Sleep Apnea Event Prediction Using Convolutional Neural Networks and Markov Chains

Rim Haidar
School of Computer Science
The University of Sydney
Sydney, Australia
rhai6781@uni.sydney.edu.au

Irena Koprinska
School of Computer Science
The University of Sydney
Sydney, Australia
irena.koprinska@sydney.edu.au

Bryn Jeffries
School of Computer Science
The University of Sydney
Sydney, Australia
bryn.jeffries@sydney.edu.au

Abstract—Obstructive sleep apnea is a breathing disorder affecting 2-4% of the adult population. It is characterized by periods of reduced breathing (hypopnea) or no breathing (apnea). Several machine learning algorithms have been proposed to automatically classify sleep apnea events, but little work has been done on predicting such events in advance, which is important for the treatment of sleep apnea, and especially for the development of auto-adjusting airway pressure devices to maintain continuous airflow during sleep.

In this paper, we propose three methods for predicting sleep apnea events, based on convolution neural networks and Markov chains. Specifically, we use data from respiratory signals (nasal flow, abdominal and thoracic) to predict apnea and hypopnea events in a 30-second period using the prior 60 seconds' data.

We evaluate the performance of the proposed methods for automatically learning the required features and predicting the sleep apnea events on a large dataset containing 48,000 examples from 1,507 subjects. The results show the effectiveness of the proposed convolutional neural network method, which achieved accuracy of 80.78% and F1 score of 80.63%. We also analyse the Markov chain rules and provide an overview of the transitions between apnea and normal events.

Index Terms—Convolution neural networks, Markov chains, Sleep apnea event prediction

I. INTRODUCTION

Obstructive sleep apnea (OSA) is one of the most common types of sleep-related breathing disorders, affecting 2–4% of the adult population [1]. It affects a person's ability to sleep well due to obstructions in the airway, causing loud snoring, choking and gasping for breath during sleep, and also daytime sleepiness and headaches. If left untreated, it can lead to serious problems such as heart attack, diabetes, depression, and early death [2].

The breathing problems are exhibited as abnormal events, most importantly *apnea* and *hypopnea*. An apnea event is defined as a complete or almost complete cessation of airflow for more than 10 seconds, while a hypopnea event is a reduction of the airflow of at least 30% of amplitude baseline for more than 10 seconds [2]. Fig. 1 shows examples of the nasal airflow signal during 30-second segments classed as normal and apnea.

The golden standard to diagnose OSA is through polysomnography (PSG) [3], where patients are monitored overnight, with multiple sensors attached to their bodies. These sensors measure signals from various channels, e.g. nasal

airflow, abdominal and thoracic effort, brain activity (EEG), heart rhythm (ECG) and eye movement. Sleep experts analyse the PSG recordings and manually mark the occurrence of apnea and hypopnea events. The severity of sleep apnea is determined by the number of these abnormal events during the night. However, manual inspection of long PSG recordings is very time consuming, expensive and subjective [4]. This motivates the application of machine learning approaches for automated analysis of PSG data.

Although many machine learning approaches have been developed to automatically *classify* and detect apnea and hypopnea events [5]–[8], few studies have focused on *predicting in advance* the occurrence of abnormal events. Predicting such events is important for the treatment of OSA, and especially for the development of auto-adjusting airway pressure devices to maintain continuous airflow during sleep [9]. Such machines can monitor the patient overnight, detect the onset of abnormal breathing events and automatically adjust the air pressure accordingly, e.g., by increasing the air pressure before anticipated events in order to prevent them. This requires the development of methods for accurate prediction of these events.

Hypothetically, an apnea or hypopnea event may be predictable because these events are in some way temporally proximate, such that the occurrence of an event increases the likelihood of more following soon after. Alternatively, there may be predictive features in the respiratory data that indicate that an apnea or hypopnea event is to follow. In the first case, a Markov chain model may be suitable to model the events, and predict future ones. In the latter, a supervised neural network model could learn the predictive features. In this work we have explored both approaches.

In our previous work [4], we proposed the use of convolutional neural network (CNN) models to classify 30-second data segments as apnea, hypopnea and normal, where the first two classification identified that they contained at least of one instance of the respective events. The CNN model was able to achieve high accuracy of up to 83.5% when using data from the nasal flow, thoracic and abdominal signals. By using only the raw data of the three signals, without any manual feature engineering, the CNN was able to learn the informative features for classifying OSA events. Therefore, in this work
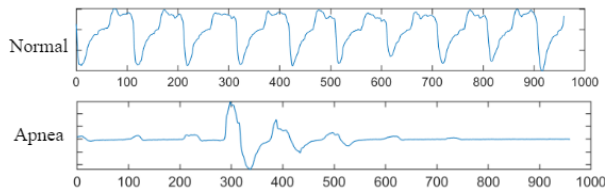
Fig. 1. Example of a 30-second segment from nasal airflow signal during Normal (N) and Apnea (A) events

we also consider CNN models but explore their efficacy in the task of learning to predict the occurrence of OSA events within a 30-second segment, using raw respiratory data from the previous 60 seconds.

