

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Obstructive Sleep Apnea Detection using Fourth Level Devices

Sheikh Shanawaz Mostafa

Supervisor: Doctor Fernando Manuel Rosmaninho Morgado Ferrão Dias

Co-Supervisor: Doctor João Paulo Salgado Arriscado Costeira

Co-Supervisor: Doctor Antonio Gabriel Ravelo García

Thesis approved in public session to obtain the PhD Degree in

Electrical and Computer Engineering

Jury Final Classification: **Pass with Distinction**

2020

UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

Obstructive Sleep Apnea Detection using Fourth Level Devices

Sheikh Shanawaz Mostafa

Supervisor: Doctor Fernando Manuel Rosmaninho Morgado Ferrão Dias

Co-Supervisor: Doctor João Paulo Salgado Arriscado Costeira

Co-Supervisor: Doctor Antonio Gabriel Ravelo García

Thesis approved in public session to obtain the PhD Degree in

Electrical and Computer Engineering

Jury Final Classification: **Pass with Distinction**

Jury

Chairperson: Doutora Isabel Maria Martins Trancoso, Instituto Superior Técnico da Universidade de Lisboa.

Members of the Committee:

Doutor Carlos Manuel Travieso González, Escuela de Ingeniería de Telecomunicación y Electrónica, Universidad de las Palmas de Gran Canaria, Espanha.

Doutor Sergi Bermúdez I Badia, Faculdade de Ciências Exatas e da Engenharia da Universidade da Madeira.

Doutora Ana Luísa Nobre Fred, Instituto Superior Técnico da Universidade de Lisboa.

Doutor Alexandre José Malheiro Bernardino, Instituto Superior Técnico da Universidade de Lisboa.

Doutor Fernando Manuel Rosmaninho Morgado Ferrão Dias, Faculdade de Ciências Exatas e da Engenharia da Universidade da Madeira.

Funding Institutions

Cofinanciado por:



agência regional para o desenvolvimento da investigação tecnologia e inovação



2020

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the *Universidade de Lisboa*.

To...
My parents

Acknowledgement.

First, I would like to give a gracious thanks to my supervisor, Dr. Morgado Dias, for his continuous guidance during these years that led to the preparation of this work. I also thank my co-supervisors, Dr. Ravelo-García, and Dr. Joao Paulo Costeira, for their scientific insights, challenges, and suggestions. Thank you to all my colleagues as well specially Fábio Mendonça and Darío Baptista. I like to thank the reviewers for their selfless efforts to make this research better. This research has been made possible with funding and donation from the list below.

“NVIDIA Corporation” for their GPU donation.

Acknowledgments to the Portuguese Foundation for Science and Technology for their support through Projeto Estratégico UID/EEA/50009/2019.

Acknowledgments to the Portuguese Foundation for Science and Technology for their support through Projeto Estratégico LA 9 - UID/EEA/50009/2013.

Acknowledgement to ARDITI - Agência Regional para o Desenvolvimento da Investigação, Tecnologia e Inovação under the scope of the Project M1420-09-5369-FSE-000001 - PhD Studentship.

Acknowledgement to MITIExcell - EXCELENCIA INTERNACIONAL DE IDT&I NAS TIC (Project Number M1420-01-01450FEDER0000002), provided by the Regional Government of Madeira.

Abstract

Obstructive sleep apnea is a common sleep disorder characterized by an interruption of one's breathing during sleep. Due to the high cost, complexity, and accessibility issues related to polysomnography, which is the gold standard test for apnea detection, it is desirable that an automation of the diagnostic test be based on a simpler method. A fourth level device, which has only one or two channels allows the study to be performed outside the sleep laboratory with a low level of intrusion to the patient. The obstruction or reduction of airflow normally decreases the blood oxygen saturation level which can be used as a marker for apnea detection. An investigation of obstructive sleep apnea showed that apnea events have progressive bradycardia, followed by abrupt tachycardia on the resumption of breathing. Thus, heart rate variability could be a complement to the oxygen saturation signal. In this PhD work, two systematic reviews were performed to analyze the already existing algorithms to detect obstructive sleep apnea. One review focused on the use of different sensors while the other review focused on deep learning, in order to discover future trends and knowledge gaps in the field. Two different methods are followed in this work: handcrafted feature-based methods (using shallow networks) and automated feature-based methods (using deep networks). With regards to the handcrafted feature-based methods, it is necessary to find a subset of features that obtain the highest accuracy for the classifier. As for the automated feature-based methods, though the features are chosen automatically during the training, the structure of the deep neural network (namely, the choice of hyperparameters) plays a crucial part in the results. Therefore, for both methods, these criteria are investigated. In most of the cases, the proposed studies can achieve better results compared with those found in the literature. Using the classified apnea events, apnea patients (were classified global classification) which consequently also performed better than those in the literature. Among the different methods implemented in this work for handcrafted feature-based methods, a combination of certain classifiers performed better than single classifiers. In relation to automated feature-based Convolutional Neural Network (CNN) methods, the multi-objective method produced better results. To reduce the high simulation time, a greedy algorithm with a very similar performance was developed. The greedy algorithm for CNN optimization achieved better results in similar databases in the literature, though it was assumed that the use of the patients' heart rate could increase the accuracy of the system implemented with SpO₂. This did not occur in the actual investigation; however, the use of the patients' heart rate increased the sensitivity of some solutions. Both the handcrafted and the automated feature-based methods achieved better global accuracy than the results in the similar database results from the literature. In conclusion, it is possible to detect apnea events as well as apnea patients using fourth level devices with good accuracy. Furthermore, these solutions are relatively easy to integrate in different wearables.

Keywords

Deep neural network, Feature selection, Optimization, Shallow neural network and Sleep apnea.

Resumo

A apneia obstrutiva do sono é um distúrbio do sono frequente, caracterizado pela interrupção da respiração durante o sono. Devido ao elevado custo, complexidade e problemas de acessibilidade relacionados com a polissonografia, o teste padrão para detecção da apneia, a automação do teste de diagnóstico com base em métodos mais simples é desejada. Um dispositivo de quarto nível, que possui apenas um ou dois canais, permite que o estudo seja realizado fora do laboratório do sono com um baixo nível de intrusão para o paciente. A obstrução ou redução do fluxo de ar tipicamente reduz o nível de saturação de oxigênio no sangue, podendo esta ocorrência ser usada como marcador para a detecção da apneia. A investigação no âmbito da apneia obstrutiva do sono mostrou que os eventos da apneia apresentam bradicardia progressiva, seguida de taquicardia abrupta na retomada da respiração. Desta forma, a variabilidade da frequência cardíaca pode ser um complemento ao sinal da saturação de oxigênio. Neste trabalho de Doutorado, duas revisões sistemáticas foram realizadas de forma a analisar os algoritmos existentes para detetar a apneia obstrutiva do sono. Uma revisão analisou o uso de diferentes sensores e a outra focou-se na aprendizagem profunda, de forma a determinar as tendências futuras e as lacunas de conhecimento no estado da arte. Duas metodologias distintas foram examinadas neste trabalho: métodos baseados em características criadas pelo investigador (usando redes neuronais) e métodos baseados na extração automática de características (usando redes profundas). Para as abordagens baseadas em características criadas pelo investigador, é necessário encontrar um subconjunto de características que obtenha a maior precisão para o classificador. Para os métodos baseados na extração automática de características, estas são escolhidas durante o processo de treino, sendo que a estrutura da rede neuronal profunda e a escolha dos hiperparâmetros desempenham um papel crucial nos resultados. Desta forma, estes critérios foram analisados para ambos os métodos. Na maioria dos casos, os estudos propostos obtiveram melhores resultados que os apresentados na literatura. A classificação dos pacientes com apneia (classificação global) foi realizada através da análise dos eventos classificados como apneia, tendo também alcançado um desempenho melhor que o apresentado na literatura. Foi verificado que nos diferentes métodos implementados neste trabalho para a análise das características criadas pelo investigador, a combinação de classificadores teve um desempenho superior aos classificadores individuais. No que concerne aos métodos baseados na extração automática de características, baseados nas Convolutional Neural Networks (CNN), foi verificado que o método multi-objetivo alcançou os melhores resultados. Todavia, resultados semelhantes foram alcançados pelo algoritmo do tipo ganancioso que foi desenvolvido para reduzir o elevado tempo de simulação necessário. Este algoritmo foi utilizado para a otimização da CNN e obteve melhores resultados quando comparando com os reportados na literatura para bancos de dados semelhantes. Embora se supusesse que a frequência cardíaca poderia aumentar a precisão do sistema

implementado com SpO₂. Tal não ocorreu na investigação realizada, no entanto, algumas soluções obtiveram uma sensibilidade superior. A precisão global alcançada pelos métodos baseados em características criadas pelo investigador e pelos métodos baseados na extração automática de características foi superior à reportada no estado da arte em bancos de dados semelhantes. Em conclusão, foi verificado que é possível detetar eventos de apneia e pacientes que sofram de apneia usando um dispositivo de nível quatro, com uma boa precisão, sendo as soluções desenvolvidas facilmente integráveis em diferentes dispositivos vestíveis.

Palavras-chave

Rede neuronal profunda, Escolha de características, Otimização, Rede neuronal, Apneia do sono

Table of Contents

<i>Acknowledgement</i>	ix
<i>Abstract</i>	xi
<i>Resumo</i>	xiii
<i>Table of Contents</i>	xv
<i>List of Figures</i>	xix
<i>List of Tables</i>	xxiii
<i>List of Acronyms</i>	xxv
1. Introduction.....	1
1.1. Introduction.....	2
1.2. Motivation.....	2
1.3. Research Hypothesis	3
1.4. Research Objectives	4
1.5. Research Contribution	5
1.6. Document Structure	9
2. Obstructive Sleep Apnea Detection.....	11
2.1. Sleep Apnea	12
2.2. Polysomnography	13
2.3. Apnea Criteria According to AASM	14
2.4. AHI Index	15
2.5. Type of Test or Device.....	16
2.6. Summary	16
3. Materials for Research	17
3.1. Databases	18
3.2. Classifier and Classifier Parameters	18
3.3. Performance Evaluation Parameters	19
3.4. Summary	20
4. Literature Review.....	21
4.1. Introduction.....	22
4.2. Obstructive Sleep Apnea Detection Approaches	22
4.2.1. Introduction.....	22
4.2.2. Based on pulse oximetry	24
4.2.3. Based on ECG.....	25
4.2.4. Based on respiration.....	28
4.2.5. Based on sound	29
4.2.6. Based on combined approaches	30

4.2.7.	Summary	33
4.3.	Detecting Sleep Apnea using Deep Learning	35
4.3.1.	Introduction	35
4.3.2.	Automatic feature learning using DVNN	36
4.3.3.	Human crafted feature learning using DVNN	37
4.3.4.	Convolutional Neural Network (CNN).....	38
4.3.5.	Recurrent Neural Network (RNN)	40
4.3.6.	Combination of multiple deep networks	41
4.3.7.	Summary	43
4.4.	Summary and Choice of Work Done in This Thesis	45
5.	Handcrafted Feature Based Method.....	47
5.1.	Introduction.....	48
5.2.	Comparison of SFS and mRMR for Oximetry Feature Selection	48
5.2.1.	Introduction	48
5.2.2.	Oximetry feature selection	49
5.2.3.	Performance of the mRMR method	52
5.2.4.	Performance of Sequential Forward Search.....	57
5.2.5.	Comparison between mRMR and SFS	61
5.2.6.	Comparison with other methods	69
5.2.7.	Summary	70
5.3.	Genetic Algorithm for Feature Selection	71
5.3.1.	Introduction	71
5.3.2.	GA based feature selection method	71
5.3.3.	Optimization result.....	73
5.3.4.	Summary	76
5.4.	Self-Configuring Classifier Combination (SC3)	76
5.4.1.	Introduction	76
5.4.2.	Self-Configuring Classifier Combination method	77
5.4.3.	Performance of SC3	80
5.4.4.	Comparison of SF and IF for MaxV SC3	83
5.4.5.	Comparison of SF and IF for WLC SC3.....	85
5.4.6.	Comparison Between MaxV and WLC	86
5.4.7.	Comparison with other work.....	88
5.4.8.	Summary	90
5.5.	Combination of SpO2 and HR using SC3	91
5.5.1.	Introduction	91
5.5.2.	HR features	92
5.5.3.	SpO2 and combination of SpO2 HRV features	95
5.5.4.	Results of epochs based classification	95
5.5.5.	Global classification.....	99
5.5.6.	Summary	100
5.6.	Summary of Handcrafted Feature Based Method.....	101
6.	Automated Feature-Based Methods.....	103
6.1.	Introduction.....	104
6.2.	Multi-Objective Architectural Hyperparameter Optimization of CNN.....	106
6.2.1.	Introduction	106
6.2.2.	Optimization of CNN hyperparameters using GA.....	107
6.2.3.	Performance of the hyperparameter optimization.....	110

6.2.4.	External database performance	115
6.2.5.	Effect of input size	115
6.2.6.	Effect of layers	116
6.2.7.	Transfer learning performance	116
6.2.8.	Comparison with the state of the art works	117
6.2.9.	Summary	119
6.3.	Greedy Based Optimization (GBO) of CNN	119
6.3.1.	Introduction	120
6.3.2.	Classifier structure	120
6.3.3.	Hyper-parameters optimization based on Greedy Algorithm	121
6.3.4.	General result of Greedy Algorithms	125
6.3.5.	Comparison between different Greedy Algorithms	127
6.3.6.	Effect of input size	131
6.3.7.	Responsiveness of the kSize and kNo	132
6.3.8.	Effect of the added CL	133
6.3.9.	Comparison with literature	134
6.3.10.	Summary	136
6.4.	Combination of SpO2 and HRV with CNN	136
6.4.1.	Introduction	136
6.4.2.	Combination of SpO2 and HR	137
6.4.3.	Transfer learning of classifiers and combination of signals	137
6.4.4.	Performance of SpO2, HRV, and SpO2+HRV	137
6.4.5.	Global classification	143
6.4.6.	Summary	145
6.5.	Summary of Automated-Feature based Methods	146
7.	Implementation	147
7.1.	Introduction	148
7.2.	Implementation	148
7.2.1.	Wearable device	149
7.2.2.	Smartphone application	152
7.3.	Summary	154
8.	Conclusion and Future Work	157
8.1.	Conclusion	158
8.2.	Future Work	159
	References	163

List of Figures

Figure 1 : Simplified block diagram of the chapters.	9
Figure 2 : No airway obstruction and airway obstruction during sleep [27] [28].	12
Figure 3 : Polysomnography for adult [55] and children [56]	13
Figure 4 : Flow of Sleep Apnea Detection Review Approaches [87].	23
Figure 5 : Flow chart of the process for article selection using PRISMA reporting style.	36
Figure 6: Oxygen saturation (SpO ₂) in a 5 minute segment.	49
Figure 7 : General pipeline for finding the best feature sub-set from the SpO ₂ for sleep apnea detection.	50
Figure 8 : Accuracy of the Physionet mRMR method.	54
Figure 9 : Sensitivity of the Physionet mRMR method.	54
Figure 10 : Specificity of the Physionet mRMR method.	55
Figure 11: Accuracy of the Ucddb mRMR method.	55
Figure 12 : Sensitivity of the Ucddb mRMR method.	56
Figure 13 : Specificity of the Ucddb mRMR method.	56
Figure 14: Accuracy of the Physionet SFS method.	58
Figure 15 : Sensitivity of the Physionet SFS method.	59
Figure 16 : Specificity of the Physionet SFS method.	59
Figure 17: Accuracy of the Ucddb SFS method.	60
Figure 18 : Sensitivity of the Ucddb SFS method.	60
Figure 19 : Specificity of the Ucddb SFS method.	61
Figure 20 : Overall performance for all the classifiers using mRMR and SFS algorithms in both databases.	63
Figure 21 : Overall selected number of features for all the classifiers using mRMR and SFS algorithms in both databases.	63
Figure 22: The general pipeline of optimization of the sleep apnea detection classifier.	71
Figure 23 : Flow chart of GA.	73
Figure 24 : Optimize structure of the system classifier using the GA and a ANN.	74
Figure 25 : Confusion matrix of a ANN with seven optimized features optimized using the GA over 100 generations.	74
Figure 26 : Non-apnea (Normal) and apnea minute with the wavelet scale coefficients chosen by a GA ANN classifier.	76
Figure 27 : Block diagram of the self-configuring classifier combination (SC3) technique.	78
Figure 28 : Accuracy (Acc), Sensitivity (Sen), and Specificity (Sp) of 1 minute, 3 minute and 5 minute LDA over the generations for the best performance objective.	81
Figure 29 : a) Cost and b) Number of features of 1 minute, 3 minute and 5 minute LDA over the generations for the best performance objective.	81

Figure 30 : Accuracy (Acc), Sensitivity (Sen), and Specificity (Spc) of 1 minute, 3 minute and 5 minute MaxV and WLC over the generations for the best performance objective.	82
Figure 31 : a) Cost and b) Number of features of 1 minute, 3 minute and 5 minute MaxV Independent Feature (IF) and Shared Feature (SF) classification combination over the generations for the best performance objective.	84
Figure 32 : a) Cost b) Number of features of 1 minute, 3 minute and 5 minute WLCISF and WLCIF classification combination over the generations for the best performance objective.	85
Figure 33 : A subtraction of the performance parameters between the MaxV and the WLC (MaxV-WLC) for the HuGCDN2008 database (Num indicates number).....	88
Figure 34 : SpO2 and HR with apnea annotation for five-minute data.....	91
Figure 35 : Cost of the 5-minute MaxVIF classification combination over the generations for the best performance objective.	96
Figure 36 : Accuracy (Acc), Sensitivity (Sen), and Specificity (Spc) of 5 minute MaxVIF over the generations for the best performance objective for the HR and HR+SpO2 signals.....	97
Figure 37 : Number of features of the 5-minute MaxVIF classification combination over the generations for the best performance objective.	97
Figure 38: Comparison of the global accuracy of HuGCDN2008 dataset with AHI calculated by medical physician (AHI G MP) and the by the CNN classifiers' AHI time in bed (AHI C TiB) for SpO2, HRV and SpO2+HRV of 1 Min input. (+) symbol is used for normal subjects (AHI<=5) and (*) is used for apnea patients.	100
Figure 39 : Simplified representation of the CNN hyperparameters optimization strategy using NSGA-II [230].	108
Figure 40 : The multi objective problem space in percentage for the 1st and the 50th generation for a) 1 minute b) 3 minute, and c) 5 minute inputs for the HuGCDN2008 database.	111
Figure 41 : The changes of the multi objectives in percentage, a) Acc, b) Sen, and c) Spc, over the generations (Gen) of the populations (Pop) for 1 minute input for the HuGCDN2008 database.	112
Figure 42 : The changes of the multi objectives in percentage, a)Acc, b)Sen, and c)Spc, over the generations (Gen) of the populations (Pop) for 3 minutes input for the HuGCDN2008 database.....	112
Figure 43 : The changes of the multi objectives in percentage, a)Acc, b)Sen, and c)Spc over the generations (Gen) of the populations (Pop) for 5 minutes input for the HuGCDN2008 database.....	112
Figure 44 : Comparison of the three inputs in the 50 th generation in 3D (Acc, Sen, and Spc) problem space.....	112
Figure 45 : 1 st Pareto front of the solutions for the HuGCDN2008. The first (1 st) and last (50) generation of a a) 1 minute b) 3 minute c) 5 minute Spo2 signal. The solution of the first generation is marked with star and the 50 th is marked with box.	113
Figure 46 : Comparison of three inputs in the 50 th generation in 2D (Sen and 1-Spc) problem space.	113

Figure 47 : The solutions for the first (1 st) and the last (50) generation of a) 1 minute b) 3 minute, and c) 5 minute SpO2 signal. The solution of first generation is marked with a star and 50 th is marked with a box. d) comparison of all of the inputs in 50 th generation.....	115
Figure 48 : Clustered-layer (CL) base scalable CNN1D structure.....	121
Figure 49 : Side by side comparison of the TT and WTT greedy methods.	123
Figure 50 : Best validation (Val) CO solution for the first layer WTT _{RE}	126
Figure 51 : Best (according to validation) test CO solution for each layer ((a), (c), (e)) and total simulation time (in hours((b), (d), (f))) for TT, RE (WTT) and FT(WTT).	131
Figure 52 : Total Mean Absolute Changes (MAC) of different objectives (in validation) with respect to unit parameter (kSize and kNo) change for 1-minute (1 Min), 3-minute (3 Min) and 5-minute (5 Min) inputs.	133
Figure 53 : Best (according to validation) training and validation (Val) CO for each layer WTT _{RE} method.	134
Figure 54 : The validation CO for the SpO2 signal using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.	139
Figure 55 : The test CO for the SpO2 signal using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.	139
Figure 56 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 1 Min input of the SpO2.	139
Figure 57 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 3 Min input of the SpO2.	140
Figure 58 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 5 Min input of the SpO2.	140
Figure 59 : The validation CO for the HRV signal using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.	140
Figure 60 : The test CO for the HRV signal using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.	141
Figure 61 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 1 Min input of the HRV.....	141
Figure 62 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 3 Min input of the HRV.....	141
Figure 63 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 5 Min input of the HRV.....	142
Figure 64 : The validation CO for the combined signals (SpO2+ HRV) using the 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.	142
Figure 65 : The test CO for the combined signals (SpO2+HRV) using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.	142
Figure 66 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for a 1 Min input of the combined signals (SpO2+HRV).....	142
Figure 67 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for a 3 Min input of the combined signals (SpO2+HRV).....	143
Figure 68 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for a 5 Min input of the combined signals (SpO2+ HRV).....	143

Figure 69 : Comparison of the global accuracy of the HuGCDN2008 dataset with the AHI calculated by a medical physician (AHI G MP) and by the CNN classifiers' AHI time in bed (AHI C TiB) for the SpO2, the HRV and the SpO2+HRV of 1 Min input. (+) symbol is used for normal subjects (AHI<=5) and (*) is used for apnea patients.	144
Figure 70 : Comparison of the global accuracy of the HuGCDN2008 dataset with the AHI calculated by a medical physician (AHI G MP) and by the CNN classifiers' AHI time in bed (AHI C TiB) for the SpO2, the HRV and the SpO2+HRV of 3 Min input. (+) symbol is used for normal subjects (AHI<=5) and (*) is used for apnea patients.	145
Figure 71 : Comparison of the global accuracy of the HuGCDN2008 dataset with the AHI calculated by a medical physician (AHI G MP) and by the CNN classifiers' AHI time in bed (AHI C TiB) for the SpO2, the HRV and the SpO2+HRV of 5 Min input. (+) symbol is used for normal subjects (AHI<=5) and (*) is used for apnea patients.	145
Figure 72 : Overall architecture of the proposed system.	149
Figure 73 : User with the watch around their wrist, and the SpO2 sensor placed on their index finger.	150
Figure 74 : User with the watch around their wrist, and the SpO2 sensor placed on their index finger.	150
Figure 75 : Encloser for the battery and microprocessor board (main body of the watch). ...	151
Figure 76 : Top part of the finger clip.	151
Figure 77 : Bottom part of the finger clip encloser for the sensor.	151
Figure 78 : a) Sign in page, and b) choice between collecting and viewing data.	153
Figure 79 : a) Recorded data and b) Results of recorded data.	153
Figure 80 : a) Connection of the sensor b) Putting on watch c) Turning on the sensor.	154
Figure 81 : a) User scan for device connected to the sensor and b) STOP-Bang Questionnaire.	154

List of Tables

Table 1 : Brief advances of OSA according to Allan I. Pack [29].	12
Table 2 : Classifiers used in the work.	19
Table 3 : Evaluation of the Analyzed Algorithms [87].	31
Table 4 : Summary of the database information: The database, year of publication, number of subjects, used signals, window size and type of classifiers(A=apnea, H= hypopnea, N=Normal, S= Severity, O=obstructive, G=Global or OSA Severity) used by selected papers (According to year).	42
Table 5 :Performance of the different works.	42
Table 6: List of SpO2 features.	51
Table 7 : Comparison of the selected features for both databases using the mRMR method (best values are marked in bold).	57
Table 8 : Comparison of selected features for both databases using the sequential forward search method (desired values are marked in bold).	57
Table 9 : mRMR feature rank for the Physionet database.	63
Table 10 : mRMR feature rank for the Ucddb database.	64
Table 11 : SFS feature sequence for the Physionet database.	64
Table 12 : SFS feature sequence for the Ucddb database.	66
Table 13 : Features' importance by frequency (number of times the features are chosen) for the mRMR methods.	67
Table 14 : Features importance by frequency (number of times feature are chosen) for SFS methods.	68
Table 15 : Comparison of sleep apnea detection approaches with SpO2.	70
Table 16 : Comparison of sleep apnea detection approaches.	74
Table 17 : Self configuring classifier combination results for the HuGCDN2008 database using two fold cross validation for different inputs.	82
Table 18 : Self configuring classifier combination results for cross database (trained with HuGCDN2008 and tested with AED) comparison for different inputs.	83
Table 19: Selected Features and Classifiers for Different SC3.	86
Table 20 :Comparison with other works.	89
Table 21: List of HR features.	93
Table 22 : Selected features for HR and SpO2+HR using S3C.	97
Table 23 : Compared to other works.	98
Table 24 : Chromosome decoding techniques and ranges for the CNN.	108
Table 25 : The results of both CNNs trained using the HuGCDN2008 database in two-fold.	111
Table 26 : Chosen CNN's layers and hyperparameters (layer parameters such as the input size, size of the filter, number of the filters represented as a form of [number of filter]@ [vertical width of filter]x [horizontal width of filter]x	

[number of Channels of filter]_ [vertical width of stride]x[horizontal width of stride]).	114
Table 27 : Transfer learning (TL).	117
Table 28 : Comparison with the literature (P is for proposed networks optimized by a GA and trained using the HuGCDN2008 database. TL indicates transfer learning where the proposed networks were retrained using the respective database. ^a Between two networks the best one is showed.).	118
Table 29 : Comparison of the results between the multi-objective method and two greedy based optimization (GBO) methods, the WTT and the TT optimized (Opt) classification using the HuGCDN2008 database, Cross Database(CD), and Transfer Learning (TL)	127
Table 30 : Chosen CNN's layers and hyperparameters by WTT (RE and FT).	128
Table 31: Chosen CNN's layers and hyperparameters by TT 1 minute and 5 minutes WTT, FT.....	129
Table 32 : Test performance of the HuGCDN2008 database for the SpO2, the HRV and the SpO2+HRV for 1 Min, 3 Min and 5 Min inputs.	139
Table 33 : STOP-BANG sleep apnea questionnaire	162

List of Acronyms

Abbreviations and Acronyms used in this work.

Abbreviation and Acronyms	Full form
2,3-DPG	2,3-diphosphoglycerate
AASM	American Academy of Sleep Medicine
Acc	Accuracy
AdaBoost	AdaBoost
AE	Autoencoder
AED	Apnea-ECG database
AF	Air Flow
AHI	Apnea Hyperpnea Index
AHI C TiB	AHI TiB Calculated from the Output of Classifiers
AHI G MP	AHI Calculated by The Medical Professional
AHI TiB	Apnea Hypopnea Minute Per Hours in Bed
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
App	Applications
AUC	Area Under ROC Curve
Avg	Mean
BatchN	Batch Normalization
batchnorm	Batch Normalization
bpm	Beats Per Minutes
C	Central Apnea
CART	Classification and Regression Trees
CC	Cepstrum Coefficient
CD	Cross Database
CF	Cost Function
CNN	Convolution Neural Network
CTM	Central Tendency Measure
CL	Clustered-Layer
CO	Combined Objective
conv	Convolution Layers
CoV	Coefficient of Variation
CWT	Continuous Wavelet Transform
D	Dimension
DAE	Deep Autoencoder
DBN	Deep Belief Network
DIndex	Delta Index
DL	Deep Learning
DNN	Deep Neural Network
DNN-FF	Deep Neural Network Feedforward

DO	Dropout Layers
DVNN	Deep Vanilla Neural Network
EA	Evolutionary Algorithms
EB	Epoch-Based
EBLS	Exclusion of The Borderline Subjects
ECG	Electrocardiography
EDR	ECG Derived Respiration
EEG	Electroencephalogram
EMD	Empirical Mode Decomposition
EMG	Electromyography
EOG	Electrooculogram
ES	Exhaustive Search
F_1	F_1 Score
F_{1w}	Weighted F_1 Score
Fb	Filter Bank
FC	Fully Connected
FD	Frequency Domain
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive
FT	Fine Tuning
F-LSTM	LSTM With Feature Inputs
G	Global
GA	Genetic Algorithm
GAcc	Global Classification Accuracy
GBO	Greedy Based Optimization
GE	Grammatical Evolution
Gen	Generation
GMM	Gaussian Mixture Models
GRU	Gated Recurrent Unit
HL	Hidden Layer
HMM	Hidden Markov Model
HR	Heart Rate
HRV	Heart Rate Variability
HuGCDN2008	Dataset from Sleep Unit of Dr. Negrín was Collected in Gran Canaria University Hospital
HYP	Hypopnea
Hz	Hertz
IF	Independent Feature
IHR	Instantaneous Heart Rates
IIR	Infinite Impulse Response
KNN	K-Nearest Neighbor
kNo	Kernel Number

kSize	Kernel Size
Kurt	Kurtosis
L	Linear
LDA	Linear Discriminant Analysis
LR	Logistic Regression
LSTM	Long Short-Term Memory
LZC	Lempel-Ziv Complexity
M	Mixed Apnea
Max	Maximum
MaxV	Maximum Voting
maxpooling	Maximum Pooling
MCE	Misclassification Error
MESA	Multi-Ethnic Study of Atherosclerosis
MFCC	Mel Frequency Cepstral Coefficient
MGH	Massachusetts General Hospital
MHLNN	Multiple Hidden Layers Neural Network
min	Minimum
Min	Minute
mRMR	Minimum Redundancy Maximum Relevance
MrOS	Osteoporotic Fractures in Men
NB	Naive Bayes Classifier
NCPAP	Nasal Continuous Positive Airway Pressure
NoFL	Numbers of Flexible Layers
NSGA-II	Non-Dominated Sorting Genetic Algorithm II
NSRR	National Sleep Research Resource
O	Obstructive Apnea
ODI	Oxygen Desaturation Index
OL	Output Layer
Opt	Optimized
OSA	Obstructive Sleep Apnea
OSAH	Obstructive Sleep Apnea Hypopnea
P	Polynomial
PCA	Principal Component Analysis
pH	Power of Hydrogen
PLC	Poly lactide
PM	Portable Monitor
Pop	Population
PPG	Photoplethysmogram
PPV	Precision or Positive Predictive Value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRV	Pulse Rate Variability
PSG	Polysomnography
QDA	Quadratic Discriminant Analysis

Rf	Number of Repetitions
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machines
REPTree	Reduced-Error Pruning Tree
RCNN	Combined Deep Recurrent and Convolutional Neural Networks
RE	Rough Estimation
Rec	Recordings from database
REn	Renyi Entropy
ReLU	Rectified Linear Unit
RF	Random Forest
RMS	Root Mean Square
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic Curve
RR	Inter-Beat
RR-ECG	R To R Interval From ECG
RUSBoost	Random Under Sampling Boosting
SAE	Stacked Autoencoder
SB	Subject-Based
SC3	Self-Configuring Classifier Combination
SCSMC	Sleep Center of Samsung Medical Center
Se	Sensitivity
Sen	Recall or Sensitivity
SEn	Shannon Entropy
SF	Shared Feature
SFS	Sequential Forward Selection
SHHS	Sleep Heart Health Study
Sk	Skewness
SNUBH	Seoul National University Bundang Hospital
SNUH	Seoul National University Hospital
SpAE	Sparse Autoencoder
Sp	Specificity
SpC	Specificity
SpO2	Blood Oxygen Saturation Index
SRBD	Sleep-Related Breathing Disorder
Sub	Subjects
SVM	Support Vector Machine
TD	Time Domain
TFD	Time Frequency Domain
TT	Topology Transfer Method
TEt	Test Time
TEO	Teager Energy Operator
TF	Time Frequency
TFMadCA	Median Absolute Deviation of Approximation

TFMadCD	Median Absolute Deviation of Details
TFSdCA	Standard Deviation of Approximation
TFSdCD	Standard Deviation of Details
TFSEnCA	Entropy of Approximation
TFSEnCD	Entropy of Details
TFVarCA	Variance of Approximation
TFVarCD	Variance of Details
TiB	Time in Bed
TL	Transfer Learning
TN	True Negative
TP	True Positive
TRt	Training Time
TW	Time Window
UCD, Ucdb	St. Vincent's University Hospital/University College Dublin Sleep Apnea Database
VAD	Voice Activity Detection
Val	Validation
Var	Variance
WLC	Weighted Linear Combination
WTT	Weighted-Topology Transfer Method

Chapter 1

Introduction

This chapter gives a brief overview of the motivation, objective, and hypothesis. The research done for the thesis is also presented. At the end of the chapter, the thesis structure is provided.

1.1. Introduction

Obstructive Sleep Apnea (OSA) is a common sleep disorder with a high prevalence among the adult population (4% in adult men and 2% in adult women) [1]. It was estimated that over 200 million people suffer from sleep apnea [2]. Polysomnography (PSG) is the gold standard to detect the biophysiological changes that occur during sleep[3]. It monitors different recordings such as Electroencephalogram (EEG), Electrooculogram (EOG), Electromyography (EMG), and Electrocardiography (ECG) during sleep. The number of signals being monitored can increase according to the requirements. Managing a PSG is a tedious and time-consuming task requiring the analysis of multiple signals [4]. There is a high economic cost associated with the equipment maintenance and the limited quantity of professionals [5], [6]. The use of different sensors creates an uncomfortable situation for the patient during sleep. Consequently, a simpler system with a lower number of sensors for apnea detection is desirable.

Blood oxygen saturation (SpO₂), measured by pulse oximetry, could be a suitable signal used to face this challenge since the lack of airflow caused by OSA events frequently causes repetitive oxygen desaturation [7]. SpO₂ sensors are readily usable and are appropriate for portable monitoring [8]. A significant amount of time domain [9] and frequency domain features [10] were developed by different researchers to classify apnea events, creating a vast pool of suitable features. The learning process with all the available features could have a negative effect on the performance generalization, particularly when irrelevant or redundant features are present. One of the solutions could be the combination of the best features found in all the previous works. However, combining two or more independent best features cannot guarantee a better feature set [11]. For this reason, it is necessary to find a subset of dominant and optimum features for a useful classification.

There are two main approaches that could be followed to solve the feature selection of the signal. The first approach consists of choosing the best handcrafted feature set. The other approach consists of automated features estimated by the classifier. In this thesis, both methods were taken into account with the aim of trying to find an optimum solution for apnea detection.

1.2. Motivation

Population-based studies show that OSA is a common sleep disorder. When using an apnea-hypopnea index (AHI) of 15 events/hour, it was estimated an OSA prevalence of 10% and 17% among 30-49 year-old and 50-70 year-old men, respectively. In women, this prevalence is 3% and 9% among 30-49 year-olds and 50-70 year-olds, respectively [12]. Moreover, more than 80% of apneic patients are not aware of the problem [13].

OSA has different effects on a persons' health. Sleep-Related Breathing Disorder (SRBD) was related to Hypertension [14]. Additionally, patients with coronary artery disease are associated with a worse long-term prognosis [15]. SRBD appears to contribute as a risk factor for stroke through hemodynamic and hematologic changes[16]. Each additional apnea or hypopnea per sleep hour increased the fasting insulin level and HOMA-IR by about 0.5% [17].

A national survey of adults living within the United States, conducted by WB&A Market Research and commissioned by the National Sleep Foundation, disclosed some of the effects of a sleep disorder in a person's relationship [18], where 28% of people said their intimate relationship had been affected because they were too sleepy. Thirty five percent of those surveyed note the occurrence of diverse problems in their relationship due to sleep disorders. On the other hand, 26% lose sleep due to their partner's sleep problem, 23% sleep in a separate bed bedroom or couch, 8% alter their sleep schedule, and 7% use an eye mask and/or an earplug [18].

In severe cases of apnea, it can also lead to accidents. The odds of work accidents were found to be nearly double in workers with OSA [19]. It has been recently estimated that 7% of road traffic injuries, for a population of male drivers, involved in motor vehicle accidents are attributable to OSA [20].

The gold standard for diagnosing sleep disorders requires an in-laboratory technician attending a PSG. Although this technique provides detailed and highly accurate results, it has several disadvantages [21]. The analysis with this multi-channel system is time-consuming and labor intensive [4]. It is costly and there is a limited number of professionals for Sleep Apnea diagnosis [5], [6]. For these reasons, patients not only have to pay high diagnostic fees but also have to wait a long time for their turn in waiting lists [22]. In addition to this, during the PSG test, patients have to stay in a hospital (or sleep laboratory) which is not their natural place to sleep and they must cope with multiple uncomfortable sensors which also affect the measured results.

Therefore, an easy to use, relatively cheap setup and one that does not displace the patients from their natural sleep is needed. In addition to this, automated algorithms to detect apnea can reduce the time and cost of the analysis.

1.3. Research Hypothesis

These are the hypothese used for this work:

- It is possible to detect apnea with good accuracy using a single sensor or two signals.
- The features used by a physician and those suitable for automatic classification may be different. In order to detect apnea events, physicians use some basic features, such as the percentage of oxygen desaturation. More complex features can be used by the classifiers because the classifier does not need any visual references to detect apnea events.

- Adding Heart Rate Variability (HRV) to the SpO2 signal increases the accuracy of the system. In some cases, the lack of air flow does not create oxygen desaturation. HRV can fill this void due to the dynamic of HRV in apnea events with some cyclic variations.
- Automatic feature methods (Deep classifiers) can perform better than feature based methods (shallow classifiers). The creation of handcrafted features that achieve good performance requires significant domain knowledge and the combination of two or more features does not guarantee an improvement [11]. In addition to this, the best features are sometimes dependent on the classifiers used. Deep learning can automatically learn features from raw data [23]; thus, by using deep networks these problems can be solved.

1.4. Research Objectives

Alternative diagnostic devices, in place of a PSG have been developed to address the identified issues, allowing for the monitoring of patients in their homes [24]. A proposed categorization of devices was carried out by dividing the methods into four levels [25]. The fourth level devices include one or two signals and could be used at home without the presence of an attendant. These devices use fewer sensors than a PSG and, are more comfortable for patients. In addition to this, home monitoring devices are gaining importance in health care systems and are likely to become the major diagnosis tools in the future. Compared to PSG, these devices are inexpensive and easy to self-assemble. Therefore, a fourth level device was chosen for the research.

Blood oxygen saturation, measured by pulse oximetry (SpO2), could be suitable to face this challenge since the lack of airflow caused by sleep apnea events frequently causes repetitive oxygen desaturation [7]. However, some respiratory pauses do not produce a clear pattern in the oximetry signal. This could be related to the hemoglobin dissociation curve, where short events would not be able to decrease the SpO2 percentage because a marked reduction in the partial oxygen pressure did not occur. In addition to that, pH (Power of Hydrogen), temperature and 2,3-DPG (2,3-diphosphoglycerate) levels, which are specific to each person, can displace the hemoglobin dissociation curve [10]. Investigations in the OSA process pointed out that apnea events showed progressive bradycardia, followed by abrupt tachycardia on the resumption of breathing [26]. For that reason, the HRV or the Heart Rate (HR) can be a good addition to the SpO2 signal. Another added advantage of HRV is that the pulse oximeter used for the measurement of the SpO2 can also measure the HRV. This makes it possible to record two signals with only one device.

The main goals of this thesis are aimed at detecting apnea events with simple fourth level devices providing a new contribution in bio signal analysis methods. In order to face the previously mentioned goals, the following objectives are proposed:

1. Literature review: One of the first objectives is to search and analyze previously published works to understand recent trends and knowledge gaps. One direction for the literature review is the analysis of the performance of different algorithms and methods, which use signals from multiple sources to detect OSA. In addition to this, recent publications show accuracy improvements using deep networks over shallow networks. To provide in-depth knowledge about the applicability of deep learning in the detection of sleep apnea, a second analysis of such works, assessing the performance of the presented methods, is undertaken.
2. Selection and creation of features for handcrafted feature-based methods: namely, for handcrafted feature-based methods to create and choose features to obtain a good performance. Therefore, different types of features will be created and tested to increase the performance.
3. Design of automatic feature-based methods for OSA detection: Implementing automatic feature-based methods presents significant challenges by itself. The structure and/or hyperparameters of the network are typically selected through an experimental search. Such methods require a significant amount of time as well as experience and expert knowledge for the creation of a handcrafted network structure and hyperparameters. Thus, one of the objectives of this thesis is to develop an independent algorithm capable of choosing the featureless network structure and hyperparameters without any human intervention.
4. Comparison of handcrafted and automatic feature methods: A comparison between handcrafted and automatic methods will be performed to discover the advantages and disadvantages of each method.
5. Comparison of SpO₂, HRV and SpO₂ with the HRV based methods: One of this thesis's hypotheses is that adding the HRV will increase the system performance. Thus, a comparison between SpO₂ and SpO₂ with HRV will be carried out.
6. Comparison of different input sizes of the signals: In some published works, authors discovered that longer inputs give better results. This phenomenon will be researched in detail.
7. Building a prototype in order to use the solutions developed throughout the thesis.

1.5. Research Contribution

In this section the research contributions of this work are described.

- Two systematic reviews were performed to analyze the performance of different algorithms and methods for feature based and featureless methods. These works were published in journals named 'IEEE Journal of Biomedical and Health Informatics' and 'Sensors'.
- Three works were conducted to find better features for handcrafted feature-based methods with SpO₂. Two works were published in a journal named 'Neural Computing and

Applications’ and the other one in a conference named ‘International Conference on Information, Communication and Automation Technologies’.

- Two works were carried out with SpO2 and automated feature-based methods. These works were published in journals named ‘IEEE Access’ and ‘Computer Methods and Programs in Biomedicine’
- A handcrafted feature-based work with SpO2 and the HRV was carried out and compared with SpO2 performance. For example, the handcrafted feature-based automated feature-based method with SpO2 and HRV is carried out and compared with the SpO2 performance.
- The global accuracy of the best work developed from the handcrafted feature-based methods and automated feature-based methods, for SpO2, HR, and SpO2+HR, was assessed.
- The attained epoch based, and global accuracy are higher than in any reported work in the current literature.
- A generic application (mobile application) was developed to be used with any of the methods proposed in this work.

List of publication related with the PhD studies

- Peer reviewed journal
1. SS Mostafa, D Baptista, AG Ravelo-García, G.Juliá-Serdá, F. Morgado-Dias, “Greedy based convolutional neural network optimization for detecting apnea”, Computer Methods and Programs in Biomedicine 197, 105640 (IF: 3.632, Q1)
 2. SS Mostafa, F Mendonça, F. Morgado-Dias, G.Juliá-Serdá, AG Ravelo-García, “Multi-Objective Hyperparameter Optimization of Convolutional Neural Network for Obstructive Sleep Apnea Detection”, IEEE Access, Vol:8, 129586-129599, 2020, DOI: 10.1109/ACCESS.2020.3009149, (IF: 3.745, Q1)
 3. F Mendonça, SS Mostafa, F Morgado-Dias, AG Ravelo-García, “Matrix of Lags: A Tool for Analysis of Multiple Dependent Time Series Applied for CAP Scoring”, Vol: 189, 2020, 105314. DOI: 10.1016/j.cmpb.2020.105314. (IF: 3.424, Q1)
 4. F Mendonça, SS Mostafa, F Morgado-Dias, AG Ravelo-García, “A Portable Wireless Device for Cyclic Alternating Pattern Estimation from an EEG Monopolar Derivation”, Entropy 2019, 21, 1203. DOI: 10.3390/e21121203 (IF: 3.031, Q1)
 5. SS Mostafa, F. Mendonça, G. Juliá-Serdá, F. Morgado-Dias, A. G Ravelo-García, “SC3: Self-Configuring Classifier Combination for Obstructive Sleep Apnea”, Neural Comput & Applic (2019). DOI: s00521-019-04582-2. (IF: 4.664, Q1)
 6. F. Mendonça, SS Mostafa, F. Morgado-Dias, A. G Ravelo-García, Thomas Penzel, “Sleep

- Quality of Subjects With and Without Sleep-Disordered Breathing Based on the Cyclic Alternating Pattern Rate Estimation from Single-Lead ECG”, *Physiological Measurement*, Vol: 40, Num: 10, 2019. DOI: 10.1088/1361-6579/ab4f08 (IF: 2.246)
7. SS Mostafa, F. Mendonça, F. Morgado-Dias, A. G Ravelo-García, “A Systematic Review of Detecting Sleep Apnea Using Deep Learning”, *Sensors* 2019, 19, 4934. DOI: 10.3390/s19224934. (IF: 3.031, Q1)
 8. F. Mendonça, SS Mostafa, F. Morgado-Dias, A. G Ravelo-García, Thomas Penzel, “A Review of Approaches for Sleep Quality Analysis”, *IEEE Access*, 2019, DOI: 10.1109/ACCESS.2019.2900345, (IF:3.557, Q1)
 9. F. Mendonça, SS Mostafa, F. Morgado-Dias, A. G Ravelo-García, “Sleep Quality Estimation by Cardiopulmonary Coupling Analysis”, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2018. DOI: 10.1109/TNSRE.2018.2881361. (IF: 3.972, Q1)
 10. D Baptista, SS Mostafa, L Pereira, L Sousa, F Morgado-Dias, “Implementation Strategy of Convolution Neural Networks on Field Programmable Gate Arrays for Appliance Classification Using the Voltage and Current (VI) Trajectory”, *Energies*, Vol 11, Issue 9, p 2460, 2018. DOI: 10.3390/en11092460 (IF:2.676, Q2)
 11. SS Mostafa, MA Awal, M Ahmad, and F Morgado-Dias, “Design of sEMG Based Clench Force Estimator in FPGA using Artificial Neural Networks”, *Neural Computing and Applications*, 2018. DOI; 10.1007/s00521-018-3600-4 (IF: 4.213, Q1)
 12. SS Mostafa, N Horta, AG Ravelo-García, F Morgado-Dias, “Analog Active Filter Design using a Multi Objective Genetic Algorithm”, *International Journal of Electronics and Communications*, 2018. DOI: 10.1016/j.aeue.2018.06.001 (IF: 2.46, Q1)
 13. SS Mostafa, F Morgado-Dias, AG Ravelo-García, “Comparison of SFS and mRMR for Oximetry Feature Selection in Obstructive Sleep Apnea Detection”, *Neural Computing and Applications*, 2018. DOI: 10.1007/s00521-018-3455-8 (IF: 4.213, Q1)
 14. F Mendonça, SS Mostafa, AG Ravelo-García, F Morgado-Dias, T Penzel, “Devices for Home Detection of Obstructive Sleep Apnea: A Review”, *Sleep Medicine Reviews*, 2018. DOI: 10.1016/j.smr.2018.02.004 (IF: 10.602, Q1)
 15. F Mendonça, SS Mostafa, AG Ravelo-García, F Morgado-Dias, T Penzel, “A Review of Obstructive Sleep Apnea Detection Approaches”, *IEEE Journal of Biomedical and Health Informatics*, 2018. DOI: 10.1109/JBHI.2018.2823265. (IF: 3.850, Q1)
 16. F Mendonça, SS Mostafa, F Morgado-Dias, JL Navarro-Mesa, G. Juliá-Serdá, A G Ravelo-García, “A Portable Wireless Device Based on Oximetry for Sleep Apnea Detection”, *Computing*, 2018. DOI: 10.1007/s00607-018-0624-7 (IF: 1.654, Q2)
 17. F Mendonça, A Fred, SS Mostafa, F Morgado-Dias, AG Ravelo-García, “Automatic detection of cyclic alternating pattern”, *Neural Computing and Applications*, 2018. DOI: 10.1007/s00521-018-3474-5. (IF: 4.213, Q1)

- Conferences

1. F Mendonça, SS Mostafa, F Morgado-Dias, AG Ravelo-García, “Cyclic Alternating Pattern Estimation from One EEG Monopolar Derivation Using a Long Short-Term Memory”, 2019 International Conference in Engineering Applications (ICEA), DOI: 10.1109/CEAP.2019.8883470
2. SS Mostafa, F Mendonça, A Ravelo-García, F Morgado-Dias, “Combination of Deep and Shallow Networks for Cyclic Alternating Patterns Detection”, 2018 13th APCA International Conference on Control and Soft Computing (CONTROLO), pp 98-103, 2018
3. F Mendonça, SS Mostafa, F Morgado-Dias and A Ravelo-García, “Sleep Quality Analysis with Cardiopulmonary Coupling”, International Conference on Biomedical Engineering and Applications, 2018.
4. K J Reza, SS Mostafa, M H Milon, S Khatun and M F Jamlos, “Scattering Behaviour Analysis for UWB Antenna in the Vicinity of Heterogeneous Breast Tissues”, International Conference on Biomedical Engineering and Applications, 2018.
5. F Mendonça, A Fred, SS Mostafa, F Morgado-Dias, AG Ravelo-García, “Automatic Detection of a Phases for CAP Classification”, Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM), pp 394-400, Funchal, Portugal.
6. SS Mostafa, JP Carvalho, F Morgado-Dias, A Ravelo-García, “Optimization of sleep apnea detection using SpO2 and ANN”, XXVI International Conference on Information, Communication and Automation Technologies (ICAT), pp. 1-6, Sarajevo, Bosnia-Herzegovina 2017. DOI: 10.1109/ICAT.2017.8171609
7. SS Mostafa, MA Awal, M Ahmad, F Morgado-Dias, “A Method for Designing EMG Integrator using an FPGA”, 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), pp. 1-6, Funchal, Portugal. DOI: 10.1109/IWOBI.2017.7985523
8. F Mendonça, SS Mostafa, F Morgado-Dias, J Navarro-Mesa, G Julia-Serda, A Ravelo-García “A minimally invasive portable system for sleep apnea detection” 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), pp. 1-5, Funchal, Portugal. DOI: 10.1109/IWOBI.2017.7985524
9. SS Mostafa, LN Sousa, NF Ferreira, RM Sousa, J Santos, F Morgado-Dias, M. Wány, “FPGA Implementation of Gamma Correction using a Piecewise Linear Approach for a Small Size Endoscopic Camera”, Electronic Imaging, Image Sensors and Imaging Systems 2016, pp. 1-6(6), San Francisco, USA. DOI: 10.2352/ISSN.2470-1173.2016.12.IMSE-276

- Awards
 1. Best student paper award in the 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI).
 2. 2017 Cátedra Telefónica da Universidad de Las Palmas de Gran Canaria.

1.6. Document Structure

The present document is organized into eight chapters, following the simplified block diagram presented in Figure 1, according to the outline:

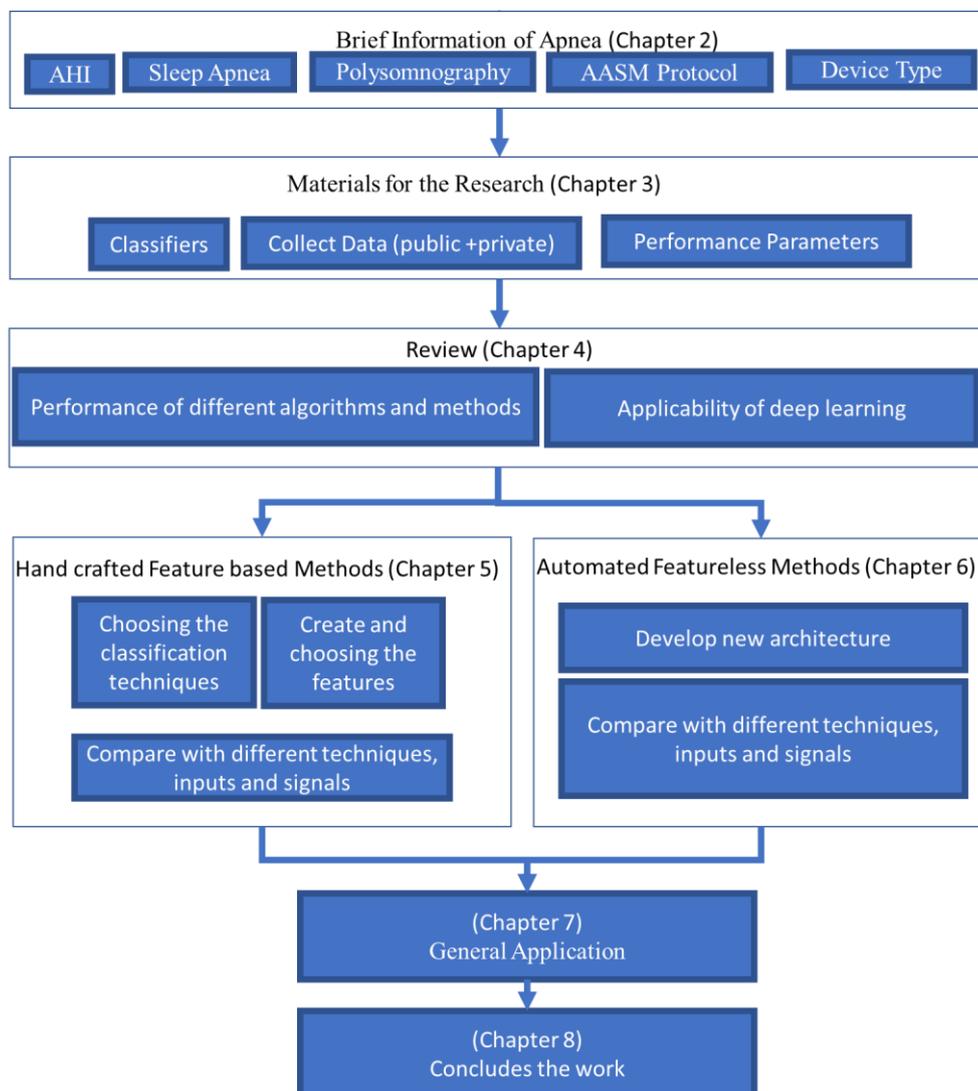


Figure 1 : Simplified block diagram of the chapters.

- ✓ Chapter 2 – A brief introduction on apnea, Polysomnography, and different definitions, and protocols related to apnea are discussed.

- ✓ Chapter 3 –The databases, the classifiers and the performance parameters are discussed.
- ✓ Chapter 4 – Related work/Literature Review (Systematic review) of the past work is presented.
- ✓ Chapter 5 – Some of the proposed research ideas with handcrafted features are presented, and an analysis is provided.
- ✓ Chapter 6 – Some of the proposed research ideas with automated features-based methods are presented, and an analysis is provided.
- ✓ Chapter 7 – One generic prototype implementing previously proposed work is presented.
- ✓ Chapter 8 –The conclusion and future proposals are presented in this chapter.

Chapter 2

Obstructive Sleep Apnea Detection

This chapter covers the definitions and basic concepts of Apnea, Polysomnography and the different protocols related to Obstructive Sleep apnea.

2.1. Sleep Apnea

Sleep apnea (also known as sleep apnoea), is a type of sleep disorder characterized by periods of shallow breathing or pauses in breathing during sleep. There are three types of sleep apnea: Obstructive (OSA), Central, and Complex. OSA occurs when the throat muscles intermittently relax and block the airway during sleep, as presented in Figure 2. Central sleep apnea occurs when the brain does not send the proper signals to the muscles that control breathing. Complex sleep apnea is present if both OSA and central sleep apnea occur in a subject.

A noticeable sign of obstructive sleep apnea could be snoring. The first modern advances of defining apnea were done in 1965 in Pickwickian syndrome where the upper airway obstruction was defined as a major pathogenetic mechanism. A brief summary of major advances is shown in Table 1. Obstructive Sleep Apnea can be characterized by two types of breathing interruption one is apnea, and the other one is hypopnea.

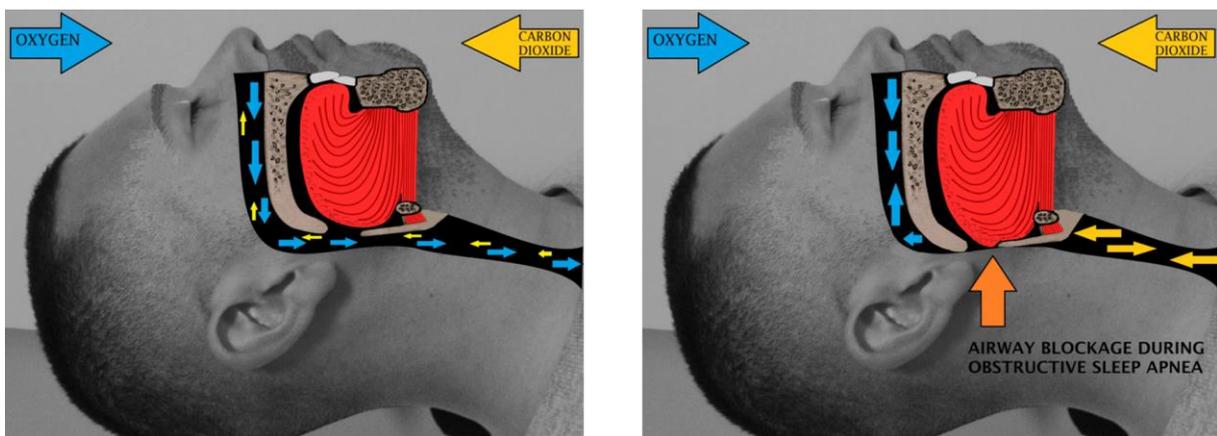


Figure 2 : No airway obstruction and airway obstruction during sleep [27] [28].

Table 1 : Brief advances of OSA according to Allan I. Pack [29].

Year	Milestones
1818, 1854	Cheyne (1818) [30] and Stokes (1854)[31] of Cheyne-Stokes' respiration was described.
1956	Alveolar Hypoventilation in obesity (Pickwickian syndrome) [32] was described.
1960, 1962	In patients with Pickwickian Periodic cessation of respiration was recognized [33][34].
1965	Airway obstruction in sleep (i.e., OSA) was recognized as the cause of cessation of respiration [35].
1971, 1974	The effectiveness of tracheostomy in patients with OSA case reports published [36][37].
1976	Pediatric sleep apnea case series [38]
1978	Unifying concept pathogenesis of OSA was described [39].
1981	Nasal CPAP[40]; specific surgery for OSA[41] were described.
1983	CO ₂ -dependent apnea threshold during sleep[42] was identified.
1988	Hypopneas have the same consequences as apneas [43] was found.
1992	Identification of neuromuscular compensation[44]; intermittent repetitive hypoxia leads to hypertension[45].
1993	High prevalence of OSA [1] was found in a study with robust epidemiologic methods.
1995	Family aggregation was shown: Cleveland Family Study[46], in Israel [47], and in relatively nonobese Scots[48].
1997	Hypertension was linked to induced obstructive apneas in dogs [49].
1998	A link with poor academic performance and high prevalence of OSA in school children was found and improvement was done using surgical treatment [50].
1999	Nasal Continuous Positive Airway Pressure's (NCPAP) efficiency in sleep apnea syndrome was identified [51].
2002	Identified that NCPAP reduces blood pressure in a randomized, placebo-controlled trial [52].

2.2. Polysomnography

PSG (Figure 3) is the gold standard for OSA diagnosis measuring multiple sensors to record the airflow, respiratory movement, SpO₂, EEG, EOG, EMG, ECG, and body position [53]. It can measure sleep stages, airflow, respiratory effort, oxygen saturation, electrocardiogram, and body position, and optional measures such as limb movement, vocalization, and carbon-dioxide level. If the cessation of airflow occurs with a concomitant respiratory muscle effort, it is called OSA, and the disorder severity is determined by the Apnea-Hypopnea Index (AHI). OSA is diagnosed if the patient has reported the indicated symptoms and presents five or more obstructive respiratory events per hour of sleep during a PSG recording[54] as defined by the American Academy of Sleep Medicine (AASM). A brief description of the signals or sensors used by PSG is given below.



Figure 3 : Polysomnography for adult [55] and children [56] .

Airflow: Oronasal airflow is one of the most direct indicators of breathing disorders to detect OSA.

Respiratory movement: In the human being, contraction of the muscle of the diaphragm and the intercostal muscles between the ribs helps to raise the ribs and expand the lungs to draw air through the inspiration process. In the expiration process the opposite happens. These movements can be used as a detecting factor of the airflow. Thus, it can be used in OSA detection.

Oxygen saturation (SpO₂): This signal measures the level of oxygen in the blood. This measurement is commonly performed using a pulse oximeter that calculates the difference between the absorption of infrared and red lights to estimate the oxygen level.

Electroencephalogram (EEG): This signal registers the electrical activity of the brain and in the PSG test, it is useful to detect the sleep stages.

Electrooculogram (EOG): Electrooculography or EOG records eye movements. It helps to detect rapid eye movement (REM) sleep.

EMG: Electromyogram (EMG) measures the muscle tension in the body, and it allows one to monitor the number of leg movements during sleep.

Electrocardiogram (ECG/EKG): This signal measures the electrical activity of the heart. It is used to monitor heart activity and detect pathologies allowing calculation of the HRV.

2.3. Apnea Criteria According to AASM

Apnea is a total blockage of the airway lasting for 10 seconds or more. According to the American Academy of Sleep Medicine (AASM) 2012 [57] the scoring of apneas follows these rules:

“1. Score a respiratory event as an apnea when both of the following criteria are met: N1, N2, N3,

a. There is a drop in the peak signal excursion by $\geq 90\%$ of pre-event baseline using an oronasal thermal sensor (diagnostic study), PAP device flow (titration study) or an alternative apnea sensor (diagnostic study).

b. The duration of the $\geq 90\%$ drop in sensor signal is ≥ 10 seconds.

2. Score an apnea as obstructive if it meets apnea criteria and is associated with continued or increased inspiratory effort throughout the entire period of absent airflow.

