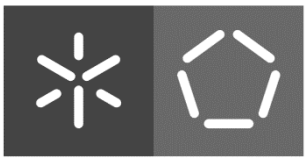
Universidade do Minho
Escola de Engenharia

Marco André Ramos Dias Prata

Using Deep Learning for Unobtrusive Sleep Stage Classification

outubro de 2018



Universidade do Minho
Escola de Engenharia

Marco André Ramos Dias Prata

Using Deep Learning for Unobtrusive Sleep Stage Classification

Dissertação de Mestrado
Mestrado Integrado em Engenharia Biomédica
Ramo de Informática Médica

Trabalho efetuado sob a orientação de
Paulo Novais
Pedro Fonseca

outubro de 2018

DECLARAÇÃO

Nome: Marco André Ramos Dias Prata

Endereço eletrónico: andreprat@gmail.com

Cartão de Cidadão: 14221158

Título da dissertação: *Using Deep Learning for Unobtrusive Sleep Stage Classification*

Orientação: Paulo Novais, Pedro Fonseca

Ano de Conclusão: 2018

Designação do Mestrado: Mestrado Integrado em Engenharia Biomédica

Área de Especialização: Informática Médica

Escola de Engenharia

Universidade do Minho, 28/10/2018

Assinatura: _____

ACKNOWLEDGEMENTS

To Pedro Fonseca, I would like to express my sincere gratitude for the amazing opportunity of coming to the Netherlands to work in Philips Research, for all the help and guidance that made me learn and professionally grow this last year.

To Mustafa Radha I would like to thank for the help, guidance and patience given throughout the year as his help was fundamental.

To professor Paulo Novais for the support, availability and guidance given along this year.

Finally, I must express my gratitude to my parents, my brother, and my soulmate for the encouragement and support given during this last year. It would not have been possible to complete this journey without them. Thank you.

ABSTRACT

Sleep represents a fundamental role to our well-being and today, as sleep disorders become more and more common, there is a growing necessity to monitor our sleep quality daily. Unobtrusive automatic sleep stage classification has made a tremendous breakthrough in this subject allowing regular users to monitor their sleep with day-to-day wearables, such as Fitbit Charge 2 tracker, contrary to the traditional manual sleep scoring based on polysomnography (PSG). Using cardiorespiratory signals to sleep stage has attracted increased attention as these signals can be obtained through unobtrusive techniques and have potential for continuous daily application.

Therefore, in this thesis, deep learning frameworks based on Long-short-memory networks (LSTMs) and Convolutional Neural Networks (CNNs) are used to sleep stage classify, either just using respiratory effort signals, for example obtained from respiratory inductance plethysmography (RIP), or using the combination of respiratory and cardiac features, often based on heart rate variability (HRV) calculated from electrocardiogram (ECG). The dataset used was the SIESTA dataset that contains a total of 294 subjects (588 PSG recordings) of which 197 are healthy subjects, 51 suffer from obstructive sleep apnea syndrome (OSA), and the remaining from a variety of sleep or sleep-related disorders. The classification problem was divided in a three-class and four-class sleep stage classification problem.

As for the results, it was obtained with respiratory data for three stages classification (Wake, rapid-eye-movement (REM) and non-REM stages) a Cohen's kappa (κ) of 0.46 for the overall pool of subjects (All), 0.50 for healthy subjects and 0.34 for OSA subjects. For four stages classification (Wake, REM, light sleep (N1/N2) and deep sleep (N3/N4) stages) it was obtained a Cohen's Kappa (κ) of 0.40 for the subject pool containing all subjects (All), 0.44 for healthy subjects and 0.31 for OSA. With cardiorespiratory data, for four stages classification, it was obtained a κ of 0.40 for the overall subject pool (All), 0.44 for healthy subjects and 0.30 for OSA subjects. With three stages, a κ of 0.46 for All subjects, 0.51 for healthy and 0.32 for OSA subjects. These results demonstrate that, with the developed frameworks, it is possible to achieve fairly good results as they are similar, in some cases moderately higher, to the current state-of-the-art but fail to generalize well, as significant differences can be found between subject types (All, Healthy and OSA).

RESUMO

O sono representa um papel fundamental no nosso bem-estar. Com o aumento de disturbios relacionados com o sono, mutio devido ao progresso técnlogico e à constante utilização de aparelhos eletronicos, existe a necessidade constante de monitorização da qualidade do mesmo. A classificação automática de estágios de sono de forma não intrusiva tem tido bastante impacto nesta matéria por permitir que utilizadores monitorizem de forma regular o seu sono através da utilização diária de "wearables", como a pulseira FitBit Charge 2, contrariamente ao método standard de classificação de estágios de sono baseado em polissomnografia (PSG). A utilização de sinais cardiorrespiratórios para classificação de estágios de sono tem ganho muito relevo devido à fácil obtenção dos mesmos através de técnicas não-invasivas e com extenso potencial para utilização continua diária.

Portanto, nesta tese, modelos deep learning baseados em Long-short-term-memory (LSTMs) e redes neuronais convolucionárias (CNNs) serão utilizados para classificação de estágios de sono, com sinais respiratórios, obtidos, por exemplo, com pletismografia respiratória por indutância (RIP) ou através da combinação de sinais respiratórios com sinais cardíacos baseados em variabilidade da frequência cardíaca (HRV) calculados através do eletrocardiograma (ECG). O dataset utilizado foi o SIESTA que contém 294 sujeitos (588 gravações PSG), dos quais 51 sofrem de sindrome de apneia obstrutiva do sono (OSA) e 94 são sujeitos saudáveis. O problema de classificação foi dividido em três classes e quatro classes de estágios de sono.

Foram obtidos, para o problema de classificação com três classes um Kappa de Cohen (κ) de 0.46 para o dataset All que contém todos os sujeitos, 0.50 para o dataset que contém apenas sujeitos saudáveis (Healthy) e 0.34 para o dataset que contém sujeitos com OSA. Para o problema de classificação de quatro classes, foi obtido um κ de 0.40 com o dataset que contém todos os sujeitos (All), 0.44 com o dataset que contém apenas sujeitos saudáveis e 0.31 com o dataset que contém sujeitos com OSA. Quanto à classificação de estágios de sono com sinais cardiorrespiratórios, com quatro classes foi obtido um κ de 0.40 com o dataset que continha todos os sujeitos (All), 0.44 com o dataset que apenas continha sujeitos saudáveis e 0.30 com o dataset que apenas continha sujeitos com OSA. Para três classes, a classificação obtida com sinais cardiorrespiratórios foi um κ igual a 0.46 com o dataset que continha todos os sujeitos (All), 0.51 com o dataset que continha apenas sujeitos saudáveis (Healthy) e 0.32 com o dataset que apenas continha pacientes com OSA. Estes resultados demonstram que, com os modelos desenvolvidos, é possível atingir resultados

moderademente satisfatórios, e, em alguns casos, ligeiramente superiores aos apresentados no estado-da-arte. No entanto, este modelos não generalizam muito bem sendo possível observar diferenças significativas entre os tipos de sujeito (All, Healthy, OSA).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

A

ANS	Autonomic Nervous System

B

BP	Blood Pressure
Bi-RNN	Bidirectional Recurrent Neural Network
Bi-LSTM	Bidirectional Long-Short-Term-Memory

C

CNN	Convolutional Neural Network

E

ECG	Electrocardiogram
EEG	Electroencephalography

F

FFT	Fast Fourier Transform

H

HRV	Heart Rate Variability
HR	Heart Rate
HF	High Frequency

L

LSTM	Long-Short-Term-Memory
LF	Low Frequency

N

NREM	Non- Rapid-Eye-Movement

O

OSA	Obstructive Sleep Apnea

P

PSG	Polysomnography
PSD	Power Spectral Density

R

RIP Respiratory Inductance Plethysmography

ReLu Rectified Linear Units

REM Rapid-Eye-Movement

RNN Recurrent Neural Network

RE Respiratory Effort

S

SWS Slow-Wave Sleep

STD Standard Deviation

STFT Short-Time Fourier Transform

T

TTS Text-To-Speech

V

VLF Virtual Low Frequency

W

WASM World Association of Sleep Medicine

1 INTRODUCTION

Sleep is, according to a simple behavioural definition, a reversible state of perceptual disengagement from and unresponsiveness to the environment [1]. Sleep represents a critical brain state and time window for the consolidation of certain types of memory, playing a fundamental role in maintaining internal homeostasis, memory consolidation, energy conservation, cognitive and behavioural performance [2], [3].

It is becoming more fundamental that we, as human species, increase our sleep quality due to our constant interaction with technology and artificial environments. Despite the advantages that this interaction might bring, it is dramatically affecting the way we sleep and sleep disorders are becoming more prominent, reaching, as the World Association of Sleep Medicine (WASM) points out, epidemic levels, that, nowadays, up to 45% of the world population suffers from sleep-related pathologies, such as insomnia, restless legs syndrome and sleep deprivation in general [4]. People often experience the symptoms without being aware of the link between these issues and their sleeping patterns [2], [5].

In order to control this phenomena, people should be empowered with the ability to easily monitor their sleep by assessing sleep quality, or sleep related problems, and to be able to adjust their sleep habits accordingly with minimum interference from an external device. To this point, the traditional sleep monitoring method, known as polysomnography (PSG) [Figure 1], is considered an obtrusive method due to the amount of electrodes and wires that have to be attached to the body during sleep. This method, even though considered the gold standard and common practice for sleep monitoring, is highly unfit for daily use has will introduce undesired sleep disturbances and can only be interpreted by highly trained sleep clinicians [6].

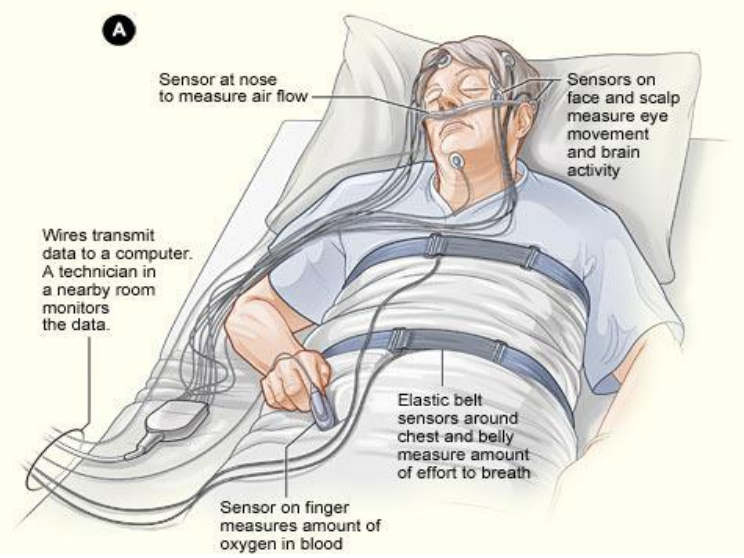


Figure 1. PSG [7].

Thus, having automatic sleep stage classification removes the human element from the equation and allows real-time sleep staging (useful for intervention studies) and remote monitoring (in-home sleep studies). Nevertheless, the acquisition of signals through PSG [7], as it was mentioned before, is quite invasive and expensive and has led to the investigation of alternative sensors and methods that allow unobtrusive sleep monitoring. This has led to the exploration of unobtrusive sleep staging using cardiorespiratory signals and body movements (actigraphy) as these signals can be meausured with unobtrusive techniques, contrary to traditional manual sleep scoring based on PSG, promising the application for personal and continuous home sleep monitoring [2], [6], [8].

1.1 Problem Description

As it was presented in previous studies [2], [9], [10], cardiorespiratory signals are considered to have great potential in the sleep analysis and sleep staging context. Even though results with cardiorespiratory signals are quite satisfactory and, almost all sleep stages can be distinguished with these signals, they still fall behind when compared with the results obtained with, for example, electroencephalography (EEG) signals obtained through obtrusive methods such as the PSG [11].
Nevertheless, there is room to improve cardiorespiratory-based sleep stage classification. In this research, the main objective focus on sleep staging, either just using respiratory effort, for example obtained from respiratory inductance plethysmography (RIP), or using the combination of respiratory and cardiac features, often based on heart rate variability (HRV) calculated from electrocardiogram

(ECG). As previous studies show, it is both possible to perform sleep staging with respiratory signals [9], [10], since human respiratory activity is associated with sleep stages throughout the night [12], and also with cardiorespiratory signals [6], [13].

In this research, there are several main differences regarding most previous work done on cardiorespiratory sleep staging:

1. The data used contains a set of healthy and disordered subjects instead of just healthy or just disordered subjects;
2. The classifier should be able to perform sleep stage classification without a priori knowledge of the condition of a subject (healthy or disordered);
3. The method of classification will be based on a deep learning framework without using manually engineered features.

1.2 Objectives and Solution Approach

One of the main research questions addressed in the present thesis is: can newly developed deep learning frameworks generalize to a specific subset of data, healthy and disordered, and still improve cardiorespiratory-based sleep staging?

As for a solution approach, one must take into consideration what type of deep learning framework would be suitable for this type of classification problem. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) are the obvious candidates to this type of task. These type of neural networks have proven to be very successful in some tasks, such as: speech recognition [14] and stock-markets [15], in the RNNs field, and image classification (face recognition, self-driving cars [16], etc.), in the CNNs field.

In this thesis, the objective would rest upon building a framework, much like others that use RNNs and CNNs combined with RNNs [11], [17] for sleep stage classification, but, unlike the previous and most state of the art, using cardiorespiratory signals to perform sleep stage classification. The potential benefit of this type of neural nets should be, when applied to cardiorespiratory signals, that it could help surpass current state of the art using the standard measurement device (PSG), enabling a home sleep-monitoring as precise as one obtained in a sleep clinic.

Regarding the objectives of this work, they are stated as follow:

1. Research on sleeping patterns and sleep architecture of a healthy adult.
    a. Respiratory and cardiorespiratory signals during sleep.

2. Improve data quality and reduce between-subject variability.
    a. Data pre-processing techniques.
3. Build a deep learning framework in order to successfully extend the existent work with cardiorespiratory data.
    a. Deep learning frameworks used on time-series data, mostly Recurrent Neural Networks (RNNs).
    b. Deep learning frameworks used to commonly distinguish features in image time-series, Convolutional Neural Networks (CNNs) combined with RNNs.

1.3 Organization of the Thesis

This report is comprised of four chapters. The first, introductory chapter, is composed by introductory novelty to sleep, unobtrusive sleep staging, problem description and objectives. In here, a resumed description of how and why unobtrusive sleep stage classification is used nowadays is presented as well it is limitations and achievements.

The following chapter, "Background", includes a theoretical description on: sleep related knowledge; cardiac and respiratory signals in different aspects of physiology; and deep learning. This chapter is the theoretical cornerstone of this report, making it indispensable prior to the reading of the remaining contents.

Chapter three, "Cardiorespiratory based sleep staging", demonstrates the findings with regard to sleep stage classification using respiratory and cardiorespiratory signals, as well as methods used, data pre-processing, dataset information and deep learning frameworks used to sleep stage classify with these signals.

The last chapter, "Conclusion" generally discusses the work presented in this thesis and answers the main research question raised before. Additionally, future work that would be interesting and promising for sleep stage classification is suggested.

2 BACKGROUND

2.1 Sleep Architecture

Sleep can clearly be defined in two separate states: rapid-eye-movement (REM) and non-rapid eye movement (NREM). Sleep is entered through NREM (pronounced non-REM) stage and NREM and REM alternate with a period of 90 to 110 minutes, in a cycle that repeats approximately four times. NREM sleep is, according to the electroencephalogram (EEG) axis, conventionally subdivided into four stages (S1-S4). The S1 and S2 stages are known as light sleep, usually with lower arousal thresholds than the other two stages. As for S3 and S4, commonly referred to as deep sleep stages or slow wave sleep, usually with higher arousal thresholds. REM stage, by contrast with NREM, is defined by EEG activation, muscle atonia, and episodic bursts of rapid eye movements, with the latter being the most commonly used marker of REM sleep phasic activity and is usually associated with muscle twitches and cardiorespiratory irregularities. Periods of marked irregularity in respiratory and heart rates are characteristic of REM sleep, in contrast with NREM sleep, during which, respiration and heart rate are regular [1], [18].

