# Tackling the Class Imbalance Problem in Multiclass Brain Signal Classification

A.H.M.T.C. Bakmeedeniya[1#], Y.M.R.D. Wepathana[2]

*Department of Information Technology*
*Advanced Technological Institute*
*Kegalle*
*Sri Lanka* [1&2]

ahmtcbakmeedeniya@sliate.ac.lk [#], ymrdilhani@sliate.ac.lk [1&2]

***Abstract***

***Brain-Computer Interface has become a heavily researched topic in the healthcare community, law enforcement, and other sectors. Some popular experimental areas in BCI are Sleep stage classifications, emotions detections, epileptic seizures, and alcoholism. Electroencephalogram(EEG) signals that are being recorded by the electrodes placed on the scalp are used widely for those experiments. A series of electroencephalogram (EEG) signal processing techniques have been developed rapidly recently. Among them, Machine Learning(ML) has become the most common development in the last decades. Class distribution of a dataset can make a significant effect on the prediction accuracy and performance of the model in ML. The imbalanced problem has become a critical issue to be solved in Machine learning. Oversampling and under-sampling are the two common mechanisms to be used in learning from the imbalanced dataset. The experiment investigates the behavior of two sampling techniques in a multiclass EEG signal classification problem. The results exposed that sampling can improve the performance of an ML model predictions in terms of both accuracy and F1- score. Accuracy improved from 0.91 to 0.93 where the F1- score increased from 0.49 to 0.72 in the rebalancing dataset.***

***Keywords**: Electroencephalogram, Oversampling, under-sampling, Rebalancing, multiclass classification*

## INTRODUCTION

In recent years, many researchers have shown their interest in Brain-Computer Interface (BCI) systems, which has resulted in many experiments and applications. BCI translate brain signal into control commands in interpreting specific human activity. Electroencephalography (EEG) is a widely used technique to measure brain activities. EEG signals record the spontaneous electrical activity of the human brain using the electrodes placed on the scalp. BCI stay popular due to some significant characteristics in EEG signals such as destructiveness, painlessness, and accurate interpretation of some brain disease. In general, EEG signals are classified according to the frequency, amplitude, shape, and position of the electrodes on the scalp. Frequency is the basic unit used to determine normal or abnormal rhythms. Brain wave frequency differs and corresponds to the behavior and mental states of the brain. These signals have a 0Hz – 100Hz frequency range (Kumar and Bhuvaneswari, 2012).

Chaabene et al. presented a comparative analysis to identify the brain state based on convolutional neural networks(CNN) (Chaabene et al., 2021). A wavelet transform-based approach by using CNN and deep learning has been presented by Budak et al. (Budak et al., 2019). A comparison of the performance of different machine learning algorithms in classifying EEG signals has been researched by (Joshi et al., 2022). (Mousa, El-Khoribi and Shoman, 2016) have experimented with differentiating the sleep stages on EEG signals using different classifiers. Principal Component Analysis has been applied for dimensionality reduction in the approach (Mousa, El-Khoribi and Shoman, 2016).

Datasets can have a significant effect on the performance of a model in machine learning. Therefore a quality dataset is always required for building a high-accuracy model. To record EEG a system consists of electrodes, amplifiers, an AID converter, and a recording device. The electrodes get the human signal through the scalp, in tum, the amplifiers progress the analog signal to expand the amplitude of the EEG signals for the AID converter to make the signal digital in a precise manner(Mousa, El-Khoribi and Shoman, 2016). Recording EEG signals is a time-consuming process that requires additional hardware and expert knowledge. Hence, the majority of the studies in BCI are using publicly available datasets. Class imbalance of a dataset is a challenging issue that needs to resolve in statistical machine learning.

Recently, there is a great interest in “Learning from imbalanced data” and new learning algorithms designed specifically for imbalanced data. One of the common approaches was to use resampling techniques to make the dataset balanced. Machine learning presents automatic approaches to resolve the class imbalanced problem using different sampling techniques depending on the features extracted using expert knowledge from the raw signal. Resampling techniques can be applied either by under-sampling or oversampling the dataset (Mohammed, Rawashdeh and Abdullah, 2020).