3. Score an apnea as central if it meets apnea criteria and is associated with absent inspiratory effort throughout the entire period of absent airflow.

4. Score an apnea as mixed if it meets apnea criteria and is associated with absent inspiratory effort in the initial portion of the event, followed by resumption of inspiratory effort in the second portion of the event. N5.

Note 1. Identification of an apnea does not require a minimum desaturation criterion.

Note 2. If a portion of a respiratory event that would otherwise meet criteria for a hypopnea meets criteria for apnea, the entire event should be scored as an apnea.

Note 3. If the apnea or hypopnea event begins or ends during an epoch that is scored as sleep, then the corresponding respiratory event can be scored and included in the computation of the apnea hypopnea index (AHI). This situation usually occurs when an individual has a high AHI with events occurring so frequently that sleep is severely disrupted and epochs may end up being scored as wake even though <15 seconds of sleep is present during the epoch containing that portion of the respiratory event. However, if the apnea or hypopnea occurs entirely during an epoch scored as wake, it should not be scored or counted towards the apnea hypopnea index because of the difficulty of defining a denominator in this situation. If these occurrences are a prominent feature of the polysomnogram and/or interfere with sleep onset, their presence should be mentioned in the narrative summary of the study.

Note 4. For alternative apnea sensors see Technical Specifications for adults A.2.

Note 5. There is not sufficient evidence to support a specific duration of the central and obstructive

components of a mixed apnea; thus, specific durations of these components are not recommended.”

AASM 2012 [57] also define the scoring of Hypopneas as
“1. Score a respiratory event as a hypopnea if ALL of the following criteria are met: N1, N2, N3.”

“a. The peak signal excursions drop by $\geq 30\%$ of pre-event baseline using nasal pressure (diagnostic study), PAP device flow (titration study), or an alternative hypopnea sensor (diagnostic study).

b. The duration of the $\geq 30\%$ drop in signal excursion is ≥ 10 seconds.

c. There is a $\geq 3\%$ oxygen desaturation from pre-event baseline or the event is associated with an arousal.

2. If electing to score obstructive hypopneas, score a hypopnea as obstructive if ANY of the following criteria are met:

a. Snoring during the event.

b. Increased inspiratory flattening of the nasal pressure or PAP device flow signal compared to baseline breathing.

c. Associated thoracoabdominal paradox occurs during the event but not during pre-event breathing

3. If electing to score central hypopneas, score a hypopnea as central if NONE of the following criteria are met:

a. Snoring during the event.

b. Increased inspiratory flattening of the nasal pressure or PAP device flow signal compared to baseline breathing.

c. Associated thoracoabdominal paradox occurs during the event but not during pre-event breathing.

Note 1. If necessary, the number of hypopneas using a definition requiring a $\geq 30\%$ drop in flow for ≥ 10 seconds that is associated with $\geq 4\%$ desaturation may additionally be reported to qualify a patient for PAP reimbursement (eg. Medicaid or Medicare patients).

Note 2. For alternative hypopnea sensors see Technical Specifications for adults A.4.

Note 3. Supplemental oxygen may blunt desaturation. There are currently no scoring guidelines for when a patient is on supplemental oxygen and no desaturation is noted. If the diagnostic study is performed while the subject is on supplemental oxygen, its presence should be mentioned in the narrative summary of the study.” [57]

2.4. AHI Index

AHI is the number of apnea hypopnea events per hour of sleep. This index is used for the categorization of apnea patients. For adults, an AHI lower than five is considered normal. OSA severity can be defined as mild ($5 \leq \text{AHI} < 15$ events/hour), moderate ($15 \leq \text{AHI} < 30$ events/hour) or severe ($\text{AHI} \geq 30$ events/hour). A brief description of the signals is given below [57].

2.5. Type of Test or Device

A portable monitor (PM) has been used as an alternative to in-laboratory PSG since it is less expensive and easy to deploy. The type of sleep apnea detection devices can be divided into four groups. Sometimes these groups are also defined as ‘Level’.

Type I: monitored standard polysomnography with a minimum of seven channels, including EEG (C4-A1 or C3- A2), EOG, chin EMG, ECG, airflow, respiratory effort, and oxygen saturation.

Type II: portable PSG is, with a minimum of seven channels, similar to Type I test except without an attendant.

Type III: modified portable sleep apnea testing, which uses a minimum of sensors: including two channels of respiratory movement, or respiratory movement and air flow, HR or ECG, and oxygen saturation.

Type IV: contains single or dual-bio parameter recording, typically including oxygen saturation or airflow [58][59].

2.6. Summary

The OSA events are defined by breathing problems caused by blocking the airways. There are four different stages, according to the AHI index.

The gold standard for apnea detection is PSG. However, according to the sensor used and implemented condition, the devices for the tests can be divided into four categories. Because of the lower number of sensors, self-assembly facility, price, and simplicity, a Type IV device is chosen for this work.

Chapter 3

Materials for Research

This chapter focuses on databases, parameters, and classifiers used in this work. The main goal is to give brief details on several concepts and information which are used in this work repeatedly.

3.1. Databases

Three databases were used in this work. Two of them were collected from Physionet and are freely available: the Physionet apnea ECG database (AED) [60], [61] and the St. Vincent's University Hospital / University College Dublin Sleep Apnea Database (UCD) [62]. A third database was collected in Hospital Universitario de Gran Canaria, Dr. Negrín. These databases are annotated with hypopnea (HYP) and central (C), obstructive (O) and mixed (M) apnea. For this work, all of the events are treated as apnea events.

The database details are as follows:

- AED has 70 recordings but only has eight SpO₂ signals available. These eight recordings range from 7 to 10 hours with minute-by-minute annotation [60], [61]. The sampling frequency is 50Hz. Like HuGCDN2008, a 3% desaturation index was used to annotate the dataset. In some sections, this database is indicated as the Physionet dataset.
- UDB [62] has 25 referred (suspected to have OSA) subjects (21 males and 4 females). This database is continuously annotated. The sampling rate of the SpO₂ signal was 8 Hz. In some sections, this database is indicated as the UCDDDB dataset. Alternatively to the previous datasets, the annotation in this dataset is continuous. A minute by minute apnea annotation was created by considering five seconds or more continuous apnea event in the annotated minute.
- The dataset from the Sleep Unit of Dr. Negrín was collected in Gran Canaria University Hospital and has 70 referred (suspected to have sleep apnea) patients (51 males and 19 females, ranging from 18 to 82 years old) which will be referred to as the HuGCDN2008 database. The subjects do not have any arrhythmia and the SpO₂ signal was sampled at 50 Hz and 16 bits resolution. The ECG digitized at 200 Hz and 16 bits resolution. The annotations were made in 30 s epochs [10]. Data were collected using the VIASYS Healthcare Inc. (Wilmington, MA, USA) device and a desaturation level of 3% with AASM 2007 [63] criterion was used for labelling the data.

3.2. Classifier and Classifier Parameters

The classification system plays a big role in biomedical signal processing and decision making. Due to its strengths and limitations, a single classification method is not suitable for all real-world problems. Most of the recent studies use experimental approaches to compare classifiers [64]. A brief description of the classifiers used in this work is presented in Table 2. These classifiers were chosen because they were reported as being suitable for OSA diagnosis in previous research works [2] [10] [22].

Table 2 : Classifiers used in the work.

Classifier	Brief Description
Artificial neural network (ANN)	A ANN is inspired by biological neurons. Though at the beginning it was closer to human the brain over time deviations from biology happened due to solving specific task. Like neurons, the node in ANN receives information and processes it by weight and bias. The theory based on the interpretation of the Kolmogorov's superposition theorem of continuous functions as an ANN [65] does not have this dependency on the training set. If the input layer consists of n_i inputs where $n_i \geq 2$, the numbers of neurons in the hidden layer should be $2n_i + 1$ [66]. Though the theory is applicable to $n_i \geq 2$, for consistency in point of view, it is also used for $n_i = 1$.
Support Vector Machine (SVM)	Vapnik proposed a supervised based (SVM) in 1992 [67], [68], [69], [70]. Apnea detection is a binary classification problem. The SVM treats apnea and normal events as two distinct classes ($y_i \in \{-1,1\}$)
k Nearest Neighbor (KNN)	The KNN makes use of the majority vote of an object's neighbors. The votes are from its k nearest neighbors (where k is a positive integer) [71]. The value of k , chosen for the best performance is dependent on data. Previous work have considered different values (k=27 [72] (ECG) or k=5(SpO2 ,ECG) [73]). For evaluation purposes in this thesis, different values of k were chosen in different works.
Linear Discriminant Analysis (LDA)	An LDA is a parametric modeling technique where the model parameters are calculated from the features needed to optimize the performance taking into account the output [74].
Naive Bayes classifier (NB)	The NB is based on Bayes theorem with an assumption of conditional independence among the features [75], [76]. In addition to that, though it conditional independence is assumed, the classifier can still work when there are dependencies among the attributes [77], [75].
Convolution Neural Network (CNN)	A CNN commonly comprises different types of layers. In this work the CNN has an input layer, convolution layers, nonlinear layers, fully connected layers, batch normalization layers, softmax layer and a class output layer. The input layer is the first layer of the network and receives the raw data. Thus, the size of the input layer is the same as the input data, thus, three input layers were tested. Different convolution kernels of the convolution layer, or filters, slide over the input providing a diversity of features that captures different local information [78]. The learning algorithm chooses the values of the filters during the training process [79]. The feature map size depends on the number of filters (sometimes referred to as depth), filter size (for 1D only the width is considered), stride (which is the number of sample points the filters slide in each step) and padding (adding zeros at the end of the data input so that filters can run on the bordering elements).

3.3. Performance Evaluation Parameters

For the performance evaluation, three parameters are used for different features and classifiers: Accuracy (Acc), Sensitivity (Se) and Specificity (Sp). These parameters are calculated from true Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) which refer to apnea correctly identified as apnea, not apnea incorrectly identified as apnea, not apnea correctly identified as not apnea and apnea incorrectly identified as not apnea, respectively.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (1)$$

$$Se(Sen) = \frac{TP}{TP+FN} \times 100 \quad (2)$$

$$Sp(Spc) = \frac{TN}{TN+FP} \times 100 \quad (3)$$

$$CO = \frac{1}{3} \times (Acc + Se + Sp) \quad (4)$$

3.4. Summary

Different databases are used for comparing the results present in different documents which are present in the literature and understanding the versatility of developed classifiers. Multiple classifiers are studied to develop and test a variety of solutions.

Chapter 4

Literature Review

This chapter focuses on previously published works and analyzes them to understand the knowledge gap and future trends.

4.1. Introduction

The four main tasks for automatic recording and analysis of sleep were identified in a review performed by Penzel and Conradt [80] in 2000. A review of computer-aided approaches for OSA diagnosis was presented by Alvarez-Estevez and Moret-Bonillo [81] covering papers from 1999 to 2013. The review analyzes sleep apnea-hypopnea syndrome screening approaches, apnea event detection and classification methods, comprehensive diagnostic systems, and commercial approaches. However, the inclusion criteria restricted the analysis to papers that used, at least, a subset of the PSG signals proposed in the AASM manual for the scoring of sleep and associated events at the time of publication, excluding the nonstandard biomedical signals for the OSA diagnosis such as ECG or pulse wave analysis. Multiple reviews have been performed with a focus on devices for home detection [82] [83] [84] and some papers focused completely on smartphone applications, for example [85][86].

Two systematic literature reviews were performed and are presented in this chapter since most of the state of the art reviews available are outdated and have a different focus than the goal of this research. The first review is about the different algorithms for OSA detection by evaluating multiple source sensors (Section 4.2 which was published in ‘IEEE Journal of Biomedical and Health Informatics’ [87]). The second is more focused on deep learning based classifiers (Section 4.3 which was published in ‘Sensors’ journal [88]).

4.2. Obstructive Sleep Apnea Detection Approaches

4.2.1. Introduction

There was no reference found in the cutting-edge reviews of a global survey of algorithms for OSA detection by evaluating multiple source sensors. Therefore, the focus of this review is the analysis of the performance of different algorithms and methods, that use signals from multiple source sensors but have not been implemented in hardware, to detect OSA. A literature review was conducted to suggest future implementations, without performing an in-depth review indicating every published paper in each field.

Flemons et al. [84] reviewed papers before 2003. As a result, a literature review covering papers published between 2003 and 2017 was undertaken. The search was conducted using the Web of Science, IEEE explorer, PubMed, and the cited literature in the included articles and various journals. The keywords employed in the search were “algorithm AND sleep apnea”, “oximetry AND apnea”, “ECG AND apnea”, “Respiration analysis AND apnea”, “snoring AND apnea”, “sound AND apnea” and “apnea AND deep”. The inclusion criterion was the presentation of an algorithm that had not been

implemented in hardware but had been verified by at least one investigation, focusing on the performance of obstructive sleep apnea classification, with data extracted from a PSG performed in a hospital or available in a database. An exclusion criterion was the absence of all the diagnostic elements analyzed in this review. Even though hundreds of papers met this criterion, a total of 84 original research articles that had the potential to be promising diagnostic tools were selected to cover multiple solutions, selecting papers with the highest results when the same method was used.

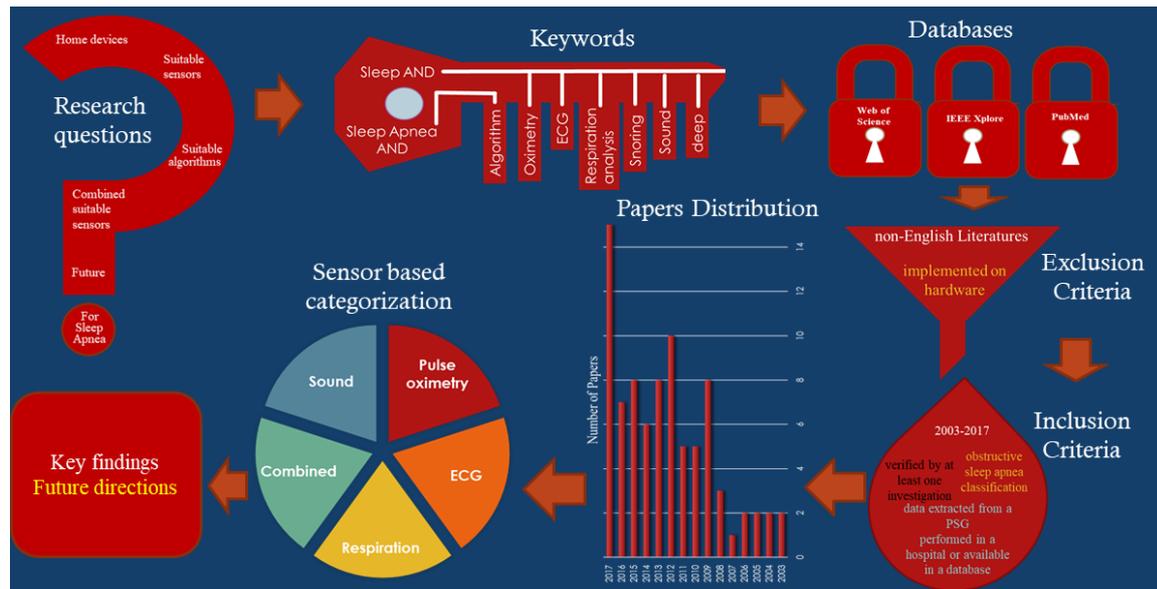


Figure 4 : Flow of Sleep Apnea Detection Review Approaches [87].

The analyzed algorithms were divided into five categories depending on the source sensor: pulse oximetry, ECG, respiration, sound, and combined approaches. A final general analysis was performed to determine the most suitable algorithms. The data were acquired directly from the papers, and the analyzed diagnostic elements were Accuracy, Sensitivity, Specificity and Area Under the Receiver Operating Characteristic Curve. Two groups of diagnostic elements were used according to the applied methodology: Subject-Based, classifying globally every subject, and Epoch-Based, classifying individually every epoch for all subjects. The Global classification specifies the accuracy of the analyzed algorithm classifying the subject.

The population used for validating the system is indicated for each paper, and the data acquisition location has been added, whether it was a hospital or a database. The databases referenced are PhysioNet apnea-ECG [61]; University College Dublin Sleep Apnea [60]; Sleep Heart Health Study [89]; MIT-BIH polysomnography database [90]; and Scaling Up Scientific Discovery in Sleep[91]. The time window used to classify the data are also presented.

Different approaches have been followed with the aim of detecting OSA. Evaluation of the reported results of the algorithms is presented in Table 3. The table is divided in five categories, corresponding

to the approaches analyzed, and each category is ordered by the year of publication. For articles that were published in the same year, the order is according to the reported global classification or EB accuracy. Taking into account the difficulties faced in drawing comparisons between methods due to the different application conditions and databases used for the experiments, a first approximation of results was performed.

4.2.2. Based on pulse oximetry

Classical oximetry analysis includes the oxygen desaturation index (ODI), cumulative time spent below a defined saturation threshold (e.g. time below 90% referred to as T90), and number of drops in the SpO₂ value below the defined baseline and signal variability (usually the delta index). A threshold approach was presented by Jung et al. [92] taking three points into consideration. Álvarez et al. [93] applied two non-linear methods to determine which would be better to improve the OSA detection capability further: the central tendency measure (CTM), and Lempel-Ziv complexity (LZC), which provides a complexity measure. It was determined that the CTM produces the best results.

A pattern classification approach was used by Marcos et al. [94], using a three stage algorithm. Four statistical pattern recognition techniques were analyzed by Marcos et al. [95], namely, the k-nearest neighbor (KNN), quadratic discriminant analysis (QDA), LDA, and logistic regression (LR). The best result was obtained using the LDA with spectral features, performed on the 0.01 to 0.033 Hz band, extracting the area enclosed in the band, the peak amplitude in the band, and the total area of the PSD. A comparison between a support vector machine (SVM) and a KNN was performed by Morales et al. [96]. The best results were provided by a KNN with five neighbors. A genetic algorithm approach was applied by Álvarez et al. [97] at the feature selection stage. The selected features were further fed into an LR classifier.

A probabilistic ANN was used by Morillo and Gross [98], with five neurons on the input layer analyzing the AHI. A three-layered (input, output, and hidden) feed-forward ANN was used as a classifier by Almazaydeh et al. [99]. It uses an ODI, a delta index and CTM as inputs. The pure linear transfer function was used as the activation function of the output layer during the training phase of the ANN. An ANN was also proposed by Álvarez et al. [100], where LZC, CTM, sample entropy, statistical moments in the time domain, spectral entropy, and the relative power in the apnea-related frequency band were used as inputs of the ANN. The same kind of ANN classifier was used by Marcos et al. [101], where normalized features were extracted from the SpO₂ data, using approximate entropy, CTM, and LZC.

Time, frequency, and time-frequency based features were analyzed by Mostafa et al. [102]. A genetic algorithm was applied for feature selection and the classification was performed by a ANN.

A combination of features was extracted from time and frequency domain statistics, the spectral characteristics were driven by PSD, and nonlinear measures were analyzed by Álvarez et al. [103]. The photoplethysmogram (PPG) signal, obtained from a pulse oximeter, was used by Lázaro et al. [104]. The algorithm looks for decreases in the amplitude fluctuations of the PPG signal, and an LDA classifies the data, using features based on the pulse rate variability (PRV). A combination of characteristics in time and frequency domains obtained from PRV and SpO2 signals were employed by Garde et al. [105], with OSA classification performed using the LDA.

An unsupervised feature learning model based on a three-layer deep auto encoder network was implemented by Mostafa et al. [106]. The first two layers were restricted Boltzmann machines, composed by an autoencoder, and the last was a soft-max layer. A Long Short-Term Memory - Recurrent Neural Network (LSTM-RNN) was employed by Pathinarupothi et al. [107] with 60 neurons in the input layer (the oximetry signal has 60 samples, each corresponding to a second of data) and 32 memory blocks, with one cell each, in the hidden layer.

4.2.3. *Based on ECG*

The analysis of ECG waveforms and ECG-derived heart rate are commonly used to detect sleep-related breathing disorders. Lin et al. [108] separated the ECG signal into four spectral components that were used as the training input for a four-layer ANN, implemented with simple neural computing elements.

A discrete wavelet transform, Symlet wavelet with order 3, was used by Khandoker et al. [22] to decompose the ECG signal into 8 levels of detailed coefficients used by a ANN for classification. Wavelet decomposition was applied by Rachim et al. [109], using the Debauches 4 wavelet, to obtain statistical features, and the Principal Component Analysis (PCA) was also performed. Then an SVM with a Gaussian radial basis function kernel was used to classify the data. The Tunable-Q factor wavelet transform was applied by Hassan [110]. The scale and feature factors from this model were then fed into the Adaptive Boosting (AdaBoost) classifier. The same kind of wavelet transform was used by Hassan and Haque [111] to decompose the EEG signal, analyzing the variance, kurtosis and skewness of the decomposition to feed as features for the Random Under Sampling Boosting (RUSBoost) classifier.

Variational mode decomposition was used by Smruthy and Suchetha [112] to decompose the ECG signal into multiple variational mode functions and the selected functions were then added together to generate the reconstructed signal. The standard deviation of the peak to peak distance and mean energy, calculated using the Teager Energy Operator (TEO), was determined from the reconstructed signal used to feed an SVM classifier. Hassan [113] used Empirical Mode Decomposition (EMD) to

generate localized time-frequency estimation, from the ECG signal. Then the mean, variance, skewness, and kurtosis were determined and used on an extreme learning machine, with a sigmoidal activation function, for a single hidden layer feed-forward ANN which classifies the data.

From the ECG-derived HR it is possible to analyze the HRV and the inter-beat (RR) interval that can be defined as the interval between successive QRS points. Quiceno-Manrique et al. [114] employed an analysis based on the HRV. The classification was performed using a KNN. Time-frequency based stochastic features were used by Martínez-Vargas et al. [115] to analyze the HRV employing linear frequency Cepstral coefficients. The highest accuracy was produced using linear label-conditioned correlation as a supervised measure of relevance, selecting seven frequency bands as features for the KNN.

An algorithm based on the analysis of cyclical variations of the HR for detection of sleep-disordered breathing, was presented by Kesper et al. [116]. The algorithm analyzes the correlation of a reference pattern, which represents a decrease in HR, with the beat-to-beat HR curve. Ravelo-García et al. [117] employed a non-linear HRV analysis using a symbolic dynamics method applied to the RR series, transforming it into a sequence of symbols. These symbols were defined through a set of rules that considered the use of three chosen thresholds. Then the classification was performed using an LR model that integrates clinical and physical variables. Zywiets et al. [118] used an LDA for classification with features based on information of four frequency bands: ULF, VLF, LF, and HF.

Time and frequency domain entropies were used by Gutiérrez-Tobal et al. [119]. This information was used as features for an LR classifier. The HRV was used by Roche et al. [120], where the wavelet transform was used to decompose the signal and the classification was performed using the Classification And Regression Trees (CART) method.

Two classifiers, LDA and QDA, were tested by Ravelo-Garcia et al. [121] and were fed with cepstrum features obtained from the RR series. The best results were obtained using a QDA. The RR interval and QRS area were derived from a single lead of the ECG signal by Mendez et al. [122]. RR intervals were also employed by Cheng et al. [123] to reconstruct a nonlinear state space. OSA classification was performed by regularized LR.

Chen and Zhang's [124] approach was to map the individual long-term RR intervals into a disease state space. A severity index was produced to represent the severity of the disease, based on the state change points, and was calculated with a general formula. This index is used as the feature for three classifiers, specifically, LDA, SVM, and LR. The last classifier produced the best results. Three classifiers were analyzed by Yilmaz et al. [125] where the RR series was obtained from a single lead of the ECG signal using an R-peak detection technique, where the R waves were differentiated by

studying the curvature and the amplitude of the ECG signal. The classifiers studied were a KNN, a QDA, and an SVM. The best accuracy was obtained using the SVM. An SVM was also used by Almazaydeh et al. [53] where the RR interval was derived from the ECG signal using an R-peak detection technique. The selected features used to feed the SVM were proposed by Chazal et al. [126] and Yilmaz et al. [125] who used a linear kernel function to map the training data into a kernel space. Additionally, the RR series was the base of the detection algorithm presented by Ravelo et al. [127]. The best results were obtained using the SVM among tested Gaussian Mixture Models (GMM) and the SVM. This classifier was also used by Travieso et al. [128], applying a kernel based on a Hidden Markov Model (HMM) over the Cepstral coefficients obtained from the RR series.

A classifier combination approach was presented by Nguyen et al. [129]. Two binary classifiers were used, an SVM and a 10 neuron hidden layer ANN. These classifiers use features from HRV and recurrence quantification analysis of the HRV. The classifier outputs were combined using a soft decision fusion rule that performs a weighted sum of the output scores. Chen et al. [130] applied the RR intervals to the signal segmentation using an iterated cumulative sum of squares algorithm that searches for the small variation changes in the time series due to OSA, and the SVM was used to classify the data. LSTM-RNN was used by Cheng et al. [131] to classify the RR intervals. The network architecture has four recurrent layers, each followed by a normalization layer, and a softmax classifier. HRV, in the form of instantaneous HR with a constant number of beats (beat window), was used by Pathinarupothi et al. [132] to feed a 2-layer stacked LSTM-RNN with two memory blocks in each layer. It was verified that a beat window of 60 beats provides the best results.

The combination of the analyses of HR and morphology of the ECG can be used to reliably detect the sleep disordered breathing, as analyzed by Penzel et al. [133] using the cardiopulmonary coupling.

EDR and RR interval signals were used by Chazal et al. [126] to obtain features for an LDA classifier that generates a discriminant value. This value was compared to a threshold to detect OSA. RR intervals and EDR signals were also used by Mendez et al. [72] with the PSD of the RR intervals series being evaluated by a bivariate autoregressive model. The features calculated from the signals were then used by the KNN classifier to categorize apnea events on a minute by minute basis. The same signals were used by Song et al. [134] to produce features that were considered to be subject-independent, implementing a learning and prediction procedure based on a discriminative HMM. OSA detection was performed using the Baum-Welch algorithm to estimate the Markov states.

A different approach was presented by Maier et al. [135], using an index based on cross-correlation, with a combination of multi-source information. It was verified that including the HR does not improve the detection accuracy. Three techniques were used by Ravelo-García et al. [136] to obtain features. In the first, the RR series was encoded into sequences of symbols, and permutation entropy

was used to distinguish different HRV patterns. The second was cepstrum analysis, obtaining cepstrum coefficients. PSD of EDR was the third, using a filter bank with equally spaced filters. These features were then used by the two tested classifiers, LR and the QDA. Both classifiers achieved similar performance, however, the QDA provided the best results. Cepstrum Coefficients, a filter bank with 34 filters (to analyze the very low, low, and high frequencies), and detrended fluctuation analysis were employed by Martín-González et al. [137] to feed the three tested classifiers: LDA, QDA, and LR. These features were obtained from the HRV. The best results were reported using the QDA.

Khandoker et al. [138] used 14 levels of Daubechies wavelets to decompose the RR and EDR signals. The result was used as the input to an SVM that classifies the OSA events. Features extracted from wavelet decomposition of HRV and EDR signals were used by Khandoker et al. [139] as inputs to the SVM classifier. The LDA classifier was also analyzed, providing similar results. HRV and EDR signals were used by Yildiz et al. [140] using 64 points of PSD (1 to 32 derived from HRV and 33 to 64 from EDR). Three SVM kernels were tested, specifically, Linear (L), Polynomial (P), and Radial Basis Function (RBF). The highest accuracy was produced by RBF using points 2, 3, 45, and 46 (selected by a hill climbing algorithm.).

4.2.4. Based on respiration

Oronasal airflow is one of the most direct indicators of breathing disorders and was used by Koley and Dey [141] to detect OSA. The features calculated from signals were then used by three binary SVM arranged in a one-against-all strategy to classify the data. The oronasal airflow, after being filtered and segmented, was also used by Koley and Dey [142], to extract time and frequency domains, as well as the nonlinear analysis features from each segment. The classification was performed in two steps using two binary SVM classifiers where the first was used to detect sleep disorders, and the second analyzed the segments marked as disorders and classified them as either apnea or hypopnea.

The Hilbert-Huang transform was applied to the nasal airway pressure signal by Caseiro et al. [143]. The LR was used as a classifier by Gutiérrez-Tobal et al. [144], using features from both airflow and respiratory rate variability (RRV) signals. Selvaraj and Narasimhan [145] analyzed the amplitude of the respiratory signal, the low pass filtered envelope of the respiratory signal (cut-off at 0.01 Hz), and the statistical dispersion of the envelope signal. Possible OSA events, determined by thresholds applied to the signals, were classified using two conditions based on the thresholds.

Daubechies wavelet was used by Minu and Amithab [146] to decompose the airflow signal, and statistical features were extracted to feed the classifying stage. Two classifiers were tested, the

AdaBoost and the Adaptive Neuro Fuzzy Inference System (ANFIS), which combines the advantages of both the neural and fuzzy classifiers. The ANFIS achieved the best performance. The same family of wavelets was employed by Avci and Akbas [147] to decompose the airflow signal. Three classifiers based on ensemble learning, namely random forest, AdaBoost, Random subspace were tested. The best results were produced by the first classifier. The airflow signal was also used by Ozdemir et al. [148] being segmented, and the energy on each segment was calculated using the TEO. Statistical features were extracted, and three classifiers were tested, the SVM, the KNN, and the linear regression, with the best performance being achieved by the SVM. The airflow signal with a sample dimensionality of 960 (30 s by 32 Hz) was fed to a deep learning classifier, specifically a CNN, by Haidar et al. [149]. The CNN architecture consists of three one-dimension convolutional layers, with each layer followed by a max-pooling layer, and, in the end, one fully connected layer with a soft-max activation function.

A different approach was presented by Thommandram et al. [150] where the respiratory effort signal (also called the RI signal) was obtained and was fed into a KNN to classify the data. Airflow and thoracic and abdominal respiratory movements were used by Maali and Al-Jumaily [151]. Wavelet decomposition was applied to the signals that were further segmented, and statistical measures were computed to produce the features that feed an SVM with a polynomial kernel. A selection of the best features subset and training data were performed interactively by a genetic algorithm.

4.2.5. Based on sound

The breathing process produces characteristic sounds that can be used to detect the presence of disorders. This principle was used by Rosenwein et al. [152]. Six features were calculated from the suspected period and were used as inputs in a binary-random forest classifier. The produced output was classified by an adaptive threshold produced for each subject's score distribution. Breathing sounds were also the base of the algorithm presented by Almazaydeh et al. [153] where Voice Activity Detection (VAD) was used to classify respiratory signals. The FFT segments of sounds were further analyzed by the VAD algorithm which compared them against the threshold, that was determined by comparing the signal value against noise. The output identifies whether the segment was a normal breath or a breathing cessation (silence). A second threshold was then used to classify the silence as either apnea or normal. Recorded respiratory sounds were used to extract spectrum features, using the FFT, by Praydas et al. [154]. The sounds were filtered and distinguished using the K-means clustering algorithm, and an SVM was employed to classify the data. Ng et al. [155] used linear predictive coding to model snore signals, and formant frequencies were extracted from the linear predictive coding spectrum. A threshold value was used to differentiate apneic and normal

snorers. A higher order statistics-based algorithm for snore sound analysis was presented by Karunajeewa et al. [156]. The pitch and total airway response waveforms were extracted from the sound signal, and features were extracted from these waveforms to feed an LR classifier.

A different approach was presented by Elisha et al. [157] where OSA was detected by analyzing particular speech signal properties. Seven GMM-based classifiers were used to classify the data. The GMM was also used by Pozo et al. [158] where the speech signals were parameterized using the Mel Frequency Cepstral Coefficient (MFCC). Benavides et al. [159] analyzed the subject's voice classifying OSA using an LDA feed with eight features. The use of tracheal sounds was analyzed by Penzel and Sabil [160], being verified that when recorded with an appropriate sensor, combining acoustic and suprasternal pressure sensors. It is possible to detect snoring, breathing and intrathoracic pressure variations. Specifically, OSA can affect the resonance produced by the upper airway, generating specific tracheal sounds. Kalkbrenner et al. [161] also analyzed the tracheal sounds, recorded by a microphone on the subject's neck.

4.2.6. Based on combined approaches

Usually, the pulse oximeter provides both the SpO₂ and HR signals, but it is also common for only SpO₂ to be considered in the OSA detection algorithm. A different approach was presented by Zamarrón et al. [162] where a combination of these two signals was used. The algorithm looks for peaks on the apnea-related frequency band of both signals to classify OSA. An algorithm that uses both EEG and oximetry was presented by Álvarez et al. [163].

A combination of oximetry and ECG was presented by Xie and Minn [73]. The features were fed to the three individual classifiers that collaborated in the final decision using a majority voting combination scheme (the chosen output class was the one on which the majority of the classifiers agree). The chosen classifiers were bagging with a Reduced-Error Pruning Tree (REPTree), AdaBoost with decision stump, and a KNN. Both SpO₂ and HR variations were analyzed by the algorithm developed by Poupard et al. [164], using the wavelet-aggregation to quantify these variations. A multi-modal approach that performs a feature-level fusion of ECG and SpO₂ signals was employed by Memis and Sert [165]. The produced signal was tested by three classifiers, specifically the Naïve Bayes, a KNN, and an SVM. The best results were achieved using the SVM with an RBF kernel.

An algorithm based on oximetry and ECG was also used by Ravelo-García et al. [10]. A combination of oxygen saturation and RR series features was used. An LDA was used to classify segments, on a minute by minute basis, as either normal or apnea. PPG-derived respiration and EDR signals were obtained by Madhav et al. [166] using EMD. OSA detection was performed by fitting an autoregressive model of the order 15 to each windowed signal.

A time delay ANN was used by Tian and Liu [167] to distinguish hypopnea and apnea from normal breathing, using the filtered airflow signal area, standard deviation, and SpO2 desaturation level as features. The combination of oximetry and tracheal sound signals was implemented by Yadollahi et al. [168]. The contribution of each calculated feature was weighted, added together, and compared with a threshold for classification.

To detect apnea Al-Angari and Sahakian [169] combined oximetry, thoracic, and abdominal respiratory effort signals with ECG signals. The data were classified using an SVM with a polynomial kernel. A combination of nasal airflow and PPG signals was used by Sommermeyer et al. [170]. The algorithm developed was capable of detecting arousals and produced the apnea hypopnea index (searching for a 90% drop in the airflow amplitude when compared to baseline for more than 10 s) and the respiratory disturbance index (searching for SpO2 drop greater than 4% or greater than 3% with arousal). The AHI, arousals index, the minimum value of SpO2 during rapid eye movement sleep, and percentage of time with SpO2 higher than 89% were used as features by Polat et al. [171] to test four classifiers, specifically, a C4.5 decision tree, a ANN, adaptive neuro-fuzzy inference system, and an artificial immune recognition system. The best results were produced by the C4.5 decision tree.

The use of both voice and facial features was presented by Espinoza-Cuadros et al.[172]. The MFCC was used for automatic speaker recognition with a GMM classifier. The sound and image features were used as inputs for a support vector regression classifier that estimates the AHI.

Table 3 : Evaluation of the Analyzed Algorithms [87].

Source sensor	Paper	Population	Data acquisition	EB Acc (%)	EB AUC (%)	EB Sen (%)	EB Spc (%)	Global classification (%)	SB AUC (%)	SB Sen (%)	SB Spc (%)	TW (s)
Oximetry	[93]	187 sub	Hospital	-	-	-	-	87	92	90	83	120
	[101]	83 sub	Hospital	-	-	-	-	86	91	91	79	-
	[95]	113 sub	Hospital	-	-	-	-	88	93	91	83	-
	[94]	129 sub	Hospital	-	-	-	-	93	95	97	79	120
	[103]	148 sub	Hospital	-	-	-	-	90	97	92	85	30
	[97]	144 sub	Hospital	-	-	-	-	87	-	92	77	-
	[99]	8 rec	Database*	93	-	88	100	-	-	-	-	-
	[104]	21 sub	Hospital	70	78	82	69	87	-	100	71	40
	[98]	115 sub	Hospital	-	-	-	-	94	96	92	96	60
	[105]	36 sub	Hospital	-	-	-	-	85	88	88	84	120
	[100]	127 sub	Hospital	-	-	-	-	90	-	94	70	-
	[106]	25 rec	Database+	85	-	60	92	-	-	-	-	60
	[96]	79 sub	Hospital	-	-	-	-	94	-	97	79	-
	[107]	8 rec	Database*	96	98	-	-	-	-	-	-	60

	[92]	92 sub	Hospital	91	-	83	89	97	99	98	95	60
	[102]	8 rec	Database*	98	-	97	99	-	-	-	-	60
ECG	[120]	147 sub	-	-	-	-	-	91	-	92	90	-
	[118]	35 rec	Database*	-	-	92	95	-	-	-	-	60
	[126]	35 rec	Database*	90	-	89	91	89	-	-	-	60
	[127]	35 rec	Database*	84	-	79	87	-	-	-	-	60
	[108]	5 rec	Database#	-	-	-	-	-	-	70	44	30
	[72]	25 rec	Database*	86	-	84	89	-	-	-	-	420
	[122]	25 rec	-	-	-	-	-	88	-	89	86	60
	[114]	35 rec	Database*	93	-	-	-	-	-	-	-	180
	[22]	16 sub	Hospital	-	-	-	-	95	-	-	-	60
	[138]	42 sub	-	-	-	-	-	93	-	-	-	-
	[139]	30 rec	Database*	93	-	90	100	100	-	-	-	60
	[125]	17 sub	Hospital	-	-	-	-	87	-	-	-	30
	[115]	35 rec	Database*	76	-	-	-	-	-	-	-	60
	[140]	60 rec	Database*	-	-	-	-	100	-	100	100	60
	[116]	35 rec	Database*	81	-	-	-	-	-	100	83	-
	[53]	32 rec	Database*	-	-	-	-	97	-	93	100	15
	[121]	35 rec	Database*	-	89	74	86	93	-	-	-	60
	[128]	35 rec	Database*	99	-	-	-	-	-	-	-	-
	[117]	97 sub	Hospital	-	-	-	-	-	94	89	83	30
	[135]	69 rec	Database*	-	-	-	-	-	93	87	88	60
	[129]	35 rec	Database*	85	-	86	83	-	-	-	-	60
	[109]	35 rec	Database*	94	-	95	93	94	-	-	-	60
	[119]	188 sub	Hospital	-	-	-	-	72	89	80	59	-
	[113]	35 rec	Database*	84	-	-	-	-	-	-	-	60
	[136]	35 rec	Database*	85	92	75	91	-	-	-	-	60
	[130]	70 rec	Database*	-	-	-	-	93	-	97	99	60
	[123]	35 rec	Database*	85	91	83	82	-	-	-	-	60
	[110]	35 rec	Database*	87	-	82	91	-	-	-	-	60
	[134]	35 rec	Database*	86	94	83	88	97	100	96	100	60
	[111]	35 rec	Database*	89	-	88	91	-	-	-	-	60
	[112]	9 rec	Database+	-	-	-	-	95	-	100	80	-
	[137]	35 rec	Database*	85	92	82	87	97	-	-	-	60
[124]	69 rec	Database*	-	-	-	-	98	-	98	100	-	
[131]	10 rec	Database*	-	-	-	-	98	-	-	-	-	
[132]	17 rec	Database*	100	-	-	-	-	-	-	-	60	
Respiration	[143]	41 sub	Hospital	-	-	-	-	-	88	81	95	300
	[151]	12 sub	-	89	-	87	90	-	-	-	-	30
	[144]	148 sub	Hospital	-	-	-	-	82	90	88	71	-
	[141]	14 rec	Database#	-	-	-	-	93	-	-	-	60
	[145]	100 rec	Database^	-	-	-	-	-	-	84	-	60
	[150]	70 rec	Database*	91	96	88	96	-	-	-	-	60
	[142]	4 sub	Hospital	82	-	86	81	96	-	-	-	-
	[147]	8 rec	Database*	99	-	-	-	-	-	-	-	60
	[148]	6 sub	-	88	-	91	77	-	-	-	-	40
	[146]	8 rec	Database*	99	-	-	-	-	-	-	-	-
	[149]	100 rec	Database~	75	-	-	-	-	-	-	-	30
Sou nd	[155]	40 sub	Hospital	-	-	88	82	-	-	-	-	-

	[158]	80 sub	Hospital	-	-	-	-	81	-	78	85	-
	[157]	87 sub	-	-	-	-	-	-	-	81	83	60
	[156]	41 sub	Hospital	-	-	-	-	90	97	89	92	-
	[153]	50 sub	-	97	-	-	-	-	-	-	-	-
	[159]	40 sub	Hospital	-	-	-	-	-	-	85	75	-
	[152]	186 sub	Hospital	-	-	-	-	86	-	-	-	-
	[154]	33 sub	Hospital	-	-	-	-	76	-	-	-	-
	[161]	10 sub	-	-	-	-	-	-	-	93	100	-
	[162]	120 sub	Hospital	-	-	-	-	89	-	94	82	-
	[167]	15 sub	-	-	-	91	86	-	-	-	-	-
	[171]	83 sub	-	-	-	-	-	95	97	92	97	-
	[163]	148 sub	Hospital	-	-	-	-	89	-	91	83	-
	[168]	66 sub	Hospital	-	-	-	-	-	95	83	91	-
	[164]	106 sub	Hospital	-	-	-	-	-	-	81	98	-
	[170]	66 sub	Hospital	-	-	-	-	-	96	90	86	-
	[73]	25 rec	Database+	82	-	84	81	-	-	-	-	60
	[169]	100 sub	Database^	82	-	70	91	95	-	92	98	60
	[172]	285 sub	Hospital	-	-	-	-	72	73	73	65	-
	[10]	70 sub	Hospital	87	92	73	92	100	-	-	-	300
	[166]	8 rec	Database*	-	-	-	-	-	-	97	-	15
	[165]	35 rec	Database*	-	-	-	-	97	-	-	-	-

* PhysioNet apnea-ECG Database

MIT-BIH polysomnography Database

+ University college of Dublin sleep apnea Database

^ Sleep Heart Health Study Database

~ Scaling Up Scientific Discovery in Sleep Database

4.2.7. Summary

Of the algorithms analyzed, the highest EB accuracy was reported by Pathinarupothi et al. [132] with algorithms based on ECG analysis by a deep network.

The maximum global classification was achieved by Ravelo-García et al. [10], Khandoker et al. [139], and Yildiz et al. [140]. All are based on ECG, however, the first also used oximetry analysis.

Mostafa et al. [102] reported the maximum EB sensitivity using an oximetry analysis and the maximum SB sensitivity was obtained by Lázaro et al. [104], using oximetry analysis, Kesper et al. [116] and Smruthy and Suchetha [112], with an ECG signal analysis.

The highest EB specificity was reported by Almazaydeh et al. [99] and Khandoker et al. [139]. The first algorithm was based on an oximetry analysis and the second on an ECG. For the SB approach, the maximum values were presented by Almazaydeh et al. [53], Song et al. [134], Chen et al. [124], who also use the ECG signal analysis, and by Kalkbrenner et al. [161] with sound analysis. Yildiz et al. [140] reported the best SB results with 100% global classification, sensitivity, and specificity using the ECG analysis.

By analyzing the algorithms based on a single source sensor it is possible to determine that ECG signals provided the highest global classification by studied population ratio. However, the majority of ECG algorithms were tested in public databases with signals that have a low level of noise contamination, which could contribute to improving the diagnostic capability of the algorithm.

Respiration based algorithms have the third highest value of this ratio and algorithms based on sound have the lowest value. Although OSA is directly related to respiration and oximetry, the algorithms based on the first sensor did not achieve as good results as those produced by the second. This is likely caused by the higher noise involved in the respiration signals. However, the algorithms based on sound are even more susceptible to noise (specifically cardiac sounds and environmental noise), and this could be the reason why they have the lowest results regarding the global classification by studied population ratio.

The combination of source sensors did not contribute to a relevant improvement of the classification capability. This could be an indicator that one of the sensors is dominating the analysis, although it could be desirable that more effort be devoted to studying the effects of the underlying physiological processes that could be quantified with different sensors. In any case, algorithms based on a single source sensor could be preferable due to their simplicity in hardware implementation. The majority of the works detect OSA employing machine learning algorithms and that an SVM, a KNN, and a ANN were the most frequently used classifiers. The domination of supervised learning could be due to the fact that OSA is a disorder with a well-established pattern that facilitates the training of the algorithms. Some methods provide an optimal performance but with a high degree of complexity which is particularly important if a hardware device is to be designed.

One key aspect has to do with the goal of obtaining a useful method with a good performance-complexity ratio. With this objective in mind, a method with a reduced number of sensors and complexity is typically of special interest.

From the overall analysis of this review it can be recognized as future directions for the research include: producing more robust OSA diagnosis tools by implementation of the presented algorithm in efficient hardware, producing more research with deep learning classifiers, capable of self-learning the features, and validating the achieved results of the algorithms by independent research groups using publicly available databases, so that the results can be reproduced. This is of a special interest in home diagnostic devices since they could be used as a first OSA diagnosis tool, leading to a considerable reduction in the diagnostics cost and waiting time for access to a sleep study. However, these devices are more susceptible to data errors caused by factors not controlled in the home of the subject. Therefore, an adaptation of the proposed algorithms to a real world environment in efficient hardware is the major challenge identified. The main gaps in the current state of the art are related to

the use of algorithms capable of self-learning the features.

4.3. Detecting Sleep Apnea using Deep Learning

4.3.1. Introduction

Although previous reviews have been performed in the field of sleep apnea detection, such as analyzing devices for home detection of obstructive sleep apnea (OSA) [21], classification methods based on respiratory and oximetry signals [26], different detection approaches [87], and detection and treatment methods [173], no review was previously performed to assess the current development of methods for detecting sleep apnea using a deep learning. In addition to that, recent publications show an accuracy improvement using deep network over shallow networks. Therefore, the main focus of this review is the analysis of such works, assessing the performance of the presented methods to provide in-depth knowledge of the applicability of deep learning in the detection of sleep apnea.

The review was performed considering the timeline between 2008 and 2018, based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) style. A systematic search was conducted on Web of Science, IEEE explorer, PubMed, ScienceDirect, and arXiv. The selected search keywords were (“sleep apnea” OR “sleep apnoea”), due to the different spellings of the word apnea, along with the AND operation and: “unsupervised feature learning”; “semi-supervised learning”; “deep belief net”; “CNN”; “convolution neural network”; “autoencoder”; “deep learning”; “recurrent neural network”; “RNN”; “long short-term memory”; and “LSTM”. A total of 255 articles were found, specifically: 93 on the Web of Science; 77 on PubMed; 51 on IEEE Xplorer; 25 on ScienceDirect; and 9 on arXiv. A total of 116 duplicate articles were removed from the list.

The title and abstract of each article were analyzed, and 19 were selected as relevant to the topic. The inclusion criteria analyzed the keywords apnea and deep network. The main exclusion criterion was non-English articles. Works that were not explicitly developed for sleep apnea detection, but could be adapted for that purpose, were also excluded. Two papers were added due to their relevance though they did not appear in the search, and two were removed despite their appearance in the search. A relevant article, found by analyzing the references of the already selected articles, was included despite not appearing in the search engines. Therefore, a total of 21 articles were selected for this review. The flow chart of the search strategy is presented in Figure 5, with n indicating the number of articles.

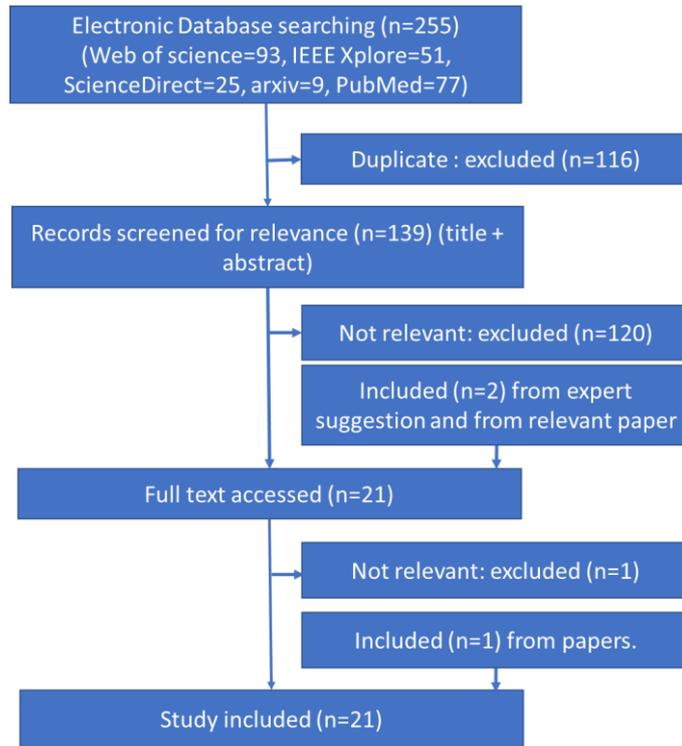


Figure 5 : Flow chart of the process for article selection using PRISMA reporting style.

The last decade was chosen for this work since most of the articles (20 articles) were published in 2017 (5 articles) and 2018 (15 articles). Only one was published in 2008. Therefore, within one year, the number of published articles was three times higher, highlighting the importance of this topic and the need for a review to consolidate the developed approaches and to point out new research lines.

There are deep networks with the final structure resembling classical neural networks with more than one hidden layer. However, sometimes, these classifiers' training strategy and layer construction are different than from classical one. These types of classifiers are mentioned in this work as Deep Vanilla Neural Network (DVNN). In addition to that, the layers between the input and output layers are named hidden layers. A typical example of a deep learning model is the feedforward deep network, or multilayer perceptron [78]. A feedforward neural network with more than one hidden layer can be considered as a deep network. In this work, a classical neural network with multiple hidden layers is indicated as a Multiple Hidden Layers Neural Network (MHLNN).

4.3.2. Automatic feature learning using DVNN

A hidden Markov model with autoencoder was used by Li et al. [174] using automatic feature learning. The implementation used 100 points of the RR series, selected by the Pan-Tompkins algorithm [175] which were passed through a median filter [130] as an input. An SAE was used for

classification and the data were divided into 50% training set (35 subjects) and test set (35 subjects). The training process was based on the mixture of unsupervised learning with fine-tuning at the end. First, a single hidden layer SAE unsupervised training was done for primary feature extraction then it was fine-tuned by using a logistic regression layer. After that, these extracted features were used as the corresponding observation vector (O_t) of a Markov model [134] which belong to two Markov states $S = \{S_N, S_A\}$ where S_N is the normal and S_A the apnea state. Then a soft decision fusion of two separate classifiers (ANN, SVM) was done based on the confidence score maximization strategy that considered the classifier quality information [129]. Two deep network structures were analyzed and the highest accuracy (83.8%) was achieved using 100 neurons on the first Hidden Layer (HL) and the second HL with 10 neurons.

A autoencoder with two HL was analyzed by Mostafa et al. [106] using the SpO2 signal resampled at 1Hz with tenfold cross validation. It was verified that the selected number of neurons has a significant impact on the results. Therefore, a grid search approach was employed, varying the number of neurons from 30 to 180, with intervals of 30 neurons, in two hidden layer DBN. The optimum number of hidden neurons (90 in the first HL and 60 in the second HL) was found by maximizing the CO. The achieved accuracy for the UCD [62] and AED [61] databases were 85.26% and 97.64%, respectively.

4.3.3. Human crafted feature learning using DVNN

Breathing sounds during sleep were analyzed by Kim et al. [176] using an MHLNN with two hidden layers (the first with 50 and the second with 25 nodes) and two dropout layers with 4 classes (normal, mild, moderate and severe). Using tenfold cross validation, windows of 2.5, 5, 7.5, and 10 seconds were tested. The 5 second window achieved the best performance. A patient wise classification was performed, with an average global accuracy of around 75%, by the MHLNN, which is slightly less than the performance attained by both an SVM and a logistics classifier.

Lakhan et al. [177] produced 17 features from an AF signal and a fully-connected neural network with layers size of 1024, 512, 256, 128, 64, 32, 16, 8, and 4 hidden nodes with a softmax function at the end. An average Acc of 83.46%, 85.39%, and 92.69 % were achieved using tenfold cross validation for three cutoff points of the AHI (5, 16, and 30) respectively.

Falco et al. [178] used Evolutionary Algorithms (EAs) with a data subsampling technique (training set consisting of 60% and test set consisting of 40% of the data) to reduce the simulation time to find the best hyperparameter of the MHLNN. The HRV was calculated from the twelve typical parameters (features) of HRV related to the frequency domain, the time domain, and the non-linear domain, which were extracted from the 1-minute segment. It was verified that 2 HLs with 23 and 24 hidden

units using ReLU as an activation function produced the highest accuracy (68.37%).

4.3.4. Convolutional Neural Network (CNN)

A CNN was mainly developed to classify images. However, some authors [179][180][181][182][149] adapted the concept by employing a one dimensional CNN (CNN1D) network for signal classification. Haider et al. [183] used 3 one dimensional signals hence producing a CNN1D with three channel inputs. Other authors [184] [185] converted the 1 dimensional signal to a 2 dimensional input to employ the two dimensional CNN (CNN2D) network directly. An analysis of both CNN1D and CNN2D was performed by McCloskey et al. [184] to assess their performance.

The signal from a single-lead ECG was analyzed by Urtnasan et al. [180] using a CNN1D with a hold-out method (the training had 63 subjects, while the test had 19 subjects). The signal was segmented into 10 second intervals, unlike the 1 minute segment used by other authors [174][106], each having 2000 sample points. The network was composed of different sizes of convolution, activation, and pooling layers, followed by dropout. The input signal was normalized by batch normalization and a REctified Linear Unit (ReLU) was employed as an activation function. Following the batch normalization and the ReLU layer, a set of convolution and pooling layers was repeated. In the end, a dropout layer followed by a fully connected layer, and a softmax activation function was used for binary classification. In between the final layer and the batch normalization layer, the set of layers was repeated. Seven CNN models with a number of layers varying from 3 to 9, with a 1-layer increment, were studied. The highest accuracy (96%) was achieved using a CNN with six layers implementing the F_1 score as a defining parameter.

Urtnasan et al. [179] also used the CNN1D for multiclass classification (normal, apnea and hyperpnea). The input of the network was 10 seconds long and contained 2000 samples. A hold-out method was used to test the model similar to what was done in a previous work [180]. The network architecture included Batch Normalization (batchnorm), convolution (conv1D), Maximum Pooling (maxpool), dropout and fully connected layers. The first layer was batchnorm followed by conv1D (20@[50 × 1]) and maxpool ([2 × 1]). Afterward, a set of variously sized conv1D, maxpool, and dropout ($p = 0.25$) was repeated and stacked, one after another, up until the final softmax layer. The 6-layer CNN achieved 90.8% mean accuracy among the classes.

Dey et al. [181] also employed a CNN1D to analyze one minute segments of a single lead ECG signal, each with 6000 samples. Unlike other implementations, it uses only convolution and fully connected layers. The pooling was performed using convolutional pooling. The authors tested the model with a different training: test dataset, from 50:50 to 20:80 with having 50:50 having the best average accuracy of 98.91%.

Binary classification (either apnea or normal) based on the nasal airflow analysis was performed by Haidar et al. [149] with a CNN1D classifier and a balanced dataset. The network consisted of three convolutional layers, each having 30 filters with $[5 \times 1]$ kernel size, 5 strides, and each followed by a max pooling layer with $[2 \times 1]$, and one fully connected layer with a soft-max activation function. It had two output nodes for each class (normal or abnormal). The activation function ReLU was chosen because of its best accuracy and fastest training time [149] by evaluating other activation functions. The model achieved an average accuracy of 75%.

The signal from a single-channel nasal pressure was analyzed with a CNN by Choi et al. [182] to detect 1 second apnea events. The database was divided into training (50 subjects), validation (25 subjects), and testing (104 subjects). It was tested using the class balance hold-out method. Overlapping windows with length ranging from 5 to 10 seconds were tested and multiple configurations of the network were analyzed, changing the number of convolution layers (1 to 3), the number of convolution filters (5, 15, 30), the kernel sizes for convolutions (4, 8, 16, 32) the strides for convolutions (1, 2, 4, 8, 16) and the strides for pooling (1, 2). It was verified that a 10 second window with three convolution layers, two maxpooling layers, and two fully connected layers achieved the highest accuracy (96.6%).

A CNN1D with three input signals was tested by Haider et al. [183], analyzing the nasal flow, the abdominal and thoracic plethysmography signals using hold-out methods with a 75% training and a 25% test dataset. Two back to back convolution layers with a subsampling layer (conv-conv-maxpooling) in a three cascading state with a final layer of a fully connected layer were studied. It was verified that the performance of the model with three channels was better than any single or double channels model, with an average accuracy of 83.5%.

McCloskey et al. [184] have also performed a multiclass classification (normal, apnea and hyperpnea), by analyzing the nasal airflow signal, normalized with 30 second epochs, with an input size of 960 samples. Three sets of conv-conv-maxpooling layers, followed by 1 fully connected layer, made the CNN1D. The first convolution layer in the set had 32 filters with a kernel size of $[3 \times 1]$, a stride of 3 and a ReLU as an activation function. The second convolution layer also had ReLU as an activation function with a kernel size of $[2 \times 1]$, and a stride of 2. The maxpooling layer kernel was $[2 \times 1]$ with a stride of 2. The output had three nodes representing 3 classes. The CNN1D achieved an average accuracy of 77.6%.

The spectrogram of the nasal airflow signal, calculated by using a continuous wavelet transform (CWT) with the analytical Morlet wavelet, was fed to a CNN2D by S. McCloskey et al. [184]. The network had two convolutional layers with ReLU activation layers afterward and one 2-D max pooling layer followed by a fully connected layer and a softmax layer with three output, each nodes

representing the three classes (normal, apnea and hyperpnea). The model achieved an average accuracy of 79.8%.

Chen et al. [185] used a CNN2D with leave one out cross validation, which has three input signals (blood oxygen saturation, oronasal airflow, and ribcage and abdomen movements) with one second annotation. A two-dimensional matrix with zero padding was created as an input to the network that consists of two convolution layers, two subsampling layers, and a fully connected layer connected to the output layer with three nodes. The multiclass classification overall accuracy was 79.61%.

4.3.5. Recurrent Neural Network (RNN)

SpO2 and IHR signals were tested by Pathinarupothi et al. [107] as an input to LSTM. The dataset was divided into 50% for training, 40% for testing, and 10% for validation. With only the SpO2 signal, the single layer, 32-memory block, LSTM, and the 32-memory block stacked LSTM achieved an AUC of 0.98. With only the IHR signal, the 32-memory block stacked LSTM achieved a 0.99 AUC for a severity detection (apnea or non-apnea). Combining both signals provided a 0.99 AUC in both single layers and stacked LSTM.

The same authors [107] also used IHR for apnea and arrhythmia classifications, with higher accuracy and F1 score of 1 [132] using a fivefold cross-validation technique. Both a single-layer and stacked layers LSTM (2 layers) were tested, and it was verified that better results were attained by the two-layer stacked LSTM. However, the single layer and 32 memory cells worked better than 2-layer stacked LSTM-RNN model.

To capture temporal information and accurately model the data Steenkiste et al.[186] used an LSTM [187] neural network. Balanced bootstrapping was employed to balance the dataset, where the entire minority class was used each time with an equal size of the majority class. These balanced datasets were used for each LSTM model, which had one LSTM layer with 3 dropout layers and with an output layer at the end. In the end, the probability of the LSTM models was aggregated into a single probability prediction per epoch by averaging. An averaged probability greater or equal to 50% was used to determine the presence of apnea. The authors also used the same LSTM network structure with a human-engineered time-domain and the frequency-domain features instead of raw respiratory signals [186]. Because it used features with LSTM, it is denoted as F-LSTM. A performance valuation was also done with three respiratory signals (abdores, thorres, and EDR) with non-temporal models and with temporal models. Both temporal models (F-LSTM, LSTM) did better than the non-temporal models (ANN, LR, RF). Among the temporal models, LSTM did better than F-LSTM in all three signals (Table 5). Although authors in the original paper detect apnea severity, in this review it was not included due to the presentation of severity being different when compared with other work

(for severity please check the Fig. 7, Fig. 8 and Fig. 9 of original work [186]; also, it is quite difficult to calculate the exact values from the figures).

A three-layered F-LSTM was used by Novak et al. [188] to calculate apnea events using HRV with features as the input. The hidden layers of the network contained five blocks, each consisting of seven memory cells, thus achieving an average accuracy of 82.1%.

Cheng et al. [131] employed a four layered LSTM to detect OSA using 20 subjects for training and 10 subjects for test and the RR-ECG signal. The network consisted of a recurrent layer and a data normalization layer, repeated four times, followed by a softmax layer, achieving an average accuracy of 97.80%.

Urtnasan et al. [189] used the normalized ECG signal with 74 subjects for training and 18 subjects for testing and six RNN layers were used to form an LSTM and a GRU. The F_w score of the LSTM and GRU were, 98.0% and 99.0%, respectively.

4.3.6. Combination of multiple deep networks

A combined deep recurrent and convolutional neural networks (RCNN) was evaluated by Biswal et al. [190], using airflow, SaO₂, chest and abdomen, and belts signals to determine the AHI. A hold-out method with 90% of data for training and 10% of data for testing was used. Both waveform representation and spectrogram representation were employed as input signals for a CNN and a combination of CNN and RNN (RCNN). The RCNN with a spectrogram representation achieved the highest accuracy (88.2% in MGH and 80.2% in SHHS).