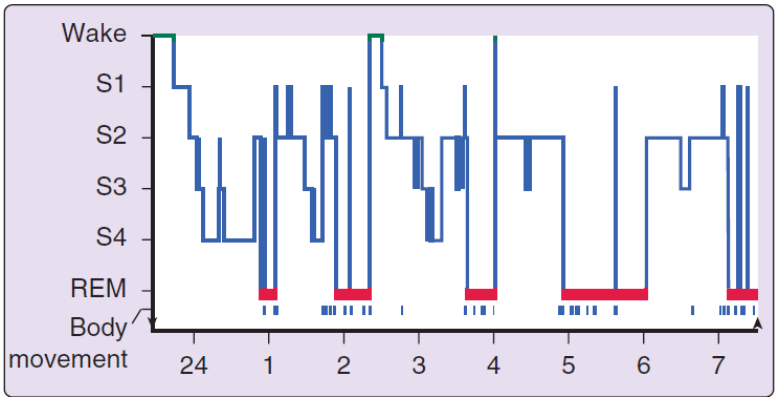


Figure 2. Hypnogram of a healthy 19-year-old man [1]

As mentioned, sleep in a healthy adult, is entered through NREM sleep, beginning with stage 1 sleep, usually persisting for only a few (1 to 7) minutes at the onset of sleep. Sleep is easily discontinued during stage 1 due to its low arousal threshold. Following stage 1, stage 2 NREM sleep, characterised by sleep spindles or K-complexes in the EEG, follows this brief episode of stage 1 sleep and continues for approximately 10 to 25 minutes. In stage 2 sleep, a more intense stimulus is required to produce arousal or awakening, and with its progress, high-voltage slow-wave activity gradually appears in the EEG [1]. Eventually, this activity meets the criteria [19] for stage 3 NREM sleep, that is, high-voltage.

Stage 3 sleep has a very short duration, just a few minutes in the first cycle as it transits to stage 4 as high-voltage slow-wave activity increases. Stage 4 NREM sleep lasts approximately 20 to 40 minutes in the first cycle, and it is identified when the high-voltage slow-wave activity comprises more than 50% of the record [1]. REM sleep in the first cycle of the night is usually short (1 to 5 minutes) as they tend to become longer across the night. The arousal threshold of REM sleep is variable throughout the entire night. NREM and REM sleep continue to alternate through the night cyclically as a series of body movements usually signal an "ascent" to lighter NREM sleep stages [1]. During the second cycle, stages 3 and 4 sleep tend to be shorter and might even disappear from later cycles, as stage 2 sleep will continuously increase occupying most of the NREM portion of the cycle [1], [18].

The average length of the first NREM-REM sleep cycle is approximately 70 to 100 minutes; the average length of the second and later cycles is approximately 90 to 120 minutes. Across the night, the average period of the NREM-REM cycle is approximately 90 to 110 minutes, where SWS predominates in the first third of the night and REM sleep the last third; Stage 1 and stage 3 sleep generally constitute 2% to 5% and 3% to 8% of sleep, respectively; Stage 2 sleep constitutes 45% to 55% of sleep and stage 4, 10% to 15% [1], [18].

Nowadays, new terminology is in place. The American Academy of Sleep Medicine (AASM) published new guidelines replacing the old terminology. According to these new guidelines, NREM sleep is now divided into 3 stages (N1-N3): N1 and N2 are used instead of stage 1 and stage 2; S3 and S4 merged into one stage, N3 [1].

2.2 Respiratory and Cardiorespiratory Sleep Staging

Sleep stages are known to be intrinsically correlated with the activity of the autonomic nervous system (ANS) [20].

When it comes to respiratory effort, it has been reported in earlier studies that some characteristics of respiration differ across sleep stages, such as respiratory frequency [12], respiratory variability and different frequency components of respiratory spectrum, enabling respiratory signals to be used in sleep staging [9].

During normal sleep, significant changes in breathing take place. During NREM sleep, breathing is remarkably regular, both in amplitude and frequency. Ventilation decreases by 13% in N2 stage and even more (15%) in slow wave sleep (N3 stage). Mean inspiratory flow is decreased, but inspiratory

duration and respiratory cycle duration are unchanged, resulting in an overall decreased tidal volume [12]. As for REM sleep, it is characterised by rapid and irregular shallow breathing [21] usually corresponding to bursts of eye movements. This breathing pattern is not controlled by the chemoreceptors but is due to the activation of the behavioural respiratory control system by REM sleep processes. So breathing during REM sleep is somewhat discordant [12], [22].

As for cardiac activity, it is characterised differently by sleep stages [2], [23]. This happens due to sympathetic and parasympathetic (or vagal) activity [2]. Between these two tones, there is a balance (the sympathovagal balance) on which one activates an action while the other suppresses it [24]. For instance, a decrease in heart rate (HR) variability is usually linked to vagal activity, while an increase is, most likely, associated with the sympathetic nerves [24]. Vagal predominance during sleep is more accentuated during NREM stages. Therefore, as we progress from wakefulness to NREM sleep, there is a noticeable decrease in HR and blood pressure (BP), which gradually become more regular, reaching their smoothest state during N3 sleep. Periods of REM and are usually associated with accelerated heart rate [1].

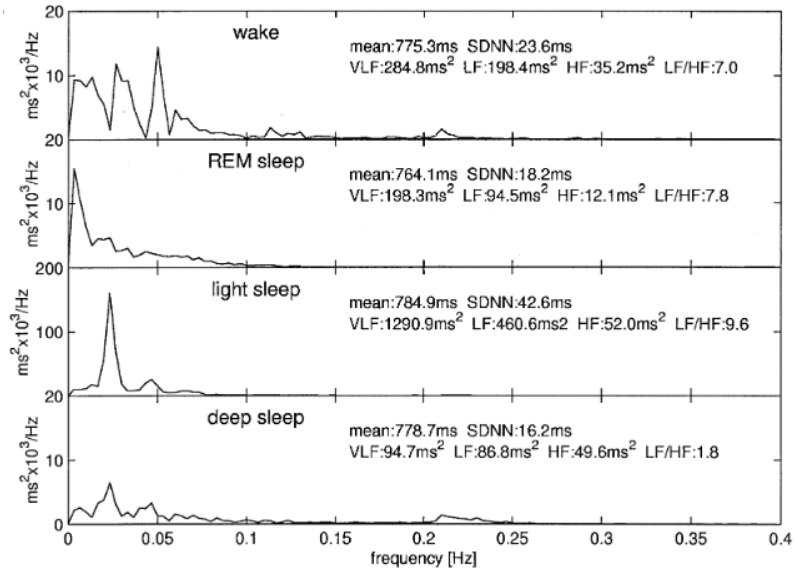


Figure 3. The figure depicts the power spectrum for light sleep, deep sleep, REM sleep, and wake for a patient with moderate sleep apnea. For the spectra the y-axis for light sleep has a different scale in order to show the pronounced virtual low frequency (VLF) peak being characteristic for sleep apnea during light sleep [25].

As it can be seen in Figure 3, and depicted in the work of [25], the inter-beat (R-R) interval time domain was analysed using the discrete Fourier transform, and it is noticeable the distinction between presented sleep stages. It was found sleep stage-specific, time domain and frequency domain changes in HR variability, particularly using spectral analysis of heart rate. Increased power in frequencies lower than 0.04 Hz band was associated with stage 2 sleep when compared to awake and slow-wave sleep states. Power in the 0.0-0.04 Hz and 0.04-0.12 Hz bands was increased in association with REM sleep when compared to non-REM sleep, and slow-wave sleep had diminished power in all frequency bands, except in the high-frequency domain, 0.15-0.4 Hz, where it appears to be the most predominant [23], [25].

As described in the literature, cardiorespiratory signals are widely used to perform sleep stage classification [2], [6], [9], [10]. However, most of the existing sleep stage classification literature focuses on data containing healthy subjects, not extending their analysis to patients with sleep disorders, such as insomnia, restless leg syndrome and obstructive sleep apnea syndrome (OSA). The latter, OSA, is arguably the most interesting for this thesis since it was assessed in the work developed by Redmond Et al. 2006 [13] and allows for a comparison with state-of-the-art on this patient group. Obstructive sleep apnea is the most common type of sleep apnea events, accounting for 84% of sleep apnea cases [26], and is caused by complete or partial obstruction of the upper airway [1]. These episodes of decreased breathing, called "apneas" (literally, "without breath"), typically last 20 to 40 seconds [27], [28]. Obstructive Sleep Apnea syndrome, OSA, it is characterised

by repetitive events of obstructive sleep apnea, shallow or paused breathing during sleep, despite the effort to breathe, and is usually associated with a reduction in blood oxygen saturation. Sleep apnea syndromes may be associated with suppression of SWS or REM sleep secondary to the sleep-related breathing problem [1]. To avoid further misconception, every time OSA is mentioned in this report henceforth, it is being referred to Obstructive Sleep Apnea Syndrome and not to the event of obstructive sleep apnea.

2.3 Deep Learning

2.3.1 Introduction

How does our brain learn? Donald Hebb, a Canadian psychologist, stated in "The organization of Behaviour" in 1949 that "When an axon of cell A is near enough to excite a cell B and repeatedly, or persistently, takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." [29]. This concept resembles the simplest way of how our brain connections work. Several scientists of Hebb's era, Locke, Hume, Mill believed that the method of learning by association was the very core of our learning foundations. In Hebb's era, neuroscientists had an approximate comprehension of how neurons worked, but he was the first to propose a mechanism in which these associations could be codified. "Neurons that fire together wire together" [29], [30].

The first artificial neuron model was proposed in 1943 by Warren McCulloch and Walter Pitts using logic gates. Its behaviour was straightforward: it turns on when the number of its active inputs surpasses a given threshold. If this threshold is one, the neuron behaves like an OR logic gate; if the threshold it is equal to the number of inputs, it behaves like an AND logic gate. Additionally, a McCulloch-Pitts neuron may prevent that other turns on, mimicking inhibitory synapses [30], [31]. What this neuron does not do is learn. For that, we have to attribute variable weights to the neuron connections.

Thus is born the Perceptron, originally invented at the end of the 1950 decade by Frank Rosenblatt, a psychologist from Cornell Aeronautical Laboratory [32]. The perceptron is a feedforward network intended to perform binary classification: a function that maps its input x (a real-valued vector) to an output value $f(x)$ (a single binary value):

$$f(x) = \begin{cases} 1 & if\ w \cdot x + b > 0 \\ 0 & otherwise \end{cases} \quad (1)$$

where w is a vector of real-valued weights, $w \cdot x$ is the dot product $\sum_{i=1}^{m} w_i x_i$, where m is the number of inputs to the perceptron and b is the bias. The bias shifts the decision boundary away from the origin and does not depend on any input value [32], [33].

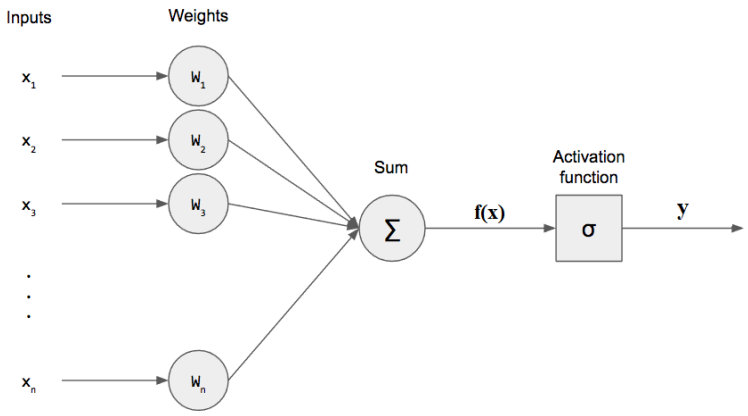


Figure 4. Artificial Neural Network (ANN).

The sum is then fed to a non-linear activation function σ that produces the output y of the neuron:

$$y = \sigma\left(\sum_{i=1}^{m} w_i x_i + b\right) = \sigma\big(f(x)\big) \quad (2)$$

From this point, a feed-forward multi-layer network can be created by chaining several neurons, as illustrated in Figure 5, $f(x) = f_3(f_2(f_1(x)))$, where f_1 is the first layer of the network (input layer), f_2 the second and f_3 the third, being the latter also called the output layer [33]. The overall length of the chain represents the depth of the model and the dimensionality of the hidden layers determines the width of the model [33].

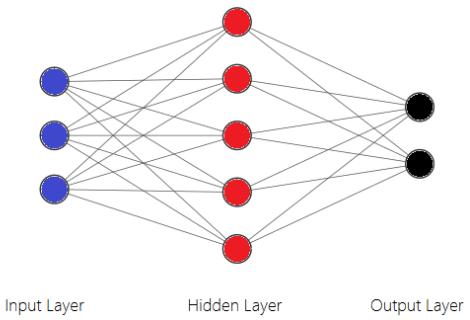


Figure 5. Feed-forward multi-layer neural network.

Usually, the input layer passes data to the hidden layers without modifying it, being the latter responsible for most of the computations. The output layer converts the hidden layer activations to an output, such as a classification. In this thesis, we will discuss convolutional networks (CNNs) (see section 2.4) and recurrent neural networks (RNNs), feedforward neural networks that are extended to include feedback connections.

2.3.2 Convolutional Neural Networks

Convolutional networks, also known as convolutional neural networks, or CNNs, much like the Perceptron, are networks inspired by biological processes, in this case, the biological visual cortex. CNNs have a neuron connectivity pattern that resembles the organisation of the visual cortex found in almost all living creatures [34], [35]. In the animal kingdom, cortical neurons respond if stimulated in a restricted region of the visual field, the receptive field. The receptive fields of different neurons are then partially overlapped covering the entire visual field extracting elementary visual features such as edges and corners [34]. This overlapping method can be mathematically approximated through convolution operations. By placing multiple receptive fields that have identical weight vectors located in different places on the image, the final result is a feature map, or activation map [33], [34]. Given an input image X, the convolutional layer convolves X with k filters {Wi}k to produce preactivation maps H.

$$H_i = W_i * X + b_i, i = 1, \dots, k \quad (3)$$

Where the symbol * denotes the convolution operation and b_i is a bias parameter [33]. These preactivation maps suffer a non-linear activation resulting in feature maps, or activation maps.

Looking at Figure 6, the first hidden layer has 16 feature maps with 5 by 5 receptive fields. In each feature map, different features are being extracted.

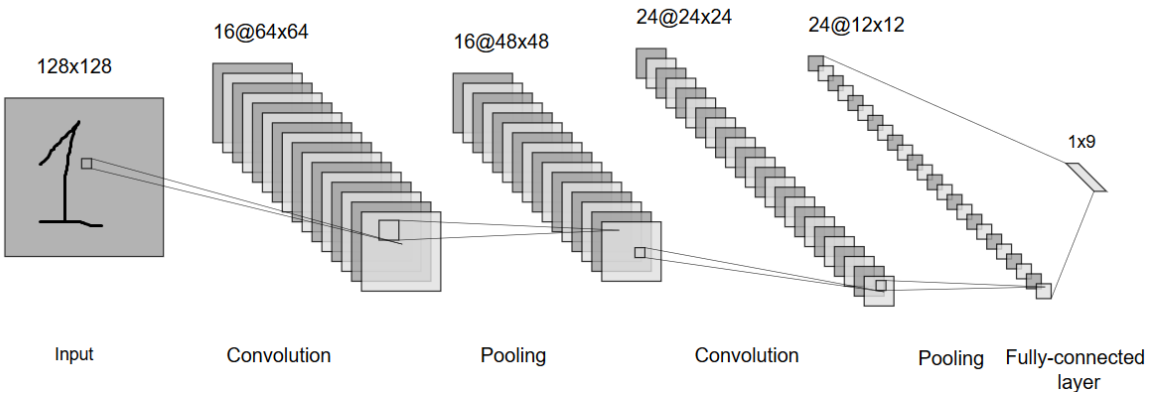


Figure 6. Example of a convolutional neural network architecture.

The architecture of these networks are usually composed by a set of convolutional and pooling (subsampling) layers (Figure 6), where convolutional layers serve as feature extractors, learning feature representations of the image, and pooling layers that perform a local averaging and subsampling of the feature maps obtained by previous layers [36].

Theoretically, layers closer to the input should learn low-level features, edges and corners, and layers close to the output should learn to combine these features to recognise more meaningful shapes [37].

2.3.3 Recurrent Neural Networks

Previous presented neural networks cannot deal with time-varying patterns, or extracting meaningful connections from time-series data, by not taking into account previous states ($t-1$), they lack persistency [38], [39]. Given a certain time-series problem, for example: classifying one sleep stage at a certain point during the night. It is unclear how a traditional neural network could use its reasoning about previous stages in the night and, accounting for past patterns, decide the present. Recurrent neural networks (RNNs) address this issue by including feedback connections [33].

