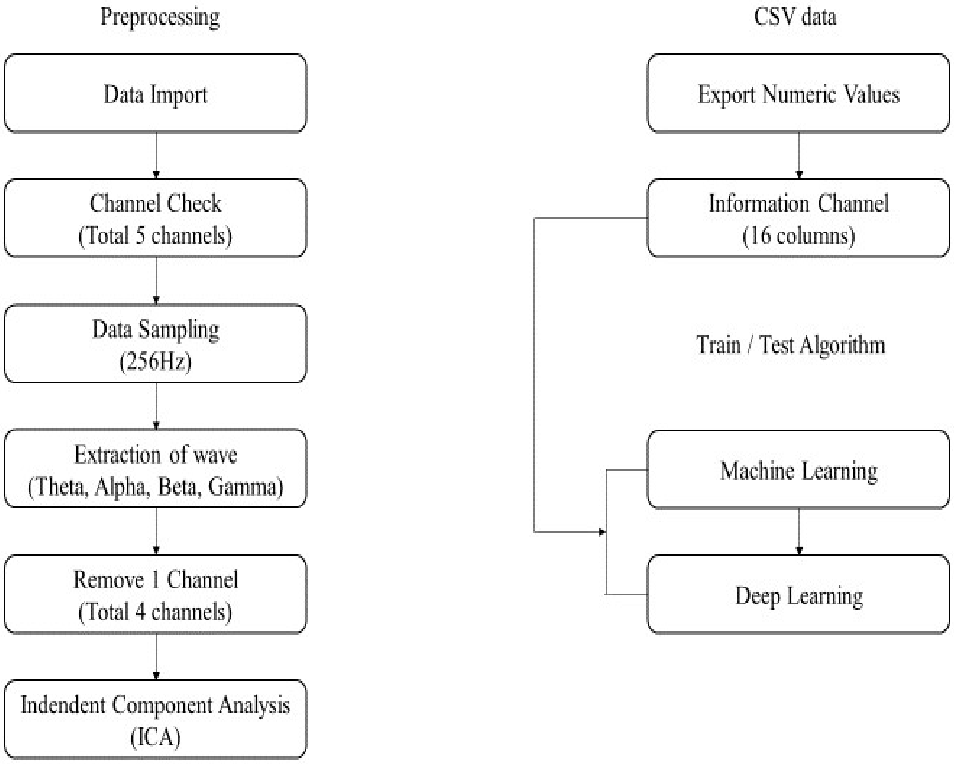

Preprocessing
Data Import
Channel Check (Total 5 channels)
Data Sampling (256Hz)
Extraction of wave (Theta, Alpha, Beta, Gamma)
Remove 1 Channel (Total 4 channels)
Indendent Component Analysis (ICA)
CSV data
Export Numeric Values
Information Channel (16 columns)
Train / Test Algorithm
Machine Learning
Deep Learning

Figure 6. Flowchart for CatBoost Model with EEG dataset

2.4 Catboost Pipeline

The CatBoost model can handle categorical variables by itself, and the basic parameters are optimized, so it goes through fewer steps than normal machine learning analyses. Other machine learning algorithms as well as boosting algorithms cannot handle categorical variables directly [13]. Therefore, it should be converted to numerical data through one-hot encoding in order to unify the range of all values from 0 to 1 through min-max normalization, such as expression(). Instead of trying all the possible hyperparameters, hyperparameter tuning through RandomizedSearchCV could be more efficient [12].

$$\mathbf{z} = \frac{\mathbf{x} - \mathbf{min(x)}}{\mathbf{max(x)} - \mathbf{min(x)}} \qquad (2)$$

2.5 Evaluation of Classifier Result

The performance of the classifier was measured by an accuracy score, which is the simple ratio of the correctly predicted observations to total observations. The accuracy score was calculated through scikit-learn.

3. Results

3.1 Catboost Results

The achievement of the classifier was measured by the accuracy score. The experiment was conducted along the CatBoost flowchart, which does not contain other procedures such as one-hot encoding or RandomizedSearchCV. The result was 87.56 %. The CatBoost algorithm visualizes the feature importance of each variable, and it showed that the gamma brain wave of the Cz channel achieved the highest score, the second one for gamma brain wave from Pz channel and the theta wave of the Pz channel achieved the lowest one. Unlike other researches mentioned above, such as [4] and [5] which just classified the targets, and got accuracy score, our research got difference in finding important features among EEG channels.