We also investigate the transitions between normal and apnea events during sleep and construct Markov chain rules [10] to summarize them. The Markov chain rules are used alone and in conjunction with a CNN classifier to help in the prediction of apnea events.

The main contributions of this work are as follows:

1) We propose three novel methods built upon CNNs and Markov chains to predict OSA events. These methods use only data from respiratory channels that can be easily and non-intrusively recorded (nasal flow, abdominal and thoracic). They also operate directly on the raw data, without manual feature engineering, in contrast to the previous approaches for sleep apnea prediction. We also provide a Markov chain method based only on the true class labels for each 30-second segment of data.

2) We conduct a comprehensive evaluation using a large dataset of about 48,000 examples from 1,507 subjects. For comparison, previous work uses small datasets of up to 64 subjects. Our results show that the CNN-only method was the most accurate method achieving accuracy of 80.7% and F1 score of 80.6%, which is a promising result for practical applications.

3) We analyse the constructed Markov chain rules and provide an overview of the transitions between apnea and normal events.

II. RELATED WORK

Many approaches have been proposed to automatically detect and classify apnea and hypopnea events, using a variety of biosignals such as electrocardiogram (ECG) [5], [8], abdominal and thoracic bands [6], and instantaneous Heart Rate (IHR) and blood oxygen saturation (SpO2) [7]. Recently there have been many deep learning approaches that have been investigated for this application [11], showing excellent classification results and ability to automatically learn features from biosignals. These techniques can be applied to raw signals, or output in some processed form such as wavelet spectrograms [12]. However, only a few studies have focused on predicting apnea events in advance.

In [9], a LAMSTAR neural network was used to predict apnea and hypopnea events 30 to 120 seconds in advance. The input signals included HRV, nasal pressure, oronasal temperature, submental EMG and electrooculography. Data segments of 30, 60, 90, and 120 seconds, labeled as apnea or hypopnea, were extracted with their preceding segments of equal duration. For each preceding segment, statistical features were calculated from the discrete wavelet transform of the signals. A separate prediction model was built for each target duration and the best results were achieved using 30 and 60 seconds leading time to predict the 30-second segment ahead. The evaluation was done using data from 64 subjects.

In [13], three different neural networks (Elman, radial basis function, and feed-forward backpropagation) were used to predict apnea events from 30 to 120 seconds ahead, based on nasal flow, abdominal and thoracic signals. However, the dataset was very small — it contained recordings from only 5 subjects. Feature engineering was performed by extracting the coefficients of the wavelet packet transformation of the leading time segment and generating a set of statistical features. The best result was obtained by the feed-forward neural network, with average AUC=0.8662 over the different leading times.

A rule-based approach for predicting sleep apnea events based on statistical features extracted from ECG recordings was proposed in [14], using differential evolution to construct the rules. It uses data from the last three minutes and predicts the occurrence of apnea event in the next minute. The approach was evaluated on data from 35 subjects, achieving accuracy of 86.2%.

There is a potential to apply existing OSA event detection techniques to instead predict future events, especially in combination with Markov models which are useful for understanding the behaviour of sleep apnea. Markov models have been used to model sleep dynamics in patients with sleep apnea. In [15] Markov chains were applied to study the occurrence of sleep apnea events. An evaluation was conducted using 30-second segments from 14 subjects. It was found that the order of the best Markov chain model depends on the sleep stage and sleep period but overall a third order Markov chain was most appropriate. In [16], a Markov process was used to understand the transitions between sleep and wake in patients with OSA. The hypnograms of 113 patients were used to build a Markov transition matrix and analyse the duration of all sleep stages.

In summary, most of the previous work on predicting apnea events do not take into consideration the transitions between apnea and normal events. In addition, the proposed models require extracting features from the leading time segments, that are then used as input to machine learning algorithms. The evaluation is also conducted on small datasets with 5–64 patients. In this work, we propose new approaches which use deep learning CNN models and Markov chains for automatically learning the required features to predict sleep apnea events in advance. In addition, we analyse the learned Markov chain rules and provide an overview of the transitions between apnea and normal events. Finally, in contrast to previous studies, our evaluation is conducted using a large dataset containing 48,000 examples from 1,507 patients.
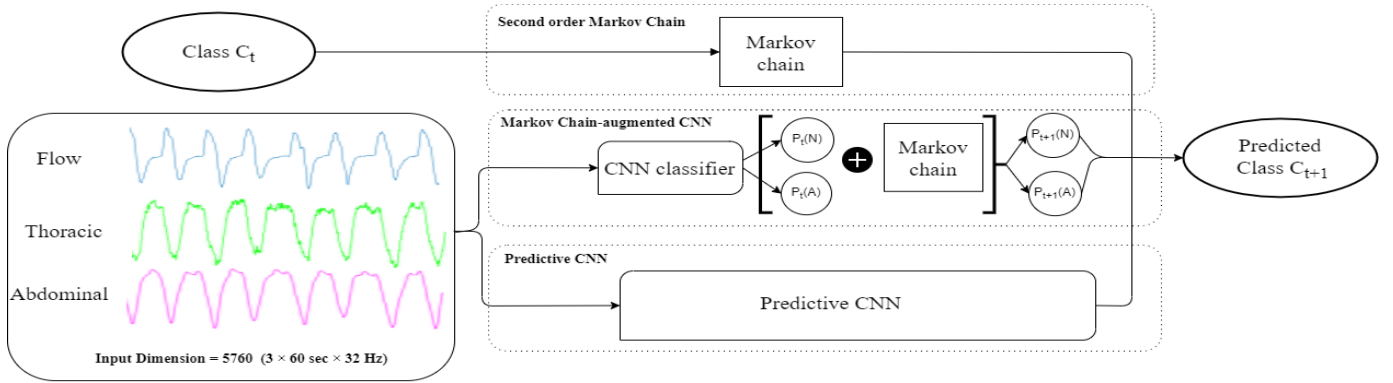
Fig. 2. Proposed approach

