INTERMEDIATE FUSION APPROACH FOR PNEUMONIA CLASSIFICATION ON IMBALANCED MULTIMODAL DATA[*]

© 2024 O.N. Ivanova, A.V. Melyokhin, E.V. Ivanova, S. Kumar, M.L. Zymbler

South Ural State University (pr. Lenina 76, Chelyabinsk, 454080 Russia)
E-mail: onivanova@susu.ru, temamel540@gmail.com elena.ivanova@susu.ru, kumars@susu.ru, mzym@susu.ru

Received: 03.08.2023

In medical practice, the primary diagnosis of diseases should be carried out quickly and, if possible, automatically. The processing of multimodal data in medicine has become a ubiquitous technique in the classification, prediction and detection of diseases. Pneumonia is one of the most common lung diseases. In our study, we used chest X-ray images as the first modality and the results of laboratory studies on a patient as the second modality to detect pneumonia. The architecture of the multimodal deep learning model was based on intermediate fusion. The model was trained on balanced and imbalanced data when the presence of pneumonia was determined in 50% and 9% of the total number of cases, respectively. For a more objective evaluation of the results, we compared our model performance with several other open-source models on our data. The experiments demonstrate the high performance of the proposed model for pneumonia detection based on two modalities even in cases of imbalanced classes (up to 96.6%) compared to single-modality models' results (up to 93.5%). We made several integral estimates of the performance of the proposed model to cover and investigate all aspects of multimodal data and architecture features. There were accuracy, ROC AUC, PR AUC, F1 score, and the Matthews correlation coefficient metrics. Using various metrics, we proved the possibility and meaningfulness of the usage of the proposed model, aiming to properly classify the disease. Experiments showed that the performance of the model trained on imbalanced data was even slightly higher than other models considered.

Keywords: multimodal model, intermediate fusion, pneumonia, deep learning, imbalanced data.

Introduction

Pneumonia is the most common diagnosis in the world among all diagnosed lung diseases. During the pandemic, this disease took the first place among all diagnosed human diseases [1]. Timely, fast, reliable detection of pneumonia can allow doctors to start using treatment as early as possible, achieve positive dynamics, improve the prognosis of the course of the disease, and, ultimately, improve the health of the population due to fewer resources. Modern deep learning technologies allow the processing of data from several modalities at once, which is very practical in the field of medicine. Indeed, in the clinic, the patient passes a lot of laboratory tests and many different types of studies. The aggregation of the results of various types of medical research was previously performed only by an experienced doctor. Now, this task can be taken over by multimodal deep learning models, at least with a recommendation and informative purpose, acting in automatic mode.

The imbalance of data in the classification is a fairly typical situation in various fields of science and practice. In medicine, the presence of a sufficient and approximately equal number of training examples for each class of multiclass classification can be considered a great success for a researcher designing a classifier model. Improbability methods in machine learning involve the desire to align data in different classes [2]. This is usually done by reducing a larger class (for example, by random deletions) or increasing a smaller class (most often by combining an

[*] The paper is recommended for publication by the Program Committee of the International Scientific Conference "Global Smart Industry Conference (GloSIC) 2023".

encoder/decoder or using GAN). In turn, probabilistic machine learning models for solving binary classification problems are weakly dependent on the balance of classes.

Logistic regression and decision trees certainly respond to class imbalance: a significant change in the value of the free term and the impurity of leaves measure relative. However, neither one nor the other change has a significant impact on the result of the prediction. In regression, the determining factor is the slope coefficient, not the intercept. Trees share samples with impurities approximately proportionally. Therefore, in the case when you do not need to use SVM or other improbability models, you can work with imbalanced classes using the unequal class weights for random forests, gradient boost, and its variations.

For multimodal data, probabilistic deep learning models are usually used. We solved the problem of binary classification by attributing multimodal information about a patient to a class of healthy people or a class of people with diagnosed pneumonia. We used [3] as the dataset. The number of patients diagnosed with pneumonia was 50. The number of healthy patients was 500. The imbalance of classes, when the smaller class is from 1% to 20% of the total capacity of the dataset, refers to a moderate degree. The imbalance of classes in our multimodal model is moderate, since the minority class accounted for 9.1% of the total data set.