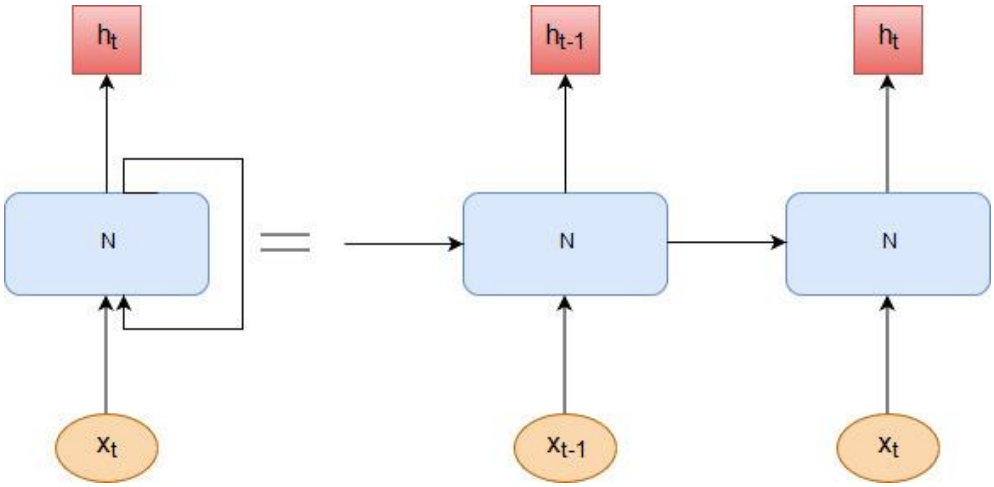


Figure 7. General structure of a regular unidirectional RNN shown (a) with a feedback connection (b) unfolded in time for two time steps.

RNNs present an ingenious way of dealing with sequential data by correlating data points that are close in the sequence [38]. Figure 7 illustrates an unfolded RNN architecture with two time-steps. A neural network, N, with an input x_t, outputs a value h_t. This chain structure makes RNNs flexible to context information, learning what to store and what to ignore, recognizing sequential patterns [40]. Eventhough RNNs present an incredible success in a variety of problems (speech recognition, translation...) they also present several drawbacks that have limited their application. Perhaps the most serious flaw of standard RNNs is that it is very difficult to get them to store information for long periods of time [41].

Storing information for more extended periods of time may be beneficial in sequence classification problems where more intrisected relations may be found. This motivated the creation of an RNN that could store more information about past states, the Long Short-Term Memory (LSTM) [42], significantly improving sequence classification problems when compared to standard RNNs.

2.3.4 Long Short-Term Memory (LSTM)

All RNNs share the same form of repeating modules. A single $tanh$ layer connecting connecting the previous state $h_{t-1}$and the present input, x_t, producing the output h_t [39]. Long Short-Term Memory (LSTM) [42] is a redesign of the simple RNN structure around special “memory cell” units. In various synthetic tasks, LSTM has been shown capable of storing and accessing information over very long timespans [43]. It has also proved advantageous in real-world domains such as speech

processing [43] and bioinformatics [42]. LSTMs are structurally similar to RNNs, except for the repeating module. Instead of having a single neural network layer, it has four, interacting with each other (Figure 8) [39], [43].

Figure 8 illustrates the complexity of operations under one LSTM cell. Brilliantly explained by

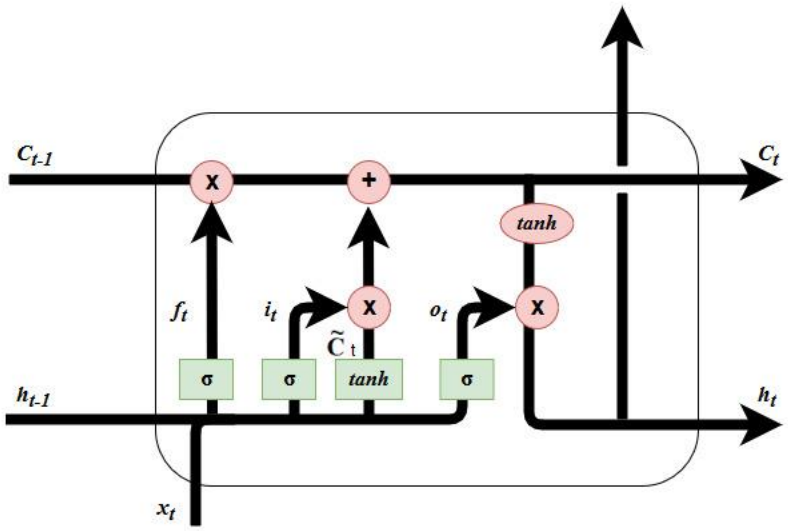


Figure 8. LSTM structure [39].

Christopher Olah in his blog "Colah's blog: Understanding LSTM Networks", LSTM structure is a combination of neural network layers (green boxes) and pointwise operations (pink circles) [39].

Unlike RNNs, LSTMs rely on the cell state, C_t, to control the flow of information through the use of gate structures, usually composed by a sigmoid neural net layer and a pointwise multiplication operation. Sigmoid layers outputs numbers between zero and one, where zero means "nothing gets through" and one the exact opposite. Three gates contol the cell state. Each gate deals with a different combination of processes [39].

The first gate, the "forget gate", decides what isn't relevant information and removes it from the cell state through the combination of the previous cell output, h_{t-1} and the present input x_t [39].

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4}$$

The second gate, "memory gate", decides what new information to store in the cell state, through the combination of two distinct processes. In the first process, a sigmoid layer, i_t, decides which

values will be updated. Second, a $tanh$ layer creates a vector of new candidate values, $\widetilde{C_t}$, that could be added to the state [39].

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (5)$$

$$\tilde{C_t} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (6)$$

The old cell state, C_{t-1}, is then updated by multiplying the old state f_t with C_{t-1} and adding $i_t * \tilde{C_t}$. These are the new candidate values, scaled by how much we decided to update each state value [39].

$$C_t = f_t * C_{t-1} + i_t * \tilde{C_t} \quad (7)$$

The third and final gate, "output gate", decides what will be the cell output, h_t, based on the filtered version of the cell state. Initially, a sigmoid layer decides what parts of the cell state are going to be outputted. Then, the cell state goes through a $tanh$ operation, making the values range between -1 and 1, and multiply it by the output of the sigmoid gate [39].

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

Even though LSTMs have greatly improved sequence classification problems, LSTMs, and RNNs, can only access information in one direction, from the past to the present ($h_{t-1} \rightarrow h_t$) [40]. It is desirable in some problems like sequence classification to exploit bidirectional access to information, from $h_{t-1} \rightarrow h_t$ and $h_{t+1} \rightarrow h_t$) [38], [44]. Bidirectional RNNs [38] combine two separate recurrent layers to scan the data forwards and backwards, providing access to all context information.

2.3.5 BIDIRECTIONAL NETWORKS

For many sequence classification tasks such as sleep stage classification, it is beneficial to have access to future as well as past context. For example, when classifying a particular sleep stage in t, it is helpful to know which one preceded it in $t-1$, and which one comes after it in $t+1$, due to temporal relation between sleep stages that, like it was mentioned before, follow a sleep cycle pattern. Standard RNNs process sequences without breakin causality [38]. Shuster and Paliwal proposed in their work, "Biderectional Recurrent Neural Networks", a structure capable of tackling the combination of future and past information [38].

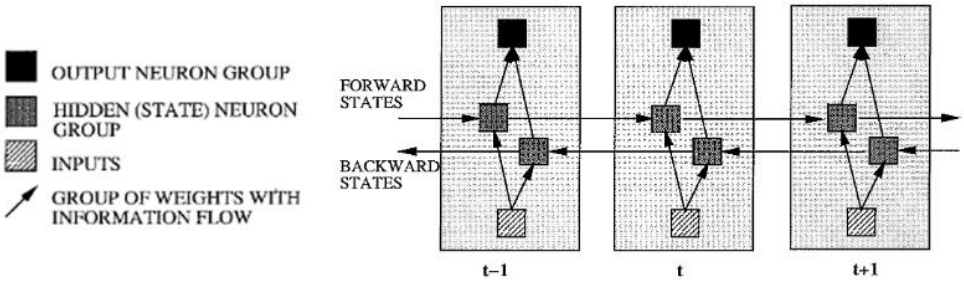


Figure 9. Structure of a bidirectional recurrent neural network [38].

By splitting the state neurons of an RNN, where one part is responsible for the forward pass (positive time direction) and the other for the backward pass (negative time direction), like presented in Figure 9 [38], [40], where the input sequence is presented to both parts in opposite directions, and the output layer are not updated until both parts process the entire input sequence [40].This structure displays improved results in various domains, by performing better than standard unidirectional RNNs.

In the scope of this thesis, Bidirectional LSTM [44] were used because they combine the benefits of long-range memory and bidirectional processing [40].

3 Methods

In this chapter, a detailed description of the approach taken towards the sleep stage classification problem defined in subsection 1.1 is presented, while the next chapters will present the results obtained, discussion and conclusion.

This chapter also includes a detailed description of the data used, pre-processing methods, deep learning frameworks and statistic procedures used to evaluate the performance of the classifiers applied to cardiorespiratory data.

3.1 DATASET

This work used the dataset collected in the SIESTA project, a project funded by the European Commission which involved several European partners. The project aimed to research the nocturnal human sleep as well as to develop and evaluate new methods of sleep analysis [45].

The SIESTA PSG dataset comprises a total of 294 subjects, where 197 are healthy individuals, 51 had been diagnosed with OSA, and amongst the rest, comprised subjects diagnosed with general anxiety disorder, depressive disorder, restless legs or Parkinson's disease. They have an average age of 52 years ± standard deviation (STD) of 17, with 126 females and 166 males. Sleep stages were manually scored on 30-s epochs by a consensus of at least two experts according to the R&K guidelines [45].

As aforementioned in subsection 1.2, the data used contains a set of healthy and disordered subjects instead of just healthy or just disordered. Therefore, it is relevant to know the sleep stage distribution across the SIESTA dataset and separate it in three groups, as it is presented in Table 1.

Table 1. Percentage of sleep stages distribution across the SIESTA dataset in distinct groups where entire night recordings were taken into account.

	Mean ± SD		
	All (%)	**Healthy (%)**	**OSA (%)**
Wake	19.98 ± 12.65	19.20 ± 11.67	18.12 ± 11.02
REM sleep	15.01 ± 5.68	15.35 ± 5.45	15.13 ± 4.92
NREM sleep	65.01 ± 9.82	65.45 ± 8.86	66.75 ± 9.50
Light sleep – N1/N2	53.83 ± 10.49	53.41 ± 9.32	59.23 ± 11.26
Deep sleep – N3	11.18 ± 6.82	12.04 ± 6.69	7.52 ± 5.69

These three groups of interest are: **All** containing all subjects, Healthy and with some type of disorder (OSA, restless legs, Parkinson's disease...); **Healthy**, just containing Healthy subjects; and **OSA**, containing patients that suffer from Obstructive Sleep Apnea syndrome.

3.2 Signal Processing

3.2.1 Respiratory Effort

In this work, two pre-processing methods were carried out on respiratory effort (RE) data and evaluated separately. These methods are going to be called Method A and Method B. The first method, Method A, consists in resampling all raw signals to the same frequency, 10 Hz. Most raw data is sampled at higher frequencies (>16 Hz), some cases even higher than 200 Hz. Then, the resampled data is filtered with low-pass and high-pass Butterworth filters, both of 3rd order with cut-off-frequencies of 0.6 and 0.05 Hz respectively, in order to eliminate high and low-frequency noise [9].

For the second method, Method B, it processes the previous method (Method A) where the baseline is filtered with a moving average filter to make it robust against motion artifacts and to remove remaining short time oscillations. Afterwards, the baseline is removed by subtracting the median peak-to-trough amplitude estimated over the entire recording. Finally, the mean, calculated over the entire recording, is also removed [6], [9].

Finally, for each 30s-epoch, a Short-Time Fourier Transform (STFT) is used to estimate the Power Spectral Density (PSD) of both pre-processed respiratory effort signals with 30-seconds FFT (Fast Fourier Transform) windows (Method A and Method B) for the frequency range between 0.1 Hz, Low Frequency (LF), and 0.5 Hz, High Frequency (HF), for a total number of 41 frequencies.

3.2.2 Cardiac

The spectrogram of the cardiac signals was computed over R-R interval time series calculated from the QRS complexes of ECG recordings. The QRS complex is the designation of three typical waveforms of an ECG, as shown in Figure 10, and the R-R interval is the time elapsed between two consecutive R waves [46].

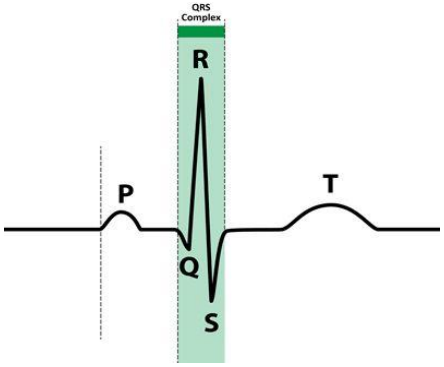


Figure 10. Typical representation of a QRS complex of an ECG recording.

The QRS complexes were detected and localised from ECG signal using a combination of a Hamilton-Tompkins detector and post-processing localisation algorithm [6].

Much like the previous subsection 3.2.1, for each 30s-epoch, a Short-Time Fourier Transform (STFT) is used to estimate the Power Spectral Density (PSD) with 30-seconds FFT windows based on the series of R-R intervals for the frequency range between virtual low frequency (VLF), 0.01 Hz, and HF, 0.4 Hz, using a total of 50 frequencies.

3.2.3 CARDIORESPIRATORY

The feature set (collection of combined frequencies of cardiac (R-R interval) and respiratory effort signals) used in this work comprises a total of 91 frequencies, expressing information about the cardiac system and the respiratory system.

As it will be clear further on, in this work four types of data were used as an input to the neural networks:

1. Resulting spectrogram from pre-processing Method A applied to respiratory effort (RE) data with 41 frequencies;
2. Resulting spectrogram from pre-processing Method B applied to respiratory effort (RE) data with 41 frequencies;
3. Pre-processing Method B applied to respiratory effort data (RE) without PSD;
4. Resulting spectrogram from pre-processing Method B applied to respiratory effort data (RE) combined with cardiac data with 91 frequencies.

3.3 Sleep Scoring, Classification & Performance Assessment

Sleep scoring is performed using supervised learning algorithms, which infer a classification output based on examples of labelled training data.

The labelled data for four-stage classification is Wake stage (W), REM stage, N1/N2 and N3 stages. Due to not being possible to distinguish very well between N1 stage from N2 stage with cardiorespiratory signals, previous works [2], [6], [9] merged N1 and N2 stages in one stage denoted as "light sleep" (N1/N2). For three-stage classification, stages N1, N2 and N3 were merged into one stage called non-REM (NREM) sleep, as the other two remained the same.

3.3.1 Classification & Performance

Classification is the process of assigning each epoch to a specific class, based on the characteristics of that epoch, which are expressed by a feature vector. In this step, there are two desired goals: building a model that 1) provides the best possible fit and that 2) is robust against variability between subjects and performs well on unseen data.

Classification requires a training phase so that the classifier can be designed and modelled to available example data, and a validation phase (in some cases, even a test phase), to estimate the error rate of the trained classifier. Training and testing samples must be different and statistically independent, in order to get reliable predictions in future classification [47].

Over time, more than one way of combining training and testing samples for error estimation has been proposed, the hold-out method and cross-validation being the standard choice in deep learning. The hold-out method is currently more used in deep learning approaches since it requires less computational power when compared to cross-validation. It consists in a simple split between training data and testing. As for cross-validation, the data is partitioned in K bins of approximately equal size and the model is trained and validated K times in a loop. The overall performance is computed based on the average performance achieved in each fold when it is part of the testing split of each cross-validation iteration. Figure 11 exemplifies a splitting of the data and training/testing according to this model.

