Individualized Sleep Stage Classification from Cardiorespiratory Features

1st Miriam Goldammer
Institute of Biomedical Engineering
Faculty of Electrical and Computer Engineering
TU Dresden
Dresden, Germany
miriam.goldammer@tu-dresden.de

2nd Lucas Weber
Institute of Biomedical Engineering
Faculty of Electrical and Computer Engineering
TU Dresden
Dresden, Germany

3rd Hagen Malberg
Institute of Biomedical Engineering
Faculty of Electrical and Computer Engineering
TU Dresden
Dresden, Germany

4th Sebastian Zaunseder
Department of Information Technology
University of Applied Sciences and Arts
Dortmund, Germany
sebastian.zaunseder@fh-dortmund.de

***Abstract*—Modern patient care aims for individualized solutions. Current machine learning techniques, in general and in the medical domain, typically incorporate big amounts of data. In fact, more data contributes to the generalizability of said techniques. However, it might conflict with the desire for individualized solutions. Our works aim at the implementation of individual solutions based on machine learning techniques. Within this contribution, we investigate the potential benefit of individualized classifiers in the context of automatic sleep staging using cardiorespiratory features.**

To that end, we performed sleep stage classification using 237 records of the Sleep Heart Health Study. For each patient, we trained an ensemble classifier that is based on a subset of the available patients. Such subsets of varying size were chosen by a modified version of sequential forward floating selection. Our results show that the individualized classifier improves classification compared to a classifier that uses all available patients by 30% (improvement in Cohen's kappa coefficient (κ) of 0.15 from 0.46 to 0.61). On average the subset used for training thereby includes five patients.

The presented contribution clearly depicts the potential of an individualized classification approach. Based on the current results, future works will try to establish metrics that can identify the most appropriate training subset in an unsupervised way.

***Index Terms*—sleep stage classification, automatic sleep staging, individualized classifier, heart rate variability, respiration**

I. Introduction

To account for individual patients' characteristics and patients' backgrounds, modern patient care aims for individualized solutions. Individualization equally affects both, diagnosis and therapy. Accurate classification or prediction, respectively, are essential tasks for diagnosis and to guide therapy. Machine learning techniques become more and more popular for such tasks even in the medical domain. A basic taxonomy on machine learning methods distinguishes model-based approaches and instance-based approaches. Model-based approaches use training instances to construct an explicit description of a target function. This description maps data to a target value, i.e. assigns a class to a query instance. Instance-based learning methods evaluate the similarity between training instances and the query instance to find the target function value for the query instance [1].

Though instance-based methods inherently incorporate individualization by their function principle, in many cases model-based approaches outperform instance-based methods in terms of classification accuracy and computational efficiency [2]. Moreover, both approaches typically exploit large datasets to estimate the target function or do comparisons, respectively. In fact, more training data allows for a better generalization and, consequently, leads to better results regarding the mean classification accuracy. However, concerning an individual patient, a specifically trained classifier might lead to better results [3] and the common concept of large training sets opposes the idea of individualization.

This paper investigates the potential of adjusting the training data for model-based classifiers to yield individualized classifiers. Figure 1 provides a problem formulation. As an exemplary application, this work directs at sleep stage classification. Conventionally, sleep staging is done by medical experts based on polysomnographic data, most importantly based on the electroencephalogram, electromyogram and electrooculogram [4]. The recording of such signals requires electrodes on the scalp and face, which substantially interfere with patients' comfort and normal sleep. Current research focuses on alternative ways for sleep stage classification. Among them, heart rate variability (HRV) and respiration were used as a basis for sleep staging [5]–[8]. Such signals can be easily acquired, even in a non-contact way [9], [10]. Sleep stage classification based on HRV and respiration most often makes use of the common training concept, namely constructing a large database and training a generalized classification model