When building multimodal models, an important issue is the fusion of modalities. Late or early fusion of modalities is used only for certain types of tasks under certain data constraints. The detection of pneumonia from X-ray images and electronic medical records of the patient's laboratory tests implies a careful choice of the fusion model due to semantic heterogeneity and the remoteness of the modals from each other. The technical issues of implementing gradient boosting in multimodal models are also non-trivial. If many single-modal models can often be significantly improved by tuning hyperparameters, then the design of the architecture of a multimodal model is a more complex process. Here, it is necessary to take into account the side effects of multimodality in the coordination and harmonization of learning outcomes, including intermediate ones. Normalization methods that are used for single-modal learning may not be sufficient. Metrics for evaluating the accuracy of multi-modality processing models are also the subject of close study by many researchers [4–6].

In this paper, we have proposed approaches to solving the listed problems in addition to the problem of detecting pneumonia.

The scientific novelty of the proposed solution is determined by the following:

1. A model of multimodal deep learning with intermediate fusion for detecting pneumonia from X-ray images and the results of clinical studies of patients is proposed.
2. It is shown that balancing imbalanced data in a multimodal dataset using probabilistic models is acceptable and does not lead to deterioration of classification results.

The practical significance of the study is as follows:

1. The proposed multimodal model of deep neural network training on moderately imbalanced classes allowed for a significant improvement in the quality of the binary classifier compared to the results of training on a single modality with medical images, and maintained comparable accuracy with the model trained on balanced multimodal data.
2. The necessity of studying and calculating various metrics of the quality of the multimodal model for the formation of an adequate multiparametric assessment of various aspects of the model is shown.

The article is organized as follows. In Section 1, we provide brief review of related works. Section 2 introduces our multimodal model of intermediate fusion. In Section 3, we discuss the results of the evaluation of the proposed model. Conclusions summarize the results of the research.

1. Literature review

Imbalanced data in multiclass classification is a typical situation in real business and production processes. In the works [7–9] various methods of working with imbalanced data in the field of medicine are proposed. All these works were devoted to solving problems of multiclass classification according to the data of one modality.

At the same time, when creating multimodal models, there are often their own peculiarities of working with data. Multimodal deep learning models are subject to such problems as data alignment between modalities, problems of mapping, translation, fusion and co-learning of modalities [10–12].

It would be logical to assume that the complexity of multimodal models increases complementarily when learning probabilistic models on imbalanced classes. However, in this study, we found out that balancing classes in a multimodal model of pneumonia detection is an optional step that can be abandoned when solving this problem.

When reviewing the modern literature, we could not find examples of the random removal of samples from multimodal models for class alignment. The simplicity of implementing this method is outweighed by a major disadvantage — when using it, valuable samples with useful information may be lost. However, other methods of majority class undersampling are quite often used by researchers to restore balance. So, in the work [13], the Tomek links search method is used. This method works well on sets with a small number of features.

The authors of [14] suggest using the Condensed Nearest Neighbor Rule. This method is often used in a situation of imbalance of classes of a strong degree ($< 1\%$).

In [15], data preprocessing was performed using one-side sampling of the majority class (or one-sided selection). Computationally, it is quite demanding. It makes sense to use this method when the dataset contains a small total number of samples.

Another type of algorithms — neighborhood cleaning rule — is suggested by [16]. The main result of this method is the removal of noise that interferes with the training of the model.

In [17], the authors propose an imbalanced multimodal model for estimating and forecasting the value of real estate objects. The proposed model is based on oversampling, that is, increasing the number of examples of a minority class. At the same time, a simple algorithm for duplicating randomly selected samples was used.

In [18], the SMOTE algorithm (Synthetic Minority Oversampling Technique) [19] was used to solve the oversampling problem. The basis of this algorithm is the generation of artificial samples. Using the nearest neighbor method, a certain number of neighbors are selected for the sample. Then the feature vectors of each pair of neighboring features are multiplied by a random value from the interval $(0, 1)$. Thus, the feature vector of an artificially created sample in the area of a minority class is calculated. Unfortunately, this approach works well only for single-modal models with well-defined class domains. In the case when the vector space of features of different classes intersects or is mixed, as is often the case when analyzing medical observation data, the use of this method leads to a deterioration in the accuracy of classification.