Under-sampling, reduce the instances from the majority class to make it a match with the minority class. It may remove possibly valuable data that can be essential for classifier models, but it is useful when you have a huge number of data. Further, under-sampling is used as a data cleaning process to remove noise (AT et al., 2016). Oversampling is a technique, which duplicates examples from the minority class to match the number of instances in the majority class (Fernández et al., 2018). Combined sampling is a technique to combine oversampling and under-sampling to improve the accuracy of the classification (Safira et al., 2021).

The models built using severely skewed datasets are inefficacious in identifying the minority classes. Further, show overfitting results for the majority classes. When comes to multiclass classification problems where more than two labels exist for a dependent, this can be more trivial.  Hence, this study investigates class-rebalancing techniques to improve the performance of models in EEG signal classification in BCI systems.

## METHODOLOGY

### *Dataset*

Most researchers have experimented with classifying sleep stages using the dataset available on https://physionet.org/. This data set is a multi-class dataset and imbalanced.  EOG and EEG signals were each sampled at 100 Hz in the sleep- cassette. There were 153 PSG recordings and the relevant Hypnogram files which have recorded from 82 healthy Caucasian subjects aged 25-101. The EEG features were recorded from Fpz-Cz and Pz-Oz electrode locations. Hypnogram files contain annotations of the sleep patterns that correspond to the PSGs. These patterns (hypnograms) consist of sleep stages W, R, 1, 2, 3, 4, M (Movement time), and ? (Not scored). The PSG files were in the format of EDF while the hypnograms were in EDF+ (Vilamala, Madsen and Hansen, 2020). This study has used the W, 1, 2, 3and 4 classes in consideration.

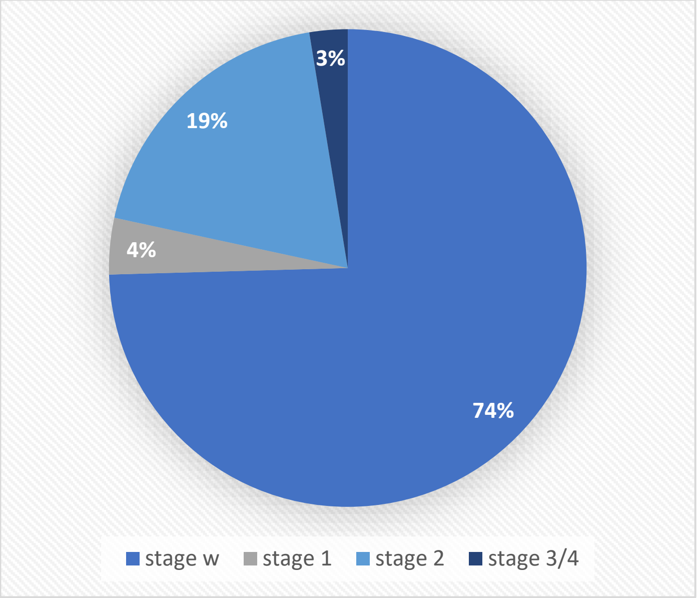

**Figure 1.** Class distribution of Original dataset

## Proposed approach

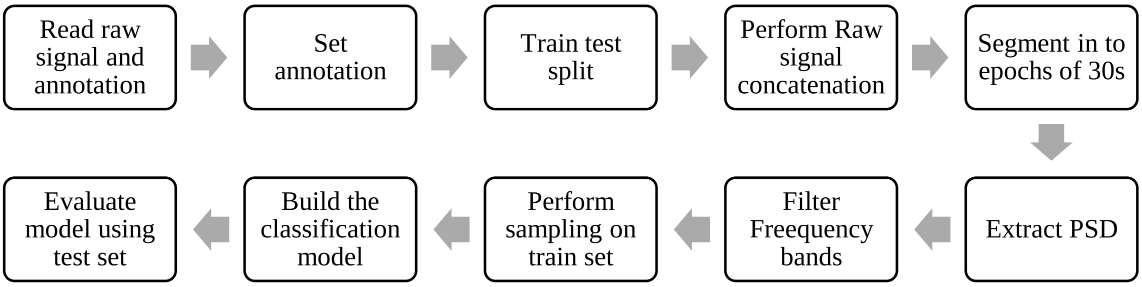
Read raw signal and annotation
Set annotation
Train test split
Perform Raw signal concatenation
Segment in to epochs of 30s
Extract PSD
Filter Freequency bands
Perform sampling on train set
Build the classification model
Evaluate model using test set

