Proceedings of Machine Learning Research 116:112–125, 2020 Machine Learning for Health (ML4H) at NeurIPS 2019

Sensor Fusion using Backward Shortcut Connections for Sleep Apnea Detection in Multi-Modal Data

Tom Van Steenkiste TOMD.VANSTEENKISTE@UGENT.BE
Dirk Deschrijver DIRK.DESCHRIJVER@UGENT.BE
Tom Dhaene TOM.DHAENE@UGENT.BE
IDLab
Ghent university - imec

Editors: Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

Sleep apnea is a common respiratory disorder characterized by breathing pauses during the night. Consequences of untreated sleep apnea can be severe. Still, many people remain undiagnosed due to shortages of hospital beds and trained sleep technicians. To assist in the diagnosis process, automated detection methods are being developed. Recent works have demonstrated that deep learning models can extract useful information from raw respiratory data and that such models can be used as a robust sleep apnea detector. However, trained sleep technicians take into account multiple sensor signals when annotating sleep recordings instead of relying on a single respiratory estimate. To improve the predictive performance and reliability of the models, early and late sensor fusion methods are explored in this work. In addition, a novel late sensor fusion method is proposed which uses backward shortcut connections to improve the learning of the first stages of the models. The performance of these fusion methods is analyzed using CNN as well as LSTM deep learning base-models. The results demonstrate a significant and consistent improvement in predictive performance over the single sensor methods and over the other explored sensor fusion methods, by using the proposed sensor fusion method with backward shortcut connections.

1. Introduction

Sleep apnea is a respiratory disorder consisting of breathing pauses, or apneaic events, during the night (Guilleminault et al., 1976). These events are either classified as obstructive sleep apnea (OSA) when the upper airway collapses or central sleep apnea (CSA) when the signals to control the breathing are disturbed. When breathing becomes shallow, but is not yet fully disturbed, it is classified as hypopnea.

Consequences of untreated sleep apnea can be severe ranging from hypertension, cardiac arrhythmia to even strokes and heart failure (Somers et al., 2008; Yaggi et al., 2005). It is estimated that 49.7% of male and 23.4% of female adults suffer from some form of sleeping related breathing disorder (Heinzer et al., 2015). However, many of these patients are unaware of their condition and suspected patients are faced with long waiting times for diagnosis due to expensive setups and a limited amount of hospital beds and trained sleep technicians. In the USA, waiting times of up to 60 months have been reported (Flemons et al., 2004).

Before diagnosis, patients are typically admitted for an overnight sleep study using a polysomnography (PSG) (Berry et al., 2016). This device measures a wide range of sensor data including respiration, heart rate, oxygen saturation and brain activity. After the recording, trained sleep technicians inspect the signals of the entire night and manually annotate the data for sleep events, including sleep apnea episodes, using a standard reference such as the AASM guidelines (Berry et al., 2016). The severity of sleep apnea is then summarized using the Apnea-Hypopnea-Index (AHI) which represents the number of apnea or hypopnea events per hour of sleep.

To assist in this process, and reduce the load on personnel, automated scoring methods have been investigated. These methods range from rule-based algorithms to machine learning approaches. The machine learning models generally use human-engineered features. However, thanks to advancements in deep learning technology, researchers have been able to use deep learning models for extracting features and detecting sleep apnea events in raw respiratory data. Examples include a CNN based model proposed by Haidar et al. (2017) which uses the nasal airflow signal and an LSTM based model proposed byVan Steenkiste et al. (2019) which was tested on the abdominal respiratory belt, thoracic respiratory belt and the ECG derived respiration signal.

Although these deep learning methods have offered an increase in detection performance and robustness, a crucial limitation is their inclusion of only a single respiratory channel whereas sleep technicians take into account a wide range of sensor data when analyzing the sleep of a patient. In addition to respiration, other important parameters include the heart rate (Snyder et al., 1964) and oxygen saturation (Berry et al., 2016). To further improve the performance and reliability of these deep learning models, it is crucial that data from the other recorded signals is included into the detection model.