III. PROPOSED APPROACH

In this paper we propose and compare three approaches for predicting the presence of OSA events in a 30-second segment using data from the preceding 60 seconds. We use the three signals commonly collected in PSG systems for monitoring respiration: nasal flow, abdominal and thoracic effort. We simplify the classification task to determining whether a segment demonstrates normal breathing (N), or contains apnea or hypopnea events (A). These methods are illustrated in Fig. 2 and described below.

A. **Second order Markov Chain:** This method uses only a Markov chain prediction model. Given the type of event of the 60-second leading time segment, it predicts the type of event of the next 30-second segment. It uses the true class labels of the two 30-second leading time segments and transitional probabilities estimated from the training data (see Fig. 6). This method serves as a baseline for the following two methods. Since it requires the true class labels for the preceding data, it is unsuitable as an automated prediction tool.

B. **Markov Chain-augmented CNN:** This method uses a CNN classification model, a Markov chain and a probabilistic rule. The CNN model takes as input the raw signal data from the 60-second leading time segment to predict the probability of the type of event for this period. An aggregated version of the above Markov chain[1] of the transition probabilities for the subsequent segment's class are estimated from the training data. The rule of total probability is then used to combine the CNN and Markov chain probabilities and predict the type of event for the next 30-second segment. This approach therefore automates the classification of events in the observed data, but still relies upon statistical transition probabilities to predict the class of the next segment.

C. **Predictive CNN:** Similar to the previous method, a CNN takes as input the raw data from the 60-second leading time segment, but is trained to predict the type of event (N or A) for the next 30-second segment. It therefore

[1]As shown later, the support for some events was too low to allow a full 2nd order Markov Chain to be used in conjunction with a multi-class CNN.

avoids relying upon the statistical transition probabilities entirely.

These methods are each described in turn below.

A. Second Order Markov Chain

Markov chains [10] are probabilistic models that describe the possible transitions from one event to another, where the probability of a certain event depends only on the outcome of the previous event. A second-order Markov chain uses the state of the previous two instances to predict the probability of the next instance. In this particular application we take the known true class labels (A for the occurrence of an apnea or hypopnea event, N for normal breathing) for each 30 second segment of data in a sequence of 90 seconds' respiration, and calculate the probability of transitioning from each possible pair (AA, AN, NA, NN) to either of the possible final outcomes (A or N). Fig. 5 shows the Markov chain diagram and transitional probabilities for our data. More details about it are presented in the Section V. Using this model, a prediction can be made for new data by picking the outcome class with the highest probability, given the labels for the prior two segments.

B. Markov Chain-Augmented CNN

CNNs are one of the main types of deep neural networks. They are mostly used in computer vision and image applications showing impressive results. In the last few years they have also been applied to other types of data, e.g. for weather forecasting [17], demonstrating excellent results. The main advantage of CNNs is their ability to learn informative features from high dimensional data without manual feature engineering.

CNNs include two main types of layers: convolutional and max pooling. Convolutional layers consist of a set of filters which slide through the input height and width and convolve with the weight matrix of the filter to compute the feature maps. Convolutional nodes in same filter share the same set of weights. Each convolutional layer has a pre-specified size of kernels and strides which controls how the filters move through the input and between the layers. Max pooling layers are usually placed after the convolutional layers, their main objective is to down-sample the input features by computing
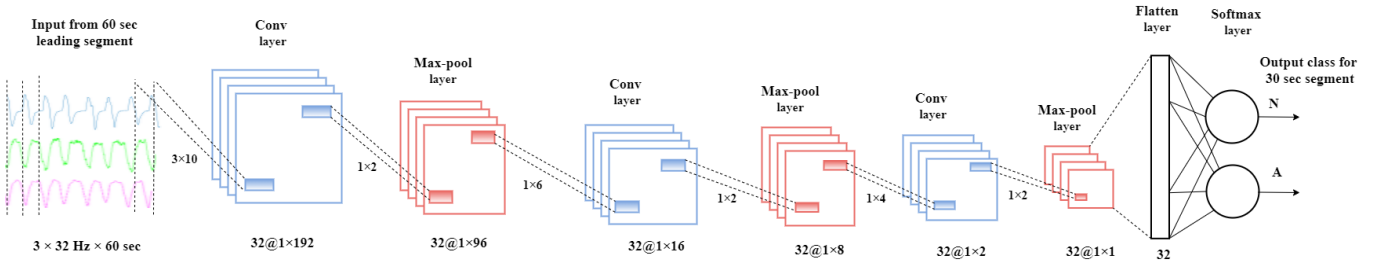
Fig. 3. Architecture of the predictive CNN model described in Section V-C. It uses data from three signals (flow, thoracic, abdominal) for the 60-second leading time segment as input and outputs the probabilities for Normal (N) and Apnea (A) events for the next 30-second segment

