university of groningen / faculty of science and engineering

Measuring Sleep in the Intensive Care Unit using Machine Learning

Bachelor's Project Thesis

Amber Schippers, a.m.schippers@student.rug.nl,
Supervisors: Dr M.A. Wiering & Dr S. Belur Nagaraj & L. Reinke & Prof Dr A.R. Absalom

Abstract: Sleep abnormalities occur frequently in Intensive Care Unit (ICU) patients, resulting in adverse effects on their health. It is important that their sleeping patterns are well understood in order to improve their sleep. Overnight electroencephalography (EEG) recordings are used to analyze sleeping patterns. The international norm is manual scoring by sleep experts following the American Academy of Sleep Medicine (AASM) criteria. When the severely disrupted EEG patterns of ICU patients were previously scored, low agreement between scorers was found. This suggests that the current standard for sleep analysis may not extend to ICU acquired data. Over a period of three years, 61 critically ill ICU patients have been monitored using EEG in the UMCG hospital in Groningen. Three machine learning algorithms (logistic regression, multi-class support vector machines and random forests) are trained on EEG recordings acquired from healthy subjects, and then tested on both the healthy recordings and EEG patterns acquired from ICU patients. The results show that the algorithms perform significantly worse on ICU subject data than they do on healthy subject data. This suggests that the current standard for sleep analysis is less suitable for the analysis of ICU patients' sleep.

1 Introduction

Research suggests that patients on the Intensive Care Unit (ICU) exhibit severely disrupted sleeping patterns compared to healthy individuals (Pisani, Friese, Gehlbach, Schwab, Weinhouse, and Jones, 2015). For instance, ICU patients experience 6.2 awakenings per hour and the majority of their sleep is spent in light sleep stages (Friese, Diaz-Arrastia, McBride, Frankel, and Gentilello, 2007). This has many consequences on one's health, including negative effects on physiology, the respiratory system, the cardiovascular system and the immunological system (Delaney, Van Haren, and Lopez, 2015). These sleep disturbances are also thought to contribute to delirium. This is a common complication in the ICU, causing issues regarding attention and consciousness, while also being associated with loss of physical function (Flaherty, 2007). All of these health implications caused by disturbed sleep may contribute to extended recovery times, which in turn may contribute to an increase in patient morbidity (Delaney et al., 2015). This is why it is important to analyze and understand sleep in order to improve it.

Causes of sleep disturbances in the ICU are thought to be medications, light, noise and patient care interactions amongst other things (Kamdar, Needham, and Collop, 2012).

Using overnight electroencephalography (EEG) the neural activity of the brain, observable as oscillating electrical potential on the scalp, can be recorded.

The current standard for sleep analysis is manual scoring of these EEG waves by sleep experts following the American Academy of Sleep Medicine (AASM) criteria. The EEG signals are split into time segments of 30 seconds, called epochs. Each epoch is assigned a sleep stage. The AASM defines five different stages, namely stage W (wakefulness), stage REM (rapid eye movement), stage N1 (non-REM1), stage N2 (non-REM2) and stage N3 (non-REM3) (Iber, Ancoli-Israel, Chesson, and Quan, 2007).

This process is called sleep staging. Because it is all done manually, it is tedious, expensive and

time consuming. It is used for things such as the diagnosis of sleep related disorders and getting to know more about the mechanisms and functions of sleep.

When EEG recordings from ICU patients were previously scored, it was found that there was low agreement between scorers compared to when healthy people's EEG recordings were scored. This suggests that there was more ambiguity and uncertainty, and that the current method for scoring might therefore be less suitable for ICU patients. This leads to the question; is the current standard for sleep scoring equally applicable to ICU patients? Based on the literature that suggests that ICU patients' sleep is different from healthy people's sleep, and that the agreement between scorers of ICU acquired EEG signals is low, the hypothesis is that the current standard of sleep analysis is less suitable for ICU patients.