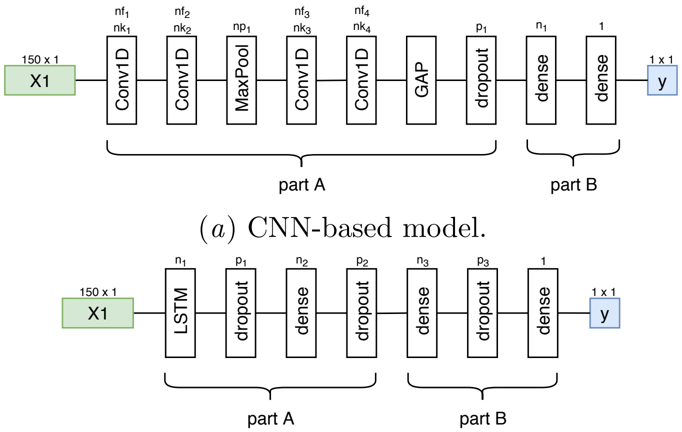
Combining data from multiple sources is known as sensor fusion and has been successfully applied across a wide range of medical use-cases including activity recognition (Gravina et al., 2017; Münzner et al., 2017), sleep quality analysis (Peng et al., 2007) and sleep detection (Chen et al., 2017). When the data is combined in the beginning of the model, the method is denoted as early fusion. When, on the other hand, the data is combined at a later stage, the method is known as late fusion. In deep learning, the early fusion method consists of a multi-input model where all data is simply concatenated. The late fusion method consists of separate branch networks for each input, that are combined at a later stage.

In this work, the performance of these two sensor fusion approaches is analyzed and a novel late fusion method using backward shortcut connections is proposed for combining data from the abdominal respiratory belt, thoracic respiratory belt, heart rate and oxygen saturation into a single deep learning model. The performance and robustness is verified using a CNN (LeCun et al., 1990) as well as an LSTM (Hochreiter and Schmidhuber, 1997) based model.

In Section 2, the base single-input deep learning models are presented. Then, in Section 3, the various sensor fusion methods are discussed and the novel approach is presented. In Section 4, the experimental setup is detailed and in Section 5 the results of the experiments are presented and discussed. Finally, conclusions are drawn in Section 6.

2. Single Input Methods

Several deep-learning architectures for automated sleep apnea detection have been proposed. Haidar et al. (2017) suggest a CNN based approach using data from the nasal airflow sensor. Van Steenkiste et al. (2019) suggest an LSTM based approach which was validated for data from the abdominal respiratory belt, thoracic respiratory belt and ECG derived respiration. However, data from these multiple sensors has not yet been combined into a single model. Based on these works, two baseline deep learning models are defined as shown in Figure 1.



(a) CNN-based model.

(b) LSTM-based model.

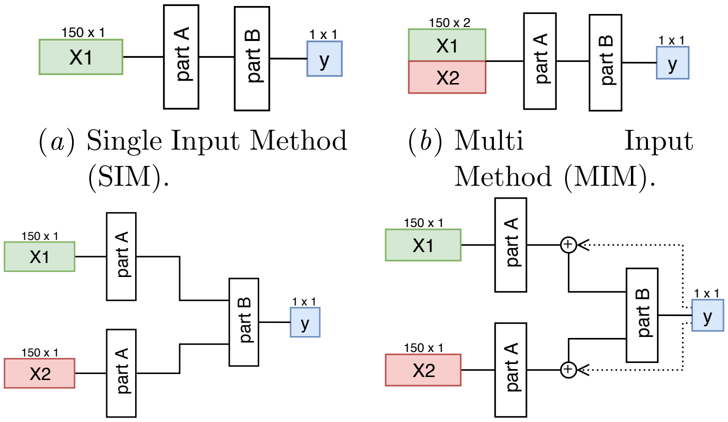
Figure 1: State-of-the-art baseline CNN and LSTM models for the automated detection of sleep apnea events in raw respiratory data. The models are split into two parts to be used by the sensor fusion methods. Here, $X1$ represents the raw respiratory data input and y represents the binary sleep apnea prediction label.

The CNN-based model in Figure 1(a) consists of two convolutional layers (LeCun et al., 1990) followed by a maxpooling layer, two more convolutional layers and a Global Average Pooling (GAP) layer. A dropout layer is added for regularization (Srivastava et al., 2014) and two dense layers provide the final predictions of the model. All layers have *relu* activations except for the final dense layer which has a *sigmoid* activation for binary prediction. The number of filters in the first two convolutional layers nf_1 and nf_2 is 100 whereas the number of filters in the final two convolutional layers nf_3 and nf_4 is 160. All convolutional layers have a kernel size nk_1, nk_2, nk_3 and nk_4 of 10. The max pooling layer has a pool size of np_1 of 3 and the dropout probability p_1 is 50%.