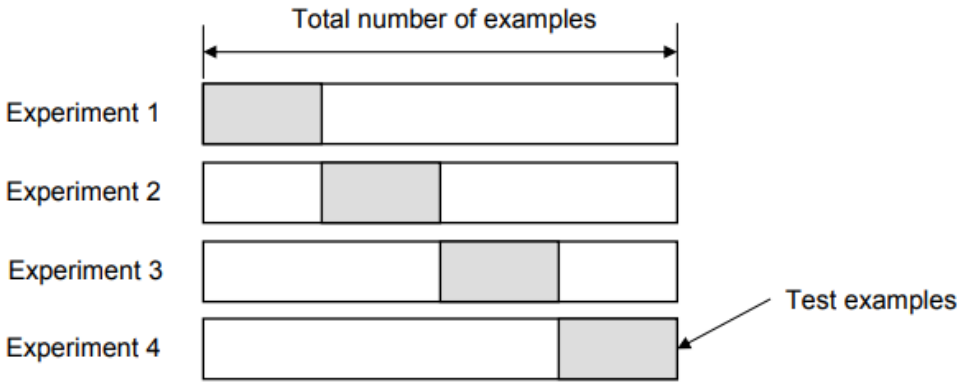


Figure 11. Sample K-fold cross-validation with K=4.

The idea is that, on each experiment, K-1 folds are used for training and the remaining one for testing. The main advantage of this method is that at the end of the training all examples have been used for both training and testing. Cross-validation, when compared to hold-out, will be less sensitive to the partitioning of the data set and more accurate. For this thesis, a 4-fold cross-validation method was used in all experiments. It is also relevant to mention that since each subject had two sessions (two recorded nights), the same subject was always part of the same dataset during cross-validation. In other words, when separating the subjects for training and validation, both recordings of each subject would exclusively be part only of the training or of the validation set, but never of both.

3.3.2 Performance Assessment

The performance will be assessed based on metric evaluation and statistical analysis. In the metric performance, accuracy and Cohen's Kappa (κ) [48] will be used to evaluate the results. Accuracy is the ratio of correct versus all predictions made. This is the most common evaluation metric for classification problems, but it can be misleading in the case of an imbalance in the number of observations in each class (which often happens) and when prediction errors are not equally important. In this specific case, sleep stages are never equally distributed. As described in the Dataset section, N1/N2 occupies more than 50% of the night. The accuracy of a trivial sleep stage which would assign N1/N2 to each epoch would reach at least 50%, giving a misleading picture of the overall performance of the classifier. Cohen's kappa coefficient of agreement, or κ for short, measures inter-rater agreement for qualitative items and it is generally thought to be more robust than simpler percentage agreement calculations as it takes into account the possibility of the

agreement occurring by chance. Cohen's kappa formula, presented in the following equation, where PO is the overall proportion of observed agreement and PE is the overall proportion of agreement expected by chance.

$$\kappa = \frac{\mathrm{PO} - \mathrm{PE}}{1 - PE} \tag{10}$$

By expecting some agreement to randomly occur and subtracting it from the observed agreement, kappa becomes robust to the problem of class imbalance. Kappa values range from -1 to 1 where κ value of one means perfect agreement and a κ value of negative one means total disagreement, whereas a κ value of zero means that the agreement is equal to chance. Table 2 offers an interpretation organised in ranges [49].

Table 2. Cohen's kappa agreement.

Kappa	***Agreement***
<0	Poor
0 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.	Almost Perfect

To complement the interpretation of the accuracy and Cohen's kappa results, statistical analysis is conducted in this thesis using the Wilcoxon signed-rank test and Wilcoxon rank-sum test. These are non-parametric statistical hypothesis tests used to compare two samples in order to assess whether their population mean ranks differ [50]. Wilcoxon signed-rank test is used for paired samples and Wilcoxon rank-sum test for unpaired (independent) samples [50]. These tests can be used as an alternative to the paired and unpaired Student's t-test.

For this practical work, two lists of Cohen's Kappa of values will be used to statistically compare two subject pools (All versus Healthy), two data types of the same subject pool (All with pre-processing Method A versus All with pre-processing Method B), Cardiorespiratory data versus Respiratory.

3.4 Deep Learning Frameworks (Models)

Before moving to the explanation of the classification frameworks, their design and why these type of architectures were used for sleep staging, there are a few deep learning concepts that should be explained beforehand.

3.4.1 Concepts

Concepts like LSTMs, CNNs, Bidirectional, and some activations layers (sigmoid, tahn) where already described in chapter 2.3. Others like batch, epoch, iteration, loss, backpropagation, optimisation algorithms, regularization concepts, belong to the core of, not only deep learning but to machine learning in general. These concepts are crucial to understand the functioning of all deep learning frameworks presented in this thesis.

Batch size, epoch and iteration are among the simplest to explain: Epoch represents a complete forward and backward pass done by the network on the entire dataset; Batch size represents the number of training samples inside a batch, whereas a batch is a division, or part, of the training data; an iteration represents the number of batches needed for the network to complete one epoch. Batch size is a significant topic of discussion since it can have some impact on the overall performance of the network. Bigger batch sizes tend to be less computational efficient [51]. Furthermore, in the context of this thesis, training epoch should not be confused with a sleep epoch which is a 30-second segment which served as the basis for visual scoring of sleep stages.

Backpropagation is an algorithm that uses backward propagation of the error calculated by a loss function to successively update the neural network weights. Loss, or objective function, quantifies the difference between the neural network output and the desired output. The quantification of this error directly results in less error in the next iteration as it is back-propagated to update the neural network weights in such way that predictions successively generate less error. As for the optimisation algorithms, they exist to minimise or maximise an objective function (loss).

A central problem in deep learning is how to make an algorithm that will perform well not just on the training data, but also on new unseen data. In order to reduce predictions errors, some techniques known as regularization techniques, are designed to reduce the test error at the expense of increased training error. In this work, Dropout regularisation [52] and L2 regularization were both used in all frameworks.

3.4.2 System Architecture

Presented in this subsection is a representation of the overall system architecture that comprises all frameworks implemented in this thesis. Figure 12 illustrates the overall system architecture.

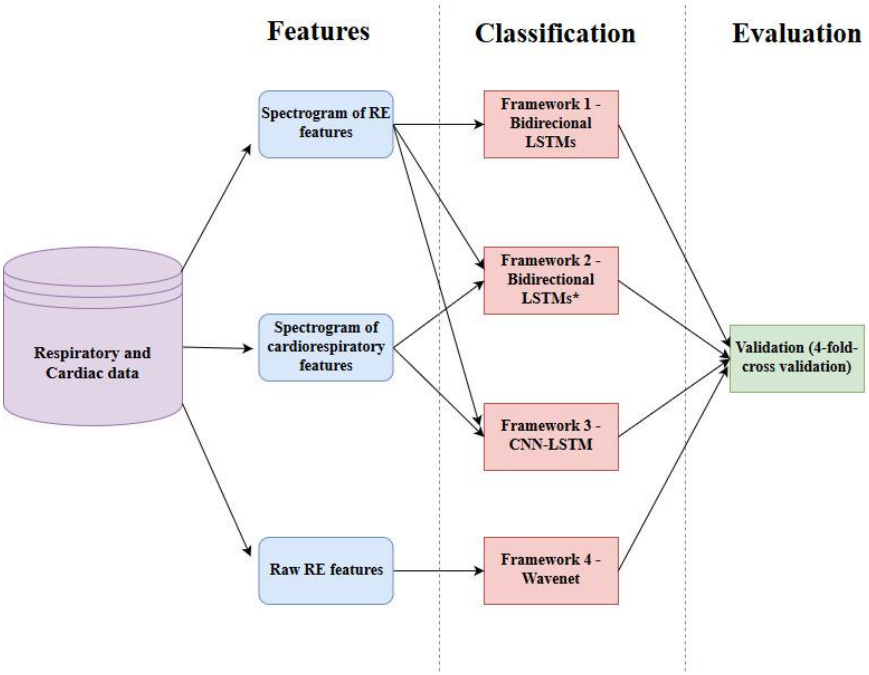


Figure 12. Analytic pipeline of this work.

This system architecture presented above summarizes the structure of the system. The initial phase of the system is comprised by the dataset that contains respiratory and cardiorespiratory data. The second phase is constituted by "Features", processed data and unprocessed data. The second phase, the classification, is constituted by the frameworks used for classification in this thesis. The last phase is constituted by the evaluation of each of these classifiers.

The input of each classifier is defined as:

1. Spectrogram of respiratory effort (RE) features: comprised by spectrograms obtained with pre-processing Method A or pre-processing Method B with respiratory effort (RE) data, both with 41 frequencies;
2. Spectrogram of cardiorespiratory features: spectrogram from pre-processing Method B applied to respiratory effort data (RE) combined with cardiac data with 91 frequencies;
3. Raw RE features: respiratory effort data (RE) pre-processed with Method B without PSD.

Following, necessary notations are presented in Table 3.

Table 3. Mathematical notations used in this thesis.

Symbol	***Definition***	***Dimensionality***
x_t, x_t*, x_t**	Spectrogram feature	$n_i \times 41$, $n_i \times 41 \times 10^*$, $n_i \times 91 \times 10^{**}$
j_t,j_t*	Spectrogram feature	$n_i \times 82$, $n_i \times 50^*$
X_t	Raw time series	$N_i \times 1$
⊕	Concatenation	

3.4.3 BI-LSTM – FRAMEWORK 1

The presented framework on Figure 13 is based on several approaches found across literature, that combine the powerful RNNs, in this case, LSTMs, with Bidirectional layers [53]–[56].

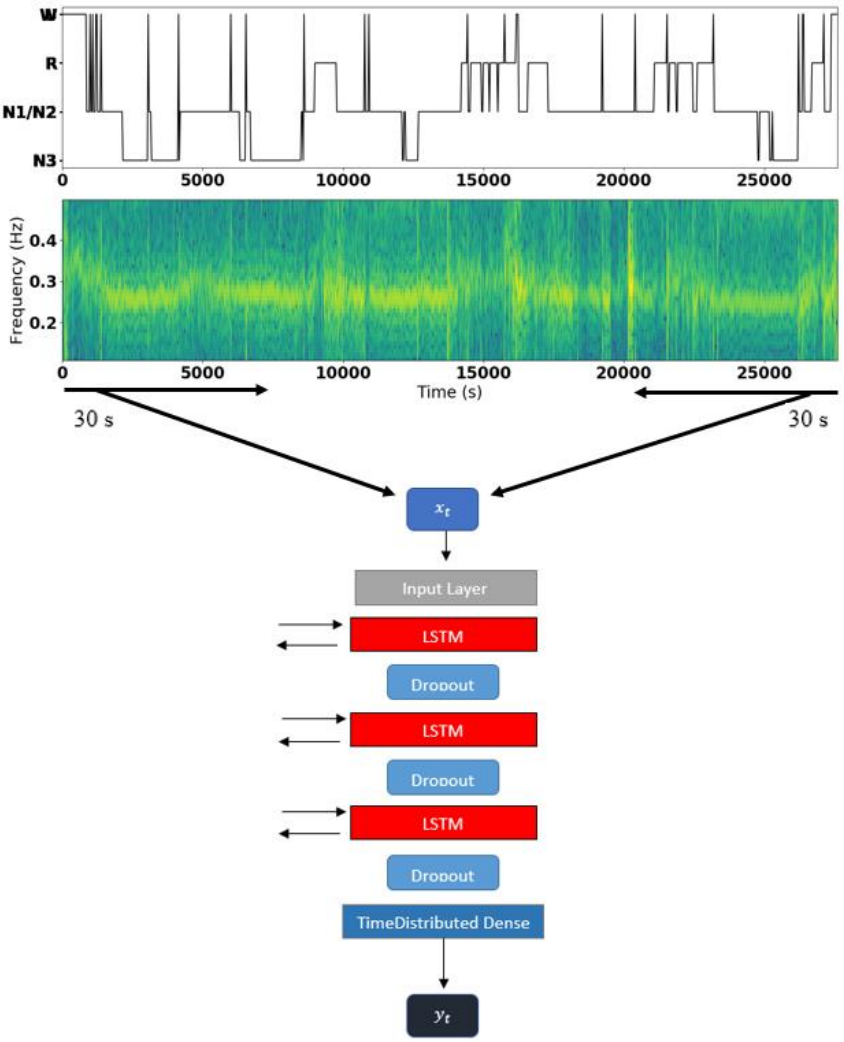


Figure 13. Ground truth label and spectrum obtained on the same healthy subject; Diagram of the LSTM framework 1.

This framework can learn the temporal dependency of x_t with a three stacked Bi-LSTM combined with a densely connected layer (Perceptron) at the end of the model. Each LSTM layer is Bidirectional, each pass containing 64 units, totalling in 128 final units. The output, y_t (equation X as described in section LSTM) is a TimeDistributed softmax. This wrapper (TimeDistributed) applies a layer to every temporal slice of an input, resulting in each 30-second segment getting the same treatment. The input x_t, is a spectrogram with n_i 30-second segments directly aligned with the ground truth labels.

$$y_t = Softmax\left(W_{yh}h_t + b_y\right) \tag{11}$$

where h_t is the currently hidden state value, and y_h is the output from the previous hidden state.

Several versions of this framework were experimented, with more and fewer LSTM layers, with BatchNormalization [57] and Dense as input layers in order to reduce variance was carried out but unsuccessfully and will not be described further in this thesis.

Parameters of training:

- Dropout rate: [0.2, 0.5,0.5]
- Weight Regularization L2: [0.001]
- Number of Hidden units per LSTM layer: [64,64,64]
- Batch size: [Number of sessions/2]
- Optimizer: Adam

3.4.4 Bi-LSTMx2 – Framework 2

For the second framework, its creation was based in the previous model and its limitation of only accepting one input at a time, not enabling the possibility of combining two different signals. This model was also created to tackle the problem of imbalanced classes. As it was mentioned, N1/N2 occupy roughly 50% of the night, and deep sleep (N3) stages are the least common. To compensate for this imbalance, it was necessary to introduce more information about deep sleep stages. The 30-second FFT window method previously used does not capture slow transitions, mainly deep sleep transitions, and to improve that it was experimented an alternative method so that slow transitions may be more captured. With the alternative method, a 120-seconds FFT window was calculated, instead of 30-seconds, with 90-seconds overlapping. The resulting spectrogram, $j_t \in$ Rni×82 is presented in figure 20. As it can be seen, the spectrogram looks more "smooth" as slower variations are more captured instead of fast ones.

As for the framework, it takes two inputs: x_t, already explained in previous framework, a spectrogram with 30-seconds segments and j_t, the spectrogram with 120 seconds size segments with 90-seconds overlap, explained at the beginning of this subsection.

There is a noticeable difference between the first framework and the second. Now, the inputs are separated in two cells (three stacked Bi-LSTMs in each). These cells are slightly different, one has a more direct approach, same as in framework 1, and the other applies a light treatment or de-noising layer with a Perceptron layer (Dense) with sigmoid activation to the input. In the end, both results from each cell are concatenated and introduced in the decision layer, a TimeDistributed softmax

layer to generate the final sleep stage prediction y described in equation 11, as it is equal to the first framework (Figure 13).

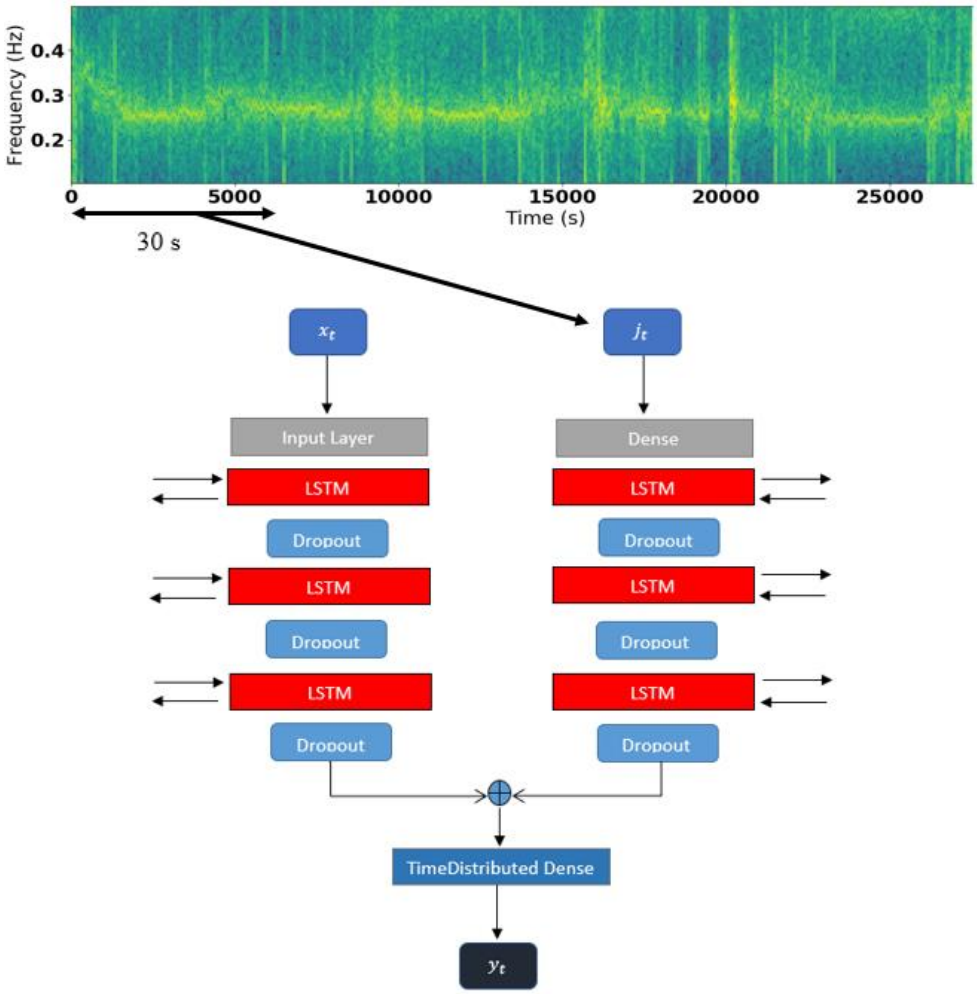


Figure 14. 120s based spectrogram (j_t) and Framework 2 structure.