3.2. Other Machine Learning Model Result

Just like the CatBoost algorithm, we evaluated the achievement of the classifier by the accuracy score. The model we used were Random Forest classifier, Gradient Boosting classifier, XG Boosting classifier, KNN classifier, Decision Tree classifier, Gaussian NB, and Logistic Regressor. The random forest classifier showed 82.22 %, gradient boosting classifier showed 78.89 %, XG Boosting classifier yielded 78.22 %, and the least accurate one was the logistic regressor, which showed 41.11 %.

3.3. Deep Learning Model Results

For comparing our model to the deep learning models, we constructed a deep neural network (DNN), long short term memory (LSTM), and bi-directional LSTM(Bi – LSTM) model. It showed that the highest accuracy was through the DNN model with a 60 % accuracy. Next was the bi-directional LSTM model with 63.11 % and the least accurate one was LSTM with 51.33 %. The average accuracy score of the deep learning models were relatively lower than he machine learning models.

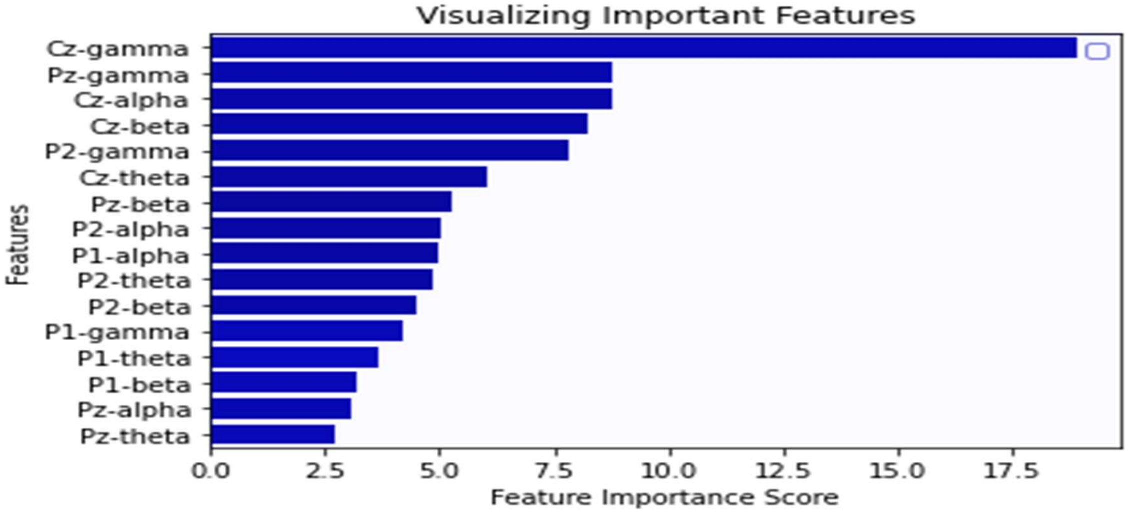


Figure 7. Feature importance score on detecting olfactory impairment

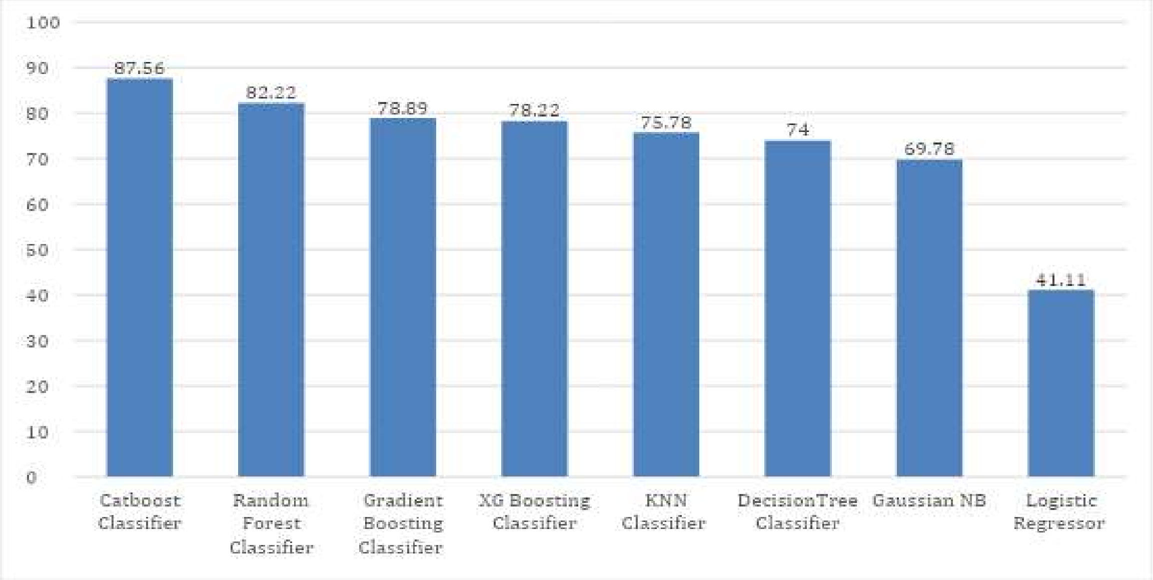


Figure 8. Accuracy of various machine learning models

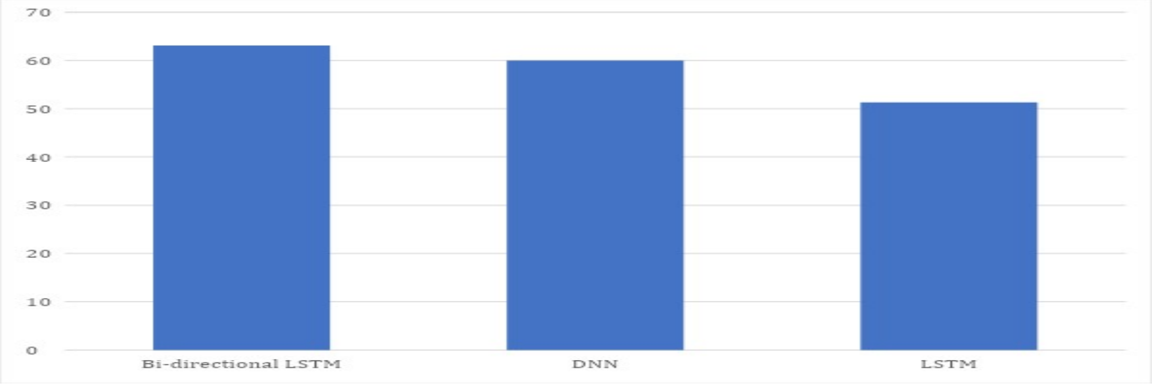


Figure 9. Accuracy of various deep learning models

4. DISCUSSION

4.1 Principal Finding

This research supports the possibility of combining CatBoost based machine learning algorithms with EEG signals as a diagnostic tool in olfactory impairment detection. THe CatBoost classifier model with 48 subjects showed the highest accuracy among the other machine learning and deep learning models at 87.56 %. As we succeeded in discriminating olfactory impairment with machine learning and deep learning models, it shows us the possibility of developing an objective diagnostic tool. Therefore, further research using machine learning and deep learning models should be undertaken. In addition to the EEG data used in our research, other types of body signals could be used. With EEG data, we can prevent patient factors that lower the reliability of diagnosis. For instance, a patient could be malingering for insurance money and as the current diagnosis methods are subjective ones, doctors or insurance companies cannot prevent malingering. This results in losses to insurance companies. However, if EEG data is used, patients can be prevented from malingering. Lastly, the CatBoost algorithm shows the feature importance, and our suggested model shows that gamma waves at the Cz channel have the highest feature importance. This result means that gamma waves at the Cz channel have the greatest influence on olfactory impairment. Even though explainable Artificial Intelligence (XAI) has been developed to show feature importance in deep learning models, it is now in the early stages so that it might take some time to identify the highest impact feature in a model.