The LSTM-based model in Figure 1(b) consists of an LSTM layer (Hochreiter and Schmidhuber, 1997) followed by three groups of dropout and dense layers. The number of LSTM nodes n_1 is 50 and their activation function is the *tanh* function. The first two dense layers consist of n_2 and n_3 number of nodes equal to 25 with a *relu* activation function. The final dense node has a *sigmoid* activation function for binary prediction. All dropout probabilities p_1, p_2 and p_3 equal 20%.

3. Sensor Fusion Methods

Sensor fusion refers to the combination of data from multiple sensors into one single decision model. This can be achieved at three different levels: The data level, the feature level and the decision level (Gravina et al., 2017; Liggins II et al., 2017). In deep learning, features can be automatically learned and extracted from raw data. In early fusion models, the data is passed into the model as a single input matrix by concatenating all input data. In late fusion models, the deep learning architecture is split up into several branches. The different possible configurations are shown in Figure 2.



(a) Single Input Method (SIM).

(b) Multi Input Method (MIM).

(c) Branched Fusion Method (BFM).

(d) Branched Fusion Method with Backward Shortcut Connections (BFM-SC).

Figure 2: Examples of different types of methods for fusion of multiple sensors. In the examples, data from two sensors is used.

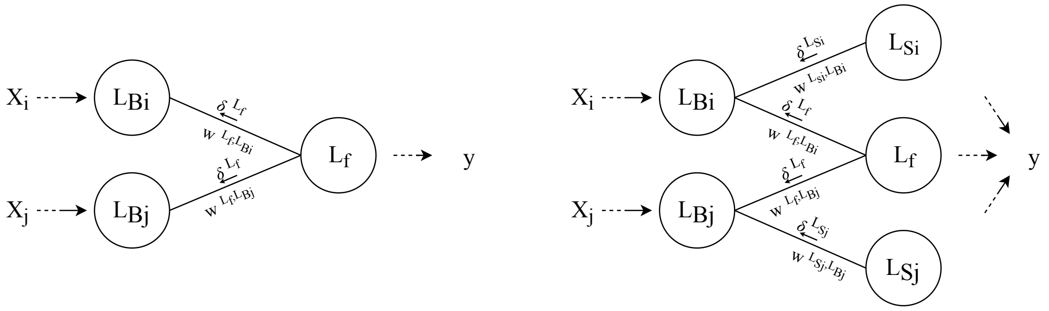
Figure 2(a) represents the Single Input Method (SIM) as presented in Section 2. The original deep learning model is split up into a *part A* representing the first stages and a *part B* representing the later stages as demonstrated in Figure 1. The SIM can only use a single input signal. However, the method can easily be extended to include multiple inputs as shown in Figure 2(b). This method is denoted as the Multi Input Method (MIM) and consists of the same architecture as the SIM but with a concatenated input matrix. Using this early fusion approach, the model is able to learn features that are correlated across the sensors. However, there is no intrinsic incentive for the model to extract as much information as possible form each input sensor individually as training and optimization procedures only focus on the combined prediction. The MIM has, for example, been applied for activity recognition (Münzner et al., 2017) based on CNN networks and sleep detection (Chen et al., 2017) based on LSTM networks using human-engineered features.

In Figure 2(c), the deep learning architecture is split up into several branches for what is denoted as the Branched Fusion Method (BFM). In this late fusion setup, correlations are being learned at the feature level instead of at the data level. This type of fusion model is often used to fuse data from multiple streams in deep learning architectures. It has, for example, been used in activity recognition (Münzner et al., 2017) based on CNN networks, sleep quality monitoring (Peng et al., 2007) based on SVM models, and driver activity anticipation (Jain et al., 2016) and medical procedure monitoring (Bernal et al., 2018) both based on LSTM networks.

