PerSleep: A Visual Analytics Approach for Performance Assessment of Sleep Staging Models

H. S. Garcia Caballero[1] , A. Corvò[1] , F. van Meulen[1] , P. Fonseca[2] , S. Overeem[1,3] , J. J. van Wijk[1] and M. A. Westenberg[1]

[1]Eindhoven University of Technology, The Netherlands
[2]Philips Research, The Netherlands
[3]Sleep Medicine Center Kempenhaeghe, The Netherlands

Abstract
Machine learning is becoming increasingly popular in the medical domain. In the near future, clinicians expect predictive models to support daily tasks such as diagnosis and prognostic analysis. For this reason, it is utterly important to evaluate and compare the performance of such models so that clinicians can safely rely on them. In this paper, we focus on sleep staging wherein machine learning models can be used to automate or support sleep scoring. Evaluation of these models is complex because sleep is a natural process, which varies among patients. For adoption in clinical routine, it is important to understand how the models perform for different groups of patients. Moreover, models can be trained to recognize different characteristics in the data, and model developers need to understand why and how performance of the different models varies. To address these challenges, we present a visual analytics approach to evaluate the performance of predictive models on sleep staging and to help experts better understand these models with respect to patient data (e.g., conditions, medication, etc.). We illustrate the effectiveness of our approach by comparing multiple models trained on real-world sleep staging data with experts.

CCS Concepts
• *Human-centered computing → Visual analytics;*

1. Introduction

Machine Learning (ML) has increased in popularity in the medical domain [NCB19] due to its success in tasks such as segmentation, classification and anomaly detection. One example is sleep medicine, where models have been proposed to score sleep stages and support sleep diagnosis [SDWG17, SOO*18, PAC*19]. These advancements bring opportunities to automate such time-consuming [IRV14], tedious and subjective tasks typically conducted by specialists. Assuring a good performance of such models is crucial for somnologists to safely rely on them.

Generally, the evaluation of ML models for sleep staging is complex for three reasons. First, sleep is a natural process that runs and evolves over *time*. When predictions are produced by a model, errors can occur at different periods of the sleep. The location of these errors is crucial because it can bias the diagnosis of sleep diseases (e.g., non-REM parasomnias usually occur in the first third of the night). Second, common statistics (e.g., accuracy, F-measure, etc.) only provide a coarse-grained perspective of the performance [ZWM*18]. A closer look at predictions and patients is necessary to better evaluate ML models. Finally, wrong predictions can suggest that fragmented sleep occurs. This can potentially be misinterpreted as a sleep disorder. Inherently, all these problems can be different across *groups of patients*. Sleep varies among patients due to

physiological reasons (e.g., age, medication, etc.). Therefore, models can be faulty in generalizing among different groups of patients. A more personalized approach would be beneficial to better understand the behavior of ML models among different groups.

In recent years, the availability of different forms of data has enabled the construction of ML models that exploit different characteristics of the data. In particular, we observe a trend towards usage of so-called surrogate devices such as smart watches, phones, etc. as the source of data for ML models [MOW18, FLW*20]. In the case of sleep staging, surrogate devices can be used to track the sleep of patients for longer periods and in less-intrusive manners than polysomnography. Usually, models consuming surrogate data output a smaller set of sleep stages than those trained on polysomnography due to less detailed data (e.g., high detail electroencephalogram (EEG) vs. low detail actigraphy and heart rate). In general, models trained on different data can fail in recognizing situations such as sleep fragmentation, arousals, etc. Analyzing and comparing models for sleep staging with heterogeneous sources of data can provide valuable insights to experts.

To the best of our knowledge, no approaches have been presented yet to conduct performance analysis in this sort of scenario. To this end, we present PerSleep, a visual analytics approach that aids ML experts in sleep staging to assess the performance of the models
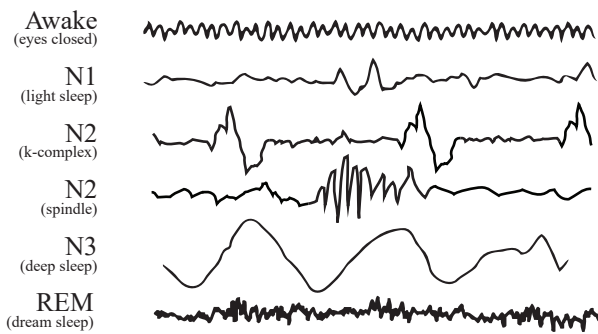
Figure 1: *Examples of EEG waves and their corresponding sleep stage. Time and amplitude scales are different in each example.*

they employ. Our main contribution is the first visual analytics approach to evaluate and compare performance of two models in sleep staging. Multiple hypnograms can be visualized simultaneously to make quick comparisons of the same target hypnogram for different models. In comparison with state-of-the-art approaches in performance analysis, ours does not solely focus on the input data of the model but also the patient data. The novelty of our work lies in the application of visual analytics in sleep staging rather than the design of new visual idioms. Furthermore, we hope that our work provides a useful example for the assessment of complex models for judging time series data for varying populations, like neurological brain disorders [AAAA20] such as epilepsy and autism, or physiological disorders like heart failure detection [KKE*21]. We present a use case on real-world data to demonstrate our approach. The use case was conducted with three experts. Results, limitations and generalization are discussed. Finally, we provide directions for future work for performance evaluation in sleep staging.

2. Medical Background

When a patient is believed to be suffering from a sleep disorder, a *polysomnography* (PSG) is often done, an electrophysiological recording of sleep and sleep-related events overnight. In a PSG, brain activity (measured by EEG), muscle activity (measured by EMG) and eye movements are recorded to assess sleep structure. In addition, other aspects are recorded such as body movements and breathing patterns. Afterwards, the recording is evaluated and annotated on an *epoch-by-epoch* basis. *Epochs* represent time-fixed periods (typically 30 seconds duration), which can then be analyzed by technicians in order to assign a sleep stage. The process of assigning sleep stages to epochs is called *sleep staging*, which was invented in the 1960's. During sleep staging, sleep is scored epoch-by-epoch as one of the five disjoint categories according to the American Academy of Sleep Medicine (AASM) [oSM*07]: *Wakefulness*, *N1*, *N2*, *N3* and *REM*. Sleep stages are characterized by specific physiological properties, which are based on consensus criteria. Figure 1 depicts some characteristics that can be observed in the EEG signals of a PSG. *Wakefulness* with the eyes closed is usually characterized by *alpha* waves (8-13*Hz*) in the EEG produced by the occipital lobe of the brain, while sleep stage *N1* often presents *theta* waves (4-7*Hz*). Other stages are characterized by in-

teractions of multiple physiological stimuli that result in the presence of EEG phenomena like k-complexes, spindles, or sawtooth-like waves. The sequence of annotated sleep stages during sleep is visually represented by a *hypnogram*, which is analyzed by a somnologist to understand the sleep pattern of a patient.