4.2 Limitations and Further Considerations

Our model shows an accuracy of 87.56% in detecting olfactory impairment. However, there are some limitations in our model. First, the sample of data was insufficient. Subjects with olfactory impairment were not very common, so it was difficult to obtain enough data. As we could not have enough data, deep learning models such as DNN, LSTM, and Bi-LSTM could not perform well enough compared to other machine learning models [14]. Secondly, even though olfactory impairment can be divided into normosmia, hyposmia, and amosmia, we were not able to classify these three conditions. As there were not enough data available, we just divided them into patients or not, a binary classification. Lastly, as we only had CSV data, we could not apply the CNN model to our EEG data. If there were more subjects and we had sufficient EEG signal graph data, we could have used a CNN model. We could have also applied an Hybrid Neural Network (HNN) model combined with the CNN + LSTM or CNN + Bi-LSTM model. The HNN model could be appropriate for this research because the EEG data was time-series, so LSTM could be efficient for classification [15]. Therefore, for further research, collecting a large dataset should be considered as a top priority. When collecting a dataset, the percentage of normal people and people with olfactory impairment should be considered to prevent the dataset from leaning toward normal data. If there is a significant difference in the amount of data between them, the result of the model is likely to be unreliable with the possibility of overfitting [16].

5. CONCLUSION

Our research shows the possibility of further research into classifying olfactory impairment by combining a model that consists of AI based models and EEG data. We suggest that the CatBoost classifier could work effectively on EEG data. The accuracy of the CatBoost classifier was higher than the other machine learning and deep learning based models. The model also provides a feature importance graph that allows us to know that the gamma wave on the Cz channel had the highest effect on olfactory impairment. Furthermore, our research has strength on showing a possibility of EEG data for diagnosing olfactory impairment. As we took the first step of the field, other researches could follow and develop ours. There exists a limitation in that a larger dataset is required. In the future, further experiments should be conducted to obtain more data from olfactory impairment subjects so that further research can be performed with a greater variety of AI based models with EEG data.

REFERENCES

[1] Normosmia/Hyposmia/Anosmia. Encyclopedia of Neuroscience. :2900–.

[2] Bell MA, Cuevas K. Using EEG to Study Cognitive Development: Issues and Practices. Journal of Cognition and Development. 2012;13(3):281–94.

[3] AMR's hiring [Internet]. Allied Market Research. [cited 2021Apr23]. Available from: https://www.alliedmarketresearch.com/

[4] Jeon Y, Kim S, Choi H-S, Chung YG, Choi SA, Kim H, et al. Pediatric Sleep Stage Classification Using Multi-Domain Hybrid Neural Networks. IEEE Access. 2019;7:96495–505.

[5] Li Y, Xiao S, Li Y, Li Y, Yang B. Classification of Mild Cognitive Impairment from multi-domain features of resting-state EEG. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020;

[6] Ni Z, Yuksel AC, Ni X, Mandel MI, Xie L. Confused or not Confused? Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics. 2017;

[7] Wang X-W, Nie D, Lu B-L. Emotional state classification from EEG data using machine learning approach. Neurocomputing. 2014;129:94–106.

[8] Roohi-Azizi M, Azimi L, Heysieattalab S, Aamidfar M. Changes of the brain's bioelectrical activity in cognition, consciousness, and some mental disorders. Medical Journal of the Islamic Republic of Iran. 2017;31(1):307–12.

[9] Vezhnevets A, Barinova O. Avoiding Boosting Overfitting by Removing Confusing Samples. Machine Learning: ECML 2007. :430–41.

[10] Dorogush, A.V., Ershov, V., & Gulin, A. CatBoost: gradient boosting with categorical features support. 2018;ArXiv, abs/1810.11363.

[11] Bergstra, J. & Bengio, Y. Random Search for Hyper-Parameter Optimization.. J. Mach. Learn. Res. 2012; 13, 281-305.

[12] Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016;

[13] Feng S, Zhou H, Dong H. Using deep neural network with small dataset to predict material defects. Materials & Design. 2019;162:300–10.

[14] Song X, Liu Y, Xue L, Wang J, Zhang J, Wang J, et al. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. Journal of Petroleum Science and Engineering. 2020;186:106682.

[15] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data. 2019;6(1).