the maximum value of the set of adjacent convolutional nodes that are connected to them. A max pooling layer operates on each feature map independently to form a new set of the same number of pooled feature maps.

In previous work [4], [18], we showed that CNNs can successfully learn the features required to detect apnea events from PSG data. We used flow, abdominal and thoracic signals with CNNs to classify 30-second segments into normal and apnea, achieving accuracy of 83.7% and outperforming existing methods.

In this proposed method we use a CNN to estimate the probability of each class ($P_t(A)$ and $P_t(N)$) for the preceding 60 seconds, and use these as input to the equations (1) and (2) with Markov chain probabilities to estimate the probability of the next 30-second segment ($P_{t+1}(A)$ and and $P_{t+1}(N)$), again choosing the highest probability class.

For consistency with our third method (Section III-C), we first build a CNN prediction model that classifies the 60-second leading time segment. Due to the low support of apnea events, we aggregated the true class labels of both 30-second segments, classing any combination involving an apnea event (NA, AN, AA) as a single class A, and all other instances as N. The CNN therefore produced as output the probabilities for each class, $P(A)$ and $P(N)$. This dispenses with the need for the true class labels used in the previous method, and provides a probabilistic estimate rather than a binary classification.

We then built a Markov chain and estimated the conditional probabilities $P(A|A)$, $P(A|N)$, $P(N|A)$ and $P(N|N)$ from the training data, where the prior state is for the previous 60 seconds and the posterior class is for the next 30 seconds. Fig. 6 shows the Markov chain diagram and transitional probabilities for our data.

The probabilities from the CNN and Markov chain were combined using the rule of total probability:

$$P_{t+1}(A) = P_t(A) \times P(A|A) + P_t(A) \times P(A|N) \quad (1)$$

$$P_{t+1}(N) = P_t(N) \times P(N|N) + P_t(N) \times P(N|A) \quad (2)$$

where $P_t(A)$ and $P_t(N)$ are the probabilities produced by the CNN classifier at time t for the 60-second leading time segment; $P(A|A)$, $P(A|N)$, $P(N|N)$ and $P(N|N)$ are the conditional probabilities extracted from the Markov chain, and $P_{t+1}(A)$ and $P_{t+1}(N)$ are the probabilities of the two events at time $t + 1$ for the target 30-second segment.

The class corresponding to the higher probability, $P_{t+1}(A)$ or $P_{t+1}(N)$, is the predicted class for the next 30-second segment.

C. Predictive CNN

As with the above method (Section III-B), a CNN is constructed to use the raw flow, abdominal and thoracic signals from the previous 60-second segment. However, this method makes no use of the class labels of the preceding data during training, and instead is trained to predict the type of event for the next 30-second segment directly, using the posterior class true label as the training data for its output.

IV. DATA

We used the Multi-Ethnic Study of Atherosclerosis (MESA) dataset which was collected by the National Sleep Research Resource (NSRR) [19]. It includes a PSG recordings for 2,056 participants, collected through a sleep study between 2010 and 2011. Each patient has a full night recording of at least 7 hours including the following signals: nasal airflow, abdominal, thoracic, EEG, and ECG. To label the apnea events, each recording (32 Hz) was divided into 30-second segments and manually marked by sleep experts as normal, obstructive apnea, hypopnea.

Our analysis included only subjects for whom there were occurrences of obstructive apnea or hypopnea events; subjects without these events were excluded. This left us with 1,507 out of the initial 2,056 subjects.

For our experiments, we selected a balanced dataset of 47,959 segments, consisting of 23,080 normal and 23,078 apnea segments. A segment was classified as apnea if it included apnea or hypopnea events occurring for more than 10 seconds. In the 23,078 apnea segments, half contained apnea events and the other half contained hypopnea events.

