# The effect of intra-patient sample correlation in EEG based Sleep Stage Classification

Gabriel-Danut Matei Sébastien Travadel Philippe Blanc
Andrea Goldstein-Piekarski

April 2022

**Abstract**

Automated sleep stage scoring has seen numerous improvements over time, as a result of both increased availability of public data and of increased computational power leading to the use of more complex models. In the race for attaining the highest possible accuracy, a number of highly practical considerations such as reproducibility, explainability and usability are being ignored. In this paper, we analyze how the inherent correlation in the data can be exploited to obtain reasonable performance with a number of simple models.

***Keywords***— Sleep Stage Scoring, Classification, Correlation, Dimensionality reduction, Undersampling, Bayes error rate

## 1 Introduction

In medical settings, a patient's biological signals are a simplified way of monitoring health evolution over a shorter or longer period of time. Examples include information regarding oxygen levels, blood pressure during an operation or eeg, emg and eog signals in the case of sleep studies. In these cases, the purpose could either be to closely monitor and anticipate problems in real-time or to derive a more simplified representation of a lengthier process.

Specifically, in the case of a sleep stage study [1] (which involves a patient spending a full night of sleep in laboratory conditions), the goal is to split the entire period of sleep into 30 second intervals (called epochs) and assign to each one of 5 possible labels (or sleep stages): Wake, Stage 1, Stage 2, Stage 3 or Rapid Eye Movement (REM). By analysing the length and alternation of these distinct stages, a doctor can detect the presence of a wide range of pathological conditions such as insomnia or narcolepsy.

Traditionally, the labelling (or scoring) is a task that is accomplished manually by trained professionals, which typically can score an 8-hour long sleep recording in about 2 to 3 hours. Nowadays, numerous automated models exist that get good performance on widely available datasets such as the Montreal

Archive of Sleep Studies (MASS) [2], with the State of the Art neural network approach [3] reaching 87% accuracy, which is comparable to human performance.

Behind these remarkable results however, there might lie a phenomenon that has not been properly documented in the literature: the fact that current neural network architectures might be needlessly complex for the amount of data they typically train on. Because the number of patients in even the largest datasets is small (under 200), if the samples coming from the same patient are correlated enough, the total number of independent training points might be only a fraction of what it appears to be at a first glance.

The paper is structured in the following way:

- Section 2 presents the evolution of automated sleep scoring and the effect of sample correlation on statistical inference in a medical context.
- Section 3 analyzes the patient sample correlation through a KNN method, and describes the creation of a new "uncorrelated" dataset followed by training a number of simple machine learning models on this dataset
- Section 4 presents a performance comparison between the 3 best models trained on the "uncorrelated" dataset, a baseline majority-class classifier, and an estimation of the Bayes Optimal Classifier.
- Section 5 shows why a reduction in accuracy (compared to the State of the Art), might be preferable, as simpler models tend to come with their own suite of advantages
- Finally, Section 6 is a summary of the paper and includes further questions regarding the subject at hand that might be worth investigating

## 2 Related work

The task of automated sleep scoring is not a new one, and different approaches have been tried over the course of the years.

The first models that were created were relatively simple and relied more heavily on feature engineering [5] [6] [7]; this was because of the low amount of available data (usually less than 10 patients). As time passed, more encompassing datasets were created and made publicly available, which allowed for more complex machine learning models (such as neural networks).

Currently, most of the top performing models are deep neural networks [8] that have gradually shifted away from manual feature engineering in favor of a more simple preprocessing of data. In particular, the state of the art utilizes an end-to-end learning approach, meaning that even the filtering steps involved in the preprocessing step are dictated by the training data.