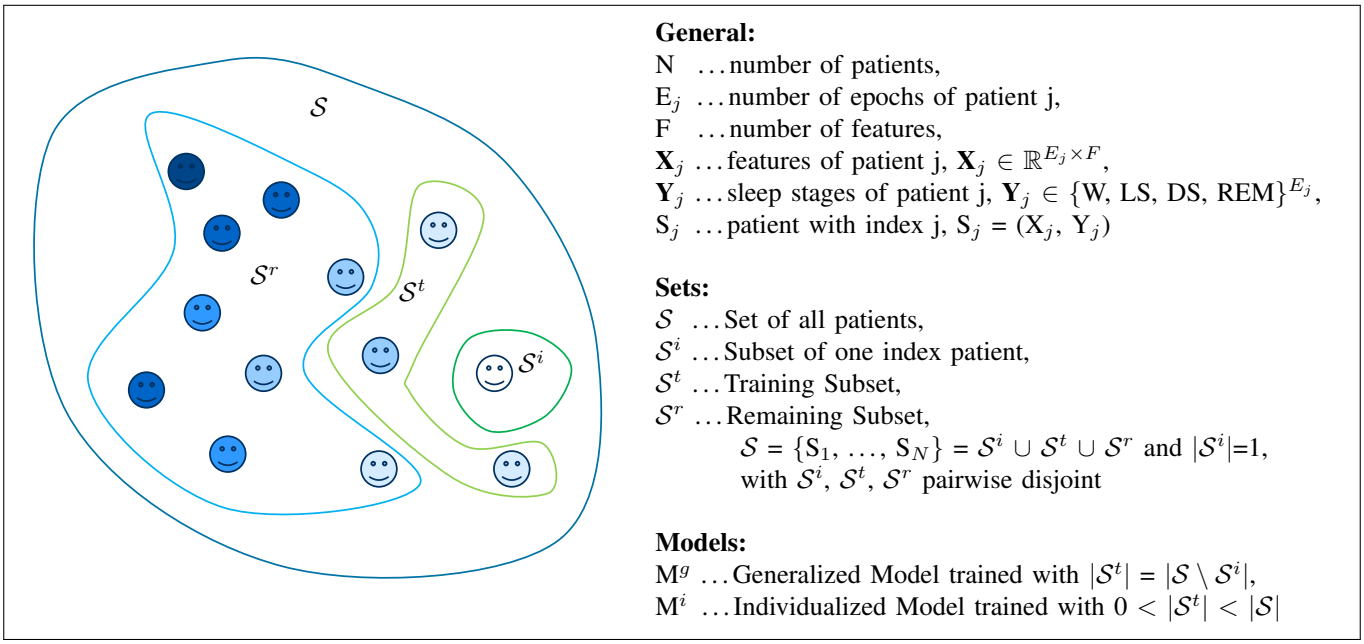

Fig. 1. Explanation of subsets used for classification. This contribution investigates if there is a subset $\mathcal{S}^t$ that optimizes the classification for a single patient (called index patient).

on it. To evaluate the potential of an individualized classifier, we compare this common concept of training to an approach which forms an individual training subset by a sequential floating search.

The remainder of the work is structured as follows. Section II details the used data and its processing. The latter comprises preprocessing, feature extraction, classification and evaluation including the method for training subset selection. In section III we give quantitative results. Section IV discusses the results and outlines the meaning of our findings regarding the fully automated construction of individualized classifiers. Section V closes with our conclusions.

II. Methods

A. Data

We used data from the first part of the Sleep Heart Health Study (SHHS) [11]. The database contains 5804 subjects. We removed patients suffering from acute cardiovascular diseases as well as those showing an Apnea-Hypopnea Index greater than 5%. Such criteria left 263 recordings. Of those, 26 were not usable because of a poor QRS detection performance (see next section for details), leaving 237 recordings for the analysis.

Each recording features polysomnographic data together with reference sleep stage annotations for each epoch of 30 s according to Rechtschaffen and Kales. We combined sleep stages S1 and S2 to light sleep and S3 and S4 to deep sleep, leading to 4 classes, namely wake (W), light sleep (LS), deep sleep (DS) and REM sleep (REM). The first and the last 5 minutes of all records were discarded because they often contained strongly corrupted signals. Figure 2 shows the resulting class distribution over all 237 recordings.