For each of the 47,959 selected segments, we also extracted the previous two 30-second segments as shown in Fig. 4. Therefore, the final dataset consists of raw data from flow, abdominal and thoracic signals for a set of 90-second samples, together with the true class labels for each 30-second segment. The dataset was divided into 75% training dataset and 25% hold-out test dataset. The training dataset consists of 36,421 samples, 17,311 ending in segments classed normal and 17,310 ending in segments classed as apnea. The test dataset consists
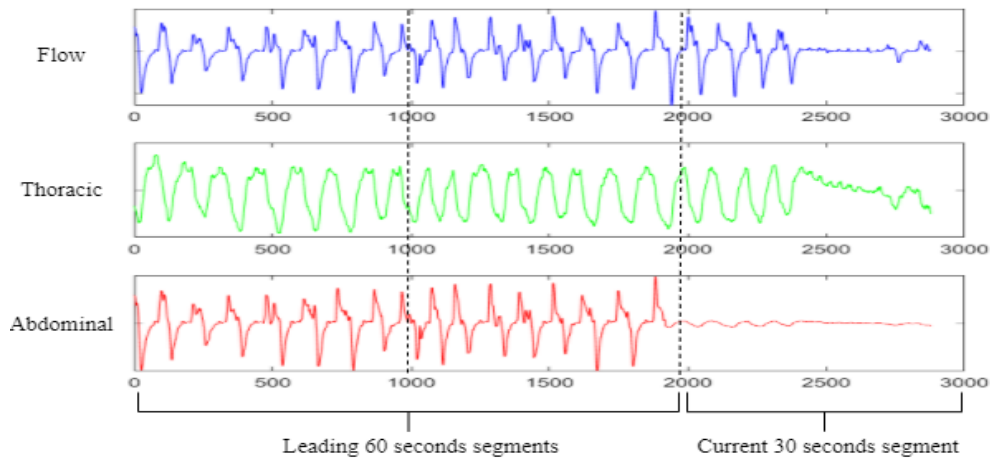
Fig. 4. Example of a 30-second segment of respiratory signal data containing an apnea event, along with the preceding 60 seconds.

of 11,538 samples (5,770 ending in normal and 5,768 ending in apnea segments).

V. EXPERIMENTS

To compare the three methods proposed in Section III, models for each were constructed using the 75% training dataset, and then evaluated using the 25% test set. In this section, we discuss the construction of each model.

A. Second Order Markov Chain

This method is considered as the baseline method, so instead of building a classifier to classify the 60-second segments into Normal or Apnea events, we used the true label of the 60-second segments to predict the next 30-second segment. Fig. 5 shows the transition probabilities learned from the 75% training dataset (Table I) that are then aggregated in Fig. 6 to construct the Markov chain; more details about the aggregation and Markov chain construction are provided in Section V-B1. It should be noted that situations where an apnea event was followed by a 30-second period without further events (AN) were very likely to result in another apnea event. This demonstrates the value of a 2$^{\text{nd}}$ order Markov Chain, since these situations would not be captured by an ordinary (1$^{\text{st}}$ order) model.

Based on Fig. 6, most of the normal prior segments have the highest probability to stay as normal in the next 30-second segment (67.52%) and the same applies for apnea events (81.01%). This baseline method therefore establishes the rule that a 60-second period of no apnea or hypopnea events is predicted to be followed by another 30 seconds without events, while any other situation (in which an apnea or hypopnea event has occurred in the past 60 seconds) is predicted to be followed by another apnea/hypopnea event.

B. Markov Chain-Augmented CNN

This method consists of two components, the CNN classifier and the Markov model, both of which were trained individually on the 75% training dataset. The models were then combined using the probabilistic rules (1) and (2) to obtain the
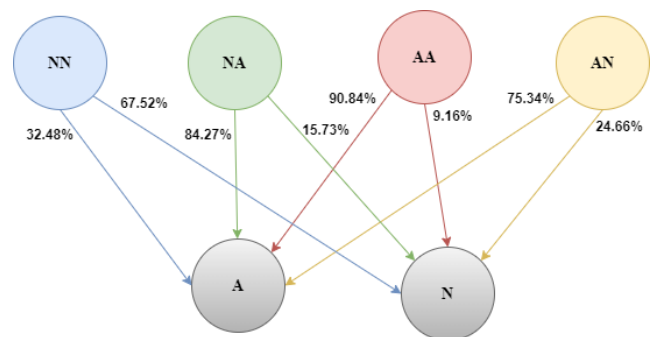

Fig. 5. Probability distribution showing the transitions between the 60-second leading time segments Normal Normal (NN), Normal Apnea (NA), Apnea Apnea (AA), Apnea Normal (AN), to the next 30-second segment (Apnea, Normal)

predictions for each sample in the 25% holdout test dataset. The probabilities $P_t(A)$ and $P_t(N)$ used from the CNN in these equations were taken as the values of the last softmax layer of the CNN model.

Details for the two components are described separately below.

1) Constructing the Markov model: The Markov chain used in this method shows the probability for an apnea and normal event in the next 30-second segment, given the sequence of events in the leading 60-second segment.

The Markov chain is constructed using the class labels of the 60-second leading segment and the 30-second target segment, estimating the probabilities from the training dataset. Since the original dataset consists of 30-second labeled segments, we re-labelled the 60-second segment by aggregating the two 30-second segments. This was done as follows:

- Label the 60-second segment as Apnea (A), if the sequence of events is Apnea then Apnea (AA), or Normal then Apnea (NA), or Apnea then Normal (AN).
- Label the 60-second segment as Normal (N), if the sequence of events is Normal then Normal (NN).