**Figure 2.** Proposed approach

Two third of the data set were taken into training and the rest were taken as the test set. As the proposed approach in Figure 1 shows the raw signals of both train and test sets were annotated using the given annotation file in the dataset. The annotated raw signals were concatenated and Fpz-Cz channels were retrieved. Then the raw signal was segmented into epochs of the 30s and extracted the Power Spectral Density (PSD) relevant to the EEG feature. Filter the frequency band from 0.5Hz-30 Hz because the sleep stages are falls in that frequency range. The events; Sleep stage W, Sleep stage 1, Sleep stage 2, Sleep stage 3 and 4 were extracted as the targets. As represents, the class distribution is severely skewed. Hence, the calculated PSD s for the training epochs were then resampled to get equal class distribution. A Classification algorithm was applied to the resampled dataset with the relevant targets for training the model. The trained model was tested using the PSD s of testing epochs.

## Rebalancing

The study used the combined sampling of oversampling and under-sampling. Synthetic Minority Oversampling Technique, now widely known as SMOTE is an oversampling technique, which is considered one of the most influential data preprocessing/sampling algorithms in machine learning and data mining. The basis is to carry out an interpolation among neighboring minority class instances (Fernández et al., 2018). The study used SMOTE for oversampling the dataset. SMOTE can generate noisy samples by interpolating new points between marginal outliers and inliers. This issue can be solved by cleaning the space resulting from over-sampling by using Tomek Link.

Tomek Link (T-Link) is a method of under-sampling. It is considered an enhancement of the Nearest-Neighbor Rule (NNR)(AT et al., 2016). The T-Link method can be used as a data cleaning technique, which can be combined with an oversampling such as SMOTE.

## Classification

Classification model built using Support Vector Machine (SVM) classifier with the Radial Basis kernel Function. Support Vector Machine (SVM) is a classification method where a training data set representing two different classes is projected into a high dimensional space through a kernel function. Not only it has a better theoretical foundation, but practical comparisons have also shown that it is superior to the ANN(AT et al., 2016).

## Evaluation

Evaluation of the model performed in terms of Accuracy as well as F score. Accuracy represents the count or the fraction of the correct prediction over the total number of samples.

$$Accuracy = \frac{TP+TN}{TN+FP+FN+TP} \tag{1}$$

Where True Negatives (TN) represent the amount of correct prediction of negative class. False positive will give the amount of incorrectly predict negative class. False Negatives (FN) and True Positives (TP) represent the number of incorrect predictions and amount of correct predictions of positive class respectively.

F-Measure represents a balance between Precision and Recall hence this can be taken as the most important matrix in the performance evaluation of a model. Though accuracy represents the number of correct predictions it is not enough to use only the accuracy scores in model evaluation. Some models are skewed to predict one type of class label though it gives higher accuracy. So another set of the matrix is required other than the accuracy. Precision is the ability of the classifier not to label as positive a sample that is negative. Recall is another important matrix in model evaluation in a classification problem as it exposes the ability of a model to capture the actual positive values as positives.

$$F1 = 2 * \frac{P*R}{P+R} \quad (2)$$

$$P = \frac{TP}{TP+FP} \quad (3)$$

$$R = \frac{TP}{TP+FN} \quad (4)$$

F scores are calculated for each class separately and then calculate as a micro average for comparison. Micro average calculate metrics globally by counting the total true positives, false negatives, and false positives.