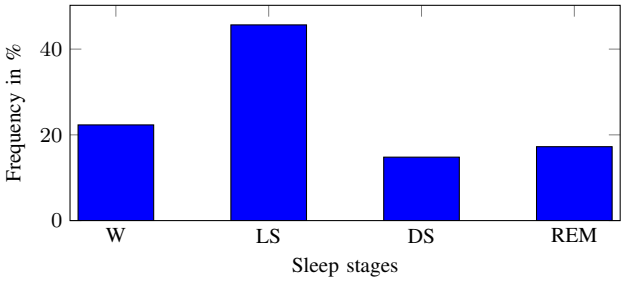

Fig. 2. Distribution of sleep stages in the data.

B. Signal processing

Signal processing covered the extraction of (filtered) beat-to-beat (RR) and breath-to-breath (BB) interval time series and feature extraction from such time series.

Time series extraction: To extract RR intervals, we detected QRS complexes from the ECG using the QRS detector by Afonso et al. [12] in its implementation from [13]. To account for outliers in the RR interval series introduced by wrong detections or arrhythmic events, we filtered the RR intervals by an iterative algorithm according to [14]. Briefly, the iterative method always considers three adjacent RR intervals, calculates a regression between the two outer ones and estimates the central one. The central RR interval is regarded as an outlier if its deviation from the estimated value exceeds 15%. Within each iteration, the algorithm considers all available RR intervals and removes outliers before the next iteration. The algorithm stops if no more outliers are present. The filtered RR series served as input to the calculation of HRV features. We discarded records if more than 25% of

initially found RR intervals were removed by the adaptive filter. If the number of filtered RR intervals was less than $d \cdot \frac{70\,\mathrm{bpm}}{2}$ or more than $d \cdot \frac{70\,\mathrm{bpm}}{0.5}$ (with d the length of the recording in seconds), the respective record was also discarded. Both criteria account for irregular RR intervals, which are likely to hinder a meaningful HRV analysis. In total 26 of 263 recordings were removed based on that criteria.

To extract the BB intervals, single respiratory cycles were detected in the respiration signal by the method `respdetect` provided in [13]. The BB interval time series served directly as input to the feature extraction without further processing or exclusion criteria.

Feature extraction: Our feature extraction is based on a sliding window approach at a step width of 30 s between analysis time instants t_a (which allows to score sleep stages each 30 s). For each t_a, we considered the filtered RR intervals and BB intervals over windows of length t_w 30 s, 90 s and 180 s centered on the current time instant t_a[1].

From each window, we derived four common HRV features [15], namely the average RR interval, the normalized high and low frequency powers (P_{LFnu}, P_{HFnu}) as well as the power in the very low frequency range P_{VLF} normalized to the average RR. The calculation was done using the BiosigToolbox [13]. We employed a regression model to yield an estimate of the power spectral density and determined said features from it. In addition to those features, which we further denote as *raw features*, we calculated so-called Δ-features for each raw feature. These Δ-features refer to the difference between a feature's value at the time t_a and its value at time $t_a - \frac{t_w}{2}$ or $t_a - t_w$, respectively. Δ-features thus capture linear trends in the used features. If the number of detected RR intervals in one of the windows belonging to a time instant t_a was less than $t_w \cdot \frac{70\,\mathrm{bpm}}{2}$ or larger than $t_w \cdot \frac{70\,\mathrm{bpm}}{0.5}$ (with t_w in seconds), no HRV features were calculated but the last calculated features were reused.

The respiration was assessed by two raw features, namely the mean respiratory interval and the standard deviation of respiratory cycles normalized to the mean interval. Again, we calculated Δ-features the same way we did for HRV features. If the number of detected respiratory cycles within a window fell below four, no respiration features were calculated but the last calculated features were reused.

As a last feature, we used the time index of the current window to indicate the time of the night. Overall, the procedure yields 49 features (4 raw HRV features per time window, 3 time windows, 2 Δ-features for each raw feature; 2 raw respiratory features per time window, 2 time windows, 2 Δ-features for each raw feature; 1 time index feature).