A different approach was presented by Banluesombatkul et al. [191], achieving 79.45% of global accuracy (detecting extremely severe OSA subjects from normal subjects) by combining CNN1D, LSTMs and MHLNN (in original work it was defined as a deep neural network (DNN)) to detect sleep apnea from a 15 second window using a tenfold cross validation method. This structure was used for automatic extraction of the features using the CNN1D with 256, 128, and 64 units, where each convolution layer was followed by a batch normalization layer and ReLU was used as an activation function. Then a LSTM, with 128, 128, and 64 units, respectively, and a recurrent dropout of 0.4, was then stacked to extract temporal information. At the end of the network, an MHLNN (with fully connected layers) was stacked with layers of size 128, 64, 32, 16, 8, and 4 hidden nodes followed by a softmax function for the classification.

Table 4 : Summary of the database information: The database, year of publication, number of subjects, used signals, window size and type of classifiers(A=apnea, H= hypopnea, N=Normal, S= Severity, O=obstructive, G=Global or OSA Severity) used by selected papers (According to year).

Paper	Year	Database	Recordings	Sensors/ Signals	Window size (seconds)	Classification Type
[188]	2008	AED [61]	70	[HRV-ECG]	60	A/N
[149]	2017	MESA	100	[Nasal airflow]	30	OA/N
[106]	2017	AED [61]	8	[SpO2]	60	OA/N
		UCD[62]	25	[SpO2]	60	A/N
[132]	2017	AED [61]	35	[IHR-ECG]	60	G
[107]	2017	AED [61]	35	[IHR-ECG]	60	OA/N,G
		AED [61]	8	[SpO2]	60	OA/N,G
[131]	2017	AED [61]	35	[RR-ECG]	-	OA/N
[181]	2018	AED [61]	35	[ECG]	60	OA/N
[184]	2018	MESA[91]	1,507	[Nasal airflow]	30	A/H/N
[176]	2018	SNUBH [176]	120	[Breathing sounds]	5	G
[180]	2018	SCSMC82[180]	82	[ECG]	10	OA/N
[185]	2018	UCD[62]	23	[SpO2, oronasal airflow, and ribcage and abdomen movements]	1	OA/H/N
[183]	2018	MESA[91]	1,507	[Nasal airflow, Abdominal and thoracic plethysmography]	30	OA/H/N
[178]	2018	AED [61]	35	[HRV ECG]	60	OA/N
[182]	2018	SNUH [182],	179	[Nasal pressure]	10	AH/N, G
		MESA[91]	50	[Nasal pressure]	10	AH/N, G
[191]	2018	MrOS (Visit 1)[192]	545	[ECG]	15	G
[177]	2018	MrOS (Visit 2)[192]	520	[Airflow]	-	G
[190]	2018	MGH	10 000	[Airflow, respiration (chest and abdomen belts), SpO2]	1	G
		SHHS [193]	5804	[Airflow, respiration (chest and abdomen belts), SpO2]	1	G
[186]	2018	SHHS-1[89]	2100	[Respiratory signals (chest and abdomen belts), EDR]	30	A/N
[179]	2018	SCSMC86[179]	86	[ECG]	10	OA/H/N
[189]	2018	SCSMC92	92	[ECG]	10	A/H/N, AH/N
[174]	2018	AED [61]	70	[RR – ECG]	60	OA/H/N,G

Table 5 :Performance of the different works.

Paper	Classifier Type	Sen/Recall (%)	Spc (%)	Acc (%)	Others
[177]	MHLNN (AHI 5)	80.47(G)	86.35(G)	83.46(G)	-
	MHLNN (AHI 15)	85.56(G)	86.96(G)	85.39(G)	-
	MHLNN (AHI 30)	93.06(G)	90.23(G)	92.69(G)	-
[178]	MHLNN	-	-	68.37	-
[176]	MHLNN	-	-	75(G)	-
[174]	SAE	88.9	88.4	83.8	AUC 0.86.9
	SAE	100(G)	100(G)	100(G)	-
[106]	DAE(UCD)	60.36	91.71	85.26	CO 79.1
	DAE(AED)	78.75	95.89	97.64	-
[179]*	CNN1D	87	87	90.8	PPV 87%, $F1_w$
[180]*	CNN1D	96	96	96	$F1_w$ 0.96
[181]	CNN1D	97.82	99.20	98.91	PPV 99.06%, NPV 98.14%
[182]	CNN1D	81.1	98.5	96.6	PPV 87%, NPV 97.7%
	CNN1D (AHI 5)	100(G)	84.6(G)	96.2(G)	PPV 95.1%, NPV 100%, F_1 0.98(G)
	CNN1D (AHI 15)	98.1(G)	86.5(G)	92.3(G)	PPV 87.9%, NPV 97.8%, F_1 0.93(G)
	CNN1D (AHI 30)	96.2(G)	96.2(G)	96.2(G)	PPV 89.3%, NPV 98.7%, F_1 0.93(G)
[149]	CNN1D	74.70	-	74.70	PPV 74.50%
[183]	CNN1D-3ch	83.4	-	83.5	PPV 83.4%, F_1 83.4
[184]	CNN1D	77.6	-	77.6	PPV 77.4%, F_1 77.5
	CNN2D	79.7	-	79.8	PPV 79.8%, F_1 79.7

[185]	CNN2D		-	79.6	-
[107]	LSTM(SpO2)	92.9	-	95.5	AUC 0.98, PPV 99.2%
	LSTM(IHR)	99.4	-	89.0	AUC 0.99%, PPV 82.4%
	LSTM(SpO2+IHR)	84.7	-	92.1	AUC 0.99%, PPV 99.5%
	LSTM(IHR)	99.4(G)			
[132]	LSTM(IHR)	-	-	100(G)	F_1 1(G)
[131]	LSTM	-	-	97.08	-
[188]	fLSTM	85.5	80.1	82.1	-
[186]	fLSTM(abdores)	57.9	73.9	71.1	AUC 71.5, PPV 33.0%
	LSTM (abdores)	62.3	80.3	77.2	AUC 77.5, PPV 39.9%
	fLSTM (thorres)	62.9	77.2	74.7	AUC 76.9, PPV 36.8%
	LSTM(thorres)	67.8	76.5	75	AUC 79.7, PPV 37.7%
	fLSTM(EDR)	48.8	60.8	58.7	AUC 57.6, PPV 21.1%
	LSTM(EDR)	52.1	61.8	60.1	AUC 58.8, PPV 22.1%
[189]	LSTM	98	98	98.5	F_{1w} 98.0
	GRU	99	99	99.0	F_{1w} 99.0
[190]	RCNN(MGH)	-	-	88.2(G)	-
[191]	CNN1D-LSTM-MHLNN	77.60(G)	80.10(G)	79.45(G)	F_1 79.09(G)

* The authors used alternative definition of true positive (detection of normal events) compare with the definition provided by Baratloo et al. [194]. So, in this table for binary classifier comparing with other authors their Sen could be treated as Spc and vice versa. If nothing is indicated in the paper, then an assumption was made that the authors did use the definition provided in Baratloo et al. [194].

4.3.7. Summary

The systematic literature review synthesized and summarized the published deep classification methods for sleep apnea detection. From the selected 21 studies, the main findings are as follows:

It was verified that a significant number of papers were published in the last two years, indicating a strong interest in the research community with regards to this topic. The comparison between the deep networks and parameter choice of the deep network is still part of ongoing research and a very relevant topic. In addition to that, which sensor or signal is most suitable for apnea detection is still being questioned.

The ECG sensor based signal was the most commonly used, which could be justified as indicated by Mendonça et al. [87], that for a single source sensor, ECG signals provided the highest global classification. However, sleep apnea is directly related to respiration. Thus, this higher accuracy with ECG signals could happen due to the use of public datasets that are less affected by noise [87]. For the works based on a single sensor, Pathinarupothi et al. [107] achieved the best results using the SpO2 signal comparing IHR calculated from ECG. Thus, the universality of better ECG signals performance is not true. However, a direct performance comparison between the works is not fair for this review because of the use of different classifiers and different databases.

It was verified that using more than one signal from the sensors improves the predictive capability of the models, as reported by Haidar et al. [183]. This is understandable because the gold standard of sleep apnea tests uses several signals. However, the main research goal of most of the works is to achieve a respectable result using fewer sensors.

Most works with deep networks outperformed the shallow networks except for T. Kim et al.'s [176] work. In their work, the deep network performs slightly worse than the shallow network. However, they use a deep network with human engineered features. Similar kinds of work where authors [177] used features with the deep network, MHLNN, outperformed classical machine learning techniques. So for T. Kim et al.'s [176] work, it may be a feature selection process or hyperparameter choice of the deep network.

The CNN was the most commonly used classifier and approaches based on both CNN1D and CNN2D were presented. However, it was not possible to indicate which are the most suitable classifiers since the testing conditions were different in all the works. Even so, McCloskey et al. [184] compared both and verified that 2-D spectrogram images of the nasal airflow performed better than raw 1-D signal with a CNN. A similar conclusion was attained by Biswal et al. [190] where the RCNN with spectrogram representation achieved a higher accuracy. Analyzing the three works from Urtnasan et al. using CNN1D [179] [180] RNN [189] where they have collected the data from the same hospital it was possible to verify that the RNN outperforms the CNN. However, more research is needed to reach a definitive conclusion. The same type of conclusion can be achieved by analyzing the works that have employed the LSTM and the GRU.

Hyperparameters optimization is also a problem in deep network implementation. Some works [106] [179][180] [106] have verified that just blindly increasing the number of layers or neurons in the hidden layers did not increase the performance. Most of the works chose their hyperparameters with an educated guess or by trial and error methods. Others used a predefined search space and attempted to find the best solution [179][180][106]. A possible alternative solution was presented by Falco et al. [178], where an evolutionary algorithm was used to choose the hyperparameters.

For performance purposes, the dominating methodologies were hold-out and cross-validation methods. Hold-out does not test all the datasets. It is understandable that due to the long simulation time and the assumption of having the same effect due to a significant number of examples, many authors do not choose the cross-validation method when using deep learning. On the other hand, cross-validation of event-based apnea detection techniques is frequently used without ensuring subject independency (or this information was not mentioned specifically in the paper), which is essential to assess the generalization capability of the model. Some authors used dataset balancing methods or specific parameters to solve the class imbalance problem. It was also not clear for some of the works presented if the test dataset was balanced or not, which should not be done since it will change the natural distribution of data and, consequently, derail the generalization of the model. To have an unbiased test, a form of cross-validation with subject independence could be suggested as an optimal choice for future research.

There are two main classification strategies, the event-by-event or epoch by epoch approach and

global classification. Most of the works concentrated on event-by-event classification and eight works used global classification considering the OSA severity classification. However, it is possible to do a global classification from event-by-event classification methods by using a threshold approach as indicated by Pathinarupothi et al. [107]. This observation is considered extremely relevant for further research since it will allow the methods to be used for clinical diagnosis.

4.4. Summary and Choice of Work Done in This Thesis

In this chapter two literature reviews were done to understand the trend and knowledge gap of sleep apnea. The former one (Section 4.2) focused on different algorithms and methods, which use signals from multiple source sensors but have not been implemented in hardware, to detect OSA. Papers were selected between 2003 and 2017 and a total of 84 original research articles were analyzed after the inclusion and exclusion criteria. The highest EB accuracy was based on an ECG analysis by a deep network where in the case of global accuracy the same trend also followed with some solution using SpO₂ or combination of SpO₂ and ECG heart rate. The reported maximum EB sensitivity was achieved with both oximetry analysis and ECG analysis. The single source works with ECG signal were tested in public databases provided the highest global classification. The public databases with signals might be successful due to a low level of noise contamination. Though OSA is directly related to respiration, oximetry achieved better results than respiration. The probable cause is the higher noise involved with the respiration signals. A multiple source sensors combination did not improve the system relevance which could be an indicator of a dominating sensor. The majority of the works detect OSA by employing supervised machine learning algorithms.

The second literature review (Section 4.3) focused on the deep learning-based classification of OSA performed between 2008 and 2018. A high concentration of recent publication shows a strong interest in deep learning. Like previous reviews (Section 4.2) ECG signals provided the highest global classification for a single source sensor. However, the main research goal of most of the works is to achieve a respectable result using fewer sensors. More than one signal from the sensors improves the predictive capability of some models. Most works with deep networks outperformed the shallow networks. Among deep classifiers, the CNN was the most commonly used classifier.

One key aspect of a relevant method is one with a good performance-complexity ratio. Therefore, a method with a reduced number of sensors and complexity is typically of special interest. Both literature reviews showed an optimistic result using oximetry or SpO₂. Comparing to complexity, price and ease of use oximetry performed better than other sensors. However, the highest accuracy was achieved by ECG signal-based methods or heart rates derived from an ECG signal. That is why a SpO₂ signal will be used as primary signal where the different methods will be tested (Section 5.2, Section 5.3, Section 5.4, Section 6.2 and , Section 6.3). After reaching the decision of the best

methods, an ECG derived heart rate will also be tested to understand the effectiveness of the ECG signal (Section 5.5, Section 6.4). When it comes to the classification feature creation and selection are problems for classical shallow networks. Therefore, different feature creation and selection methods with different classifiers will be tested in this work (Section 5.2, Section 5.3, Section 5.4). From the research published before one could notice the real promise of deep classifiers in the field of apnea classification (Section 4.3). That is why this work will test a deep learning for apnea detection side by side with classic classification methods (Chapter 6). One of the problems of deep learning classifier is the problem of systematic structure hyperparameter searching, which will be also investigated (Section 6.2, and Section 6.3). Along with these, effects of more than one signal will be tested. The combined effects of signals will be tested using the SpO2 and heart rate derived from an ECG (Section 5.5 and Section 6.4). For clinical diagnosis, global accuracy is important. Therefore, a threshold based global accuracy will be implemented on successful solutions (Section 5.5.5 and Section 6.4.5). At the end a general implementation will be done to accommodate any solutions tested in this thesis (Chapter 7).

Chapter 5

Handcrafted Feature Based Method

This chapter develops and analyzes different handcrafted features as well as classifiers. A combined approach of features and classifier combination is also proposed in this chapter.

5.1. Introduction

This chapter analyzes the performance of different handcrafted features and feature based classifiers. For classifying apnea events a vast pool of suitable features was developed by different researchers. The learning process with all the available features could have a negative effect on the performance generalization, particularly when irrelevant or redundant features are present. One of the solutions could be the combination of the best features of all the previous works. However, combining two or more independent best features cannot guarantee a better feature set [11]. Therefore, it is important to find a subset of dominant and optimum features to achieve the best classification. To address these issues, different feature selection techniques and classifications were indicated and tested besides proposing new features. The tested features, classification selection techniques and a proposed technique were first tested on SpO₂ signals (Section 5.2, Section 5.3, and Section 5.4). After choosing the best methods, the HR and the combination of SpO₂ and HR (SpO₂+HR) were tested (Section 5.5). Finally, the global classification performance was also checked (Section 0).

Section 5.2 was published in “Neural Computing and Applications” [195] and compares different features with different methods and classifiers. Section 5.3 was published in the “2017 XXVI International Conference on Information, Communication and Automation Technologies (ICAT)” [102] and presents the same features used in Section 5.2 with a different feature selection technique. The performance of the classification combination technique was also checked in Section 5.4, published in “Neural Computing and Applications” [196].

5.2. Comparison of SFS and mRMR for Oximetry Feature Selection

5.2.1. Introduction

A popular signal is SpO₂, measured by pulse oximetry which has some added advantages over other sensors. It is more portable [8] and can be used in hypopneas, where the drop in oxygen saturation is caused by a reduction in the airflow due to apnea event (Figure 6). Different features can be created using SpO₂. However, a different feature set provides a difference in performance in the classification process. A high-dimensional feature space creates problems for the classifier called the curse of dimensionality. Fortunately, a feature selection process can solve this problem. Feature selection is the process of selecting a subset of prominent features for the use in the classification model construction. Feature selection techniques can simplify the models to make them easier to interpret by researchers or users, shorten training times, and enhance generalization by reducing overfitting problems. In this work, minimum Redundancy Maximum Relevance (mRMR) and

Sequential Forward Search (SFS) feature selection algorithms are used to find proper features and apnea classification. These two feature selection processes are chosen because mRMR and SFS have different ways of choosing features: one uses mutual information without considering the classifiers (mRMR); the other one uses the wrapper methods (SFS) which consider the classifiers.

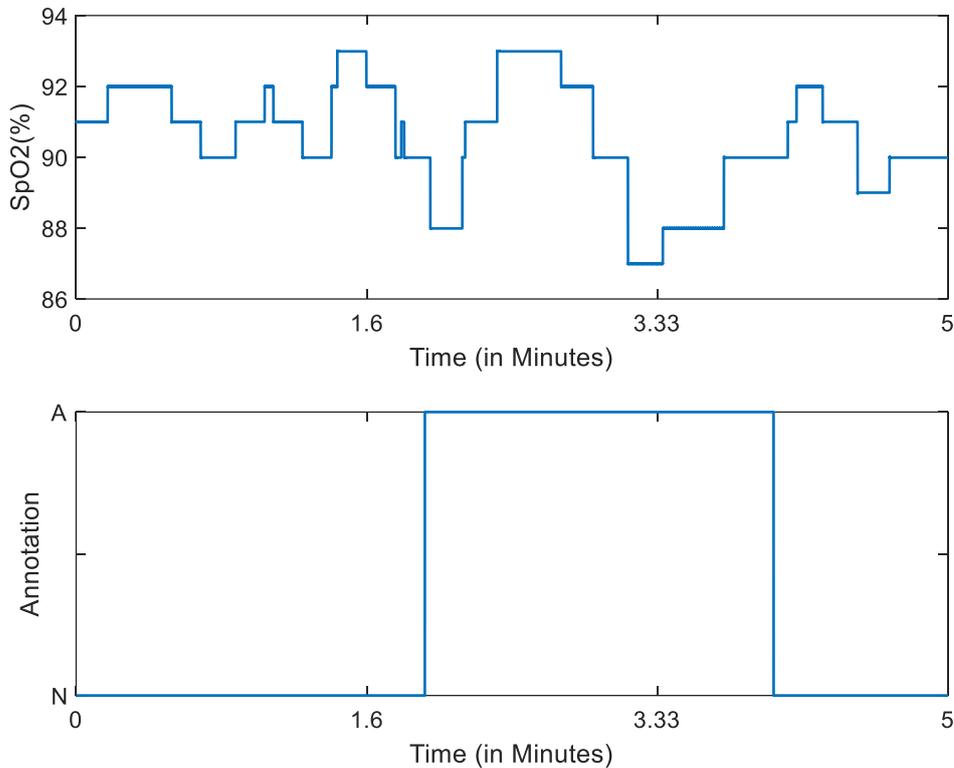


Figure 6: Oxygen saturation (SpO2) in a 5 minute segment.

5.2.2. Oximetry feature selection

The proposed work is focused on choosing the best feature set for different classifiers (ANN, SVM_L, SVM_P, SVM_RBF, NB, LDA, and KNN) using two different feature selection methods. These feature selection methods are SFS and mRMR and the general flowchart of the implemented model is presented in Figure 7. One of the focuses of this work is to compare the performance of these two feature selection methods for obstructive sleep apnea classification using the SpO2 signal. The advantages and disadvantages of the implementation of both algorithms are also discussed.

Data are collected from two well-known databases (the Physionet Database and the Ucddb Database). The first step is the pre-processing and data preparation (Figure 7 Dataset to node 1) which includes segmentation, unwanted data removal, annotation linking with segmented SpO2 signals, and the performance of feature extraction (Table 6). After the features are extracted, the variables are chosen with two different feature selection methods, mRMR (Figure 7 dashed line) and SFS (Figure 7

dotted line). For both feature selection techniques and for all the classifiers, datasets were divided into two parts, one for training and one for testing. To make the system independent from the effect of the subject the training subjects and test subjects are different. This is done by random sub-sampling choosing the test subjects and training subjects fifty times randomly ($r = 50$), and by making an average of the accuracies in the different iterations. The average results are used to select the best features for each classifier and method (Figure 7 node 3 to node 4). In the flowchart, the logical 1 and 0 indicate true and false, respectively. The j_{max} is the j^{th} feature for which the feature set has maximum accuracy, N is the total number of features, and r is the repetition number. A brief description of each step is given below.

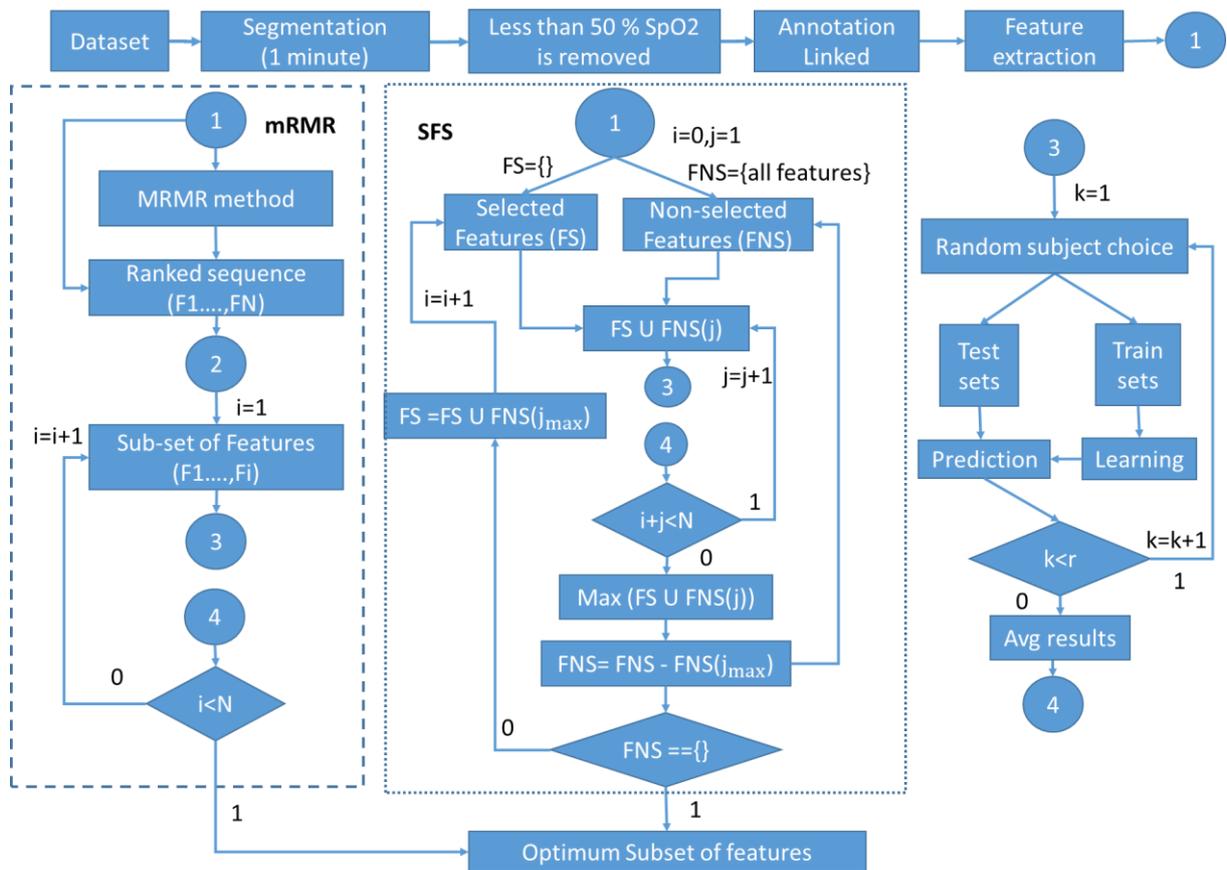


Figure 7 : General pipeline for finding the best feature sub-set from the SpO2 for sleep apnea detection.

Preprocessing: The Ucddb and the PhysioNet datasets sampling frequencies are 8 Hz and 50 Hz, respectively. For a uniform test, the Physionet sampling frequency is resampled to 8 Hz. This type of resampling is common in apnea detection [2], [106].

The Physionet dataset is annotated every minute by the physician whereas the Ucddb database is annotated according to the presence of the events in a continuous way. For feature extraction and classification, the continuous SpO2 data are segmented into one-minute intervals and linked with the

annotations of normal or apnea events. According to the definition of OSA, an apnea event should last at least ten seconds [63]. In the case of the continuously annotated Ucddb, an apnea event with, for example, 10 seconds could be divided into two adjacent minutes each having a fewer amount of apnea event time than what is needed to be identified as an apnea minute. To solve this problem, in the presence of five or more consecutive seconds of apnea or hypopnea, the minutes are treated as apnea [73],[106].

Any segmented minute with SpO2 values of less than 50% are considered as artifacts and hence removed from the analysis [73],[106].

SpO2 Features: Features are the distinctive attributes of the signal that are used by the classifier to identify the classes. A total of 61 features were studied in this work, due to their reported successful performance in the classification of apnea events [102] [195]. These features can be broadly divided into three categories: Time Domain (TD) (13 features); Frequency Domain (FD) (20 features); and time-frequency domain (28 features), where a Daubechies 3 wavelet of decomposition level of six was used for the Time Frequency Domain (TFD) analysis. The TF details coefficients ($CD1(n), CD2(n), \dots, CD6(n)$) and the approximations coefficients $CA6(n)$, which were used to extract features. A description of the features is listed in Table 6.

Table 6: List of SpO2 features.

Type	Feature Number	Feature	Details about Features
TD	1	Mean (Avg)	$Avg = \frac{1}{n} \sum_{i=1}^n x_i$ where n is the number of data points of the signal x .
TD	2	Variance (Var)	$Var = \frac{1}{n} \sum_{i=1}^n (x_i - Avg)^2$
TD	3	Coefficient of Variation (CoV)	$CoV = \frac{\sqrt{Var}}{Avg}$
TD	4	Skewness (Sk)	$Sk = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - Avg)^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - Avg)^2 \right)^{3/2}}$
TD	5	Kurtosis ($Kurt$)	$Kurt = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - Avg)^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - Avg)^2 \right)^2}$
TD	6	Root Mean Square (RMS)	$RMS = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i)^2}$
TD	7	Maximum (Max)	$Max = maximum(x)$
TD	8	Minimum (Min)	$Min = minimum(x)$
TD	9	Shannon Entropy (SEn)	$SEn = -\sum_{i=1}^n p(i) \ln(p(i))$ Where $p(i)$ is the probability of a specific event occurrence.
TD	10	Renyi Entropy (REn)	$REn = \frac{1}{1-q} \ln(\sum_{i=1}^n p(i)^q)$
FD	11	Twenty equally spaced filters to form a Filter Bank (Fb) [10] [197] where, the Δ_m is the bandwidth of the m^{th} filter with a window U and center frequency b_m , and N is a number of samples. For the first filter bank ($Fb1$) $m = 1$.	$Fb1 = \frac{\sum_{k=b_1-\Delta_1}^{b_1+\Delta_1} \left \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}} \right ^2}{\sum_{k=0}^{N-1} \left \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}} \right ^2} U_{\Delta_1}$

FD	12	Second filter bank (<i>Fb2</i>)	$Fb2 = \frac{\sum_{k=b_2-\Delta_2}^{b_2+\Delta_2} \left \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}} \right ^2 U_{\Delta_2}}{\sum_{k=0}^{\frac{N}{2}-1} \left \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}} \right ^2}$
FD
FD	30	Twentieth filter bank (<i>Fb20</i>)	$Fb20 = \frac{\sum_{k=b_{20}-\Delta_{20}}^{b_{20}+\Delta_{20}} \left \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}} \right ^2 U_{\Delta_{20}}}{\sum_{k=0}^{\frac{N}{2}-1} \left \frac{1}{N} \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi nk}{N}} \right ^2}$
TFD	31	Entropy of the level 6 approximation of wavelet (<i>TFSEnCA6</i>)	$TFSEnCA6 = SE_n(CA6)$
TFD	32	Variance of the level 6 approximation of wavelet (<i>TFVarCA6</i>)	$TFVarCA6 = \frac{1}{n} \sum_{i=1}^n (CD6_i - Avg(CA6))^2$
TFD	33	Standard deviation of the level 6 approximation of wavelet (<i>TFSdCA6</i>)	$TFsCA6 = \sqrt{\frac{\sum_{i=1}^n (CA6_i - Avg(CD6))^2}{N-1}}$
TFD	34	Median absolute deviation of the level 6 approximation of wavelet (<i>TFMadCA6</i>)	$TFMadCA6 = median(CA6)$
TFD	35	Entropy of the level 6 details of wavelet (<i>TFSEnCD6</i>)	$TFSEnCD6 = SE_n(CD6)$
TFD	34	Variance of the level 6 details of wavelet (<i>TFVarCD6</i>)	$TFVarCD6 = \frac{1}{n} \sum_{i=1}^n (CD6_i - Avg(CD6))^2$
TFD	37	Standard deviation of the level 6 details of wavelet (<i>TFSdCD6</i>)	$TFsCD6 = \sqrt{\frac{\sum_{i=1}^n (CD6_i - Avg(CD6))^2}{N-1}}$
TFD	38	Median absolute deviation of the level 6 details of wavelet (<i>TFMadCD6</i>)	$TFMadCD6 = median(CD6)$
TFD	44	Entropy of the level 5 details of wavelet (<i>TFSEnCD5</i>)	$TFSEnCD5 = SE_n(CD2)$
TFD	45	Variance of the level 5 details of wavelet (<i>TFVarCD5</i>)	$TFVarCD5 = \frac{1}{n} \sum_{i=1}^n (CD2_i - Avg(CD2))^2$
TFD	46	Standard deviation of the level 5 details of wavelet (<i>TFSdCD5</i>)	$TFsCD5 = \sqrt{\frac{\sum_{i=1}^n (CD2_i - Avg(CD2))^2}{N-1}}$
TFD	47	Median absolute deviation of the level 5 details of wavelet (<i>TFMadCD5</i>)	$TFMadCD5 = median(CD2)$
TFD
TFD	55	Entropy of the level 1 details of wavelet (<i>TFSEnCD1</i>)	$TFSEnCD1 = SE_n(CD1)$
TFD	56	Variance of the level 1 details of wavelet (<i>TFVarCD1</i>)	$TFVarCD1 = \frac{1}{n} \sum_{i=1}^n (CD1_i - Avg(CD1))^2$
TFD	57	Standard deviation of the level 1 details of wavelet (<i>TFSdCD1</i>)	$TFsCD1 = \sqrt{\frac{\sum_{i=1}^n (CD1_i - Avg(CD1))^2}{N-1}}$
TFD	58	Median absolute deviation of the level 1 details of wavelet (<i>TFMadCD1</i>)	$TFMadCD1 = median(CD1)$
TF	59	Central Tendency Measure (<i>CTM50</i>) [93] where $r = 0.58$ is selected from previous research [93].	$CTM = \frac{\sum_{i=1}^{n-2} \partial d_i}{n-2}$ $\partial(d_i) = \begin{cases} 1 & \text{if } [(x_{i+2} - x_{i+1})^2 + (x_{i+1} - x_i)^2]^{0.5} < 0.58 \\ 0 & \text{otherwise} \end{cases}$
TF	60	Delta Index (<i>DIndex</i>) [198] [199]	$DIndex = \frac{1}{n} \sum_{i=1}^n \left \frac{\partial(SpO_2)}{\partial t} \right $ (12 second intervals)
TF	61	Oxygen saturation Index (<i>ODI3</i>) [93] [200]	∂ is the change in SpO2; n=number of intervals; and t=time. $ODI3 = \text{Number of falls} < ((.03 * (\text{Avg of first 3 minutes})))$

5.2.3. Performance of the mRMR method

Physionet dataset: For the Physionet dataset with the mRMR method, the highest accuracy is achieved by the SVM classifier with a linear kernel (SVM-L) (Table 7). However, the best sensitivity is achieved by SVM-RBF and the best specificity is achieved by the LD classifier. The lowest number of features is achieved by the ANN and NB classifiers.

There are no significant differences in performance between the different classifiers. The NB and ANN classifiers with the lowest number of features are 1.72% and 1.56% less accurate than the SVM-L classifier, which uses 25 times more features. Though the SVM-L achieved the highest accuracy with 50 features, from a practical point of view it is not necessary to use this high number of variables. With 16 features the same classifier presents an accuracy of 96.76% and with 3 features 95.16% is achieved. Thus, 3.33 times more features are used to increase 0.13% accuracy, and 16.67 times more features are used to increase 1.73% accuracy (Figure 8).

The accuracy rises for the KNN classifier at the beginning of the process of the addition of features (2 features). The accuracy remains stable until 12 features are added and then a sudden drop is observed among all the KNN classifiers. These first two features are the most effective and this trend of sharp accuracy rises by just adding two features is common in all the classifiers (Figure 8). The sensitivity and specificity are showed in Figure 9 and Figure 10.

UCDDB database: The performances in Acc, Se, and Sp achieved with Ucdadb are poorer than with the Physionet database (Figure 11, Figure 12, and Figure 13). The highest accuracy is achieved by an ANN with the lowest number of features. SVM-RBF, KNN1, and KNN3 perform poorly compared to other classifiers. However, these classifiers need a lower number of features. The highest specificity (Sp) of 98.33% is presented by an NB, which obtains an accuracy of 80.39 % which is not too distant from the highest achieved accuracy of 81.95% by a ANN with poor sensitivity. On the other hand, with poor accuracy (Acc), the SVM-RBF presents the highest sensitivity.

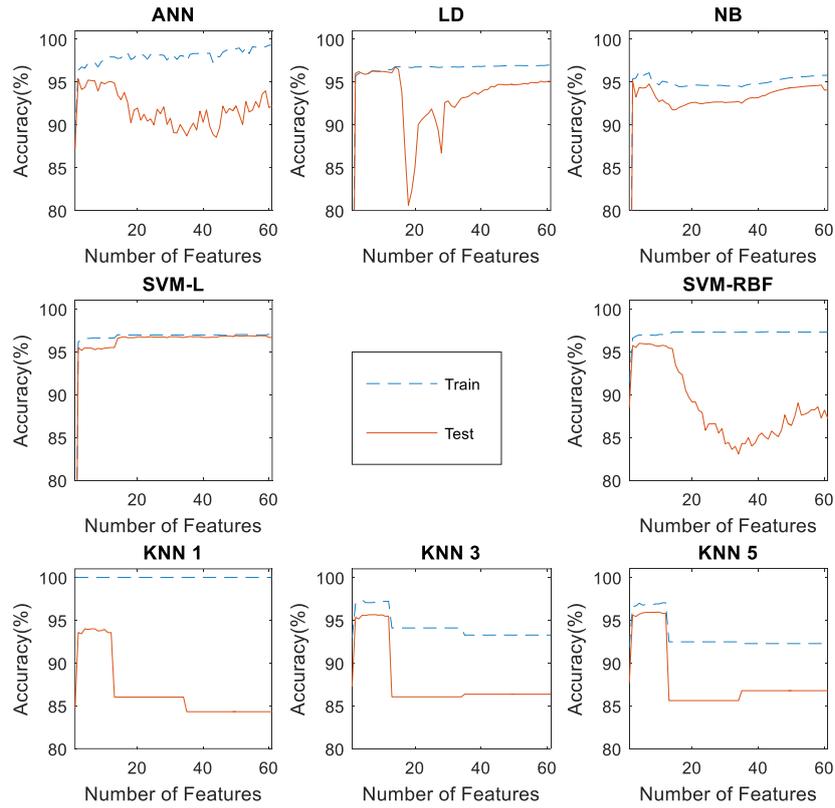


Figure 8 : Accuracy of the Physionet mRMR method.

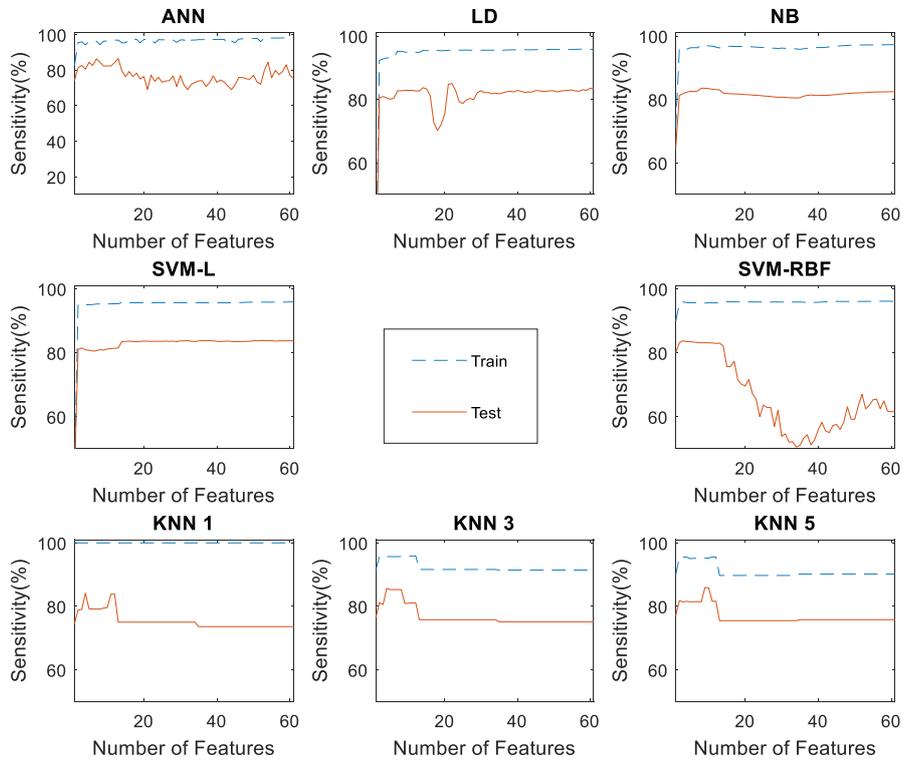


Figure 9 : Sensitivity of the Physionet mRMR method.

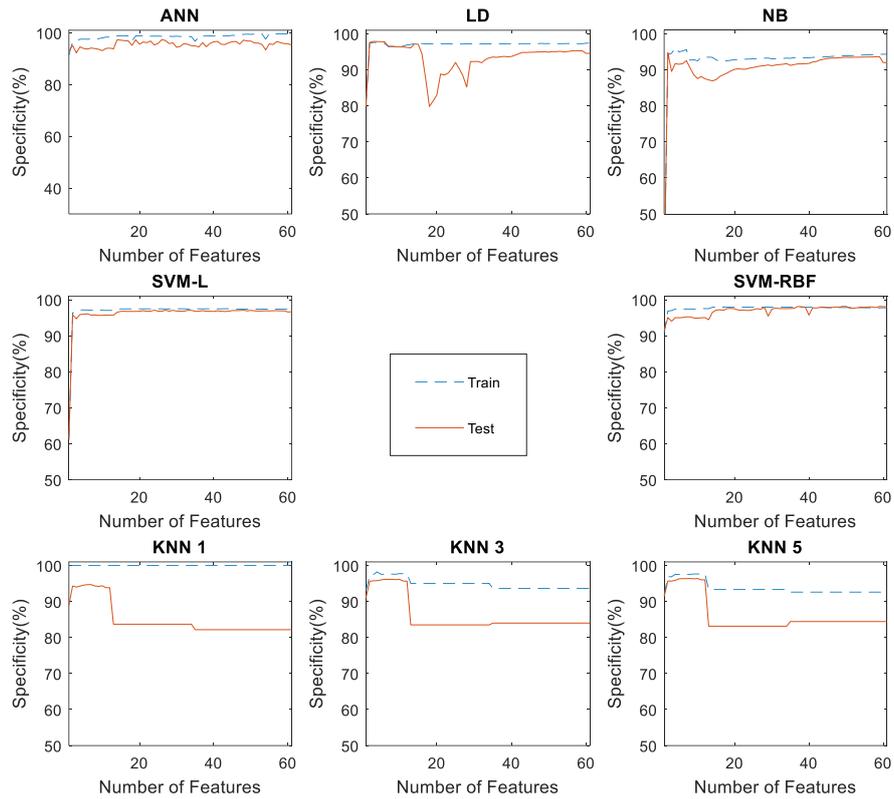


Figure 10 : Specificity of the Physionet mRMR method.

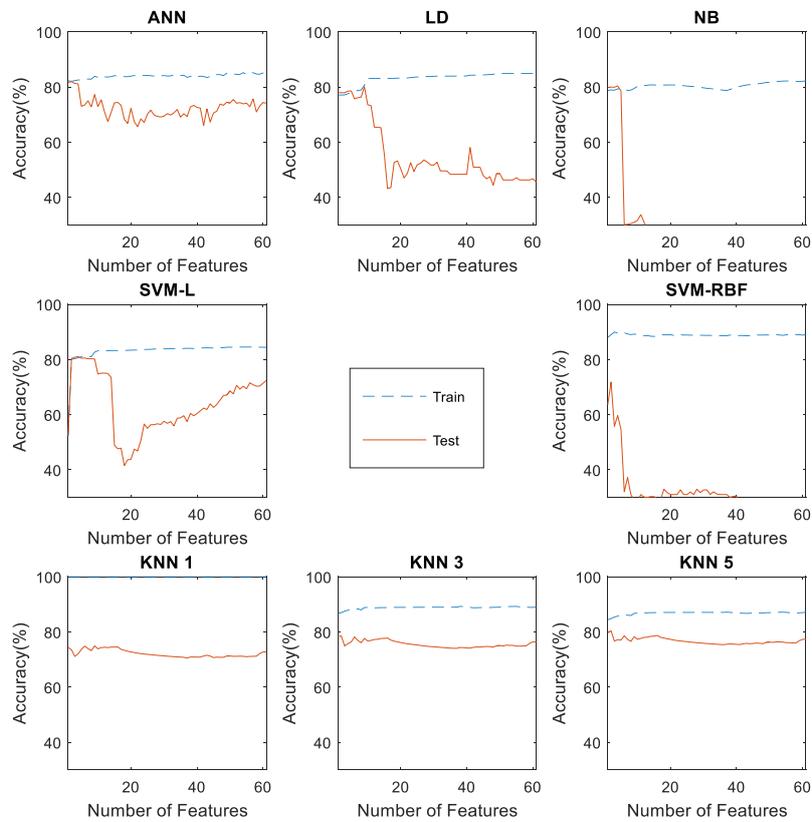


Figure 11: Accuracy of the Ucd db mRMR method.

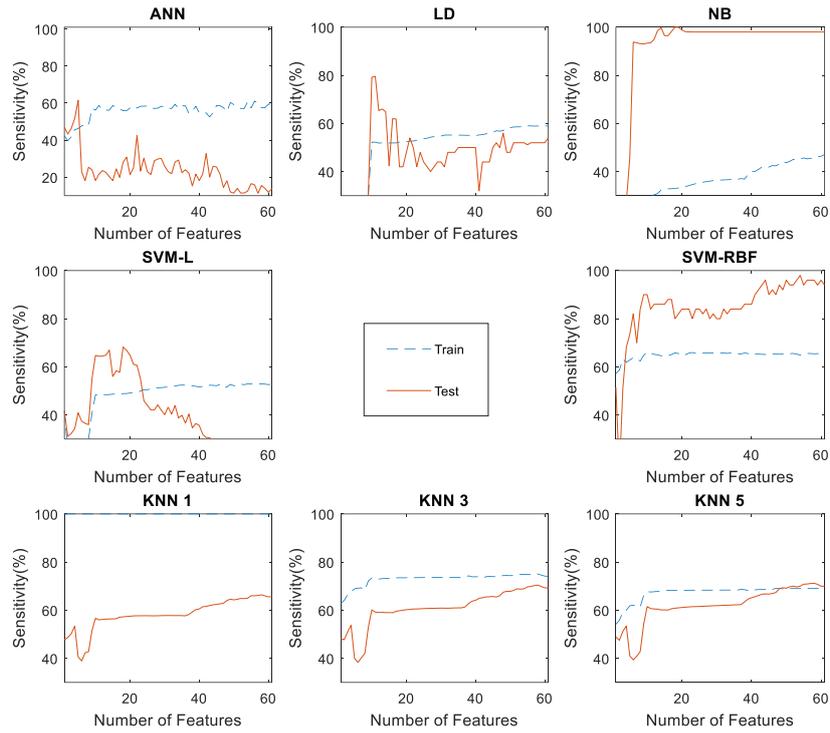


Figure 12 : Sensitivity of the Ucdb mRMR method.

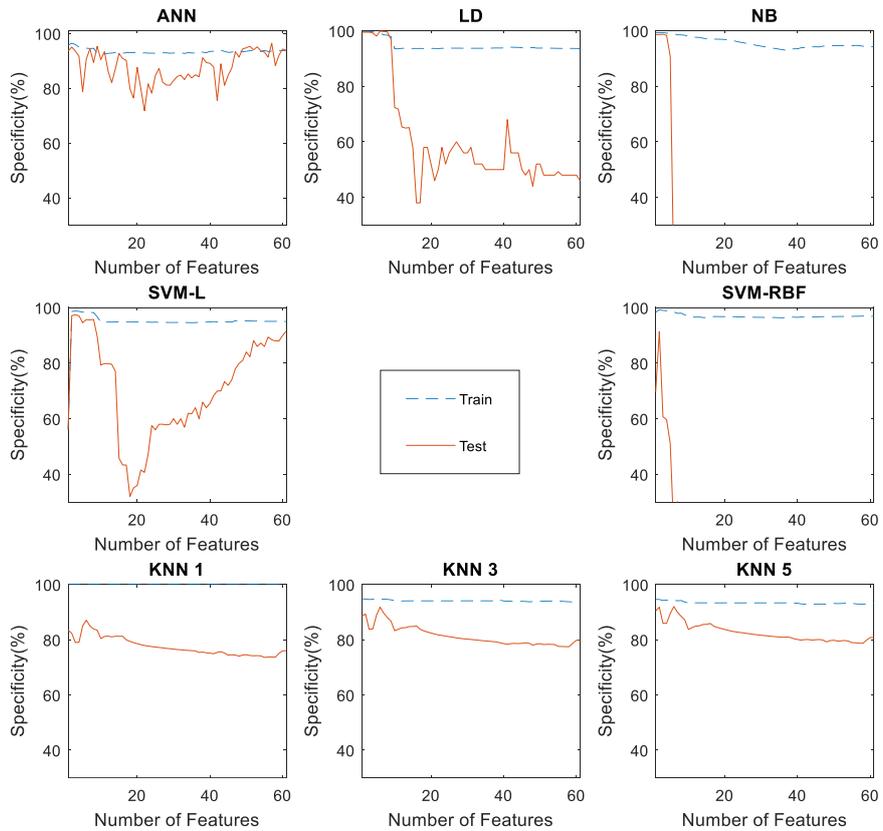


Figure 13 : Specificity of the Ucdb mRMR method.

Table 7 : Comparison of the selected features for both databases using the mRMR method (best values are marked in bold).

Database	Classifier	No of features	Performance [%]		
			Se	Sp	Acc
Physionet Database	ANN	2	81.41	95.11	95.33
	LD	14	83.68	97.19	96.75
	NB	2	81.26	94.73	95.17
	KNN1	4	84.11	94.27	94.00
	KNN3	8	85.28	96.16	95.65
	KNN5	10	85.87	96.40	95.95
	SVM-L	50	83.76	97.03	96.89
	SVM-RBF	4	95.69	95.02	95.99
Ucddb Database	ANN	2	43.31	95.03	81.95
	LD	9	31.39	96.93	80.10
	NB	4	27.17	98.33	80.39
	KNN1	9	51.60	83.43	74.96
	KNN3	2	47.70	89.25	78.56
	KNN5	2	47.51	91.76	80.42
	SVM-L	4	34.32	97.04	81.04
	SVM-RBF	2	58.57	91.37	71.83

5.2.4. Performance of Sequential Forward Search

Compared to the mRMR method, the SFS presents a much smoother training and test curve (Figure 8 to Figure 13 vs. Figure 14 to Figure 19) The highest accuracy of 97.38% is achieved in the Physionet database using a SFS method by the SVM-L classifier with 20 features. A ANN with 10 features and an NB with 14 features present the highest sensitivity and specificity, respectively (Table 8).

In case of the Ucddb database, the highest accuracy, sensitivity, and specificity are achieved by the LD (Acc 83.27%) with 9 features, the SVM-RBF (Se 74.06%) with 10 features, and the KNN3 (Sp=94.45) with 9 features, respectively (Table 8, Figure 17, Figure 18, and Figure 19).

Table 8 : Comparison of selected features for both databases using the sequential forward search method (desired values are marked in bold).

Database	Classifier	No of features	Performance [%]		
			Se	Sp	Acc
Physionet	ANN	10	84.71	96.17	96.13
	LD	21	84.28	97.73	97.19
	NB	14	84.71	94.91	96.48
	KNN1	18	82.58	94.94	95.29
	KNN3	43	82.07	96.35	96.19

Ucddb Database	KNN5	5	83.02	96.76	96.53
	SVM-L	20	84.57	97.28	97.38
	SVM-RBF	32	84.59	96.50	96.83
	ANN	3	44.30	95.39	82.40
	LD	9	61.78	91.03	83.27
	NB	9	59.83	91.58	83.18
	KNN1	33	59.06	81.10	75.11
	KNN3	9	42.11	94.45	81.24
	KNN5	6	40.67	95.74	81.81
	SVM-L	11	49.57	93.54	82.35
	SVM-RBF	10	74.06	81.39	78.86

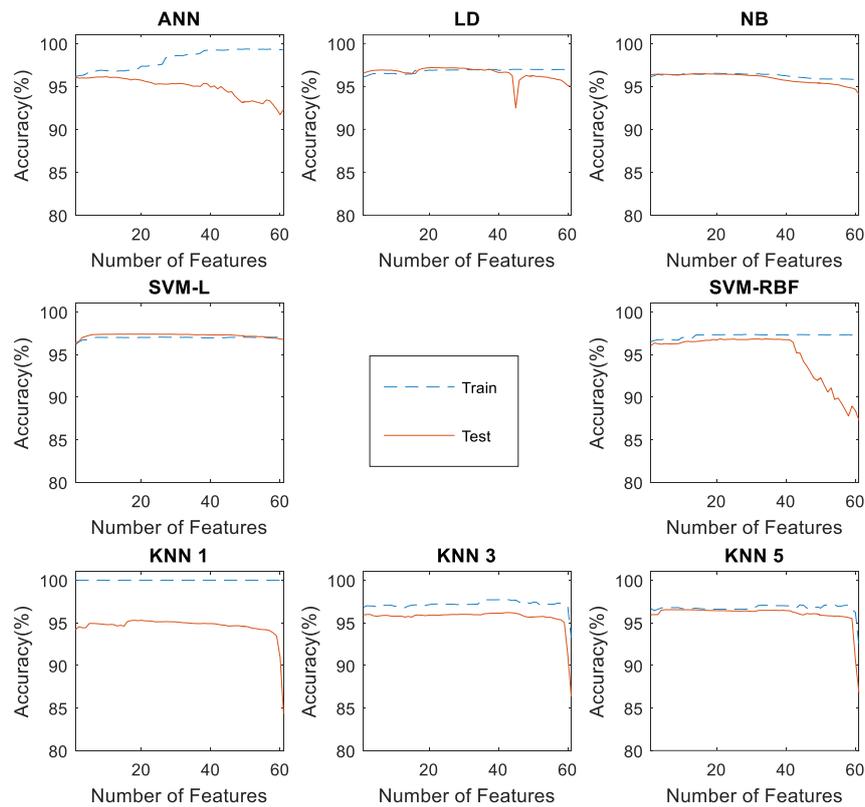


Figure 14: Accuracy of the Physionet SFS method.

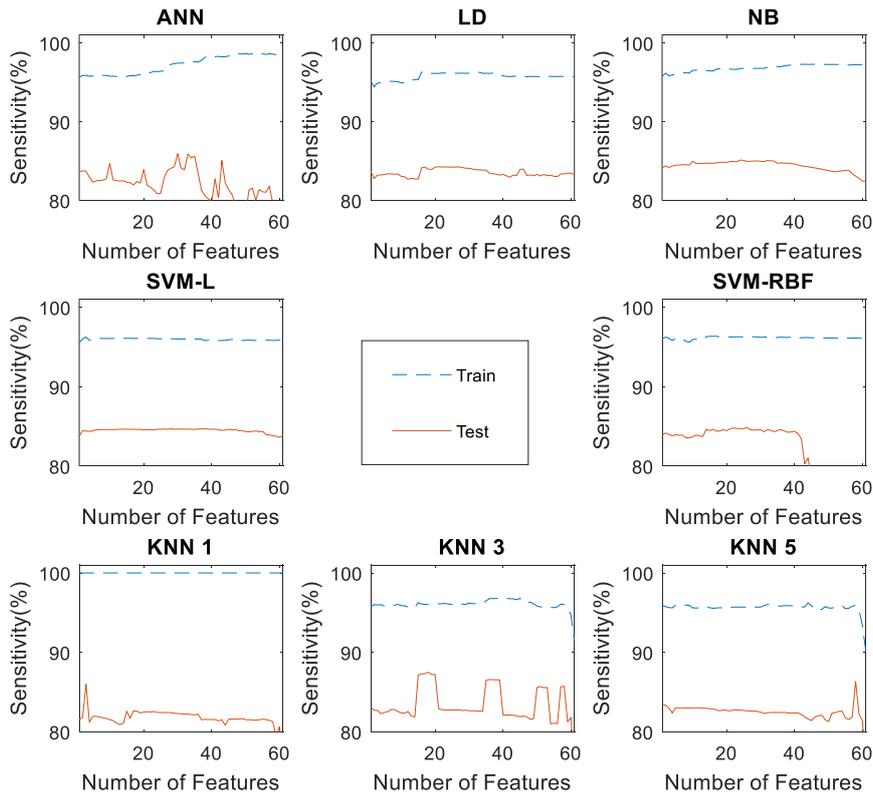


Figure 15 : Sensitivity of the Physionet SFS method.

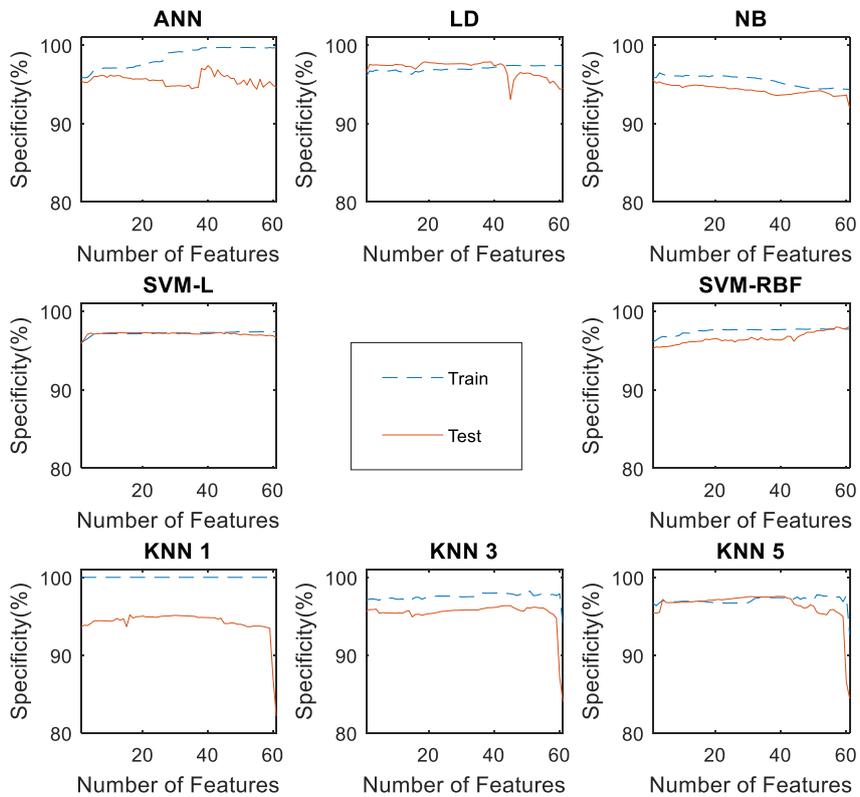


Figure 16 : Specificity of the Physionet SFS method.

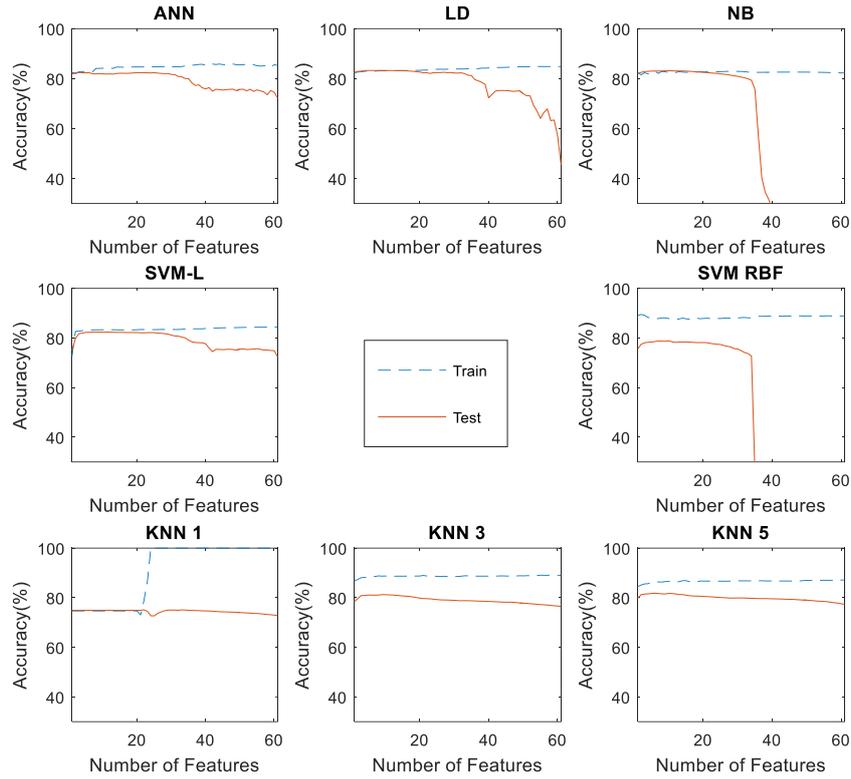


Figure 17: Accuracy of the Ucddb SFS method.

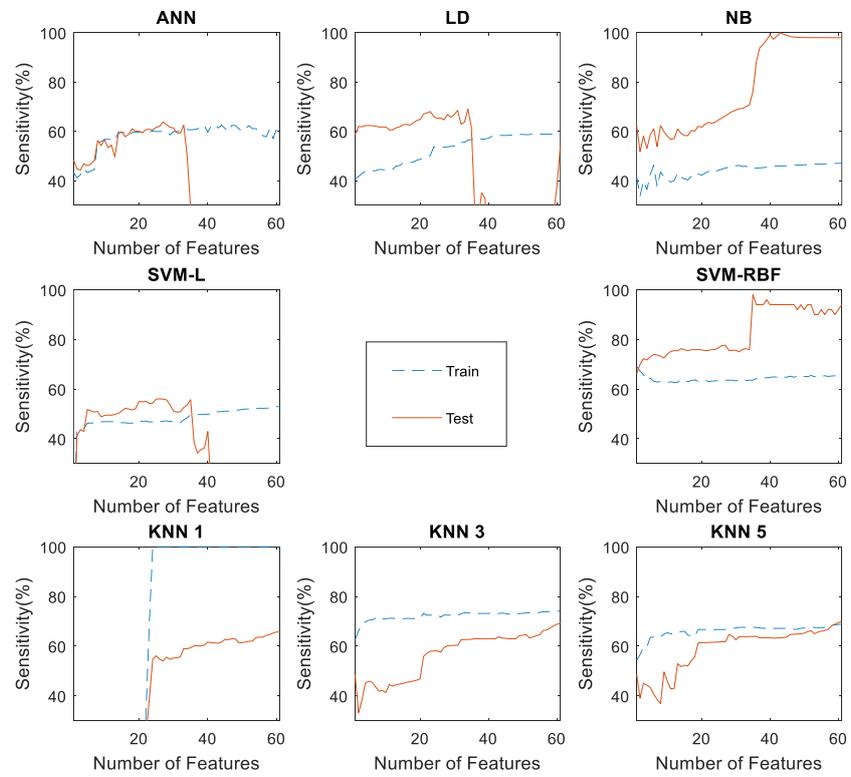


Figure 18 : Sensitivity of the Ucddb SFS method.

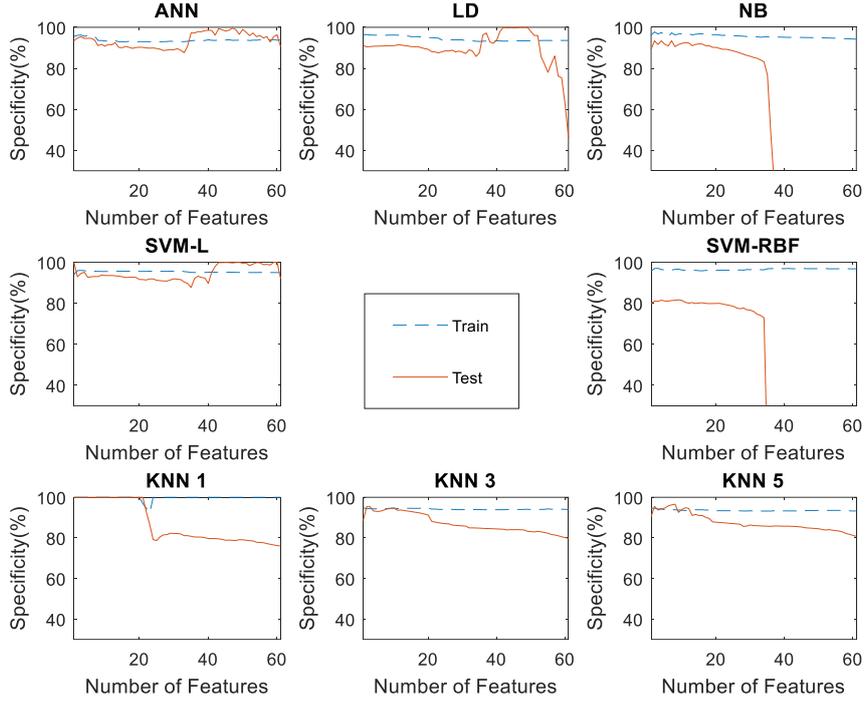


Figure 19 : Specificity of the Ucdb SFS method.