In order to research this, three machine learning algorithms are utilized, namely logistic regression, multi-class support vector machines and random forests. The goal is to make these classifiers predict the sleep stage for each epoch. They will be trained on EEG signals acquired from healthy people that are scored by AASM standards. The classifiers get tested separately on the same healthy acquired data and data collected from ICU patients. They are then evaluated on their accuracy, F_1 score and Cohen's kappa against the annotations provided by the sleep experts. Then, comparisons are made between the results to determine whether the classifiers perform differently. If it is the case that the classifiers tested on EEG signals from ICU patients perform worse than the same classifiers tested on EEG signals from healthy people, the current standard might be less applicable to analyze the sleep of ICU patients.

Many machine learning algorithms for classifying sleep have already been developed. They are trained on data sets containing thousands of subjects and they have a level of performance comparable to human experts (Biswal, Sun, Goparaju, Westover, Sun, and Bianchi, 2018). The goal of this thesis, however, is not to make the best performing algorithm. Instead it is to find out how fruitful applying the current standard to critically ill patients is.

2 Methods

2.1 EEG datasets

The data was collected as part of a joint project of the UMCG hospital in Groningen with Philips Research in Eindhoven. Over a period of three years, 61 critically ill patients and 10 healthy subjects have been monitored in the ICU of the UMCG hospital. Seven channels were recorded (F3-A2, C3-A2, C4-A1, O1-A2, EOGL-A2, EOGR-A2, EMGL-EMGR) at a sampling frequency of 256 Hz. Only four of those channels were used (F3-A2, C3-A2, C4-A1, O1-A2), because those are the ones used by the sleep expert when sleep staging (Iber et al., 2007). The option to reduce the amount of channels to just the best performing one(s) was considered, but this was not done for the reasons that it would stray from the original sleep scoring method, the data would be smaller, the channels' errors were similar to each other and optimal performance was not the goal. The EEG recordings were then scored according to AASM criteria by two different sleep experts. Any epochs that did not receive a label were excluded from the data. The epochs were originally scored into six stages; W (wake), N1 (non-REM1), N2 (non-REM2), N3 (non-REM3), REM (rapid eye movement) and movement time. This last stage is merged with the 'wake' stage, because movement time is typically not used in the AASM classification criteria (Iber et al., 2007).

The datasets combined result in over one thousand hours of EEG data, which can be split into more than 125,000 epochs over four channels to train and test the classifiers on. From these epochs, 11,382 correspond to the healthy dataset and 113,910 belong to the ICU dataset. Figure 2.1 shows what a single epoch looks like.

The distribution of sleep stages in both datasets is shown in Figure 2.2. The data from healthy subjects were only recorded overnight, while the data from ICU subjects were recorded all day. This is done, because 40 to 50% of the total sleep time in an ICU takes place during daytime (Hilton, 1976; Aurell and Elmqvist, 1985). This explains the different distribution in the 'wake' stage. Another factor that contributes to unequal distributions between the datasets is that REM sleep is often suppressed in ICU patients (Kamdar et al., 2012)