## RESULTS AND DISCUSSION

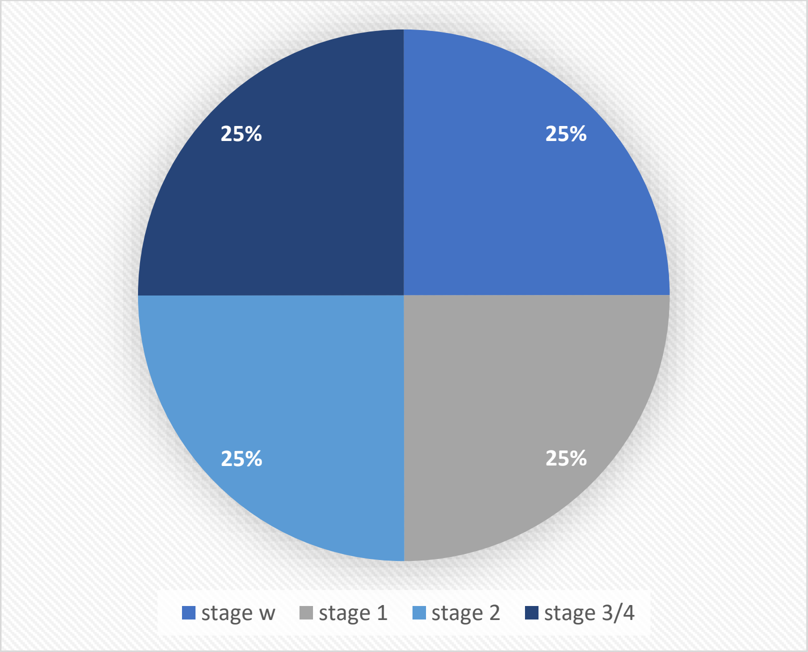

**Figure 3.** Class distribution after resampling

Originally 2516 amounts epochs were extracted and as Figure 2 depicts 74% are from the "stage 1" class type among them. Only 3%, 4%, and 19% instances are included from the rest of the class respectively. Which is heavily unbalanced and will show overfitting results for the majority class. In contrast, Figure 3 shows that the resampled dataset distributes the proportions of instances equally as 25% among the classes. 1875, 1873, 1872, and 1874 amount of epochs were resampled for "stage w", "stage 1", "stage 2" and "stage 3/4" respectively.

The accuracy of the prediction was figure 0.91 in the original dataset whereas 0.93 in the resampled dataset. Though the given accuracy is high in original dataset it shows poor F1-score. The average F1-score for all the classes was 0.49 and 0.72 in original and resampled data sets respectively as elaborated in the Figure 4. F1 score has increased by 0.23 after rebalancing the dataset.

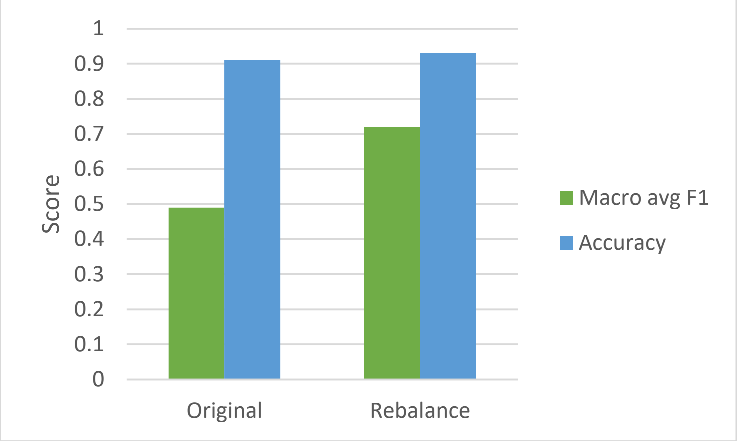

**Figure 4.** Score comparison

F1 scores presented in Table 1 show that the model trained with the original dataset has given low F1 scores for the classes “stage 1” and “stage3/4” as opposed to the class “stage w” and “stage 2”. The lowest F-score has been given for “stage 1” in the original dataset which had the least amount of instances. “stage3/4” classes have also given a low F-score of 0.07 whereas the original dataset contains a low amount of instances as shown in Figure 2. The model built using rebalanced dataset showed improved F1 scores for “stage 1” and “stage 3/4" as 0.23 and 0.82.

**Table 1:** F1-Score comparison of classes

| | **stage w** | **stage 1** | **stage 2** | **stage 3/4** |
|---|---|---|---|---|
| **Original** | 0.99 | 0.06 | 0.83 | 0.07 |
| **Rebalanced** | 0.99 | 0.23 | 0.86 | 0.82 |