C. Used Classifiers

As classifiers, we used Random Forests (RF) and k-Nearest Neighbors (kNN). RF is a model-based approach that uses decision trees to create an ensemble classifier. It combines bagging and feature subset selection in order to yield diversity over the ensemble. Our RF was trained with 30 trees by `fitcensemble` from MATLAB's Statistics and Machine Learning Toolbox. The k-Nearest Neighbors algorithm is an instance-based classifier. It compares the epoch to be classified to all single epochs in the training set. It assigns a class based on the k most similar epochs' majority class. We used the kNN implementation `fitcknn` by MATLAB's Statistics and Machine Learning Toolbox and evaluated the performance for different k.

[1]Respiration is considered over 90 s and 180 s only

D. Evaluation strategy

To evaluate the performance of individualized models, we performed a leave one out cross validation, so that each patient S_j was index patient once, i.e. member of the subset $\mathcal{S}^i$ (see figure 1 for the used notation). For each subject, we computed a generalized model using RF and kNN (denoted as gRF and gkNN) and the individualized model using RF (denoted as iRF). The comparison was done based on Cohen's kappa coefficient (κ).

Generalized Model: For the generalized model, all subjects except the index patient were used for training, i. e. $\mathcal{S}^r = \emptyset$ and $\mathcal{S}^t = \mathcal{S} \setminus \mathcal{S}^i$.

Individualized Model: The individualized model assumes that not all subjects in $\mathcal{S}$ hold valuable information to establish a model for an index patient. Accordingly, there should exist an individualized training subset $\mathcal{S}^t$ that significantly improves classification quality compared to the generalized model. To prove this hypothesis is the central goal of this contribution.

We sought the individualized training subset by a modified sequential forward floating search (SFFS) approach as described in algorithm 1. SFFS typically is used to select a subset of features which optimize the classification accuracy. To that end, SFFS iteratively adds and removes features while optimizing a criterion function. From this wrapper function, the concept of feature selection can readily be transferred to select a subset of subjects. In contrast to the well known SFFS by Pudil et al. [16], we did not try to find an optimal subset for each subset size but applied a heuristic to yield the locally best subset with fewest training subjects. Therefore, the algorithm only adds or removes subjects that improve the classification result independent of the subset size, i. e. always compared to the last best subset. The search stops as soon as adding and removing any subject does not improve the classification result. For computational reasons, we restricted the maximum number of subjects to ten, i.e. the search was also stopped when the training subset comprised ten patients.

Evaluation measure: We evaluate the results with Cohen's κ [17] for each index patient. κ is calculated as:

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

where p_o denotes the proportion of correct classifications and p_e denotes the proportion of correct classifications that is expected by chance. κ of 0 is the expected result for guessing the classes without previous knowledge, whereas κ of 1 constitutes a perfect classifier as described in Table I.

TABLE I
INTERPRETATION OF κ [18]

κ	Classification quality
< 0	worse than chance
$0.01 - 0.20$	slight
$0.21 - 0.40$	fair
$0.41 - 0.60$	moderate
$0.61 - 0.80$	substantial
$0.81 - 0.99$	almost perfect

Algorithm 1 SFFS, notation according to figure 1

```
K = NaN( |S|, 1 )
for S_n in S do
    S^i = {S_n}
    S^t = ∅
    S^r = S \ {S_n}
    kappa_max = 0
    R = NaN( |S^r|, 1 )
    for S_j in S^r do
        S^temp = S^t ∪ {S_j}
        M = model( S^temp )
        R_j = kappa( M.predict( S^i ) )
    end for
    jm = index of max(R)
    if R_jm > kappa_max then
        kappa_max = R_jm
        S = S_j ∈ S^r
        S^t = S^t ∪ {S}
        S^r = S^r \ {S}
        while 1 do
            R = NaN( |S^t|, 1 )
            for S_j in S^t do
                S^temp = S^t \ {S_j}
                M = model( S^temp )
                R_j = kappa( M.predict( S^i ) )
            end for
            jm = index of max(R)
            if R_jm > kappa_max then
                kappa_max = R_jm
                S = S_j ∈ S^t
                S^t = S^t \ {S}
                S^r = S^r ∪ {S}
            else
                break
            end if
        end while
    else
        K_n = kappa_max
    end if
end for
```