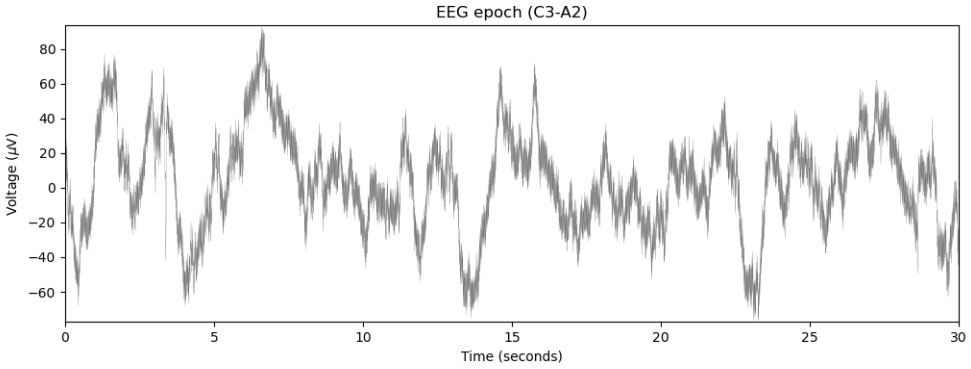

Figure 2.1: One epoch from the C3-A2 channel

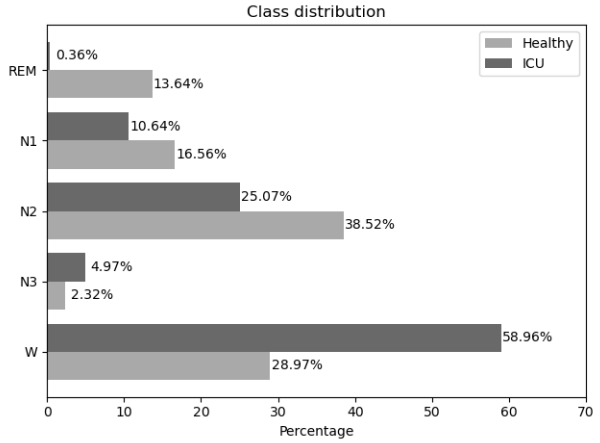

Figure 2.2: The class distributions in percentages of the datasets with five classes

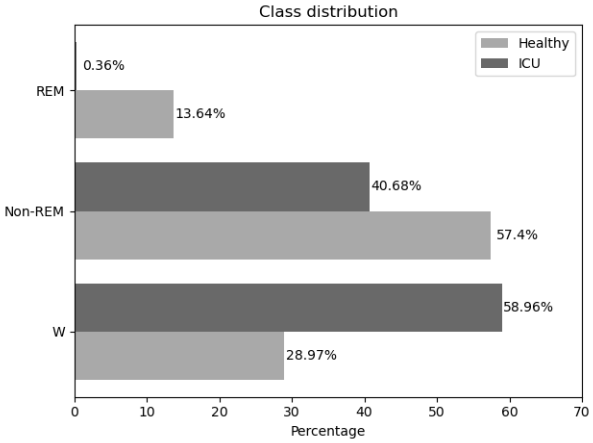

Figure 2.3: The class distributions in percentages of the datasets with three classes

2.2 Classification targets

Each epoch is labeled as one of five stages; W, N1, N2, N3 and REM. When training and testing the classifiers, it was found that many prediction errors were made in classifying the non-REM stages. In order to improve the performance of the algorithms, the non-REM stages were merged. This results in three target labels, namely W, non-REM and REM. Even though optimal performance is not the goal of this thesis, it is important for the algorithms to perform well as they need to resemble the human scorer in order to find out how they would perform on ICU acquired data. Hence, it is now formulated as a three-class classification problem. The new class distributions are shown in Figure 2.3.

2.3 Data preparation

The data need to be converted into features that the machine learning algorithms can use. First, a bandpass filter between 0.1 and 30 Hz is applied to get the required frequency range. In order to transform the data to the frequency domain, a Short-time Fourier transform is applied using Hanning window to minimize spectral leakage. Twenty-two spectral features are extracted for each epoch from the four different channels that are used. These features are the power in the (sub-)bands delta (p_δ), theta (p_θ), alpha (p_α), spindle (p_σ), lower beta ($p_{\beta L}$), and upper beta ($p_{\beta U}$) bands, total power (p_τ), normalized by total power - $\frac{p_\delta}{p_\tau}$, $\frac{p_\theta}{p_\tau}$, $\frac{p_\alpha}{p_\tau}$, $\frac{p_\sigma}{p_\tau}$, $\frac{p_{\beta L}}{p_\tau}$, $\frac{p_{\beta U}}{p_\tau}$, normalized by delta band - $\frac{p_\theta}{p_\delta}$, $\frac{p_\alpha}{p_\delta}$, $\frac{p_\sigma}{p_\delta}$,

$\frac{p_{\beta L}}{p_\delta}$, $\frac{p_{\beta U}}{p_\delta}$, and normalized by theta band $\frac{p_\alpha}{p_\theta}$, $\frac{p_\sigma}{p_\theta}$, $\frac{p_{\beta L}}{p_\theta}$, $\frac{p_{\beta U}}{p_\theta}$. With the delta band representing signals between 0.1 and 4 Hz, theta representing 4 to 8 Hz, alpha representing 8 to 12 Hz, spindle representing 12 to 16 Hz and beta representing 16 to 30 Hz.

For classifiers that use the distances between data, feature scaling is recommended. It normalizes the range of all features so that each feature contributes a proportional amount. Hence, scaling is applied to the features for the logistic regression and support vector machines classifiers which will be explained in the upcoming section.