A disadvantage of the BFM is that the separate branches often do not reach their full potential. There is no incentive for each of the branches to perform at their full potential. This can be derived by inspecting the error function used during the backpropagation step as defined in Equation 1 where L_{B_i} represents the final layer of branch i, L_f represents the first layer of the fusion part, $w^{L_f, L_{B_i}}$ represents the weights between the branch and the fusion part, z^{L_f} represents the weighted input to the fusion part and σ is the activation function.

$$\delta_{BFM}^{L_{B_i}} = ((w^{L_f, L_{B_i}})^T \delta^{L_f}) \odot \sigma'(z^{L_f}) \tag{1}$$

When one of the branches is able to capture the required behavior, the δ_o at the output node will be near zero. This will be propagated via δ^{L_f} to $\delta^{L_{B_i}}$ for each branch i, resulting in small updates to the branches and the network in general. For the branch or branches influencing this output the most, this is to be expected and desired as the output is close to the target value. For the other branches, however, this can critically limit the learning capacity as they are only subjected to small weight updates from that point forward. The first branch or set of branches to accurately predict the events will lead the model and the potential benefit of including the other sensors is lost.



(a) BFM.

(b) BFM-SC.

Figure 3: Error propagation in the BFM and BFM-SC. Here, L_{B_i} is the final layer of the ith branch, L_f is the first layer of the fusion part and L_{S_i} is the first layer of the shortcut connection for branch i.

In Figure 2(d), a novel approach is proposed which uses backward shortcut connections to enhance the learning capacity of each separate branch: the Branched Fusion Method with Shortcut Connections (BFM-SC). Shortcut connections have been used before, e.g. as skip connections in residual neural networks (He et al., 2016). In that setting, the connections are used to pass the input forward to the next layer. Here, they are used to pass the target back to the different branches. This incentivizes each part of the model, i.e. the separate branches and the fusion layer, to work at maximum performance. The backward shortcut connections can be implemented by adding a separate *part B* for each input branch, each with its own output. The branch thus provides input to its own *part B* and the fusion *part B*. This results in two different error measures being propagated back to the branch as visualized in Figure 3. The error $\delta^{L_{B_i}}$ for branch i then consists of the combination of the shortcut connection error $\delta^{L_{S_i}}$ and the original fusion part error δ^{L_f} as demonstrated in Equation 2.

$$\delta^{L_{B_i}}_{BFM-SC} = ((w^{L_f,L_{B_i}})^T \delta^{L_f}) \odot \sigma'(z^{L_f})/2 + ((w^{L_{S_i},L_{B_i}})^T \delta^{L_{S_i}}) \odot \sigma'(z^{L_{S_i}})/2 \tag{2}$$

When one or more of the branches of the BFM-SC take the lead, resulting in a small δ_o and thus a small δ^{L_f}, the other branches still get weight updates via their respective $\delta^{L_{S_i}}$ and hence they can continue learning. The output of these branches will continue to change, causing updates to the weights $w^{L_f,L_{B_i}}$ connecting the branch to the fusion part. Consequently, the importance of the branch in the fusion output continuously changes while the performance is increasing and branches that take longer to reach their full performance can still be included into the model. When the BFM-SC is deployed and used, the additions of to the backward shortcut connections can be removed and the model functions as a regular BFM.

4. Experimental Setup

To analyze and compare the performance of the different methods, data from the Sleep-Heart-Health-Study-1 database (Quan et al., 1997) is used which contains PSG recordings of 5804 adults of age 40 and older. From this database, five non-overlapping datasets, or groups, of 100 patients each are created. In order to accurately assess the performance for sleep apnea detection in hospital settings, there are no specific inclusion criteria and patients are simply sequentially included. Hence, the first group of patients corresponds to the first 100 patients in the SHHS-1 database. Each of the groups is split in 30 patients for training, 20 patients for validation and 50 patients for testing. An overview of patient characteristics is provided in Table 1. For privacy, age is represented as a categorical variable with five representing patients between 35 and 44 years of age and ten representing patients 85 year or older. Each step represents an increase of 10 years.