Parameters of training:

- Dropout rate: [0.2, 0.5,0.5]
- Weight Regularization L2: [0.001]
- Number of Hidden units per LSTM layer: [64, 64, 64]
- Number of Hidden units in Dense layer: [16 or 8]
- Batch size: [Number of sessions/2]
- Optimizer: Adam

3.4.5 CNN-LSTM – FRAMEWORK 3

By combining a CNN with an RNN (LSTM), we can obtain a hybrid model which is able to extract features present in the spectrogram and preserve the long-term temporal relationship present in cardiorespiratory data [56], [58].

In this framework, a 2D CNN first processes the spectrogram over all non-overlapping 30-second windows of data to learn feature representation. Here, "spatial" features are extracted. This feature representation is then concatenated with the original input values and passed to an LSTM model, which learns the temporal dependency present of the spatial feature extracted by the CNN.

Following, it is presented the overall system architecture of this framework.

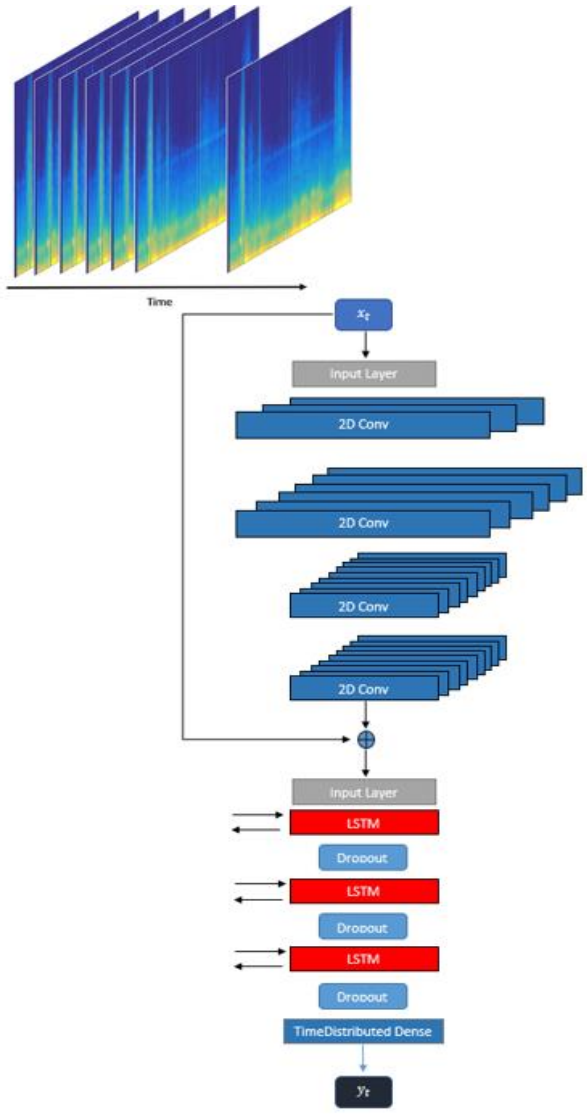


Figure 15. Architecture of framework 3.

The respiratory and cardiorespiratory ($\boldsymbol{X_t}$) were also represented in the frequency domain using FFTs. The spectrogram was obtained by segmenting each 30-seconds epoch in 10 sub-epochs that are 3-seconds long. Further, the resulting spectrogram $x_t * \in \mathbb{R}^{n_i \times 41 \times 10}$, where each spectrogram has a dimension of 41 × 10, and n_i is the number of 30-second windows in the entire signal. Pictured above is represented the resulting input tensor to the network, a stack of 2D images (spectrograms) [58]. The same method was used to obtain $x_t **$ which represents the spectrogram of the cardiorespiratory signal.

As for the framework, it takes the input $x_t{}^*$, or $x_t{}^{**}$, through a BatchNormalization layer, followed by a TimeDistributed 2D convolution layer with a kernel size of (5, 3) and stride of (2, 1). Following, a max-pooling layer with a (2, 1) pooling and a Rectifier Linear Unit (ReLu) [59] as non-linear activation layer. After that, a 1 x 1 convolution is used to deepen the model, followed by another ReLu activation layer. Two convolution layers with kernel size of (3, 3) and strides (2, 1) followed by a regularization layer (Dropout). Then, the $x_t{}^*$ is dimensionally reduced to $x_t * \in \mathbb{R}^{n_i \times 410}$, $x_t ** \in \mathbb{R}^{n_i \times 910}$ by a Flatten layer, as so is the output of the convolutions. These two tensors are then concatenated and the resulting tensor is given to a familiar three stacked Bi-LSTM layers, already explained previously, producing the output $\boldsymbol{y_t}$ described in equation 11.

Parameters of training:

- Dropout rate: [0.2, 0.5, 0.5]
- Weight Regularization L2: [0.001]
- Number of Hidden units per LSTM layer: [64, 64, 64]
- Filter size in CNNs: [(5×4), (1×1), (3×3), (3×3)]
- Optimizer: RMSProp

3.4.6 WAVENET – FRAMEWORK 4

Over the last few years, several systems have been developed to generate speech from text. Thus, generating speech from text became a common and desirable feature, thanks to the popularity of softwares, such as: Apple's Siri, Microsoft's Cortana, Google Assistant, among others [60]. Deep Mind proposed in 2016 a deep generative model for raw audio waveforms, "WAVENET: A GENERATIVE MODEL FOR RAW AUDIO" [61], capable of generating realistic-sounding human-like voices through the sampling of real human speech modelling, directly, waveforms [60]. This model

is able to create accurate models with different voices, accents and tones of the input correlating with the output (i.e, if it is trained with German people, it produces German speech) [60]–[62].

Based on the work of van Den Oord Et al, the WaveNet model is very similar to PixelCNNs [63], where the conditional probability distribution is modelled by a set of convolutional layers. No pooling layers are present in this network, and the output of this model has the same temporal dimensionality as the input. Also, the model outputs a categorical distribution related to the next value. This is achieved with a softmax layer and is optimized to maximize the loglikelihood of the data related to the parameters [61]. The WaveNet is mainly described in six components, such as:

- Dilated Causal Convolutions: The principal advantage of a WaveNet are casual convolutions. These convolutions guarantee that the model cannot violate the order that the data is modelled;
- Softmax distributions: Softmax distributions tend to work better to modeling the conditional distributions over the individual audio samples, even when the data is implicitly continuous. A reason for this is that the categorical distribution is more flexible and can easily model arbitrary distribution, since it does not make assumptions about the shape.
- Gated Activation Units: The same gated activation units that are used on PixelCNN are used here. Also, it was observed that non-linearity works significantly better than rectified linear activation functions, when audio signals are the input;
- Residual and Skip Connections: Both these types of connections are used throughout the network. The main objective of this is to speed up the convergence of models and enable training of deeper models;
- Conditional Wavenets: Given an additional input, the wavenets can model the conditional distribution of the audio. Also, conditioning the model on alternative input variables, the wavenet can generate audio with the required characteristics.
- Context Stacks: Several ways can be used to increase the receptive field size of a wavenet, such as: the number of dilation stages, the quantity of layers, the size of the layers, better dilation factors and/or a combination of all. Another approach is to use separate, smaller context stack to process a great part of the audio signal and conditions a larger wavenet that processes a smaller part of an audio signal. Also, it can use several context stacks with varying lengths and number of hidden units. Finally, context stacks with bigger receptive fields have fewer units per layer and can have pooling layers running at a lower frequency.

Some tests made with this technique showed that it can outperform the already existing Google text-to-speech (TTS) systems, despite that it is less credible than actual human speech [61] and it requires much more computational processing power when compared to other TTS systems in real world applications [64].

To the context of sleep stage classification, the WaveNet is attractive as it obtains relevant information from raw signals. Other frameworks here presented have as an input a spectrogram and some relevant respiratory effort features might be lost with this processing method. Therefore, the idea of using pre-processed with Method B raw respiratory effort signals without a spectrogram ($\boldsymbol{X_t}$) was to obtain important information from the raw signal that in other models was being lost.

Raw Respiratory effort signals are introduced in framework 4, for each patient they can be represented by Xt ∈ R Ni × 1 where Ni is the number of points in the signal. As for ni it stands for the number of 30-second windows in the entire recording. Credit to the implementation of this model to Bas Veeling GitHub repository "WaveNet implementation in Keras"[1].

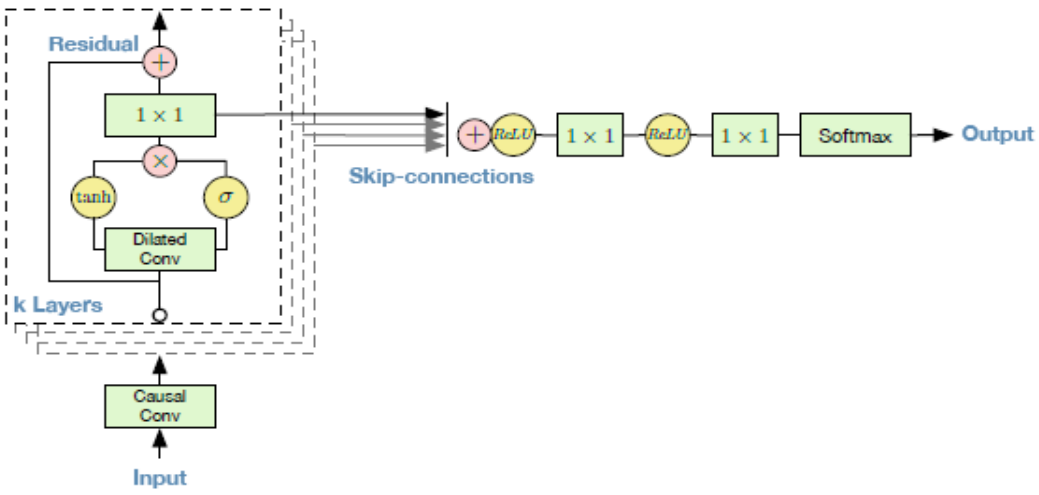


Figure 16. Representation of the Wavenet model [61].

As for this framework performance, it is fairly poor and random as, in terms of accuracy, never surpasses the 50% threshold suggesting that it is randomly picking. With this framework, no relevant result was obtained. Values of Kappa were around 0.05 and its loss never decreased (Figure 20, Appendix B), proving the fact that it was severely underperforming, and it did not learn anything from

[1] https://github.com/basveeling/wavenet

sleep data used as input. Therefore, no results or further discussion will be conducted on this framework.

4 RESULTS

Following the guidelines described in Section 3.1.3, all recordings presented in this chapter were evaluated in four and three-stage classification tasks using the four frameworks described in subsection 3.1.4 and 4-fold cross-validation. The results were separately analysed in two sets, Method A and Method B data. Inside these, the subject pool was then divided and analysed in three classes, like aforementioned in subsection "Dataset": All, containing all subjects (Healthy and disordered); Healthy, just containing Healthy subjects; and OSA, containing patients that suffer from Obstructive Sleep Apnea syndrome.

The overall presentation of the results section is separated in:

- Results obtained with each framework;
- Head-to-head comparison of results obtained with each framework;
- Comparison of RE versus Cardiorespiratory sleep staging.

There is a statistical component, aforementioned in section 3.2.2, that was calculated in order to infer significant statistical differences between frameworks, data types and between subject pools. Specifically, it was used a Wilcoxon signed-rank test and a Wilcoxon rank-sum test. The first, Wilcoxon signed-rank test, was used to statistically compare paired samples, in this case, data pre-processed with Method A and data pre-processed with Method B (Healthy Method A and Healthy Method B, All Method A and All Method B...), Framework 1 and Framework 2, for example. The second, Wilcoxon rank-sum test, was used to statically compare unpaired (independent) samples, in this case, All Method A data and OSA Method A data. The latter was only conducted between same processing type, e.g. between All Method A and OSA Method A and never between All Method A and Healthy Method B, for example.

4.1 Framework 1 – Results & Performance of Respiratory Data

As illustrated in Figure 17, and similarly in all tested subsets, the maximum average validation performance is obtained between ~200-300 epochs, with an average Kappa of 0.39 for four stages classification and a Kappa of 0.45 for three stages classification. Also noticeable in figure 17, is a plateau in performance after 300 epochs and substantially decreasing afterwards. This suggests that, in some cases, it is pointless to train the model for more than 300 epochs as it starts to slightly overfit, as it can be seen by the increase in training/validation accuracy curves. Afterwards, validation loss starts to increase, and with it, also the training accuracy while the validation accuracy decreases. The validation loss stays mostly flat during a long period of training due to weight regularisation

factors (L2 regularisation), it is noticeable that the model can train for more extended periods without severally overfit when such methods are applied.

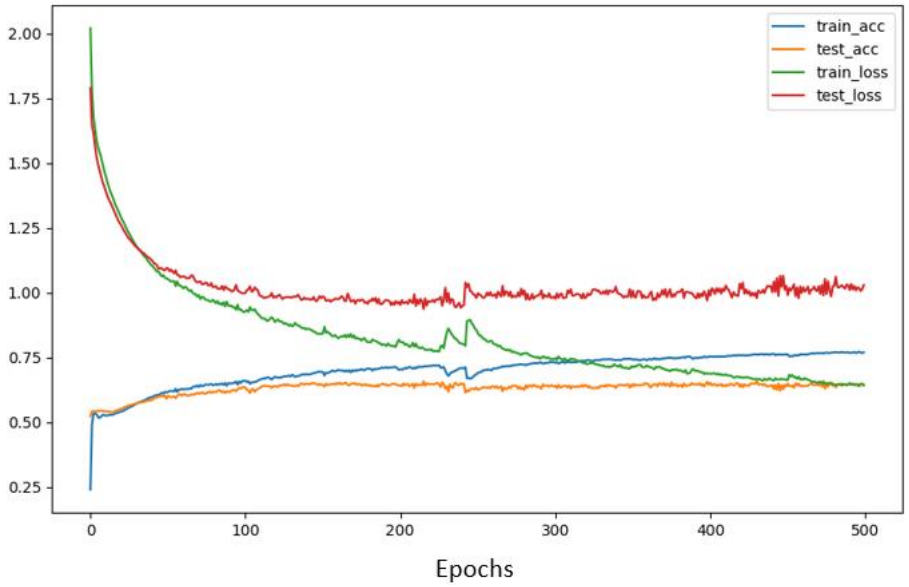


Figure 17. Framework 1 training performance.

Thus, the results obtained with this framework related to the subject pool presented previously are the following:

Table 4. Respiratory effort results obtained using the first model for four stages classification. Significantly different than OSA and All for Healthy ([a] $p < .05$, [b] $p < .01$ and [c] $p < .001$) after a Wilcoxon rank-sum test.

4 Stages	***Method A***			***Method B***		
Groups	**All**	**Healthy**	**OSA**	**All**	**Healthy**	**OSA**
Accuracy	0.64±0.11	0.65±0.11	0.65±0.10	0.65±0.11	0.66±0.10	0.62±0.12
Kappa	0.39±0.18[a]	0.41±0.18	0.32±0.18[c]	0.39±0.17[b]	0.43±0.16	0.29±0.17[c]

Table 5. Respiratory effort results obtained using the first model for three stages classification. Significantly different than All Method B for All Method A ([a] $p < .05$) after a Wilcoxon signed-rank test. Significantly different than OSA and All for Healthy ([b] $p < .01$, and [c] $p < .001$) after a Wilcoxon rank-sum test.