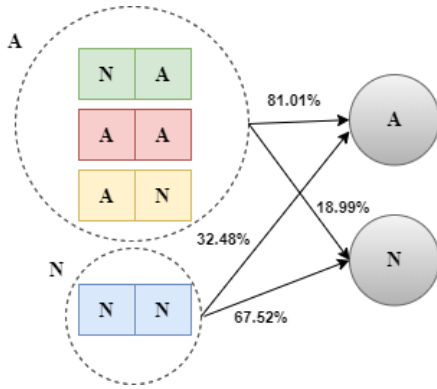In other words, a 60-second leading segment is labeled as

Fig. 6. Aggregated Markov chain showing the transitions between 60-second leading time segments to the next 30-second segment. Both lead and following segments are classed Apnea (A) if an apnea or hypopnea event occurs, otherwise Normal (N).

Apnea if any of its two 30-second segments includes an apnea event; otherwise it is labelled as Normal.

After labeling, the 75% training dataset consists of 60-second segments (22,124 normal and 12,497 apnea) while the 25% test dataset consists of 7,400 normal and 4,138 apnea events. Table I shows the sequence events for the training dataset which we used to build the Markov model, where $P(A|A) = 10124/12497 = 81.01\%$, $P(N|N) = 14938/22124 = 67.52\%$, $P(N|A) = 2373/12497 = 18.99\%$, $P(A|N) = 7186/22124 = 32.48\%$. The resulting model is shown in Fig. 6.

TABLE I
TRAINING DATASET EVENTS SEQUENCE TRANSITIONS BETWEEN
60-SECOND LEADING TIME SEGMENT AND THE NEXT 30-SECOND
SEGMENT

Leading 60 seconds		Next 30 seconds		
		Normal	*Apnea*	*Total*
	AN	1490	4553	6043
Apnea	NA	699	3746	4445
	AA	184	1825	2009
Normal	NN	14938	7186	22124
Total		17311	17310	34621

2) CNN classifier parameter selection: The classifier input for the CNN is the raw data of the three biosignals (flow, abdominal, thoracic) of the 60-second leading time segments where the input includes 5,760 features (3 signals × 60 seconds × 32 Hz). The classifier has two output nodes, corresponding to the two classes (Normal or Apnea) of the 60-second segment, where the softmax function gives that probabilities of each class.

CNNs include several hyperparameters that need to be selected and optimized, e.g., number of convolutional and max pooling layers, number of filters, kernel size, number of strides for each convolutional layer, pool size for each max pooling layer; type of activation function, optimizer and loss function. In our previous work [4], we have built a CNN model to classify apnea events as normal, obstructive apnea and

hypopnea, for 30-second segments of flow, abdominal, thoracic signals, which has shown good performance. We therefore modified this CNN structure by first doubling the input size, and reducing the number of output classes from three to two.

The hyperparameters were then manually tweaked, varying each parameter in around its initial value, to find the the parameter set with with the best average classification accuracy when evaluated using 10-fold cross-validation. The final structure of the CNN model is shown in Table II.

During training, a dropout [20] of 0.5 was used to avoid overfitting. The Adam optimizer [21] was used to minimize the categorical cross entropy function. The number of training epochs was set to 100 with a batch size of 100. All convolution layers used the Relu activation function.

C. Predictive CNN

As with the model in Section V-B2 the CNN prediction model takes as input the raw data of the flow, abdominal and thoracic signals from the preceding 60 seconds. However, the output of the CNN model is the probability for the two classes, Normal and Apnea, for the next 30-second segment.

To select the best architecture for the predictive CNN, 10-fold cross-validation was evaluated with different parameter combinations using the 75% training dataset. The model with best average accuracy was selected and evaluated on the holdout test dataset. The final selected architecture is shown in Fig. 3. It consists of three convolutional layers, three max pooling layers, and one softmax layer with two nodes, outputting the probability for class Normal and Apnea respectively for the next 30-second segment. The final predicted class is the one corresponding to the node with highest probability. Each of the convolutional layers is followed by a max pooling layer with a pool size of 2. The number of filters for the three convolutional layers was set to 32 with ReLu activation function. The kernel size for each of the three layers is as follows: Conv1 (3×10) with strides=10, Conv2 (1×6) with stride=6, Conv3 (1×4) with stride=4. The first layer kernel height was set to 3 so the filter moves through the three signals at the same time and combines the features.

During training, a dropout of 0.5 was used to avoid overfitting. The Adam optimizer was used to minimize the categorical cross entropy function. The number of training epochs was set to 300 with a batch size of 100.