III. RESULTS

Figure 3 shows the results for the different classifiers. For gRF, a mean κ of 0.46 is achieved with a standard deviation of 0.11. For gkNN, k equal to 25 (see figure 4) yields the best

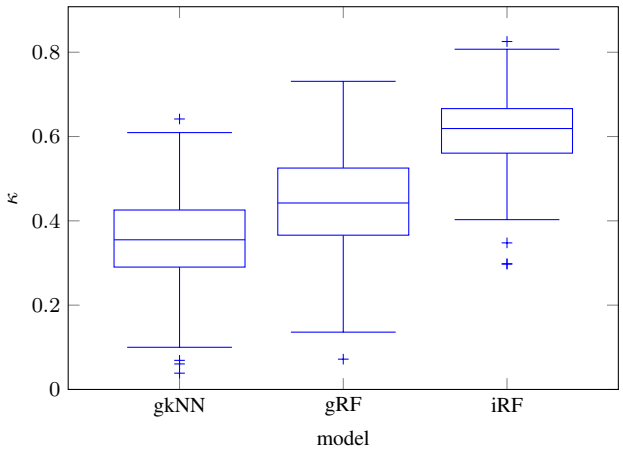

Fig. 3. κ distribution for different classifiers.

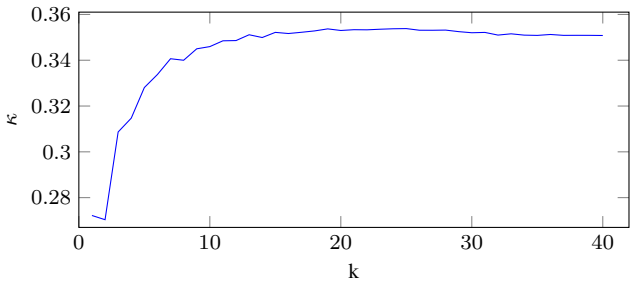

Fig. 4. Mean κ of gkNN for different number of epochs k.

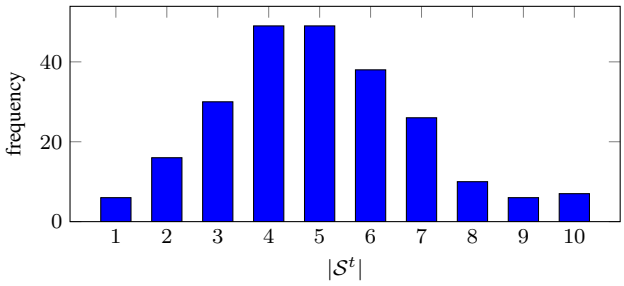

Fig. 5. Distribution of found training subset sizes

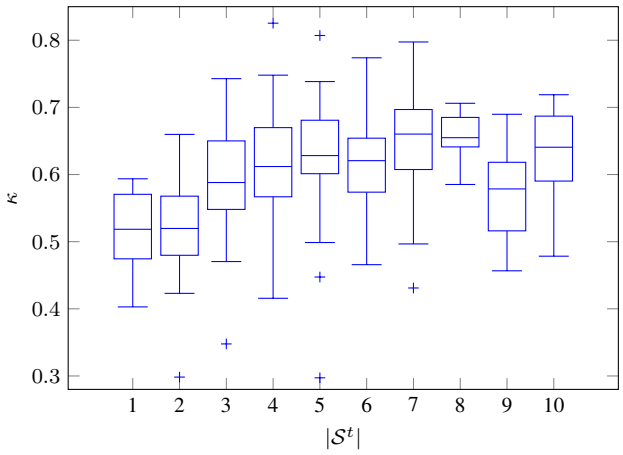

Fig. 6. κ results against size of training subset for iRF.