3 Stages	***Method A***			***Method B***		
Groups	**All**	**Healthy**	**OSA**	**All**	**Healthy**	**OSA**
Accuracy	0.75±0.11	0.76±0.10	0.71±0.11	0.74±0.10	0.76±0.10	0.71±0.07
Kappa	0.45±0.20[a, b]	0.49±0.19	0.33±0.18[c]	0.44±0.20[c]	0.49±0.19	0.31±0.18[c]

Table 4, presented above, express the results across the three subjects pool. As it was mentioned previously, in order to evaluate the performance of the framework in three-class task (WRN), classes Deep sleep (N3) and Light sleep (N1/N2) were merged in a single non-REM class (N). Analysing the performance of the framework in table four and five, we see that the classification performance rises substantially, to a Kappa of 0.45 and an accuracy of 0.74%, when compared to four classes. It is clear that, as expected, results are far better in three stages classification than in four. As for the subject pool, the best results belong to the Healthy subjects and the worse to OSA subjects, either for four stages or three stages classification. The overall performance is given by the All subject pool, being the maximum 0.40 Kappa. As it can also be seen, there is no significant difference between data treatment types, between Method A and B.

As for the statistical analysis of the overall performance, significant differences were found between subjects types. Inside the same data type, e.g. OSA from Method A versus Healthy from Method A, all p values were lower than 0.05, especially between Method B data in which, either for three stages and four stages classification, all values of p were lower than 0.01. This portraits that the framework can make a clear distinction between subject types as it exists a significant difference between each. Between data types, e.g. between Method A and Method B data, for three stages there is a tenuous statistical difference between All Method A subjects and All Method B subjects were $p < 0.05$. As for Healthy and OSA subjects, the p values are 0.478 and 0.097, respectively. As for four stages classification, the case is similar to three stages except that no significant statistical difference is found, and p values are 0.406 for All subjects, 0.065 for healthy subjects and 0.083 for OSA subjects.

4.2 Framework 2 – Results & Performance of Respiratory Data

This framework, when compared to framework 1, takes a larger number of epochs to achieve the maximum average validation performance, taking even longer to start overfitting. As expected, the overall behaviour of this framework is similar to the previous framework presented. The maximum average validation performance is also obtained between ~200-300 epochs, with an average Kappa of 0.40 for four stages classification and a Kappa of 0.46 for three stages classification. Similar to framework 1, there is a plateau in performance after ~350 epochs and substantially decreasing afterwards.

In the following tables, the framework 2 results are presented:

Table 6. Respiratory effort results obtained using the second model for four stages classification. Significantly different than All Method B for All Method A ([a] $p < .05$) after a Wilcoxon signed-rank test. Significantly different than OSA and All for Healthy ([b] $p < .01$, and [c] $p < .001$) after a Wilcoxon rank-sum test.

4Stages	***Method A***			***Method B***		
Groups	**All**	**Healthy**	**OSA**	**All**	**Healthy**	**OSA**
Accuracy	0.66±0.11	0.67±0.10	0.64±0.10	0.65±0.11	0.66±0.11	0.64±0.09
Kappa	0.40±0.18[a,b]	0.44±0.17	0.31±0.17[c]	0.39±0.19[c]	0.43±0.18	0.28±0.15[c]

Table 7. Respiratory effort results obtained using the second model for three stages classification. Significantly different than OSA and All for Healthy ([a] $p < .01$, and [b] $p < .001$) after a Wilcoxon rank-sum test.

3Stages	***Method A***			***Method B***		
Groups	**All**	**Healthy**	**OSA**	**All**	**Healthy**	**OSA**
Accuracy	0.75±0.10	0.77±0.09	0.70±0.09	0.74±0.21	0.76±0.10	0.70±0.09
Kappa	0.46±0.20[a]	0.50±0.19	0.34±0.17[b]	0.45±0.21[b]	0.50±0.19	0.34±0.18[b]

After analysing the tables above, it can be said that there're some significant differences between the subject pool, similar to framework 1. Results are better in three stages classification. On the other hand, there is a more noticeable difference between data types classification where Method A data presents better results.

Regarding statistical analysis, the same methods used in framework 1 were carried out to evaluate the framework 2. Inside the same data type, e.g. All from Method A versus Healthy from Method A, all *p* values were lower than 0.01 especially between Method B data in which, either for three stages and four stages classification, all values of *p* were lower than 0.001 showing a more significant difference between subject types.

Between data types, e.g. Method B data versus Method A data, for four stages there is significant statistical difference between All Method B subjects and All Method A subjects were $p < 0.05$. As for Healthy and OSA subjects, the *p* values are 0.072 and 0.061, respectively. As for three stages classification, the case is similar to 4 stages expect that no significant statistical difference is found, and *p* values are 0.676 for All subjects, 0.712 for healthy subjects and 0.690 for OSA subjects.

4.3 Framework 3 – Results & Performance of Respiratory Data

As illustrated in Figure 18, the maximum average validation performance is obtained between ~150-250 epochs, with an average Kappa of 0.30 for four stages classification and a Kappa of 0.35 for three stages classification. Also noticeable in figure 18, is a "smoothing" on the validation accuracy and loss curve due to a substantial decrease to the learning rate, enabling the model to train for longer and achieving better results. Even with the implementation of an adaptable learning rate during the training phase, it can be seen a plateau in performance after ~175 epochs. This may suggest that the learning rate of the model is not optimal and, even after a substantial reduction to the same, this model never outperforms the previous.

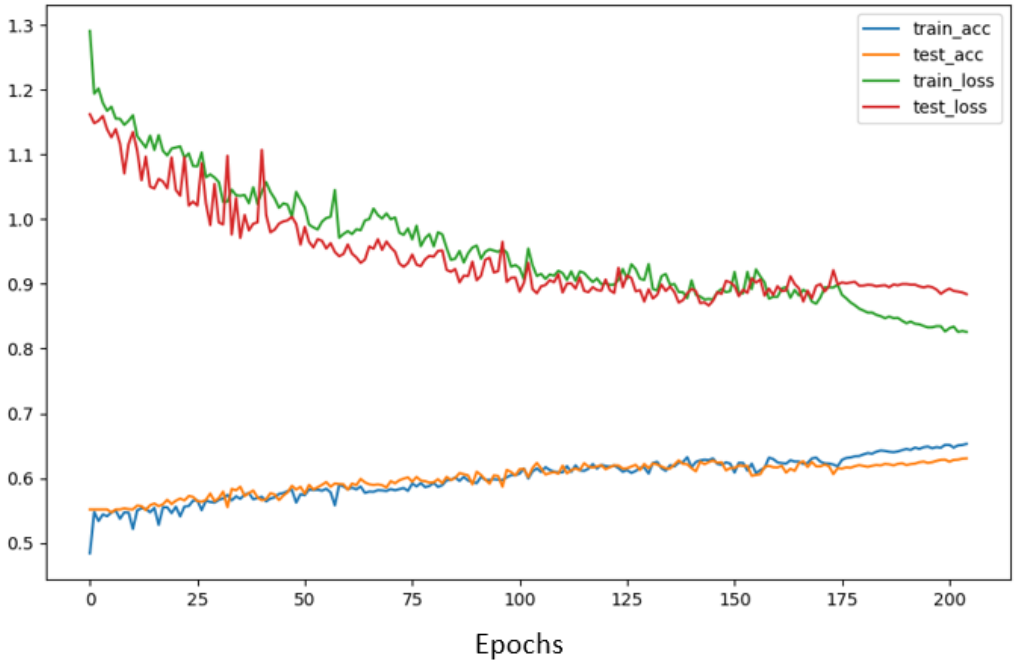


Figure 18. Framework 3 training performance.

Even with regularisation methods applied to almost every layer in the model, it is more difficult to train these type of networks, and the results fall behind both frameworks previously tested. The best result, similar to the previous designs, is achieved for three stages classification with an overall performance of 0.32 Kappa and 71% of accuracy, being Healthy the best subject pool with 0.35 Kappa and 72% of accuracy.

Table 8. Respiratory effort results obtained using the third model for four stages classification. Significantly different than OSA from Method A for OSA from Method B ([a] $p < .01$) after a Wilcoxon signed-rank test. Significantly different than OSA and All for Healthy ([b] $p < .05$, and [c] $p < .001$) after a Wilcoxon rank-sum test.

4 Stages	***Method A***			***Method B***		
Groups	**All**	**Healthy**	**OSA**	**All**	**Healthy**	**OSA**
Accuracy	0.62±0.10	0.62±0.10	0.62±0.10	0.62±0.10	0.62±0.09	0.63±0.10
Kappa	0.30±0.17[b]	0.33±0.17	0.20±0.12[c]	0.31±0.17[b]	0.33±0.16	0.23±0.14[a, c]

Table 9. Respiratory effort results obtained using the third model for three stages classification. Significantly different than All from Method B for All from Method A ([a] $p < .001$) and significantly different than Healthy from Method B for Healthy from Method A ([b] $p < .01$) after a Wilcoxon rank-sum test. Significantly different than OSA and All for Healthy ([c] $p < .05$, and [d] $p < .001$) after a Wilcoxon rank-sum test.

3 Stages	***Method A***			***Method B***		
Groups	**All**	**Healthy**	**OSA**	**All**	**Healthy**	**OSA**
Accuracy	0.71±0.09	0.72±0.09	0.69±0.09	0.71±0.09	0.71±0.09	0.69±0.08
Kappa	0.32±0.19[a,c]	0.35±0.19[b]	0.21±0.14[d]	0.30±0.18[c]	0.33±0.17	0.19±0.13[d]

Analyzing the tables above, there are some significant differences between the subject pool, for OSA subject pool even more noticeable. As it so happened in previous frameworks, results are better in three stages classification, but not that much better as it happened in other models here tested. There is a clear distinction, even more noticeable than the other frameworks, between OSA subject pool performance and Healthy, or All, for three stages and four stages classification. In both cases, the results round the 0.20 Kappa and the best result for this group of subjects falls on four stages classification, rather than three stages as it so happened for other subject pool or frameworks here tested.

Regarding statistical analysis, the same methods used in framework 1 and 2 were carried out to evaluate framework 3. Between subject types, inside same data type, e.g. OSA from Method A versus All from Method A, all p values were lower than 0.05, and even lower when OSA versus the other two types ($p < .001$). Method A data showed lower values of p for four stages classification and higher for three stages. Method B data, on the other hand, showed lower values of p for stages classification and higher for four stages.

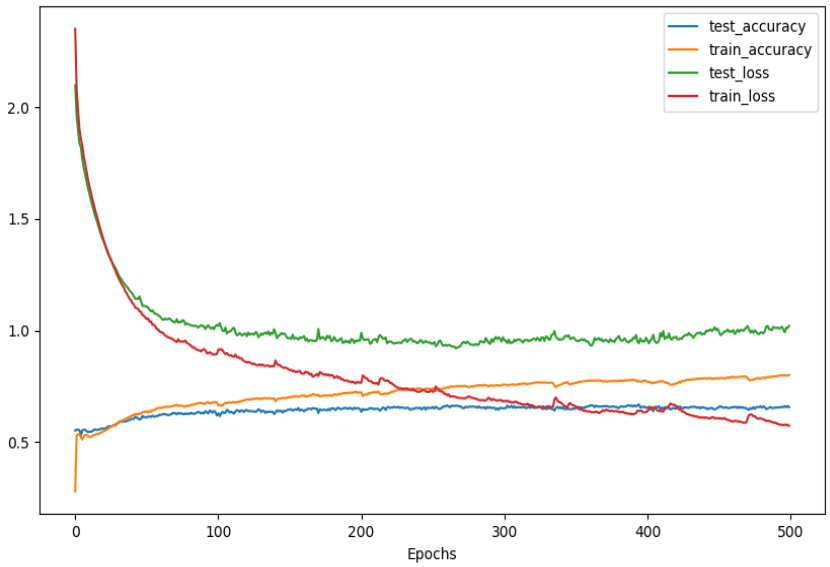
Between data types, e.g. Method A data versus Method B data, for four stages there is significant statistical difference between OSA Method B and OSA Method A where $p < 0.01$. As for Healthy and All subjects, the p values are 0.533 and 0.054, respectively. For three stages classification, the case

is the opposite to 4 stages, where between All from Method B and All from Method A, as well as between Healthy from Method B and Healthy from Method A data, there is significant statistical difference where p values are lower than 0.001 and 0.01, respectively.

4.4 Cardiorespiratory Data Performance

To train and evaluate the performance of cardiorespiratory data, framework 2 and 3 were used. Framework 2 was used because it is capable of combining two different signals and apply them different settings. Framework 3 was also used in order to find out whether these type of networks, CNNs combined with LSTMs, would perform better with more features, cardiac features combined with respiratory features, as input instead of spectrograms with just respiratory effort information.

a.



b.

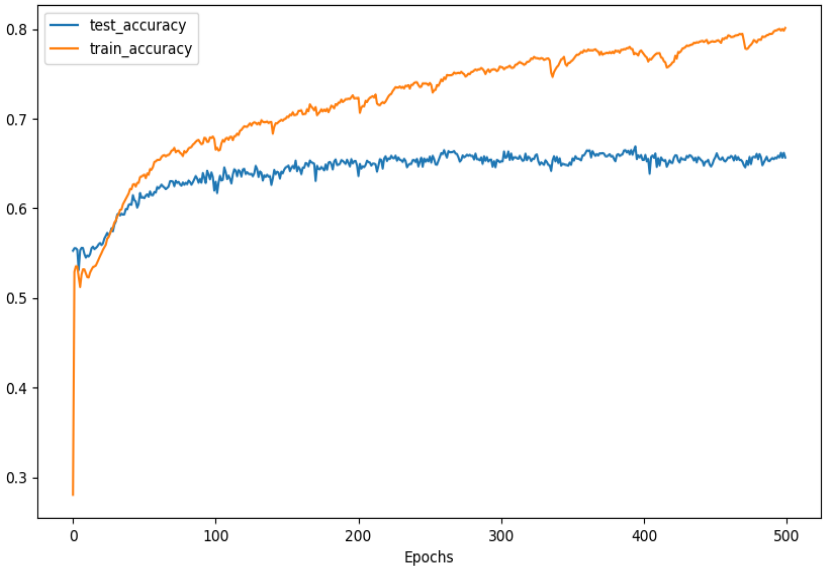


Figure 19. Overall performance of the framework on cardiorespiratory data. (a) Overall Performance of framework 2 with cardiorespiratory data (1 Fold); (b) Performance of the same framework for training and validation accuracy.

As it can be seen in Figure 19, the performance with framework 2 is similar to the one obtained with respiratory data alone. The maximum average validation performance is also obtained between ~200-300 epochs, with an average Kappa of 0.41 for four stages classification and a Kappa of 0.46 for three stages classification. Similar to framework 2, there is a plateau in performance after ~350 epochs and substantially decreasing afterwards.

The results presented next in table ten and eleven were obtained using the second framework.

Table 10. Cardiorespiratory data performance obtained using the second model for four stages classification. Significantly different than OSA and All for Healthy ([a] $p < .01$, and [b] $p < .001$) after a Wilcoxon rank-sum test.

4 Stages	***Cardiorespiratory (Framework 2)***		
Groups	**All**	**Healthy**	**OSA**
Accuracy	0.66±0.11	0.67±0.11	0.65±0.09
Kappa	0.40±0.18[a]	0.44±0.17	0.30±0.15[b]

Table 11. Cardiorespiratory data performance obtained using the second model for three stages classification. Significantly different than OSA and All for Healthy ([a] $p < .001$) after a Wilcoxon rank-sum test.

3 Stages	***Cardiorespiratory (Framework 2)***		
Groups	**All**	**Healthy**	**OSA**
Accuracy	0.75±0.11	0.77±0.10	0.69±0.08
Kappa	0.46±0.20[a]	0.51±0.18	0.32±0.16[a]

Regarding statistical analysis, the same methods used in all previous frameworks were carried out to evaluate cardiorespiratory performance. Between subject types, inside same data type, e.g. OSA Method A versus All Method A, all p values were lower than 0.01, suggesting a good generalisation regarding performance as it exists between subject types significant statistical difference.

It was also assessed cardiorespiratory performance with framework 3. The results are relatively worse than those obtained with framework 2 and follow the same pattern as respiratory did with this framework. With an overall κ of 0.25 for four stages classification, it is possible to state that the results fall greatly behind of those achieved with framework 2, as it also happened with respiratory data. For this reason, results obtained with this framework for cardiorespiratory data will not be subject of statistical analysis as they greatly lack behind the best obtained for this task.