For each patient, the abdominal respiratory belt (abdores), thoracic respiratory belt (thorres), heart rate (HR) and oxygen saturation (SaO2) are extracted in addition to the human-based annotations of OSA, CSA and hypopnea. All recorded signals are resampled to 5Hz. To reduce noise and enable the deep learning models to focus on the relevant aspects of the signals, the data is passed through a basic preprocessing stage. The respiratory belt signals are filtered using a fourth-order low-pass zero-phase-shift Butterworth filter with a cut-off frequency of 0.7Hz to extract relevant respiratory information (Hettrick and Zielinski, 2006). Next, they are normalized using the 5% and 95% percentile per patient. Sensor artifacts in the HR and SaO2 signals are removed through linear interpolation of missing values. The HR signal is normalized based on a minimum and maximum heart rate of 50 and 105 beats per minute respectively. The SaO2 signal is normalized between an

Table 1: Patient characteristics for the different groups used in this study. The AHI metric demonstrates that group 2 has the most sleep apnea events per hour of sleep.

group	AHI [events/h]	BMI [kg/m^2]	gender [% male]	age [category]
0	12.8 ± 13.5	27.0 ± 4.4	37 ± 48	7.1 ± 1.1
1	13.3 ± 16.4	27.5 ± 5.7	56 ± 50	6.7 ± 1.3
2	19.6 ± 17.3	27.3 ± 4.4	40 ± 49	7.2 ± 1.1
3	13.7 ± 12.2	27.2 ± 4.9	52 ± 50	7.0 ± 1.3
4	10.9 ± 10.0	26.8 ± 4.7	55 ± 50	7.3 ± 1.3

80% and 100% saturation interval. The three annotation signals OSA, CSA and hypopnea are combined into a single binary indicator signal.

The preprocessed data are split up into epochs of 30 seconds with a stride of 1 second. If, at any point during the epoch, there was a human-based annotation of apnea, the entire epoch is labeled as apnea-positive. Apnea epochs are less common than normal sleep leading to an imbalanced data problem. For training, the dataset is balanced via random balanced sampling. For testing and validation, the dataset is not altered to accurately represent the real-life problem case.

The generated datasets are used to benchmark the performance of the different methods presented in Section 3. Each of these methods is tested using a CNN-based model as well as using an LSTM-based model. For each of the four input signals, a separate CNN and LSTM model with the SIM is constructed. The four signals are also combined to generate a CNN and LSTM model with the MIM, BFM and BFM-SC. This results in 14 different model configurations being analyzed. Each configuration is trained using the binary-crossentropy loss function with an Adam optimizer (Kingma and Ba, 2014) with learning rate set to 0.001 for 25 epochs. Gradients are clipped at 0.5 to improve learning. Early stopping based on the validation set is applied (Morgan and Bourlard, 1990; Caruana et al., 2001).

To accurately assess the performance, the different model configurations are trained and evaluated five times using a different group of patients. In addition, the robustness of the configurations is assessed by training and evaluating the models an additional four times on the first group of patients. This results in a total of nine experiments for each configuration.

The primary endpoint of this study is the performance of the model for detecting sleep apnea events. As these experiment deal with an unbalanced dataset, the area under the precision-recall curve (AUPR) is used as is recommended in literature (Kotsiantis et al., 2006; He and Garcia, 2009). To test the statistical significance of any improvement of the BFM-SC over the other configurations, a paired T-test is used. In addition, the robustness of the improvement of the BFM-SC over the different other configurations is analyzed using a heteroscedastic T-test over the five repetitions on the first group of patients. For each experiment, a significance level of $p < 0.05$ is required.

5. Results and Discussion

All experiments showed convergence before the 25 epoch threshold. The results of the five runs to test the performance of the models are shown in Table 2 and Table 3 for the CNN-based and LSTM-based configurations respectively. These results demonstrate that the BFM-SC approach consistently performs better than the other approaches. Figure 4 summarizes the performance of the different approaches.

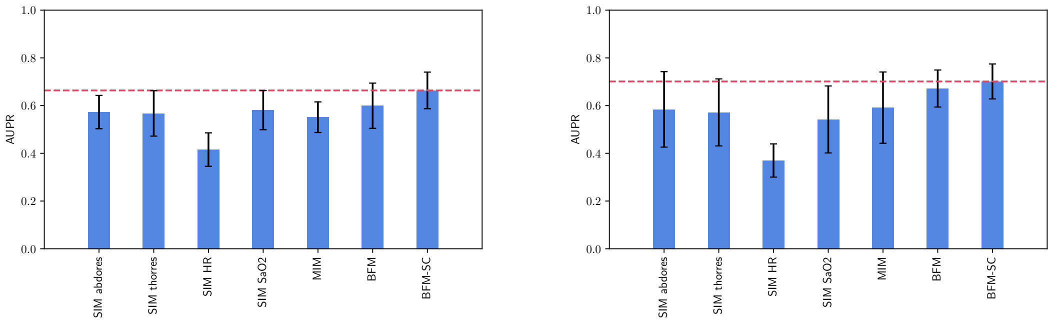
Table 2: AUPR scores for the different tested approaches using the CNN-based model. The BFM-SC outperforms all other configurations.