The work [20] uses the ACM (Adaptive Synthetic Minority Oversampling) algorithm, which is a modification of SMOTE. The authors propose to perform the generation of artificial neighbor feature vectors only within the cluster. This modification of the oversampling algorithm is applicable to scattered classes, but greatly slows down the system, since it actually solves both the clustering problem and the sampling problem. There is also a requirement for the presence of clusters in the source data.

The authors of the paper [21] have demonstrated that the use of the oversampling algorithms described above is advisable only for binary classification, or if, in a multiclass classification, all minority classes differ from the majority by the same amount. In conditions where all classes (more than two) are imbalanced to varying degrees, the authors suggest using the ADASYN algorithm. This algorithm assumes calculating the coefficient of the distribution of sample weights within a minority class in order to generate artificial samples similar to the most important samples for learning. The coefficient is calculated based on the analysis of distances to samples of the majority class, which is a measure of complexity for training the model.

Thus, many researchers suggest balancing classes before training multimodal models. Such preprocessing takes a lot of time and resources. At the same time, the training of modalities is still carried out by probabilistic models that are insensitive to imbalance. Our idea is to eliminate the class balancing stage when building the architecture of a multimodal model of binary classification of medical data on lung diseases.

2. Methods

2.1. Data preprocessing

To test our hypothesis, we had to prepare two versions of the dataset — the imbalanced and balanced ones. The imbalanced dataset contained information about 50 patients with diagnosed pneumonia and 500 healthy people. We would like to make a reservation that by healthy people we meant patients who were represented in the same medical multimodal dataset, with diagnosed brain diseases, primarily with sleep disorders of various nature.

The dataset consisted of two modalities — X-ray images and the results of laboratory tests of patients' blood and urine. Images in the frontal position of the earliest registration were selected for each patient. As it is correct, for patients with pulmonary diseases, X-ray examinations were carried out repeatedly, with the preservation of all images in the medical history. With the course of treatment and the development of the disease, changes in the condition of the lungs were displayed on later images. Therefore, we took only primary images to train the model.

With regard to clinical data, the dataset in question had some redundancy for solving a particular problem of detecting pneumonia. In general, there are 1630 parameters of clinical analyses in the dataset. The reduction of the number of parameters occurred in several stages. First, all parameters that were not found in all patients were removed. After this operation, 523 parameters of clinical analyses remained. At the second stage, we made maps of the frequency of parameters and removed those that mainly occurred only in patients of a certain class. At the same time, we were guided by a threshold of 80%: if the parameter was observed in 80% of healthy and 80% of sick people, then such a parameter fell into our sample.

In medical practice, the patient receives an appointment for the delivery of parameters repeatedly. Usually, when a patient is hospitalized, the main indicators of blood and urine are studied. Further, with a primary and clarified diagnosis, the doctor prescribes additional studies. Therefore, at the last, third stage, we selected for consideration only those tests that were taken from patients at the beginning of hospitalization, thus eliminating disease-specific tests from consideration. In the end, we left 49 parameters of medical tests.

We obtain a balanced dataset through the sampling procedure, which is described below.

2.2. Data sampling

Imbalanced classes were the source data for sampling for us. To obtain balanced classes, we used the method of generating artificial samples of the greatest importance — ADASYN [22]. First, we used the SVM method to visualize class boundaries in two-dimensional space and found out that a number of samples of the minority class are quite close to the samples of the majority class, sometimes even mixed. That is why, of all the sampling methods, we chose and applied the ADASYN method in order to set the samples at the borders with greater weight and generate most of the artificial samples of the minority class in such "mixed" zones. This allowed us to make the boundaries between classes clearer. Also, we compared the predictive ability of the proposed model with the single-modal models described in [7–9]. In all the works, references to the source codes were given and the authors' permission to use their code to conduct experiments on other datasets was posted. All the work was processed only X-ray images and determined the presence of pneumonia.

2.3. Multimodal model of intermediate fusion

The general architecture of the proposed multimodal model is shown in figure. To process the modality of chest X-ray images, we chose the MobileNetV2 model. To process clinical data, we chose the Attention model presented in the Keras framework. The fusion of modalities was carried out according to an intermediate type. In the final metamodel, four embeddings were submitted for input — two from each modality. The first embedding of the clinical data modality came from the last learning layer of the transformer model. The second embedding moved from the last layer after pooling and compaction. The embeddings of the convolutional neural network used to train the classification of X-ray images were received, respectively, after the first and last block of convolutional layers.