Whereas the classical approach (which involved complex preprocessing and simple models) was prone to underfitting, the modern approach (simple preprocessing and complex models) might be headed in the opposite direction, that of overfitting. Overfitting can be mitigated through a number of methods

that constrain the complexity of the model (including dropout, L2 regularization and batch normalization), but the best way is to provide sufficient data samples. However, as stated in [4], correlation between observations (which naturally arises from repeated sampling of the same patient) that is unaccounted for leads to a loss of confidence in the computed statistics. Consequently, this can lead to an overestimation of both the training set's representativity, and the classifier's performance on data coming from other patients.

# 3 Methods

## 3.1 Problem statement

The problem of classifying 30 second intervals of biological signals (EEG, EMG and EOG) can be formally presented in the following way.

We start with 3 stochastic processes

$$\{X_{eeg}(t)\}_{t\in T}, \{X_{emg}(t)\}_{t\in T}, \{X_{eog}(t)\}_{t\in T} \in R$$

and a sampling frequency $f \in N^+$. For each $i \in N^+$ we define a subset of discrete indices

$$S_i = [30 \cdot f \cdot i, 30 \cdot f \cdot (i+1)] \subset T$$

which will identify each separate epoch. For each $i \in N^+$ we also have an associated label $y_i \in \{0, 1, 2, 3, 4\}$ corresponding to the epoch's sleep stage.

The goal is to find a function that minimizes

$$C(f(\{X_{eeg}(t), X_{emg}(t), X_{eog}(t)\}_{t\in S_i}), y_i), i \in N^+$$

where $C$ is a classification cost function.

Depending on the data acquisition method, one or more of these signals might be missing, but the one that is the most important for correct scoring is the EEG signal.

## 3.2 Montreal Archive of Sleep Studies (MASS) Dataset

The MASS dataset [2] is a collection of whole night sleep recordings taken from a diverse range of patients that includes, among others, EEG, EMG and EOG data. It is divided into 5 cohorts and totals 200 patients with ages between 18 and 76, approximately equally distributed among the 2 genders.

Because of data corruption reasons, in our experiemnts we only used 150 recordings. In addition to this, we decided to focus only on the EEG channel as it is the most informative. Naturally, integrating the other available channels would lead to an across the board increase in the reported performance metrics.

### 3.3 Data representation

Initially, the data is made up of non-overlapping 30 second intervals sampled 100 times each second (downsampled from 256 Hz). This means 3000 features for each sample.

Even intervals that are part of the same class can vary quite a lot visually. For this reason, we compute a spectrogram to gain frequency information, which is the same approach used by the state of the art. After this step, a sample has shape 30 x 130.

We further average the frequency information over the whole 30-second interval, which leaves us with only 130 features. This completely removes time information from the sample, but offers a better summary of the whole epoch.

### 3.4 Correlation between observations

To demonstrate that samples from the same patient are fundamentally correlated, we proceed in the following way.

First, we choose a random patient from our dataset (which we will call *reference patient*) and we set aside the last 80% of his epochs for testing.

Then, we will create and train 2 models:

1. **General KNN** - trained on sampled data from 10 other random patients, totaling 500 training samples
2. **Patient KNN** - trained on the first 20% epochs of the reference patient, totaling 200 training samples

We repeat this procedure 150 times by sequentially choosing each patient as the reference patient. If the epochs are relatively independent, we expect the General KNN which was trained on more numerous and diverse data to outperform the Patient KNN.

As can be seen from Figure 1, the opposite is, in fact, true. The average accuracy of the General KNN is 52.86%, while the average accuracy of the Patient KNN is 65.93%.

### 3.5 Creating an uncorrelated training set

First, we will randomly split our 150 patients into a group of 120 reserved for training models and a group of 30 reserved for testing.

Because observations coming from the same patient are so correlated, we will create a simplified training set of only 10 samples / class / patient. This will ensure that the data we use for training is as representative and as diverse as possible.