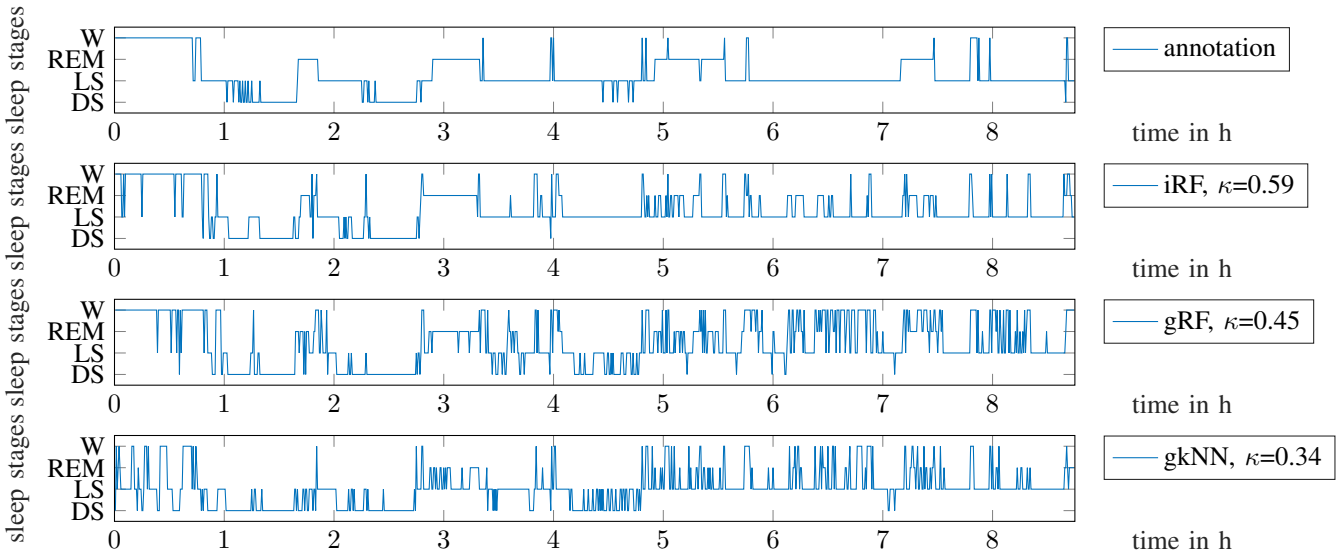

Fig. 7. Hypnogram of patient 193 from different classifiers.