5.2.5. Comparison between mRMR and SFS

The comparison between the two methods is carried out in terms of time requirement, performance parameters, and the number of features required.

Time performance: The time to calculate the performance of one specific feature set is the sum of the Training Time (TR_t) and Test Time (TE_t) multiplied by the number of folds or the number of Repetitions (R_f) the classifier is trained and tested by. The total search time of the feature set is the multiplication of the number of feature sets by the sum of the training and testing times plus the pre-training and testing times. If there are F_n number of features, then for SFS we have $F_n(F_n + 1)/2$ numbers of feature sets and mRMR has F_n number of feature sets. Though the SFS does not imply have preprocessing steps, the mRMR method has feature ranking steps. If that takes FR_t time, then the time ratio between the two methods is

$$\frac{SFS_t}{mRMR_t} = \frac{R_f \times (TR_t + TE_t) \times \frac{F_n(F_n+1)}{2}}{FR_t + (R_f \times (TR_t + TE_t) \times F_n)} \approx \frac{F_n+1}{2} \quad (5)$$

The training and testing of each set take most of the time in the search for the best set of features. Thus, if this condition $FR_t \ll (R_f \times (TR_t + TE_t) \times F_n)$ is applied to the ratio, it can be seen that SFS is approximately $(F_n + 1)/2$ times more time-consuming.

In addition to that, if parallel processing is considered, the two methods act differently. In both cases, the repetition of the training and testing process is susceptible to be parallelized. However, in

the search phase, the SFS cannot be parallelized because of its sequential nature in which its current state depends on the previous state. On the other hand, after the ranking sequence of the feature set, it is possible to parallelize the training and test portion of the mRMR method, which is the most time-consuming part. If we consider that all the feature sets take the same time to training and test and the user provides infinite parallel resources to run, then the mRMR method can have F_n parallel computations. If the time ratio of the previous calculation is considered, then the mRMR method can have a theoretical speed $F_n \times (F_n + 1)/2$ times faster than the SFS. However, in practice, it is a shorter time, because the feature sets take different amounts of time to finish and the parallel system depends on the maximum time instead of the average time. In addition to that, deploying and gathering information and communication between the parallel processes require time.

For the mRMR method, a feature ranking is needed though the time consumption is small. To sum it up, if the user has access to parallel resources, the mRMR method is a great choice. Even in the scenario that the user does not have access to parallel processing, from the time consumption point of view, the mRMR method is better than the SFS.

Performance parameters: Performance of Acc, Se, and Sp are presented in Figure 20 for all the feature selection methods and datasets. It can be seen that, for the Physionet database, the median sensitivity of the mRMR method is 83.94% while median sensitivity of the SFS is 84.43%. In addition to this, the maximum and minimum of the mRMR is 10.98% higher and 0.81% lower compared to the SFS. The mRMR method presents lower median specificity (Sp) (95.64%) than median specificity (95.4%) gain by the SFS. The accuracy (median value: SFS 96.51% mRMR 95.37%) also follows a similar trend. In the case of the Ucddb database, SFS presents a higher median Se (SFS: 54.31%, mRMR: 45.41%) and median Acc (SFS: 82.08%, mRMR: 80.25%) than the mRMR method. On the other hand, with a 93.4% median Sp mRMR method has higher Sp than that of the SFS.

Required number of features: Regarding the number of features and considering both databases, the mRMR method presents a median of 4 features while the SFS presents 10. The high number of features for the SFS is required for Physionet with a median of 19, and a median of 9 features is required for the Ucddb dataset. On the other hand, mRMR presents 6 features with Physionet and 3 features with Ucddb (Figure 21). Details of the selected features and the rank of the features are listed in Table 9 to Table 12.

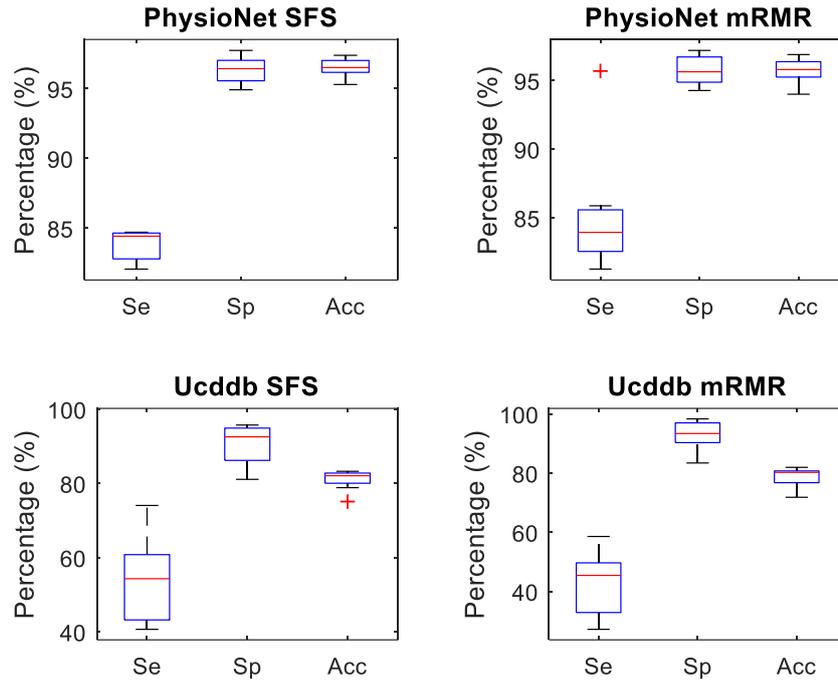


Figure 20 : Overall performance for all the classifiers using mRMR and SFS algorithms in both databases.

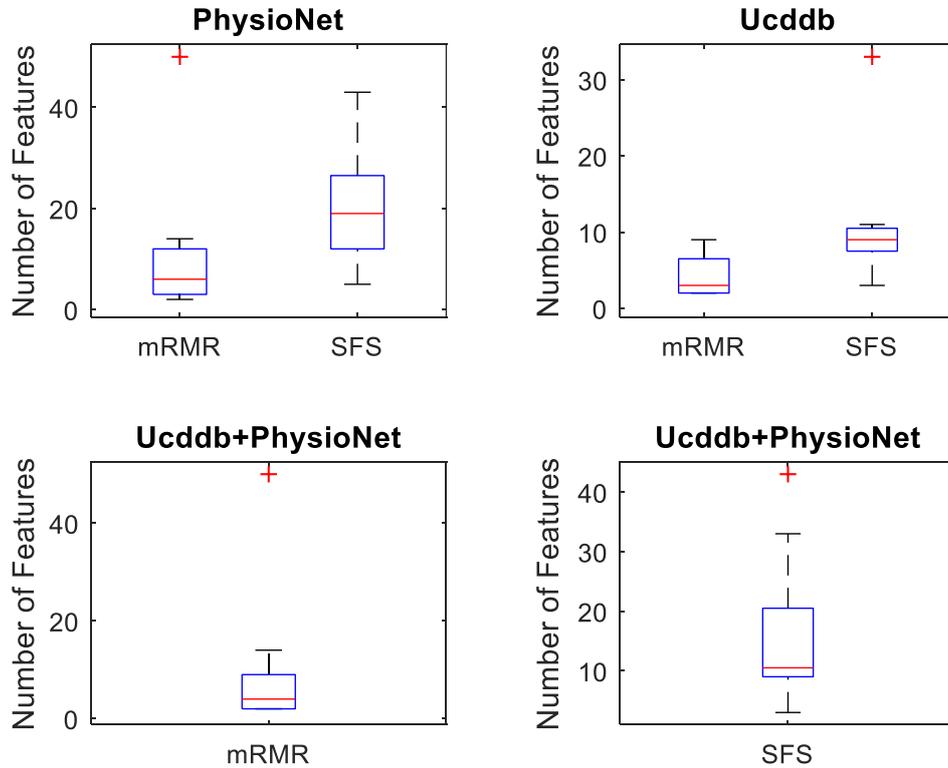


Figure 21 : Overall selected number of features for all the classifiers using mRMR and SFS algorithms in both databases.

Table 9 : mRMR feature rank for the Physionet database.

Rank	1	2	3	4	5	6	7

Feature Number	35	47	1	2	39	43	3
Rank	8	9	10	11	12	13	14
Feature Number	5	6	7	8	9	10	11
Rank	15	16	17	18	19	20	21
Feature Number	12	13	14	15	16	17	18
Rank	22	23	24	25	26	27	28
Feature Number	19	20	21	22	23	24	25
Rank	29	30	31	32	33	34	35
Feature Number	26	4	27	28	29	30	31
Rank	36	37	38	39	40	41	42
Feature Number	32	33	34	36	37	38	40
Rank	43	44	45	46	47	48	49
Feature Number	41	42	44	45	46	48	49
Rank	50	51	52	53	54	55	56
Feature Number	50	51	52	53	54	55	56
Rank	57	58	59	60	61		
Feature Number	57	58	59	60	61		

Table 10 : mRMR feature rank for the Ucddb database.

Rank	1	2	3	4	5	6	7
Feature Number	43	4	47	35	51	1	55
Rank	8	9	10	11	12	13	14
Feature Number	39	2	3	5	6	7	8
Rank	15	16	17	18	19	20	21
Feature Number	9	10	11	12	13	14	15
Rank	22	23	24	25	26	27	28
Feature Number	16	17	18	19	20	21	22
Rank	29	30	31	32	33	34	35
Feature Number	23	24	25	26	27	28	29
Rank	36	37	38	39	40	41	42
Feature Number	30	31	32	33	34	36	37
Rank	43	44	45	46	47	48	49
Feature Number	38	40	41	42	44	45	46
Rank	50	51	52	53	54	55	56
Feature Number	48	49	50	52	53	54	56
Rank	57	58	59	60	61		
Feature Number	57	58	59	60	61		

Table 11 : SFS feature sequence for the Physionet database.

Sequence Rank	Classifiers							
	ANN	KNN1	KNN3	KNN5	LD	NB	SVM-L	SVM-RBF
1	52	3	53	53	53	53	53	52
2	53	59	56	54	3	49	23	61
3	5	56	52	61	49	2	58	39
4	41	53	59	11	43	52	2	40
5	57	49	57	38	38	57	43	37
6	40	52	48	48	5	48	50	53

7	59	48	45	52	4	50	49	36
8	4	58	58	56	36	54	16	32
9	43	54	54	3	59	44	24	47
10	51	57	50	23	61	14	22	7
11	56	50	44	57	41	43	32	58
12	39	44	46	58	39	55	19	55
13	35	46	38	50	52	58	21	51
14	55	45	3	13	48	46	5	11
15	36	2	2	46	54	56	56	10
16	47	11	11	44	11	51	25	1
17	58	38	49	24	40	7	30	35
18	42	17	37	12	58	4	59	3
19	38	37	13	15	51	32	12	49
20	61	20	17	22	8	47	4	48
21	37	42	15	14	47	3	14	41
22	50	41	22	59	57	61	18	54
23	48	24	20	25	50	45	13	42
24	45	12	12	16	45	27	48	57
25	44	15	16	17	44	42	45	6
26	49	26	28	21	46	37	55	38
27	9	13	18	18	56	38	57	8
28	6	18	41	19	55	33	54	9
29	31	16	42	20	35	8	27	31
30	1	21	40	26	42	34	26	50
31	54	22	19	40	37	41	47	34
32	10	19	21	60	7	16	29	59
33	8	25	24	45	33	59	51	46
34	32	40	14	49	13	13	28	2
35	46	29	7	41	34	40	44	45
36	2	30	61	9	9	10	52	56
37	7	55	23	42	1	5	46	43
38	11	27	25	27	31	23	8	44
39	16	14	9	28	17	25	15	33
40	33	28	26	29	10	12	20	60
41	13	23	27	30	2	19	17	5
42	34	9	29	37	6	21	11	16
43	23	60	30	36	32	28	33	4
44	14	47	60	43	60	24	3	24
45	15	32	36	55	14	26	38	19
46	60	36	4	51	16	22	34	29
47	24	61	5	47	21	29	42	13
48	17	33	43	4	22	30	41	17
49	22	51	51	7	19	18	61	28
50	28	43	35	5	24	17	1	12
51	3	34	47	8	29	20	7	26
52	27	4	39	1	26	15	9	30
53	30	5	55	2	28	9	6	15
54	25	7	6	6	30	1	31	22

55	18	8	1	35	23	6	10	18
56	26	1	8	39	12	31	37	14
57	12	6	34	34	15	36	35	27
58	20	39	33	33	27	39	40	25
59	29	35	32	32	20	35	36	21
60	19	10	10	10	18	11	39	23
61	21	31	31	31	25	60	60	20

Table 12 : SFS feature sequence for the Ucddb database.

Sequence Rank	Classifiers							
	ANN	KNN1	KNN3	KNN5	LD	NB	SVM-L	SVM-RBF
1	43	11	43	43	61	3	11	2
2	59	12	8	31	50	43	46	35
3	5	13	58	59	49	59	50	61
4	35	14	59	9	3	35	42	36
5	39	15	47	7	53	5	61	33
6	4	16	9	10	5	4	5	32
7	47	17	1	6	41	47	44	43
8	50	18	31	1	38	50	48	34
9	40	19	5	57	45	2	58	39
10	46	20	10	8	4	32	56	3
11	51	21	35	47	58	36	36	38
12	56	22	6	28	34	60	4	37
13	54	23	39	58	59	61	40	40
14	33	24	23	5	37	39	52	50
15	55	25	29	30	46	48	59	41
16	49	26	7	4	54	44	41	52
17	2	27	26	35	60	33	39	48
18	36	28	25	39	57	52	54	46
19	34	29	22	32	42	46	49	44
20	48	30	30	29	33	40	32	42
21	53	59	32	23	39	34	57	56
22	37	60	4	17	40	38	38	47
23	32	7	36	26	47	56	35	53
24	61	43	27	19	48	37	37	49
25	52	2	17	22	55	49	2	54
26	44	4	60	20	36	45	45	45
27	45	5	2	27	35	53	53	57
28	38	32	55	50	44	42	60	5
29	57	8	28	55	43	54	47	60
30	41	35	24	60	51	41	55	55
31	42	9	20	2	32	57	51	4
32	3	39	50	18	31	55	43	58
33	58	61	38	25	56	58	34	51
34	8	47	19	24	2	51	33	59
35	60	36	21	21	52	31	3	16
36	31	37	56	15	10	10	10	1
37	10	40	18	41	7	6	7	12

38	6	1	16	40	8	1	1	15
39	1	38	15	14	11	7	6	9
40	7	50	13	16	28	8	8	6
41	9	41	14	42	12	9	31	8
42	11	42	12	13	23	11	28	19
43	28	48	52	12	22	12	26	18
44	20	56	40	36	21	27	25	31
45	17	46	42	56	16	24	27	11
46	12	55	41	61	26	15	30	7
47	29	44	11	11	18	21	15	10
48	14	57	48	48	13	13	16	21
49	22	51	46	52	14	18	29	17
50	19	52	37	38	19	19	24	20
51	21	58	57	46	29	14	13	14
52	16	53	51	37	24	17	22	26
53	15	6	44	51	6	16	18	22
54	25	33	49	49	1	20	20	24
55	18	49	61	53	27	30	23	27
56	23	10	33	44	30	23	19	25
57	13	3	53	54	15	25	17	13
58	27	54	34	33	20	22	21	23
59	24	34	3	34	17	26	14	28
60	26	45	45	45	25	28	12	29
61	30	31	54	3	9	29	9	30

Comparison of the chosen features: In the case of the Physionet dataset, features 35 and 47 are ranked first with a frequency of 8 (Table 13) and both are ranked second on the Ucddb database. In the Ucddb database, the features 4 and 43 ranked as first with the same frequency (Table 13) and ranked third and sixth in the Physionet database. Overall, the best independent features which ranked number one with the frequency of 12 are 35, 43, and 47 (Table 13).

Table 13 : Features' importance by frequency (number of times the features are chosen) for the mRMR methods.

Independent of classifiers and dependent on feature selection method						
Position/rank	Dependent of Database			Independent of Database		
	Physionet		Ucddb			
	Frequency	Feature/ Features	Frequency	Feature/ Features	Frequency	Feature/ Features
1	8	35, 47	8	4, 43	12	35, 43, 47

2	6	1, 2	4	35, 47	9	4
3	4	3, 5, 39, 43	2	1, 2, 39, 51, 55	8	1, 2
4	3	6, 7			6	39
5	2	8-11			4	3, 5
6	1	4, 12-34, 36-38, 40-46, 48-50			3	6, 7
7					2	8-11, 51, 55
8					1	12-34, 36-38, 40-42, 44-46, 48-50

Compared to the mRMR method, the SFS classifier has more control in the feature choosing so there is more variety of features chosen by different classifiers than with the mRMR method. Feature 53 ranks first in the Physionet database, whereas features 43 and 5 rank first in Ucdadb database. This contributed to ranking second position feature number 43 in the database independent list (Table 14). This trend is also true for number one ranked feature 59 in these databases.

Regardless of the database classifiers or methods, the top ten features are 43, 35, 47, 2, 4, 5, 3, 59, 1 and 39. It is interesting to note that on the one hand time-frequency features perform well, while on the other hand, some simple time features also have a great impact. However, this ranking is general.

Table 14 : Features importance by frequency (number of times feature are chosen) for SFS methods. Independent of classifiers and dependent on feature selection method

Position/rank	Dependent of Database		Independent of Database			
	Physionet		Ucdadb			
	Frequency	Feature/ Features	Frequency	Feature/ Features		
1	8	53	6	5, 43	11	59
2	6	49, 52, 54, 58, 59	5	59	10	43
3	5	11, 38, 48, 50, 57	4	61	9	5, 53
4	4	2, 3, 40, 41, 43, 61	3	2, 3, 9, 35, 50	8	50, 58, 61
5	3	4, 5, 44, 46, 51, 56	2	4, 7, 8, 11, 31, 32, 36, 39, 47, 58	7	2, 3, 11, 49

6	2	7-9, 12, 14, 16,17, 19, 21-25, 30, 32, 36, 37, 39, 42, 45, 47, 55	1	1, 10, 12-30, 33, 34, 38, 41, 42, 44-46, 48, 49, 53, 56, 60	6	38, 48, 52, 54
7	1	1, 6, 10, 13, 15, 18, 20, 26-29, 31, 34, 35			5	4, 9, 41, 57
8					4	7, 8, 32, 35, 36, 39, 40, 44, 46, 47, 57
9					3	12, 14, 16, 17, 19, 21-25, 30, 31, 42, 45, 51
10					2	1, 10, 13, 15, 18, 20, 26-29, 34, 37, 55
11					1	6, 33, 60

5.2.6. Comparison with other methods

A comparison between the proposed method and others proposed in the literature was carried out with some limitations due to the lack of common features and datasets (Table 15). Sensitivity, specificity, and accuracy are converted to percentages for a common comparison. Some previous works present a similar or smaller number of features [2], [99], [10], [73]. However, the accuracy was lower compared to the method proposed in this work. Xie et al. [73] used a ANN and the PhysioNet database with the same type of features and with similar results. A automated features based work carried out with the deep auto encoder network [106] obtaining an accuracy higher than the investigated classifiers. However, in that study of features creation was not carried out and the difference is irrelevant when compared with the SFS LD classifier for Ucddb (1.99%) and with the SFS LD classifier for Physionet (0.26%).

Table 15 : Comparison of sleep apnea detection approaches with SpO2.

Method	Approach	Classifier	Database	No of features	Performance [%]		
					Se	Sp	Acc
[2]	SpO2	SVM	Own	7			90
[99]	SpO2	ANN	[61]	3	87.5	100	93.3
[10]	SpO2	LDA	Own	19	75.6	91	86.5
[73]	SpO2	Combined Classifier	[62]	39	83.55	81.25	81.81
[106]	SpO2	DAE	[62]	NA	60.36	91.71	85.26
	SpO2	DAE	[61]	NA	78.75	95.89	97.64
mRMR	SpO2	SVM-L	[61]	50	83.76	97.03	96.89
SFS	SpO2	SVM-L	[61]	20	84.57	97.28	97.38
mRMR	SpO2	ANN	[62]	2	43.31	95.03	81.95
SFS	SpO2	LD	[62]	9	61.78	91.03	83.27

5.2.7. Summary

This work provides improvements to most of the existing methods with the SpO2 signal used for the classification of apnea events. The performances of 61 individual SpO2 features were obtained. From the results, the use of a subset of features improved the performance when compared to a set with all the features combined. Most of the features chosen by the classifiers are defined in the time-frequency domain which supports previous research [21]. In addition to that, some basic features such as the variance have a strong influence over all the classifiers and methods.

In the test, the SVF-L has better accuracy in the Physionet database. On the other hand, the ANN and the LD did better in the Ucdadb database. KNN classifiers show an overfitting problem, so in the feature selection process, it is advised to use KNN with caution, especially with a low number of neighbors.

Clearly, the feature selection process is not independent, but is rather an algorithm, database and classifier dependent problem.

The databases used are also annotated differently. Physionet was annotated by apnea minutes, whereas the Ucdadb was continuously annotated. Due to the different hypothesis of different classifiers, this results in different importance of the features.

In an attempt to compare the mRMR and the SFS, it can be concluded that in terms of computation, the mRMR is the most suitable option. The SFS uses a higher number of features in most of the cases. The accuracy differences between both algorithms are not significant. Both methods suffer from low

sensitivity due to the prevalence of normal segments compared to apnea events. This problem could be solved by using a multi-objective approach [106] or penalization [73].

The performance can also be increased by the combination of classifiers [73] [201]. Adding more signals such as an ECG [201] can also increase the performance of the system. Different types of feature selection methods such as the Floating Forward Feature Selection Genetic Algorithm and, the Hill Climbing Algorithm [202] can be tested for performance assessment.

5.3. Genetic Algorithm for Feature Selection

5.3.1. Introduction

Inspired by natural selection, Genetic Algorithms (GA) can be used in optimization techniques. In the GA the population changes over time to maximize or minimize a parameter according to specific rules [203], [204]. In this case, the GA is trying to minimize the classification error. The classification systems play a big role in biomedical signal processing and decision making. Because of their learning capabilities and their successful use in the literature, Artificial neural networks (ANN) are used for classification. The same feature created in Section 5.2 is used to benchmark the results.

5.3.2. GA based feature selection method

Data were collected from the PhysioNet web site [205]. For this research, the Apnea-ECG database is used. After pre-processing, features are extracted from the dataset. These features are used along with an ANN for optimization of the system. The general pipeline of the methods is shown in Figure 22.

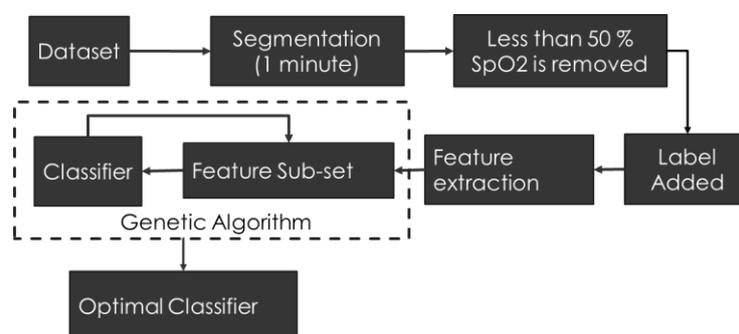


Figure 22: The general pipeline of optimization of the sleep apnea detection classifier.

Preprocessing: The dataset is annotated every minute by the physician and the annotation file is available with the recordings. For classification and feature extraction purposes, the continuous SpO2 data were segmented in a one-minute interval and linked with annotated Apnea or non-Apnea events.

The minutes in which a blood saturation level was lower than fifty percent were removed from the analysis because they are considered as artifacts [73].

GA steps: GA run an iterative loop that repeats the steps in Figure 23, considering that the individual population is a binary string and measuring the fitness and stopping criteria.

1) Initialize Population: The GA starts with a randomly generated initial population in which each individual solution is represented by a chromosome.

2) Evaluation: Each chromosome of the population is evaluated according to the pre-defined fitness function.

3) Stop Condition: The stopping criteria are checked to end the algorithm. If the criteria are not met, then the algorithm proceeds to the next step, or else it stops there.

4) Elitism: Elitism is applied to bypass high fitness valued individual's chromosomes to the next generation without alteration. This is done so that solutions do not degrade over the generation.

5) Parents Selection: Parents are selected according to a binary tournament selection process.

6) Crossover: It exchanges parts of the parent's chromosome information to their children or offspring.

7) Mutation: After the application of crossover, mutation is done to prevent premature convergence to local optima.

8) New population: The children created from the previous steps, form a new population. This new population is evaluated against the fitness function.

As shown in Table 6, the system primarily has 61 features. The challenge is to achieve good accuracy with a minimum error optimization. All the features are used to make chromosomes. A population with random features is generated. Using a binary GA, its evolution selection, crossover and mutation are carried out as presented in Figure 23.

The fitness function is defined by the ANN Misclassification Error ($MCE=(100-Acc)/100$). The GA has to minimize the MCE. The stopping condition used in this work is 100 generations and the population size was 50. The mutation and crossover rate were 0.1 and 0.8 respectively. The number of elites was two.

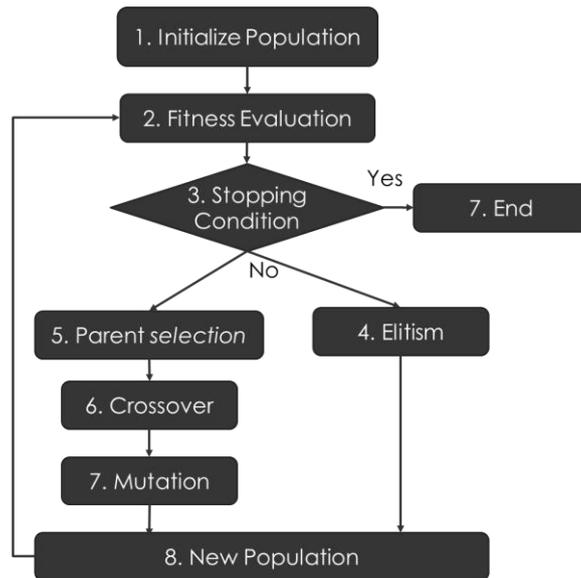


Figure 23 : Flow chart of GA.

5.3.3. Optimization result

The optimization of the system is done using a GA and a ANN. The system primarily has 61 features. To achieve good accuracy with a minimum error optimization is the challenge. All the features are used as chromosomes. In the beginning, a population with random features is generated. Using binary GA evolution optimization is achieved through the selection, crossover and mutation are as presented in Figure 23.

The GA is able to optimize the number of features and also to decrease the MCE. The best solution found by the GA was with seven features. The final structure of the optimized system is shown in Figure 24 and the training and testing results are shown in Figure 25.

The dataset has 1457 apnea events and 2278 non apnea events. All the events are randomly divided into three sections (70% for training, 15 % for validation, and 15 % for testing). In Figure 25, it can be seen that among 1009 apnea events 968 were correctly recognized in the training session. In the case of validation and testing, it was 210 among 219 and 221 among 229, respectively. The sensitivity and specificity of the system are high in all the cases. In the test, the specificity is 2% higher than the sensitivity. In the training and the test, it is 1.5% and 3.5% respectively. The total accuracy of the system is 97.2%, where the test accuracy is 97.7%, the training accuracy is 96.9% and the validation accuracy is 98.0%. The ANN with GA feature selection process result is compared with those of existing works (Table 16). The GA with a ANN is not only able to optimize the number of features but is also able to increase the accuracy (97.7%) when compared to other methods. The Genetic algorithm (GA) proposed this section [102] obtained an improvement of 0.32% accuracy compared to an SFS SVM-L with fewer features (For details please look into section 5.2). In the same work, the

authors have a 0.49% accuracy improvement compared with the mRMR SVM-L. However, the ANN with GA was not implemented with subject independency [102].

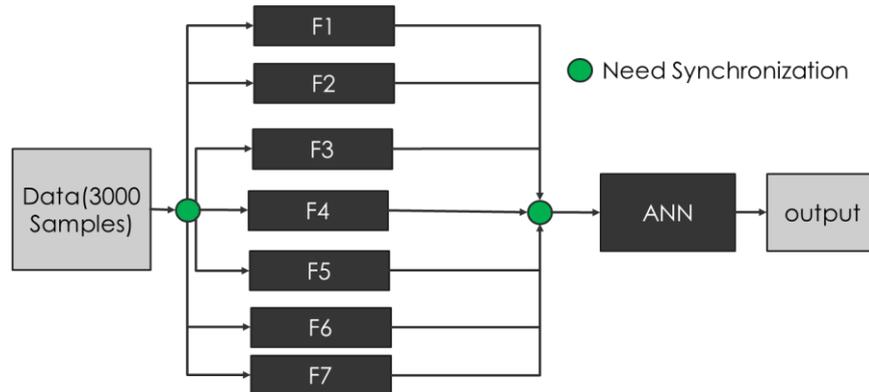


Figure 24 : Optimize structure of the system classifier using the GA and a ANN.



Figure 25 : Confusion matrix of a ANN with seven optimized features optimized using the GA over 100 generations.

Table 16 : Comparison of sleep apnea detection approaches.

Ref.	Approach	Classifier	Database	No of features	Performance [%]		
					Se	Sp	Acc
[114]	ECG	KNN3	[61]	30	NA	NA	92.67
[163]	SaO2+EEG	LR	Own	5	91	83.3	88.5

[108]	EEG	ANN	[90]	12	69.64	44.44	NA
[206]	HRV	LDA	Physionet Database	30	90.8	92.7	88.31
[72]	HRV	KNN27	[61]	4	83.90	88.50	85
[73]	SpO2 and ECG	KNN5	[62]	39	79.75	85.89	84.80
[99]	SpO2 signal	ANN	[61]	3	87.5	100	90.3
[126]	Respiration	LD	[61]	13	NA	NA	91
[125]	RR-interval	KNN5	Own	3	NA	NA	89.4
		QAD		3	NA	NA	94.5
		SVM		3	NA	NA	94.5
[53]	ECG signal	SVM-L	[205]	10	92.9	100	96.5
[2]	SpO2	SVM	Own	7	NA	NA	90
[117]	ECG	Additive logistic regression	Own		88.71	82.86	
[10]	SpO2	LDA	Own	19	75.6	91	86.5
mRMR[195] (Section 5.2)	SpO2	SVM-L	[61]	50	83.76	97.03	96.89
SFS[195] (Section 5.2)	SpO2	SVM-L	[61]	20	84.57	97.28	97.38
Proposed[102]	SpO2	ANN	[61]	7	96.5	98.5	97.7

The features selected are Variance (F1), Root mean square (F2), 'filter20' (F3), Daubechies 3 wavelet's Shannon entropy of approximation coefficients level 6 (CA6) (F4), Standard deviation detail coefficients of level 6 (CD6) (F5), variances and standard deviation detail coefficients level 3 (CD3) (F6, F7). The variance and 'filter20' are quite dominating features and these were also selected using 50 repeated random sub-sampling validations methods with a linear discriminant analysis (LDA) classifier and a different database[10]. Other features that are implemented in this work and selected by the GA algorithm are the wavelet base. In wavelet transform, the signal or the approximation coefficients at each level are filtered and create approximation coefficients and detail coefficients. The sampling frequency of 50 Hz is so that the highest frequency of the original SpO2 signal can be 25 Hz. Keeping that in mind, the selected scales have frequency band of 0 Hz-0.3907 Hz (CA6), 0.3907 Hz-0.7813 Hz (CD6) and 3.125 Hz-6.25 Hz (CD3). From Figure 26, it can be seen that all the features chosen by the GA have a significance difference between apnea and non-apnea events.

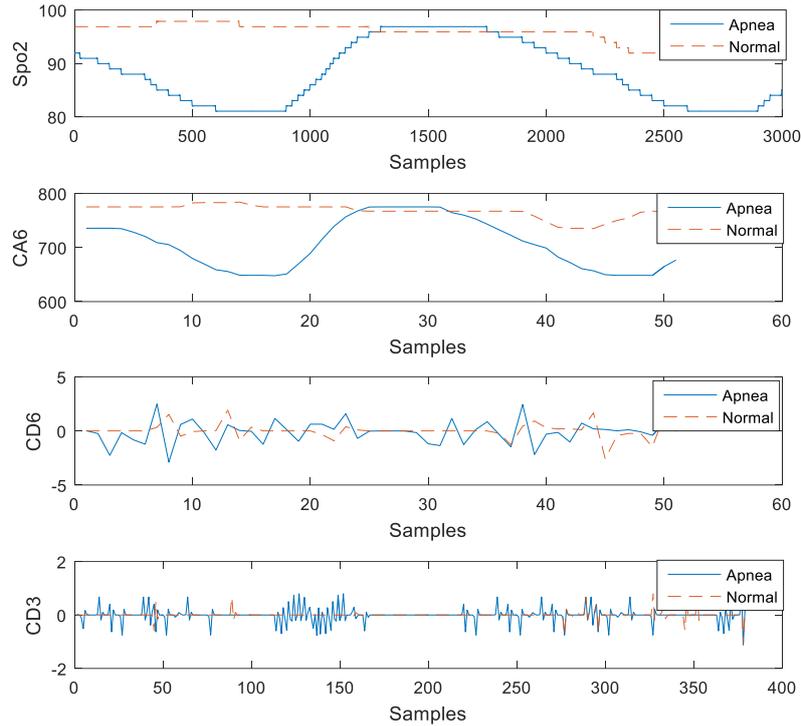


Figure 26 : Non-apnea (Normal) and apnea minute with the wavelet scale coefficients chosen by a GA ANN classifier.

5.3.4. Summary

In this section, a successful optimization of sleep Apnea detection is carried out. It was shown that the accuracy achieved by combining the ANN with the GA is good. The features selected by the GA algorithm are mostly time-frequency (four among seven). Only two time features and one frequency feature are selected. From this selection procedure, it is understandable that apnea events have most of the information in time-frequency space. Compared to previous work it improves the accuracy. In the future, this work can be tested on different databases and classifiers.

A limitation of the methods used here is that the structure of the selected MLP is calculated using a rule of thumb and experimental search which can be replaced, for example, by a GA. Furthermore, the experiment is not subject independent. These ideas are to be held under consideration in future work.

5.4. Self-Configuring Classifier Combination (SC3)

5.4.1. Introduction

The conventional approach for event classification is based on a single classifier, such as an SVM [2] [195], a ANN [99], and an LDA [10]. However, Xie et. al. [73] verified that a combination of classifiers can enhance the performance of the model when compared to the performance of individual classifiers. Therefore, a classification combination technique was proposed and employed

in this work to address the issue of OSA detection from a single signal, addressing the classical issues associated with this approach, specifically the selection of the most relevant classifiers and the features for each classifier. Hence, an automated algorithm to choose the feature and construct a combined classifier without human intervention was developed for this work.

The classification was performed with the combination of different classifiers outputs to create a single output. Considering the classifiers used in the state of the art [87], seven classifiers were tested: a ANN (the number of neurons in the hidden layer was selected to be $2n_i + 1$ [66] [195] [102], where n_i is the number of inputs); an SVM with a linear kernel (SVM-L); an SVM with a Gaussian Radial Basis Function kernel (SVM-RBF); an SVM with a polynomial kernel (SVM-P); a k-Nearest Neighbor (KNN); an LDA; and a Naive Bayes Classifier (NB).

Two combination strategies were tested. The first one was Maximum Voting (MaxV), and the other was a weighted method. The first combination strategy employed three classifiers since three is the lowest number needed to break a tie in a majority voting. The same number of classifiers was also used in the weighted method for a fair comparison.

5.4.2. *Self-Configuring Classifier Combination method*

A parallel combination method was employed as a combination strategy where each classifier was generated and trained in parallel. Also, there are homogeneous learners (using only one type of classifier) selected by the GA, providing a greater margin for optimization, and heterogeneous learners (using multiple types of classifiers). Though the system is designed for heterogeneous learners, the algorithm does not prohibit homogeneous learners, however. A block diagram of the Self-Configuring Classifier Combination (SC3, can be read as ‘S’ triple ‘C’) system is shown in Figure 27.

Two input features selection strategies were implemented. The first one was classifier dependent, where each classifier was allowed to choose its own feature set independently (shown using the dashed line in Figure 27). The second one, shared features among classifiers ($F_n = F_a$) (shown using the dotted line in Figure 27). Since a set of 61 individual features was used in this work, the Shared Feature (SF) selection methods have to choose from 61 features. Conversely, the Independent Feature (IF) selection method can select 61 features (Table 6) for each classifier (the method employed 3 classifiers in each configuration), allowing 183 (61×3) possible choices.

A GA was employed to solve the optimization problem. Each chromosome had 61 or 183 bits for feature selection depending, on the feature selection method. A total of 7 classifiers were tested and they were represented by 3 bits variables. For the weighted method, three sets of 7 bits were used for

the weight of each classifier. Therefore, a MaxV SF needs 70 ($61+3\times 3$) bits and the MaxV IF needs 192 ($61\times 3+3\times 3$) bits to represent the problem space. On the other hand, a weighted linear combination (WLC) SF classifier method needs 91 ($61+3\times 3+7\times 3$) bits, and the WLC IF one needs 213 ($61\times 3+3\times 3+7\times 3$) bits.

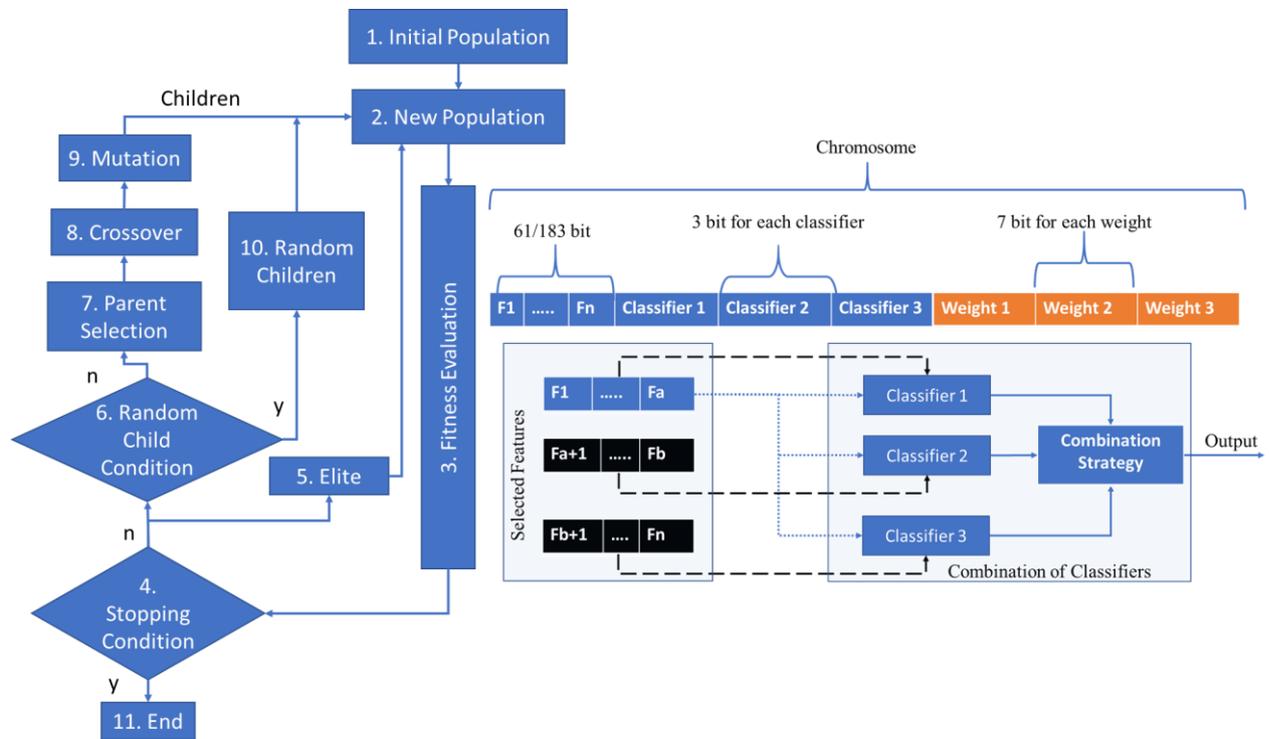


Figure 27 : Block diagram of the self-configuring classifier combination (SC3) technique.

The objective of the SC3 was to design a classification process by reducing the Cost Function (CF). This model flowchart considers:

1. Initial population: The GA started with an initial population with random chromosomes. For the first-generation, the initial population was treated as a new population. The population of each implementation was equal to the number of problem space bit size (chromosome size).
2. New population: For the first generation the new population was equal to the initial population. For other generations, the new population was the combination of the elite and the new children.
3. Fitness evaluation: The outputs of the classifiers are combined using two combination strategies: Majority Voting and Weighted Method independently described below.

Majority Voting: MaxV reaches a decision using a voting method. If the majority of the classifiers vote for one output the system chooses that output to be the system output. MaxV combines the classifiers by

$$\text{Out} = \begin{cases} 1 & \text{if } \sum_{i=1}^M O_i(x) \geq \lfloor M/2 \rfloor + 1 \\ 0 & \text{else} \end{cases} \quad (6)$$

where x is the input, and $\text{Out}_i \in \{0,1\}$ is the crisp output of i^{th} classifier. This method provides an accurate class label when at least $\lfloor M/2 \rfloor + 1$ classifiers give correct classifications. For this work M was selected to be 3 [207].

Weighted Method: Fumera et al. [208] showed that a linear combination of the weighted averaging of classifier's output outperforms a conventional averaging process. Therefore, a WLC was also used to combine the classifiers [209]. However, finding the optimal weights for each classifier is still an open problem [208] [210]. In this work, the optimal weights for each classifier were found by making the weights as a part of the problem variables that were optimized by the GA. The final output of WLC is

$$\text{Out} = \begin{cases} 1 & \text{if } \sum_{i=1}^M \frac{w_i}{\sum_{j=1}^M w_j} O_i(x) \geq 0.5 \\ 0 & \text{else} \end{cases} \quad (7)$$

where w_i is the weight value of the i^{th} classifier, ranging between 0 to 127 assuming that these numbers were enough for the weight resolution and that these numbers could be translated to 7 bits. The algorithm has $M = 3$ classifiers. Therefore, each classifier needs its own 7 bits to represent the weights which results in 21 more bits to represent the optimization problem in the GA compared to MaxV.

The output of the combined classifiers was used for performance and the fitness evolution. Due to the previous success of the combination objective (CO) in the case of unbalance datasets [106] [211] a CO was used for the fitness evaluation where CO is the average of the accuracy (Acc), sensitivity (Sen) and specificity (Spc).

$$CO = \frac{1}{3} (Acc + Sen + Spc) \quad (8)$$

Therefore, the cost function (CF) was derived from CO . The CF used in this work was

$$CF = 100 - CO \quad (9)$$

The fitness evaluation was responsible for training and testing the designed classifier. The data were divided into two sets for a two-fold method. For the fitness evaluation, the average of the two-fold CF was used.

4. Stopping condition: If 4 or more consecutive generations of random children were created according to step 6 or the maximum number of generations (100) was reached, the algorithm stops.
5. Elite: The best performing chromosomes who were elites in the previous generation were passed on to the next generation without mutation or crossover. For this work, the two best members of the population compose the elite.
6. Random children condition: If the average evaluation parameter of the generation and best evaluation parameter had a difference less than or equal to 0.1 percent, then instead of

crossover or mutation, a set of random children was created. Therefore, diversity was introduced in the generation.

7. Parent selection: A tournament selection process was used to select both parents. In this work, a tournament size of 10 was used.
8. Crossover: Crossover, sometimes called recombination, was used to produce children. A single point crossover was used for this work with a probability of 0.9.
9. Mutation: The genetic diversity of the chromosomes was created by a mutation which alters the values of the chromosomes. The probability of the mutation used in this work was 0.1.
10. Random Children: Random children generation was similar to the initial population where the chromosomes were created randomly.
11. End: When one of the termination conditions occurs, the algorithm stops.

5.4.3. Performance of SC3

To understand the effectiveness of the classifier combination, a commonly used single classifier (LDA) [87] with a GA based feature selection is estimated using the HuGCDN2008 database. The Acc, Sen, and Spc of each generation are presented in Figure 28, while the optimized cost function and the number of the features are shown in Figure 29. For the LDA, the best *CF* at the termination (after 21 generations) was 18.06%, 18.22%, and 19.17% for, 1 minute, 3 minute and 5 minute, respectively. Both 1 minute and 5 minute solutions have 37 features, and their Acc (85.58% for 1 minute, 85.34% for 3 minute), Sen (68.91% for 1 minute, 68.99% for 3 minute) and Spc (91.35% for 1 minute, 91.02% for 3 minute) are similar. On the other hand, the 5 minute solution has 34 features with increased Spc (92.46%), decreased Sen (64.74%) and similar Acc (85.29%). From these results it is possible to conclude that the increased input size does not have a significant positive impact on the results. The best *CF* of the LDA based solution was 18.06% with a 1 minute input size.

The SC3 technique was implemented using the HuGCDN2008 database. A subject independent two-fold method was applied to avoid bias. Two strategies, MaxV and WLC, were tested with IF for each classifier and SF among the classifiers. The effect of the input size (signal length) was studied for three scenarios, specifically, the same features with 1 minute (3000 samples), 3 minutes (9000 samples), and 5 minutes (15000 samples). The self-configuring classifier combination technique reached the stopping criteria at 21 generations (Figure 30, Figure 28) and the final results are summarized in Table 17. From the HuGCDN2008 dataset, each combined classifier has two trained classifiers because of the two-fold method. These two classifiers were tested on the AED dataset, and the average results are presented in Table 18. Every solution of the proposed self-configuring classifier combination is better than the LDA respect to *CF*. The self-configuring classifier technique achieved a similar Acc when compared with the single classifier system; however, the SC3 results are

more balanced. The main achievement of the SC3 is increased sensitivity (apnea events) without sacrificing the accuracy of the system, thus providing a relevant model for clinical analysis.

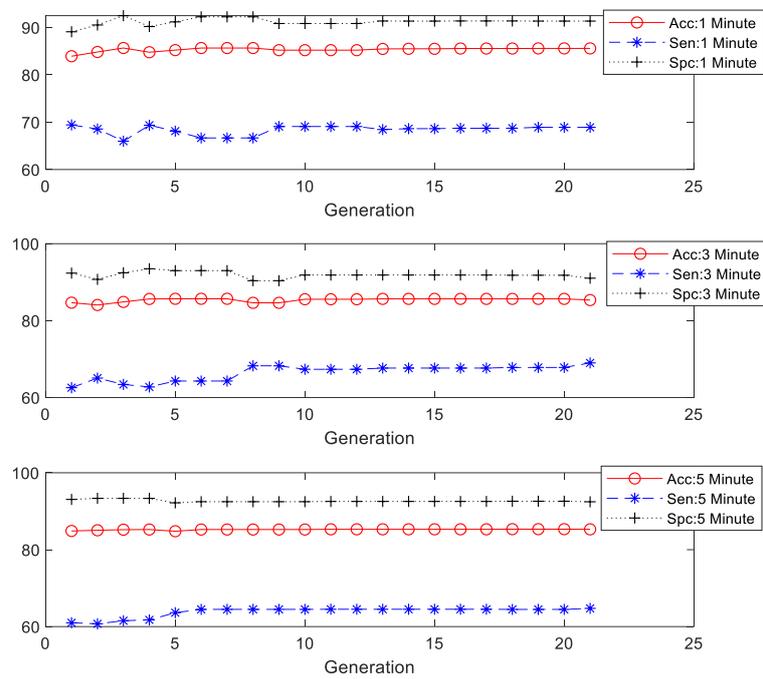


Figure 28 : Accuracy (Acc), Sensitivity (Sen), and Specificity (Spc) of 1 minute, 3 minute and 5 minute LDA over the generations for the best performance objective.

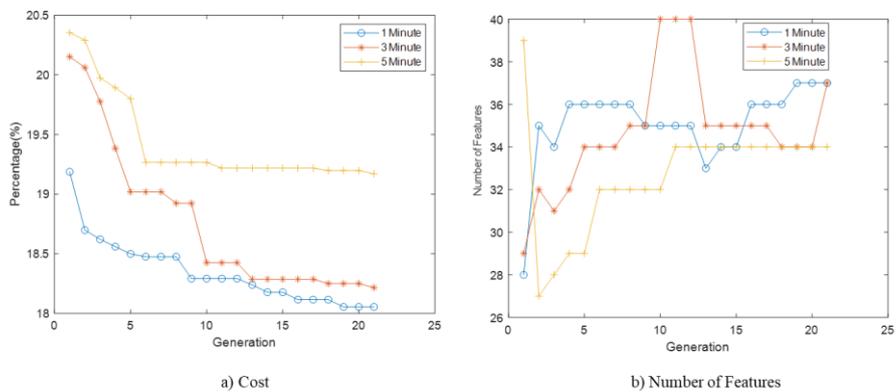


Figure 29 : a) Cost and b) Number of features of 1 minute, 3 minute and 5 minute LDA over the generations for the best performance objective.

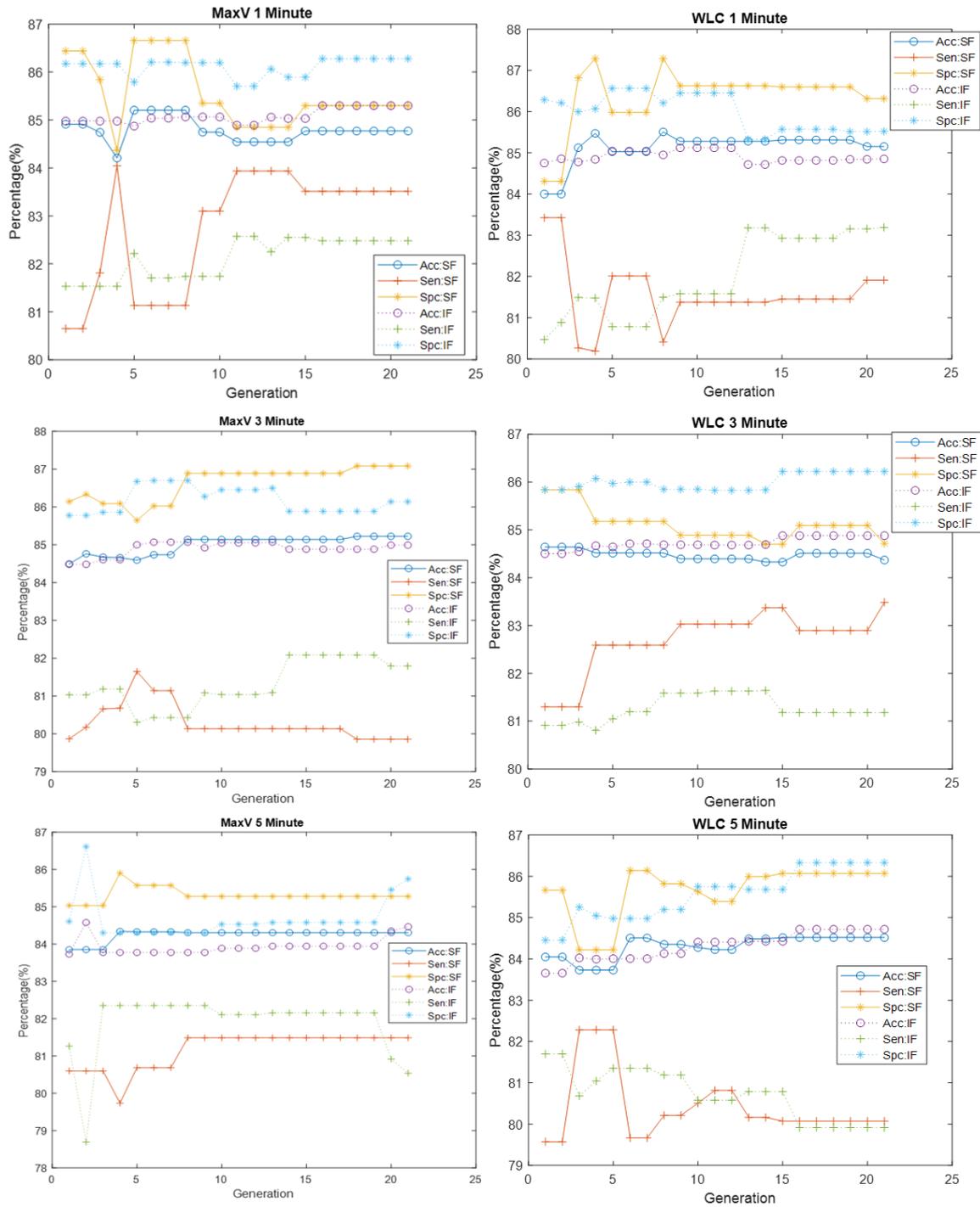


Figure 30 : Accuracy (Acc), Sensitivity (Sen), and Specificity (Sp) of 1 minute, 3 minute and 5 minute MaxV and WLC over the generations for the best performance objective.

Table 17 : Self configuring classifier combination results for the HuGCDN2008 database using two fold cross validation for different inputs.

Classifier	Input size (Seconds)	No of features	Sen	Sp	Acc	CF
MaxV SF1	60	34	83.51	85.30	84.77	15.47
MaxV SF3	180	27	79.86	87.08	85.22	15.95
MaxV SF5	300	34	81.49	85.28	84.3	16.31
MaxV IF1	60	98	82.48	86.28	85.30	15.31

MaxVIF3	180	90	81.79	86.14	84.99	15.69
MaxVIF5	300	89	80.54	85.74	84.46	16.42
WLCSF1	60	34	81.91	86.31	85.15	15.54
WLCSF3	180	29	83.49	84.71	84.37	15.81
WLCSF5	300	30	80.07	86.07	84.52	16.45
WLCIF1	60	96	83.19	85.52	84.85	15.48
WLCIF3	180	99	81.18	86.22	84.88	15.91
WLCIF5	300	89	86.33	79.91	84.72	16.35
LDA	60	37	68.91	91.35	85.58	18.06
LDA	180	37	68.99	91.02	85.34	18.22
LDA	300	34	64.74	92.46	85.29	19.17

Table 18 : Self configuring classifier combination results for cross database (trained with HuGCDN2008 and tested with AED) comparison for different inputs.

Classifier	Input size (Seconds)	No of features	Sen	Spc	Acc	CF
MaxVSF1	60	34	98.11	86.98	91.33	7.86
MaxVSF3	180	27	96.83	86.94	90.92	8.44
MaxVSF5	300	34	97.58	68.31	80.09	18
MaxVIF1	60	98	95.02	88.35	90.95	8.56
MaxVIF3	180	90	95.41	84.66	88.99	10.31
MaxVIF5	300	89	79.56	86.17	83.51	16.92
WLCSF1	60	34	97.39	86.50	90.75	8.45
WLCSF3	180	29	97.31	83.08	88.81	10.27
WLCSF5	300	30	96.71	87.35	91.12	8.27
WLCIF1	60	96	96.71	87.73	91.23	8.11
WLCIF3	180	99	94.23	84.56	88.45	10.92
WLCIF5	300	89	92.27	87.88	89.65	10.07

5.4.4. Comparison of SF and IF for MaxV SC3

The MaxV classification technique considers the maximum number of outputs of the classifiers. Thus, if two or more classifiers agree on an output, it produces that output as the final output. All of the simulations terminate in the 21st generation because of the stopping condition for the HuGCDN2008 database (Figure 30).

Among the six variations of the developed self-configuring classifier combination, the 1 minute IF had the lowest *CF* of 15.31%, whereas 5 minute IF had the highest *CF* of 16.42% (Figure 31 a). The MaxV SF (MaxVSF) with three different inputs produced three different classifier combinations for 1 minute (SVM_L, KNN, and SVM_P), 3 minute (SVM_L, NB, and LD) and 5 minute (SVM_L, SVM_P, and LD) (Table 19). The MaxVSF 1 minute (MaxVSF1), the MaxVSF 3 minute (MaxVSF3) and the MaxVSF 5 minute (MaxVSF5) stopped in 21 generations with a *CF* of 15.47%, 15.95% and 16.31%, respectively. All MaxVSF have an SVM_L classifier. Following the same naming rule as MaxVSF1, MaxVIF1 (SVM_L, LD, and SVM_P), MaxVIF3 (SVM_P, SVM_L, and LD), and MaxVIF5 (SVM_RBF, KNN, and LD) achieved the goal with a *CF* of 15.31%, 15.69%, and 16.42%, respectively. The Acc, Sen, and Spc of each SC3 is shown in Figure 30.

Figure 31 shows the cost function of the best solution for each generation of the MaxV implementation. By analyzing the figure, it is noticeable that both the shared and independent features with 1 minute achieved the best performance, followed by the 3 minute and 5 minute performance. In

the beginning, the best CF for 1 min was 15.77% for IF and 16% for SF. In the case of 3 minute, the starting difference was 0.25% where IF was 16.25% and SF was 16.5%. For 5 minute the starting difference was almost the same, namely 16.8% (for IF) and 16.84% (for SF).

Through all of the simulations, the SC3 was able to keep the cost function stable or reduce it in each generation. This stability was possible by the elites where the two best solutions from the previous generation were carried onto the next generation without any modification. In the case of the three minute window, the IF solution was always better than the shared one. On the other hand, for 1 minute window, the role of the best cost function switched between the IF and the SF. However, in the end, the IF performed better than the SF but this observation was not true for the 5 minute. After generation 5, the IF was never able to surpass SF. This result contradicts the theory that the best performance was obtained by combining different feature sets with different individual classifiers [212] [210]. Theoretically, it is simple to understand that the SF system is a subset of the IF system. There is always a solution in the IF system which can mirror the SF solution. These results might occur due to the number of generations that were used by the models since the IF system has more variables to decide the solution which makes the problem more challenging to solve.

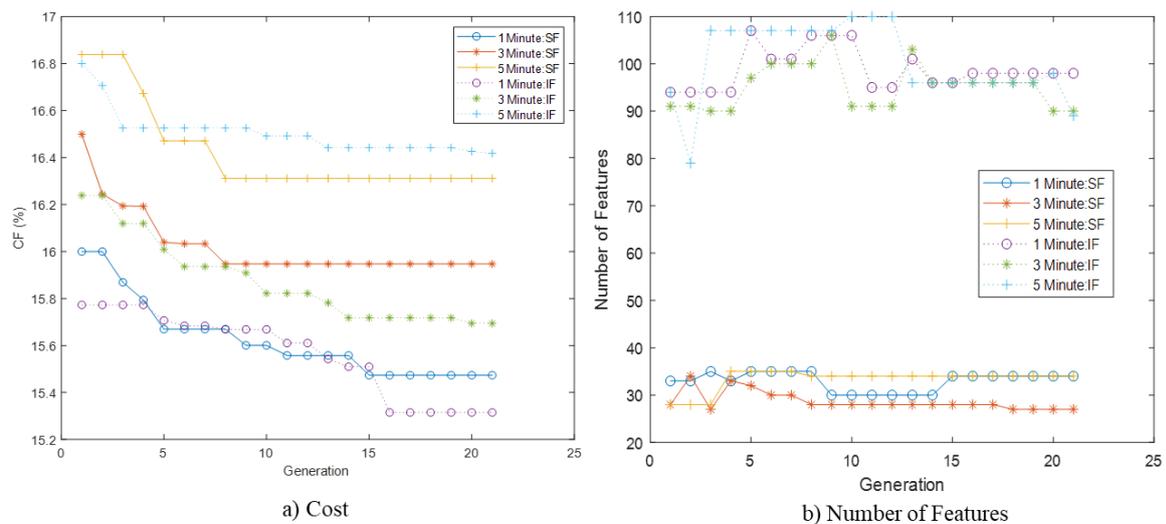


Figure 31 : a) Cost and b) Number of features of 1 minute, 3 minute and 5 minute MaxV Independent Feature (IF) and Shared Feature (SF) classification combination over the generations for the best performance objective.

Figure 31 b) shows the number of features used by each variation of the SC3. The figure indicates the total number of features. The IF solutions have more features than the SF ones. The lowest number of features (27 features) was selected by the 3 minute SF while both the 1 minute and 5 minute SF used 34 features to achieve their CF . On the other hand, all of the IF solutions have almost 3 times more features than the SF. 1 minute, 3 minute, and 5 minute IF solutions, the SC3 used 98, 90 and 89 features, respectively.

5.4.5. Comparison of SF and IF for WLC SC3

Instead of counting the outputs of the classifiers directly, the WLC multiplies the crisp output of each classifier by a weight, and if the weighted sum of the output favors one output over the others then that output is chosen. For WLCSF1, the combination of classifiers was SVM_P, LD, and SVM_L; for WLCSF3 the combination of classifiers was NB, SVM_L, and KNN and for WLCSF5 the combination of classifiers was SVM_L, LD, and SVM_P. In comparison to these WLCIF1 used LD, SVM_L, and SVM_P as classifiers; WLCIF3 used LD, SVM_L, and SVM_L as classifiers; and WLCIF5 used LD, SVM_L, and SVM_P as classifiers to create the final classification combination. The Acc, Sen, and Spc are presented for the respective SC3.

As can be seen in Figure 32, except for the 5 minutes WLC, the other four SC3 start the generation between 16.07% and 16.25% *CF*. In the end, the 1 minute WLCIF had the lowest *CF* of 15.48% (best), followed by 15.81% of the 3 minute WLCSF in second place. The WLCIF methods achieved a better solution than the WLCSF for the 1 minute and 5 minute cases. However, for the 3 minute window the WLCSF attained the best results (Figure 32 a).