We will also reduce the dimensionality of each sample with one PCA per class, retaining $\approx 90\%$ of the variance with only 7 principal components. Since samples from different classes have strongly differing shape, we will create one PCA model for each class, in an attempt to project our data into each class' representation. This leads us to a total of 35 features per observation. This

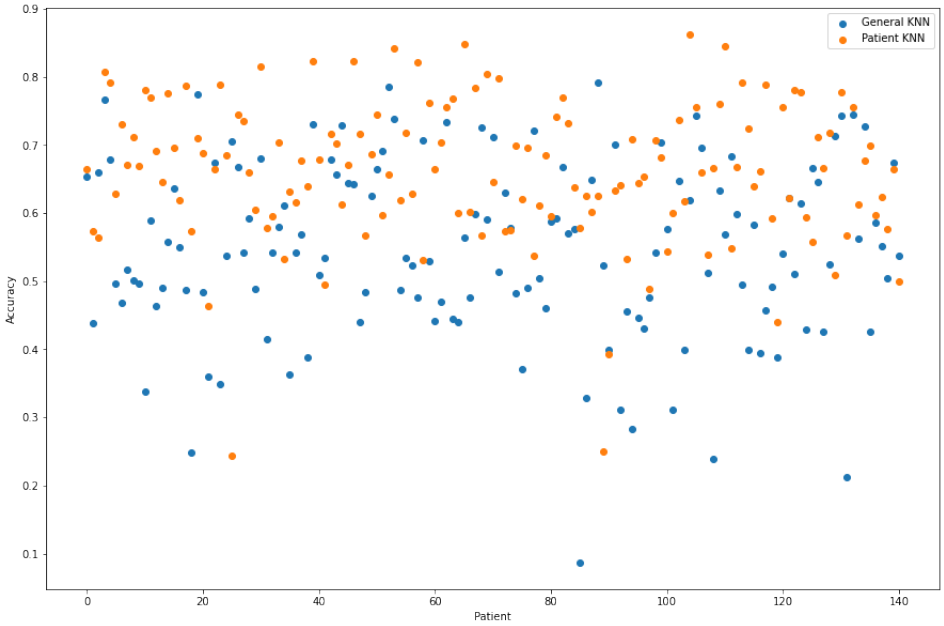


Figure 1: Accuracy comparison by patient for the two types of KNN

drastic reduction in the number of features ensures that we do not fall prey to the curse of dimensionality, especially considering the size of our reduced dataset.

The data will be used to train and validate a number of simple classifiers, followed by a final evaluation on the held out test set.

The way the entire data is distributed into training, validation and test sets is illustrated in Figure 2. Note that we separate the training data used for the PCA from the training data used for the classifiers.

After the top 3 models are determined using the validation accuracy, the validation set is reincorporated into the classifier training set for a final fit. Lastly, these 3 models are evaluated on the test set.

### 3.6 Evaluating the performance of basic classifiers

Because of the limited size of the newly defined dataset, we limit our exploration of classifiers to only the simplest among them: Naive Bayes, KNN, Decision Tree, Random Forest, SVM, LDA, QDA and MLP. The top 3 of these models are then selected and their hyperparameters are optimized. Note that we do not employ extensive hyperparameter search methods as we want to emphasize results obtained by relatively simple means.

Furthermore, we would like to put into context the performance of the best classifiers. For this reason, we train 2 additional models that will help us ap-