Generally, somnologists look for patterns in the hypnogram in terms of overall presence of and transitions between sleep stages. These patterns have clinical meaning, i.e., they can be indicators of sleep disorders. For instance, *fragmented hypnograms* present many transitions between sleep stages occurring in short time intervals, resulting in a fragmented sleep pattern. Such pattern can be indicative of a sleep disorder such as insomnia and narcolepsy.

Progress made in ML in this area has brought the opportunity for hospitals to switch to automated methods, which can be used to score PSGs. To this end, models need to be robust and reliable to assure the validity of the outcome they output. To support their assessment, better understanding of how models perform on clinical data is essential, and with our work we aim to contribute to that.

3. Problem Definition

It is difficult to develop ML models with a high accuracy and reliability in sleep staging. Furthermore, assessing the performance of a model is non-trivial. The result of an automated process is a hypnogram, rather than just statistics on individual epochs, and the overall quality of a hypnogram is hard to assess. Also, the performance of a model can depend on characteristics of the patients, such as age and gender. In general, models in sleep staging do not consider the demographics of the patients and focus on the physiological signals such that the model can generalize from these. This is due to two reasons. First, if demographics such as age and sleep disorders were to be considered by the model, it would require an immense amount of representative data of all combinations of age groups, and each sleep disorder, requiring thousands of participants. Obtaining this large amount of data is challenging and time consuming as it often involves patients being recorded, for at least one entire night, with many sensors in a sleep center. Second, it may be that the sampling done when selecting the patients introduced bias due to specific features (i.e., artifacts) of such selected group being hooked on by the model. These may not be true for other patients belonging to that category and not included in the sampling. The goal of this project is therefore to develop a visualization to enable experts to evaluate and understand the performance of ML methods for producing hypnograms, also in relation to the properties of patients.

The data used in performance assessment of sleep staging models is multivariate as it combines patient's data of different nature (e.g., demographics, clinical information and physiological records) and the model's data (e.g., predictions and probabilities). The dataset depicting patient data is a *table* with an undetermined number of attributes. Typically, age (quantitative), gender (categorical) and other comorbidities (categorical) are part of the patient's data. In general, both categorical and ordered attributes can be part of the patient's data. Moreover, each physiological record of the patient is a *field*, where each cell depicts the physiological measurements for a given point in time. In sleep staging, the sampling frequency is uniform. In most situations, this field dataset is fed to

the ML model to predict a *list* of sleep stages. Each sleep stage is a categorical value representing one of the possible classes defined by the AASM. Similarly, a list of probabilities is also produced by the model, where each probability is a quantitative value. Aggregations are often used to summarize performance data. For example, a confusion matrix is a *table* where both items and attributes depict sleep stages. Each cell of this table contains a quantitative value.

The system should support various levels of detail for the evaluation of the performance, where each level leads to its own questions, and enable smooth transitions between these:

L1 Based on individual epochs: What is the *probability* of the model for a given prediction? What was the *input data* for a given epoch?

L2 Based on individual hypnograms: What are the *main differences* between two models? How did the *confidence* of the model fluctuate over the entire night?;

L3 Based on aggregate results across large sets of patients: What are the *scores* for aggregate statistics? Are there *correlations* between data attributes?;

Furthermore, the expert must be able to split the set of patients into cohorts, specific subgroups, based on their properties, and compare the performance for cohorts and focus on specific cohorts, to answer questions such as *how does one subgroup compare to another?* and *are there groups of patients that have similar performance indicators?*. Also, rather than focusing on just a single model, the expert should be enabled to compare multiple models using different data and/or ML models where the *test* model depicts the one to be evaluated (e.g., neural network) and the *reference* model acts as ground truth (e.g., manual scoring).

From the previous levels of detail and questions, we derive the following tasks:

T1 Explore the distribution of patients in terms of attributes. This provides an overview of what sort of distribution an attribute follows for the entire group of patients. Visualizing such distributions can help to detect odd behaviors in our model. [**L3**]

T2 Find correlations between data attributes. Correlations are important to gain insights into the behavior of the model. For example, it may be the case that our model performs worse for patients that are old and take a specific medication. Hence, experts should be enabled to perform selections on attributes to generate and validate hypotheses. [**L3**]

T3 Analyze the performance of a model. Summarized statistics such as accuracy or kappa, only give a glimpse of the whole picture. Instead, an exploratory process is needed to gain insights into several factors that usually are intertwined. For example, accuracy value can be low and yet the clinical interpretation of both test and reference hypnograms be the same. [**L3**, **L2**]

T4 Compare hypnograms. Visualizing the hypnograms for both test and reference models is crucial to understand whether the performance of the model is good enough for medical purposes. When inspecting the epochs of a patient, the approach must enable experts to select portions for closer inspection. Input data should be provided to contextualize an epoch. [**L2**, **L1**]

Tasks **T1** and **T2** shall also be performed for groups of patients.

4. Related Work

In this section, we provide a review on previous work on performance analysis, time series and sleep analysis.

Performance Analysis. In performance analysis, predictions are the core element to be investigated. Generally, they are generated in combination with a set of probabilities that indicate how likely the prediction is to be of a certain class. A common approach in performance analysis is to explore the entire set of probabilities to find possible outliers. Most approaches visualize probabilities grouped by predicted class. ModelTracker [ACD*15] does not stratify predictions in classes because it just considers binary classification. Squares [RAL*17] is an extension from ModelTracker to support multiclass analysis. It makes use of histogram-like visualization to show probability distributions. They explicitly divide these into two groups: labeled and predicted class. The authors use color encoding to depict situations where both labeled and predicted class agree or disagree. Our work follows a similar approach as Squares, but using a somewhat simplified visual encoding to present probabilities. Moreover, we complement it with a confusion matrix that is used to explore specific cases in more detail by means of interaction.

Boxer [GBYH20] is a system that assists experts in developing and assessing classifiers. They address multiclass classification by means of interactive views that are formed by standard visualizations. It allows experts to layout views in *boxes* such that it gives different perspectives of the data, resulting in a flexible analysis of the performance of the classifiers. While Boxer considers multiple classifiers at the same time, our approach focuses on two models to maximize contrast and highlight differences. Furthermore, Boxer does not handle time-series data. In addition, our approach is model-agnostic within the sleep staging domain.

Time Series. A common visualization approach consists of displaying the input data of a model together with the predictions. It enables the exploration of the input features of a model. In the sleep staging domain, the input data is temporal. This data has received little attention as most works in ML literature target either multidimensional [ZWM*18], text [SGB*19] or image data [PHVG*18]. All these approaches provide interaction to select subsets of predictions to explore the whole (or partial) input space.

Some work has been done in understanding ML models where time is a component. RetainVis [KCK*19] and DPVis [KAS*20] utilize neural networks and continuous-time hidden Markov models respectively, whereas our work is model agnostic. They stratify patients into different groups defined beforehand, wherein our approach any data attribute can be used for this purpose. Finally, RetainVis and DPVis take feature contribution as key in their designs. Our work, however, does not take feature contribution into account and just focuses on performance indicators and patient data.