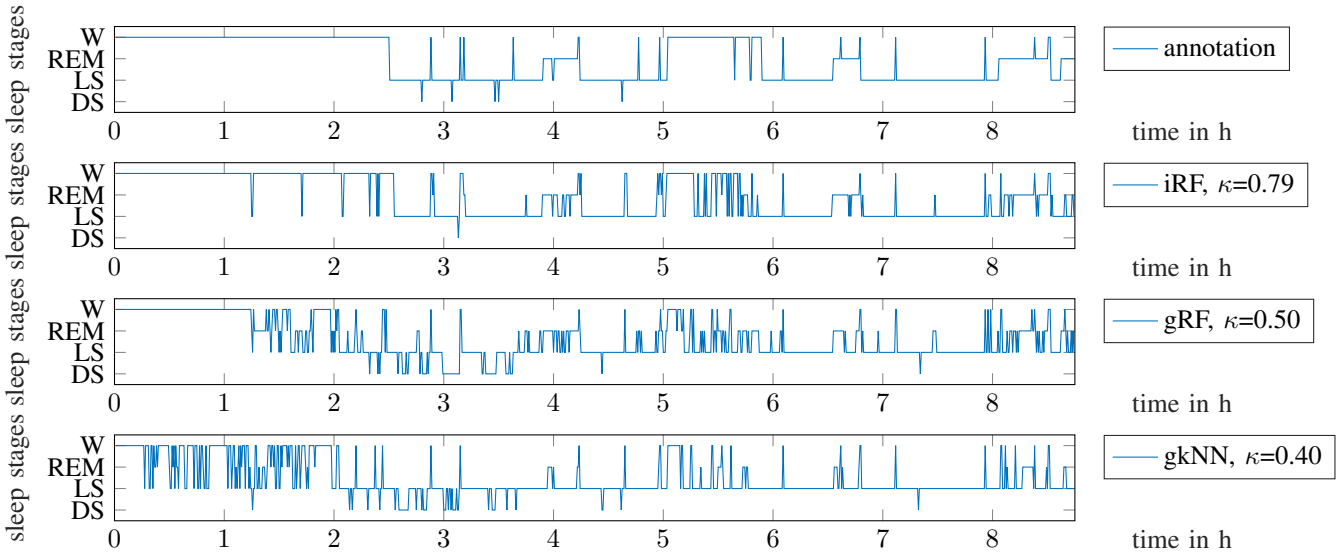

Fig. 8. Hypnogram of patient 35 from different classifiers.

mean κ of 0.35 with a standard deviation of 0.11. In contrast, the resulting mean κ of iRF is 0.61 with a standard deviation of 0.09.

While $|\mathcal{S}^t|$ is 236 for both gRF and gkNN, the mean training subset size of iRF is only five patients (see figure 5).

Figure 6 shows that for $|\mathcal{S}^t|$ between 3 and 10, the median κ is consistently greater than 0.57.

The four hypnograms in figure 7 are for the same night of one patient. The record exemplarily illustrates the effect of an increase in κ by 0.15, which iRF yields averaged over all records. In this example, the iRF classifier is the most stable in detecting the light sleep from hours 5.5 to 7 and the only one not to wrongly detect deep sleep from hour 4 to 5.

An even wider range of different classification qualities is displayed in figure 8 and underlines the potential of the proposed approach. A κ of 0.79 is an exceptionally good result for the iRF. Even though there are some problems with detecting the REM phase at hour 4 and the wake phase from hours 5 to 6, the similarity between the iRF hypnogram and the annotation is striking.

IV. Discussion

Interpretation: Our results show that an individualized training subset can distinctly improve classification results over a larger training set (we yield an improvement of more than 30%). We found subset sizes smaller than ten being large enough to meet the stopping criterion and improve the results for almost all patients. As the results for $|\mathcal{S}^t|$ of size one

and two are apparently worse than for larger subset sizes (see figure 6), we assume that for such patients no more appropriate training patients were available.

Even in terms of absolute values, our results are remarkable. For gRF, the results comply well with other works that make use of HRV and respiration to classify sleep stages. Such works also yield κ in the range up to 0.5 [5]–[8] (note that HRV and respiration cannot yield an accuracy as high as electroencephalographic features can yield). iRF, however, yields a classification accuracy beyond what is obtained by other works. Moreover, regarding the absolute classification accuracy, we assume the individualized classifier to bear more potential: compared to other works in the field, the used database of 237 patients is large. However, the individualized classifier relies on identifying matching training subjects. There are no design rules on minimum database sizes, but a larger database naturally will provide a better base for the individualization as pursued here. Generalized classifiers, in turn, are likely to saturate in their classification accuracy so that increasing the database size will have a smaller effect.

Limitations: In a way, we overfit the training subset to each index patient, because we use classification accuracy on the index patients to guide the subset selection. Consequently, the found κ might overestimate the true obtainable performance. However, first, we do not use the index patients features, which would mean a higher risk of overfitting. Second, the found improvement of 30% compared to the general classifier is substantial. We doubt that overfitting would allow such fundamental improvement. Third, we modified the original SFFS algorithm in order to stop the selection immediately if no single subject improves the results. According to SFFS's original description, adding subjects even if they do not improve the results instantaneously would be possible, which would promote overfitting. Fourth, frequent subset sizes of 1 and 2 could hint at overfitting. However, as figure 5 shows, such subset sizes rarely occur and yield poor results. The found sizes, in turn, seem to be of reasonable size. Consequently, our results show that individualization is possible from a small patient set, and without training data from the individual itself.

V. Conclusion

The presented work proved the potential of individualized classifiers to improve classification by using small training subsets rather than all available data. However, at the current state our methods can not be applied to test data in an unsupervised fashion, i. e. this contribution focuses on the principle benefits of an individualization. The automatic, unsupervised creation of such training subsets for previously unknown patients is the final goal of our work.