2.4 Training algorithms

The data are classified in Python 3 using the scikit-learn package. As the datasets contain labels, supervised learning techniques are used to classify the epochs. Multiple classifiers are used to see how consistent the results are. The hyperparameters for the classes were chosen using randomized search cross validation, which randomly takes a set from the candidate hyperparameters, trains models for all of the sets made and compares their fitness via cross validation. The hyperparameters for each classifier are specified in the Appendix. To train and test on the EEG features collected from healthy subjects, leave-one-out cross-validation is used; nine out of ten subjects' sets were used for training and the remaining set was used for testing. The 'test set' was rotated amongst all healthy subjects' sets until they were all tested. This was not necessary for classifying the EEG features collected from ICU patients, because the training and testing sets do not overlap. The three following classification algorithms were used.

Logistic regression: This classifier works by using a sigmoid function to estimate probabilities of a certain class occurring for a certain linear combination of the independent variables or in this case, features (Theil, 1969). Each class is assigned a probability between 0 and 1, with the sum of all probabilities being 1. Multinomial logistic regression is used as there are three classes and they are not ordered. Instead of the loss function that would be used in binary logistic regression, the cross-entropy loss function is used.

Support vector machines: the objective of a binary support vector machine is to compute a hyperplane that divides the classes from each other by separating the data points belonging to one class from the other (Cortes and Vapnik, 1995). The aim when finding this hyperplane during training is to maximize the distance between the hyperplane and the nearest data points. This is done in order to have greater confidence when testing.

If your data is not linearly separable, a linear hyperplane will not do well in separating the data points into the correct classes. This is why different kernel functions can be chosen when computing the hyperplane. A radial basis function (RBF) kernel is used in this research as it gave a higher accuracy than a polynomial or sigmoid kernel.

Because there are three classes in this case, a multi-class support vector machine classifier is needed. In order to support three classes, a combination of three independent binary classification tasks is made. Using the one-vs-rest (OVR) approach, one class is trained against the 'rest'. A combination of these binary classifiers is then used.

Random forests: random forests is an ensemble learning method based on decision trees (Breiman, 2001). These decision trees work by splitting into different 'branches'. At each split, if the data of that particular test iteration meet a certain condition, it follows one branch and otherwise it will follow another. This repeats itself until a 'leaf' is found. Every leaf belongs to a certain class and if that leaf is reached, the decision tree will choose that class as its prediction. The majority vote out of all trees is chosen as the final classification of that test iteration of the algorithm. The trees have low correlation with each other so if one errs, the others will not necessarily err too.

These classifiers were chosen because they are common and applicable to this type of problem, yet they are still different in the way they work. Since four EEG channels were used, each tested epoch gets classified by the algorithms four times. The majority vote is chosen as the final classification for each epoch. If there is no single majority, a random class from the majority group is chosen.

2.5 Performance measures

Besides the classification accuracy (Eq. 2.1), three other performance metrics were utilized. This is done because of the so-called accuracy paradox; if there is a class imbalance and the classifier predicts the most common class very often, even wrongly so, the accuracy will still be high. As is visible in Figure 2.3, there is indeed a class imbalance. The accuracy is described as the proportion of correct classifications out of all classifications.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

The F_1 score (Eq. 2.4) is a performance metric that combines precision (Eq. 2.2) and recall (Eq. 2.3) in a harmonic mean. Precision asks the question; out of the predicted positive values, how many are actually positive? It is useful when the cost of false positives is high. Recall asks the question; out of all actual positives, how many were predicted as positive? It is useful when the cost of false negatives is high.

Because the F_1 score is a harmonic mean of these metrics, it is more suitable for an uneven class distribution than the classification accuracy is.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$F_1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.4}$$

Cohen's kappa (κ) (Eq. 2.5) was also determined. It is used to measure agreement among raters, in this case the classifier algorithm and the sleep staging expert. The number can vary between -1 and 1, with $\kappa = 1$ being perfect agreement among the raters, $\kappa = 0$ being no agreement among the raters other than what would be expected by chance and $\kappa = -1$ being that the raters are in complete disagreement.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.5}$$

Where:

- p_o = relative observed agreement among raters (or accuracy)
- p_e = hypothetical probability of chance agreement, calculated by the probabilities of the classes appearing

The accuracy and F_1 score metrics were both calculated individually for each class and in total. The totals were computed by taking the true positives, true negatives, false positives and false negatives of all classes summed up. Because multiple testing iterations had to be done for the healthy dataset, the means of the iterations were taken. In order to accurately record the F_1 score, each class needed to be present in the tested data or the precision and recall would result in zero. In the ICU dataset, not every person's data had a REM sleep annotation. To make sure the requirement for the F_1 score was still met, the data was randomly split into ten groups, with each group incorporating a data file that contained REM sleep annotations.