Sleep Analysis. The work of Combrisson et al. [CVE*17] presents a visualization tool to help technicians to manually score hypnograms. Automated techniques are used in their work to detect characteristic features in EEG signals. Their approach emphasizes the detected features in the polysomnography such that technicians can make better informed decisions when scoring the hypnogram. Although we do not aim to manually score hypnograms nor detect features in the EEG, we do make use of hypnograms in our
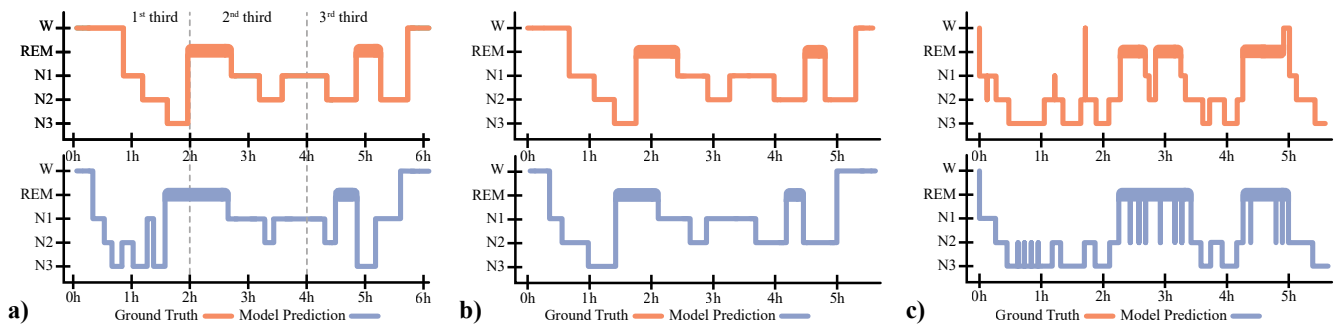
Figure 2: *Illustration of three example problems in performance analysis for sleep staging: error grouping (a), global performance (b), and fragmentation (c). In (a), most of the misclassifications occur in the first and last third of the night, whereas the second third is the most accurate. (b) shows shifted predictions where the model predicts nearly the same global pattern but slightly earlier in time, resulting in low accuracy. Last, in (c) we see many transitions (*fragmented hypnogram*) between REM to N2 that are not present in the ground truth.*

approach. Hypnograms are the *de facto* standard of visualizing the sequence of sleep stages, and are also used in several other approaches [FvGR*20, PAC*19]. Our work focuses on assessing the performance of models, whereas the aforementioned work addresses the design of ML models from a pure ML perspective focusing on the inner workings of models.

In earlier work, we introduced V-Awake [GCWGvW19], a visual analytics system to help sleep technicians finding potential misclassifications from deep learning models in sleep staging. V-Awake was received positively by the domain experts, but rather than fixing errors afterwards, they also aimed to develop better models, and they missed systems to assess the quality of these. This led to the work presented here. PerSleep would be used when designing a model. Afterwards, V-Awake would be used in a real-world scenario to find possible misclassifications.

In summary, the major contribution of our work is the first system that focuses on the evaluation of the performance of models for sleep staging. Sleep staging is very special and important, and we must carefully design something ad hoc for that. It differs from other domains due to the incorporation of a time domain (epochs) and the need for flexibility to define subgroups. Currently, systems do not consider both characteristics simultaneously and generally focus on predictions independently from each other. Also, sleep staging is a great example of a ubiquitous pattern, not only for health applications, but also many other domains like epilepsy and autism [AAAA20], and heart failure detection [KKE*21].

5. PerSleep

The tasks defined in Section 3 steered the design of PerSleep. In this section, we introduce its main components (see Figure 3).

5.1. Model Selection

The model selection component (Figure 3A) enables users to import new models (containing scoring data) to PerSleep by clicking the plus button. These models can be easily selected now to allow for quick switching between distinct models. For instance, when exploring the hypnogram of a patient, it is interesting to switch

between different models to verify if differences arise in terms of performance. In PerSleep, we compare two models: a *test* and a *reference* model. When both models are selected, all the relevant performance information will be displayed accordingly.

5.2. Patient Data

The *patient data* component consists of four views to provide an overview of the distribution of our entire population of patients and mechanisms to select and create groups of patients.

5.2.1. Patient Attribute Views

After discussing with the experts in sleep staging their needs, we opted for two linked views to present the patient's data in two ways: a barchart plot (Figure 3B1) and a parallel coordinate plot (PCP, Figure 3B2). Experts can decide what attributes to show at each time, focusing the analysis in specific parts of the entire dataset.

The barchart plot is useful to gain an understanding of the data distribution quantitatively (task **T1**). The PCP can be used to understand correlations in the data attributes (task **T2**). We use curves as a solution for the crossing problem [GK03]. The PCP provides a mechanism to have the same axis ranges for a set of selected attributes. This is useful to make direct comparisons between attributes (e.g., accuracy and kappa) when they have different value ranges. PerSleep uses colors to depict patients belonging to groups of interest, i.e., created groups.

Some of the data attributes presented in our approach are computed dynamically when a model is selected. These attributes aim to summarize the information contained in the hypnogram and are meant to help detecting scenarios like those depicted in Figure 2:

Performance metric per third of the night This attribute can be used to inspect how the accuracy and kappa fluctuate during different parts of the night (see Figure 2a). Experts in the sleep domain usually divide the night into three equal parts to carry out their analysis (e.g., non-REM parasomnias usually occur in the first third of the night).

Pair-wise alignment metric The Smith-Waterman sequence alignment [SW*81] is applied in order to verify how well test

Figure 3: *The views B1-B4 are patient-related, where B1 and B2 depict views where patients can be filtered based on the selections performed. B4 displays the patients that match the current selection. Groups of patients can be created, modified and selected in B3. C1-C3 shows performance information, where C1 depicts our probability plot, C2 a confusion matrix and C3 complements information displayed in the confusion matrix. In D1 the physiological data view shows the hypnograms for a selected patient. Alternatively, extra models can be visualized together with those the selected in A as shown in D2. Finally, D3 shows the signals that represent the input data.*

and reference model align. This helps to detect situations in which the accuracy metric is poor but the overall pattern of both hypnograms is somewhat similar (see Figure 2b).

Transitions and Sleep Fragmentation Index (SFI) The number of transitions as well as the SFI [MFK*00] can be useful to detect situations in which a model does not recognize sleep fragmentation adequately (see Figure 2c).

Sometimes, it is interesting to incorporate new data attributes. For instance, the expert may want to know how many transitions there are for a certain stage to verify a hypothesis generated during exploration. To this end, our approach enables the creation of these attributes on the fly, which are then saved for subsequent exploratory sessions. To do so, users need to manually code how the values of these new attributes are computed by using the JavaScript language. A dialog can be opened by clicking on the cog icon present above the barchart plot and the PCP. This mechanism is meant to be used by users with knowledge on scripting languages, which is usually the case for ML experts.