The deep concatenation method was used for fusion, after which the metamodel created a fully connected level using the ReLU method and collapsed it using the SoftMax method.

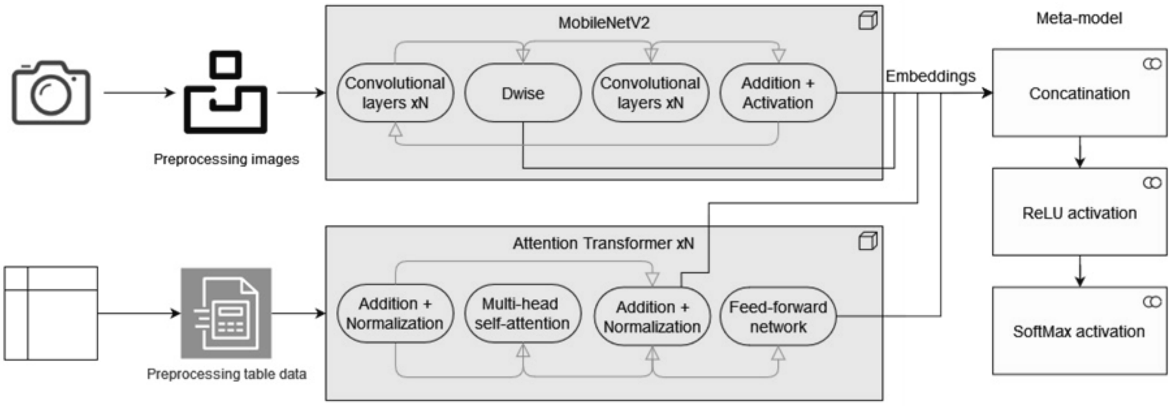

Fig. Architecture of the proposed multimodal classification model

We applied the classical loss function to self-tune the model:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p(P_i)\right) + (1 - P_i)\log\left(1 - p(P_i)\right). \quad (1)$$

2.4. Model validation

To test the accuracy of the model and the hypothesis as a whole, we compared the results not only between the two proposed models with balanced and imbalanced classes, but also with models that are publicly available for binary classification of healthy people and people with diagnosed pneumonia [7–9]. The presented models were single-modal and could only process lung X-rays.

We chose several metrics for training because classical single metrics cannot specify all the aspects of the multimodal data. The evaluation of the quality of the multimodal model is carried out both individually for each class (using the precision and recall metrics), and integrally, for all classes (metrics F1, PR AUC, and Matthews correlation coefficient).

The classical Accuracy metric for evaluating the accuracy of the model is useless in a situation with imbalanced classes. This metric will tend to inflate values for incorrectly defined negative classes, which will lead to low predictive ability.

The Precision metric can be interpreted as the proportion of correctly classified objects of a positive class, while the Recall metric shows what proportion of objects of a positive class out of all objects of a positive class the algorithm has found. An important feature of these metrics is that they are calculated not on the basis of predicted estimates, but on the basis of predicted classes.

There are several metrics called F_β which harmonize two metrics above — Precision and Recall. When evaluating the effectiveness of multimodal models, the F1 metric is traditionally used, which equally takes into account the importance of Precision and Recall.

For our task, it is more important to correctly identify the present disease than its absence. Thus, the F1 metric is suitable for our case, since we are very interested in the correct definition of a positive class.

The ROC AUC metric should not be used in our case. This metric is good when we have well-balanced classes and care about both true positive and true negative prognosis. In the case of 1:10 balance between classes, the false positive rate for highly imbalanced datasets is pulled down due to a large number of true negatives.

Unlike ROC AUC, PRAYS metric, like F1, focuses mainly on the positive class (Precision and True Positive Rate), pays less attention to the frequent negative class in imbalanced data. Therefore, this metric is also an adequate choice for our case.

Another popular metric for evaluating a model with imbalanced classes is the Matthews correlation coefficient. This coefficient is more complete compared to the F1 metric. The fact is that when using the F1 and PR AUC metrics, the key is the statement about which class is interpreted as positive — minority or majority. When exchanging labels, the value of the F1 metric will change, it will need to be recalculated. At the same time, the F1 metric completely ignores the true negative rate, which gives a certain limitation in the interpretation of the results of the model. The Matthews metric does not have these disadvantages, so it is advisable to use it in our case.