3 Results

The results of applying the performance metrics discussed in the previous section can be found in Table 3.1, Table 3.2 and Table 3.3. The tables contain the mean of each performance metric over all subjects of the individual datasets rounded to two decimals. The confusion matrices for each algorithm are shown in Figures 3.1-3.6. The accuracies and F_1 scores for individual classes can be found in Table 3.5.

For every classifier and every performance metric, the numbers for the classifiers tested on healthy subjects are higher than the ones for the classifiers tested on ICU subjects. The accuracies for classifiers tested on healthy subject data range from approximately 0.80 to 0.82, while the accuracies for classifiers tested on ICU subject data range from approximately 0.58 to 0.60. Similar differences are found in other performance metrics (0.72-0.74

against 0.42-0.44 for F_1 score, 0.61-0.64 against 0.25-0.27 for Cohen's kappa).

Cohen's kappa can be interpreted according to the range it falls in. This is shown in Table 3.4. The κ values for classifiers tested on healthy data fall in the lower end of the 'substantial' range, while the κ values for classifiers tested on ICU data fall in the 'fair' range.

The results of the individual classes found in Table 3.5 show that the REM class tends to be predicted relatively poorly. This is even more prominent in the F_1 score of the ICU data tested classifiers.

Three independent samples t-tests were conducted on both the classification accuracies and F_1 scores. The independent samples t-test was chosen because the samples were collected independently of each other and the difference between the means will be tested. For classification accuracy, the healthy-tested classifiers of logistic regression ($M_{LR} = 0.80$, $SD_{LR} = 0.09$), support vector machines ($M_{SVM} = 0.81$, $SD_{SVM} = 0.11$) and random forests ($M_{RF} = 0.82$, $SD_{RF} = 0.08$) were tested against the ICU-tested classifiers of logistic regression ($M_{LR} = 0.60$, $SD_{LR} = 0.13$), support vector machines ($M_{SVM} = 0.58$, $SD_{SVM} = 0.14$) and random forests ($M_{RF} = 0.56$, $SD_{RF} = 0.13$). These tests revealed a significant difference between testing on healthy data and testing on ICU data across all classifiers; logistic regression $t(69) = 4.1$, $p<.001$, support vector machines, $t(69) = 4.2$, $p<.001$, and random forests, $t(69) = 5.0$, $p<.001$.

For the F_1 score, another three independent samples t-test were performed. The healthy-tested classifiers of logistic regression ($M_{LR} = 0.72$, $SD_{LR} = 0.11$), support vector machines ($M_{SVM} = 0.74$, $SD_{SVM} = 0.14$) and random forests ($M_{RF} = 0.73$, $SD_{RF} = 0.12$) were tested against the ICU-tested classifiers of logistic regression ($M_{LR} = 0.42$, $SD_{LR} = 0.07$), support vector machines ($M_{SVM} = 0.41$, $SD_{SVM} = 0.08$) and random forests ($M_{RF} = 0.41$, $SD_{RF} = 0.08$). These tests revealed a significant difference between testing on healthy data and testing on ICU data across all classifiers; logistic regression $t(18) = 7.5$, $p<.001$, support vector machines, $t(18) = 7.0$, $p<.001$, and random forests, $t(18) = 7.5$, $p<.001$.