The barchart and the PCP highly rely on brushing and linking to perform selections. When a selection is made in one of the views, the other is updated with the same selection. Having both perspectives (i.e., quantitative and correlative) linked helps the experts to better understand the global context of their data. More precisely, having visual feedback in the barchart when performing a selection

in the PCP aids to understand whether the data selected follows a different distribution compared to the entire population or not.

5.2.2. Group Manipulation View

In Figure 3B3 the *group manipulation view* enables creation and manipulation of groups of patients. Created groups and their number of patients are shown in this view, which helps addressing **T1** and **T2** for groups of patients. Every group is identified with a name and a color that are assigned when created. In our system, we assume non-overlapping groups. Once patients have been added to a group, the PCP component updates accordingly, color coding each patient with the group color that it belongs to.

PerSleep provides a mechanism to create groups automatically using DBSCAN [EKS*96]. The clustering technique is fed with the normalized confusion matrices for each patient. Normalization is done *per patient* according to the total number of epochs. The aim of the clustering is to help experts in finding groups of patients that have similar model performance. The technique takes two parameters: *minPts* and ε. We set *minPts* to be 2 as we are interested in groups that contain at least two patients. ε can be adapted to enable exploration of different cluster outputs. We use an euclidean distance as the distance metric.

Groups can be selected on demand to explore performance data exclusively for those patients contained in the selected group.

5.2.3. Population View

The *population view* (see Figure 3B4) contains a table-like view that displays descriptive information of the recording of the patients and a summary of the disagreement for the selected models. We opt to explicitly encode the differences between the two sequences [GAW*11] by computing the epochs in which they disagree. This can be utilized to spot regions of disagreement, which partly supports task **T4**. Our visual encoding guarantees visibility [Mun14] of the disagreements. Moreover, visual aids are added to depict three partitions of each sequence to ease comparison between patients with different sleep duration. Recordings can have different lenghts, which are shown numerically. For each patient, the visualization of the recording is stretched over the full width of the column to ease comparison. This is motivated by discussion with experts who needed to understand if errors present any pattern at a night level. The population view also indicates the groups to which patients belong. It is done by color coding the icon button placed on the left side of this view.

Users can sort the table of patients based on data attributes. This helps to quickly find patients for which a model performs the worst or the best. This view provides ways to select a patient for further exploration. Once a patient is selected, the patient data view, performance view and physiological data view change accordingly to show the information for this patient. We opt to keep the selected patient always visible by creating a visual duplicate on the top of the list. This is very helpful for experts to be aware of the patient that was selected, even when scrolling through the list. The *group view* updates to indicate the group to which this patient belongs by changing its visual encoding (see Figure 3B3). Moreover, patients can be unchecked such that they are not included in a group upon group creation. This gives a fine-grained method to exclude patients from the selections performed in the *patient attribute views*.

5.3. Performance View

To support task **T3** we introduce the performance view, which is composed of two components: *probability view* and *confusion matrix view*. They are aimed to be used together to gain insights into the performance of the model.

The *probability view* depicts a multi-axis view that conveys information about the probabilities for every epoch (see Figure 4). Each axis depicts the probability for each sleep stage, which are divided in ten bins of equal size ranging from 0 to 100. We distinguish between two categories: *agreements* and *disagreements*, which are placed on the right and left of each axis respectively. Each category features a vertical barchart depicting the number of predictions for a specific probability. The agreements represent the *true positives*, whereas the disagreements can represent either *false negatives* or *false positives* based on the choice of the user. *False negatives* for class *C* in a multiclass problem are those samples where the *reference* model classified as *C*, whereas the *test* model classified otherwise. Similarly, *false positives* are samples where the *test* model classified as *C* but the *reference* model did differently.

This view is linked with other components and updates when a selection is performed to accordingly show the aggregated values. Also, the expert can select specific bins in this view to generate a
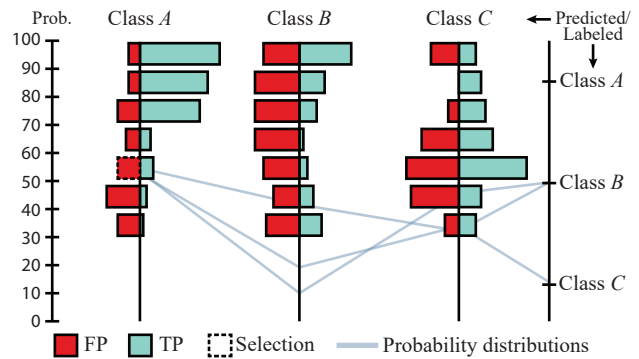

Figure 4: *Glyph design of probability view. A multi-axis depicts the distribution of* true positives *and* false positives *per class. A line plot reveals the probability distribution for each prediction contained in a selected bin. The right axis depicts the labeled class.*

line plot in the background (see Figure 3C1). It displays a double histogram for each class where each bin depicts the probability of a prediction. For each histogram, the right side shows agreements between both models (i.e., true positives and true negatives), whereas the left side depicts disagreements (i.e., false positives, false negatives or a combination of both). Moreover, the right-most axis depicts either the labeled or predicted class depending on whether the user is interested in *false positives* or *false negatives*, respectively. This complements the information shown in the *confusion matrix view* to give a better overview of the performance of the model.

The *confusion matrix view* (see Figure 3C2) provides a performance summary for every sleep stage. It is a standard way of presenting performance in multi-class scenarios. Although it can present information for any number of distinct classes, it is a good practice to keep this number low. In general, we deal with up to five distinct classes in sleep staging, thus clutter is not an issue. Previous work [GCWGvW19] also used a confusion matrix to guide the user to find potential misclassifications when ground truth is not available. In PerSleep, ground truth is available. Therefore, our confusion matrix shows the actual data rather than an estimate.

Sleep staging is, by definition, an imbalanced problem [SDWG17] with higher frequency of class *N2*. This poses the problem of getting biased insights when visually inspecting the confusion matrix. We provide an interactive confusion matrix where experts can decide how to display the data in the cells. Four possibilities are available: whole numbers, percentages, precision and recall. Percentages, precision and recall values are shown as ratios in our approach. When selecting recall or precision in the confusion matrix, the visual encoding adapts to better convey that numbers are *normalized* per column or per row. The confusion matrix can convey a general message on where accuracy and error occur more often. However, it is difficult to get an idea on the relative distributions of errors. For this reason, we accompany the confusion matrix with a double, vertical bar chart (Figure 3C3). The top side shows the *true positives* and *true negative* epochs (i.e., matching epochs), whereas the bottom side displays either *false negatives*, *false positives* or a combination of both (i.e., mismatching epochs).

Next to the confusion matrix, accuracy and kappa statistics are shown. These are understood by most experts in sleep staging.

Experts can interact with the cells of the confusion matrix. When one is selected, it filters out those patients that have at least one epoch in the selected category. For example, experts can select patients where *REM* is confused with *N3*. This particular situation is of interest since *REM* and *N3* are conceptually very different.