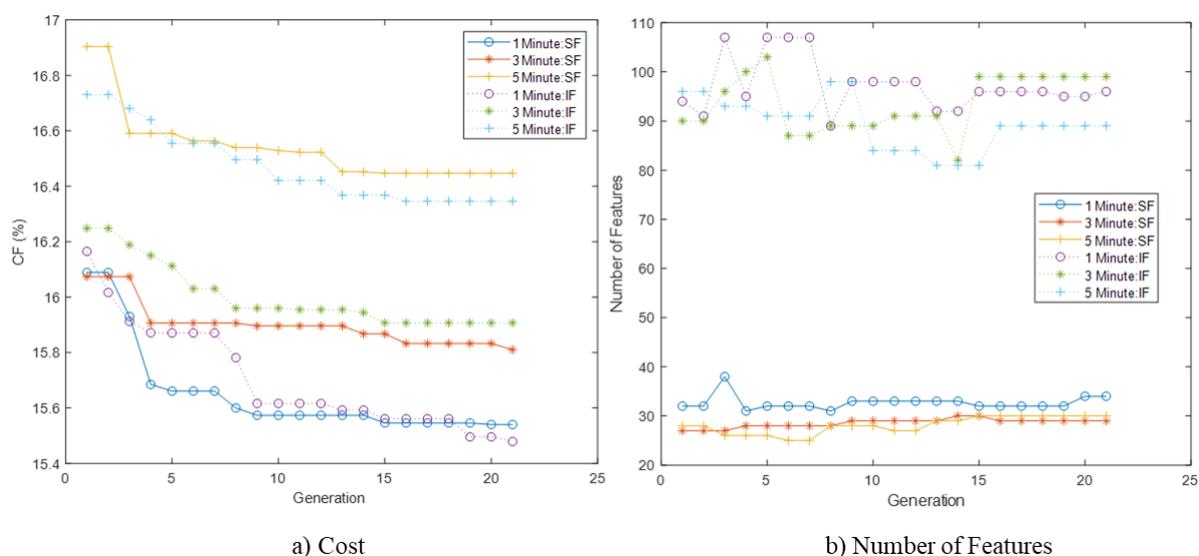


Figure 32 : a) Cost b) Number of features of 1 minute, 3 minute and 5 minute WLCSF and WLCIF classification combination over the generations for the best performance objective.

The WLCIF had a higher number of features in all generations (Figure 32 b). The lowest number of features (29) was used by the 3 minute WLCSF while the highest number of features (99) was used by the 3 minute WLCIF. The WLCSF5 used 30 features while the 1 minute window used 34. The WLCIF with the 5 minute' window used 89 features and the 1 minute window used 96 features.

5.4.6. Comparison Between MaxV and WLC

The number of bits needed for the WLC was always higher when compared to the MaxV in the SC3. The weight of each classifier is responsible for the extra bits in the WLC. In this work, three classifiers were used with 7 bits to represent the weights. Thus, for a similar problem, the WLC had 21 more bits for each chromosome than the MaxV.

The best *CF* (15.31%) was accomplished by MaxVIF1, which also attained the highest Acc (85.30%) even though it does not have the highest Sen (86.33%) or Spc (87.08%) which were attained by WLCIF5 and MaxVSF3, respectively. MaxVIF1 has a balanced performance, with 82.48% Sen and 86.28% Spc. For the SF, MaxVSF1 achieved the best *CF*, and in the case of IF, MaxVIF1 has the best *CF*. The lowest number of features and SF was achieved by MaxVSF3, while for IF, both MaxVIF5 and WLCIF5 (used 89 features each, please refer to Table 19).

Table 19: Selected Features and Classifiers for Different SC3.

Strategy (Num of Features)	Classifiers (Num of Features)	Features
MaxVSF1 (34)	SVM_L	Avg, Var, CoV, Sk, Kurt, Min, REp, Fb1, Fb3, Fb5, Fb7, Fb8, Fb10, Fb11, Fb12, Fb14, Fb15, Fb16, TFVarCA6, TFMadCA6, TFVarCD6, TFSdCD6, TFMadCD6, TFSdCD5, TFSEnCD4, TFVarCD4, TFVarCD3, TFMadCD3, TFSdCd2, TFSEnCD1, TFVaeCD1, TFSdCD1, CTM50, ODI3
	KNN	
	SVM_P	
MaxVSF3 (27)	SVM_L	Var, Kurt, RMS, Min, SEn, Fb9, Fb12, Fb15, Fb17, Fb18, TFSEnCA6, TFVaeCA6, TFMadCA6, TFMadCD6, TFVarCD5, TFSEnCD4, TFVarCD4, TFSdCD4, TFMadCD4, TFSEnCD3, TFVarCD3, TFMadCD3, TFSdCD2, TFMadCD2, TFVarCD1, TFMadCD1, ODI3
	NB	
	LD	
MaxVSF5 (34)	SVM_L	Avg, Var, Kurt, RMS, Max, SEn, Fb2, Fb3, Fb4, Fb8, Fb9, Fb10, Fb12, Fb13, Fb14, Fb17, Fb18, Fb19, TFSEnCA6, TFVarCA6, TFSdCA6, TFVarCD6, TFSdCD5, TFSEnCD5, TFMadCD4, TFVarCD3, TFSdCD3, TFMadCD3, TFSEnCD2, TFVarCD2, TFSdCD2, TFSEnCD1, CTM50, DIndex
	SVM_P	
	LD	
WLC5F1 (34)	SVM_P	Var, CoV, Kurt, RMS, REp, SEn, Fb1, Fb3, Fb6, Fb7, Fb10, Fb14, Fb15, Fb17, Fb18, Fb19, TFSEnCD6, TFVarCA6, TFMadCA6, TFSdCD6, TFMadCD6, TFSEnCD5, TFSdCD5, TFSEnCD4, TFVarCD4, TFSdCD4, TFSEnCD3, TFVarCD3, TFMadCD3, TFVarCD2, TFSdCD2, TFSEnCD1, TFVarCD1, DIndex
	LD	
	SVM_L	
WLC5F3 (29)	NB	Kurt, RMS, Fb1, Fb2, Fb4, Fb5, Fb6, Fb7, Fb8, Fb10, Fb11, Fb12, TFVarCA6, TFSdCA6, TFMadCA6, TFSEnCD6, TFVarCD6, TFMadCD6, TFSEnCD5, TFMadCD5, TFSdCD4, TFSdCD3, TFMadCD3, TFSEnCD2, TFVarCD2, TFCVarD1, CTM50, ODI3, DIndex
	SVM_L	
	KNN	
WLC5F5 (30)	SVM_L	Var, Kurt, RMS, Min, Ren, Fb2, Fb3, Fb7, Fb10, Fb11, Fb12, Fb20, TFSEnCA6, TFVarCA6, TFSdCD6, TFSEnCD5, TFSdCD5, TFSEnCD4, TFVarCD4, TFSdCD4, TFSEnCD3, TFVarCD3, TFMadCD3, TFSEnCD2, TFVarCD2, TFMadCD2, TFVarCD1, TFMadCD1, CTM50, DIndex
	NB	
	ANN	
MaxVIF1 (98)	SVM_L (39)	Avg, Sk, Kurt, RMS, Max, REp, Fb1, Fb2, Fb3, Fb4, Fb5, Fb6, Fb7, Fb9, Fb10, Fb11, Fb12, Fb13, Fb14, F18, Fb19, Fb20, TFSEnCA6, TFVarCA6, TFMadCA6, TFSEnCD6, TFMadCD6, TFMadCD5, TFSEnCD3, TFVarCD3, TFSdCD3, TFSEnCD2, TFVarCD2, TFSdCD2, TFSdCD1, TFMadCD1, CTM50, ODI3, DIndex
	LD (25)	Var, CoV, RMS, Max, Min, REp, SEn, Fb3, Fb7, Fb10, Fb11, Fb13, Fb16, Fb17, Fb20, TFVarCA6, TFSdCD6, TFMadCD5, TFVarCD4, TFMadCD4, TFSEnCD3, TFVarCD3, TFSdCD3, TFSEnCD2, TFSdCD1
	SVM_P (34)	Avg, Kurt, RMS, Max, Min, REp, Fb1, Fb4, Fb10, Fb12, Fb13, Fb14, Fb16, Fb19, Fb20, SEnCA6, TFCa6, TFSdCA6, TFVarCD6, TFSdCD6, TFSEnCD5, TFVarCD5, TFSdCD5, TFMadCD5, TFVarCD4, TFMadCD4, TFSEnCD3, TFVarCD3, TFSdCD3, TFMadCD3, TFSEnCD2, TFVarCD1, TFSdCD1, DIndex
MaxVIF3 (90)	SVM_P (32)	Kurt, RMS, Min, REp, SEn, Fb1, Fb5, Fb6, Fb7, Fb8, Fb12, Fb13, Fb15, Fb16, Fb17, Fb18, Fb19, TFSEnCA6, TFVarCA6, TFSEnCD6, TFVarCD6, TFSEnCA5, TFSdCA5, TFMadCA5, TFSdCD4, TFMadCD4, TFSEnCA3, TFVarCD3, TFSdCD2, TFSEnCD1, TFSdCD1, ODI3
	SVM_L (27)	Var, Kurt, REp, SEn, Fb2, Fb8, Fb9, Fb11, Fb12, Fb16, Fb19, TFSdCA6, TFSEnCD6, TFVarCD6, TFSdCD6, TFMadCD6, TFVarCA5, TFMadCA5, TFVarCD4, TFSdCD4, TFSEnCA3, TFSEnCD2, TFSdCD2, TFVarCD1, TFSdCD1, ODI3, DIndex,
	LD (31)	Var, CoV, Sk, Kurt, Max, Min, REp, SEn, Fb4, Fb5, Fb7, Fb8, TFSEnCA6, TFVarCA6, TFSdCA6, TFMadCA6, TFSdCD6, TFVarCA5, TFMadCA5, TFVarCD4, TFSdCD4, TFMadCD4, TFSdCD3, TFVarCD2, TFSdCD2, TFMadCD2, TFSEnCD1, TFVarCD1,

		TFSdCD1, ODI3, DIndex
	SVM_RBF (29)	Avg, Var, CoV, RMS, Max, REn, SEn, Fb1, Fb4, Fb5, Fb7, Fb9, TFSEnCA6, TFVarCA6, TFMadCA6, TFSEnCD6, TFVarCD6, TFVarCD4, TFSdCD4, TFMadCD4, TFSEnCA3, TFSdCD3, TFMadCD3, TFSEnCD2, TFVarCD1, TFSdCD1, TFMadCD1, CTM50, DIndex
MaxVIF5 (89)	KNN (31)	Kurt, RMS, Min, Fb1, Fb2, Fb3, Fb5, Fb6, Fb, Fb10, Fb15, Fb16, Fb17, Fb19, Fb20, TFSdCA6, TFMadCA6, TFSEnCD6, TFVarCD6, TFMadCD6, TFSEnCA5, TFMadCA5, TFSEnCD4, TFMadCD4, TFSEnCA3, TFVarCD3, TFSdCD3, TFSEnCD2, TFMadCD2, TFVarCD1, TFSdCD1
	LD (29)	Avg, Var, Kurt, Max, Min, REn, SEn, Fb1, Fb7, Fb8, Fb9, Fb11, Fb12, Fb13, Fb17, Fb19, TFSEnCA6, TFSdCD6, TFMadCD6, TFSEnCA5, TFSdCA5, TFMadCA5, TFSdCD3, TFMadCD3, TFMadCD2, TFSEnCD1, TFVarCD1, CTM50, DIndex
	LD (36)	Var, CoV, RMS, REn, SEn, Fb1, Fb6, Fb7, Fb8, Fb9, Fb10, Fb13, Fb14, Fb16, Fb17, Fb18, Fb19, TFSEnCA6, TFVarCA6, TFVarCD6, TFSdCD6, TFMadCD6, TFSEnCA5, TFVarCA5, TFSdCA5, TFMadCA5, TFVarCD4, TFSdCD4, TFMadCD4, TFSdCD3, TFSEnCD2, TFVarCD2, TFVarCD1, TFSdCD1, CTM50, DIndex
WLCIF1 (96)	SVM_L (30)	Avg, Sk, Kurt, RMS, Max, REn, Fb2, Fb4, Fb6, Fb10, Fb11, Fb13, Fb16, Fb19, TFVarCA6, TFSdCA6, TFSEnCD6, TFVarCD6, TFSdCD6, TFMadCD6, TFSEnCA5, TFVarCA5, TFSdCA5, TFVarCD4, TFSdCD4, TFMadCD4, TFMadCD3, TFSdCD2, CTM50, ODI3
	SVM_P (30)	Avg, CoV, RMS, Fb1, Fb4, Fb5, Fb8, Fb13, Fb14, Fb15, Fb16, Fb17, Fb20, TFSEnCA6, TFVarCA6, TFMadCA6, TFSdCD6, TFMadCD6, TFSEnCA5, TFVarCA5, TFSdCA5, TFVarCD4, TFVarCD3, TFSdCD3, TFMadCD3, TFSEnCD2, TFVarCD1, TFSdCD1, TFMadCD1, ODI3
	LD (31)	Avg, CoV, Sk, SEn, Fb1, Fb4, Fb5, Fb6, Fb7, Fb12, Fb16, Fb19, Fb20, TFSEnCA6, TFSdCA6, TFMadCA6, TFSEnCD6, TFVarCD6, TFMadCD6, TFSEnCA5, TFSdCA5, TFVarCD4, TFMadCD4, TFSEnCA3, TFSdCD3, TFMadCD3, TFSEnCD2, TFVarCD2, TFSEnCD1, ODI3, DIndex
WLCIF3 (99)	SVM_L (32)	Avg, CoV, Kurt, RMS, Max, REn, SEn, Fb1, Fb3, Fb7, Fb8, Fb10, Fb11, Fb16, Fb19, Fb20, TFSEnCA6, TFVarCA6, TFSEnCD6, TFSdCD6, TFMadCD6, TFSEnCA5, TFMadCA5, TFSEnCD4, TFVarCD4, TFSdCD4, TFSEnCA3, TFVarCD3, TFVarCD2, TFMadCD1, ODI3, DIndex
	SVM_L (36)	Var, CoV, Sk, RMS, Max, Min, SEn, Fb2, Fb3, Fb5, Fb9, Fb10, Fb11, Fb12, Fb13, Fb14, Fb15, Fb19, Fb20, TFVarCA6, TFSEnCD6, TFMadCD6, TFSEnCA5, TFSdCA5, TFSEnCD4, TFSdCD4, TFVarCD3, TFSdCD3, TFMadCD3, TFSEnCD2, TFVarCD2, TFSdCD2, TFMadCD2, TFMadCD1, CTM50, DIndex
	LD (29)	Avg, Var, RMS, Max, REn, SEn, Fb2, Fb3, Fb5, Fb6, Fb12, Fb14, Fb17, Fb18, Fb20, TFSEnCA6, TFMadCA6, TFSEnCD6, TFVarCD6, TFSdCD6, TFSEnCA5, TFSdCA5, TFMadCA5, TFSEnCD4, TFSEnCA3, TFMadCD3, TFMadCD2, CTM50, Dindex
WLCIF5 (89)	SVM_L (27)	Var, Kurt, Max, Min, Fb2, Fb6, Fb10, Fb12, Fb17, Fb18, TFSEnCA6, TFSdCA6, TFMadCA6, TFVarCD6, TFSdCD6, TFSEnCA5, TFVarCA5, TFVarCD4, TFSdCD4, TFMadCD4, TFSEnCA3, TFSdCD3, TFSEnCD2, TFMadCD2, TFVarCD1, CTM50, DIndex
	SVM_P (33)	Var, Sk, Kurt, Min, SEn, Fb2, Fb3, Fb6, Fb7, Fb8, Fb9, Fb10, Fb11, Fb13, Fb16, Fb18, Fb20, TFSEnCD6, TFVarCD6, TFMadCD6, TFSEnCD4, TFVarCD4, TFSdCD4, TFSEnCA3, TFVarCD3, TFSdCD3, TFVarCD2, TFSdCD2, TFVarCD1, TFSdCD1, TFMadCD1, CTM50, DIndex

Regarding the number of chosen features, SF 1 minute and IF 5 minute (the WLC and the MaxV) selected a similar number of features. On the other input sizes, the difference is small. The WLC has a higher accuracy in a SF 1 minute, a SF 5 minute and an IF 5 minute. The maximum difference between the two methods regarding the Acc is lower than 0.5% (for the IF) and 0.9% (for the SF). For the Spc, a maximum of 5.83% and 2.3% difference occurred in the IF and the SF, respectively. For the Sen, the values are 5.79% and 3.63%, respectively.

For further analysis, a side by side comparison of the subtracted performance parameter between the MaxV, and the WLC is presented in Figure 33. The negative values in the figure indicate that the WLC has a higher parametric value, while positive values indicate that the MaxV has a higher parametric value. There are 4 parametric values of three different input sizes; therefore, a total of twelve parametric values for SF and IF were compared. Among them, the MaxV has a better parametric value, in case SF 5 parameters and in case of IF 6 parameters are better, where in both cases 1 parameter is equal to both the SF and the IF. So overall, the MaxV performs better than the

WLC.

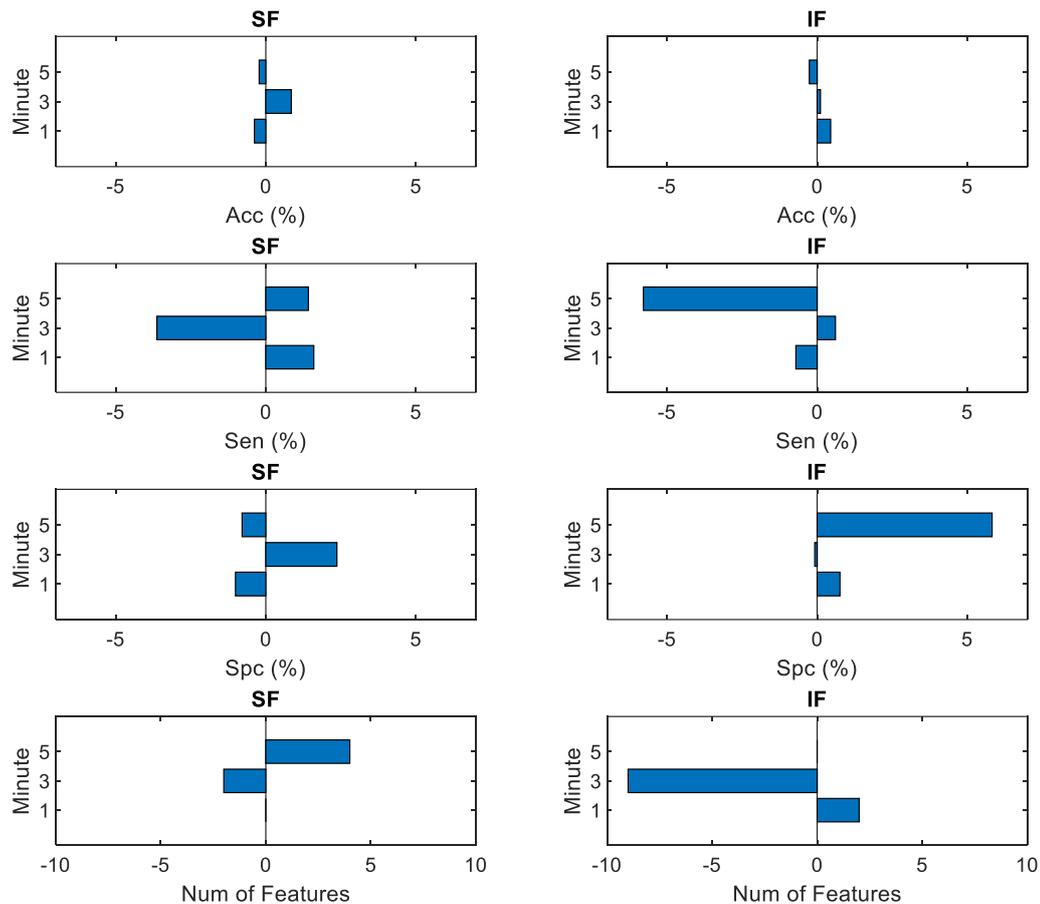


Figure 33 : A subtraction of the performance parameters between the MaxV and the WLC (MaxV-WLC) for the HuGCDN2008 database (Num indicates number).

5.4.7. Comparison with other work

The SC3 algorithm was implemented in the HuGCDN2008 database to find an optimized combined classifier without human intervention. Due to the unbalanced dataset, the cost function was chosen to balance the Acc, Sen and Spc. The best cost function value was achieved by MaxVIF1, which has an accuracy of 85.30%. This is higher than most of the performances reported in most of the works presented in the literature [185] [106] [195] [73] [190].

A comparison with other works is shown in Table 20. Two works presented in the literature, [2] [190], attained a better accuracy than MaxVIF1. Zhang et al. [2] used an SVM with seven features and it was tested on 40 patients conducted in a hospital in East Asia. Biswal et al. [190] used a combination of deep recurrent and convolutional neural networks in the Massachusetts General Hospital (MGH) dataset having better accuracy than the proposed system. Biswal et al. [190] worked in the Sleep Heart Health Study (SHHS) database, having a lower accuracy than the proposed system

and thus indicating a possible data dependency of the classification system. Other deep network implementations include a Deep Auto Encoder (DAE) [106] and a CNN with a two dimensional input (CNN2D) [185], which present lower accuracy than the proposed system. However, their works were implemented in the UCD database [62]. All of the classifiers using the AED [61] dataset presented a better accuracy than the proposed system which was tested on the HuGCDN2008 database (Table 20), thus supporting the database independency of the proposed method.

If a direct comparison is carried out then only one work is available in the literature using the HuGCDN2008 database with a SpO2 [10], reporting an average Acc of 86.5%. Using the SpO2 and the HRV, Acc was 86.9%. These results are 1.2% and 1.6% higher than MaxVIF1, respectively. However, the main accomplishment of the SC3 algorithm was the balanced results of MaxVIF1 where the Sen was 81.91% while the reported Sen by Ravelo-García et al. [10] was 75.6% (SpO2) and 73.4% (SpO2 and HRV).

The best value for the *CF* (8.56%) when considering the HuGCDN2008 dataset was achieved by MaxVIF1. For the AED dataset the best *CF* was 7.86% and it was achieved by MaxVSF1. It is also relevant to note that MaxVSF1 achieved an average Acc, Spc and Sen of 91.33%, 86.98% and 98.11%, respectively. Some works have employed deep learning techniques, such as LSTM [107] and DAE [106], which present the risk of overfitting the classifier due to the lower number of subjects in the AED database. On the contrary, shallow classifiers, such as SVM-L [195], are able to achieve a 97.38% accuracy. The best matching technique was used by Mostafa et al. [102] where the authors used the GA to find the best features having an accuracy of 97.7% which is the highest among the literature using the AED database. However, their implementation was not subject independent. This proposed implementation is not only subject independent but also database independent, as the AED database was an unseen dataset for the classifiers.

Table 20 : Comparison with other works.

Ref.	Signal	Classifier	Database	Recordings or Patients	Input size (Seconds)	No of features	Sen	Spc	Acc
[99]	SpO2	ANN	AED [61]	8	60	3	87.5	100	90.3
[195] (Section 5.2)	SpO2	SVM-L	AED [61]	8	60	50	83.76	97.03	96.89
[195] (Section 5.2)	SpO2	SVM-L	AED [61]	8	60	20	84.57	97.28	97.38
[102] (Section 5.3)	SpO2	ANN	AED [61]	8	60	7	96.5	98.5	97.7
[106]	SpO2	DAE	AED [61]	8	60	-	78.75	95.89	97.64
[107]	SpO2 + IHR	LSTM	AED [61]	8	60	-	84.7	-	92.1
[107]	SpO2	LSTM	AED [61]	8	60	-	92.9	-	95.5
[2]	SpO2	SVM	Own	40	150	7	-	-	90
[190]	SpO2 + airflow + respiration	RCNN	MGH	10000	1	-	-	-	88.2
[190]	SpO2 + airflow + respiration	RCNN	SHHS	5804	1	-	-	-	80.2
[73]	SpO2 + ECG	Bagging.RepTree	UCD [62]	25	60	39	79.75	85.89	84.80
[73]	SpO2	Bagging.RepTree	UCD [62]	25	60	39	78.23	84.25	82.79
[195] (Section 5.2)	SpO2	ANN	UCD [62]	25	60	2	43.31	95.03	81.95
[195]	SpO2	LD	UCD [62]	25	60	9	61.78	91.03	83.27

(Section 5.2)									
[106]	SpO2	DBN	UCD [62]	25	60	-	60.36	91.71	85.26
[185]	SpO2 + oronasal airflow + ribcage and abdomen movements	CNN2D	UCD[62]	23	1	-			79.6
[10]	Spo2	LDA	HuGCDN2008	70	60,300	19	75.6	91.00	86.5
[10]	Spo2 + HRV	LDA	HuGCDN2008	70	60,300	33	73.4	92.3	86.9
p	SpO2	MaxVIF1	HuGCDN2008	70	60	98	82.48	86.28	85.30
p	SpO2	MaxVSF1	AED [61]	8	60	34	98.11	86.98	91.33

5.4.8. Summary

The model proposed in this work employs a self-configuring classification combination method, which was able to choose the most relevant features and classifier structure automatically. Two well established methods, specifically the maximum voting and the positive weighting methods were used for classifier combination. The methods were tested in the models with shared features, for all classifiers, and independent features, for each classifier. It was verified that the maximum voting method with independent features for each classifier attained the best performance and all self-configuring classification combination models outperform the LDA based on a single classifier. The model was also able to achieve significant and well-balanced results, despite the unbalanced dataset, advocating the potential application in clinical diagnosis.

Another relevant fact is that the results were attained using both subject and database independence. The trained classifiers were able to detect the apnea event of different subjects (subject independence) from dissimilar datasets (database independence). The proposed system achieved similar performance on the two datasets that were analyzed, and it was verified that a longer input size, or data length, does not always improve the results (the best performance was achieved by using a 1 minute input). It was also verified that the difference between independent and share features is significant, however, the independent feature based self-configuring classification combination requires more bits to define the problem, hence increasing the complexity of the system.

A possible improvement in the results could be attained by changing the stopping criteria to run the simulation for a longer period to perceive if any improved solutions could be possible. Another alternative is to add another signal, such as HRV, to the models which is put into consideration for the future work along with the testing of different types of fusion and classifier ensembles (Section 5.5).

It was verified that the combined methods can perform better than a common classifier with a small increase in complexity. Conversely, the complexity of the combined classifiers is lower than a deep network and, unlike these networks, it does not require a large amount of data to train the model. This is especially relevant in domains where the data are scarce.

5.5. Combination of SpO2 and HR using SC3

5.5.1. Introduction

The obstruction or reduction of airflow normally decreases the blood oxygen saturation level which can be used as a marker for apnea detection. On the other hand, due to respiratory sinus arrhythmia [213] which is the modulation of the HR in respiration, the HR can be used for apnea detection (Figure 34). Additionally, in some cases, a noticeable reduction in the partial oxygen pressure does not occur. Investigations of obstructive sleep apnea showed that apnea events have progressive bradycardia, followed by abrupt tachycardia on the resumption of breathing. So, heart rate variability is a good complement to the oxygen saturation signal. Some works in the literature used HR and SpO2+HR for apnea detection (for more details please refer to Section 4.2.3 and Section 4.2.6). Therefore, using the HR or both SpO2 and HR may improve the diagnosis of OSA in patients at little additional cost. To test this idea, the performance of the combination of SpO2 and HR is tested in this section. Because of the success of the proposed SC3 methods mentioned previously (Section 5.4), the MaxVIF classification technique of the SC3 method is used in this section.

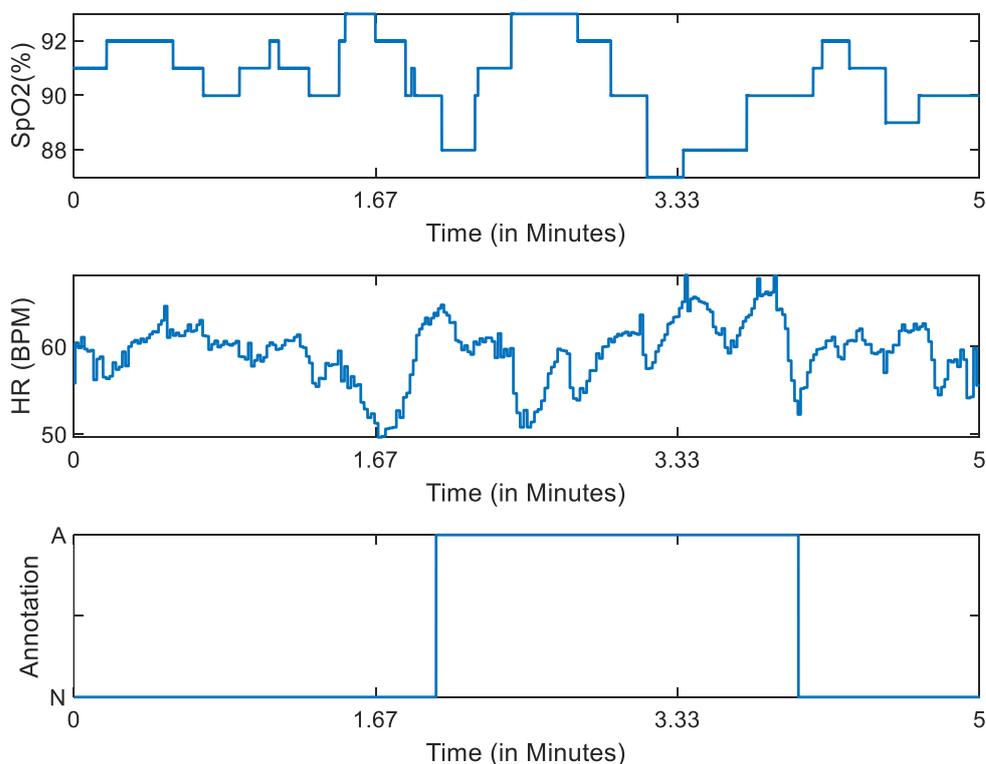


Figure 34 : SpO2 and HR with apnea annotation for five-minute data.

5.5.2. HR features

From the ECG-derived HR, it is possible to analyze the HRV and the inter-beat (RR) interval that can be defined as the interval between successive QRS points. Chen and Zhang [124] mapped the individual long-term RR intervals into a disease state space. A set of 72 recurrence quantification analysis features from the HRV was used by Nguyen et al. [129]. Chen et al. [130] applied the RR intervals for signal segmentation using an iterated cumulative sum of squares algorithm that searches for the small variation changes in time series due to OSA. The combination of the analyses of the HR and the morphology of the ECG can be used to reliably detect sleep disordered breathing, as analyzed by Penzel et al. [133] using the cardiopulmonary coupling.

Almazaydeh et al. [53] derived the RR interval from the ECG signal using an R-peak detection technique and used different time based features from the RR-interval such as the mean, standard deviation, NN50 among others. These are common types of features and are used by almost all of the works presented, such as by Chazal et al. [126]. Yilmaz et al. [125] used a linear kernel function to map the training data into kernel space using 6 features from the RR. Additionally, the RR series was the base of the detection algorithm presented by Ravelo et al. [127].

Three techniques were used by Ravelo-García et al. [136] to obtain the HRV features. First of all, the RR series was encoded into sequences of symbols, and the permutation entropy and symbolic dynamics [117] were used to distinguish different HRV patterns. The second was the cepstrum analysis, thus obtaining cepstrum coefficients. Besides this work the authors [136] used Cepstral coefficients in another work [121]. Besides them, Martínez-Vargas et al. [115] and Travieso et al. [128] used Cepstral from the RR series. The PSD of the EDR was the third technique, using a filter bank with equally spaced filters [136]. These features were then used by the two tested classifiers, an LR and a QDA. Both classifiers achieved similar performance, however, the QDA provided the best results. Cepstrum Coefficients, a filter bank with 34 filters (to analyze the very low, low and high frequency), and detrended fluctuation analysis were employed by Martín-González et al. [137] with LDA, QDA, and LR classifiers reporting that the QDA as better. Quiceno-Manrique et al. [114] employed an analysis based on HRV using KNN with 30 dynamic features based on ten linearly distributed filters. When it comes to frequency based features besides Cepstrum Coefficients and filter bank, Zywiets et al. [118] used four frequency bands: the ULF, the VLF, the LF, the HF with an LDA classifier. The quotient of different frequency bands was used by Kesper et al. [116].

The wavelet transform of the HRV was used by Roche et al. [120]. Time and frequency domain entropies were used by Gutiérrez-Tobal et al. [119]. Khandoker et al. [138] used 14 levels of Daubechies wavelets to decompose the RR and the EDR signals. The result was used as an input to an SVM that classifies the OSA events. Features extracted from the wavelet decomposition of the HRV

and the EDR signals were used by Khandoker et al. [139] as inputs to the SVM classifier. The LDA classifier was also analyzed, providing similar results. The HRV and EDR signals were used by Yildiz et al. [140] using 64 points of the PSD (1 to 32 derived from the HRV and 33 to 64 from the EDR). Three SVM kernels were tested, specifically, the linear, the polynomial, and the Radial Basis Function (RBF). The highest accuracy was produced by the RBF using points 2, 3, 45, and 46 (selected by a hill climbing algorithm.).

The thesis aims is to develop a simple and accurate classification. Therefore, only ECG derived HR is used instead of ECG specific features. A list of features from the literature and an idea calculated from the RR series are tested in this section (Table 21). The variance of the HRV changes with the signal length. The suggested length was five minutes from Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology [214]. Because of this, a five minute signal length is used for the HR analysis.

Table 21: List of HR features.

Feature	Details about Features
1 Minimum (<i>Min</i>)	$Min = \text{minimum}(x)$
2 Skewness (<i>Sk</i>)	$Sk = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - Avg)^3}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - Avg)^2} \right)^3}$
3 Mean (<i>Avg</i>) [53]	$Avg = \frac{1}{n} \sum_{i=1}^n x_i$ where n is the number of data points of the signal x .
4 Coefficient of Variation (<i>CoV</i>)	$CoV = \sqrt{\frac{Var}{Avg}}$
5 Kurtosis (<i>Kurt</i>)	$Kurt = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - Avg)^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - Avg)^2 \right)^2}$
6 Maximum (<i>Max</i>)	$Max = \text{maximum}(x)$
7 Shannon Entropy (<i>SEn</i>)	$SEn = -\sum_{i=1}^n p(i) \ln(p(i))$ Where $p(i)$ is the probability of a specific event occurrence.
8 Renyi Entropy (<i>REn</i>)	$REn = \frac{1}{1-q} \ln(\sum_{i=1}^n p(i)^q)$
9 Variance (<i>Var</i>)	$Var = \frac{1}{n} \sum_{i=1}^n (x_i - Avg)^2$
10 Root mean square (<i>RMS</i>)	$RMS = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i)^2}$
11 Twenty equally spaced filters to form a Filter bank (<i>Fb</i>) [10] [140] [197] where, the Δ_m is the bandwidth of m^{th} filter with a window U and center frequency b_m and N is a number of samples. For the first filter bank (<i>Fb1</i>) $m = 1$.	$Fb1 = \frac{\sum_{k=b_1-\Delta_1}^{b_1+\Delta_1} \left \frac{1}{N} \sum_{n=0}^{N-1} Sat(n) e^{-\frac{j2\pi k n}{N}} \right ^2 U_{\Delta_1}}{\sum_{k=0}^{N-1} \left \frac{1}{N} \sum_{n=0}^{N-1} Sat(n) e^{-\frac{j2\pi k n}{N}} \right ^2}$
12 Second filter bank (<i>Fb2</i>)	$Fb2 = \frac{\sum_{k=b_2-\Delta_2}^{b_2+\Delta_2} \left \frac{1}{N} \sum_{n=0}^{N-1} Sat(n) e^{-\frac{j2\pi k n}{N}} \right ^2 U_{\Delta_2}}{\sum_{k=0}^{N-1} \left \frac{1}{N} \sum_{n=0}^{N-1} Sat(n) e^{-\frac{j2\pi k n}{N}} \right ^2}$
.....
32 Twentieth filter bank (<i>Fb20</i>)	$Fb20 = \frac{\sum_{k=b_{20}-\Delta_{20}}^{b_{20}+\Delta_{20}} \left \frac{1}{N} \sum_{n=0}^{N-1} Sat(n) e^{-\frac{j2\pi k n}{N}} \right ^2 U_{\Delta_{20}}}{\sum_{k=0}^{N-1} \left \frac{1}{N} \sum_{n=0}^{N-1} Sat(n) e^{-\frac{j2\pi k n}{N}} \right ^2}$ 34 filter is used [137]
31 Entropy of level 11 approximation of wavelet (<i>TFSEnCA11</i>)	$TFSEnCA6 = SEn(CA6)$

33	Variance of level 11 approximation of wavelet (TFVarCA11)	$TFVarCA6 = \frac{1}{n} \sum_{i=1}^n (CD6_i - Avg(CA6))^2$
34	Standard deviation of level 11 approximation of wavelet (TFSdCA11)	$TFSdCA6 = \sqrt{\frac{\sum_{i=1}^n (CA6_i - Avg(CD6))^2}{N - 1}}$
35	Median absolute deviation of level 11 approximation of wavelet (TFMadCA11)	$TFMadCA6 = median(CA6)$
36	Entropy of level 1 details of wavelet (TFSEnCD1)	$TFSEnCD1 = SE_n(CD1)$
37	Variance of level 1 details of wavelet (TFVarCD1)	$TFVarCD1 = \frac{1}{n} \sum_{i=1}^n (CD1_i - Avg(CD1))^2$
38	Standard deviation of level 1 details of wavelet (TFSdCD1)	$TFSdCD1 = \sqrt{\frac{\sum_{i=1}^n (CD1_i - Avg(CD1))^2}{N - 1}}$
39	Median absolute deviation of level 1 details of wavelet (TFMadCD1)	$TFMadCD1 = median(CD1)$
40	Entropy of level 2 details of wavelet (TFSEnCD2)	$TFSEnCD2 = SE_n(CD2)$
41	Variance of level 2 details of wavelet (TFVarCD2)	$TFVarCD2 = \frac{1}{n} \sum_{i=1}^n (CD2_i - Avg(CD2))^2$
42	Standard deviation of level 2 details of wavelet (TFSdCD2)	$TFSdCD2 = \sqrt{\frac{\sum_{i=1}^n (CD2_i - Avg(CD2))^2}{N - 1}}$
43	Median absolute deviation of level 2 details of wavelet (TFMadCD2)	$TFMadCD2 = median(CD2)$
....
75	Entropy of level 11 details of wavelet (TFSEnCD11)	$TFSEnCD6 = SE_n(CD6)$
76	Variance of level 11 details of wavelet (TFVarCD11)	$TFVarCD6 = \frac{1}{n} \sum_{i=1}^n (CD6_i - Avg(CD6))^2$
77	Standard deviation of level 11 details of wavelet (TFSdCD11)	$TFSdCD6 = \sqrt{\frac{\sum_{i=1}^n (CD6_i - Avg(CD6))^2}{N - 1}}$
78	Median absolute deviation of level 11 details of wavelet (TFMadCD11)	$TFMadCD6 = median(CD6)$
79-108	30 Cepstrum Coefficient (CC) [137].	$c(\tau) = real(F^{-1}(\log(F(x))))$. Triangular 50% overlapping filter is used to calculate CC.
109	Ratio between power in frequency band [0.026Hz, 0.06 Hz] and power in frequency band [0.06Hz, 0.25 Hz] in RR series[215].	$dvi_pow = (\text{power in frequency band [0.026Hz, 0.06 Hz]}) / (\text{power in frequency band [0.06Hz, 0.25 Hz]})$.
110	Power in the frequency band from 0 Hz up to 0.0033 Hz [10].	ULF
111	Power in the frequency band from 0.0033 Hz up to 0.04 Hz [10] [216] [217].	VLF.
112	Power in the frequency band from 0.04 Hz up to 0.15 Hz [10] [217].	LF.
113	Power in the frequency band from 0.15 Hz up to 0.4 Hz [10] [217].	HF.
114	Ratio of LF and HF [10] [217].	$LF_HF = LF/HF$.
115	Ratio between LF and total power (pTot) [10].	$LF_P = LF / pTot$.
116	Ratio of HF and total power (pTot) [10].	$HF_P = HF / pTot$.
117	Ratio between VLF and total power (pTot) [10].	$VLF_P = VLF / pTot$.
118	ULF/P: ratio between ULF and total power (pTot) [10].	$ULF_P = ULF / pTot$.
119	Ratio of sum of ULF, VLF and total power (pTot) [10].	$ULFVLF_P = (ULF + VLF) / pTot$.
120	Ratio of sum of ULF, VLF, LF and total power (pTot) [10].	$ULFVLF_P = (ULF + VLF + LF) / pTot$.
121	LF in normalized units, $LF / (P - VLF) \times 100$ [10].	$LF_n = LF / (pTot - VLF) \cdot 100$.
122	HF in normalized units, $HF / (P - VLF) \times 100$ [10].	$HF_n = HF / (pTot - VLF) \cdot 100$.
123	Percentage of RR-interval differences greater than 10 ms [216].	pNN10
124	Percentage of RR-interval differences greater than 20 ms [10] [216].	pNN20
125	Percentage of RR-interval differences greater than 50 ms [10] [216].	pNN50
126	Percentage of RR-interval differences greater than 100 ms [10].	pNN100
127	Percentage of RR-interval differences greater than 120 ms [10].	pNN120
128	Percentage of RR-interval differences greater than 150 ms [10].	pNN150
129	First the RR intervals are divided into 4 four codes Sn	WPSUM13

{0,1,2,3}.

$$Sn = \begin{cases} 0 & \text{if } x_{rr} > Avg_{rr} \text{ and } x_{rr} \leq (1+a)Avg_{rr} \\ 1 & \text{if } x_{rr} > (1+a)Avg_{rr} \\ 2 & \text{if } x_{rr} > (1-a)Avg_{rr} \text{ and } x_{rr} \leq Avg_{rr} \\ 3 & \text{if } x_{rr} \leq (1-a)Avg_{rr} \end{cases}$$

Then words are formed using 3 symbols. The percentage of these words containing the symbols “1” and “3” is indicated as WPSUM13 [117] [10].

130	Same process as 129 features. The percentage of words contains code ‘0’ and ‘2’ [10].	WPSUM02
131	Forbidden words (low probability less than 0.001) of length 3 [10].	FORBWORDS
132	Shannon entropy of the words [10].	FWSHANNON
133	Renyi entropy (using $q = 0.25$) for the words [10].	FWRENYI0_25
134	Renyi entropy (using $q = 4$) for the words [10].	FWRENYI4
135	Variability (standard deviation) of the time series calculated from the transformed sequence of words [10]. The transformation is S_T where n_{13} is the number of symbols corresponding to 1 or 3 in the word and $S_{13}(\omega_i)$ represents the word having either ‘1’ or ‘3’ in the first symbol for $i = 1, 2, 3 \dots$	$WSDVAR = \sqrt{\left(\frac{\sum_{i=1}^N (S_T(\omega_i) - \frac{1}{N} \sum_{i=1}^N S_T(\omega_i))^2}{N-1} \right)}$ $S_T(\omega_i) = \begin{cases} 3 & \text{if } n_{13}(\omega_i) = 3 \text{ and } S_{13}(\omega_i) = '1' \\ 2 & \text{if } n_{13}(\omega_i) = 2 \text{ and } S_{13}(\omega_i) = '1' \\ 1 & \text{if } n_{13}(\omega_i) = 1 \text{ and } S_{13}(\omega_i) = '1' \\ 0 & n_{13}(\omega_i) = 0 \\ -1 & \text{if } n_{13}(\omega_i) = 1 \text{ and } S_{13}(\omega_i) = '3' \\ -2 & \text{if } n_{13}(\omega_i) = 2 \text{ and } S_{13}(\omega_i) = '3' \\ -3 & \text{if } n_{13}(\omega_i) = 3 \text{ and } S_{13}(\omega_i) = '3' \end{cases}$
136	Probability that the word “000000” occurs for a specific thresholding code S_{th} used with a threshold th of 5 ms.	POLVAR5
	$S_{th,i} = \begin{cases} '0' & \text{if } RR_i - RR_{i-1} < th \\ '1' & \text{if } RR_i - RR_{i-1} \geq th \end{cases}$	
137	Probability that the word “000000” occurs for a specific thresholding code S_{th} used with a threshold th of 10 ms.	POLVAR10
138	Probability that the word “000000” occurs for a specific thresholding code S_{th} used with a threshold th of 20 ms.	POLVAR20
139	Probability that the word “111111” occurs for a specific thresholding code S_{th} used with a threshold th of 5 ms.	PHVAR5
140	Probability that the word “111111” occurs for a specific thresholding code S_{th} used with a threshold th of 10 ms.	PHVAR10
141	Probability that the word “111111” occurs for a specific thresholding code S_{th} used with a threshold th of 20 ms.	PHVAR20

5.5.3. SpO2 and combination of SpO2 HRV features

To have an equal input for both signals the HR signal five-minute features of SpO2 signal are used. SpO2 and SpO2+HR are carried out by combining features of the SpO2 (Table 6) and the HR (Table 21).

5.5.4. Results of epochs based classification

The MaxVIF classification technique of the SC3 (Section 5.4) was considered since it previously attained significant results. A continuous improvement of the cost function over the generation is

visible for both HR, and HR+SpO2 (Figure 35) and the simulations terminate in the 21st generation for the HuGCDN2008 database because of the stopping condition. The optimum solution for the HR signal is with the combination of a ANN with two SVM_RBF. During these generations, the accuracy is improved from 73.21% to 75.25% for the HR (Figure 36). In the case of Sen, it was improved from 69.26% to 72.81%, while for Spc, it was improved from 74.75% to 76.17% (Figure 36). The improvement of the cost function as well as of Acc, Sen and Spc also happened with SpO2+HR. The Acc, Sen and Spc improvements over these 21 generations from were 84.27% to 85.09%, from 77.98% to 78.90%, and from 86.50% to 87.21%, respectively (Figure 36).

From Figure 37, it can be that the number of total features for the best solution for each generation fluctuates over the generations. The total number of features changes between 198 to 241 for the HR signal and between 277 to 308 for the SpO2+HR signals (Figure 37). The optimum number of features for the HR signal was 222 which were distributed between a ANN and two SVM_RBF classifiers 72, 77, and 73 respectively (Table 22). In the case of SpO2+HR, the total number of features of the optimum solution was 290, where 93 features were used from the SpO2 signal and 197 from the HR signal (Table 22). Among the classifiers, the ANN used 105 features and the two SVM_L used 93 and 92 features. The selected features for the HR and the SpO2+HR signals are shown in Table 22.

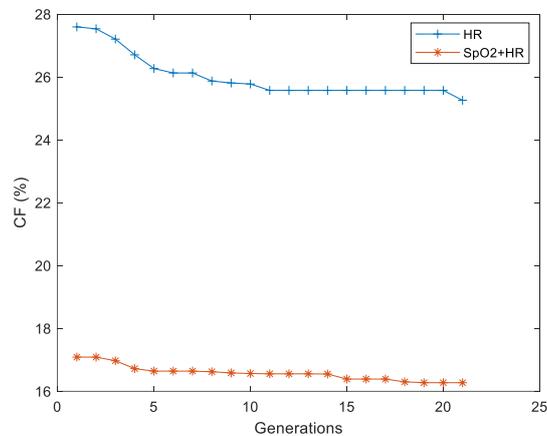


Figure 35 : Cost of the 5-minute MaxVIF classification combination over the generations for the best performance objective.

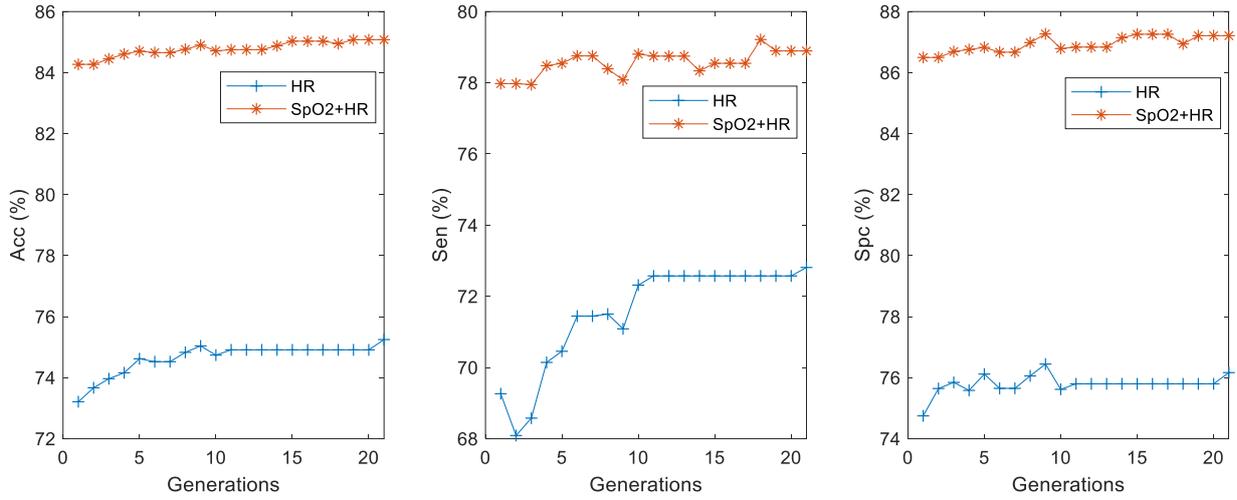


Figure 36 : Accuracy (Acc), Sensitivity (Sen), and Specificity (SpC) of 5 minute MaxVIF over the generations for the best performance objective for the HR and HR+SpO2 signals.

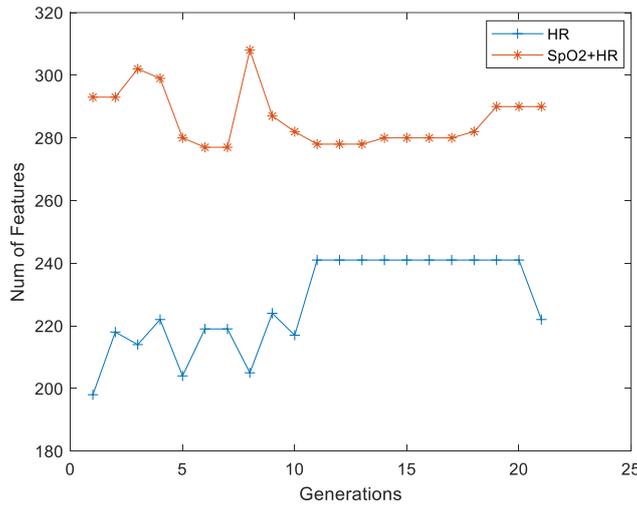


Figure 37 : Number of features of the 5-minute MaxVIF classification combination over the generations for the best performance objective.

Table 22 : Selected features for HR and SpO2+HR using S3C.

Signals	Classifiers	Nub of Features	Features
HR	ANN	72	mean, variance, CoV, kurtosis, max, renyi v2, shannon entropy, filter1, filter2, filter5, filter6, filter8, filter9, filter14, filter15, filter17, filter18, filter19, MAD_CA11, SD_CD1, Var_CD2, SD_CD2, MAD_CD2, Var_CD3, SEn_CD4, Var_CD4, MAD_CD4, SD_CD5, MAD_CD5, Var_CD7, MAD_CD7, SEn_CD8, Var_CD8, SD_CD8, MAD_CD8, SEn_CD9, MAD_CD9, Var_CD10, MAD_CD10, Var_CD11, SD_C11, Rcc2, Rcc3, Rcc7, Rcc8, Rcc9, Rcc10, Rcc15, Rcc16, Rcc18, Rcc23, Rcc24, Rcc25, Rcc28, Rcc30, dvi_pow, VLF, HF, LF_P, VLF_P, ULFVLF_P, LFn, pNN20, pNN50, pNN120, WPSUM13, WPSUM02, FWRENY10_25, FWRENYI4, WSDVAR, POLVAR5, POLVAR20
	SVM_RBF	77	mean, kurtosis, rms, min, renyi v2, shannon entropy, filter1, filter4, filter8, filter9, filter11, filter15, filter16, filter17, filter18, filter19, Var_CA11, MAD_CA11, Var_CD1, MAD_CD1, SEn_CD2, MAD_CD2, MAD_CD3, SEn_CD4, SEn_CD5, SD_CD5, SD_CD6, MAD_CD6, Var_CD7, SD_CD7, SEn_CD8, Var_CD8, MAD_CD8, SEn_CD9, Var_CD9, SD_CD9, MAD_CD9, SEn_CD10, Var_CD10, MAD_CD10, SEn_CD11, SD_CD11, MAD_CD11, Rcc1, Rcc5, Rcc6,

			Rcc7, Rcc8, Rcc10, Rcc11, Rcc12, Rcc13, Rcc15, Rcc19, Rcc20, Rcc23, Rcc24, Rcc25, Rcc26, Rcc28, dvi_pow, ULF, LF_HF, LF_P, HF_P, ULF_P, LFn, pNN20, pNN150, WPSUM13, WPSUM02, FWSHANNON, FWRENYI4, WSDVAR, POLVAR10, PHVAR5, PHVAR20.
	SVM_RBF	73	mean, skewness, kurtosis, max, min, renyi v2, shannon entropy, filter3, filter6, filter7, filter10, filter11, SEn_CA11, Var_CA11, SD_CA11, MAD_CA11, SEn_CD1, SD_CD1, SEn_CD2, SD_CD2, SEn_CD3, SEn_CD4, SD_CD4, MAD_CD4, SEn_CD5, Var_CD5, MAD_CD5, SD_CD7, SEn_CD8, Var_CD8, SEn_CD9, Var_CD9, SD_CD9, Sen_CD10, MAD_CD10, Var_CD11, SD_CD11, MAD_CD11, filter15, filter16, filter18, Rcc1, Rcc3, Rcc8, Rcc9, Rcc10, Rcc11, Rcc13, Rcc16, Rcc17, Rcc19, Rcc22, Rcc23, Rcc27, dvi_pow, LF, HF, LF_P, HF_P, VLF_P, ULF_P, ULFVLF_P, pNN10, pNN20, pNN100, WPSUM02, FWSHANNON, FWRENYI0_25, POLVAR5, POLVAR20, PHVAR5, PHVAR10, PHVAR20
	ANN	105	SpO2(33) CoV, kurtosis, max, min, filter2, filter3, filter5, filter8, filter9, filter12, filter15, filter17, filter20, SEn_CA6, Var_CA6, SD_CA6, MAD_CA6, SEn_CD6, Var_CD6, MAD_CD6, SEn_CD5, Var_CD5, MAD_CD4, Var_CD3, SD_CD3, MAD_CD3, SEn_CD2, Var_CD2, MAD_CD2, Var_CD1, MAD_CD1, CTM50, DIndex. HR (72) CoV, kurtosis, rms, min, renyi v2, shannon entropy, filter7, filter13, filter16, filter18, filter19, filter20, MAD_CA11, SEn_CD1, Var_CD1, MAD_CD1, Var_CD2, MAD_CD2, Var_CD3, SD_CD3, SD_CD4, MAD_CD4, SEn_CD6, SD_CD6, SEn_CD7, Var_CD7, SD_CD7, SEn_CD8, Var_CD8, SD_CD8, MAD_CD8, Var_CD9, MAD_CD9, SEn_CD10, Var_CD10, SD_CD10, SEn_CD11, MAD_CD11, Rcc1, Rcc4, Rcc7, Rcc8, Rcc9, Rcc10, Rcc11, Rcc12, Rcc14, Rcc16, Rcc17, Rcc22, Rcc23, Rcc24, Rcc25, Rcc30, VLF, HF, LF_P, HF_P, ULF_P, ULFVLF_P, ULFVLF_P, HFn, pNN20, pNN150, WPSUM13, FWSHANNON, FWRENYI4, WSDVAR, POLVAR10, POLVAR20, PHVAR5, PHVAR10.
	SVM_L	93	SpO2(29) variance, skewness, kurtosis, rms, renyi v2, filter1, filter2, filter5, filter9, filter11, filter12, filter14, filter17, filter18, filter19, SEn_CA6, Var_CD6, MAD_CD6, SEn_CD5, SEn_CD4, Var_CD4, SD_CD4, MAD_CD4, SEn_CD2, Var_CD2, SD_CD2, SEn_CD1, MAD_CD1, DIndex. HR (64) mean, max, min, filter3, filter4, filter5, filter6, filter7, filter8, filter9, filter13, filter15, filter16, filter17, filter19, filter20, MAD_CA11, SEn_CD1, Var_CD1, Var_CD2, SD_CD2, MAD_CD2, SEn_CD3, Var_CD3, SD_CD3, MAD_CD3, MAD_CD4, Var_CD5, SD_CD5, MAD_CD6, SEn_CD7, MAD_CD8, SEn_CD10, Var_CD10, SD_CD10, SEn_CD11, MAD_11, Rcc4, Rcc9, Rcc11, Rcc13, Rcc14, Rcc16, Rcc18, Rcc22, Rcc24, Rcc28, Rcc30, HF, LF_HF, VLF_P, ULF_P, ULFVLF_P, ULFVLF_P, pNN20, pNN100, pNN120, pNN150, WPSUM13, WPSUM02, FORBWORDS, FWSHANNON, POLVAR5, POLVAR10.
	SVM_L	92	SpO2(31) variance, kurtosis, rms, min, renyi v2, filter1, filter3, filter6, filter8, filter11, filter16, filter17, SEn_CA6, Var_CA6, MAD_CA6, SD_CD6, MAD_CD6, SD_CD5, MAD_CD5, SEn_CD4, Var_CD4, SD_CD4, Var_CD3, SD_CD3, MAD_CD3, SD_CD2, SEn_CD1, Var_CD1, SD_CD1, CTM50, DIndex. HR (61) mean, CoV, kurtosis, renyi v2, filter2, filter3, filter6, filter8, filter9, filter10, filter13, filter14, filter16, SEn_CD1, Var_CD1, SD_CD1, SEn_CD2, SD_CD2, SD_CD3, SD_CD4, MAD_CD4, SEn_CD5, Var_CD5, MAD_CD5, SEn_CD6, SD_CD6, SD_CD7, MAD_CD8, Var_CD9, MAD_CD9, SEn_CD10, Var_CD10, MAD_CD10, Var_CD11, MAD_CD11, Rcc4, Rcc5, Rcc8, Rcc9, Rcc11, Rcc14, Rcc17, Rcc23, Rcc24, Rcc28, Rcc29, dvi_pow, HF, LF_P, VLF_P, ULFVLF_P, pNN20, pNN50, pNN120, pNN150, WPSUM02, FORBWORDS, FWRENYI4, POLVAR10, PHVAR5.

SpO2+HR

Table 23 : Compared to other works.

Ref.	Signal	Classifier	Database	Recordings or Patients	Input size (Seconds)	No of features	Sen	Spc	Acc
[99]	SpO2	ANN	AED [61]	8	60	3	87.5	100	90.3
[195] (Section 5.2)	SpO2	SVM-L	AED [61]	8	60	50	83.76	97.03	96.89
[195] (Section 5.2)	SpO2	SVM-L	AED [61]	8	60	20	84.57	97.28	97.38
[102] (Section 5.3)	SpO2	ANN	AED [61]	8	60	7	96.5	98.5	97.7
[196] (Section 5.4)	SpO2	MaxVSF1	AED [61]	8	60	34	98.11	86.98	91.33
[106]	SpO2	DAE	AED [61]	8	60	-	78.75	95.89	97.64

[107]	SpO2 + IHR	LSTM	AED [61]	8	60	-	84.7	-	92.1
[107]	SpO2	LSTM	AED [61]	8	60	-	92.9	-	95.5
[2]	SpO2	SVM	Own	40	150	7	-	-	90
[190]	SpO2 + airflow	RCNN	MGH	10000	1	-	-	-	88.2
[190]	SpO2 + airflow + respiration	RCNN	SHHS	5804	1	-	-	-	80.2
[73]	SpO2 + ECG	Bagging.RepTree	UCD [62]	25	60	39	79.75	85.89	84.80
[73]	SpO2	Bagging.RepTree	UCD [62]	25	60	39	78.23	84.25	82.79
[195]	SpO2	ANN	UCD [62]	25	60	2	43.31	95.03	81.95
(Section 5.2)									
[195]	SpO2	LD	UCD [62]	25	60	9	61.78	91.03	83.27
(Section 5.2)									
[106]	SpO2	DBN	UCD [62]	25	60	-	60.36	91.71	85.26
[185]	SpO2 + oronasal airflow + ribcage and abdomen movements	CNN2D	UCD[62]	23	1	-			79.6
[10]	Spo2	LDA	HuGCDN2008	70	60,300	19	75.6	91.00	86.5
[10]	Spo2 + HRV	LDA	HuGCDN2008	70	60,300	33	73.4	92.3	86.9
[196]	SpO2	MaxVIF1	HuGCDN2008	70	300	89	80.54	85.74	84.46
(Section 5.4)									
P	HR	MaxVSF5	HuGCDN2008	70	300	222	72.81	75.25	76.17
P	SpO2+HR	MaxVSF5	HuGCDN2008	70	300	290	78.90	87.21	85.08

5.5.5. Global classification

The true definition of the AHI is the number of apnea and hypopnea events per hour of sleep. However, window based, or epoch-based apnea and hypopnea classifiers might not be able to detect the true AHI; instead, they detect the number of apnea and hypopnea in the window. Therefore, the window based method is limited to the window size. In this work, a 1-minute annotation with a sliding window is used. This restriction is also posed because of the data set annotation. Thus, instead of events, an apnea hypopnea minute per hour in bed (AHI TiB) is implemented to detect apnea patients. This technique is applied in the literature for apnea patients detection otherwise called the global classification [10][169]. The accuracy of the global classification (GAcc) is varies from 96.67% to 100% in different databases and techniques.