Future work: So far, we did not further analyze the training subsets chosen by SFFS. This is a main task for our future works. We will use the found training subsets to train neighborhood algorithms, which identify individualized training subsets in an unsupervised manner. To that end, we plan to analyze the relationship and similarity between the index patient, its training subset and remaining patients concerning features, distribution of sleep stages and demographics.

References

[1] T. Mitchell, *Machine Learning (Mcgraw-Hill International Edit)*. McGraw-Hill Education (ISE Editions), 1997.

[2] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, no. 3, pp. 249–268, 2007. [Online]. Available: http://www.informatica.si/index.php/informatica/article/view/148/140

[3] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric," *PLoS One*, vol. 10, no. 5, pp. 1–13, 2015.

[4] "The AASM Manual for the Scoring of Sleep and Associated Events," Tech. Rep., 2007.

[5] M. Aktaruzzaman, M. Migliorini, M. Tenhunen, S. L. Himanen, A. M. Bianchi, and R. Sassi, "The addition of entropy-based regularity parameters improves sleep stage classification based on heart rate variability," *Med. Biol. Eng. Comput.*, vol. 53, no. 5, pp. 415–425, 2015.

[6] F. Ebrahimi, S. K. Setarehdan, J. Ayala-Moyeda, and H. Nazeran, "Automatic sleep staging using empirical mode decomposition, discrete wavelet transform, time-domain, and nonlinear dynamics features of heart rate variability signals," *Comput. Methods Programs Biomed.*, vol. 112, no. 1, pp. 47–57, oct 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23895941

[7] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort." *Physiol. Meas.*, vol. 36, no. 10, pp. 2027–40, oct 2015. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/26289580

[8] S. J. Redmond and C. Heneghan, "Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, 2006.

[9] S. Zaunseder, A. Henning, D. Wedekind, A. Trumpp, and H. Malberg, "Unobtrusive acquisition of cardiorespiratory signals," *Somnologie*, vol. 21, no. 2, pp. 93–100, jun 2017. [Online]. Available: http://link.springer.com/10.1007/s11818-017-0112-x

[10] S. Zaunseder, A. Trumpp, D. Wedekind, and H. Malberg, "Cardiovascular assessment by imaging photoplethysmography - a review." *Biomed. Tech. (Berl).*, vol. 63, no. 5, pp. 617–634, oct 2018. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29897880

[11] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl, "The Sleep Heart Health Study: Design, Rationale, and Methods," *Sleep*, vol. 20, no. 12, 1997. [Online]. Available: https://academic.oup.com/sleep/article/20/12/1077/2749934/The-Sleep-Heart-Health-Study-Design-Rationale-and

[12] V. X. Afonso, W. J. Tompkins, T. Q. Nguyen, and S. Luo, "ECG beat detection using filter banks." *IEEE Trans. Biomed. Eng.*, vol. 46, no. 2, pp. 192–202, feb 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9932341

[13] C. Vidaurre, T. H. Sander, and A. Schlögl, "BioSig: The Free and Open Source Software Library for Biomedical Signal Processing." *Comput. Intell. Neurosci.*, vol. 2011, pp. 935 364–935 376, 2011. [Online]. Available: https://doi.org/10.1155/2011/935364

[14] D. Wichterle, J. Simek, M. T. La Rovere, P. J. Schwartz, A. J. Camm, and M. Malik, "Prevalent low-frequency oscillation of heart rate: novel predictor of mortality after myocardial infarction." *Circulation*, vol. 110, no. 10, pp. 1183–90, sep 2004. [Online]. Available: http://dx.doi.org/10.1161/01.CIR.0000140765.71014.1C

[15] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology., "Heart rate variability: standards of measurement, physiological interpretation and clinical use." *Circulation*, vol. 93, no. 5, pp. 1043–65, mar 1996. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/8598068

[16] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[17] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, apr 1960. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/001316446002000104

[18] A. Viera and J. Garrett, "Understanding interobserver agreement: The kappa statistic," *Family Medicine*, vol. 37, no. 5, pp. 360–363, 5 2005.