5.4. Physiological Data View

The *physiological data view* is shown in Figure 3D1, D2 and D3 and addresses task **T4**. This component provides a close look at the predictions of the test model, the ground truth of the reference model, and the signals for a single patient. Restricting to a single patient is motivated by discussions with the somnologist who stated that visualizing hypnograms simultaneously for many patients would rather obfuscate the analysis. By default, the hypnograms of the selected models are displayed. However, the expert may alternatively choose to visualize the hypnogram of other models (Figure 3D2) for quick comparisons.

This view features a piano-roll visualization for both test and reference scored hypnograms (see Figure 5a). It provides a visual overview on how similar they are. Both hypnograms follow a linear representation, with relative scale and unified layout [BLB*16]. Sometimes, deviations can be subtle and difficult to spot. To this end, we propose a combination of juxtaposition and explicit codification of differences [GAW*11]. The *difference view* follows the same principles as in the *population view*. The piano roll encodes sleep stages with position and color. Experts can switch to a single color piano-roll, which is closer to the representations they currently use in the sleep domain, or customize the colors for each sleep stage, which updates the entire interface of PerSleep. Our previous work [GCWGvW19] featured a similar view, however it was restricted to just one hypnogram to enable experts to spot and brush interesting patterns to find misclassifications.

When probability data is available, experts can choose to visualize it together with the hypnogram (see Figure 3D1). This provides an overview on how the probabilities fluctuate, potentially signaling situations in which the model was not sure about a prediction. Similar to our previous work [GCWGvW19], the probability is encoded in the background of the hypnogram. Two modes are provided: predicted sleep stage or chosen sleep stage. The first projects the probability of the class that is predicted in a certain epoch, whereas the second enables experts to visualize the probability of a sleep stage for all the epochs. This is useful to generate hypotheses about what the model detects in the input data.

Basic demographic information, such as age and sex, is shown in this view when a patient is selected. These demographics are important for experts in order to correctly interpret the hypnogram. The sleep of young and old people is different, for example. Patient ID and file name are shown for completeness such that the expert can quickly identify the patient being visualized and from which file the PSG is taken. Experts can explore patient data in full detail by clicking on the medical notes icon, which opens a new window where the full data is listed in tabular form.
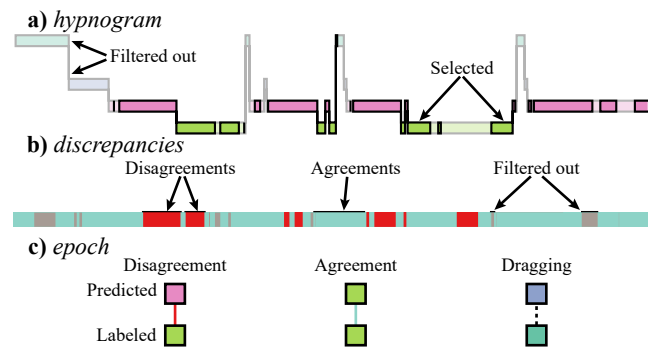

Figure 5: *Glyph design used in* physiological data view. *a) depicts the way we visualize hypnograms when selections are done in the* performance view*; b) shows the way we encode the differences between predicted and labeled hypnograms, where the colored stripes indicate disagreement and different opacity is used to depict current selection; c) presents the slider used to select a specific epoch.*

A widget on the upper left brings the set of signals. The expert can select any available signal to be shown below the hypnograms (Figure 3D3). This provides a flexible mechanism to deal with models that have different input signals (e.g., with and without respiratory information). Every signal is shown through a time window, which can be configured by the expert. Navigation is enabled in *epoch* units. The expert can set the length of the signal in seconds to be displayed, giving more flexibility to explore the epochs surrounding a prediction.

Two main interactions are provided. The first one concerns selections of parts of the hypnograms with focus+context. To this end, we rely on brushing over the *difference view*. It facilitates the selection of disagreeing, or interesting in general, fragments. The second interaction enables quick navigation to the input data for a specific epoch. This can be done by dragging a slider over the hypnograms. The design of the slider provides hints indicating the class of the test and the reference models. A line connecting both ends encodes whether they both classes agree. Examples can be seen in Figure 5c.

5.5. Complexity and Scalability

PerSleep has been implemented as a web app that entirely runs on *client* side. This means that data never goes out of the local machine of the user, and everything remains in the local storage of the web-client. This decision was made to ensure there are no privacy issues with the data being analyzed. However, it also introduces limitations that need to be addressed.

The visualization of the input data besides the outcomes is important for performance evaluation. In sleep staging, many physiological signals are recorded, resulting in a large amount of data. In order to enable a smooth data exploration, we make use of WebAssembly [HRS*17] to run EDFlib, a C library that reads EDF [KVR*92] files efficiently. With this approach, we achieve a nearly native speed, resulting in a smooth and viable data exploration.

Another consideration goes for the computation of data attributes. As discussed in previous sections, our approach provides
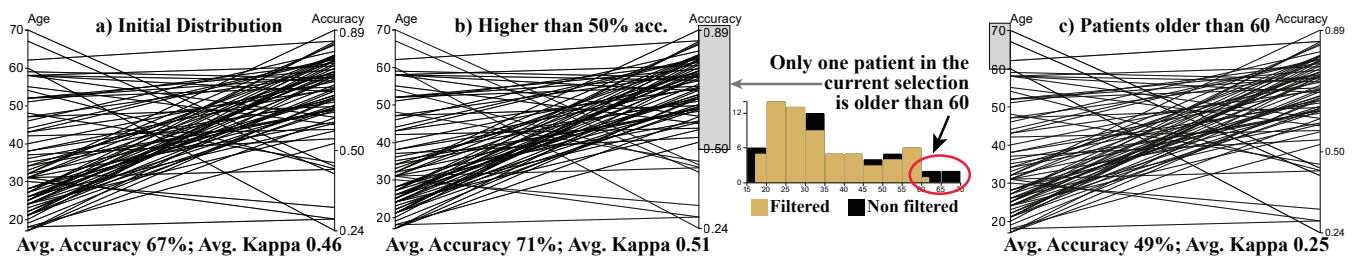
Figure 6: *a) initial set of patients for ECG model; b) selection with accuracy higher than 50%; c) selection of patients older than 60.*

mechanisms to define new attributes. These are recomputed on demand every time a new model has been imported. This ensures that PerSleep remains responsive during the exploratory phase.

We have run experiments with 236 patients and 463,090 epochs. In other experiments based on real-world data, we handled smaller amounts of data in terms of patients and, thus, number of epochs. Our approach has been able to handle all our experiments adequately and no concerns arose from our users. We have observed that the *probability view* tends to be computationally most demanding of our system. However, we have not experienced any significant impact on the interactive performance for real-world use cases.