The HRV based work performed by Martín-González et al. [137] used 30 subjects from the AED database (achieved Acc of 96.67%) and 39 subjects from the HuGCDN2014 database (achieved 87.18% of Acc). To achieve these results, the borderline subjects were removed. This Exclusion of the Borderline Subjects (EBLS) is also followed by other researchers in order to improve the GAcc such as the work of Ravelo-García et al. [10] which achieved 81.8% of GAcc using the HRV. In the same work other results of GAcc were achieved 71% using the HRV without the EBLS; 94.3% and 97% using the SpO2 without the EBLS and with the EBLS, respectively; 91.4% and 100% using the combination of SpO2 and HRV without the EBLS and with the EBLS, respectively. In another work, using 66 subjects, the results are: 91.89% for the SpO2, 91.89% for the HRV and 100% for SpO2+HRV [197]. Al-Angari et al. [197] have achieved a GAcc of 95% (SpO2), 78% (HRV), 87% (respiratory), and 95% (SpO2+ HRV+ respiratory) using SVM and 100 subjects from the SHHS database. Biswal et al. [190] achieved 80.2%, and 88.2% of GAcc with RCNN classifier using the

SHHS database and the MGH database, respectively. In case of a minute-by-minute apnea detection, an AHI index of more than 60 is theoretically impossible. Therefore, there are works that excluded subjects with $AHI \geq 60$ when the GAcc is calculated [169].

This work presents the comparison between the AHI calculated by the medical professional (AHI G MP) of the database and the AHI TiB calculated from the output of the classifiers (AHI C TiB) (Figure 38). Since the current standard of the AHI index for detecting apnea patients is 5, AHI TiB = 5 without any removal of subjects (EBLS) is used in this work. In the case of the HuGCDN2008 database this applies to all of the seventy subjects. The SpO2 signal achieves 97.14% (with $R^2(R2) = 0.8956$), the HR method achieves 84.29% (with $R^2(R2) = 0.7621$) and SpO2+HR achieves 97.14% (with $R^2(R2) = 0.8956$) for 5Min using the HuGCDN2008 database. The closest work found in the literature compared with this work was carried out by Ravelo-García et al. [10], which uses the same HuGCDN2008 database with 35 test subjects. Without the EBLs and using the SpO2 signal, Ravelo-García et al.'s [10] work achieved a GAcc of 91.4% compared to our 97.14%; for HR 71.4% compared to our 84.29% and for SpO2+HR 94.3% compared to our 97.14%. Therefore, a better global accuracy is achieved compared with similar database.

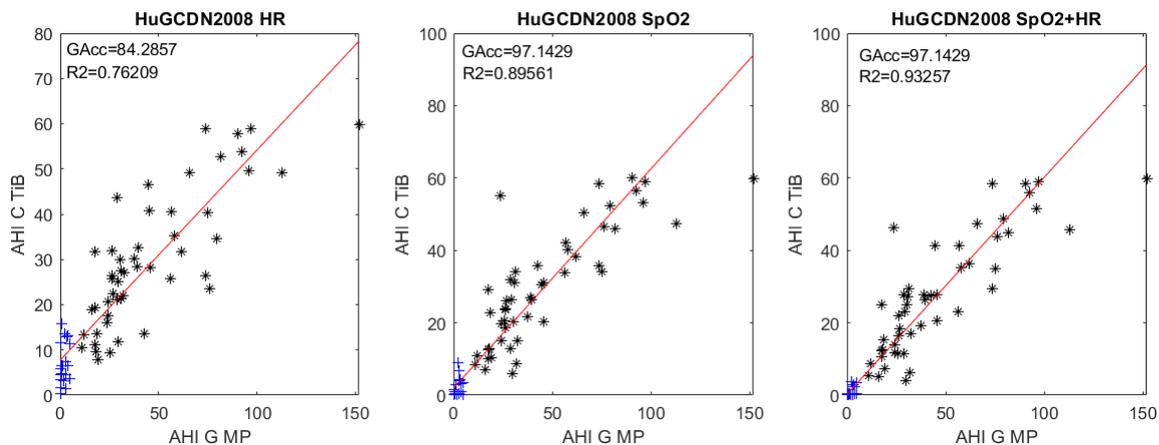


Figure 38: Comparison of the global accuracy of HuGCDN2008 dataset with AHI calculated by medical physician (AHI G MP) and the by the CNN classifiers' AHI time in bed (AHI C TiB) for SpO2, HRV and SpO2+HRV of 1 Min input. (+) symbol is used for normal subjects ($AHI \leq 5$) and (*) is used for apnea patients.

5.5.6. Summary

As was previously verified, the maximum voting with independent features for each classifier attained the best performance of all self-configuring classification combination models. This model was used in this section to test the HR and SpO2+HR performance. Although they either outperform similar kinds of work or achieved almost the same performance, the results of the proposed methods

are more balanced. A well-balanced classifier is preferable in certain cases to an extremely sensitive one to apnea or normal events.

In the case of global classification, the performance of the SpO2 signal is better than that of the HR and SpO2+HR. The result is better than in similar kinds of work in the literature.

5.6. Summary of Handcrafted Feature Based Method

This chapter mainly focused on feature creation, selection and choosing of shallow classifiers. Initially, first filter methods (mRMR) of classification with a wrapper method (SFS) were tested. A subset of features improved the performance when compared to a set with all the features combined. Most of the features chosen by the classifiers are defined in the time-frequency domain. In terms of computation, the mRMR is the most suitable option. An SFS uses a higher number of features in most of the cases. The accuracy difference between both algorithms are not significant. Both methods suffer from low sensitivity due to the prevalence of normal segments compared to apnea events. Due to the different hypotheses of different classifiers, they result in different importance for the features. To increase the sensitivity of GA method, a ANN was tested and achieved a better solution. Like previous solutions the time-frequency feature did better. The combination of classifiers can also increase performance. Through the combination of feature selection with combination of classifiers, a selection process was developed named the self-configuring classification combination method. When compared with a single classifier with the feature selection method, the SC3 performs better. Two voting methods the MaxV and the WLC, with two feature selection techniques were tested using SC3. For SpO2, it was verified that maximum voting with independent features for each classifier attained the best performance. SC3 classifiers were able to achieve subject independence and database independence with a balanced performance. Some research suggested adding more signals such as an ECG can increase the performance of the system. Therefore, performance of the HR derived from the ECG and SpO2+HR (HR derived from ECG) were tested using SC3 technique. However, the HR and SpO2+HR were not able to outperform the SpO2 signals performance. A similar trend of results also follows for global classification.

Chapter 6

Automated Feature-Based Methods

This chapter describes automated feature-based methods. Two hyperparameter optimization methods for CNN classifiers is proposed and tested with different signals.

6.1. Introduction

This chapter describes the automated feature-based methods applied in this work. The problem of building a reliable system using any sensor is mainly two-fold, firstly, in finding the best features that describe the apnea events, secondly, in using these features to detect the apnea accurately. Multiple researchers have analyzed different time domain [9] and frequency domain characteristics [10], thus, creating a vast pool of suitable features. The creation of handcrafted features that achieve good performance requires significant domain knowledge. Furthermore, it is becoming significantly more challenging to find a new set of features that can achieve higher performance, since combining two or more features does not guarantee an improvement in performance [11]. Therefore, a large number of features needed to be sorted according to relevance in order to increase accuracy. Various techniques were employed in this work to address this problem (minimum Redundancy Maximum Relevance (mRMR), Sequential Forward Search (SFS) [195] and Genetic Algorithms (GA) [102]). However, these techniques most of the time, do not guarantee that the best features have been selected. Additionally, to that, the best features are sometimes dependent upon the classifiers used. Deep learning has the ability to automatically learn features from raw data [23]. Thus, by using deep network these problems can be solved.

A deep CNN is one of the most successful deep networks that are inspired by the vision system. Traditionally, a CNN is designed for two-dimensional (2D) images as input with different channels [218]. However, it can also be used for one-dimensional (1D) signals with a single-channel [219] [180]. In most of the cases, automated detection of obstructive sleep apnea events using a CNN performs better when compared with shallow classifiers [177]. Some authors used nasal airflow [184] or a combination of SpO₂, oronasal airflow, and ribcage and abdomen movements [185], and then converted these one dimensional signals into a two dimensional input to employ the two dimensional CNN (CNN2D) directly for apnea detection. A one-dimensional CNN (CNN1D) is a good alternative requiring far less preprocessing (does not need to convert 1D to 2D) for 1D signals. Electrocardiogram (ECG) [179][180][181] and nasal pressure or airflow [182][149] signals have been used with a CNN1D for OSA classification. Haider et al. [183] employed three one dimensional signals (nasal airflow, abdominal and thoracic plethysmography) to feed a CNN1D with three channel inputs for OSA detection. Following this line of research, the SpO₂ signal which is 1D in nature was selected to be directly fed to a CNN1D without any dimension transformation stage. For details about automated feature-based apnea detection and the use of CNN in the literature please refer to Section 4.3.

Implementing a CNN, however, presents significant changes. The structure and/or hyperparameters of the network are typically selected through an experimental search. Such methods requires a

significant amount of time as well as experience and expert knowledge for the creation of a handcrafted network structure and hyperparameters [220]. A possible alternative is to use evolutionary algorithms, such as a GA, to solve the structural optimization problem. The algorithm starts with a random individual generation and uses mutation and crossover over a defined number of generations to achieve the optimized solution by optimizing the fitness function. Zhining et al. [221] designed a genetic convolutional neural network model based on a random sample and found it has a better performance than a CNN in the MNIST data set. Evolutionary algorithms also achieved significant success in the configuration of topologies [222] and in the connection of convolution layers [223]. Regarding the selection of the network hyperparameters, an asynchronous evolutionary approach was successfully used on a Titan supercomputer [224]. Furthermore, a neuro-evolution was able to construct large, accurate networks from trivial initial conditions while searching through a large space without experimenter participation [222]. By combining Dynamic Structured Grammatical Evolution (DSGE) with GA Assuncao et al. [225] were able to achieve better results without resorting to prior knowledge. Grammatical Evolution (GE) was also used for handwritten digit recognition [226] as well as human activity recognition [227]. A DNN with a GA to optimize the predictive accuracy, named EvoDeep, was developed by Martin et al. [228]. This concept of using GA for choosing the best network was also successfully extended to transfer learning [229].

Unbalanced data are also a common issue in sleep apnea detection, having an insufficiency of one class (apnea) level and prevalence of another class (normal) level. Thus, a single objective technique (which was applied in the previously mentioned applications) commonly tries to maximize the accuracy (either directly or indirectly), leading to a biased classifier, since an increase in accuracy can sacrifice the sensitivity (apnea events detection) that is related to the less prevalent class. To solve this problem, a multi-objective method (Section 6.2) and a combined approach (Section 6.3) are used for a different test.

Therefore, the primary objectives of this chapter are:

- To design an automatic feature-based sleep apnea events detection algorithm using a CNN for a SpO2 signal, an HR signal and HR+SpO2;
- To develop independent algorithms capable of choosing the CNN structure and hyperparameters without any human intervention, using balanced optimization.
- To analyze the effects of input sizes, database dependencies, and the layer size in the classification.

To achieve the desired objectives different algorithms were developed for the automatic creation and architectural hyper-parameterization of the CNN. The first one is a multi-objective optimization (Section 6.2) with three input sizes. Afterward, another faster algorithm is developed (Section 6.3) to reduce the optimization time. Finally (in Section 6.4), the HR. SpO2+HR as well as effectiveness of adding an extra fully connected layer, and a dropout layer, are tested.

6.2. Multi-Objective Architectural Hyperparameter Optimization of CNN

6.2.1. Introduction

Because of the successful implementation of Evolutionary Algorithms (EAs) to optimize the hyperparameters of a deep network for sleep apnea detection with HRV by Falco et al. [178], a GA was employed in this work for hyperparameter optimization of the CNN.

To solve unbalanced data, the designed model addresses this issue by simultaneously considering the Acc, Sen, and Spc in a multi-objective problem. The Non-dominated Sorting Genetic Algorithm II (NSGA-II) [230] was selected for this work due to the large success it has in other areas, such as filter design [231], and water distribution systems [232]. The proposed system used a CNN with NSGA-II algorithms to solve the multi-objective problem by choosing a suitable structure which can achieve the goal (a balanced result). Three different databases were used to create and test the system. Due to the non-uniformity between the datasets, such as sampling frequency and annotation methods, a normalization preprocess was performed. To have all of the databases at the same sample rate the UDB [62] was resampled at 50Hz. In this work, apnea events were detected with one minute epochs (as employed by AED [61]). Therefore, the annotation for the HuGCDN2008 database was produced by labeling the minute as apnea minute if any or both of its 30 second windows were annotated as an apnea by the physician. For the UDB [62] if 10 or more seconds, in a minute, were annotated as apnea by the physician then the one minute epoch was labeled as apnea. The input sizes of 1 minute, 3 minutes (with 2 overlapping minutes), and 5 minutes (with four overlapping minutes) were created considering the central minute as the one that defines the label. Therefore, taking into consideration the selected sampling frequency the 1 minute (60 seconds), 3 minute (180 seconds), and 5 minute (300 seconds) windows had 3000, 9000 and 15000 sampling points, respectively.

In brief, to achieve the desired objectives an algorithm was developed for the automatic creation and architectural hyper parameterization of the CNN using one database with three input sizes. Later another two databases (for a total of three) with three input sizes were tested in two different settings: direct implementation and transfer learning implementation, and the results were analyzed. Therefore, the primary objectives of this work are:

- To design an automatic sleep apnea events detection algorithm using a CNN and a SpO2 signal.
- To develop an independent algorithm capable of choosing the CNN structure and the hyperparameters without any human intervention, using the multi-objective optimization.
- To analyze the effects of the input sizes, the database dependencies and the layer size on the

classification.

6.2.2. Optimization of CNN hyperparameters using GA

The CNN's hyperparameters were optimized using a multi-objective genetic algorithm named NSGA-II [230][233]. The multi-objective technique was used in order to have an equally better performance in all the objectives, contrary to what is present in the single-objective optimization method [231]. A Multi-objective optimization technique optimizes a vector (O) of objective functions (in this work Acc, Sen, and Spc) and the optimization consists of finding V (in this case, the hyperparameters) which maximizes $O(v)$ represented as

$$O(v)=(O_1, O_2(v), \dots, O_k(v)) \quad (10)$$

$$\text{subject to: } \mathcal{Y}(v)=(y_1(v), y_2(v), \dots, y_j(v)) \quad (11)$$

$$v_i^l \leq v_i \leq v_i^u \quad (12)$$

where v is the vector of design variables in V parameter space with N elements with upper bound v_i^u and lower bound v_i^l , $y(v)$ is the objective space and $O(v)$ is the vector representation of the objective functions that has to be maximized [234].

A simplified representation of the implementation strategy is presented in Figure 39 where all inputs of layers are represented as x and the outputs as y . For every generation (Gen) the chromosomes of each population (Pop) (P_t) were generated using mutation and crossover with the information needed to create a CNN. Then, it was translated to the CNN structure and parameters using the decoding methods indicated in Table 24. After twofold training and testing, the next generation population (P_{t+1}) was chosen according to the Pareto fronts and the crowding distance using Acc, Sen, and Spc of the two-fold test.

The implemented NSGA-II [230] technique can be described in 11 steps:

Step 1: A parent population P_0 with the size of N is randomly generated.

Step 2: The system converts the chromosome to a CNN. A fixed input (3000 neurons for 1 minute, 9000 neurons for 3 minutes and 15000 neurons for 5 minutes) layer and output layer (fully connected (FC) layer with 2 outputs, a softmax layer and a class output) are always present (fixed), regardless of the structure chosen by the GA algorithm. The GA algorithm was only allowed to choose the number of the layers between fixed layers, the type of layers, the size of kernels, the pooling sizes, the stride and the number of neurons of a fully connected layer. A real coded chromosome that ranges from 0 to 1 was used in this work. However, different types of parameters for the CNN had different ranges, thus proper decoding was done according to Table 24. First the generated chromosomes were scaled between the defined ranges and then a ceiling function was used to obtain natural numbers. To reduce the number of possible solutions, hence reducing the simulation time, two different types of layer

combinations were used. The first was a convolution layer with a ReLU layer and a batch normalization defining it as ConvX. The second one was maximum pooling, indicated by MaxP. The ConvX layer has three cascaded functions, performing convolution with the input and a defined kernel (k) then a batch normalization, and finally a ReLU, indicated together by f_{bd} in the Figure 39. To prevent losing too much information in each layer, a back to back MaxP layer (cascaded MaxP layer) was replaced by a ConvX layer.

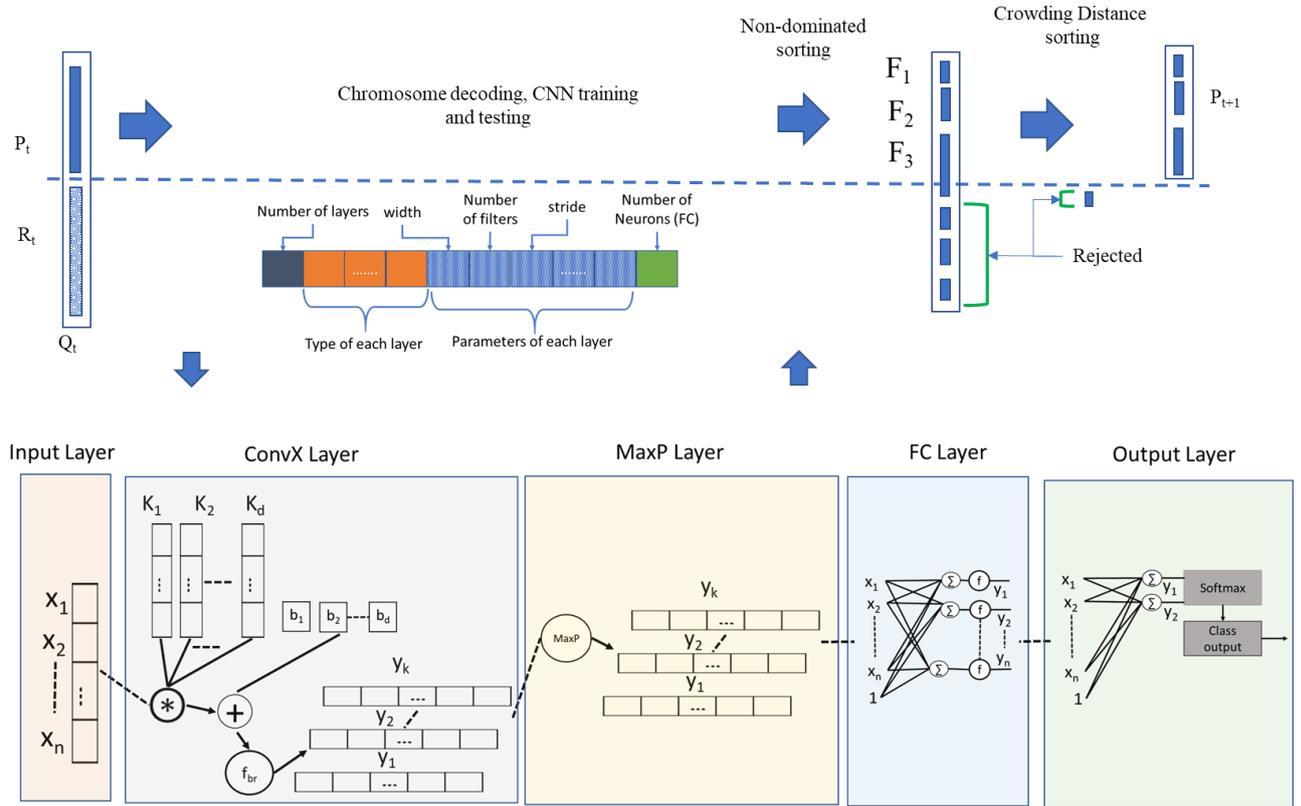


Figure 39 : Simplified representation of the CNN hyperparameters optimization strategy using NSGA-II [230].

Table 24 : Chromosome decoding techniques and ranges for the CNN.

Type	Number of Position	Quantity	Range
1	1	Number of layers	1-10
2	Number of max flexible layers	Type of each layer	0-1(ConvX)-(MaxP)
3	Number of highest parameters*Number of max layers	Each layer parameter	
3.1	1	Number of filters	1-15 (ConvX)
3.2	1	Filter size (Width)	3-9(ConvX), 2-5 (MaxP)
3.3	1	Stride	1- filter size
4	1	FC neuron number	20-200

Step 3: After generating the CNN structure (hyperparameters), the network was trained using the ADAM algorithm [235] with two fold methods. The HuGCDN2008 database has 70 subjects which

were divided into 35 subjects in for the training and test set. Subject independence between the training and test sets was ensured by not mixing the subjects data between the sets. An initial learning rate of 0.03 was employed during the training and for every 10 epochs the learning rate drop factor was 0.1. The batch size was 1024 and the data were shuffled in every epoch. An average of the Acc, Sen, and Spc was calculated to be used as objective parameters.

Step 4: Using the objective parameters a non-dominant sorting was performed for sorting the parent population, where $P = P_0$ [230].

FIRST NON-DOMINATED SORT (P)

```

for each  $p \in P$ 
   $S_p = \emptyset, n_p = 0$ 
  for each  $q \in P$ 
    if  $p$  dominates  $q$  ( $p < q$ ) then
       $S_p = S_p \cup \{q\}$ 
    else if ( $q < p$ ) then
       $n_p = n_p + 1$ 
  if  $n_p = 0$  then  $p_{rank} = 1$ 
   $\mathcal{F} = \mathcal{F}_1 \cup \{q\}$ 
 $i = 1$ 
while  $\mathcal{F}_i \neq \emptyset$ 
   $Q = \emptyset$ 
  for each  $p \in \mathcal{F}_i$ 
    for each  $q \in S_p$ 
       $n_q = n_q - 1$ 
    if  $n_q = 0$  then
       $q_{rank} = i + 1$ 
       $Q = Q \cup \{q\}$ 
   $i = i + 1$ 
 $\mathcal{F}_i = Q$ 

```

where domination count is n_p , S_p is the set of p dominating solutions, q is the member of S_p , and Q is the list of zero dominated q .

Step 5: Simulated binary crossover [236] and polynomial mutation were used to create new offspring population (Q) of size N .

Step 6: A combined population of R_t was created using the offspring Q_t and the parent population P_t . Thus, the size of R_t becomes $2N$.

Step 7: Fast non-dominanted sort was used to sort the entire population in the same was as in Step 4.

Step 8: Calculate the crowding distance using the method defined by Deb et al. [230].

CROWDING DISTANCE ASSIGNMENT (\mathcal{T})

$l = |\mathcal{T}|$

for each i , set $\mathcal{T}[i]_{distance} = 0$

$\mathcal{T} = \text{sort}(\mathcal{T}, m)$

$\mathcal{T}[1]_{distance} = \mathcal{T}[l]_{distance} = \infty$

For $i = 2$ to $(l - 1)$

$\mathcal{T}[i]_{distance} = \mathcal{T}[i-1]_{distance} + (\mathcal{T}[i+1].m - \mathcal{T}[i-1].m) / (f_m^{max} - f_m^{min})$

where \mathcal{T} is a non-dominated set, $\mathcal{T}[i].m$ is the m^{th} objective function value of the i^{th} individual in the set \mathcal{T} , f_m^{max} and f_m^{min} are the maximum and minimum of m^{th} objective function.

Step 9: The combined population, Q_t , was sorted according to a non-dominant sort and crowding distance. If the population size (from first F_1 to last F_l front) was greater than N then a crowded-comparison operator, $<_n$, was used in descending order to populate the population size until N from F_l and others ($F_{>l}$) are discarded. The partial order $<_n$ was given by

$$i <_n j \text{ if } (i_{rank} < j_{rank}) \text{ or } ((i_{rank} == j_{rank}) \text{ and } (i_{distance} > j_{distance}))$$

Step 10: Keep the number of elements (N) from the sorted list and increased the number of generations.

Step 11: Repeat Step 5 onwards until the termination condition (50 generations were produced) was met.

6.2.3. Performance of the hyperparameter optimization

The algorithm was implemented in MATLAB and run on a computer with Intel Core (TM) i7-8700k processor, 64 GB RAM, and two NVIDIA GeForce GTX 1080 Ti GPUs. Two-fold runs were processed in parallel in the two GPUs and the average of the obtained results of the objective functions was computed. The optimizations were carried out with three different input sizes, with the termination condition of producing 50 generations with population size of 50, which leads to $(50(\text{Gen}) * 50(\text{Pop} = \text{offspring population} = N))$ 2500 different networks and 5000 networks to train for each input size (because a twofold method was employed).

The computation took 587.83926 hours (≈ 24.49 days), 832.04399 hours (≈ 34.67 days) and 911.226716 hours (≈ 37.97 days), respectively, for 1 minute, 3 minute and 5 minute input to finish the simulations. The original networks, trained with the HuGCDN2008 dataset, have two solutions for each network structure where hyperparameter optimization was done. These solutions are indicated in Table 25 by _F1 and _F2 for the fold one and the fold two dataset. Different input sizes are indicated as _1, _3 and _5 for 1 minute, 3 minute, and 5 minute respectively.

Table 25 : The results of both CNNs trained using the HuGCDN2008 database in two-fold.

Symbol	Test Database	Sen	Spc	Acc
CNN1DF1_1	HuGCDN2008	71.50	95.30	88.50
CNN1DF2_1	HuGCDN2008	73.60	93.10	88.60
CNN1DF1_3	HuGCDN2008	73.70	95.10	88.90
CNN1DF2_3	HuGCDN2008	74.40	94.10	89.50
CNN1DF1_5	HuGCDN2008	75.10	94.90	89.20
CNN1DF2_5	HuGCDN2008	74.40	93.90	89.40
CNN1DF1_1	AED [61]	91.64	93.36	92.65
CNN1DF2_1	AED [61]	87.89	92.54	90.63
CNN1DF1_3	AED [61]	83.77	93.62	89.58
CNN1DF2_3	AED [61]	85.64	93.36	90.20
CNN1DF1_5	AED [61]	79.03	93.62	87.64
CNN1DF2_5	AED [61]	88.58	93.67	91.58
CNN1DF1_1	UCD[62]	56.72	93.32	84.55
CNN1DF2_1	UCD[62]	64.12	90.69	84.33
CNN1DF1_3	UCD[62]	67.35	90.51	84.96
CNN1DF2_3	UCD[62]	82.51	76.30	77.79
CNN1DF1_5	UCD[62]	66.57	90.19	84.53
CNN1DF2_5	UCD[62]	85.37	60.94	66.79

In the first step, the algorithm generates a random population to ensure the diversity of the population and ranks them according to the multi objectives optimization (Acc, Sen, and Spc) [231]. Over each generation, using mutation and crossover, the algorithm was able to reach better solutions. Almost all the solutions of the 50th generation were better than the 1st Gen (Figure 40, Figure 41, Figure 42, and Figure 43). From the 50th Gen solutions, among three different inputs, it is noticeable that the 3 minute and 5 minute inputs have better results compared to 1-minute solutions. However, both of them are in a similar range (Figure 44).

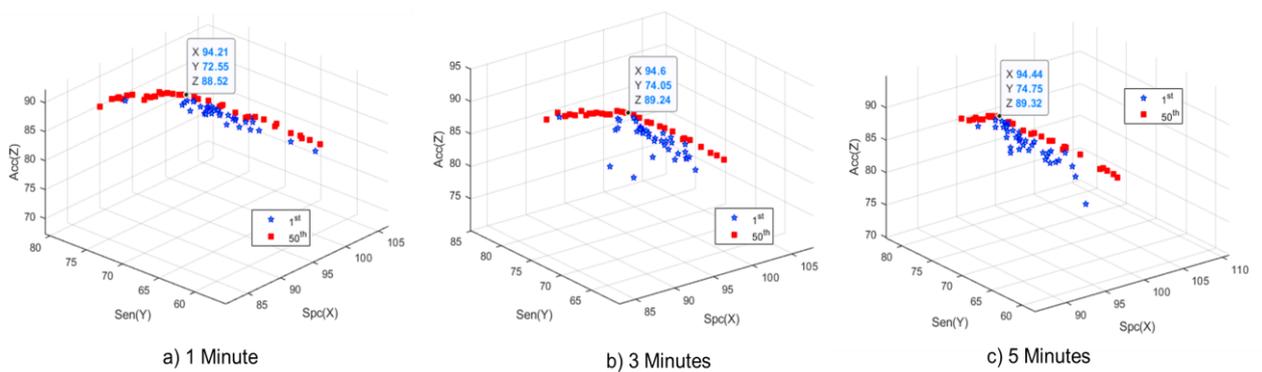


Figure 40 : The multi objective problem space in percentage for the 1st and the 50th generation for a) 1 minute b) 3 minute, and c) 5 minute inputs for the HuGCDN2008 database.

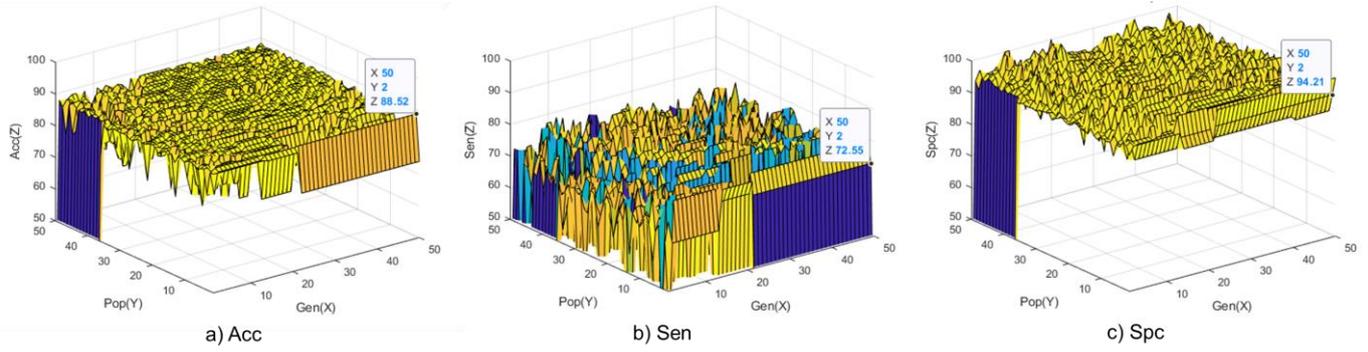


Figure 41 : The changes of the multi objectives in percentage, a) Acc, b) Sen, and c) Spc, over the generations (Gen) of the populations (Pop) for 1 minute input for the HuGCDN2008 database.

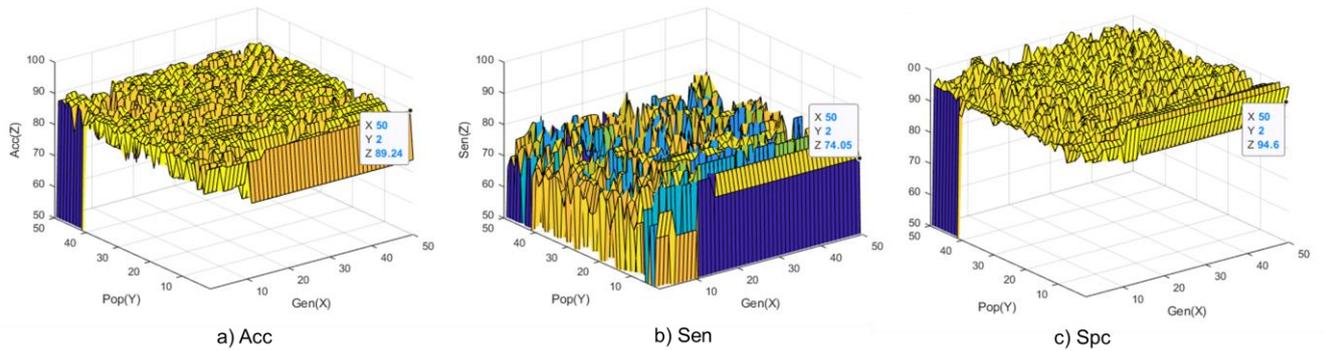


Figure 42 : The changes of the multi objectives in percentage, a)Acc, b)Sen, and c)Spc, over the generations (Gen) of the populations (Pop) for 3 minutes input for the HuGCDN2008 database.

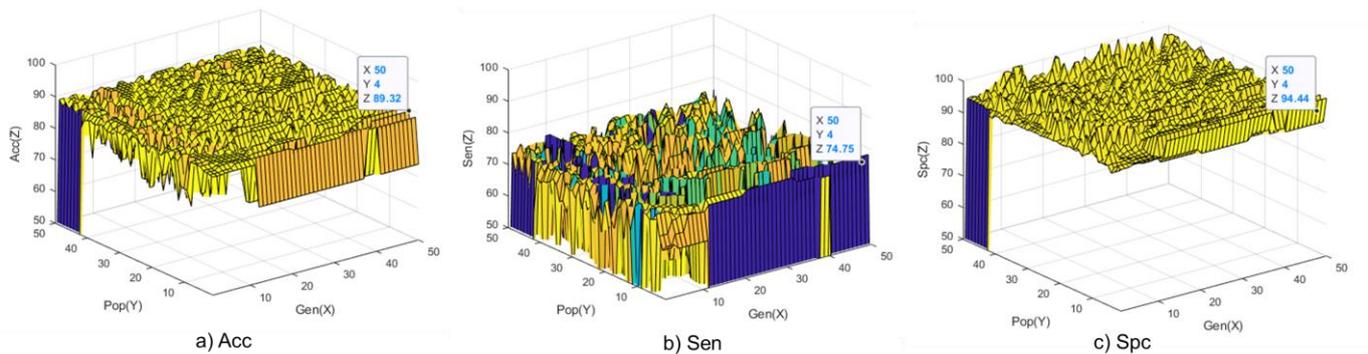


Figure 43 : The changes of the multi objectives in percentage, a)Acc, b)Sen, and c)Spc over the generations (Gen) of the populations (Pop) for 5 minutes input for the HuGCDN2008 database.

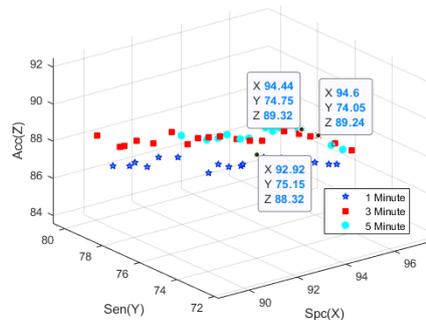


Figure 44 : Comparison of the three inputs in the 50th generation in 3D (Acc, Sen, and Spc) problem space.

The NSGA-II algorithm ranked the outputted solution according to Pareto front numbers. All the solutions on the first Pareto front are valid solutions as NSGA II does not generate a single solution but a set of Pareto non-dominated solutions. One way of assessing the classifiers' performance is the receiver Operating Characteristic (ROC) curve (Sen vs 1- Spc). Since the algorithm uses a three dimensional problem space (Figure 40 and Figure 44) a modified version of the ROC curve (with 2 dimensions), where all of the first Pareto front solutions are shown in Figure 46 and Figure 46. At the 50th generation, it was assessed that 5 minute has the best solution followed by a 3 minute and a 1 minute CNN (Figure 44 and Figure 46).

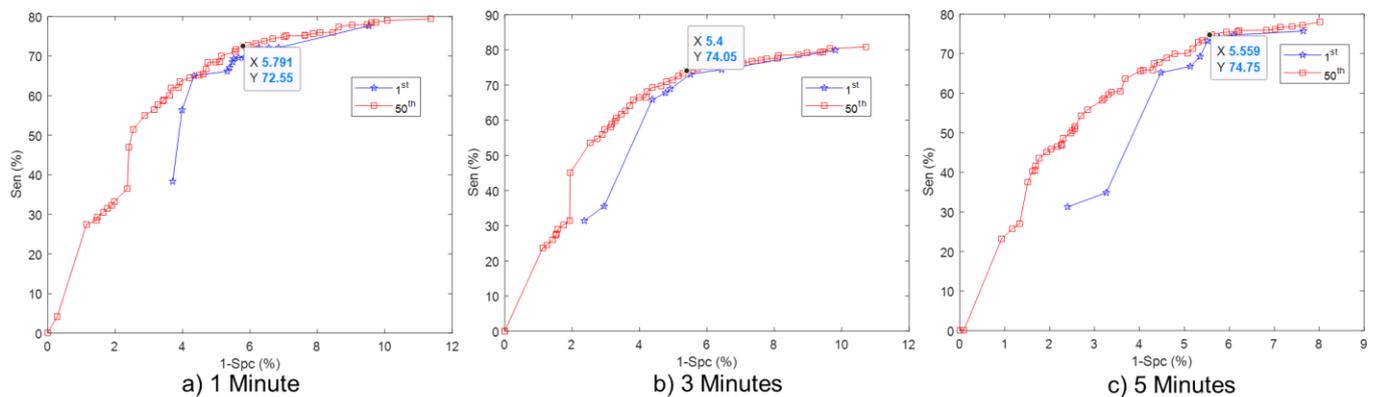


Figure 45 : 1st Pareto front of the solutions for the HuGCDN2008. The first (1st) and last (50th) generation of a) 1 minute b) 3 minute c) 5 minute Spo2 signal. The solution of the first generation is marked with star and the 50th is marked with box.

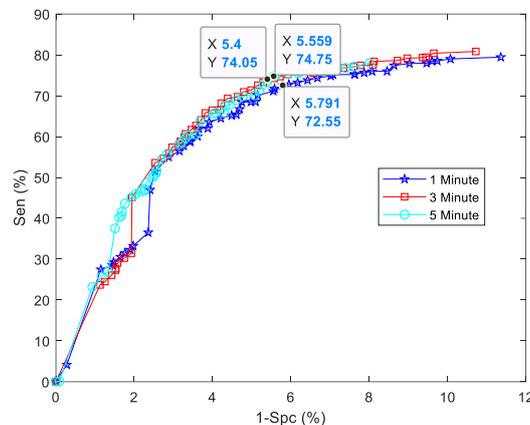


Figure 46 : Comparison of three inputs in the 50th generation in 2D (Sen and 1-Spc) problem space.

Although the algorithm was trying to solve a multi-objective optimization, NSGA-II treated the optimization variables (Acc, Sen and Spc) equally, due to a restriction with the problem space and constraints such as overlap of apnea and normal events (where the Sen and Spc are dependent and Acc is dependent on Sen and Spc). The solutions do not have equal Sen Spc and Acc. Therefore, the final solution was chosen with the highest Acc (structure in Table 26) among all the valid solutions which will also help in the comparison with the other methods presented in the literature. By using Acc as a criteria of choosing one solution over others, three solutions for a 1 minute, 3 minutes, and 5

minute inputs produced an Acc of 88.2%, 89.24% and 89.32%, a Sen of 72.55%, 74.05% and 74.75% and aSpC of 94.21%, 94.60% and 94.44% respectively (Table 28). These solutions are marked with a black dot and the values are indicated in a box in Figure 44 and Figure 46.

Table 26 : Chosen CNN’s layers and hyperparameters (layer parameters such as the input size, size of the filter, number of the filters represented as a form of [number of filter]@ [vertical width of filter]x [horizontal width of filter]x [number of Channels of filter]_ [vertical width of stride]x[horizontal width of stride]).

		1 minute (60 seconds)		3 minutes (180 seconds)		5 minutes (300 seconds)	
No.	Layer	Layer Parameters	Layer	Layer Parameters	Layer	Layer Parameters	
L1	Input	1x3000x1	Input	1x9000x1	Input	1x15000x1	
L2	Conv_1	15@1x9x1_1x5	Conv_1	15@1x9x1_1x5	Conv	6@1x7x1_1x3	
L3	Batchnorm_1	15 channels	Batchnorm_1	15 channels	Batchnorm_1	6 channels	
L4	ReLU_1		ReLU_1		ReLU_1		
L5	MaxP_1	1x5_1x5	MaxP_1	1x5_1x5	MaxP_1	1x3_1x3	
L6	Conv_2	8@1x8x12_1x3	Conv_2	12@1x8x15_1x3	Conv_2	14@1x9x6_1x9	
L7	Batchnorm_2	8 channels	Batchnorm_2	12 channels	Batchnorm_2	14 channels	
L8	ReLU_2		ReLU_2		ReLU2		
L9	Conv_3	8@1x8x12_1x3	Conv_3	9@1x8x12_1x3	MaxP_2	1x3_1x2	
L10	Batchnorm_3	8 channels	Batchnorm_3	9 channels	Conv_3	12@1x7x14_1x1	
L11	ReLU3		ReLU3		Batchnorm_3	12 channels	
L12	Conv_4	13@1x4x8_1x2	Conv_4	14@1x4x9_1x2	ReLU_3		
L13	Batchnorm_4	13 channels	Batchnorm_4	14 channels	MaxP_3	1x3_1x2	
L14	ReLU_4		ReLU4		Conv_4	15@1x5x12_1x3	
L15	FC1	117 FCL	FC1	124 FCL	Batchnorm_4	15 channels	
L16	ReLU_5		ReLU_5		ReLU_4		
L17	FC2	2 FC2	FC2	2 FC2	FC1	99 FCL	
L18	Softmax		Softmax		ReLU_5		
L19	Classoutput		Classoutput		FC2	2 FC2	
L20					Softmax		
L21					Classoutput		

The Numbers of Flexible Layers (NoFL) were 5, 5 and 7 (Figure 47) resulting in 19, 19 and 21 layers for 1 minute, 3 minute, and 5 minute CNN networks, respectively (Table 26). The layer sequence in 1 minute and 3 minute CNNs is also similar where first Conv layer (L2) and Batchnorm (L3) are the same for both. However, the remaining layers have more kernels or channels. The five minute CNN has two more flexible layers and the actual layers are in the form of Maximum pooling (L9, L13). It has the same number of conv, ReLU and Batchnorm layers as the 1 minute and 3 minute CNNs, except that the number of kernels and channels in the first layer is higher.

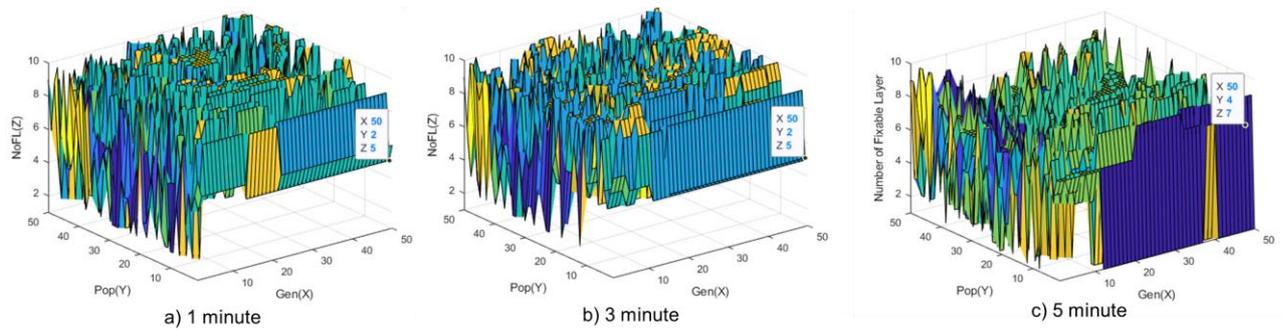


Figure 47 : The solutions for the first (1st) and the last (50) generation of a) 1 minute b) 3 minute, and c) 5 minute SpO2 signal. The solution of first generation is marked with a star and 50th is marked with a box. d) comparison of all of the inputs in 50th generation.

6.2.4. External database performance

To check the universality of the system, the trained (in the HuGCDN2008 database) CNNs were tested in the AED [61] and UCD [62] databases. The results are presented in Table 28. The performance of all three best networks with the UCD [62] database were lower than the originally trained database but were higher in the AED [61]. The highest accuracy, 92.65%, was achieved with 1 minute input in the AED [61] database. For the HuGCDN2008 and the UCD [62] databases, the main difficulty lay in the detection of short apnea events. This could be related to the fact that some respiratory pauses do not produce a clear pattern in the oximetry signal. This could be related to the hemoglobin dissociation curve where short events would not be able to decrease the SpO2 percentage because a marked reduction in the partial oxygen pressure did not occur. Additionally, pH, temperature and the 2,3-diphosphoglycerate (2,3-DPG) levels, which are specific to each person, can displace the hemoglobin dissociation curve [10].

6.2.5. Effect of input size

With an increase of the input size (from 1 to 5 minutes), the performance of the HuGCDN2008 dataset was improved slightly in Acc, from 88.52% to 89.28% and 89.32%. However, between 3 minute and 5 minute, the results were almost the same. The Sen (apnea events) was affected by the input size with an improvement of more than 2% when compared with the 1 minute and 5 minute inputs. The Spc (normal events) remains almost the same. By increasing the Sen and keeping the Spc stable, the classifiers were able to increase the Acc. A possible reason to justify why longer inputs achieve better Sen could be related to the fact that an apnea event could be present in different minutes; thus, having the information of longer apnea events increases the detection capabilities. Another reason could be, as indicated in other works [10], higher (five) minutes allow the spectral features to show more relevance. However, in other datasets (AED [61] and UCD [62]) this trend was not consistent. For the AED [61] dataset the highest Acc, of 92.65%, and Sen, of 91.64%, were

achieved by the 1 minute input and for the UCD [62] the best results, Acc of 84.96% and Sen of 67.35%, were achieved by the 3 minute input. Occasionally, one network of higher input size performs worse than one of lower input size. Therefore, the performance parameters are more dependent on the data and the training weight, than on the input size.

6.2.6. Effect of layers

Due to the success of big (deeper) networks one can assume that an increase in the number of layers can provide better results in the case of deep learning. However, this assumption is not always valid. The number of chosen layers for each solution can be seen in Figure 47. By analyzing the figures, it is possible to assess that the algorithm attempted different layer sizes to solve the problem and a better solution did not have the highest number of layers (NoFL). A similar conclusion was presented by Urtnasan et al. [180] [179] which has an occurrence of an optimum six layered CNN while testing from 3 to 9 layers.

6.2.7. Transfer learning performance

Transfer learning could be useful by using the information learned from one problem and implementing it on others. This work mainly focused on OSA detection, thus the transfer learning performance was analyzed with the AED [61] and UCD [62] databases while the main network was trained using the HuGCDN2008 database.

For transfer learning, the last three layers (L 17-19 for 1 minute, and 3 minute, L 19-21 for 5 minute, as indicated in Table 26) were removed and replaced with similar types of layers. Afterwards, they were retrained with the leave one out method (due to their low number of subjects). There were two different weighted networks for each CNN input network (generated using two-fold methods implemented in the HuGCDN2008 dataset). Because the actual training data for these transfer learning networks were coming from different dataset, there is no need to find the average of the two networks' results (as with the two-fold methods in the HuGCDN2008 dataset). Therefore, only the best networks of the HuGCDN2008 dataset (Table 25) were chosen for transfer learning.

It was verified that, in all of the cases, the transfer learned networks had better accuracy (Table 25 vs Table 27). However, there were database dependencies. In some of the cases, it was not the same original network trained with the HuGCDN2008 dataset (e.g. the CNN1DF2_1 second fold network for 1 minute in Table 25 which performed the best; however after transform learning CNN1DF1_1 did best in Table 27 for the AED dataset).

Table 27 : Transfer learning (TL).

Symbol	Database	Sen	Spc	Acc
CNN1DF1_1	AED [61]	92.04	95.78	94.24
CNN1DF2_1	AED [61]	89.25	94.65	92.42
CNN1DF1_3	AED [61]	92.79	94.56	93.83
CNN1DF2_3	AED [61]	89.87	96.78	93.93
CNN1DF1_5	AED [61]	91.49	95.52	93.86
CNN1DF2_5	AED [61]	87.76	96.61	92.96
CNN1DF1_1	UCD[62]	54.39	94.14	84.52
CNN1DF2_1	UCD[62]	58.32	93.32	84.85
CNN1DF1_3	UCD[62]	60.02	93.93	85.73
CNN1DF2_3	UCD[62]	60.38	93.90	85.79
CNN1DF1_5	UCD[62]	60.42	93.43	85.44
CNN1DF2_5	UCD[62]	60.34	93.54	85.51

6.2.8. Comparison with the state of the art works

The closest match for a comparison with this work was developed by Ravelo-García et al. [10] where the same database, HuGCDN2008, was used. A shallow classifier, linear discriminant analysis (LDA), was employed with an SpO2 signal and combination of SpO2 and HRV. The proposed work achieved 89.32% Acc with only the SpO2 5 minute's window compared to 86.5% and 86.9% with mix of 1 minute and 5 minute windows using the SpO2, and SpO2+HRV signals. The proposed implementation was also able to keep the same performance level with a 3 minute window and not sacrificing any parameters. Even the one-minute window has better Acc and Sen compared to the other works in Table 28. Our feature based approach, SC3 (Section 5.4 and Section 5.5), also used the HuGCDN2008 database. The accuracy and specificity of a multi-objectively optimized CNN is better.

For the AED [61] dataset the proposed optimized CNN achieved 92.65% Acc, 93.36 % Spc and 91.64% Sen. Though the Acc was not the best among the other implementations, it has one of the best Sen, only surpassed by an LSTM [107] and ANN [102]. However, neither of these works, [102][107], were subject independent. For the UCD [62], the transfer learning approach achieved the highest accuracy compared to the other works except the DAE [106] likewise was not subject independent. In both databases, transfer learning increases the performance parameters.

If the comparison only includes deep learning, the proposed networks achieved the best accuracy among all the subject independent implementations. This is the case even when they are compared to some implementation where more signals were employed, such as a combination of SpO2, airflow and respiration [190] or the combination of SpO2, oronasal airflow and movements (ribcage and abdomen) [185].

Table 28 : Comparison with the literature (P is for proposed networks optimized by a GA and trained using the HuGCDN2008 database. TL indicates transfer learning where the proposed networks were retrained using the respective database. ^aBetween two networks the best one is showed.).

Ref.	Signal	Classifier	Database	Recordings or Patients	Input size (Seconds)	No of features	Sen	Spc	Acc
[2]	SpO2	SVM	Own	40	150	7	-	-	90
[73]	SpO2 +ECG	Bagging.RepTree	UCD [62]	25	60	39	79.75	85.89	84.80
[73]	SpO2	Bagging.RepTree	UCD [62]	25	60	39	78.23	84.25	82.79
[99]	SpO2	ANN	AED [61]	8	60	3	87.5	100	90.3
[10]	SpO2	LDA	HuGCDN2008	70	60,300	19	75.6	91.00	86.5
[10]	SpO2+HRV	LDA	HuGCDN2008	70	60,300	33	73.4	92.3	86.9
[195]	SpO2	SVM-L	AED [61]	8	60	50	83.76	97.03	96.89
(Section 5.2)									
[195]	SpO2	SVM-L	AED [61]	8	60	20	84.57	97.28	97.38
(Section 5.2)									
[195]	SpO2	ANN	UCD [62]	25	60	2	43.31	95.03	81.95
(Section 5.2)									
[195]	SpO2	LD	UCD [62]	25	60	9	61.78	91.03	83.27
(Section 5.2)									
[102]	SpO2	ANN	AED [61]	8	60	7	96.5	98.5	97.7
(Section 5.3)									
[196]	SpO2	MaxVSF1	AED [61]	8	60	34	98.11	86.98	91.33
(Section 5.4)									
[106]	SpO2	DAE	AED [61]	8	60	-	78.75	95.89	97.64
[106]	SpO2	DAE	UCD [62]	25	60	-	60.36	91.71	85.26
[107]	SpO2+IHR	LSTM	AED [61]	8	60	-	84.7	-	92.1
[107]	SpO2	LSTM	AED [61]	8	60	-	92.9	-	95.5
[185]	SpO2+oronasal airflow+ribcage and abdomen movements.	CNN2D	UCD[62]	23	1	-			79.6
[190]	SpO2+airflow+ respiration.	RCNN	MGH	10000	1	-			88.2
[190]	SpO2+airflow+ respiration.	RCNN	SHHS	5804	1	-			80.2
[196]	SpO2	MaxVIF1	HuGCDN2008	70	60	98	82.48	86.28	85.30
(Section 5.4)									
[196]	SpO2	MaxVIF1	HuGCDN2008	70	300	89	80.54	85.74	84.46
(Section 5.4)									
(Section 5.5)	HR	MaxVSF5	HuGCDN2008	70	300	222	72.81	75.25	76.17
(Section 5.5)	SpO2+HR	MaxVSF5	HuGCDN2008	70	300	290	78.90	87.21	85.08
P	SpO2	CNN1D	HuGCDN2008	70	60	-	72.55	94.21	88.52
P	SpO2	CNN1D	HuGCDN2008	70	180	-	74.05	94.60	89.24
P	SpO2	CNN1D	HuGCDN2008	70	300	-	74.75	94.44	89.32
P ^a	SpO2	CNN1D	AED [61]	8	60	-	91.64	93.36	92.65
P ^a	SpO2	CNN1D	AED [61]	8	180	-	85.64	93.36	90.20
P ^a	SpO2	CNN1D	AED [61]	8	300	-	88.58	93.67	91.58
P ^a	SpO2	CNN1D	UCD[62]	25	60	-	56.72	93.32	84.55
P ^a	SpO2	CNN1D	UCD[62]	25	180	-	67.35	90.51	84.96
P ^a	SpO2	CNN1D	UCD[62]	25	300	-	66.57	90.19	84.53
TL ^a	SpO2	CNN1D	AED [61]	8	60	-	92.04	95.78	94.24

TL ^a	SpO2	CNN1D	AED [61]	8	180	-	89.87	96.78	93.93
TL ^a	SpO2	CNN1D	AED [61]	8	300	-	87.76	96.61	92.96
TL ^a	SpO2	CNN1D	UCD[62]	25	60	-	58.32	93.32	84.85
TL ^a	SpO2	CNN1D	UCD[62]	25	180	-	60.38	93.90	85.79
TL ^a	SpO2	CNN1D	UCD[62]	25	300	-	60.34	93.54	85.51

6.2.9. Summary

The goal of this work was to develop and test a novel fully automated hyperparameters optimization algorithm for a CNN. Consequently, significant results were attained.

Three different window sizes were also tested and it was verified that there is almost no difference between 3 minute and 5 minutes window sizes. In some cases, the 1 minute outperformed the 3 minute and 5 minute inputs. Compared to shallow networks, the developed CNNs were able to achieve a better performance with a smaller input size and without the need for hand crafted feature extraction.

It was also verified that the performance of the models with almost similar structure networks was more sensitive to training and data than the hyperparameters choice. Also, it was verified that transfer learning has a strong potential for implementation in similar domains.

One of the limitations of this work is the fact that multi objective optimization was only applied to hyperparameter optimization and not used for the training. Thus, when the transfer learning concept was implemented, the network was sacrificing Sen to achieve a better Acc. The second limitation is the population number, of only 50, which cannot ensure that the network had a strong diversity to start with. However, this issue was mitigated by the use of mutation. This work was not designed to be optimized for the layer size. Thus, even in the 50th generation, the Pop has a substantial different sized network. One way of solving this issue would be to run for more generations until stable solutions were found. Another way of achieving this could be involve an NoFL, which is one of the objectives which are under consideration for future research. It was verified in the literature, that increasing the number of signals [183] or selecting an RNN [189] could improve the results [179] [180]. Therefore, this could be investigated in the future.

6.3. Greedy Based Optimization (GBO) of CNN

The optimization technique developed for CNNs in Section 6.2 is to a certain extent time-consuming. Therefore, a faster technique is investigated in this section.

6.3.1. Introduction

A deep network classifier, capable of automatically extracting features, was used. There are different types of deep networks used to detect OSA: CNNs [179][180][181][182], DNNs [177], LSTMs [107][132][131], the Gated Recurrent Unit (GRU) [189], among others. The literature has shown encouraging results by using a CNN as a deep network classifier [179][180][181][182]. For that reason, in this work, a CNN is used for detecting apnea from the SpO₂ signal. Since the input of the CNN has just one dimension, the proposed deep network classifier is called CNN1D and is used by different works in the literature [179][219][180]. The optimal CNN structural hyperparameters for a suitable classifier are one of the most challenging topics, which need time and knowledge. Different authors used different techniques such as an evolutionary algorithm [237][238][239], a Bayesian optimization [240], an efficient framework for hyperparameters [220] and a sequentially structured search [241] [242]. To solve this limitation, in this work it is also proposed that a fast searching mechanism for optimizing hyperparameters for CNN1D be used.

Two databases were preprocessed to test the performance and the idea behind this work. A scalable structure was designed with modular cluster layers. The classifier is scaled with added cluster layers by using the proposed algorithms.

6.3.2. Classifier structure

The CNN topology proposed to detect apnea is schematically presented in Figure 48. This topology is composed of two different layers: a Clustered-layer (CL) and an Output Layer (OL). The OL is applied for the classification phase and it is composed of the following sub-layers sequentially connected: Fully Connected (FC), softmax and class output. The proposed CL is applied to extract local features at high resolutions and reduces the dimensionality of the complex features.

The CL is constituted of a set of layers methodologically arranged in order to detect, extract, and decrease the variation of layers and to facilitate the training process. In the first layer of the CL, convolution operations (ConV) are used to extract features. Convolution preserves the spatial relationship between inputs by learning features using trainable kernels. By nature, CNN systems have several layers that are sensitive to initial trainable kernels and configurations of the training algorithm. One possible reason for this sensitiveness is the distribution of the inputs between the deep layers, which may change after each mini-batch when the trainable parameters are updated. Consequently, the learning algorithm will chase a moving target. This phenomenon is technically called the internal covariate shift and it can be solved through batch normalization [243][244]. Batch normalization normalizes each input channel across a mini-batch. By using the Batch Normalization (BatchN) between convolution operations and nonlinearities (ReLU), the CNN training speeds up and

its sensitivity towards initial trainable parameters is reduced. Then, MAX-Pooling (MaxP) is applied for down-sampling feature maps, reducing sharp variations and the number of connections to the between layers [245].

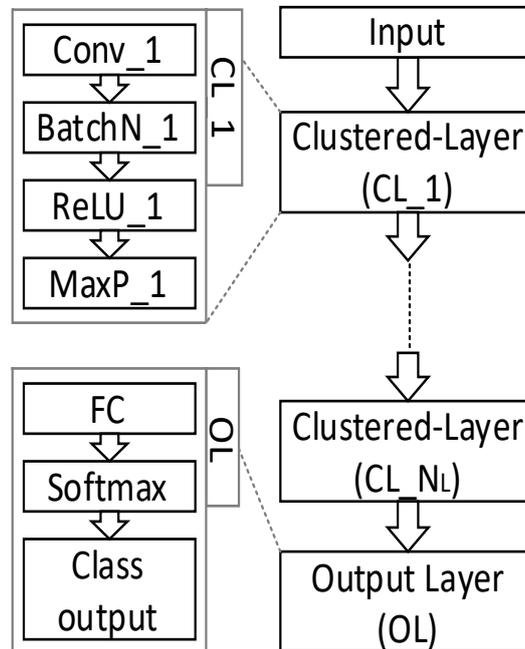


Figure 48 : Clustered-layer (CL) base scalable CNN1D structure.

6.3.3. Hyper-parameters optimization based on Greedy Algorithm

To find the best CNN hyper-parameters for a problem, expert knowledge and an enormous amount of research time are needed. In this work, a Greedy Based Optimization (GBO) algorithm with variants of topology transfers is proposed. Therefore, the objective is to minimize the searching time and to introduce reproducible procedures for a well-structured searching optimization.

The inspiration for the proposed algorithm is based on deep belief nets (DBN) developed by GE Hinton et al. [23]. As presented in DBNs, the proposed GBO algorithm optimizes the network layer by layer. While the DBN uses unsupervised learning for trainable parameters without optimizing the layers' hyper-parameters, the developed GBO algorithm uses a supervised learning with the advantage of hyper-parameters optimization for each added CL. After running the GBO algorithm, it is possible to find the optimum number of CLs. The proposed GBO algorithm is divided mainly into two different methods: The Topology Transfer (TT) method and the Weighted-Topology Transfer (WTT) method. The main methodology of both methods remains the same. However, some details about how they are trained are different from one another. The TT method saves just the topology and its weights need to be re-trained when a new CL is added. The WTT method could be divided into two sub-methods. The first sub-method, called Rough Estimation (RE), consists of saving the

topology and weights, and then, training just the new CL when it is added. The second method, called Fine Tuning (FT), consists of fine tuning by retraining only the best networks obtained by the RE.

If a CNN is composed of a sequence of CL $l = 1, 2 \dots N_L$ and, at the end, an output layer (OL_l). The optimized number of CL and its parameters need to be found. The ideal procedure is an Exhaustive Search (ES), where all possible combinations are developed to find the best CNN. The following equation gives all these combinations of the ES.

$$\sum_{l=1}^{N_L} A^l \quad (13)$$

where N_L is the maximum number of CL and $A = S_k * N_k$. N_k is the number of kernels of the user defined set kNo and S_k is the quantity of different kernels sizes of the user defined set $kSize$. For a specific kernel number and size, n and m are used (like m^{th} and n^{th}).

The proposed method assembles layers based on the GBO algorithm. It is a technique to decrease the time to search for the best number of CLs and to optimize the parameters of each CL.