	group 0	group 1	group 2	group 3	group 4
SIM abdores	0.70	0.59	0.55	0.51	0.51
SIM thorres	0.64	0.55	0.69	0.54	0.41
SIM hr	0.44	0.35	0.54	0.37	0.37
SIM SaO2	0.63	0.58	0.70	0.53	0.47
MIM	0.62	0.58	0.61	0.48	0.47
BFM	0.67	0.66	0.69	0.53	0.45
BFM-SC	**0.74**	**0.68**	**0.75**	**0.57**	**0.58**

Table 3: AUPR scores for the different tested approaches using the LSTM-based model. The BFM-SC outperforms all other configurations.

	group 0	group 1	group 2	group 3	group 4
SIM abdores	0.73	0.68	0.65	0.57	0.29
SIM thorres	0.71	0.53	0.73	0.54	0.34
SIM hr	0.36	0.37	0.50	0.29	0.33
SIM SaO2	0.65	0.57	0.71	0.31	0.46
MIM	0.72	0.63	0.72	0.57	0.31
BFM	0.76	0.67	0.76	0.61	0.56
BFM-SC	**0.78**	**0.70**	**0.78**	**0.64**	**0.61**

The variation in performance across the groups can be explained by the sensitivity of the AUPR metric to the positive-negative balance in the labels. For an accurate estimation of the performance, these scores need to be compared against the performance of a random binomial model with predictions based on the balance of positive and negative events in the training data. These scores are presented in Table 4 for each of the groups. As would be expected, there is a strong correlation with the AHI values of each group presented in Table 1.



(a) CNN-based model.

(b) LSTM-based model.

Figure 4: Summary of the performance of the different models across the different datasets. The BFM-SC approach results in an improvement over all other methods. The red line indicates the mean score of the BFM-SC.

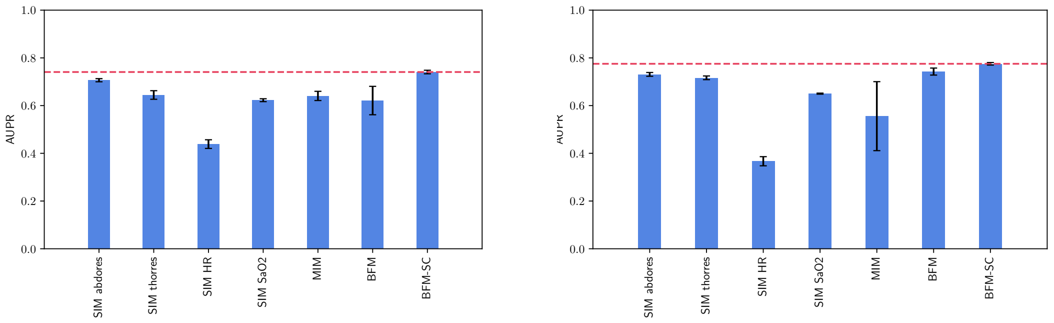
Table 4: AUPR scores for a random binomial prediction model based on the intrinsic imbalance in each training data set. The difference in balance influences the baseline performance for each group.

group 0	group 1	group 2	group 3	group 4
0.36	0.36	0.42	0.31	0.29

The robustness of the methods, analyzed by training and evaluating the different configurations five times on the first group of patients is summarized in Figure 5. Most of the configurations have robust performance except for the BFM with the CNN-based model and the MIM with the LSTM-based model. This demonstrates the impact of the different underlying base models.

The BFM-SC results in a consistent increase in performance over the single input models as well as over the other sensor fusion configurations. Statistical significance is assessed using the T-test and Table 5 demonstrates the resulting p-values for comparing the AUPR scores of each of the configurations to the AUPR scores of the BFM-SC configuration. For each experiment, the result is a $p < 0.05$, demonstrating a statistically significant improvement of the BFM-SC over the other approaches.