6. Use Case

We demonstrate our approach on a real-world multivariate sleep dataset which contains three different alternative sleep staging models, based on the use of ECG, respiratory effort and ECG+respiratory data. The data that these models consume can be extracted from surrogate devices. For example, the ECG model is fed with heart rate variability, which can be measured with most modern smartwatches. The interest of the experts in these models is to verify if they could replace the gold standard (i.e., PSG) to perform sleep studies in a less intrusive manner. The three models output 4 classes: *awake*, *N1+N2*, *N3* and *REM*. The dataset was recorded in a Dutch hospital as part of a study about sleep in 74 patients with intellectual disabilities more than 16 years old who suffered from a variety of sleep problems. Some patients present comorbidities such as heart problems and epilepsy. The patients received a PSG during routine clinical care. Patients with absent ECG channels or poor ECG and EEG were discarded from the study. Patients' attributes such as age, sex, comorbidities, primary diagnosis, whether they receive medication, etc. are available in the dataset. We show how the visualizations and interactions in our approach help experts to gain insights into this dataset. The use case was performed throughout several interactive group sessions with a somnologist, a machine learning expert and a signal processing expert. The three users are all knowledgeable about sleep and machine learning. A multidisciplinary setting is common for assessing the performance of a ML model. Each expert can add a different point of view: the somnologist provides a clinical perspective, the ML expert adds knowledge about the model and the signal processing expert includes feedback about the signals fed to the ML model.

Generally, interesting patients are those that exhibit an extremely good or bad performance compared to others. Initially, we created

a case study in our system containing the patients recorded in the previously defined study. Then, we incorporated the data from the three models, resulting in 79,573 epochs per model.

First, experts selected the ECG model as test model, and human scoring based on EEG as the reference model. We started the exploration by inspecting relationships between data attributes. Experts wanted to verify if there was some correlation between basic demographics and accuracy. For this, the experts selected age in the barchart view, and age and accuracy in the PCP. After a first inspection of the PCP, the experts observed a certain trend of lower accuracy scores for older patients. In particular, for 3 out of 4 patients (75%) that are 60 or older, the ECG model scored lower than 50% in terms of accuracy. This contrasts to patients that are younger than 60, where for just 6 out of 70 of them (9%) the model scored lower than 50% accuracy (Figure 6). This finding was interesting, as there is no data on age dependency of alternative sleep staging models. We explored the same set of patients with the respiratory model. In this case, only 1 out of 4 patients scored lower than 50% accuracy, which may indicate that the ECG model is not reliable for older patients. A closer inspection of this patient revealed that the ECG model was not able to detect any *REM* stage, and most of times the model confused *N3* with *N1/N2*.

Epilepsy appears to be more prevalent among people with intellectual disabilities. In fact, it is believed that up to one-fifth of the population with intellectual disabilities also suffer from epilepsy. Epilepsy certainly has an impact on sleep abnormalities. In particular, it can increase sleep latency (i.e., time to fall asleep), sleep fragmentation, awakenings and stage shifts [MR01]. However, in addition, epileptic activity in the EEG can make it more difficult to annotate sleep stages. For these reasons, the experts wanted to understand how the ECG model was performing on these patients. To this end, the experts created two groups: patients with and without epilepsy. These groups resulted in 28 and 46 patients respectively. We observed that average accuracy and kappa values were very similar for both groups, which suggests that the ECG model is not performing differently for patients with epilepsy.

To gain more insights into the epilepsy patients, the experts sorted the patient list by accuracy to focus on the top and bottom cases. In particular, we found one case with low accuracy (44%) which contained only ECG artifacts (Figure 7 top). Despite this, the model was still able to classify some sleep stages. The experts switched the test model to check how the respiratory and ECG+respiratory models performed. The respiratory model, whose data does not seem to contain artifacts, scored slightly lower (42%)
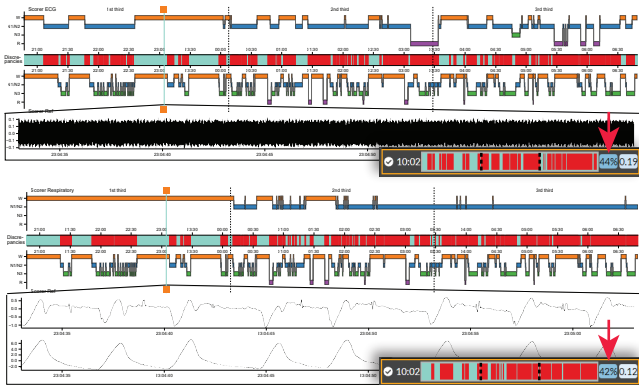
Figure 7: *Patient (ID 51) with diagnosed epilepsy. The ECG model (top row) performed slightly better than respiratory model (bottom row) although the former used data with many artifacts as input, which is shown for epoch 275 and a 30 seconds window.*
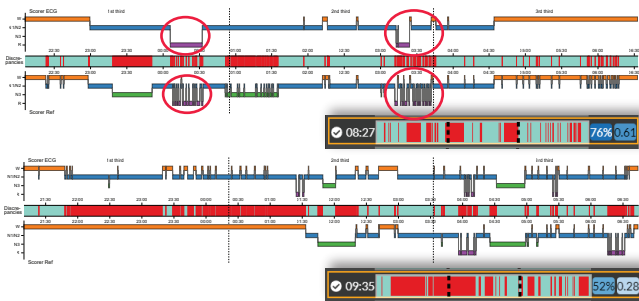

Figure 8: *Examples of misleading statistics. The top row (ID 114) presents a case in which the test ECG model does not recognize* REM *fragmentation (see red circles). Bottom row (ID 133) have similar clinical interpretation (i.e., similar overall pattern) despite the low accuracy and kappa of the test model.*

than the ECG (Figure 7 bottom). However, it was interesting to note that the combination of both (ECG+respiratory) provided a slightly better performance (48%). Experts looked at comments made by the sleep technicians by clicking on the medical notes icon. It was clear that the technician scoring the reference sleep recording had trouble identifying some sleep stages while it was easy to distinguish between *wake* and *sleep*. It was difficult to decide between *N1/N2* and *N3* in this patient. After reading these annotations, the experts suggested that the ECG model may be providing a better scoring than the human EEG based scoring.

From this moment of the analysis, experts focused on selecting patients from the full patient list to obtain an idea on common problems. Experts detected two common situations: *REM* misclassifications and *N3* fragmentation. Regarding *REM* misclassifications, the experts proposed that this could be caused by sleep apnea (often occurring in this stage), or autonomic dysfunction leading to abnormal ECG patterns in *REM*. By inspecting the data, they were able to verify that a breathing disorder was indeed diagnosed for some of these patients.

During this part of the analysis, experts found cases in which aggregated metrics (accuracy and kappa) were absolutely misleading. In particular, they found cases where kappa and accuracy were high, but the clinical interpretation of the two hypnograms would be very different (Figure 8 top) due to the absence of sleep fragmentation. Similarly, we found some cases in which kappa and accuracy were low, but the clinical interpretation of the test model would most likely be the same as the reference model (Figure 8 bottom) because the overall pattern of transitions looks alike for the test and the reference model.