Firstly, CNNs with just one CL and OL are trained for N_k different kernels sizes and S_k different kernel numbers. Then, the structure with the best performance is saved. From the previously found structure, a new CL_l is added between the previous CL_{l-1} and the OL_l . The new structure of the added CL is searched in a similar way to that previously described. This procedure is carried out for N_L clustered layers given the number of combinations when the proposed method is used.

$$N_L * A \quad (14)$$

In contrast to the ES method, the proposed GBO method allows reducing the number of combinations for finding the best CNN and, consequently, the time for searching the best CNN. The gain over all simulated networks is quantified by

$$G_{ES/GBO} = \frac{\sum_{l=1}^{N_L} A^l}{N_L * A} = \frac{A \sum_{l=0}^{N_L-1} A^l}{N_L * A} = \frac{A^{N_L-1+1}-1}{N_L(A-1)} = \frac{A^{N_L-1}}{N_L(A-1)} \quad (15)$$

Assuming that $A^{N_L} \gg 1$ and $N_L > 1$ since it is a deep learning structure, the above can be simplified to

$$G_{ES/GBO} \approx \frac{A^{N_L}}{N_L(A-1)} = \frac{(S_k * N_k)^{N_L}}{N_L(S_k * N_k - 1)} \quad (16)$$

The TT and WTT methods take $Data = \{X_{train}, Y_{train}, X_{val}, Y_{val}, X_{test}, Y_{test}\}$ as input, where X_{train} and Y_{train} are the training data and the training annotation, X_{val} and Y_{val} are the validation data and the validation annotation and X_{test} and Y_{test} are the test data and the test annotation. For both methods $S_k = n(kSize)$ and $N_k = n(kNo)$ which are the cardinality of the user specified $kSize$ and kNo set. For all the combination of S_k and N_k , A , the TT and WTT methods perform their specific steps. For the first CL ($l = 1$), the TT and WTT do not have any difference. From the second added CL, the difference is in the network creation and the training process.

The TT method has six stages executed sequentially (Figure 49): layer creation, layer concatenation, the training process and the testing process:

<pre> TT (<i>Data</i>, N_L, <i>kSize</i>, <i>kNo</i>) $S_k = n(kSize)$ and $N_k = n(kNo)$ for $l = 1$ to N_L for $m = 1$ to S_k for $n = 1$ to N_k //Layer Creation: If $l > 1$ $Net_{l,m,n}^f \leftarrow$ Topology ($BNet_{l-1}$) – [BOL_{l-1}] $Net_{l,m,n}^{CL} \leftarrow CL_{l,m,n}[+]OL_l$ Else $Net_{l,m,n}^f \leftarrow IL$ $Net_{l,m,n}^{CL} \leftarrow CL_{l,m,n}[+]OL_l$ end //Layer Concatenation: $Net_{l,m,n} \leftarrow Net_{l,m,n}^f[+]Net_{l,m,n}^{CL}$ //Training Process: $Net_{l,m,n} \leftarrow$ train($Net_{l,m,n}$, X_{train}, Y_{train}, X_{val}, Y_{val}) //Testing Process: $P_l \leftarrow test(Net_{l,m,n}, X_{test}, Y_{test})$ end end $BNet_l \leftarrow Best(Net_l)$ end //Best Network Selection: $Net_{Best} \leftarrow Best(BNet)$ Return Net_{Best} </pre>	<pre> WTT (<i>Data</i>, N_L, <i>kSize</i>, <i>kNo</i>, <i>Type</i>) $S_k = n(kSize)$ and $N_k = n(kNo)$ for $l = 1$ to N_L for $m = 1$ to S_k for $n = 1$ to N_k //Layer Creation: If $l > 1$ $Net_{l,m,n}^f \leftarrow$ Topology & weight ($BNet_{l-1}$) – [BOL_{l-1}] $Net_{l,m,n}^{CL} \leftarrow CL_{l,m,n}[+]OL_l$ Else $Net_{l,m,n}^f \leftarrow IL$ $Net_{l,m,n}^{CL} \leftarrow CL_{l,m,n}[+]OL_l$ end //Training Process: $X_{train}^f \leftarrow Net_{l,m,n}^f(X_{train})$ // feature extraction $X_{val}^f \leftarrow Net_{l,m,n}^f(X_{val})$ // feature extraction $Net_{l,m,n}^{CL} \leftarrow$ train($Net_{l,m,n}^{CL}$, X_{train}^f, Y_{train}, X_{val}^f, Y_{val}) //Layer Concatenation: $Net_{l,m,n} \leftarrow Net_{l,m,n}^f[+]Net_{l,m,n}^{CL}$ //Testing Process: $P_l \leftarrow test(Net_{l,m,n}, X_{test}, Y_{test})$ end end $BNet_l \leftarrow Best(Net_l)$ end If <i>Type</i> == 'RE' //Best Network Selection $Net_{Best} \leftarrow Best(BNet)$ ElseIf <i>Type</i> == 'FT' for $l = 1$ to N_L //Re Training Process $BNet_l^r \leftarrow retrain(BNet_l, X_{train}, Y_{train})$ //Testing Process $P_l^r \leftarrow test(BNet_l^r, X_{test}, Y_{test})$ end //Best Network Selection $Net_{Best} \leftarrow Best(BNet^r)$ end Return Net_{Best} </pre>
---	---

Figure 49 : Side by side comparison of the TT and WTT greedy methods.

i) Layer creation: A partial network, $Net_{l,m,n}^f$ is created from the previous best network by

removing the output layer, OL_{l-1} . Furthermore, in this stage, the partial network $Net_{l,m,n}^{CL}$ is created by the concatenation of a new $CL_{l,m,n}$ and a new OL_l . In the algorithms the '[+]' symbol is used for concatenate and the ']-[' symbol is used for reverse concatenate.

ii) Layer concatenation: A network $Net_{l,m,n}$ is created by the concatenation of $Net_{l,m,n}^f$ and $Net_{l,m,n}^{CL}$.

iii) Training process: $Net_{l,m,n}$ is trained with training data X_{train} , and its annotation, Y_{train} , and is validated with validation data X_{val} , and its annotation Y_{val} . In this process, the training is done for N_k different kernels with S_k different size kernels. In total, $S_k * N_k$ different networks are trained in this stage.

iv) Testing process: the $S_k * N_k$ different networks are tested.

v) The procedures described above are repeated N_L times. vi) In the end, the best network is chosen from N_L different networks.

As mentioned before, the WTT method has two sub-methods (Figure 49): the RE and the FT($type = \{ 'RE', 'FT' \}$).

i) Layer creation: A partial network $Net_{l,m,n}^f$ is created from the previous best network by removing the output layer, OL_{l-1} . Also, in this stage the partial network $Net_{l,m,n}^{CL}$ is created by the concatenation of a new $CL_{l,m,n}$ and a new OL_l .

ii) Training process: only the partial network $Net_{l,m,n}^{CL}$ is trained in this stage. Firstly, the feature map X_{train}^f and X_{val}^f are obtained from the simulation of the partial network $Net_{l,m,n}^f$. Then, X_{train}^f along with Y_{train} and X_{val}^f along with Y_{val} are used for training the partial network $Net_{l,m,n}^{CL}$. In total, $S_k * N_k$ different networks are trained in this stage.

iii) Layer concatenation: the network $Net_{l,m,n}$ is created by the concatenation between $Net_{l,m,n}^f$ and $Net_{l,m,n}^{CL}$.

iv) Testing process: the $S_k * N_k$ different networks are tested.

v) The procedures described above are repeated N_L times.

vi) if the $type == 'RE'$ then the best network is chosen from N_L different networks. Elseif $type == 'FT'$ then

vi.a) all N_L best networks from RE, $BNet_l$ are re-trained.

vi.b) all N_L re-trained networks are tested.

vi.c) The best network is chosen from all the resulting networks.

In case of the TT method, CNNs with just one CL and one OL are trained which takes a time, T , defined by

$$T = A * t \quad (17)$$

where, t is the average time to train each CL. Taking into account that the TT method keeps the topology, its weights need to be re-trained when a new CL is added. If it is assumed that there is a linear relationship between cluster number (l) and the training time, then the total time to execute the TT method for N_L clustered layers is given by

$$T_{TT} = T + 2T + 3T + \dots + N_L T = T \sum_{l=1}^{N_L} l = T \left(\frac{N_L(N_L+1)}{2} \right) \quad (18)$$

On the other hand, the calculation of the time to perform RE ($T_{WTT,RE}$) is the sum of the training time of the latest CL, (T) because the weights of previous layer are kept. The expression for computing the $T_{WTT,RE}$ value is given by

$$T_{WTT,RE} = N_L * T \quad (19)$$

The time to execute the FT ($T_{WTT,FT}$) is given by the addition of the time taken to perform RE and retraining the best networks resulting from the RE.

$$T_{WTT,FT} = T_{WTT,RE} + T_{fine} \quad (20)$$

where T_{fine} is the time to perform the additional steps of the FT. Since the fine tuning's steps consists on retaining only the best networks from one to N_L CL, the T_{fine} calculation is given by

$$T_{fine} = t_{FT} + 2t_{FT} + 3t_{FT} + \dots + N_L t_{FT} = t_{FT} \sum_{l=1}^{N_L} l = t_{FT} \left(\frac{N_L(N_L+1)}{2} \right) \quad (21)$$

Taking into account the $T_{WTT,FT}$ and T_{fine} , $T_{WTT,FT}$ can be simplified to

$$T_{WTT,FT} = N_L * T + t_{FT} \left(\frac{N_L(N_L+1)}{2} \right) \quad (22)$$

From $T_{WTT,FT}$ and T_{TT} , it is possible to presume that the $T_{WTT,FT}$ and $T_{WTT,RE}$ are lower than the T_{TT} . The time gain $G_{TT/FT}$ and $G_{TT/RE}$ are

$$G_{TT/FT} = \frac{T_{TT}}{T_{WTT,FT}} = \frac{T \left(\frac{N_L(N_L+1)}{2} \right)}{N_L * T + t_{FT} \left(\frac{N_L(N_L+1)}{2} \right)} = \frac{S_K * N_K * t \left(\frac{N_L(N_L+1)}{2} \right)}{S_K * N_K * t + t_{FT} \left(\frac{N_L(N_L+1)}{2} \right)} = \frac{S_K * N_K (N_L+1)}{2S_K * N_K + N_L + 1} | t_{FT} = t \quad (23)$$

$$G_{TT/RE} = \frac{T_{TT}}{T_{WTT,RE}} = \frac{T \left(\frac{N_L(N_L+1)}{2} \right)}{N_L * T} = \frac{(N_L+1)}{2} \quad (24)$$

6.3.4. General result of Greedy Algorithms

The algorithm was tested using MATLAB running on a windows 10 operating system with an Intel Corei7-8700K CPU, with 32 GB of RAM and two NVIDIA TITAN Xp GPUs. For training the networks, the Adam learning algorithm with 512 mini batch size and 400 maximum epochs were

used. Each training fold ran in a separate GPU. The other parameters of the training algorithm are 200 validation frequency, 0.1 learning rate drop factor, and 10 epochs learn rate drop period. In each epoch, the data were shuffled. For TT and RE, the initial learning rate is 0.001 and, for FT the initial learning rate is 0.00001.

To find an optimal solution and structure of the CNN for each added CL, the Kernel Size (kSize) is varied from 3 to 9 with a step size of 2 and Kernel Number (kNo) is varied from 5 to 55 with a step size 10. For each CL, the kSize and kNo are chosen from the best CO which is the average of Acc, Sen and Spc. An example using the RE method and 1-minute as input at the first CL, in which the kSize and the kNo are selected as 7 and 35, respectively to get a CO of 77.23%, which is the maximum value (Figure 50). This step is repeated until the maximum CLs size of 10 is reached.

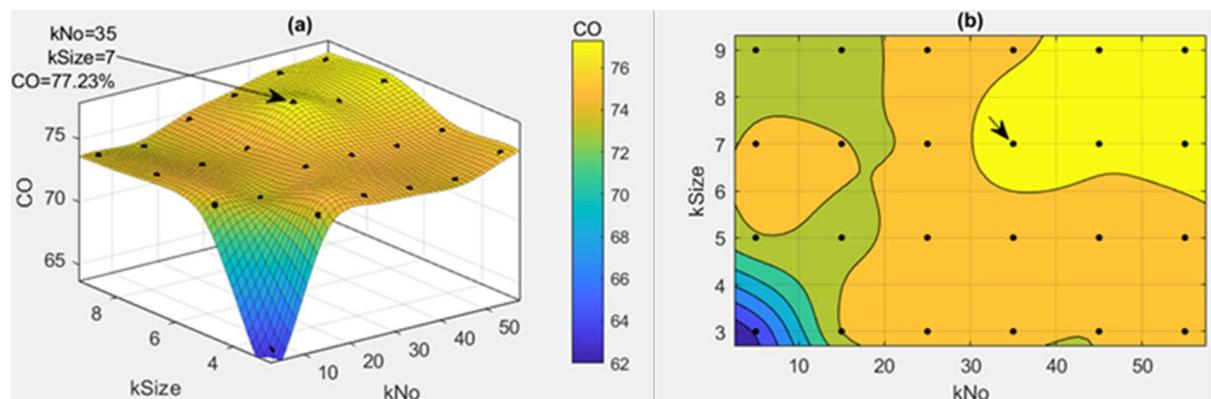


Figure 50 : Best validation (Val) CO solution for the first layer WTT_{RE}.

At first, the TT and the WTT (RE and FT) of the GMO methods were tested for a one minute input. After comparing the performance vs optimization time, only the WTT was tested for the effects of the input size. The details are discussed in section 6.3.5 and section 6.3.6.

The final optimized (Opt) classification using the HuGCDN2008 dataset for greedy methods and the multi-objective method are shown in Table 29. If the absolute performance of the multi-objective and GBO is compared, then the multi-objective method did better than the GMO. Even though multi-objective method was not developed to use CO it performs better on that parameter compared to the GMO. A difference occurs because the GBO starts with finding the structure with the best performance for just one CL. Then a new CL is added, and its best structure is searched. This framework does not produce the best optimal structure with the combination (which can be achieved by an extensive search with the cost of the simulation time). It has high a chance of producing local a optimal solution. However, the difference between the two methods' performance parameters is not that high. The main motivation behind the GBO methods was in reducing the searching time for optimization. If a 1 minute input is considered, compared to the multi objective method, the GBO TT method is 16.68 times faster, the GBO WTT_{RE} is 72.39 times faster, and the WTT_{FT} is 47.6 times

faster. Whereas GBO take half or one third of a day the multi-objective methods take almost a month. If amount of the data are increased, the optimization time will increase also. Therefore, it depends on the user which algorithm is most suitable for his or her task.

Table 29 : Comparison of the results between the multi-objective method and two greedy based optimization (GBO) methods, the WTT and the TT optimized (Opt) classification using the HuGCDN2008 database, Cross Database(CD), and Transfer Learning (TL) .

Type	TEST DATABASE	INPUT SIZE (IN MINUTE)	OPT CL	TIME (IN HOURS)	SEN	SPC	ACC	CO
Muti-objective (Section 6.2)	HuGCDN2008	1	-	587.84	72.55	94.21	88.52	85.09
Muti-objective (Section 6.2)	HuGCDN2008	3	-	832.04	74.05	94.60	89.24	85.96
Muti-objective (Section 6.2)	HuGCDN2008	5	-	911.23	74.75	94.44	89.32	86.17
Muti-objective (CD) (Section 6.2)	AED	1	-	-	91.64	93.36	92.65	92.55
Muti-objective (CD) (Section 6.2)	AED	3	-	-	85.64	93.36	90.20	89.73
Muti-objective (CD) (Section 6.2)	AED	5	-	-	88.58	93.67	91.58	91.28
Muti-objective (TL) (Section 6.2)	AED	1	-	-	92.04	95.78	94.24	94.02
Muti-objective (TL) (Section 6.2)	AED	3	-	-	89.87	96.78	93.93	93.53
Muti-objective (TL) (Section 6.2)	AED	5	-	-	87.76	96.61	92.96	92.44
GBO (TT)	HuGCDN2008	1	9	29.87	69.51	94.90	88.24	84.22
GBO (WTT _{RE})	HuGCDN2008	1	8	8.12	70.09	93.98	87.71	83.93
GBO (WTT _{RE})	HuGCDN2008	3	9	11.78	71.71	93.93	88.13	84.59
GBO (WTT _{RE})	HuGCDN2008	5	9	21.06	73.64	93.80	88.49	85.31
GBO (WTT _{FT})	HuGCDN2008	1	8	12.34	69.68	94.22	87.78	83.89
GBO (WTT _{FT})	HuGCDN2008	3	9	25.18	71.47	94.07	88.17	84.57
GBO (WTT _{FT})	HuGCDN2008	5	10	34.27	73.27	93.57	88.23	85.02
GBO (WTT _{RE,CD})	AED	1	8	-	83.84	89.51	87.18	86.84
GBO (WTT _{RE,CD})	AED	3	9	-	88.94	90.38	89.79	89.70
GBO (WTT _{RE,CD})	AED	5	9	-	75.51	94.95	86.95	85.80
GBO (WTT _{RE,TL})	AED	1	8	-	87.45	96.30	92.65	92.13
GBO (WTT _{RE,TL})	AED	3	9	-	92.36	97.08	95.14	94.86
GBO (WTT _{RE,TL})	AED	5	9	-	87.63	95.17	92.07	91.62

6.3.5. Comparison between different Greedy Algorithms

Considering the HuGCDN2008 dataset, the best performance of the CO for 10 CLs and a 1-minute input is achieved (Table 30 and Table 31): in the ninth CL for the TT method (CO for validation 89.40%; CO for testing 84.22%), in the eighth layer for the RE (CO for validation 88.48%; CO for testing 83.92%), and in the eighth layer for the FT (CO for validation 88.67%; CO for testing 83.89 %).

Table 30 : Chosen CNN’s layers and hyperparameters by WTT (RE and FT).

No.	1 minute (RE, FT)		3 minutes (RE, FT)		5 minutes (RE)	
	Layer	Layer Parameters	Layer	Layer Parameters	Layer	Layer Parameters
L1	Input	1x3000x1	Input	1x9000x1	Input	1x15000x1
L2	Conv_1	35@1x7x1_1x1	Conv_1	15@1x5x1_1x1	Conv_1	25@1x9x1_1x1
L3	BatchN_1	35 channels	BatchN_1	15 channels	BatchN_1	25 channels
L4	ReLU_1		ReLU_1		ReLU_1	
L5	MaxP_1	1x2_1x2	MaxP_1	1x2_1x2	MaxP_1	1x2_1x2
L6	Conv_2	55@1x9x35_1x2	Conv_2	15@1x7x15_1x1	Conv_2	25@1x7x25_1x1
L7	BatchN_2	55 channels	BatchN_2	15 channels	BatchN_2	25 channels
L8	ReLU_2		ReLU_2		ReLU_2	
L9	MaxP_2	1x2_1x2	MaxP_2	1x2_1x2	MaxP_2	1x2_1x2
L10	Conv_3	15@1x7x55_1x1	Conv_3	35@1x9x15_1x1	Conv_3	5@1x5x25_1x1
L11	BatchN_3	15 channels	BatchN_3	35 channels	BatchN_3	5 channels
L12	ReLU_3		ReLU_3		ReLU_3	
L13	MaxP_3	1x2_1x2	MaxP_3	1x2_1x2	MaxP_3	1x2_1x2
L14	Conv_4	5@1x7x15_1x1	Conv_4	25@1x3x35_1x1	Conv_4	5@1x3x5_1x1
L15	BatchN_4	5 channels	BatchN_4	25 channels	BatchN_4	5 channels
L16	ReLU_4		ReLU_4		ReLU_4	
L17	MaxP_4	1x2_1x2	MaxP_4	1x2_1x2	MaxP_4	1x2_1x2
L18	Conv_5	5@1x9x5_1x1	Conv_5	15@1x3x25_1x1	Conv_5	5@1x9x5_1x1
L19	BatchN_5	5 channels	BatchN_5	15 channels	BatchN_5	5 channels
L20	ReLU_5		ReLU_5		ReLU_5	
L21	MaxP_5	1x2_1x2	MaxP_5	1x2_1x2	MaxP_5	1x2_1x2
L22	Conv_6	45@1x9x5_1x1	Conv_6	5@1x5x15_1x1	Conv_6	15@1x9x5_1x1
L23	BatchN_6	45 channels	BatchN_6	5 channels	BatchN_6	15 channels
L24	ReLU_6		ReLU_6		ReLU_6	
L25	MaxP_6	1x2_1x2	MaxP_6	1x2_1x2	MaxP_6	1x2_1x2
L26	Conv_7	5@1x7x45_1x1	Conv_7	25@1x3x5_1x1	Conv_7	35@1x7x15_1x1
L27	BatchN_7	5 channels	BatchN_7	25 channels	BatchN_7	35 channels
L28	ReLU_7		ReLU_7		ReLU_7	
L29	MaxP_7	1x2_1x2	MaxP_7	1x2_1x2	MaxP_7	1x2_1x2
L30	Conv_8	5@1x3x5_1x1	Conv_8	15@1x7x25_1x1	Conv_8	45@1x9x35_1x1
L31	BatchN_8	5 channels	BatchN_8	15 channels	BatchN_8	45 channels
L32	ReLU_8		ReLU_8		ReLU_8	
L33	MaxP_8	1x2_1x2	MaxP_8	1x2_1x2	MaxP_8	1x2_1x2
L34	FC	2	Conv_9	35@1x3x15_1x1	Conv_9	5@1x5x45_1x1

L35	Softmax	BatchN_9	35 channels	BatchN_9	5 channels
L36	Classoutput	ReLU_9		ReLU_9	
L37		MaxP_9	1x2_1x2	MaxP_9	1x2_1x2
L38		FC	2	FC	2
L39		Softmax		Softmax	
L40		Classoutput		Classoutput	

Table 31: Chosen CNN’s layers and hyperparameters by TT 1 minute and 5 minutes WTT, FT.

		1 minute (TT)		5 minutes (WTT, FT)	
No.	Layer	Layer Parameters	Layer	Layer Parameters	
L1	Input	1x3000x1	Input	1x15000x1	
L2	Conv_1	55@1x9x1_1x1	Conv_1	25@1x9x1_1x1	
L3	BatchN_1	55 channels	BatchN_1	25 channels	
L4	ReLU_1		ReLU_1		
L5	MaxP_1	1x2_1x2	MaxP_1	1x2_1x2	
L6	Conv_2	45@1x7x55_1x2	Conv_2	25@1x7x25_1x1	
L7	BatchN_2	45 channels	BatchN_2	25 channels	
L8	ReLU_2		ReLU_2		
L9	MaxP_2	1x2_1x2	MaxP_2	1x2_1x2	
L10	Conv_3	25@1x3x45_1x1	Conv_3	5@1x5x25_1x1	
L11	BatchN_3	25 channels	BatchN_3	5 channels	
L12	ReLU_3		ReLU_3		
L13	MaxP_3	1x2_1x2	MaxP_3	1x2_1x2	
L14	Conv_4	45@1x9x25_1x1	Conv_4	5@1x3x5_1x1	
L15	BatchN_4	45 channels	BatchN_4	5 channels	
L16	ReLU_4		ReLU_4		
L17	MaxP_4	1x2_1x2	MaxP_4	1x2_1x2	
L18	Conv_5	5@1x5x45_1x1	Conv_5	5@1x9x5_1x1	
L19	BatchN_5	5 channels	BatchN_5	5 channels	
L20	ReLU_5		ReLU_5		
L21	MaxP_5	1x2_1x2	MaxP_5	1x2_1x2	
L22	Conv_6	45@1x7x5_1x1	Conv_6	15@1x9x5_1x1	
L23	BatchN_6	45 channels	BatchN_6	15 channels	
L24	ReLU_6		ReLU_6		
L25	MaxP_6	1x2_1x2	MaxP_6	1x2_1x2	
L26	Conv_7	15@1x9x45_1x1	Conv_7	35@1x7x15_1x1	
L27	BatchN_7	15 channels	BatchN_7	35 channels	

L28	ReLU_7		ReLU_7	
L29	MaxP_7	1x2_1x2	MaxP_7	1x2_1x2
L30	Conv_8	35@1x3x15_1x1	Conv_8	45@1x9x35_1x1
L31	BatchN_8	35 channels	BatchN_8	45 channels
L32	ReLU_8		ReLU_8	
L33	MaxP_8	1x2_1x2	MaxP_8	1x2_1x2
L34	Conv_9	45@1x5x35_1x1	Conv_9	5@1x5x45_1x1
L35	BatchN_9	45 channels	BatchN_9	5 channels
L36	ReLU_9		ReLU_9	
L37	MaxP_9	1x2_1x2	MaxP_9	1x2_1x2
L38	FC	2	Conv_10	5@1x7x5_1x1
L39	Softmax		BatchN_10	5 channels
L40	Classoutput		ReLU_10	
L41			MaxP_10	1x2_1x2
L42			FC	2
L43			Softmax	
L44			Classoutput	

The CO difference between these three methods is less than 1% (0.35). As we can see in Table 29 and Figure 51, the same is true for Acc, Sen, and Spc. The time to finish the different solutions is also presented in Figure 51 and Table 29 .

To optimize the CNN, 29.87 hours were needed if the TT method is used, 8.12 hours were needed if the WTT_{RE} method is used, and 12.34 hours were needed if the WTT_{FT} method is used (Table 29). The time needed for executing the TT method is triple the time needed for executing the WTT_{FT} method (~ 3.68 times). The time for executing the WTT_{FT} method is 1.52 times more than the time for executing the WTT_{RE} method. However, in both cases, the performance gain is low compared with the execution time.

There is some difference between the practical gain and the gain deduced theoretically. These differences occur due to i) the assumed that the time of each added CL is incremented linearly, while this behavior is not linear as can be seen in Figure 51 ; and ii) the previous theoretical calculation only considered the time of training. The practical time calculation is the sum of the time needed to create networks, save networks, load the data, extract features, save the relevant results.

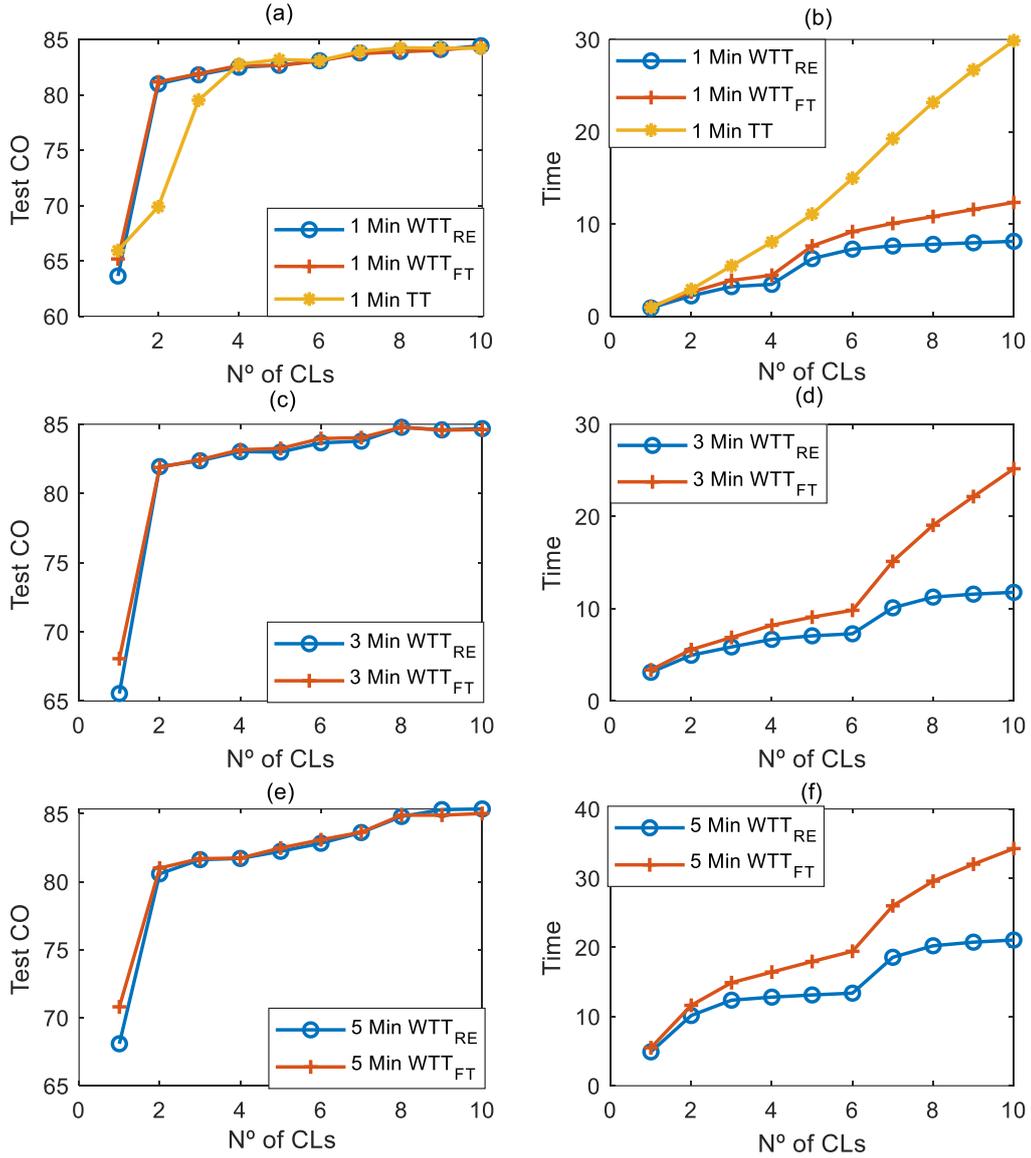


Figure 51 : Best (according to validation) test CO solution for each layer ((a), (c), (e)) and total simulation time (in hours((b), (d), (f))) for TT, RE (WTT) and FT(WTT).

6.3.6. Effect of input size

To understand the effect of different input sizes, three inputs are tested: a 1-minute, a 3-minute, and a 5-minute input. These input sizes are used because of the previous success in the literature [87]. Even though the TT method takes more time than the WTT, there is no significant difference in performance between them (Table 29). Thus, only the WTT_{RE} and the WTT_{FT} methods are tested further.

From Table 29 , the test CO for the best solution using the WTT_{FT} and the WTT_{RE} are: 83.89% and 83.93% for 1-minute, 84.57% and 84.59% for 3-minute and 85.02% and 85.31% for 5-minute inputs, respectively, in case of the HuGCDN2008 database. The validation CO for the best solution using the WTT_{FT} , and the WTT_{RE} are: 88.67% and 88.48% for 1-minute, 87.99% and 88.09% for 3-minute and

87.66% and 87.49% for 5-minute inputs, respectively. The performance difference between the test and validation is due to the training and test data.

Furthermore, for the HuGCDN2008 database Table 29 shows that the optimal CNN of 5-minute input size is achieved with 10 CLs for the WTT_{FT} and 9 CLs in the case of the WTT_{RE} methods. In the case of 1-minute and 3-minute input sizes, the corresponding number of CLs are 8 and 9, respectively. The difference between the CO, Acc, Sen, and Spc is less than 1% for the WTT_{FT} and the WTT_{RE} methods.

6.3.7. Responsiveness of the $kSize$ and kNo

For each CL, the $kSize$ and kNo parameters are changed. The responsiveness of the parameters is analyzed for these parameters. The responsiveness of a parameter is defined by the objective function responses due to the changes in the parameter. Consider the variations of the objective CO by changing the $kSize$ defined by

$$C_{kSize}^{CO} = \frac{dCO}{dkSize} \quad (25)$$

To understand the effect of one parameter on another parameter (kNo) is kept constant. However, for each added CL there are several solutions. That means, for a positive or negative variation of the CO a similar variation in C_{kSize}^{CO} arises. The total amount of responsiveness for $kSize$ (MAC_{kSize}^{CO}) is given by

$$MAC_{kSize}^{CO} = \frac{1}{N_L} \sum_{l=1}^{N_L} \frac{1}{M_{kSize}} \sum_{m=1}^{M_{kSize}} abs(C_{kSize}^{CO}) \quad (26)$$

where kNo is a constant, and M_{kSize} is the number of C_{kSize}^{CO} values for each added CL. Similarly, the total amount of responsiveness for kNo (MAC_{kNo}^{CO}) is given by (where $kSize$ is constant):

$$MAC_{kNo}^{CO} = \frac{1}{N_L} \sum_{l=1}^{N_L} \frac{1}{M_{kNo}} \sum_{m=1}^{M_{kNo}} abs(C_{kNo}^{CO}) \quad (27)$$

A similar approach to MAC_{kSize}^{CO} and MAC_{kNo}^{CO} is implemented and plotted in Figure 52 for Acc, Sen and Spc. From Figure 52, it can be seen that every input size and every objective are more sensitive to $kSize$ than to kNo . These results allow one to justify the reason that a short step and range is used for $kSize$ (range from 3 to 9 and step 2) and a big step and range is used for kNo (range from 5 to 55 and step 10).

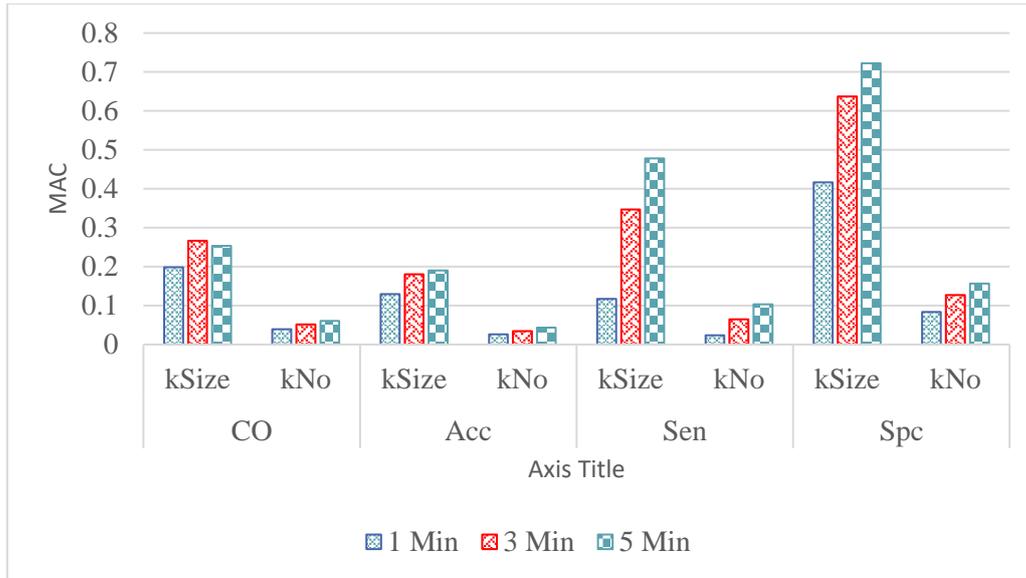


Figure 52 : Total Mean Absolute Changes (MAC) of different objectives (in validation) with respect to unit parameter (kSize and kNo) change for 1-minute (1 Min), 3-minute (3 Min) and 5-minute (5 Min) inputs.

6.3.8. Effect of the added CL

After optimizing each CL with the best kSize and kNo, a new CL is added to obtain the best number of CLs. Figure 53 shows the effect of adding CLs on different performance parameters in the case of the WTT_{RE} . The performance parameters variations are computed by subtracting the current layer performance from the previous one. The addition of the second CL has the highest CO increase: 8.57% for 1-minute, 9.58% for 3-minute and for 6.22% 5-minute. However, adding more layers does not always increase performance. For example, it can be seen that for all inputs sizes the performance is reduced by adding the tenth CL.

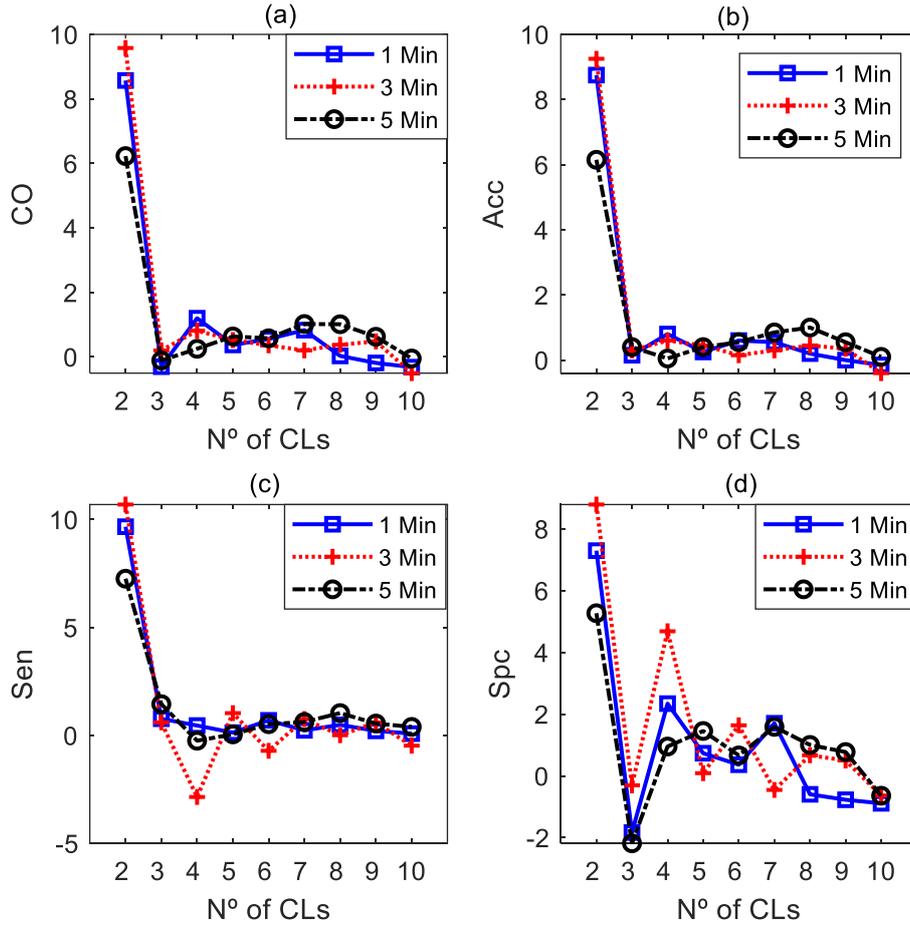


Figure 53 : Best (according to validation) training and validation (Val) CO for each layer WTT_{RE} method.

6.3.9. Comparison with literature

Among the tested greedy based methods (TT , WTT_{RE} , and WTT_{FT}) the attained performance is similar. However, the execution time between them is different. The WTT_{RE} method is the best if the execution time vs performance is considered. Therefore, the CNN structures found by the WTT_{RE} method are tested on a different dataset named AED [61] because it is an available and commonly used dataset in the literature. The HuGCDN2008 dataset is used for determining the best CNN structure.

A two-fold validation technique is used to determine the best CNN structure resulting in two classifiers with the same structure but different trainable parameters. In this work the first-fold classifiers are chosen for comparison.

The classifiers are tested in two different methods using the AED dataset. The first method, called the Cross Database (CD), consists of feeding the new data from the AED dataset into the optimized. WTT_{RE} classifiers (Table 29). The second method, called Transfer Learning (TL), consists of deleting

the last layers of the classifiers and re-training the classifier with the new AED dataset. This last method is performed using the leave-one-out method with 50% for training and validation data and with the same training parameters as the WTT_{RE} .

Different techniques are proposed in the literature to detect apnea events. Xie et al. [73] use Bagging RepTree to detect an apnea minute using SpO₂ and a combination of SpO₂ with an ECG for 1-minute input size. An Acc of 82.79% with SpO₂ and 84.80% with SpO₂+ECG is reached using the UCD database [62]. Other researchers using only SpO₂ signal from the UCD database achieved an Acc of 81.95% with a ANN [195], 83.27% with an LD[195] and 85.26% with a DAE [106]. Some of these solutions have low sensitivity rates, such as 43.31% [195].

Another study using SVM classifiers and 100 subjects from the SHHS database reaches an Acc of 80.1% and a Sen of 60.9% with SpO₂ and an Acc of 82.4% and a Sen of 69.9% with respiratory signals+ECG+SpO₂ [169]. Several classifiers using the AED database can be found in the literature, such as: an ANN achieving an Acc of 90.3% [99], a DAE achieving an Acc of 97.64% [106] and linear SVM(SVM-L) achieving an Acc of 96.89% and 97.38% [195]. Using the CNN and the SpO₂ signal for children, Vaquerizo-Villar et al. achieved 93.6% Acc, 56.5% Sen, and 96.7% Spc [246]. An accuracy of 90.8% was achieved by ECG and CNN [179].

Three databases from the University Hospital of Gran Canaria Dr. Negrin, are available in the literature: HuGCDN2014 (77 subjects) [137], HuGCDN2004 (66 subjects) [197], and HuGCDN2008 (70 subjects) [10]. LDA classifiers and an HRV signal (or RR series) achieve an Acc of 81.96% (HuGCDN2014), 81.18% (HuGCDN2004) [197] and 79.4% (HuGCDN2008), respectively. Using the same classifier with SpO₂ signal instead, the Acc reaches 76.88% (HuGCDN2004) [197], and 86.5% (HuGCDN2008) [10]. The combination of the SpO₂ and the HRV gives an Acc of 82.68% (HuGCDN2004) [197] and 86.9% (HuGCDN2008) [10].

In this work, CNN1D classifiers optimized with the WTT_{RE} method are proposed as a novelty. Classifiers optimized through this method using the HuGCDN2008 database achieve an Acc of 87.71% for a 1-minute input (1Min), 88.13% for 3-minute (3Min) and 88.49% for 5-minute (5Min) input (Figure 51, and Table 29). These results show an improvement in performance over the work carried out by Ravelo-García et al. [10]. Testing the same CNN1D classifiers with the TL using the AED dataset, an Acc of 92.65% for 1-Min, 95.14% for 3-Min and 92.07% for 5-Min are obtained. Using the same dataset in the literature, 95.5% is obtained with an LSTM[107], 97.64% is obtained with a DAE [106] and 97.38% is obtained with a linear SVM [195]. Comparing this work's with the literature, the highest difference is 2.5%. This difference is justified by the inexistence of a subject independent test in the DAE [106] and the LSTM [107] classifiers and the low Sen in the SVM-L [195].

The reasons behind the low sensitivity might be related to the definition of OSA. It is known that OSA events are not always accompanied by desaturation. The main definition is associated with a certain decrease in airflow regardless of the presence or absence of hemoglobin desaturation [247]. In addition to that, the HuGCDN2008 dataset has patients with other diseases that can cause a decrease in the SpO₂ during sleep without apnea occurring such as asthma, chronic obstructive pulmonary disease, fibrosis among others, which can contribute to a false positive.

6.3.10. Summary

The main novelty of this work is to detect sleep apnea using a CNN1D optimized through the greedy based optimization method. The CNN1D developed and optimized with the greedy based optimization method presents better performance than the similar works found in the literature. Within the greedy based optimization algorithm, three different variants are presented: TT, WTT_{RE} and WTT_{FT} . Considering the balance between the result of the execution time and the performance gain of the three variants, the WTT_{RE} method is the best.

The greedy based optimization method consists of searching for the best number of CL and its parameters. In the beginning, the added CL has the tendency of improving the performance significantly. At a certain point, this insertion has a marginal increase. On the other hand, each CL is optimized through the kNo and kSize parameters. During this optimization, it has been concluded that kSize presents more responsiveness than kNo.

Different input sizes are tested to understand the effects. The best performance is obtained using 5Min input size for the HuGCDN2008 dataset with the WTT_{RE} method. For the AED dataset with the WTT_{RE} method, the best performance is obtained using a 3 Min input size. In general, there is a relationship between the input size and performance. To verify this relationship, more research with several databases needs to be done in the future.

As future work, an improvement of the proposed approach could be obtained by adding more signals.

6.4. Combination of SpO₂ and HRV with CNN

6.4.1. Introduction

In some of the literature discussed before (in Section 4.2.6 and Section 5.5.1) a combination of two or more signals can improve the performance. Therefore, the main purpose of the work that resulted

in this section is to reevaluate the HR or the HRV signal and the SpO₂+HR signals. The effect of fully connected layers and a drop out layer on the CNN are also investigated.

6.4.2. Combination of SpO₂ and HR

It is possible to have an HR with oxygen saturation using the SpO₂ sensor without any extra or at most a little cost. Both signals are tested. However, the HuGCDN2008 database does not have the HR from the SpO₂ sensor. In the literature, it is possible to find a high correlation between the PPG-HRV, and the ECG-HRV [248]. To check the performance, the RR interval calculated from the ECG was used. Thereafter, the RR interval is converted to calculate HRV. Nevertheless, it should be noted that this introduces some uncertainty/error in the signals since the HR is not even collected in the same place of the body.

The HuGCDN2008 database has 70 subjects, which are divided into 35 training and 35 test subjects. A two-fold cross validation method is used for validation. For each fold from the training set a validation set is created for training and for a stopping criterion for the training. The data are prepared with a 1 minute (1 Min), 3 minute (3 Min), and 5 minute (5 Min) window with one minute sliding window. Because the HuGCDN2008 database is annotated in 30 seconds, if one or both of the 30 seconds in the minute is annotated as apnea that minute is annotated as apnea or else it is annotated as normal. The middle minute's annotation considers the annotation of the 3-Min and 5-Min windows.

6.4.3. Transfer learning of classifiers and combination of signals

The CNN1D classifiers developed in the previous section (Section 6.3) are used for testing the performance of the signals. Additionally, to understand the effect of the fully connected layers and the drop out layer, the last layers (the FC, the softmax, and the classout) are replaced by two Dropout Layers (DO), two fully connected layers (FC), a softmax, and a classout. The parameters of these two layers are found out using an extensive search (grid search). The DO is varied from 0.1 to 0.8 with an increment of 0.1. The number of neurons of the FC layer is varied from 50 to 200 with an increment of 50.

6.4.4. Performance of SpO₂, HRV, and SpO₂+HRV

A two-fold method was employed in the HuGCDN2008 database for the SpO₂, the HRV, and the SpO₂+HRV evaluation, considering 1 Min, 3 Min and 5 Min inputs. The developed networks were

trained and tested varying the DO and the FC. Since there are 32 combinations for each classifier, the selection was based on the maximum validation CO of the two-fold.

When the SpO2 signal was used as an input, the maximum validation CO (85.16%) was achieved at 0.6 DO and 150 FC for 1 Min; 0.2 DO and 150 FC achieved a maximum 86.34% for 3 Min and 0.8 DO and 200 FC achieved a maximum 85.56% at for 5 minute. From Figure 55, it is visible that the maximum CO position in the validation and the test is different. This occurred due to the different data and subjects in the validation set and the testing set. The test Acc, Sen, and Spc of the extensive search for best DO and FC are showed in Figure 56 (for 1 Min), Figure 57 (for 3 Min), and Figure 58 (for 5 Min) and the final results are showed in Table 32.

For the HRV, the maximum validation CO for a 1 Min input is 75.34% and it was achieved with 0.5 DO and 100 FC; for a 3 Min input this value is 76.5% with 0.3 Do and 150 FC; and for a 5 Min input it is 79.86% with 0.2 DO and 150 FC (Figure 59). Unlike the SpO2 signal, the HRV has an incremental improvement with respect to the input size. However, because of the same reason as the SpO2 signal, the HRV signal also has a different maximization point in the validation and the test space (Figure 60). The test Acc, Sen and Spc of the extensive search for the best DO and FC are illustrated in Figure 61 (for 1 Min), Figure 62 (for 3 Min), and Figure 63 (for 5 Min) and the final results are shown in Table 32.

By combining the two signals (SpO2+HRV), the maximum validation CO for 1 Min is 86.20% with 0.3 DO and 200 FC; for 3 Min it is 84.66% with 0.8 DO and 150 FC; and for 5 Min it is 85.93% with 0.3 DO and 150 FC (Figure 64, Figure 65). The same property is shown in the validation (Figure 64, Figure 65) and the test (Figure 65) CO for SpO2+HRV as with the SpO2 and the HRV. The test Acc, Sen, and Spc of the extensive search for the best DO and FC are shown in Figure 66 (for 1 Min), Figure 67 (for 3 Min), and Figure 68 (for 5 Min) and the final results are revealed in Table 32.

From Table 32 it is clear that for an epoch based classification the highest accuracy of 85.78% was achieved by the 1 Min SpO2 and the 3 Min SpO2+HR. The highest sensitivity was also achieved by the 1 Min SpO2. However, the best sensitivity was achieved by the 5 Min SpO2+HR. If the number of signals and the size of input are considered and results considered as well then the 1 Min SpO2 performed quite well as it reaches the highest accuracy and sensitivity with just one signal.

Adding the DO and the FC layer on top of the classifier developed in the previous section (Section 6.3) increases the sensitivity of the classifiers. However, the multi-objective objective methods' (Section 6.2) overall performance was better even after adding the extra layer to the classifier optimized using the greedy methods (Section 6.3).

Table 32 : Test performance of the HuGCDN2008 database for the SpO2, the HRV and the SpO2+HRV for 1 Min, 3 Min and 5 Min inputs.

Type	TEST DATABASE	INPUT SIZE			SEN	SPC	ACC	CO	GACC	R ²
		(IN	DO	FC						
		MINUTE)								
SpO2	HuGCDN2008	1	0.6	150	77.87	88.53	85.78	84.06	92.86	0.9247
SpO2	HuGCDN2008	3	0.2	150	82.06	86.69	85.47	84.74	97.14	0.91966
SpO2	HuGCDN2008	5	0.8	200	82.93	86.79	85.76	85.16	97.14	0.91052
HRV	HuGCDN2008	1	0.5	100	63.99	72.63	70.41	69.01	74.29	0.3958
HRV	HuGCDN2008	3	0.3	150	63.89	79.80	75.69	73.13	75.71	0.66898
HRV	HuGCDN2008	5	0.2	150	64.82	80.01	75.80	73.60	80	0.54921
SpO2+HRV	HuGCDN2008	1	0.3	200	79.02	88.01	85.63	84.23	91.43	0.9335
SpO2+HRV	HuGCDN2008	3	0.8	150	83.79	86.42	85.78	85.34	97.14	0.91138
SpO2+HRV	HuGCDN2008	5	0.3	150	84.86	85.68	85.60	85.37	95.71	0.92235

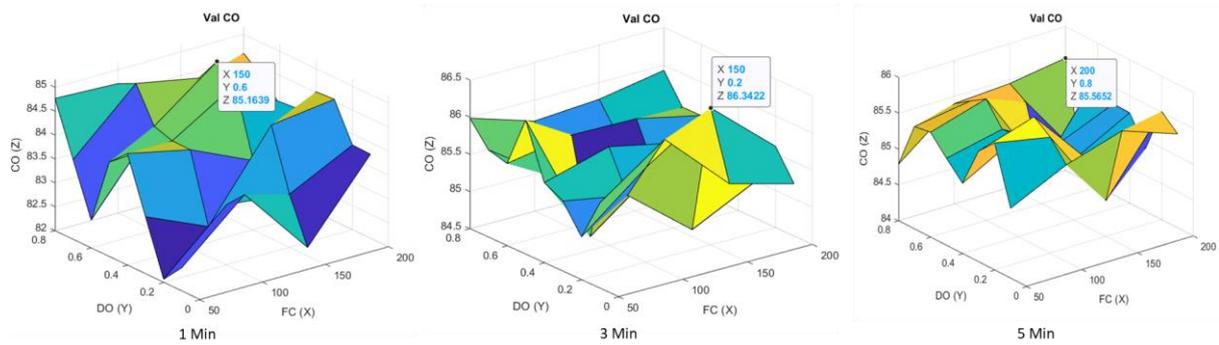


Figure 54 : The validation CO for the SpO2 signal using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.

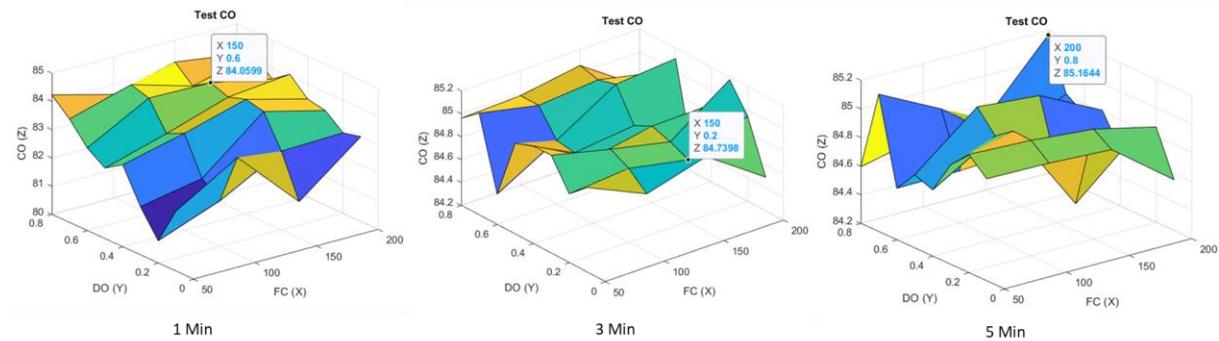


Figure 55 : The test CO for the SpO2 signal using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.

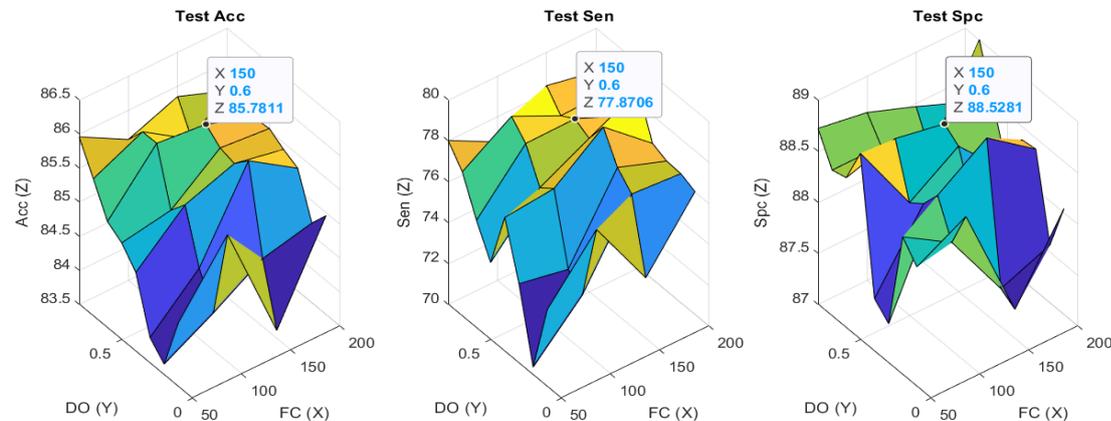


Figure 56 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 1 Min input of the SpO2.

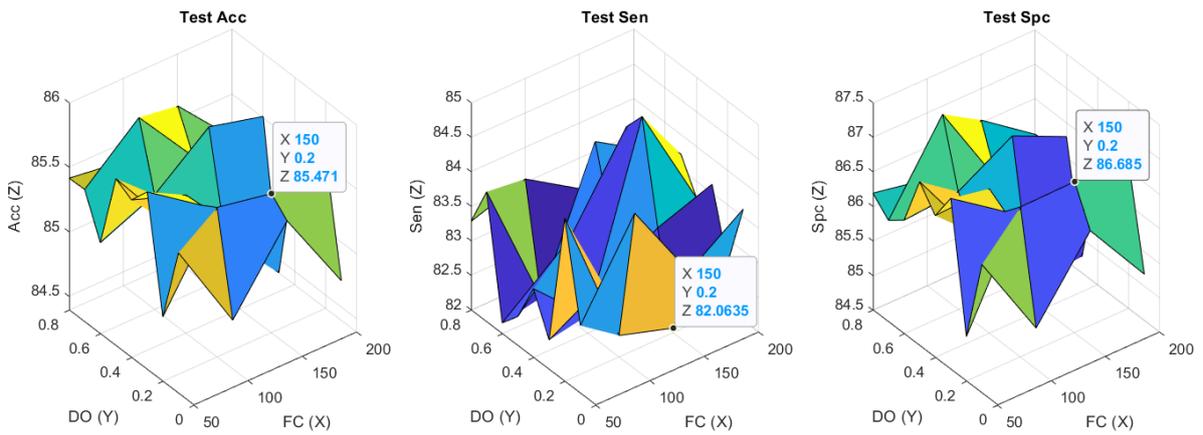


Figure 57 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 3 Min input of the SpO₂.

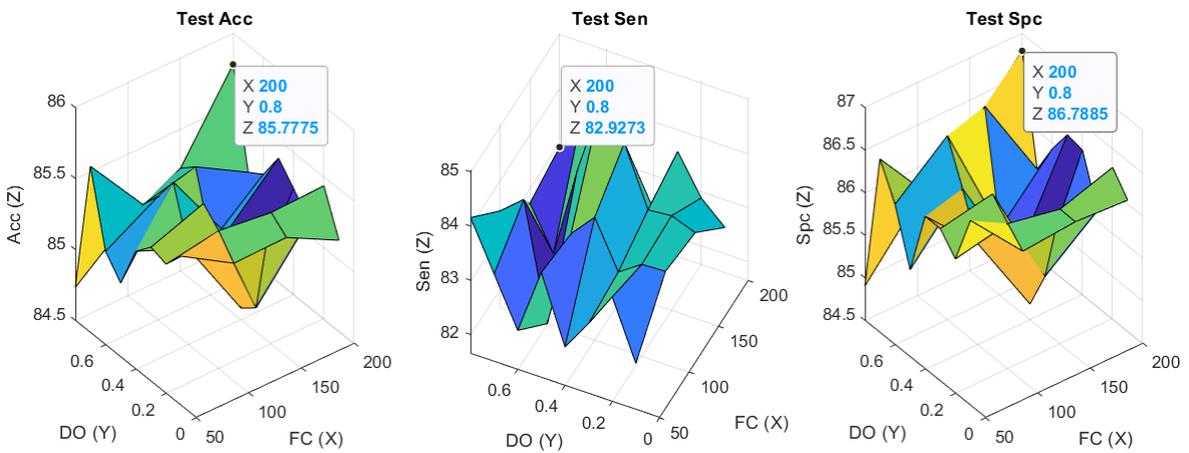


Figure 58 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 5 Min input of the SpO₂.

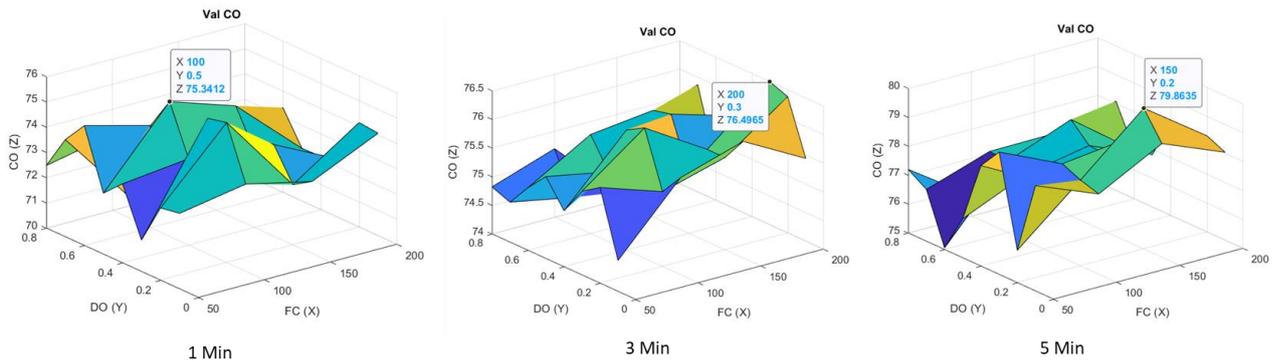


Figure 59 : The validation CO for the HRV signal using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.

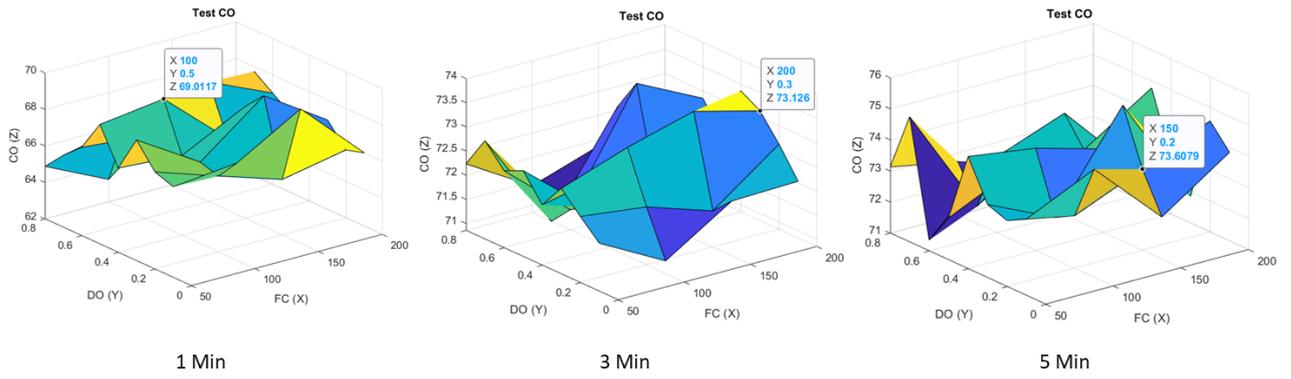


Figure 60 : The test CO for the HRV signal using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.

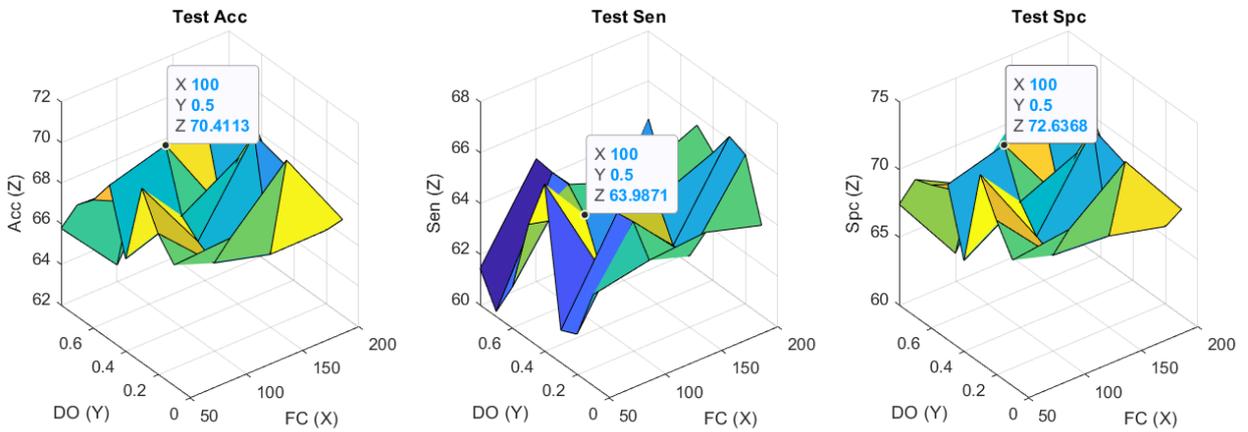


Figure 61 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 1 Min input of the HRV.