	Healthy	ICU
Accuracy	0.80	0.60
Precision	0.75	0.39
Recall	0.70	0.46
Macro F1-score	0.72	0.42
Cohen's kappa	0.61	0.27

Table 3.1: Logistic regression

	Healthy	ICU
Accuracy	0.81	0.58
Precision	0.76	0.39
Recall	0.72	0.44
Macro F1-score	0.74	0.41
Cohen's kappa	0.63	0.25

Table 3.2: Support vector machines

	Healthy	ICU
Accuracy	0.82	0.56
Precision	0.75	0.40
Recall	0.73	0.43
Macro F1-score	0.74	0.41
Cohen's kappa	0.64	0.26

Table 3.3: Random forests

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

Table 3.4: Interpretation of Cohen's kappa (Landis and Koch, 1977)

		Logistic regression		Random forests		Support vector machines	
		Healthy	ICU	Healthy	ICU	Healthy	ICU
Accuracy	Wake	0.81	0.53	0.87	0.63	0.84	0.62
	Non-REM	0.87	0.73	0.88	0.53	0.86	0.56
	REM	0.49	0.41	0.45	0.39	0.57	0.39
F_1 score	Wake	0.80	0.64	0.82	0.70	0.82	0.68
	Non-REM	0.85	0.60	0.87	0.52	0.85	0.53
	REM	0.51	0.03	0.50	0.02	0.56	0.02

Table 3.5: Individual classes

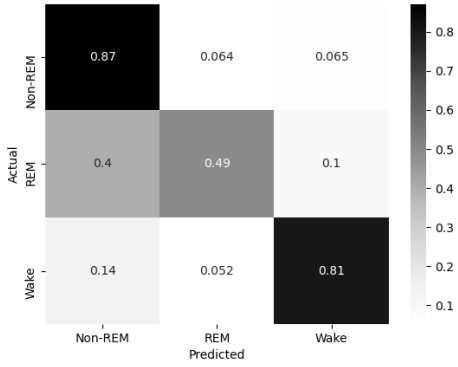

Figure 3.1: Confusion matrix for the logistic regression classifier tested on healthy subject data

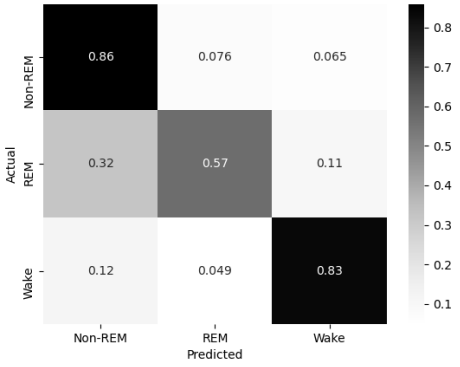

Figure 3.3: Confusion matrix for the support vector machines classifier tested on healthy subject data

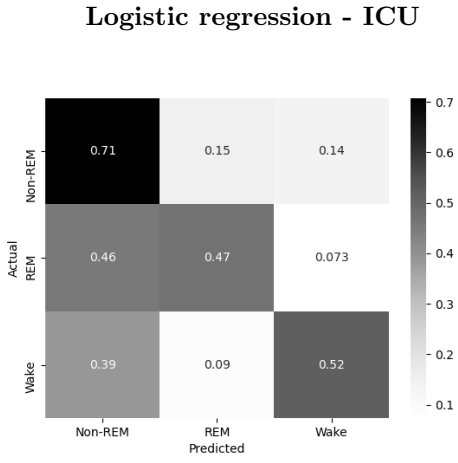

Figure 3.2: Confusion matrix for the logistic regression classifier tested on ICU subject data

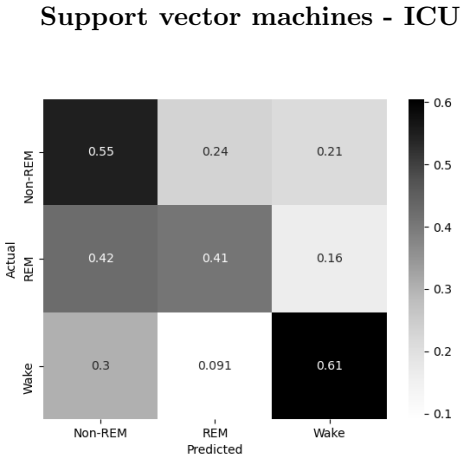

Figure 3.4: Confusion matrix for the support vector machines classifier tested on ICU subject data