3. Results and discussion

We have run the full training cycle of the multimodal model five times. Here we present the averaged results of the model (Tab. 1). To generate balanced classes, we used the ADASYN method.

Table 1. Comparison of accuracy and losses in different models

Model	**Test accuracy, %**	**Test loss, %**
Single modality CXR [1]	93.0	4.81
Single modality CXR [3]	92.8	3.93
Single modality CXR NSGANETV2 [4]	93.5	4.11
Proposed model on balanced classes	95.9	2.14
Proposed model on imbalanced classes	96.0	2.11

As can be seen from the above data, the accuracy of the model using artificial balancing of classes and in the original representation of classes is almost the same, even with a slight advantage in favor of imbalanced classes. As we expected, the use of probabilistic deep learning models embedded in modern transformers and convolutional networks for binary classification allows us to work without pre-calibration of imbalanced classes in the presence of a sufficient number of samples. This result could be due to the nature and power of the dataset used, as well as the successful architecture of the multimodal model. We will not dare to assert the possibility of extrapolating this statement in relation to other tasks, models and datasets. However, the potential possibility of such elimination of one of the steps of data preprocessing may become a decisive factor when choosing models in conditions of limited time and computational capabilities of researchers. Table 2 presents the results of calculating metrics for evaluating the accuracy of the proposed multimodal model on imbalanced classes.

Table 2. Comparison of metrics on the proposed model

Metric **Dataset**	**ROC AUC**	**PR AUC**	**F1 score**	**Matthews coefficient**
Train	0.9827	0.9627	0.9577	0.9583
Test	0.9780	0.9480	0.9501	0.9455

The ROC AUC metric, as expected, has lower values, as it has a lower sensitivity to the minority class. This metric gives a false sense of the high accuracy of the model, but does not describe the real predictive ability of the model. The remaining metrics, PR AUC, F1 score, Matthews coefficient, give a more adequate assessment of the accuracy of the model on imbalanced data.

Conclusions

In this study, we tested the hypothesis that imbalanced data in a multimodal deep learning model for binary classification of the definition of pneumonia is not always necessary to undergo a balancing and alignment procedure.

For the dataset under consideration with a moderate degree of class imbalance, we found out that the generation of artificial samples with subsequent affixing of their correspondence to analogues in other modalities is optional. The training of a multimodal deep learning model by probabilistic algorithms for processing individual modalities is even slightly higher in accuracy than the results of the same model on artificially balanced data.

Excluding the data balancing step from data preprocessing before training the model can significantly increase the learning rate, and at the same time solves the problem of matching artificially generated samples in various modalities.

We conducted a study of the predictive ability of the model using several metrics. For imbalanced classes, the F1 and Matthews coefficient metrics showed a more adequate assessment of the accuracy of the proposed multimodal model.

This study is supported by the Russian Science Foundation regional grant No. 23-21-10009.