TABLE II
CNN STRUCTURE FOR MARKOV CHAIN-AUGMENTED CNN METHOD

Classifier structure
Conv (32@3 × 3), Conv (32@1 × 2)
Max-pool (pool size = 2)
Conv (32@3 × 3), Conv (32@1 × 2)
Max-pool (2)
Conv (32@3 × 3), Conv (32@1 × 2)
Max-pool (pool size = 2)
Flatten()
Softmax hidden layer (number of neurons = 2)

Method	Event	Recall	Precision	F1 score	Accuracy
2nd order Markov Chain	Normal	86.76	67.65	76.02	
	Apnea	58.50	81.54	68.21	
	Average	72.63	74.59	72.07	72.62
Markov Chain-augmented CNN	Normal	86.03	70.53	77.51	
	Apnea	64.04	**82.08**	71.95	
	Average	75.037	76.31	74.73	75.04
Predictive CNN	Normal	**89.46**	**76.23**	**82.32**	
	Apnea	**87.24**	72.09	**78.94**	
	Average	**81.73**	**80.78**	**80.63**	**80.78**

VI. RESULTS AND DISCUSSION

From the Markov chain transitions in Fig. 5 and Fig. 6, we can see that if the 60-second leading time segment includes an apnea or hypopnea event, there is a high (81.01%) probability that the next 30-second segment will also include an apnea or hypopnea event. This suggests that apnea and hypopnea events tend to occur in episodes that take time to abate, confirming the expectation that they are temporally proximate. Moreover, based on Fig. 5 and Table I we can see that even when an apnea segment is followed by a normal segment (AN), there is a 75.34% chance that another apnea segment will follow. Table I also shows that the total number of 30-second apnea segments is 17,310 where 26.3% had AN in their 60-second leading time, 21.64% had NA, 10.54% had AA and 41.51% had NN. This suggests that the least probable sequence of events is AA in the leading time, then A in the next 30 seconds. This is understandable as having 3 apnea events within 90 seconds is considered a rare case and usually happens in patients with severe sleep apnea.

After applying the three methods on the 25% hold-out test dataset, the efficiency of each method was evaluated based on how many apnea or normal 30-second segments were predicted correctly using the information from the prior 60 seconds. We calculated precision, recall, F1 score and accuracy to evaluate and compare the performance of each method. The results are shown in Table III.

The best results were obtained by the predictive CNN, achieving accuracy of 80.78%, followed by the Markov chain-augmented CNN with accuracy of 75.04% and then the true-label based Markov chain with accuracy of 72.62%. The predictive CNN also outperformed the other methods in terms of all recall and F1 scores. Interestingly the Markov chain-augmented CNN performed best for precision in predicting apnea segments. This may be relevant for systems where falsely predicting such events could have important repercussions. However, it is more likely that missing a coming event has more consequence, so the much higher apnea recall of the predictive CNN is the more important measure.

The high accuracy of all the methods, considerable above the baseline of 50% for our class-balanced dataset, indicates that the occurrence of apnea events can be predicted in advance. Since both CNN models are fast to execute and require no manual input, they can be used in real time applications in conjunction with preventative techniques (such as an adaptive CPAP system) to help reduce the number of apnea events occurring during sleep.

From an application perspective, a perfect system would predict and prevent all apnea and hypopnea events. It is therefore worth comparing how the methods perform when only considering sequences where no such events occur in the preceding 60 seconds. Our test dataset contained 7,400 such instances, with 32% leading to an apnea segment. Naturally the Markov chain method predicted no apnea segments in any cases. The predictive CNN detected the apnea segments with 67.5% recall and 80.0% precision, which the Markov chain-augmented CNN achieved 55.4% recall and 71.1% precision.

The key difference between the two CNN models was the target variable on which each model was trained. Since the predictive CNN was trained with the next segment's true class, it would learn features likely to be indicative of a coming apnea or hyponea event. By contrast, the Markov chain-augmented CNN depends upon the temporal proximity of such events to make successful predictions. The fact that the former method gave such superior performance suggests that the respiratory channels contain information that can be used to predict apnea and hypopnea events before they occur.

The strong performance of the predictive CNN is also very promising given that it did not involve any manual feature engineering. The predictive CNN model was able to automatically extract and combine 5,760 input features, and these features are likely to be less prone to bias between subjects — by comparison, the transition probabilities used in the Markov chain methods are known to vary significantly between individuals depending upon the severity of a person's OSA condition [15].

VII. CONCLUSION

In this work, we proposed three new methods for predicting the occurrence of apnea events in advance, based on CNNs and Markov chains. Specifically, using a 60-second leading time, the task was to predict the type of event (normal or apnea) in the next 30-second segment. The proposed methods use data from three respiratory channels that are easily and non-intrusively recorded (nasal flow, thoracic and abdominal) and do not require any feature engineering.

The performance was evaluated on a large dataset containing 48,000 examples from 1,507 subjects. We also provided an overview of the transitions between apnea and normal events by analysing the constructed Markov chain rules. The best result was obtained by predictive CNN method, giving an accuracy of 80.78% and F1 score of 80.63%. This is a very promising result showing the potential of CNN for use in practical applications for patient monitoring and reducing the number of apnea events. Although some previous approaches have reported higher accuracies, our results are achieved with

a smaller set of input signals and tested on a much larger number of subjects' data.

The success of our methods suggests that apnea events can be predicted in advance of their occurrence using features from respiratory data, and that they are not predictable simply due to temporal correlation.

VIII. FUTURE WORK

A challenge for any deep learning approach for use in a medical application is the black-box nature of such models. We wish to investigate the features learned by the predictive CNN, to better understand what characteristics of the leading signals are predictive of an apnea event.