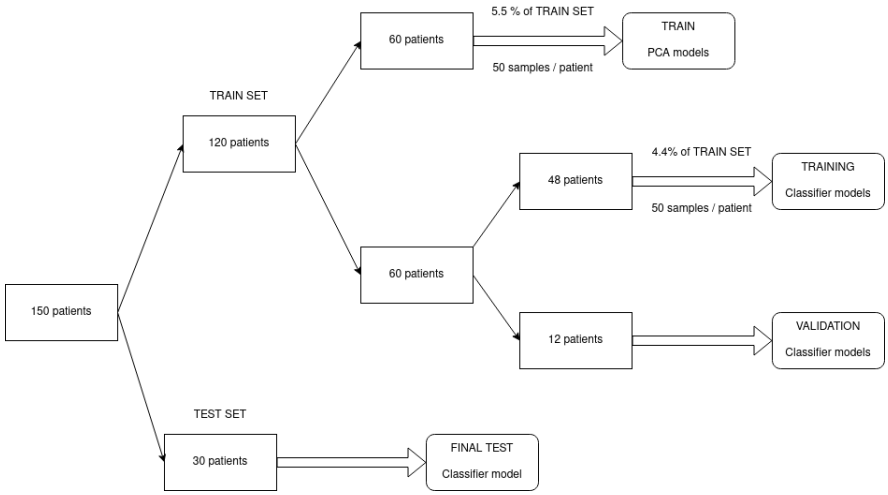


Figure 2: Data split for evaluating the simple classifiers

proximate the lower and upper bound respectively on the accuracy.

For the lower bound, an easy choice would be to train a baseline "dummy" classifier that labels all epochs as the most frequent class.

For the upper bound, the theoretically optimal classifier is the Bayes Optimal Classifier, with its associated Bayes Error Rate. It is the best model we can obtain if we already know the true distribution of the population, assumption which is often unrealistic. Fortunately, [9] offers a way to obtain a lower bound on the Bayes error rate ($\rho$) by using the error rate of a 1NN classifier ($\alpha$):

$$\rho \geq \frac{\alpha}{2}$$

By using the lower bound on the error rate we easily obtain an upper bound on the accuracy.

# 4 Results

After extensive training and validation, the 3 most performant models are the MLP, the SVM and the 7NN. Specifically, when trained on our dataset made up of 10 samples / class / patient, their test set accuracies are 68%, 69% and 63% respectively.

Admittedly, even though the value 10 proved to be a reasonable choice for our purposes, it is rather arbitrary. For this reason, we decided to plot the results of the training and subsequent testing process based on the number of samples / class / patient that were used in the creation of our dataset. The final results can be seen in Figure 3. Boosting was used to counteract the effects of the random sampling from the available epochs.

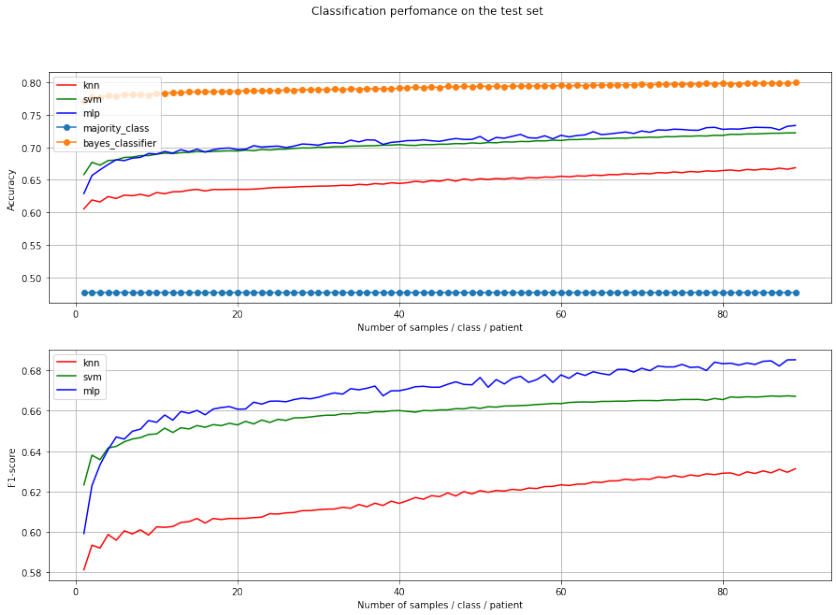


Figure 3: Test set results

# 5 Discussion

## 5.1 Interpretation

Figure 3 offers a great number of insights related to our models.

First of all, their performance is significantly better than the "dummy" classifier, even when trained on only a couple of samples / class / patient. Even though highly specific, the observations coming from each patient manage to contribute enough to our classifiers so that they generalize.