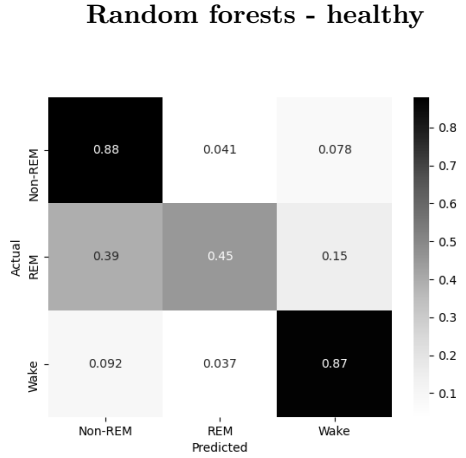

Figure 3.5: Confusion matrix for the random forests classifier tested on healthy subject data

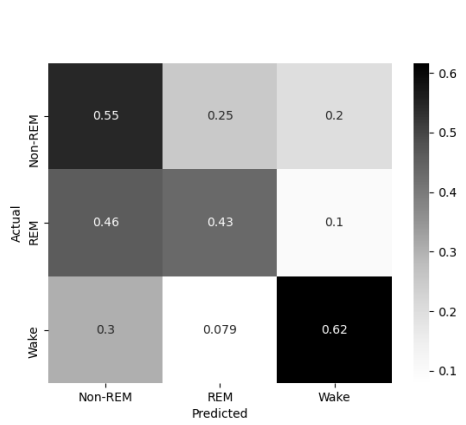

Figure 3.6: Confusion matrix for the random forests classifier tested on ICU subject data

4 Discussion

This study attempted to determine whether the current standard for sleep analysis is just as applicable to ICU patients as it is to healthy individuals. In order to research this, three machine learning classifiers trained on EEG signals collected from healthy people were tested on both EEG signals obtained from healthy people and EEG signals obtained from ICU patients. The result of comparing the performance of the classifiers when tested on one group against the same classifier when tested on the other, show that the classifiers tested on ICU data perform significantly worse.

These findings suggest two things. Firstly, as was already found in previous research, ICU patient's sleep is different from healthy people's sleep. Many factors contribute to this, a lot of research has been done and efforts have been made to improve it (Friese et al., 2007; Friese, 2008; Delaney et al., 2015).

Secondly, and more importantly, the classifiers and sleep experts have a significantly lower agreement with each other when classifying ICU sleep compared to classifying healthy sleep. This suggests uncertainty and ambiguity for ICU sleep analysis. This in turn implies that the current standard to analyze sleep, which is based on healthy people, is less applicable to ICU patients' sleep.

Something that stands out from the results is that they are similar across all algorithms. The reason for this is unclear. Something else that stands out is how poorly REM sleep is classified, both in healthy sleep and in ICU sleep. As can be seen in the class distributions pictured in Figure 2.3 in the methods section, REM sleep is the smallest class in the healthy/training set. Therefore, the algorithms have trained less on the REM class than others. The class performance of REM sleep is also much worse in the ICU data trained classifiers than the healthy data trained classifiers. This might have to do with the fact that ICU patients have a lack of REM sleep, because REM sleep duration increases as duration of sleep increases and ICU patients sleep for short periods of time (Aurell and Elmqvist, 1985; Delaney et al., 2015). As the REM sleep stages are so short, they might be difficult to classify.

In hindsight it would have been good to make use of a neural network in addition to the classifiers. This method is often used for sleep classification for several reasons; compared to several other conventional classifiers, it is robust (Schaltenbrand, Lengelle, and Macher, 1993), good with new data (Robert, Guilpin, and Limoge, 1998) and it performs well in discriminating the REM sleep stage (Robert, Guilpin, and Limoge, 1997) which the classifiers in this research did not perform well on. It might therefore have produced different results. Alternatively, more REM stage training data could have been provided so that the classifiers would perform better on this particular sleep stage.

For future research, it would be interesting to

find out how ICU sleep should be analyzed instead. As clarified in the introduction, this is important, as when a good method for sleep analysis for ICU patients is found, a way to improve their sleep might be found.

As was also clarified in the introduction, previous studies have created machine learning algorithms for sleep classification that perform as well as human experts do (Biswal et al., 2018). It might be interesting to research how well these classifiers would perform when trained and tested on EEG signals from ICU patients. If this method would work well, it would likely also be less time-consuming and expensive than manual sleep staging.