An added advantage of the BFM-SC approach is that it can be used to improve the interpretability of the results. Such interpretability and explainability is crucial to enable the deep learning methods to be used in clinical practice. The backward shortcut connections that are used during training can be removed at evaluation time. However, they can also be used to get a prediction per input sensor on top of the general fusion-based prediction. Figure 6 demonstrates the performance of the individual branches in the BFM-SC setting for the detection of sleep apnea. The performance is close to that of a SIM model for each



(a) CNN-based model. (b) LSTM-based model.

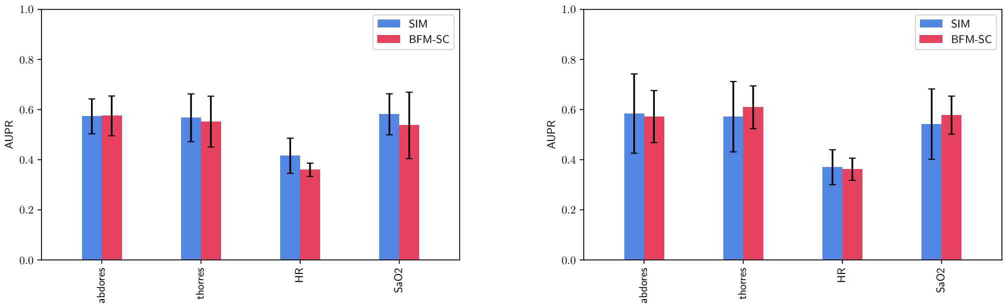
Figure 5: Summary of the robustness of the methods assessed by training and evaluating each configuration five times on the same dataset. The red line indicates the mean score of the BFM-SC.

Table 5: Resulting p-values of the statistical significance tests using a paired T-test for performance analysis and a heteroscedastic T-test for robustness analysis. All experiments demonstrate a significant improvement of the BFM-SC over the original approaches.

	CNN		LSTM	
	performance	robustness	performance	robustness
SIM abdores	0.016	0.001	0.049	0.000
SIM thorres	0.009	0.001	0.013	0.000
SIM hr	0.000	0.000	0.000	0.000
SIM SaO2	0.003	0.000	0.010	0.000
MIM	0.000	0.000	0.037	0.039
BFM	0.014	0.017	0.001	0.013

corresponding input. These extra outputs can lead to additional insights into the model decision and into which parameters are contributing to the detection.

There are some limitations to this study that warrant careful interpretation of the results. Although the performance of the BFM-SC and the other approaches was tested on five non-overlapping datasets, the analysis is still based on a single database. Validating these results on other sleep apnea databases could provide more weight to the posed conclusions. However, the SHHS-1 database is general and there is not a lot of variation in PSG measurements across other databases. The generalizability of the BFM-SC approach should also be tested on other use-cases than sleep apnea. In addition, the hyperparameters of the different models and configurations were based on the current state-of-the-art in literature.



(a) CNN-based model.

(b) LSTM-based model.

Figure 6: Performance of the individual branches in the BFM-SC for the prediction of sleep apnea events, compared to individual SIM models. The performance of each branch is close to the performance of a separately trained SIM model.

A complete analysis would require the hyperparameters to be optimized for each model, configuration and dataset. This would result in a significant computational cost. Initial experiments demonstrated that small changes to these parameters have limited impact on the results as the models have sufficient learning capacity in their current setting.

The proposed BFM-SC can be extended in several different ways. At the moment, each of the input signals is required to have the same sampling frequency. However, some signals are more informative at a higher sampling frequency whereas the learning process with other signals could be more efficient at lower frequencies. Further research should analyze methods of incorporating multi-frequency data into a BFM-SC setting. Moreover, the performance of the current configurations is impacted by the hard labeling strategy based on human annotations. The exact start and endpoints of an apnea event is not precise. Further research should include smooth labeling. Finally, the current BFM-SC setup does not enable any data-level correlations to be learned. An interesting extension to the current work is the combination of data-level fusion and feature-level fusion.

6. Conclusion

Sleep apnea is typically diagnosed using multi-sensor data from a polysomnography device. As personnel is limited and waiting times are long, automated scoring methods are being developed. Deep learning algorithms have demonstrated good performance for sleep apnea detection in a single channel of respiratory data. However, these approaches ignore relevant and important markers in other signals. Various types of sensor fusion methods have been presented for medical use-cases and for other domains. In this work, the different types of sensor fusion were analyzed and a novel method based on backward shortcut connections was presented. It was demonstrated that the BFM-SC shows a significant and consistent improvement for automated sleep apnea detection in multi-sensor data. These results enable a more reliable automated sleep apnea detection method and present an opportunity for improving sensor fusion methods in other domains.