References

1. COVID-19 and vascular disease. EBioMedicine. 2020. Aug. Vol. 58. P. 102966. DOI: 10.1016/j.ebiom.2020.102966.
2. Problems of Training Sets Formation in Machine Learning Tasks. Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics. 2016. Vol. 16. P. 15–24. DOI: 10.14529/ctcr160302.
3. Soenksen L.R., Ma Y., Zeng C., *et al.* Code for generating the HAIM multimodal dataset of MIMIC-IV clinical data and x-rays. 2022. DOI: 10.13026/3F8D-QE93.
4. Qiu S., Chang G.H., Panagia M., *et al.* Fusion of deep learning models of MRI scans, Mini–Mental State Examination, and logical memory test enhances diagnosis of mild cognitive impairment. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring. 2018. Jan. Vol. 10, no. 1. P. 737–749. DOI: 10.1016/j.dadm.2018.08.013.
5. Parcalabescu L., Frank A. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models Tasks. 2022. DOI: 10.48550/ARXIV.2212.08158.
6. Bakalos N., Voulodimos A., Doulamis N., *et al.* Fusing RGB and Thermal Imagery with Channel State Information for Abnormal Activity Detection Using Multimodal Bidirectional LSTM. Cyber-Physical Security for Critical Infrastructures Protection. Springer International Publishing, 2021. P. 77–86. DOI: 10.1007/978-3-030-69781-5_6.
7. Sarada N., Rao K.T. A Neural Network Architecture Using Separable Neural Networks for the Identification of "Pneumonia" in Digital Chest Radiographs. International Journal of e-Collaboration. 2021. Jan. Vol. 17, no. 1. P. 89–100. DOI: 10.4018/ijec.2021010106.
8. Vashisht S., Sharma B., Lamba S. Using Support Vector Machine and Generative Adversarial Network for Multi-Classification of Pneumonia Disease. 2023 4th International Conference for Emerging Technology (INCET). IEEE, May 2023. DOI: 10.1109/incet57972.2023.10170180.
9. Yadav P., Menon N., Ravi V., Vishvanathan S. Lung-GANs: Unsupervised Representation Learning for Lung Disease Classification Using Chest CT and X-Ray Images. IEEE Transactions on Engineering Management. 2023. Aug. Vol. 70, no. 8. P. 2774–2786. DOI: 10.1109/tem.2021.3103334.
10. Fang M., Peng S., Liang Y., *et al.* A Multimodal Fusion Model with Multi-Level Attention Mechanism for Depression Detection. SSRN Electronic Journal. 2022. DOI: 10.2139/ssrn.4102839.

11. Cai S., Wakaki R., Nobuhara S., Nishino K. RGB Road Scene Material Segmentation. Computer Vision – ACCV 2022. Springer Nature Switzerland, 2023. P. 256–272. DOI: 10.1007/978-3-031-26284-5_16.

12. Msuya H., Maiseli B.J. Deep Learning Model Compression Techniques: Advances, Opportunities, and Perspective. Tanzania Journal of Engineering and Technology. 2023. June. Vol. 42, no. 2. P. 65–83. DOI: 10.52339/tjet.v42i2.853.

13. Pereira R.M., Costa Y.M., Jr. C.N.S. MLTL: A multi-label approach for the Tomek Link undersampling algorithm. Neurocomputing. 2020. Mar. Vol. 383. P. 95–105. DOI: 10.1016/j.neucom.2019.11.076.

14. Tang B., He H., Zhang S. MCENN: A variant of extended nearest neighbor method for pattern recognition. Pattern Recognition Letters. 2020. May. Vol. 133. P. 116–122. DOI: 10.1016/j.patrec.2020.01.015.

15. Xin L., Mou T. Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification. Wireless Communications and Mobile Computing / ed. by C.-H. Wu. 2022. July. Vol. 2022. P. 1–13. DOI: 10.1155/2022/9555790.

16. Aridas C.K., Karlos S., Kanas V.G., *et al.* Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers Under Imbalanced Data Sets. IEEE Access. 2020. Vol. 8. P. 2122–2133. DOI: 10.1109/access.2019.2961784.

17. Li Y., Branco P., Zhang H. Imbalanced Multimodal Attention-Based System for Multiclass House Price Prediction. Mathematics. 2022. Dec. Vol. 11, no. 1. P. 113. DOI: 10.3390/math11010113.

18. Mathew R.M., Gunasundari R. An Oversampling Mechanism for Multimajority Datasets using SMOTE and Darwinian Particle Swarm Optimisation. International Journal on Recent and Innovation Trends in Computing and Communication. 2023. Mar. Vol. 11, no. 2. P. 143–153. DOI: 10.17762/ijritcc.v11i2.6139.

19. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 2002. Vol. 16. P. 321–357. DOI: 10.1613/jair.953.

20. Siriseriwan W., Sinapiromsaran K. Adaptive neighbor synthetic minority oversampling techniqueunder 1NN outcast handling. Songklanakarin Journal of Science and Technology (SJST). 2017. Vol. 39. P. 5. DOI: 10.14456/SJST-PSU.2017.70.

21. Alhudhaif A. A novel multi-class imbalanced EEG signals classification based on the adaptive synthetic sampling (ADASYN) approach. PeerJ Computer Science. 2021. May. Vol. 7. P. 523. DOI: 10.7717/peerj-cs.523.

22. He H., Bai Y., Garcia E.A., Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of the International Joint Conference on Neural Networks, IJCNN 2008, part of the IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong, China, June 1-6, 2008. IEEE, 2008. P. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.