6.1. Sleep Fragmentation

Sleep fragmentation is one of the problems depicted in Figure 2. One of the causes of sleep fragmentation is medication. To analyze how our model copes with this, the experts selected three attributes in the PCP view: transitions in the reference model, medication and transitions in the test model. Immediately, they observed two things: medication seems to have an influence in the total number of transitions in the reference model but not in the test model; and the test model produced a significantly lower number of transitions for all the patients (see Figure 9). The latter may indicate that the model is *smoothing* the resulting hypnograms.

In order to verify this hypothesis, the experts created two groups of patients. The first group contained patients with a low number of transitions (≤ 150) whereas the second had a high number of transitions (> 150) according to the reference model. Initially, they noticed that both groups had different performance metrics (70% vs 60% accuracy; 0.5 vs 0.4 kappa). Next, they wanted to gain more insights into the behavior of the model for each group. They started the analysis by selecting the group with many transitions and sorting the patient list by accuracy. After inspecting several patients with different accuracy values, they discovered that sleep fragmentation was indeed not correctly detected in the vast majority of the cases. They repeated the process for the other group of patients. They found out that the majority of patients did not present a very fragmented hypnogram, but they found a few cases that presented fragmentation and yet the model produced a smoother version that omitted such fragmentation. This may indicate that the model is not able to capture the transitions between sleep stages as a human scorer would.
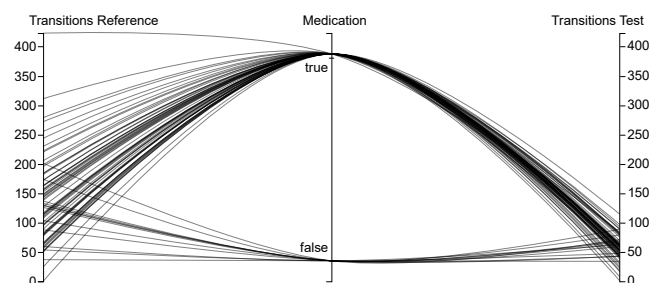

Figure 9: *PCP comparing transitions in test and reference model against medication. As can be seen, the test model seems to smooth the number of transitions for patients with and without medication.*

7. Discussion

The use case presented in Section 6 shows how PerSleep can be used for an exploratory analysis to evaluate the performance of a ML model for sleep staging. This is done by leveraging the proposed design to explore the data from different perspectives. Tight linking between patient data and epoch data provides the somnologist and the ML experts a mechanism to generate and test hypotheses that would otherwise require a lot of manual work.

The definition of performance for sleep staging is ill-defined due to the coarse-grained nature of traditional metrics that are often provided when evaluating sleep staging classifiers. According to the somnologist and ML expert: *"There was a clinical message out of this. It shows that the clinical interpretability of a surrogate or target hypnogram definitely is not fully determined by the kappa and accuracy numbers"*. This claim strongly supports the need for visual analytics systems that help in performing exploratory analysis of performance in sleep staging.

The current workflow for performance analysis in sleep staging is rather cumbersome. It often involves several steps performed in different platforms and software. Among the steps, we can find visual tasks like analysis of hypnograms, etc. Our approach unifies all the steps and provides mechanisms to link elements in a visual manner. A remark from the somnologist after the use case was: *"In papers, you very often look either at the group level or some illustrated hypnograms. One of the merits of PerSleep resides in the ability to analyze very quickly individual recordings (i.e., hypnograms) so one could search for specific reasons why discrepancies happen, which can really be different from subject to subject"*.

In general, working with clinicians limits the choices when designing a visual analytics system. For example, we found that the PCP was on the edge of complexity for the doctor. This motivates using simple visualizations. Clinicians are used to some graphical representations (in this case hypnograms). These traditions must be respected and included in the final design. Otherwise, the system may become unusable for clinicians.

As for other approaches, ours has limitations. For instance, ours does not deal with techniques that use inputs such as images. This affects the generalization of our approach. Also, we heavily rely on prior knowledge from the expert on the models being analyzed such as the input data, or the set of patients (e.g., healthy or not). Additionally, the creation of new data attributes relies on prior knowledge with scripting languages, which can be a problem for experts on sleep medicine that lack of a more technical background. Finally, our approach does not provide mechanisms to distinguish between findings that may not be statistically significant. We however believe this is alleviated because the users can rely on their expertise to decide whether the findings are representative or not.

7.1. Approach Generalization

Our approach can be generalized to other domains. In the medical field, it can be applied in epilepsy prediction [SMMHH20]. In this context, ML models are used to detect seizures from EEG data. Hence, this domain shares many similarities with sleep staging: predictions are collected per patient, which also has multivariate data, and they are sequential. The data abstractions of the epilepsy and the sleep staging domains are alike.

In epilepsy detection, ML models are trained to, at least, detect two different classes: *seizure*, and *non-seizures*. These classes can be split into more specialized ones to characterize the severity of the epilepsy: *simple-partial*, *complex-partial*, *generalize convulsive* and *generalize non-convulsive* seizures. The former two epilepsy seizures happen in one hemisphere of the brain, whereas the latter two happen in the whole brain. Usually, the duration of a seizure ranges from seconds to minutes. Users in this field would be interested in correctly recognizing the type of seizures to determine the severity of the epilepsy attack and correlate the predictions of the model with other clinical data. Our approach was designed to address the tasks stated in Section 3. Except **T4**, which involves comparing hypnograms, the remaining tasks would still be suitable in this field. Additionally, locating the origin of the seizure is important in epilepsy detection. In this regard, the EEG (i.e., the brain electrodes) already gives some clues on where the seizure took place. Therefore, it would be necessary to add a new task. This would help in locating the origin of the seizure.

8. Conclusions and Future Work

We presented a novel approach to evaluate the performance of ML models for sleep staging. It combines different visual and interactive components to enable experts to conduct their exploratory analysis. In contrast to related work, we address a problem that involves predictions over time in combination with patient data, which cannot be dealt with using current approaches. A use case has been presented to demonstrate our approach where we describe the main discoveries made by experts during exploration.

In principle, a similar approach like ours can be used to any situation where complex dynamic signals have to be judged for the state of the object of interest. In our case, we took care to understand the needs of our collaborators and carefully tuned the system accordingly. This may be simply the way to go, but it is also intensive. The design of a generic, flexible system that enables similar functionality without programming is still an open challenge. As for future work, we aim to extend the clustering capabilities of our approach and evaluate different techniques and distance metrics. Finally, analyzing the uncertainty of a model would be greatly interesting. This is motivated by the work of Stephansen et al. [SOO*18] where they used uncertainty to model a narcolepsy classifier, raising expectations for the discovery of other sleep-related disease markers. Incorporating this information in a visual analytics system like PerSleep could help discovering new markers.