However, this method does come with an important assumption, namely that different ICU patients' sleep can all be classified in the same way. This leads to another interesting question; can all ICU patients be put into a single group when it comes to sleep classification? There might be differences depending on age, amount of time since they have been admitted to the ICU, type of illness or degree of illness.

References

J. Aurell and D. Elmqvist. Sleep in the surgical intensive care unit: continuous polygraphic recording of sleep in nine patients receiving postoperative care. *Br Med J (Clin Res Ed)*, 290(6474): 1029–1032, 1985.

S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi. Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association*, 25(12):1643–1650, 2018.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

L. J. Delaney, F. Van Haren, and V. Lopez. Sleeping on a problem: the impact of sleep disturbance on intensive care patients-a clinical review. *Annals of intensive care*, 5(1):1–10, 2015.

J. H. Flaherty. Delirium. In James E. Birren, editor, *Encyclopedia of Gerontology (Second Edition)*, pages 359–368. Elsevier Science, 2007.

R. S. Friese. Sleep and recovery from critical illness and injury: a review of theory, current practice, and future directions. *Critical care medicine*, 36(3):697–705, 2008.

R. S. Friese, R. Diaz-Arrastia, D. McBride, H. Frankel, and L. M. Gentilello. Quantity and quality of sleep in the surgical intensive care unit: are our patients sleeping? *Journal of Trauma and Acute Care Surgery*, 63(6):1210–1214, 2007.

B. A. Hilton. Quantity and quality of patients' sleep and sleep-disturbing factors in a respiratory intensive care unit. *Journal of advanced nursing*, 1(6):453–468, 1976.

C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. *Westchester, IL: American academy of sleep medicine*, 1, 2007.

B. B. Kamdar, D. M. Needham, and N. A. Collop. Sleep deprivation in critical illness: its role in physical and psychological recovery. *Journal of intensive care medicine*, 27(2):97–111, 2012.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

M. A. Pisani, R. S. Friese, B. K. Gehlbach, R. J. Schwab, G. L. Weinhouse, and S. F. Jones. Sleep in the intensive care unit. *American journal of respiratory and critical care medicine*, 191(7): 731–738, 2015.

C. Robert, C. Guilpin, and A. Limoge. Comparison between conventional and neural network classifiers for rat sleep-wake stage discrimination. *Neuropsychobiology*, 35(4):221–225, 1997.

C. Robert, C. Guilpin, and A. Limoge. Review of neural network applications in sleep research. *Journal of Neuroscience methods*, 79(2):187–193, 1998.

N. Schaltenbrand, R. Lengelle, and J. P. Macher. Neural network model: application to automatic

analysis of human sleep. *Computers and Biomedical Research*, 26(2):157–171, 1993.

H. Theil. A multinomial extension of the linear logit model. *International economic review*, 10(3): 251–259, 1969.

A Appendix

The hyperparameters used for the classifiers can be found in the following tables.

parameter	value
penalty	l1
dual	False
tol	1e-4
C	10
fit_intercept	True
intercept_scaling	1
class_weight	None
random_state	None
solver	saga
max_iter	10000
multi_class	multinomial
verbose	0
warm_start	False
n_jobs	None
l1_ratio	None

Table A.1: Hyperparameters for logistic regression

parameter	value
C	1000
kernel	rbf
degree	3
gamma	0.001
coef0	0.0
shrinking	True
probability	False
tol	0.001
cache_size	200
class_weight	None
verbose	False
max_iter	-1
decision_function_shape	ovr
break_ties	False
random_state	None

Table A.2: Hyperparameters for support vector machines

parameter	value
n_estimators	800
criterion	gini
max_depth	80
min_samples_split	10
min_samples_leaf	4
min_weight_fraction_leaf	0.0
max_features	auto
max_leaf_nodes	None
min_impurity_decrease	0.0
min_impurity_split	None
bootstrap	True
oob_score	False
n_jobs	None
random_state	None
verbose	0
warm_start	False
class_weight	None
ccp_alpha	0.0
max_samples	None

Table A.3: Hyperparameters for random forests