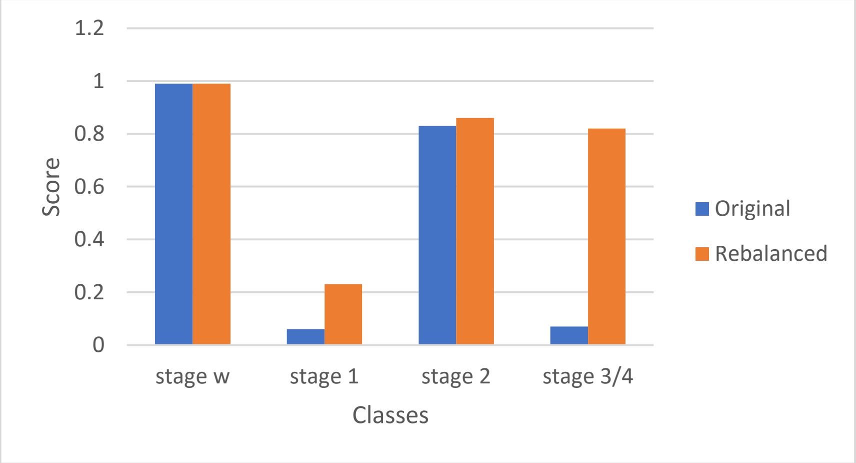

**Figure 5.** F1-Score comparison of classes

## CONCLUSION

Raw EEG signal processing is used in most of the experiments in BCI systems. It is widely used in clinical setups to diagnose brain diseases. Machine learning plays a vital role in classifying EEG signals. Learning from an imbalanced dataset has been wide open for research as the imbalanced dataset is a challenging issue in statistical machine learning. Models built with an imbalanced dataset may be biased to predict the majority class. This paper presented an analysis of the impact of rebalancing an EEG raw signal dataset. SMOTE and Tomek Link are used for Oversampling and under-sampling respectively and SVM is used for classification. Accuracy and F1-score were used for the evaluation of the models. As Figure 5 shows rebalanced dataset has improved the F1 score of all the classes with low amount of epochs extracted while accuracy remains nearly 0.9. It can be concluded that, rebalancing improves the F-scores of a EEG signal classification model.

## REFERENCES

AT, E., Aljourf, M., Al-Mohanna, F. and Shoukri, M., 2016. ‘Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method’, *Global Journal of Technology and Optimization*, 01(S1). doi: 10.4172/2229-8711.s1111.

Budak, Ümit & Bajaj, Varun & Akbulut, Yaman & Atila, Orhan & Sengur, Abdulkadir., 2019, ‘An effective hybrid model for EEG-based drowsiness detection’, *IEEE Sensors Journal*, 19(17), pp. 7624–7631. doi: 10.1109/JSEN.2019.2917850.

Chaabene, Siwar., Bouaziz, Bassem., Boudaya, Amal., Hökelmann, Anita., Ammar, Achraf., and Chaar, Lotfi., 2021, ‘Convolutional neural network for drowsiness detection using eeg signals’, *Sensors*, 21(5), pp. 1–19. doi: 10.3390/s21051734.

Fernández, A., Garcia, Salvador., Herrera, Francisco., and Chawla, Nitesh V., 2018, ‘SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary’, *Journal of Artificial Intelligence Research*, 61, pp. 863–905. doi: 10.1613/jair.1.11192.

Joshi, A., Kamble, A., Parate, A., Parkar, S., Puri, D., and Gaikwad, C., 2022, ‘Drowsiness Detection using EEG signals and Machine Learning Algorithms’, *ITM Web of Conferences*, 44, p. 03030. doi: 10.1051/itmconf/20224403030.

Mohammed, R., Rawashdeh, J. and Abdullah, M., 2020 ‘Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results’, *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, pp. 243–248. doi: 10.1109/ICICS49469.2020.239556.

Mousa, F. A., El-Khoribi, R. A. and Shoman, M. E., 2016, ‘A Novel Brain Computer Interface Based on Principle Component Analysis’, *Procedia Computer Science*, 82(1), pp. 49–56. doi: 10.1016/j.procs.2016.04.008.

Mohd Noor, Nor Safira Elaina & Ibrahim, Haidi & Lah, Muhammad & Abdullah, Jafri., 2021, Improving Outcome Prediction for Traumatic Brain Injury From Imbalanced Datasets Using RUSBoosted Trees on Electroencephalography Spectral Power. *IEEE Access*, PP. 1-1. 10.1109/ACCESS.2021.3109780.

Vilamala, A., Madsen, K. H. and Hansen, L. K., 2020, ‘DEEP CONVOLUTIONAL NEURAL NETWORKS FOR INTERPRETABLE ANALYSIS OF EEG SLEEP STAGE SCORING Technical University of Denmark Danish Research Centre for Magnetic Resonance’, (659860).