4.5 FRAMEWORK 1 VERSUS 2

To access this performance comparison, much like previously done in different sub-chapters, Cohen's kappa and Wilcoxon signed-rank test will be used to statically compare two frameworks. The best overall performance, performance on All subject pool, of each framework was selected.

Focusing on the first two models, they are compared by accessing the overall performance of each existent subset: All Framework 1 versus All Framework 2, Healthy Framework 1 versus Healthy Framework 2, etc.

The results are the following:

Table 12. Wilcoxon signed-rank test p-values on best Performance framework 1 vs framework 2 on respiratory data.

	Framework 1 versus Framework 2	
p-values	**3Stages**	**4Stages**
All	0.502	<0.05
Healthy	0.513	<0.05
OSA	0.499	0.199

Although regarding Kappa and accuracy the differences are small, framework 2 had the best results and, as it is shown in Table 12, the differences are statistically significant.

4.6 FRAMEWORK 1 & 2 VERSUS 3

An overall performance assessment between framework 1 and 3 and between framework 2 and 3 is presented in the next table.

As it can be seen, substantial significant differences are found in both cases, proving that the first two frameworks severely outperform framework 3.

Table 13. Framework 1 versus Framework 3.

	Framework 1 versus Framework 3	
p-values	**3Stages**	**4Stages**
All	<0.001	<0.001
Healthy	<0.001	<0.001
OSA	<0.001	<0.001

Table 14. Framework 2 versus Framework 3.

	Framework 2 versus Framework 3	
p-values	**3Stages**	**4Stages**
All	<0.001	<0.001
Healthy	<0.001	<0.001
OSA	<0.001	<0.01

4.7 CARDIORESPIRATORY VERSUS RESPIRATORY

In this subsection, much like the two previous, two methods are compared. In this case, cardiorespiratory sleep staging versus Respiratory sleep staging. Using the best performance in both cases, the results were then evaluated by the same method as the previous comparisons.
Presented next is the result of the statistical comparison:

Table 15. Cardiorespiratory versus Respiratory with Framework 2.

	Cardioresp versus Resp	
p-values	**3Stages**	**4Stages**
All	0.078	0.842
Healthy	0.052	0.915
OSA	0.170	0.405

The results, as it can be seen on table 20, show that there is no significant statistical difference as all p-values are higher than 0.05.

5 Discussion

5.1 FRAMEWORK 1

This framework has some limitations, but it was obtained with it a reasonably good overall performance. Even though it only takes one spectrum as an input, narrowing down the outcome predicted with just a couple of frequencies, this framework proves that the Bi-LSTMs are powerful enough to retrieve relevant information towards sleep stages.

When looking to the overall statistical analysis of the framework, presented in "Framework 1 performance" subchapter, it is possible to say that the model did not generalised well, as significant differences were found between subject types. Nevertheless, this model still poses as a potential candidate for future studies of the same type. Looking into more detail in this statistical analysis, framework 1 did not significantly present a noticeable change between Method A data and Method B data except for All subject pool from Method A were it presented significant differences when compared with its homonymous from Method B. Two limitations may cause this not significant difference:

- The processing Method B, when compared to A, may be removing important feature information and thus the not so significant difference between Methods;
- The framework may not be capturing enough information about all sleep stages, or the number of features is not enough so that the model can generalise;

To sustain some of the claims previously presented about this framework, a comparison with state-of-the-art methods with similar frameworks is mandatory in order to understand what failed and what can be perfected. The standard in respiratory effort based sleep staging, the work of Long in "Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging" presents a Cohen's kappa of 0.41 for a pool of healthy subjects [10]. Even though Long used traditional machine learning classifiers, the results obtained are very similar to the ones obtained with this deep learning framework. This may suggest that either the pre-processing methods, or features used, are not enough or were incorrectly performed, or that the framework used may not perform as well as it was expected. A similar framework to framework 1 is presented in the work of Radha Et al. 2018 [65] "LSTM knowledge transfer for HRV-based sleep staging" where a three stacked Bi-LSTM is used to obtain current state-of-the-art results in unobtrusive sleep stage classification for four stages classification with a presented Cohen's kappa of 0.67. It is difficult to compare the results here presented with Mustafa's work as sleep stage classification is done with different data, as in this framework 1 HRV features derived from ECG are not used. This may lead to

the conclusion fact that framework 1 may be well suited to unobtrusive sleep stage, including with just respiratory effort data. It can also be inferred, even though it is not a linear concept, that with a more extensive feature selection procedure, or with different pre-processing methods, done on respiratory effort data, results would likely improve.

5.2 Framework 2

It was expected a slight improvement with this framework when compared to the previous on respiratory data and cardiorespiratory data. It proved that it could improve sleep stages classification by applying more pre-processing methods to respiratory signals, obtaining good results with all subjects, especially OSA and Healthy subjects where the literature is more focused.

The overall performance of this framework with respiratory data is significantly superior to the previous and the other two frameworks hereby tested. Introducing more data related to deep sleep (N3), even though only accounting for 13% average of the night, it was expected a modest increase in performance, as it was seen in the results presented. With an overall performance of 0.40 Kappa, 0.44 for healthy subjects and 0.31 for OSA subjects, this is the best result presented in this thesis as it performs slightly better than the previous framework, which subsequently, already reached state-of-the-art results and in some cases, performed slightly better. These results also portrait that this framework cannot generalise very well as significant differences are found between subject types, similar to framework 1.

As for cardiorespiratory data, the values were almost identical to the ones obtained with respiratory data, and no significant statistical difference was found between them. Better results were expected with the latter than with respiratory as different data was being combined giving the model the extent possibility of better generalisation. Even though more features were added in this framework, this did not directly translate in an increase of performance. Cardiac data introduced was not pre-processed as extensively as respiratory effort data and relevant features may not have been extracted as in similar works [6], this probably lead to an increase of noise and irrelevant data resulting in a framework that underperformed probably due to the quality of the data than rather the architecture of the framework.

Even though it is not possible to directly compare this network with any from current state-of-the-art content, it is possible to conclude for respiratory effort data that this framework, combined with more pre-processing, significantly improved the results.

5.3 Framework 3

The primary purpose of using this framework was that with CNNs, the feature extraction done by the framework would significantly improve contributing to a better generalisation. Therefore, it was expected overall improvement in performance when compared to previous LSTM-based frameworks on respiratory effort data and cardiorespiratory data.

This framework captures more information as it combines convolutional neural nets with recurrent ones. Feature maps retrieved from the CNN should have contained more information in order to help to distinguish between sleep stages. The results prove otherwise. This method may be failing due to excess of redundant information or that does not capture relevant features from unobtrusive data. When creating the spectrogram, it was used a 3-second FFT window in a 30-second epoch. In sleep, information has very few fast variations that can be captured in such small windows, leading to the increase of redundant information, and possibly irrelevant noisy data, as most 3-second windows will contain the same information within 30 seconds. Another problem may be associated, just like in previous frameworks, related to the quality of features of both respiratory effort and cardiorespiratory data, where pre-processing methods were not extensively carried out and not so relevant features were used.

Concluding, no relevant positive conclusion was drawn from the implementation of this framework. It is only possible to infer that this type of data is not well suited, or not well processed, to serve as input to this framework, as similar frameworks have proved to work in sleep stage classification problems [54]–[56].

5.4 Overall Performance of Frameworks and Pre-Processing Methods

Framework 2 is the framework that presented the best overall performance when compared to the other two presented in this work. The ground-truth for this comparison is how each of these frameworks performed on respiratory data, the common denominator.

Framework 1 performed almost as well as framework 2 with just some minor differences regarding four stages classification where framework 2 significantly performed better. Framework 3

performance is by far the worst of the three frameworks, where no relevant improvements were registered.

As for pre-processing methods, the common denominator for comparison is respiratory data overall performance. It is not very clear which pre-processing method, A or B stands out more as both perform equally well. Even though, the best result with respiratory effort for four stages is obtained with Method A, All subjects, and it is significantly different from Method B, it is not possible to infer that Method A is better than Method B as Method B also performed as well or, in some cases, even better than Method A.

5.5 Cardiorespiratory versus Respiratory Data

Cardiorespiratory data has been widely used and extremely positive results are being accomplished with it [6], [66]. In this work, it was expected an increase in performance when using this type of data, as previously explained in framework 2 and 3 performance. What was concluded is that no significant difference in performance was obtained by combining respiratory with cardiac data. As it was also stated, this is probably caused by the lack of pre-processing methods applied to cardiac data as relevant features may not have been extracted, and when combined with respiratory effort data, the effect was the opposite of the expected.

Nevertheless, it is known that cardiorespiratory significantly improved sleep stage classification problems as it is seen, for example, in Long 2014b where they used respiratory features to sleep stage with four stages, obtaining a $\kappa = 0.41$ and, in Fonseca 2015, the same authors, combined respiratory with cardiac features and obtained a κ of 0.49 for four stages classification, increasing the overall classification by 0.07 Kappa.

5.6 Comparison with state-of-the-art

In literature, only a few studies focused on using just respiratory effort signals to perform sleep stage classification with four stages and the results presented here in this work are amongst the best.

Following next is a table that presents all state-of-the-art results for sleep stages classification using respiratory effort and cardiorespiratory signals.

Table 16. Performance comparison with state-of-the-art.

Reference	*Modalities*	*N*	*Age*	*Average κ*	*Average Accuracy*
WRLD					
This work	RIP	197(Healthy)	52 ± 17	0.44	0.67
This work	RIP	55 (OSA)	52 ± 17	0.31	0.64
This work	RIP	294 (All)	52 ± 17	0.40	0.66
This work	RIP, ECG	197(Healthy)	52 ± 17	0.44	0.67
This work	RIP, ECG	55 (OSA)	52 ± 17	0.30	0.65
This work	RIP, ECG	294 (All)	52 ± 17	0.40	0.66
Fonseca 2015	RIP, ECG	48	41.3±16.1	0.49	0.69
Long 2014[2]	RIP	48	41.3±16.1	0.41	0.65
Willemen et al. 2014	RIP, ECG, ACT	85 (36 subj.)	22.1 ± 3.2	0.56	0.69
WRN					
This work	RIP	197(Healthy)	52 ± 17	0.50	0.77
This work	RIP	55 (OSA)	52 ± 17	0.34	0.70
This work	RIP	294 (All)	52 ± 17	0.46	0.75
This work	RIP, ECG	197(Healthy)	52 ± 17	0.51	0.77
This work	RIP, ECG	55 (OSA)	52 ± 17	0.32	0.69
This work	RIP, ECG	294 (All)	52 ± 17	0.46	0.75
Fonseca 2015	RIP, ECG	48	41.3±16.1	0.56	0.80
Long 2014[2]	RIP	48	41.3±16.1	0.48	0.77
Redmond et al. 2006	RIP, ECG	37 (OSA)	46.7±10.4	0.32	0.67
Redmond et al. 2007	RIP, ECG	31	42.0 ± 7.4	0.45	0.76

As it can be seen, this work presents some results that reach, and in some cases surpass, the presented state-of-the-art results that used just respiratory data in four and three stages classification.

[2] X. Long *et al.*, "Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging," *Physiol. Meas.*, vol. 35, no. 12, pp. 2529–2542, Dec. 2014.

It is to be noted that, this work used a pool of subjects that contained healthy and disordered patients (All) for training the model, as none of the previous did, and without manually engineered features, as also none of the previous did. Long Et al. [10] presented a work with manually engineered features with four and three stages classification, being this the best result with just respiratory signals so far (κ = 0.41 and accuracy = 0.65). This work results, (κ=0.44 and accuracy = 0.67) slightly surpass the previous. Redmond Et al. [13] presented in its work, a three stages classification problem with OSA subjects using cardiorespiratory data, results of κ = 0.32 and accuracy of 0.67. This work also outperforms the previous with κ = 0.34 and accuracy of 0.69 for three stages classification using just respiratory data and, for four stages classification, κ = 0.31 and accuracy of 0.64, almost reaching the same values for three stages classification. Regarding cardiorespiratory sleep staging, this work results fall behind on most recent results. It is also relevant to mention that the average age of the individuals present in this dataset is superior when compared with datasets used in state-of-the-art (Table 16, “Age”). Even though, it was not tested if this was a negative contributing factor to the results.

6 CONCLUSIONS

This work addressed the performance of cardiorespiratory data for sleep stages classification using four and three stages. It was shown that without manually engineered features and a dataset containing healthy and disordered subjects, it is possible to develop a sleep stage classifier with a moderate performance using either respiratory effort signals, or cardiorespiratory signals. It was also proved that it is possible to build a framework capable of generalizing well when classifying different types of subjects, as it was presented in this work with OSA, healthy and the overall pool of subjects, Raw, containing healthy and disordered patients.

Hence, in this work it is proposed a method that can sleep stage classify based on a dataset containing healthy and disordered subjects using either respiratory effort data, or cardiorespiratory data without manually engineered features. This process comprised two steps: first, data processing; second, build the framework that can be used in both data;

The new classification framework was evaluated in a k-cross-validation method (k=4), on a dataset containing 294 subjects, totalling in 588 PSG recordings of which, 51 of OSA subjects and 197 of healthy subjects. For respiratory data, an overall improvement was achieved with this method when compared with state-of-the-art results in three stages (Wake, REM and non-REM) and four stages classification (Wake, REM, N1/N2 and N3). It was obtained a Cohen's Kappa (κ) of 0.40 for the subject pool containing all subjects, 0.44 for healthy subjects and 0.31 for OSA subjects with four stages classification. For three stages classification, the performance obtained was a κ of 0.46 for the overall pool of subjects (Raw), 0.50 for healthy subjects and 0.34 for OSA subjects. For cardiorespiratory data, no improvement was made to the state-of-the-art results. For four stages classification, it was obtained a κ of 0.40 for the overall subject pool (Raw), 0.44 for healthy subjects and 0.30 for OSA subjects. With three stages, a κ of 0.46 for Raw subjects, 0.51 for healthy and 0.32 for OSA subjects.

In summary, the proposed framework outperforms the current state-of-the-art results when using only respiratory effort and introduces new ones for overall classification with a pool of subjects containing healthy and disordered.

6.1 Future Work

As future work, it would be advantageous to implement more features, manually engineered ones, in respiratory effort and cardiac data. It may be logic that, if with a simple system it was possible to outperform current state-of-the-art values, with more features the same method would probably

achieve even better results. Another method would be combining actigraphy and cardiorespiratory data, as it was done previously in Domingues Et al. 2014 and Willemen Et al. 2014, as this may prove to be beneficial as presented in their work.

Another method, this one related to deep learning instead of data related, would be using adversarial training, as it is the new way forward with neural networks. Due to the lack of time, this part was placed on hold and no further investigation was conducted.

REFERENCES

[1] M. H. Kryger, T. (Tom) Roth, and W. C. Dement, *Principles and practice of sleep medicine*. Saunders/Elsevier, 2011.

[2] X. Long, "On the analysis and classification of sleep stages from cardiorespiratory activity."

[3] M. P. Walker, T. Brakefield, A. Morgan, J. A. Hobson, and R. Stickgold, "Practice with Sleep Makes Perfect: Sleep-Dependent Motor Skill Learning," *Neuron*, vol. 35, no. 1, pp. 205–211, Jul. 2002.

[4] A. Wade, N. Zisapel, and P. Lemoine, "Prolonged-release melatonin for the treatment of insomnia: targeting quality of sleep and morning alertness," *Aging health*, vol. 4, no. 1, pp. 11–21, Feb. 2008.

[5] "2018 WORLD SLEEP DAY® PRESS RELEASE |." [Online]. Available: http://worldsleepday.org/2018-press-release. [Accessed: 28-Nov-2018].

[6] P. Fonseca, X. Long, M. Radha, R. Haakma, R. M. Aarts, and J. Rolink, "Sleep stage classification with ECG and respiratory effort," *Physiol. Meas.*, vol. 36, no. 10, pp. 2027–2040, Oct. 2015.

[7] "The Polysomnogram Test for Sleep Apnea." [Online]. Available: https://www.sleep-apnea-guide.com/polysomnogram.html. [Accessed: 28-Nov-2018].

[8] J. Paalasmaa, M. Waris, H. Toivonen, L. Leppakorpi, and M. Partinen, "Unobtrusive online monitoring of sleep at home," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 3784–3788.

[9] X. Long, J. Foussier, P. Fonseca, R. Haakma, and R. M. Aarts, "Analyzing respiratory effort amplitude for automated sleep stage classification," *Biomed. Signal Process. Control*, vol. 14, pp. 197–205, Nov. 2014.