Acknowledgment

References

[AAAA20]   ALTURKI F. A., ALSHARABI K., ABDURRAQEEB A. M., ALJALAL M.: Eeg signal analysis for diagnosing neurological disorders

using discrete wavelet transform and intelligent techniques. *Sensors 20*, 9 (2020), 2505. 2, 4

[ACD*15] AMERSHI S., CHICKERING M., DRUCKER S. M., LEE B., SIMARD P., SUH J.: Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 337–346. 3

[BLB*16] BREHMER M., LEE B., BACH B., RICHE N. H., MUNZNER T.: Timelines revisited: A design space and considerations for expressive storytelling. *IEEE Transactions on Visualization and Computer Graphics 23*, 9 (2016), 2151–2164. 7

[CVE*17] COMBRISSON E., VALLAT R., EICHENLAUB J.-B., O'REILLY C., LAJNEF T., GUILLOT A., RUBY P. M., JERBI K.: Sleep: an open-source python software for visualization, analysis, and staging of sleep data. *Frontiers in Neuroinformatics 11* (2017), 60. 3

[EKS*96] ESTER M., KRIEGEL H.-P., SANDER J., XU X., ET AL.: A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (1996), vol. 96, pp. 226–231. 5

[FLW*20] FOZOONMAYEH D., LE H. V., WITTFOTH E., GENG C., HA N., WANG J., VASILENKO M., AHN Y., WOODBRIDGE D. M.-K.: A scalable smartwatch-based medication intake detection system using distributed machine learning. *Journal of Medical Systems 44*, 4 (2020), 1–14. 1

[FvGR*20] FONSECA P., VAN GILST M. M., RADHA M., ROSS M., MOREAU A., CERNY A., ANDERER P., LONG X., VAN DIJK J. P., OVEREEM S.: Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population. *Sleep* (04 2020). zsaa048. 4

[GAW*11] GLEICHER M., ALBERS D., WALKER R., JUSUFI I., HANSEN C. D., ROBERTS J. C.: Visual comparison for information visualization. *Information Visualization 10*, 4 (2011), 289–309. 6, 7

[GBYH20] GLEICHER M., BARVE A., YU X., HEIMERL F.: Boxer: Interactive Comparison of Classifier Results. *Computer Graphics Forum* (2020). doi:10.1111/cgf.13972. 3

[GCWGvW19] GARCIA CABALLERO H. S., WESTENBERG M. A., GEBRE B., VAN WIJK J. J.: V-awake: A visual analytics approach for correcting sleep predictions from deep learning models. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 1–12. 4, 6, 7

[GK03] GRAHAM M., KENNEDY J.: Using curves to enhance parallel coordinate visualisations. In *Proceedings on Seventh International Conference on Information Visualization, 2003. IV 2003.* (2003), IEEE, pp. 10–16. 4

[HRS*17] HAAS A., ROSSBERG A., SCHUFF D. L., TITZER B. L., HOLMAN M., GOHMAN D., WAGNER L., ZAKAI A., BASTIEN J.: Bringing the web up to speed with webassembly. *SIGPLAN Not. 52*, 6 (June 2017), 185–200. URL: https://doi.org/10.1145/3140587.3062363, doi:10.1145/3140587.3062363. 7

[IRV14] IMTIAZ S. A., RODRIGUEZ-VILLEGAS E.: Recommendations for performance assessment of automatic sleep staging algorithms. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2014), pp. 5044–5047. doi:10.1109/EMBC.2014.6944758. 1

[KAS*20] KWON B. C., ANAND V., SEVERSON K. A., GHOSH S., SUN Z., FROHNERT B. I., LUNDGREN M., NG K.: Dpvis: Visual analytics with hidden markov models for disease progression pathways. *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. Early access. doi:10.1109/TVCG.2020.2985689. 3

[KCK*19] KWON B. C., CHOI M.-J., KIM J. T., CHOI E., KIM Y. B., KWON S., SUN J., CHOO J.: Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 299–309. doi:10.1109/TVCG.2018.2865027. 3

[KKE*21] KWON J.-M., KIM K.-H., EISEN H. J., CHO Y., JEON K.-H., LEE S. Y., PARK J., OH B.-H.: Artificial intelligence assessment for early detection of heart failure with preserved ejection fraction based on electrocardiographic features. *European Heart Journal-Digital Health 2*, 1 (2021), 106–116. 2, 4

[KVR*92] KEMP B., VÄRRI A., ROSA A. C., NIELSEN K. D., GADE J.: A simple format for exchange of digitized polygraphic recordings. *Electroencephalography and Clinical Neurophysiology 82*, 5 (1992), 391–393. 7

[MFK*00] MORRELL M. J., FINN L., KIM H., PEPPARD P. E., SAFWAN BADR M., YOUNG T.: Sleep fragmentation, awake blood pressure, and sleep-disordered breathing in a population-based study. *American Journal of Respiratory and Critical Care Medicine 162*, 6 (2000), 2091–2096. 5

[MOW18] MA J., OVALLE A., WOODBRIDGE D. M.-K.: Medhere: A smartwatch-based medication adherence monitoring system using machine learning and distributed computing. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2018), IEEE, pp. 4945–4948. 1

[MR01] MÉNDEZ M., RADTKE R. A.: Interactions between sleep and epilepsy. *Journal of Clinical Neurophysiology 18*, 2 (2001), 106–127. 8

[Mun14] MUNZNER T.: *Visualization analysis and design.* CRC press, 2014. 6

[NCB19] NICHOLS J. A., CHAN H. W. H., BAKER M. A.: Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews 11*, 1 (2019), 111–118. 1

[oSM*07] OF SLEEP MEDICINE A. A., ET AL.: The aasm manual for the scoring of sleep and associated events: rules, terminology and technical specifications. *Westchester, IL: American Academy of Sleep Medicine 23* (2007). 2

[PAC*19] PHAN H., ANDREOTTI F., COORAY N., CHÉN O. Y., DE VOS M.: 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 (2019), 400–410. 1, 4

[PHVG*18] PEZZOTTI N., HÖLLT T., VAN GEMERT J., LELIEVELDT B. P., EISEMANN E., VILANOVA A.: Deepeyes: Progressive visual analytics for designing deep neural networks. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 98–108. 3

[RAL*17] REN D., AMERSHI S., LEE B., SUH J., WILLIAMS J. D.: Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2017), 61–70. 3

[SDWG17] SUPRATAK A., DONG H., WU C., GUO Y.: Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 25*, 11 (2017), 1998–2008. 1, 6

[SGB*19] STROBELT H., GEHRMANN S., BEHRISCH M., PERER A., PFISTER H., RUSH A. M.: Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 353–363. 3

[SMMHH20] SIDDIQUI M. K., MORALES-MENENDEZ R., HUANG X., HUSSAIN N.: A review of epileptic seizure detection using machine learning classifiers. *Brain Informatics 7*, 1 (2020), 1–18. 10

[SOO*18] STEPHANSEN J. B., OLESEN A. N., OLSEN M., AMBATI A., LEARY E. B., MOORE H. E., CARRILLO O., LIN L., HAN F., YAN H., ET AL.: Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature communications 9*, 1 (2018), 1–15. 1, 10

[SW*81] SMITH T. F., WATERMAN M. S., ET AL.: Identification of common molecular subsequences. *Journal of Molecular Biology 147*, 1 (1981), 195–197. 4

[ZWM*18] ZHANG J., WANG Y., MOLINO P., LI L., EBERT D. S.: Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 364–373. 1, 3