We intend to explore prediction of apnea and hypopnea events with longer leading times, and with an interval of time between the monitored and predicted segments, to increase the opportunity for a medical device to preemptively intervene. We also hope to develop models that use other promising biosignals, such as heart activity, which may be easier to monitor in a home environment.

While our Markov-chain augmented CNN approach was less effective than the predictive CNN, it may be possible to combine these approaches to make a model more tailored to an individual. This could be done by constructing a Markov chain model of the respiratory events exhibited by the patient, and using these to weight the predictions from the CNN.

REFERENCES

[1] L. J. Epstein, D. Kristo, P. J. Strollo *et al.*, "Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults." *Journal of Clinical Sleep Medicine*, vol. 5, no. 03, pp. 263–276, 2009.

[2] M. R. Mannarino, F. Di Filippo, and M. Pirro, "Obstructive sleep apnea syndrome," *European Journal of Internal Medicine*, vol. 23, no. 7, pp. 586–593, 2012.

[3] W. T. McNicholas, "Diagnosis of obstructive sleep apnea in adults," *Proceedings of the American Thoracic Society*, vol. 5, no. 2, pp. 154–160, 2008.

[4] R. Haidar, S. McCloskey, I. Koprinska, and B. Jeffries, "Convolutional neural networks on multiple respiratory channels to detect hypopnea and obstructive apnea events," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[5] D. Dey, S. Chaudhuri, and S. Munshi, "Obstructive sleep apnoea detection using convolutional neural network based deep learning framework," *Biomedical engineering letters*, vol. 8, no. 1, pp. 95–100, 2018.

[6] Y.-Y. Lin, H.-t. Wu, C.-A. Hsu, P.-C. Huang, Y.-H. Huang, and Y.-L. Lo, "Sleep apnea detection based on thoracic and abdominal movement signals of wearable piezoelectric bands," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 6, pp. 1533–1545, 2017.

[7] R. K. Pathinarupothi, E. S. Rangan, E. Gopalakrishnan, R. Vinaykumar, K. Soman *et al.*, "Single sensor techniques for sleep apnea diagnosis using deep learning," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 2017, pp. 524–529.

[8] A. M. da Silva Pinho, N. Pombo, and N. M. Garcia, "Sleep apnea detection using a feed-forward neural network on ECG signal," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2016, pp. 1–6.

[9] J. A. Waxman, D. Graupe, and D. W. Carley, "Automated prediction of apnea and hypopnea, using a LAMSTAR artificial neural network," *American Journal of Respiratory and Critical Care Medicine*, vol. 181, no. 7, pp. 727–733, 2010.

[10] J. G. Kemeny and J. L. Snell, *Markov chains*. Springer-Verlag, New York, 1976.

[11] S. S. Mostafa, F. Mendonça, A. G Ravelo-García, and F. Morgado-Dias, "A systematic review of detecting sleep apnea using deep learning," *Sensors*, vol. 19, no. 22, p. 4934, 2019.

[12] S. McCloskey, R. Haidar, I. Koprinska, and B. Jeffries, "Detecting hypopnea and obstructive apnea events using convolutional neural networks on wavelet spectrograms of nasal airflow," in *22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, ser. Lecture Notes in Computer Science, vol. 10937. Springer, 2018, pp. 361–372.

[13] Y. Maali and A. Al-Jumaily, "Multi neural networks investigation based sleep apnea prediction," *Procedia Computer Science*, vol. 24, pp. 97–102, 2013.

[14] I. De Falco, G. De Pietro, and G. Sannino, "On finding explicit rules for personalized forecasting of obstructive sleep apnea episodes," in *2015 IEEE International Conference on Information Reuse and Integration*. IEEE, 2015, pp. 326–333.

[15] N. Z. M. Saat, A. A. Jemain, K. Ibrahim *et al.*, "Statistical model of the occurrence of sleep apnea," *Journal of Applied Sciences*, vol. 10, no. 9, pp. 752–758, 2010.

[16] J. Kim, J.-S. Lee, P. Robinson, and D.-U. Jeong, "Markov analysis of sleep dynamics," *Physical Review Letters*, vol. 102, no. 17, p. 178104, 2009.

[17] A. Haidar and B. Verma, "Monthly rainfall forecasting using one-dimensional deep convolutional neural network," *IEEE Access*, vol. 6, pp. 69 053–69 063, 2018.

[18] R. Haidar, I. Koprinska, and B. Jeffries, "Sleep apnea event detection from nasal airflow using convolutional neural networks," in *Proceedings of 24th International Conference on Neural Information Processing (ICONIP)*, ser. Lecture Notes in Computer Science, vol. 10638. Springer, 2017, pp. 819–827.

[19] D. A. Dean, A. L. Goldberger, R. Mueller, M. Kim, M. Rueschman, D. Mobley, S. S. Sahoo, C. P. Jayapandian, L. Cui, M. G. Morrical *et al.*, "Scaling up scientific discovery in sleep medicine: the National Sleep Research Resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, 2016.

[20] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] D. Kinga and J. B. Adam, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.