[10] X. Long *et al.*, "Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging," *Physiol. Meas.*, vol. 35, no. 12, pp. 2529–2542, Dec. 2014.

[11] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

[12] N. J. Douglas, D. P. White, C. K. Pickett, J. V Weil, and C. W. Zwillich, "Respiration during sleep in normal man.," *Thorax*, vol. 37, no. 11, pp. 840–4, Nov. 1982.

[13] S. J. Redmond and C. Heneghan, "Cardiorespiratory-Based Sleep Staging in Subjects With Obstructive Sleep Apnea," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 485–496, Mar. 2006.

[14] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[15] T.-J. Hsieh, H.-F. Hsiao, and W.-C. Yeh, “Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm,” *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2510–2525, Mar. 2011.

[16] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, “DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving.” pp. 2722–2730, 2015.

[17] X. Zhang, W. Kou, E. I.-C. Chang, H. Gao, Y. Fan, and Y. Xu, “Sleep stage classification based on multi-level feature learning and recurrent neural networks via wearable device,” *Comput. Biol. Med.*, vol. 103, pp. 71–81, Dec. 2018.

[18] M. A. Carskadon and W. C. Dement, “Chapter 2-Normal Human Sleep : An Overview.”

[19] A. Kales, *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Washington DC: United States Government Printing Office, 1968.

[20] H. J. Burgess, J. Trinder, Y. Kim, and D. Luke, “Sleep and circadian influences on cardiac autonomic nervous system activity.,” *Am. J. Physiol.*, vol. 273, no. 4 Pt 2, pp. H1761-8, Oct. 1997.

[21] D. R. Smith and T. Lee-Chiong, “Respiratory Physiology During Sleep,” *Sleep Med. Clin.*, vol. 3, no. 4, pp. 497–503, Dec. 2008.

[22] J. W. Kantelhardt, T. Penzel, S. Rostig, H. F. Becker, S. Havlin, and A. Bunde, “Breathing during REM and non-REM sleep: correlated versus uncorrelated behaviour,” *Phys. A Stat. Mech. its Appl.*, vol. 319, pp. 447–457, Mar. 2003.

[23] B. V. Vaughn, S. R. Quint, J. A. Messenheimer, and K. R. Robertson, “Heart period variability in sleep,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 94, no. 3, pp. 155–162, Mar. 1995.

[24] Z. Shinar, S. Akselrod, Y. Dagan, and A. Baharav, “Autonomic changes during wake–sleep transition: A heart rate variability based approach,” *Auton. Neurosci.*, vol. 130, no. 1–2, pp. 17–27, Dec. 2006.

[25] T. Penzel, J. W. Kantelhardt, L. Grote, J. Peter, and A. Bunde, “Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea,” *IEEE Trans. Biomed. Eng.*, vol. 50, no. 10, pp. 1143–1151, Oct. 2003.

[26] T. I. Morgenthaler, V. Kagramanov, V. Hanak, and P. A. Decker, “Complex Sleep Apnea Syndrome: Is It a Unique Clinical Syndrome?,” *Sleep*, vol. 29, no. 9, pp. 1203–1209, Sep.

2006.

[27] "THE INTERNATIONAL CLASSIFICATION OF SLEEP DISORDERS, REVISED Diagnostic and Coding Manual," 1990.

[28] L. Spicuzza, D. Caruso, and G. Di Maria, "Obstructive sleep apnoea syndrome and its management," *Ther. Adv. Chronic Dis.*, vol. 6, no. 5, pp. 273–285, Sep. 2015.

[29] D. 0 Hebb, "The Organization of Behavior A NEUROPSYCHOLOGICAL THEORY."

[30] P. Domingos, *The master algorithm : how the quest for the ultimate learning machine will remake our world*. .

[31] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[32] F. Rosenblatt, *The perceptron, a perceiving and recognizing automaton : (Project Para)*. Buffalo, NY: Cornell Aeronautical Laboratory, 1957.

[33] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning."

[34] Y. LeCun and Y. Bengio, "Convolutional Networks for Images, Speech, and Time-Series Parsing View project Oracle Performance for Visual Captioning View project," 1997.

[35] D. Hubel and T. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, 1962.

[36] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," Jul. 2012.

[37] K. Fukushima and S. Miyake, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition," Springer, Berlin, Heidelberg, 1982, pp. 267–285.

[38] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," 1997.

[39] C. Olah, "Understanding LSTM Networks – colah's blog." [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed: 28-Nov-2018].

[40] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[41] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies."

[42] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[43] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning Precise Timing with LSTM Recurrent Networks," *J. Mach. Learn. Res.*, vol. 3, no. Aug, pp. 115–143, 2002.

[44] A. Graves, S. Fernández, and J. Schmidhuber, “Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition,” pp. 799–804, 2005.

[45] G. Klosh *et al.*, “The SIESTA project polygraphic and clinical database,” *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 51–57, 2001.

[46] P. Fonseca, R. M. Aarts, X. Long, J. Rolink, and S. Leonhardt, “Estimating actigraphy from motion artifacts in ECG and respiratory effort signals,” *Physiol. Meas.*, vol. 37, no. 1, pp. 67–82, Jan. 2016.

[47] A. K. Jain, “Advances in Statistical Pattern Recognition,” in *Pattern Recognition Theory and Applications*, Berlin, Heidelberg: Springer Berlin Heidelberg, 1987, pp. 1–19.

[48] J. Cohen, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.,” *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968.

[49] A. Viera and J. Garrett, “Understanding Interobserver Agreement: The Kappa Statistic,” 2005.

[50] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bull.*, vol. 1, no. 6, p. 80, Dec. 1945.

[51] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, “On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima,” Sep. 2016.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” 2014.

[53] I. N. Yulita, M. I. Fanany, and A. M. Arymuthy, “Bi-directional Long Short-Term Memory using Quantized data of Deep Belief Networks for Sleep Stage Classification,” *Procedia Comput. Sci.*, vol. 116, pp. 530–538, Jan. 2017.

[54] H. Dong, A. Supratak, W. Pan, C. Wu, P. M. Matthews, and Y. Guo, “Mixed Neural Network Approach for Temporal Sleep Stage Classification,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 2, pp. 324–333, Feb. 2018.

[55] W. Chen *et al.*, “Multimodal Ambulatory Sleep Detection.,” ... *IEEE-EMBS Int. Conf. Biomed. Heal. Informatics. IEEE-EMBS Int. Conf. Biomed. Heal. Informatics*, vol. 2017, pp. 465–468, Feb. 2017.

[56] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, “A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 4, pp. 758–769, Apr. 2018.

[57] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” Feb. 2015.

[58] S. Biswal *et al.*, “SLEEPNET: Automated Sleep Staging System via Deep Learning,” Jul. 2017.

[59] X. Glorot, A. Bordes, and Y. Bengio, “Deep Sparse Rectifier Neural Networks,” 2011.

[60] J. Kahn, “Google’s DeepMind Achieves Speech-Generation Breakthrough - Bloomberg,” 2016. [Online]. Available: https://www.bloomberg.com/news/articles/2016-09-09/google-s-ai-brainiacs-achieve-speech-generation-breakthrough. [Accessed: 28-Nov-2018].

[61] A. Van Den Oord *et al.*, “WAVENET: A GENERATIVE MODEL FOR RAW AUDIO.”

[62] D. Coldewey, “Google’s WaveNet uses neural nets to generate eerily convincing speech and music | TechCrunch.” [Online]. Available: https://techcrunch.com/2016/09/09/googles-wavenet-uses-neural-nets-to-generate-eerily-convincing-speech-and-music/. [Accessed: 28-Nov-2018].

[63] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, “Pixel Recurrent Neural Networks,” Jan. 2016.

[64] BBC NEWS, “Adobe Voco ‘Photoshop-for-voice’ causes concern - BBC News.” [Online]. Available: https://www.bbc.com/news/technology-37899902. [Accessed: 28-Nov-2018].

[65] M. Radha, P. Fonseca, M. Ross, A. Cerny, P. Anderer, and R. M. Aarts, “LSTM knowledge transfer for HRV-based sleep staging,” Sep. 2018.

[66] T. Willemen *et al.*, “An Evaluation of Cardiorespiratory and Movement Features With Respect to Sleep-Stage Classification,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 2, pp. 661–669, Mar. 2014.

APPENDICES

A – RESULTS FRAMEWORK 1

A.1 METHOD A

Table 17. Classification obtained with Method A data for 4 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.39 ±0.18	0.64± 0.11	0.53± 0.15	0.60 ± 0.15	0.55 +0.14
Healthy	0.41 ±0.18	0.65± 0.11	0.55± 0.14	0.62 ± 0.14	0.57±0.14
OSA	0.32 ±0.18	0.65 ± 0.10	0.48± 0.14	0.55 ± 0.14	0.50±0.14

Table 18. Classification obtained with Method A data for 3 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.45 ± 0.20	0.75 ± 0.11	0.61± 0.15	0.68 ± 0.14	0.62 ±0.13
Healthy	0.49 ±0.19	0.76 ± 0.10	0.64± 0.14	0.70 ± 0.14	0.64 ±0.13
OSA	0.33 ±0.18	0.71 ± 0.09	0.54± 0.13	0.61 ± 0.14	0.54 ±0.13

A.2 METHOD B

Table 19. Classification obtained with Method B data for 4 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.39±0.17	0.65 ± 0.11	0.53± 0.14	0.61 ± 0.14	0.56 ±0.14
Healthy	0.43±0.16	0.66 ± 0.10	0.56± 0.13	0.63 ± 0.13	0.58 ±0.13
OSA	0.29 ±0.17	0.62 ± 0.12	0.46± 0.14	0.54 ± 0.15	0.49 ±0.15

Table 20. Classification obtained with Method B data for 3 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.44±0.20	0.74 ± 0.10	0.60± 0.14	0.67 ± 0.14	0.61 ±0.13
Healthy	0.49±0.19	0.76 ± 0.10	0.63± 0.14	0.70 ± 0.14	0.63 ±0.13
OSA	0.31±0.18	0.71 ± 0.07	0.52± 0.12	0.60 ± 0.12	0.53 ±0.12

B – RESULTS FRAMEWORK 2

B.1 METHOD A

Table 21. Classification obtained with Method A data for 4 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.40 ± 0.18	0.66 ± 0.11	0.54 ± 0.15	0.61 ± 0.14	0.56 ±0.14
Healthy	0.44 ± 0.17	0.67 ± 0.10	0.57 ± 0.14	0.63 ± 0.14	0.58 ±0.13
OSA	0.31 ± 0.17	0.64 ± 0.10	0.48 ± 0.14	0.54 ± 0.14	0.50 ±0.14

Table 22. Classification obtained with Method A data for 3 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.46 ± 0.20	0.75 ± 0.10	0.61 ± 0.14	0.68 ± 0.13	0.62 ±0.13
Healthy	0.50 ± 0.19	0.77 ± 0.09	0.64 ± 0.13	0.70 ± 0.12	0.65 ±0.13
OSA	0.34 ± 0.17	0.70 ± 0.09	0.54 ± 0.12	0.61 ± 0.12	0.56 ±0.12

B.2 METHOD B

Table 23. Classification obtained with Method B data for 4 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.39±0.19	0.65±0.11	0.53±0.15	0.61±0.15	0.54±0.14
Healthy	0.43 ± 0.18	0.66 ± 0.11	0.56 ± 0.14	0.63 ± 0.14	0.57 ±0.13
OSA	0.28 ± 0.15	0.64 ± 0.09	0.46 ± 0.13	0.54 ± 0.13	0.48 ±0.13

Table 24. Classification obtained with Method B data for 3 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.45±0.21	0.74±0.11	0.62±0.14	0.67±0.14	0.63±0.14
Healthy	0.50 ± 0.19	0.76 ± 0.10	0.65 ± 0.14	0.70 ± 0.13	0.65 ±0.13
OSA	0.34 ± 0.18	0.70 ±0.09	0.54 ± 0.13	0.60 ± 0.12	0.55 ±0.13

B.3 CARDIORESPIRATORY

Table 25. Classification obtained with Method B cardiorespiratory data for 4 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.40 ± 0.18	0.66 ± 0.11	0.54± 0.15	0.61 ± 0.15	0.56 ±0.14
Healthy	0.44 ± 0.17	0.67 ± 0.11	0.56± 0.14	0.63 ± 0.14	0.58 ±0.13
OSA	0.30 ± 0.15	0.65 ± 0.09	0.46± 0.12	0.54 ± 0.14	0.48 ±0.13

Table 26. Classification obtained with Method B cardiorespiratory data for 3 stages.

	Kappa	*Accuracy*	*F1-score*	*Precision*	*Recall*
Raw	0.46± 0.20	0.75 ± 0.11	0.62± 0.14	0.68 ± 0.15	0.63 ±0.13
Healthy	0.51 ± 0.18	0.77 ± 0.10	0.65± 0.14	0.71 ± 0.14	0.65 ±0.12
OSA	0.32 ± 0.16	0.69 ± 0.08	0.53± 0.12	0.58 ± 0.12	0.55 ±0.12

C – RESULTS FRAMEWORK 3

C.1 METHOD A

Table 27. Classification obtained with Method A data for 4 stages.

4stages	***Kappa***	***Accuracy***	***F1-score***	***Precision***	***Recall***
Raw	0.30±0.17	0.62±0.10	0.46±0.13	0.55±0.14	0.48±0.13
Healthy	0.33±0.17	0.62±0.10	0.48±0.14	0.57±0.14	0.50±0.12
OSA	0.20±0.12	0.62±0.10	0.38±0.10	0.48±0.12	0.41±0.11

Table 28. Classification obtained with Method A data for 3 stages.

3stages	***Kappa***	***Accuracy***	***F1-score***	***Precision***	***Recall***
Raw	0.32±0.19	0.71±0.09	0.52±0.13	0.62±0.14	0.52±0.11
Healthy	0.35±0.19	0.72±0.09	0.54±0.13	0.64±0.14	0.55±0.11
OSA	0.21±0.14	0.69±0.09	0.45±0.10	0.56±0.13	0.46±0.08

C.2 METHOD B

Table 29. Classification obtained with Method B data for 4 stages.

4stages	***Kappa***	***Accuracy***	***F1-score***	***Precision***	***Recall***
Raw	0.31±0.17	0.62±0.10	0.47±0.14	0.57±0.14	0.49±0.13
Healthy	0.33±0.16	0.62±0.09	0.49±0.13	0.58±0.13	0.51±0.12
OSA	0.23±0.14	0.63±0.10	0.42±0.12	0.52±0.13	0.43±0.12

Table 30. Classification obtained with Method B data for 3 stages.

3stages	***Kappa***	***Accuracy***	***F1-score***	***Precision***	***Recall***
Raw	0.30±0.18	0.71±0.09	0.50±0.12	0.60±0.14	0.51±0.11
Healthy	0.33±0.17	0.71±0.09	0.53±0.12	0.62±0.14	0.53±0.12
OSA	0.19±0.13	0.69±0.08	0.43±0.09	0.52±0.12	0.45±0.08

D – RESULTS FRAMEWORK 4

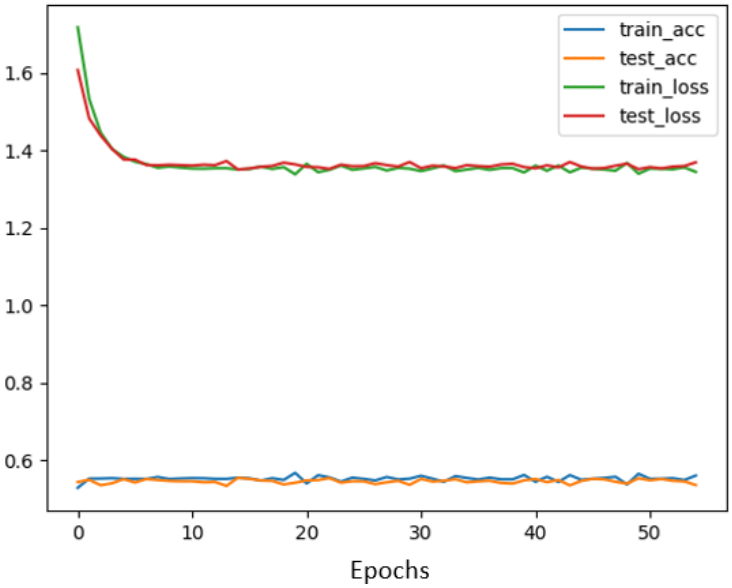


Figure 20. Framework 4 training performance.