Acknowledgments

This research received funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme.

References

Edgar A Bernal, Xitong Yang, Qun Li, Jayant Kumar, Sriganesh Madhvanath, Palghat Ramesh, and Raja Bala. Deep Temporal Multimodal Fusion for Medical Procedure Monitoring Using Wearable Sensors. *IEEE Transactions on Multimedia*, 20(1):107–118, 2018.

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Robin M Lloyd, Carole L Marcus, and Bradley V Vaughn. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications: version 2.3*. American Academy of Sleep Medicine, 2016.

Rich Caruana, Steve Lawrence, and C Lee Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in neural information processing systems*, pages 402–408, 2001.

Weixuan Chen, Akane Sano, Daniel Lopez Martinez, Sara Taylor, Andrew W McHill, Andrew J K Phillips, Laura Barger, Elizabeth B Klerman, and Rosalind W Picard. Multimodal ambulatory sleep detection. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*, pages 465–468. IEEE, 2017.

W Ward Flemons, Neil J Douglas, Samuel T Kuna, Daniel O Rodenstein, and John Wheatley. Access to diagnosis and treatment of patients with suspected sleep apnea. *American journal of respiratory and critical care medicine*, 169(6):668–672, 2004.

Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion*, 35:68–80, 2017.

Christian Guilleminault, Ara Tilkian, and William C Dement. The sleep apnea syndromes. *Ann. rev. of medicine*, 27(1):465–484, 1976.

Rim Haidar, Irena Koprinska, and Bryn Jeffries. Sleep apnea event detection from nasal airflow using convolutional neural networks. In *International Conference on Neural Information Processing*, pages 819–827. Springer, 2017.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

R Heinzer, S Vat, P Marques-Vidal, H Marti-Soler, D Andries, N Tobback, V Mooser, M Preisig, A Malhotra, G Waeber, et al. Prevalence of sleep-disordered breathing in the general population: the hypnolaus study. *The Lancet Respiratory Medicine*, 3(4): 310–318, 2015.

Douglas A Hettrick and Todd M Zielinski. Bioimpedance in cardiovascular medicine. *Encyclopedia of Medical Devices and Instrumentation*, 2006.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 3118–3125. IEEE, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Sotiris Kotsiantis, Dimitris Kanellopoulos, and Panayiotis Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.

Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

Martin Liggins II, David Hall, and James Llinas. *Handbook of multisensor data fusion: theory and practice.* CRC press, 2017.

Nelson Morgan and Hervé Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. In *Advances in neural information processing systems*, pages 630–637, 1990.

Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. Cnn-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 158–165. ACM, 2017.

Ya-Ti Peng, Ching-Yung Lin, Ming-Ting Sun, and Carol A Landis. Multimodality sensor system for long-term sleep quality monitoring. *IEEE Transactions on Biomedical Circuits and Systems*, 1(3):217–227, 2007.

Stuart F Quan, Barbara V Howard, Conrad Iber, James P Kiley, F Javier Nieto, George T O'connor, David M Rapoport, Susan Redline, John Robbins, Jonathan M Samet, et al. The sleep heart health study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

Frederick Snyder, J Allan Hobson, Donald F Morrison, and Frederick Goldfrank. Changes in respiration, heart rate, and systolic blood pressure in human sleep. *Journal of applied physiology*, 19(3):417–422, 1964.

Virend K Somers, David P White, Raouf Amin, William T Abraham, Fernando Costa, Antonio Culebras, Stephen Daniels, John S Floras, Carl E Hunt, Lyle J Olson, and Others. Sleep apnea and cardiovascular disease. *Circulation*, 118(10):1080–1111, 2008.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15 (1):1929–1958, 2014.

Tom Van Steenkiste, Willemijn Groenendaal, Dirk Deschrijver, and Tom Dhaene. Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2354–2364, 2019.

H Klar Yaggi, John Concato, Walter N Kernan, Judith H Lichtman, Lawrence M Brass, and Vahid Mohsenin. Obstructive sleep apnea as a risk factor for stroke and death. *New England Journal of Medicine*, 353(19):2034–2041, 2005.