Second of all, there are still some improvements left to be made when comparing them to the Bayes Optimal Classifier. This might be because of the relatively simple nature of the chosen models, as they might still lean on the side of underfitting.

Finally, increasing the number of samples / class / patient contributes to an increase in both accuracy and f1-score, but this improvement too small to justify the increased training time.

## 5.2 Comparison with the State of the Art

The State of the Art [3] is a sequence-to-sequence model, meaning that it takes as an input a list of 20 consecutive epochs and outputs a corresponding list of 20 labels. Unfortunately, this particularity makes it incompatible with the way we created our uncorrelated dataset, as epochs are randomly sampled and their

order is not kept. For this reason, we cannot directly evaluate its performance the same way we did with our models.

Nonetheless, we can still compare some aspects, specifically the internal complexity and computational cost. While our proposed models cannot match the State of the Art's raw accuracy, they offer a reasonable performance with only a fraction of required resources. They can be trained on more modest machines, having less parameters by two orders of magnitude and utilizing only a tenth of training data, saving on both CPU time and memory.

| | State of the Art | Proposed models |
|---|---|---|
| Parameter count | $10^4$ | $10^2$ |
| Training data | 100% | 9.9% |
| Accuracy | 87% | 70% |

# 6 Conclusion and Future Work

We started from the assumption that the data used in automated sleep scoring is fundamentally correlated, because of intra-patient similarity. We proved this assumption using 2 KNN models and we subsequently created an uncorrelated dataset, consisting of approximately 10% of the initial number of observations. Then, we trained a number of simple classifiers, the top 3 of which had accuracies between 63% and 69%. Finally, we presented the advantages of using these models instead of the State of the Art, most of them stemming from significantly reduced computational and storage costs.

Some other directions of future inquiry include: the performance improvement from adding the eog and emg channels, the effect of number of patients on the final accuracy and lastly testing on other cohorts of patients that are not part of the MASS dataset.

# Acknowledgements

My deepest thanks to Sébastien and Philippe for all of the brilliant lectures during Data Sophia and likewise for the encouragement and guidance they offered at all stages of this project.

Additionally, none of this would have been possible without the amazing introduction to sleep scoring offered by Andrea who also was the key factor in getting access to the MASS dataset.

# References

[1] Richard S Rosenberg and Steven Van Hout. The american academy of sleep medicine inter-scorer reliability program: sleep stage scoring. *Journal of clinical sleep medicine*, 9(1):81–87, 2013.

[2] Christian O'reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of sleep research*, 23(6):628–635, 2014.

[3] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Seqsleepnet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

[4] Kristin Sainani. The importance of accounting for correlated observations. *PM&R*, 2(9):858–861, 2010.

[5] Masaaki Hanaoka, Masaki Kobayashi, and Haruaki Yamazaki. Automated sleep stage scoring by decision tree learning. In *2001 conference proceedings of the 23rd annual international conference of the IEEE engineering in medicine and biology society*, volume 2, pages 1751–1754. IEEE, 2001.

[6] Guillaume Becq, Sylvie Charbonnier, Florian Chapotot, Alain Buguet, Lionel Bourdon, and Pierre Baconnier. Comparison between five classifiers for automatic scoring of human sleep recordings. In *Classification and Clustering for Knowledge Discovery*, pages 113–127. Springer, 2005.

[7] Salih Güneş, Kemal Polat, and Şebnem Yosunkaya. Efficient sleep stage recognition system based on eeg signal using k-means clustering based feature weighting. *Expert Systems with Applications*, 37(12):7922–7928, 2010.

[8] Hui Wen Loh, Chui Ping Ooi, Jahmunah Vicnesh, Shu Lih Oh, Oliver Faust, Arkadiusz Gertych, and U Rajendra Acharya. Automated detection of sleep stages using deep learning techniques: A systematic review of the last decade (2010–2020). *Applied Sciences*, 10(24):8963, 2020.

[9] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.