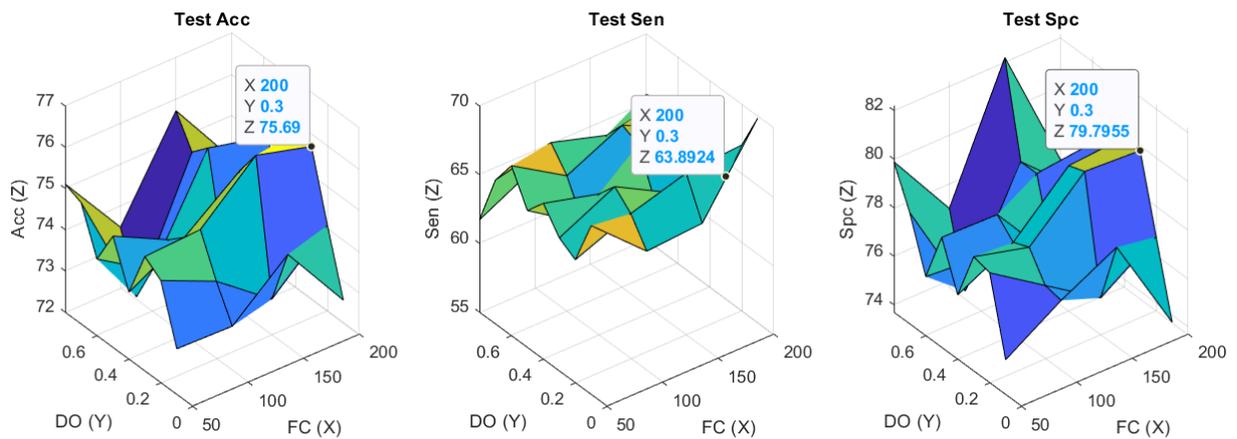


Figure 62 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 3 Min input of the HRV.

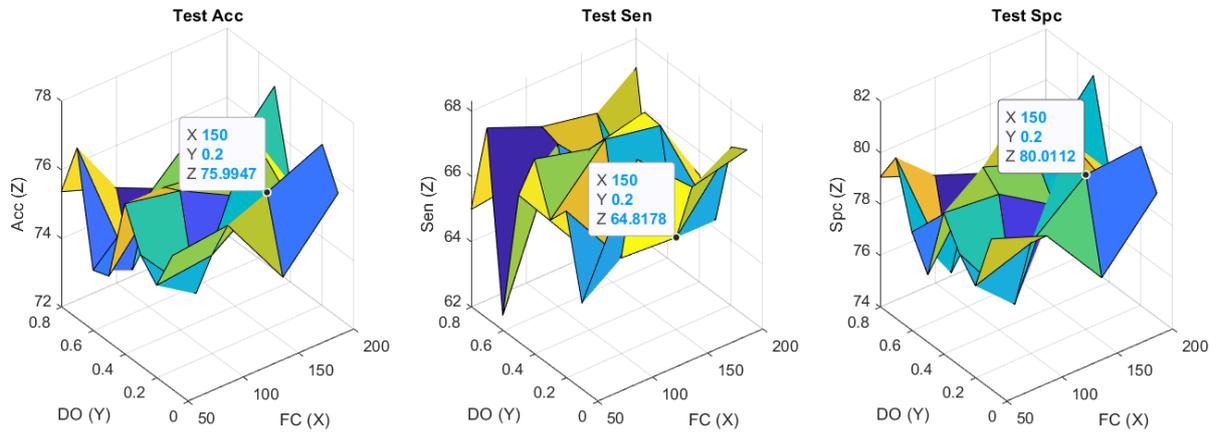


Figure 63 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for 5 Min input of the HRV.

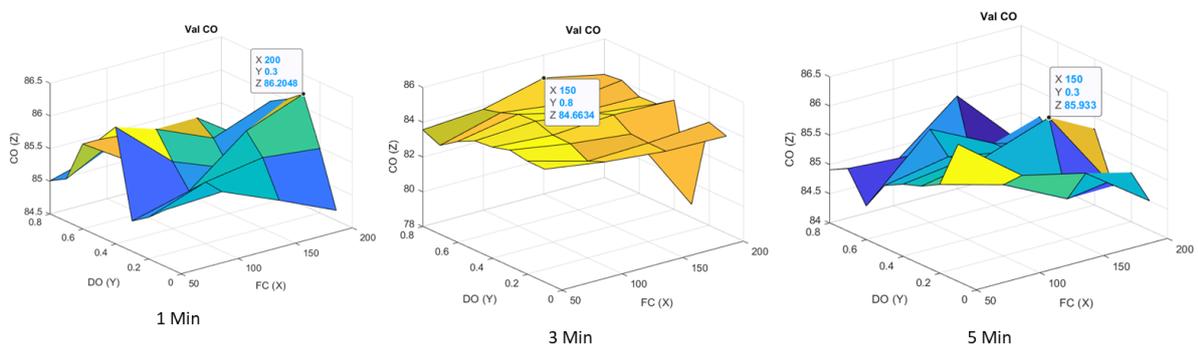


Figure 64 : The validation CO for the combined signals (SpO2+ HRV) using the 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.

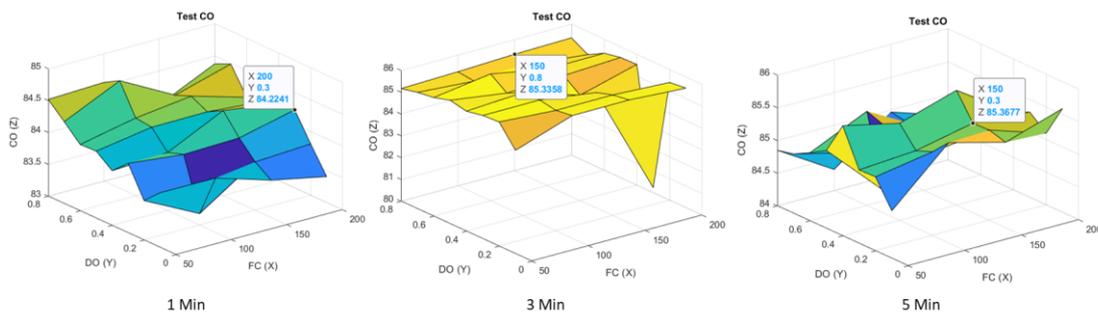


Figure 65 : The test CO for the combined signals (SpO2+HRV) using 1 Min, 3 Min, and 5 Min inputs with respect to the DO and the FC variation.

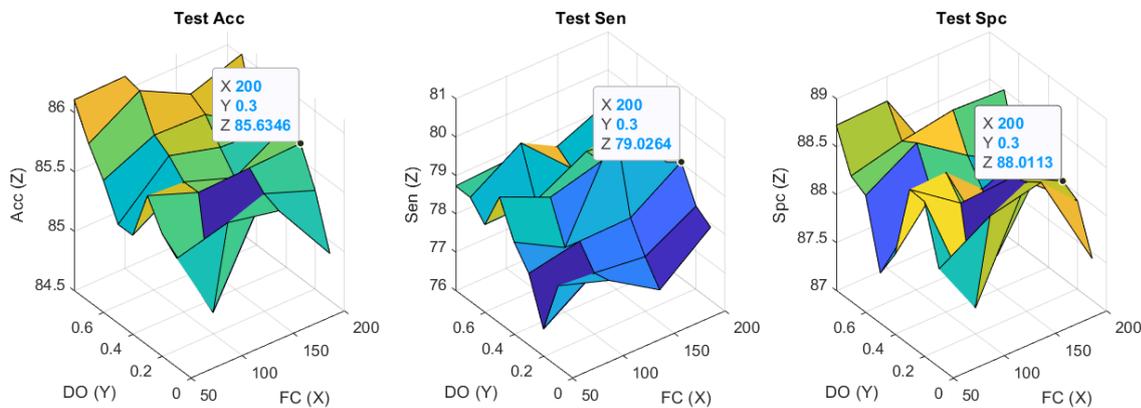


Figure 66 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for a 1 Min input of the combined signals (SpO2+HRV).

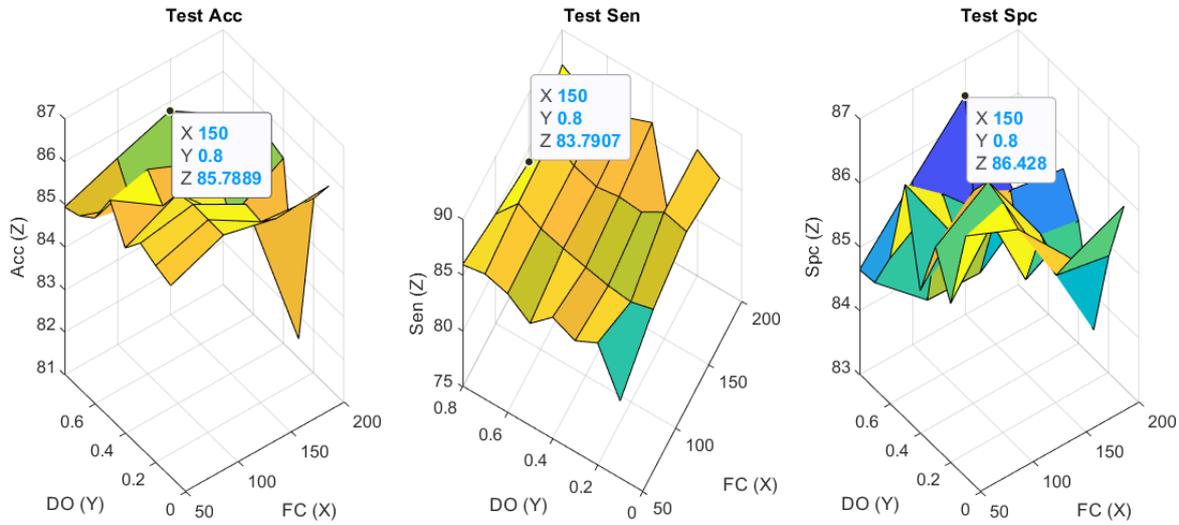


Figure 67 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for a 3 Min input of the combined signals (SpO2+HRV).

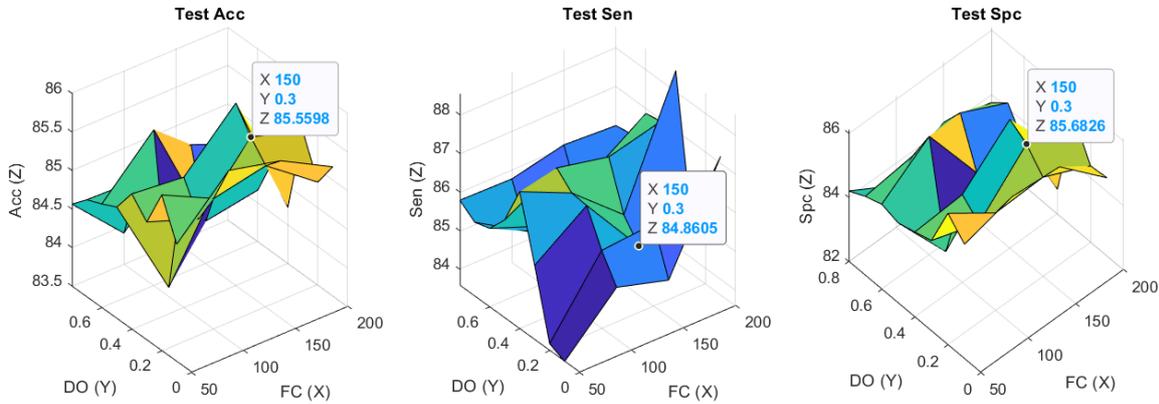


Figure 68 : The test Acc, Sen, and Spc of the extensive search of the DO and the FC for a 5 Min input of the combined signals (SpO2+ HRV).

6.4.5. Global classification

A similar work was previously carried out, with a similar dataset (HuGCDN2008), by Ravelo-García et al. [10] with 35 test subjects without the EBLs. In their work, the model achieved a GAcc of 91.4% using the SpO2 signal; 71.4% using the HR signal and 94.71% using the SpO2+HR signals [10]. This work presents the comparison between the AHI calculated by the medical professional (AHI G MP) of the database and the AHI TiB calculated from the output of the classifiers (AHI C TiB) (Figure 69, Figure 70, and Figure 71). Without any removal of subjects (EBLS), the current standard of the AHI index for detecting apnea patients is 5 (AHI TiB = 5) and is used in this work. In

case of the HuGCDN2008 all the seventy subjects were used. The SpO2 signal achieved 92.86% (with $R^2(R2) = 0.9247$), the HR method achieves 74.29% (with $R^2(R2) = 0.3958$) and the SpO2+HR achieves 91.43% (with $R^2(R2) = 0.9335$) for 1 Min input. The SpO2 signal achieves 97.14% (with $R^2(R2) = 0.9197$), the HR method achieves 75.71% (with $R^2(R2) = 0.6690$) and the SpO2+HR achieves 97.14% (with $R^2(R2) = 0.9114$) for 3 Min input using HuGCDN2008 database. The SpO2 signal achieves 97.14% (with $R^2(R2) = 0.9105$), the HR method achieves 80% (with $R^2(R2) = 0.5492$) and the SpO2+HR achieves 95.71% (with $R^2(R2) = 0.9224$) for 5 Min input using HuGCDN2008 database.

From the results it is visible that not only the epochs based classification but also the global accuracy of the HR signal-based system benefited from using longer input size. For the SpO2, a 3 Min input achieved the best result. When it comes to the combination of the SpO2 and the HR, it presents a domination of the SpO2 signals performance. Like the epochs-based classification the accuracy is mainly dependent on the SpO2 signal.

If a comparison between this work's features-based systems (Section 5.5) and automated feature based systems is compared, in both cases the highest global accuracy of 97.14% is achieved by only the SpO2 signals. In some cases, adding the HR signal with the SpO2 can reduce the accuracy.

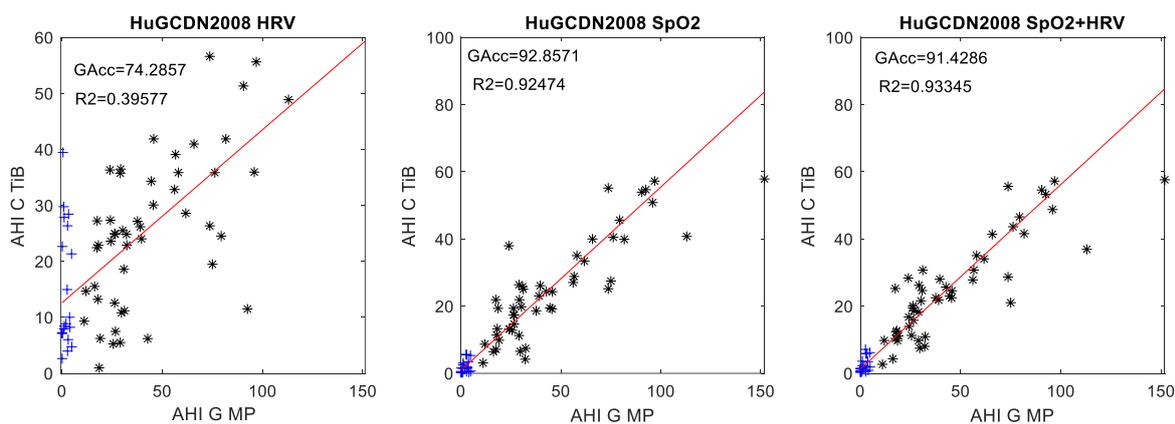


Figure 69 : Comparison of the global accuracy of the HuGCDN2008 dataset with the AHIC calculated by a medical physician (AHIG MP) and by the CNN classifiers' AHIC time in bed (AHIC TiB) for the SpO2, the HRV and the SpO2+HRV of 1 Min input. (+) symbol is used for normal subjects (AHIC ≤ 5) and (*) is used for apnea patients.

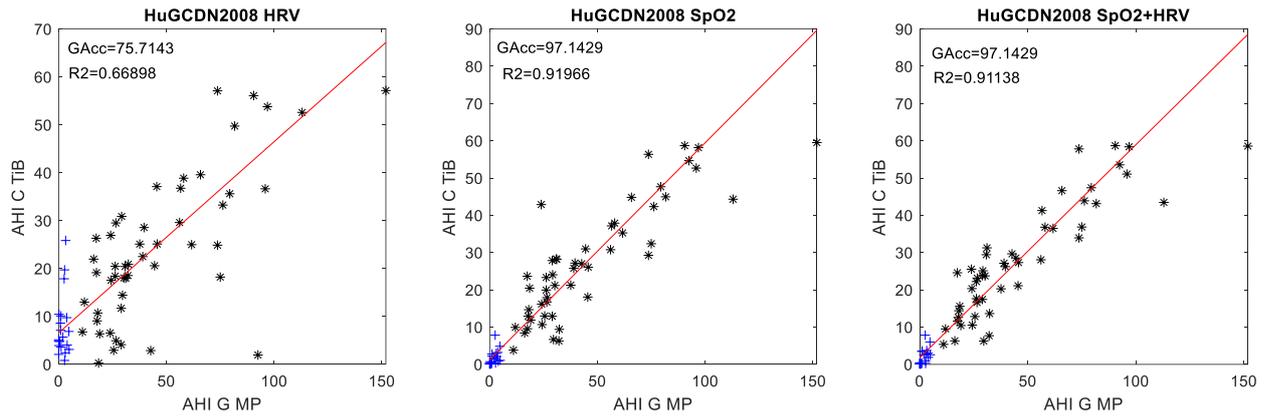


Figure 70 : Comparison of the global accuracy of the HuGCDN2008 dataset with the AHI calculated by a medical physician (AHI G MP) and by the CNN classifiers' AHI time in bed (AHI C TiB) for the SpO₂, the HRV and the SpO₂+HRV of 3 Min input. (+) symbol is used for normal subjects (AHI≤5) and (*) is used for apnea patients.

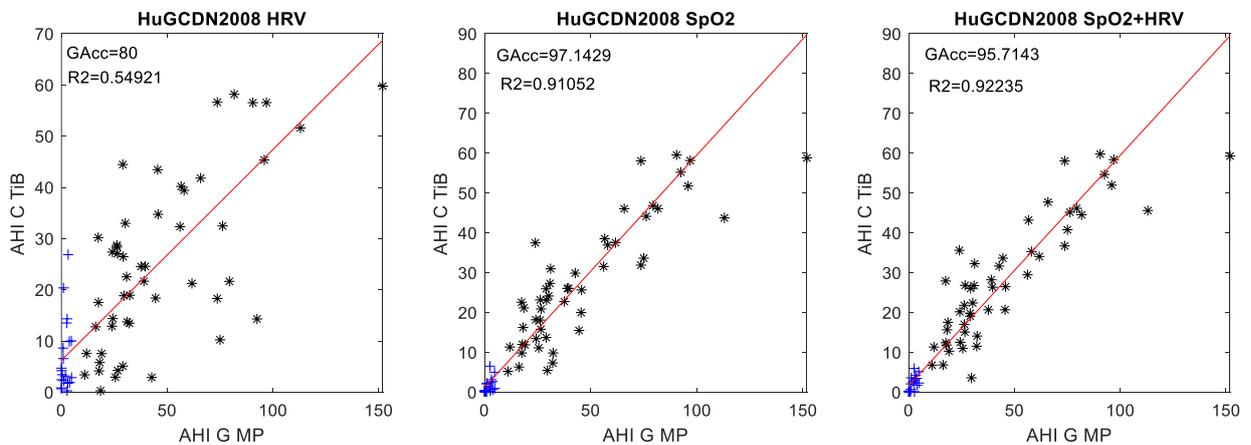


Figure 71 : Comparison of the global accuracy of the HuGCDN2008 dataset with the AHI calculated by a medical physician (AHI G MP) and by the CNN classifiers' AHI time in bed (AHI C TiB) for the SpO₂, the HRV and the SpO₂+HRV of 5 Min input. (+) symbol is used for normal subjects (AHI≤5) and (*) is used for apnea patients.

6.4.6. Summary

The goal of this section was to understand the effectiveness of applying transfer learning to the classifiers developed in the previous section. It was concluded that transfer learning is possible, and it was verified that it is also possible to use different signals to classify using transfer learning. For the HuGCDN2008 dataset the balanced results lay on the roam of the other research. In the case of global accuracy, this work performs better than the literature. However, adding an extra layer of the FC layer does not significantly improve the result. The system was able to classify the apnea subjects with high accuracy.

6.5. Summary of Automated-Feature based Methods

In this chapter, two structural hyperparameter optimizers were developed. One was based on multi-objective optimization and the second one was based on greedy methods. The multi-objective method achieved the best accuracy among the proposed methods. Compared to shallow networks, the developed CNNs were able to achieve better performance with a smaller input size and without the need for hand crafted feature extraction.

The goal of this work was to develop and test a novel fully automated hyperparameters optimization algorithm for the CNN and significant results were attained. Three different window sizes were also tested, and it was verified that there is almost no difference between the 3 minute and the 5 minute window sizes. In some cases, the 1 minute input outperform the 3 minute and 5 minute inputs.

The multi objective method and greedy optimization both have high a cross database and transfer learning capabilities. However, the multi objective method performed better than the greedy methods. The greedy methods were able to achieve almost similar performance with less optimization time.

As with handcrafted feature based methods (Section 5.5), the HR and the HR+SpO2 did not perform better than the SpO2. Similar trends for global classification also occur between the handcrafted feature based methods (Section 5.5.5) and the automatic feature based methods (Section 6.4.5).

Chapter 7

Implementation

This chapter discusses a generic system designed for the physical implementation of the developed algorithms. There are two parts of the developed system: the signal acquisition system (hardware part) and the classification, the interpretation and the display (software part).

7.1. Introduction

A large amount of customer based home devices were developed using dedicated and, frequently, expensive hardware such as a computer, which reduces the accessibility to the general population. Smart phone Applications (Apps) can solve this accessibility issue since they are widely available and used by the population. The smart phone penetrations in 2018 in top 5 countries were 82.2% in the United Kingdom, 79.3% in the Netherlands, 78.8% in the Sweden, 78.8% in the Germany and 77.0% in the United States of America [249]. Therefore, an app can serve as a solution to the apnea detection problem. The mobile apps are used for prescribing, managing, coding and billing, the diagnosis and the treatment of the patients. In recent years, the Food and Drug Administration (FDA) of the United States approved apps from different companies such as Pear Therapeutics, Apple. In Germany's new Digitalisation and Innovation Act (Digital Supply Act), apps can be prescribed by the doctor. With this trend in mind more countries will open the door for the use of apps in medical fields.

An astonishing number of 325,000 mobile health apps were available in 2017 [250]. One research work in 2016 found 51 unique sleep apps in both iOS and Google Play stores [251]. Another review work found 6 apps that capture data from the phone's sensors [252]. A common sensor among all of the apps is the accelerometer besides that the microphone, and the light sensor are quite commonly used for medical apps as well.

On the other hand, some researchers used the Bluetooth data collection system with a computer [253][254], while some used the phone oximeter[105]. Others chose the Bluetooth data collection system with a phone-based processing system. In this category, Dipti et al. used a Bluetooth pulse oximeter sensor to collect the SpO₂ signal and used the K-means and the random forests algorithm for the classification in the phone [9]. Nuria et al. also used an off-the-shelf pulse oximeter with the same configuration [255]. For pillow control, a Bluetooth based pulse oximeter with an SVM was used by Zhang et al. [2].

Mobile phone-based apnea detection gained momentum because of the high performance of mobile phones and their availability. A Bluetooth data collection system gives more flexibility and comfort. Because of this, in this work a Bluetooth data collection system with a mobile app is developed and tested.

7.2. Implementation

The implementation of the proposed system was divided into two main components, the wearable device and the smartphone application.

The system collects the SpO2 and the heartrate values in the main loop of a microcontroller system (Arduino system) and forwards the data to whichever Bluetooth device it is connected to. This process runs at 50Hz. A simple messaging protocol was developed to identify each message with a message ID. This assures that the connected devices can identify out of order or duplicate messages. The overall architecture of the proposed system is depicted in Figure 72. In the scope of this chapter, this work will refer to all these components as the watch, since it reassembles a common wristwatch. As the HR from a PPG and an ECG is almost similar for practical implementation, a PPG is chosen for easy assembly and it was possible to calculate the HR with the same sensor designed for the SpO2. However, the PPG HR could not be used in case of any specific problems where fingertip PPG recording would be impractical (peripheral hyperkinesia or tremor). As a matter of fact, in that kind of situation, this work’s finger-based apnea detection methods will also fail. A question might arise as to why this work used the HR from the ECG. This is because the ECG HR is the gold standard and this work wishes to compare its results with other similar works.

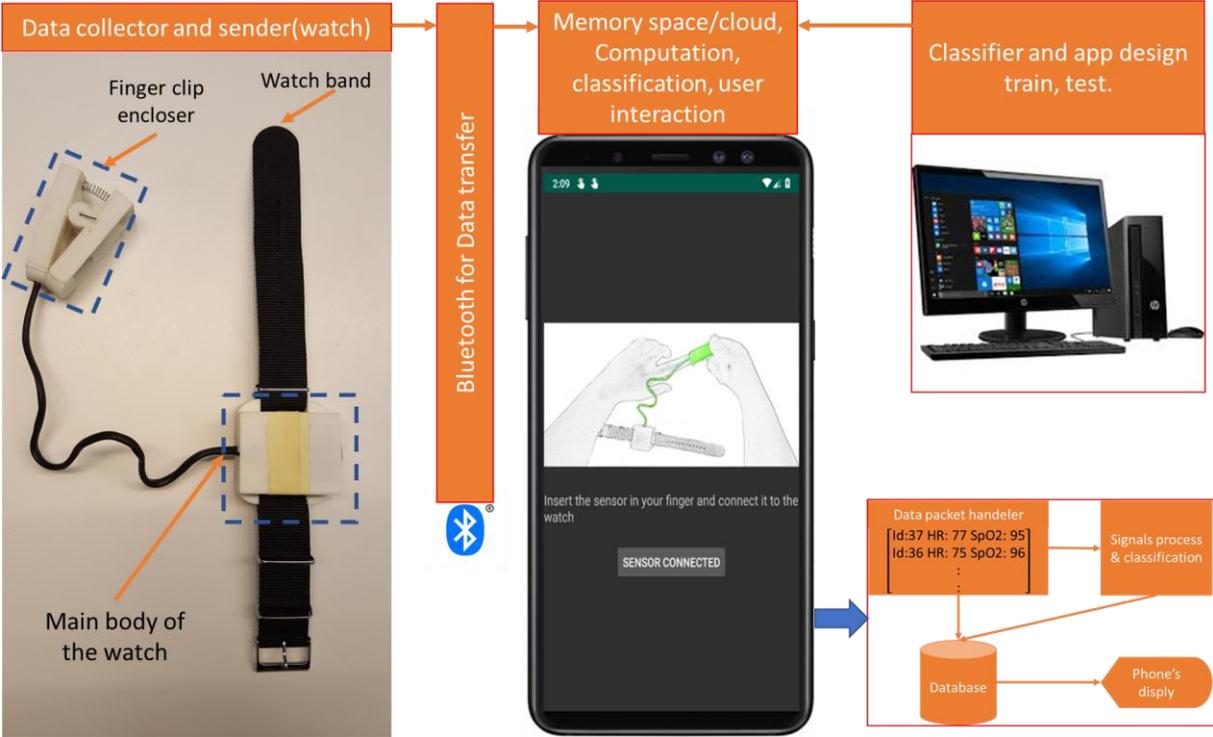


Figure 72 : Overall architecture of the proposed system.

7.2.1. Wearable device

A custom watch similar to a wearable device is developed with a SpO2 sensor (MAX 30102 which includes internal LEDs, photodetectors, optical elements, and low-noise electronics with ambient light rejection, an integrated pulse oximetry and a heart-rate monitor module) and a microprocessor (Bluno Beetle BLE Bluetooth Arduino board) (Figure 73). There are 4 main parts of the watch design, these are marked as 1 (the encloser for the battery), 2 (the encloser for the

microprocessor which covers the Bluetooth Arduino board with a custom-made PLC (polylactide) case), 3 (the top part of the finger clip), and 4 (the bottom part of the finger clip enclosure for the sensor) (Figure 73 and Figure 74). In addition to these, there is the watch band to secure the battery and microprocessor enclosure, wired to connect the sensor and microprocessor. The internal configuration is illustrated in Figure 74. A switch and an LED light allow the user to turn the watch ON/OFF and to assess its status (Figure 73). The internal configuration is illustrated in Figure 74. The current consumption of the wearable device is 37.20 mA with data transmission. A 400 mAh battery is used.

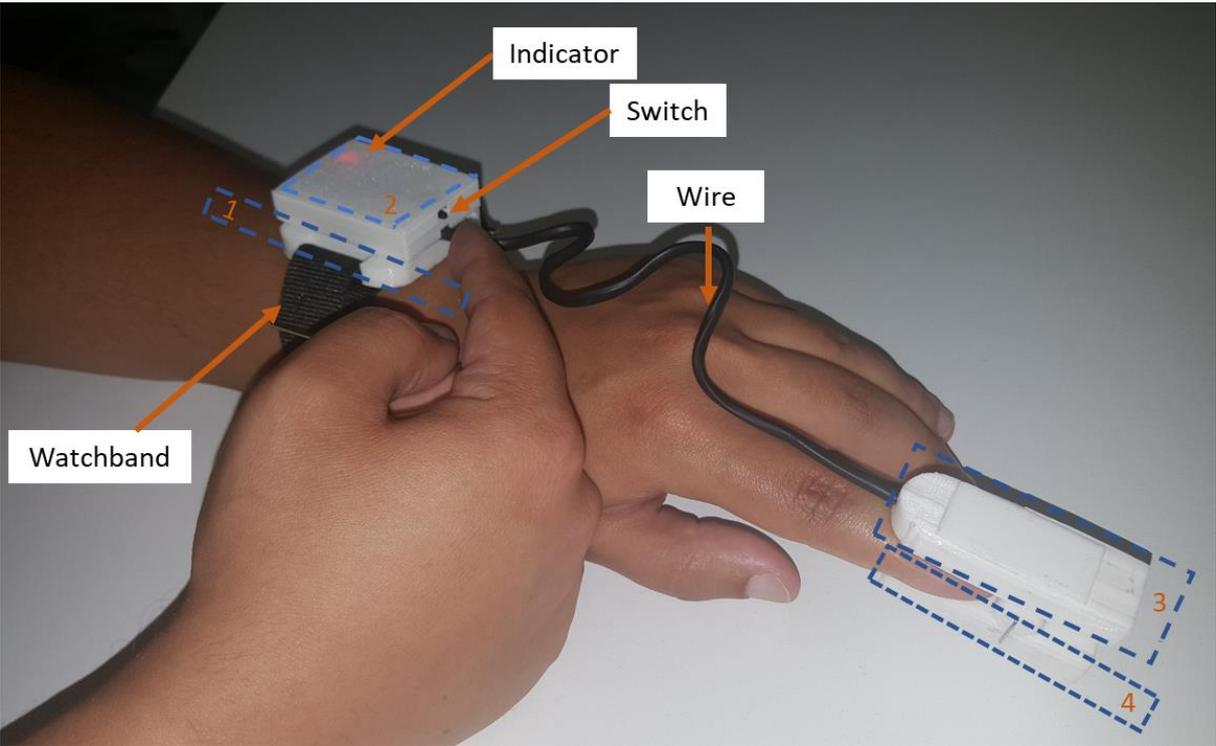


Figure 73 : User with the watch around their wrist, and the SpO2 sensor placed on their index finger.

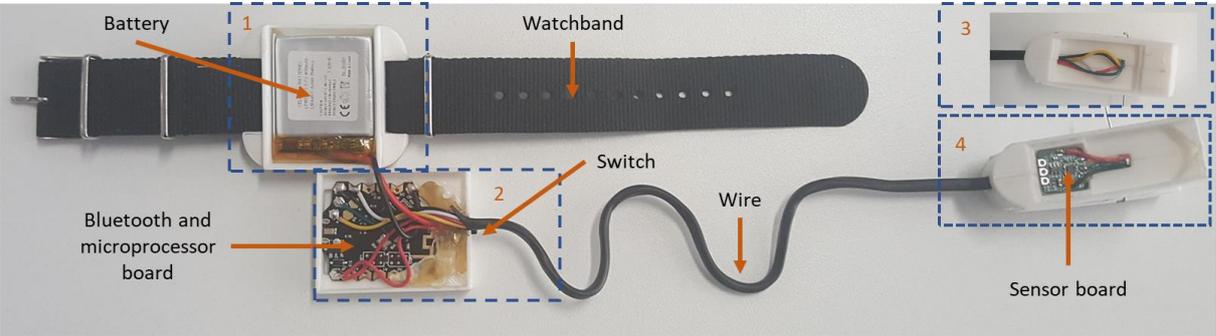


Figure 74 : User with the watch around their wrist, and the SpO2 sensor placed on their index finger.

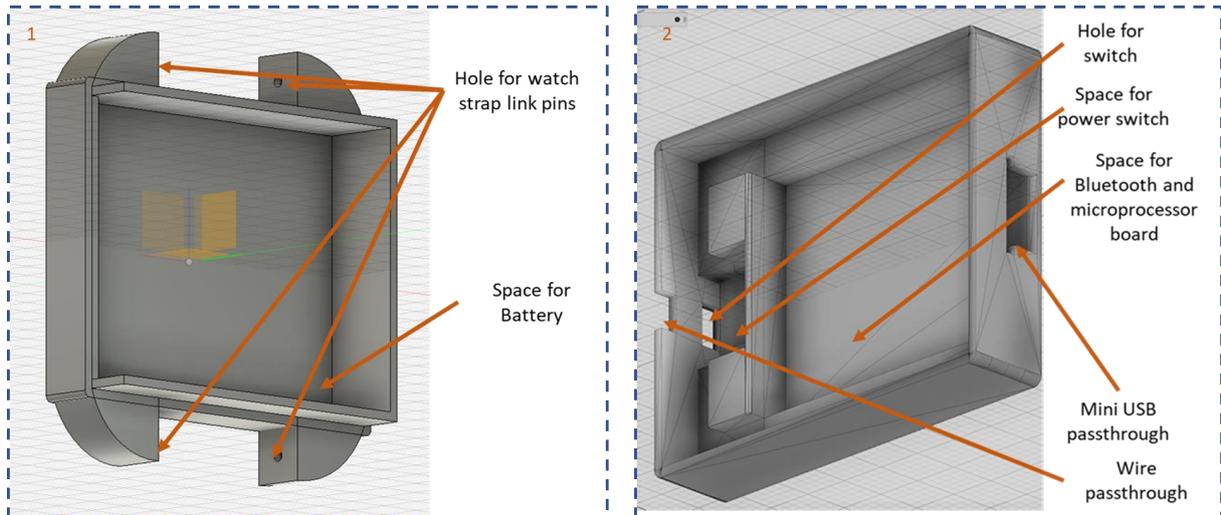


Figure 75 : Encloser for the battery and microprocessor board (main body of the watch).

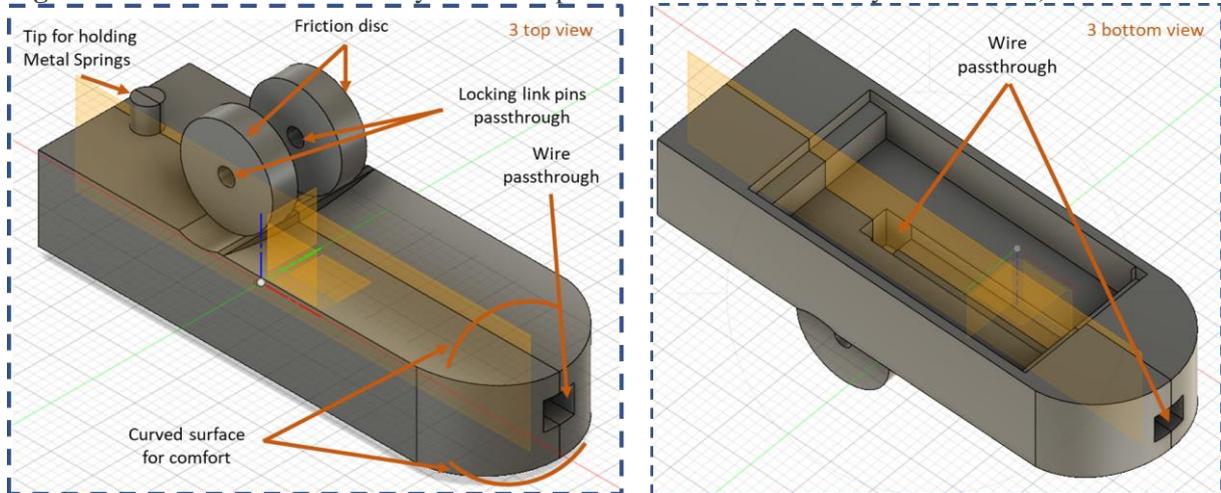


Figure 76 : Top part of the finger clip.

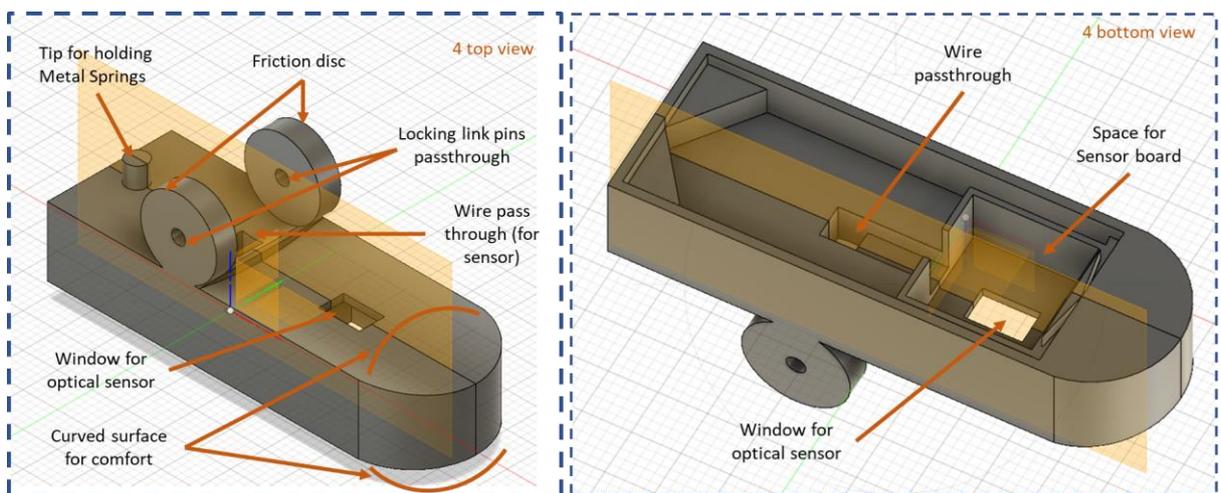


Figure 77 : Bottom part of the finger clip encloser for the sensor.

The encloser for the battery (Figure 75, marked as 1) is 3D printed. The encloser holds a lithium ion battery which is connected to the microcontroller board using a wire (Figure 74). This encloser has 4 holes for the watch straps' link pins which are used for connecting the watchstrap to the wearable device. The encloser for the microprocessor board (Figure 75, marked as 2) is also 3D printed. It has

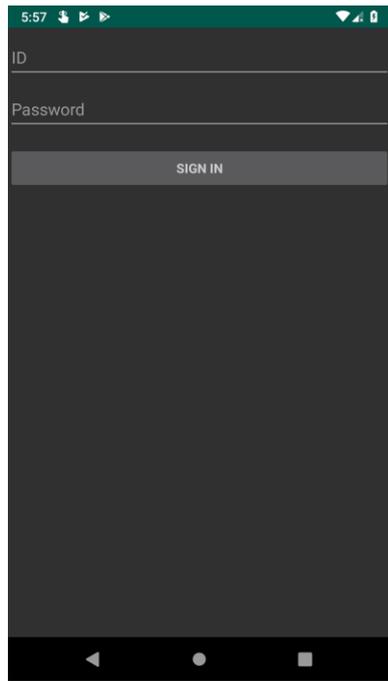
three holes for the switch, the wire passthrough and the mini USB.

The bottom part of the finger clip (Figure 77) houses the sensor board. The enclosure has a window to measure the signal from the fingertip. It has another hole for the wire passthrough to transfer the sensor data to the microprocessor. Both the top (Figure 76) and the bottom part (Figure 77) of the finger clip are connected using a locking pin and it clamps the finger using a spring. The jaw of the finger clip is made of interlocking friction discs secured with a locking pin. The spring is locked using the one holding cylinder in each part of the finger clip. The surface of the clip touching the finger is curved for maximum comfort.

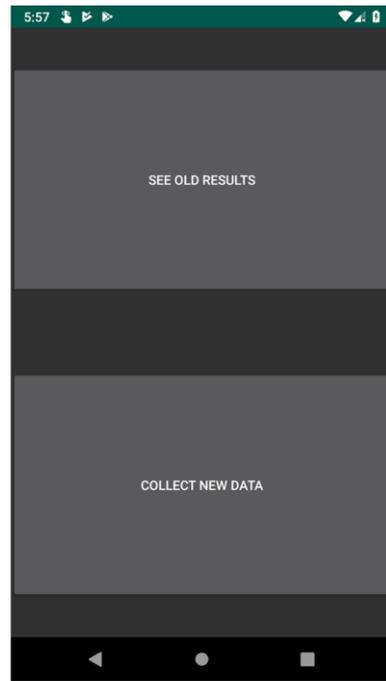
7.2.2. Smartphone application

The smartphone application (app) is the main point of the interaction with the user. It was developed using the android SDK as well as the support libraries for tasks such as displaying charts using the data collected from the sensor and implemented algorithms.

The main goal when building the mobile application was to make sure that the instructions were clear. Therefore, it was decided to follow a wizard model with simple instructions along with a picture in each step. The user is welcomed with a sign in page (Figure 78 a) where a user ID and password can be setup for secure log in. Then the choice between seeing the old recorded data and the new data collection is presented to the user (Figure 78 b). If the user wishes to see the old data, a choice among the saved data are presented (Figure 79 a) and later the results are shown according to the user's choice (Figure 79 b). However, if the user chooses to collect new data (Figure 78 b) he or she is instructed through a proper device wearing process (Figure 80). Afterward, the user is brought to a place where the user is asked to allow the searching of the Bluetooth device and connecting of the sensor (Figure 81 a). Then the user is asked to fill in the information for the STOP-Bang Questionnaire [256] before going to sleep (Figure 81 b and Table 33). Afterward, the user goes to sleep, and the application keeps running in the background. Even if the user closes the application, a process will keep receiving data from the watch. When the user is awake, he or she can open the application either by clicking on the notification or by the clicking on the application button. The user needs to state that he or she woke up and the application will present a summary of their sleep, with charts and textual information regarding the collected metrics, with a focus on OSA occurrences (Figure 79 b).

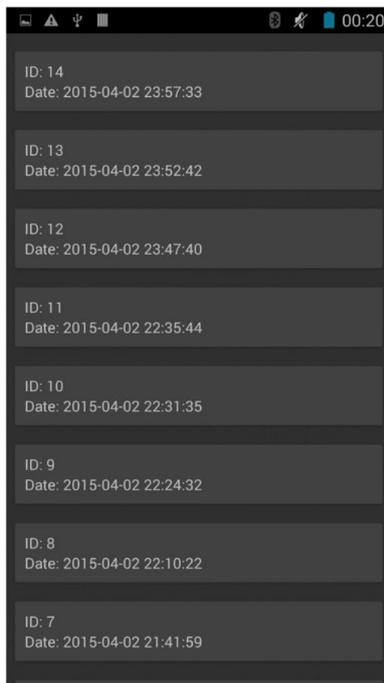


a) Sign in page



b) Choice between collecting and viewing data

Figure 78 : a) Sign in page, and b) choice between collecting and viewing data.

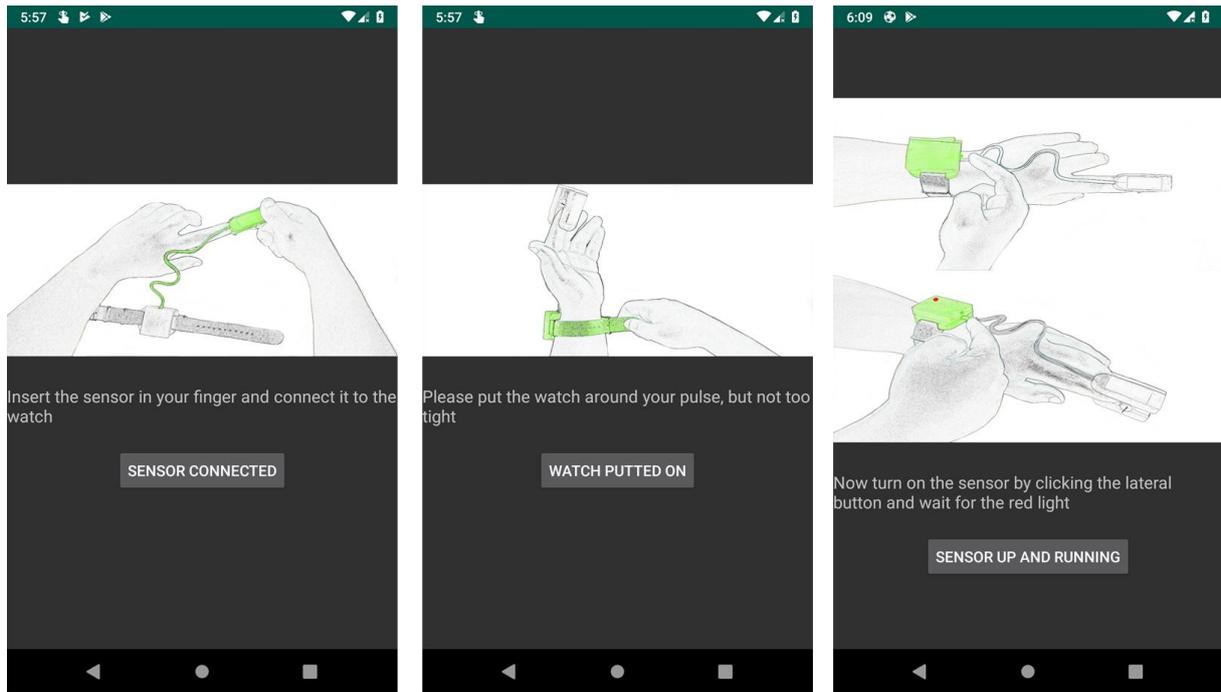


a) Recorded data



b) Results of recorded data

Figure 79 : a) Recorded data and b) Results of recorded data.

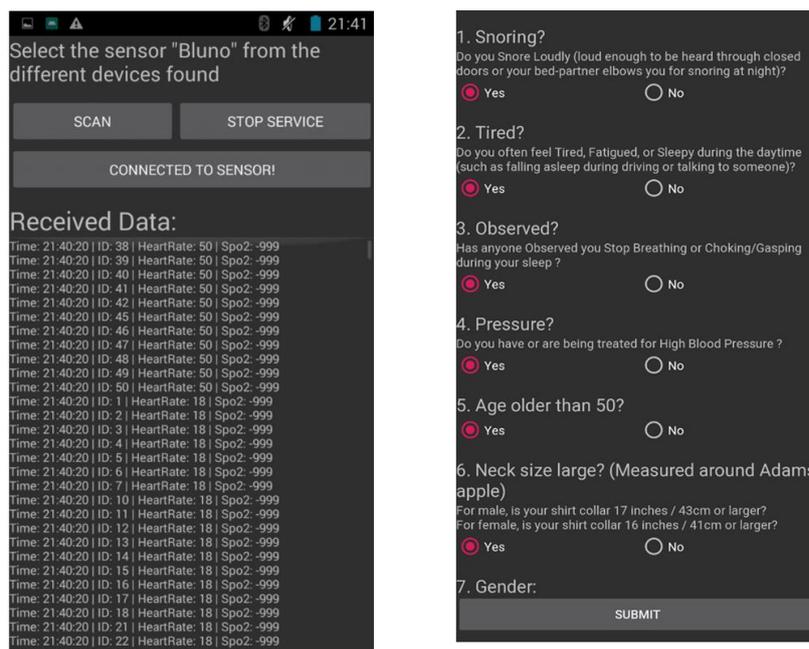


a) Connect sensor

b) Put on watch

c) Turn on the sensor

Figure 80 : a) Connection of the sensor b) Putting on watch c) Turning on the sensor.



a) User scan for device connected to the sensor

b) STOP-Bang Questionnaire

Figure 81 : a) User scan for device connected to the sensor and b) STOP-Bang Questionnaire.

7.3. Summary

The watch and the mobile application are made for a general-purpose use so that any algorithm developed can be integrated to detect the apnea event and produce the results. The app is also capable

of collecting and storing the data. Additionally, a sleep questionnaire is added for more versatility. All of the results are shown in a summarized way so that it is easily interpretable for the patients. In the future the implemented designed could be tested against the collected data with the PPG method.

Chapter 8

Conclusion and Future Work

This chapter finalizes this work, summarizing the conclusions and pointing out aspects to be developed in future work.

8.1. Conclusion

The main purposes of this work, which is devoted to detecting sleep apnea problems, were threefold: to find and analyze suitable literature; to implement and develop different algorithms to solve the existing problem; and to improve the current literature by advancing the state of art.

The review of different literature indicated that the highest accuracy in detecting sleep apnea was attained using an ECG. However, the majority of ECG algorithms were tested in public databases with potentially cleaner signals, which could contribute to improving the diagnostic capability of the algorithm. This corresponds to a trend that signals more susceptible to noise are less discriminatory towards the events producing a lower accuracy even if they are clearly connected to the apnea events such as respiration signals. Another relevant topic was the use of more sensors and signals. Though in some specific cases adding more sensors or signals provides a better accuracy, overall, the combination of the source sensors did not always contribute to a relevant improvement of the classification capability, indicating that one signal dominated the others. The review of the deep networks also observed the domination of the ECG based methods. Though in most of the cases deep networks performed better than normal shallow classifiers, this it is not always valid. A gap of hyperparameters optimization was also found in the deep network implementation.

The methods developed in this work attained a performance that was higher than most of the existing state of the art methods for the classification of apnea events. From the results it became apparent that a subset of related features would perform better than a large number of features. For the SpO2 signal, most of the important features are in the frequency or the time frequency domain. Additionally, wrapper based methods as well as a combination of classifiers [73] [201] were better for sleep apnea detection. It was found that in the literature review there was a lack of the deep classifier optimization. Thus, when it comes to automated feature-based methods, two approaches were developed. The first one was a multi-objective method that performed better than the greedy based method. However, due to time constraints, the greedy methods were implemented for testing other parts of the work. Though the classifiers were developed initially for the SpO2 signal, they worked well for the HR and also performed well on other databases (different from the one used for developing the model) showing that they have transfer learning capabilities.

From the results, it was verified (from Section 5.3) that the subject dependent classification strategy could improve the performance. However, due to the impracticality of the subject dependent classification in real world problems, subject independent tests should be done and were carried out through this work, except for Section 5.3.

In the case of combining the SpO2 and the HR, both automated and handcrafted solutions

performed worse than the SpO2 signal alone. This was verified by the initial idea that some of the high accuracy from the ECG and the SpO2+HR might be data dependent.

An AHI index of 5 was considered for the global classification methods. This approach created a disadvantage for the developed work when compared with previously published works that considered an AHI index of 10 or more. Because of the thresholding technique, the developed methods had to be more precise. Additionally, some works removed the borderline patients from their global classification, which most likely increased the accuracy of the system. This is due to the fact that when a thresholding technique is used, most misclassifications occur for the cases that are near the threshold. The global accuracy test was performed without removing any subjects, which created another layer of difficulties for the classification technique. Some works present in the literature changed the AHI index according to their data thus creating a data dependent classification method. To avoid such an occurrence, an AHI index of 5 was used. However, even with all the difficulties placed on the classification, the proposed methods performed better than in similar literature.

For the implementation strategy, a mobile app with the SpO2 and the PPG HR was developed. The implementation was generic, so it is possible to implement any classifier developed in this work to run on this implementation.

8.2. Future Work

This work can be expanded in different directions such as

- The focus in this work was on the classification accuracy. The features selected for a better classification could be used in the future to understand the undermining physiological reasons.
- Signals other than SpO2 and the HR could be tested.
- The HR for the PPG could be checked with an alternative for the ECG HR.
- Different types of classifiers and classifier combinations such as LSTM, combination CNN LSTM could be tested to verify if a better performance can be achieved.
- In relation to the implementation, alternative platforms can be tested besides phones such as microcomputers for cheaper and easier designs and FPGAs for more dedicated hardware designs.

Annex A

Table 33 : STOP-BANG sleep apnea questionnaire

STOP		
Do you SNORE loudly (louder than talking or loud enough to be heard through closed doors)?	Yes	No
Do you often feel TIRED , fatigued, or sleepy during daytime?	Yes	No
Has anyone OBSERVED you stop breathing during your sleep?	Yes	No
Do you have or are you being treated for high blood PRESSURE ?	Yes	No
BANG		
BMI more than 35kg/m ² ?	Yes	No
AGE over 50 years old?	Yes	No
NECK circumference > 16inches (40cm)?	Yes	No
GENDER : Male?	Yes	No
TOTAL SCORE		

References

- [1] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, “The Occurrence of Sleep-Disordered Breathing Among Middle-Aged Adults,” *N Engl J Med*, vol. 328, pp. 1230–1235, 1993.
- [2] J. Zhang, Q. Zhang, Y. Wang, and C. Qiu, “A Real-Time Auto-Adjustable Smart Pillow System for Sleep Apnea Detection and Treatment,” in *In Proceedings of the 12th international conference on Information processing in sensor networks (IPSN)*, 2013, pp. 179–190.
- [3] V. K. Kapur *et al.*, “Clinical Practice Guideline for Diagnostic Testing for Adult Obstructive Sleep Apnea,” *J. Clin. Sleep Med.*, vol. 1313, no. 3, pp. 479–504, 2017.
- [4] R. Agarwal and J. Gotman, “Computer-Assisted Sleep Staging,” *IEEE Trans. Biomed. Eng.*, vol. 48, no. 12, pp. 1412–1423, 2001.
- [5] D. R. Hillman, A. S. Murphy, and L. Pezzullo, “The Economic Cost of Sleep Disorders.,” *Sleep*, vol. 29, no. 3, pp. 299–305, 2006.
- [6] N. Alghanim, V. R. Comondore, J. Fleetham, C. A. Marra, and N. T. Ayas, “The Economic Impact of Obstructive Sleep Apnea,” *Lung*, vol. 186, no. 1, pp. 7–12, 2008.
- [7] T. Penzel, J. McNames, A. Murray, P. de Chazal, G. Moody, and B. Raymond, “Systematic Comparison of Different Algorithms for Apnoea Detection Based on Electrocardiogram Recordings,” *Med. Biol. Eng. Comput.*, vol. 40, no. 4, pp. 402–407, 2002.
- [8] S. Boudaoud, H. Rix, O. Meste, C. Heneghan, and C. O’Brien, “Corrected Integral Shape Averaging Applied to Obstructive Sleep Apnea Detection from the Electrocardiogram,” *EURASIP J. Adv. Signal Process.*, vol. 2007, pp. 1–12, 2007.
- [9] D. Patil, V. M. Wadhai, Snehal Gujar, K. Surana, P. Devkate, and Shruti Waghmare, “APNEA Detection on Smart Phone,” *Int. J. Comput. Appl.*, vol. 59, no. 7, pp. 15–19, 2012.
- [10] A. Ravelo-García *et al.*, “Oxygen Saturation and RR Intervals Feature Selection for Sleep Apnea Detection,” *Entropy*, vol. 17, no. 5, pp. 2932–2957, 2015.
- [11] T. M. Cover, “The Best Two Independent Measurements Are Not the Two Best,” *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-4, no. 1, pp. 116–117, 1974.
- [12] P. E. Peppard, T. Young, J. H. Barnet, M. Palta, E. W. Hagen, and K. M. Hla, “Increased Prevalence of Sleep-Disordered Breathing in Adults,” *Am. J. Epidemiol.*, vol. 177, no. 9, pp. 1006–1014, May 2013.
- [13] A. Nassir and O. Barnea, “Wireless Body-Area Network for Detection of Sleep Disorders,” in

- 27th Convention of Electrical and Electronics Engineers in Israel, *IEEEI 2012*, 2012, pp. 1–5.
- [14] L. Grote, T. Ploch, J. Heitmann, L. Knaack, T. Penzel, and J. Peter, “Sleep-Related Breathing Disorder Is An Independent Risk Factor for Systemic Hypertension,” *Am J Respir Crit Care Med*, vol. 160, no. 6, pp. 1875–1882, 1999.
- [15] T. Mooe, K. A. Franklin, K. Holmström, T. Rabben, and U. Wiklund, “Sleep-disordered Breathing and Coronary Artery Disease,” *Am. J. Respir. Crit. Care Med.*, vol. 164, no. 10, pp. 1910–1913, Nov. 2001.
- [16] V. Mohsenin, “Sleep-Related Breathing Disorders And Risk of Stroke,” *Stroke*, vol. 32, no. 6, pp. 1271–8, Jun. 2001.
- [17] S. M. I. Mary, B. Lam, M. T. N. Matthew, W. K. Lam, and K. S. L. L. Kenneth W. T. Tsang, “Obstructive Sleep Apnea Is Independently Associated With Insulin Resistance,” *Am. J. Respir. Crit. Care Med.*, vol. 165, no. 5, pp. 670–676, 2002.
- [18] “The 2005 Sleep in America poll,” *National Sleep Foundation*, 2005. [Online]. Available: https://www.sleepfoundation.org/sites/default/files/inline-files/2005_summary_of_findings.pdf. [Accessed: 24-Mar-2019].
- [19] S. Garbarino, O. Guglielmi, A. Sanna, G. Mancardi, and N. Magnavita, “Risk of Occupational Accidents in Workers with Obstructive Sleep Apnea: Systematic Review and Meta-analysis,” *Sleep*, vol. 39, no. 6, pp. 1211–8, 2016.
- [20] S. Garbarino, A. Pitidis, M. Giustini, F. Taggi, and A. Sanna, “Motor vehicle accidents and obstructive sleep apnea syndrome,” *Chron Respir Dis*, vol. 12, no. 4, pp. 320–8, 2015.
- [21] F. Mendonça, S. S. Mostafa, A. G. Ravelo-García, F. Morgado-Dias, and T. Penzel, “Devices for Home Detection of Obstructive Sleep Apnea: A Review,” *Sleep Med. Rev.*, vol. 41, pp. 149–160, Feb. 2018.
- [22] A. H. Khandoker, J. Gubbi, and M. Palaniswami, “Automated Scoring of Obstructive Sleep Apnea and Hypopnea Events Using Short-Term Electrocardiogram Recordings,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 6, pp. 1057–1067, Nov. 2009.
- [23] G. E. Hinton, S. Osindero, and Y. W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Comput.*, vol. 18, no. 7, pp. 1527–54, 2006.
- [24] T. Penzel, “Home Sleep Testing,” in *Principles and Practice of Sleep Medicine*, 6th ed., Elsevier, 2016, pp. 1610–1614.
- [25] A. L. Chesson, R. B. Berry, and A. Pack, “Practice Parameters for the Use of Portable Monitoring Devices in the Investigation of Suspected Obstructive Sleep Apnea in Adults,” *Sleep*, vol. 26, no. 7, pp. 907–13, Nov. 2003.
- [26] C. Guilleminault, S. Connolly, R. Winkle, K. Melvin, and A. Tilkian, “Cyclical Variation of the Heart Rate in Sleep Apnoea Syndrome. Mechanisms, and Usefulness of 24 H Electrocardiography as a Screening Technique,” *Lancet*, vol. 21, no. 1(8369), pp. 126–31, 1984.

- [27] Wikimedia, “New no obstruction.png.” [Online]. Available: https://commons.wikimedia.org/wiki/File:New_no_obstruction.png. [Accessed: 17-Dec-2019].
- [28] “Airway obstruction.png.” [Online]. Available: https://commons.wikimedia.org/wiki/File:Airway_obstruction.png. [Accessed: 01-Dec-2019].
- [29] A. I. Pack, “Advances in Sleep-disordered Breathing,” *Am. J. Respir. Crit. Care Med.*, vol. 173, no. 1, 2006.
- [30] J. Cheyne, “A Case of Apoplexy in Which the Fleshy Part of The Heart was Converted Into Fat,” *Dublin Hosp Rep*, vol. 2, pp. 216–223, 1818.
- [31] W. Stokes, “The Diseases of the Heart and Aorta,” *Dublin J. Med. Sci.*, vol. XVII, pp. 120–144, 1854.
- [32] C. Burwell, E. Robin, R. Whaley, and A. Bickelmann, “Extreme Obesity Associated with Alveolar Hypoventilation: a Pickwickian Syndrome,” *Am. J. Med.*, vol. 21, no. 5, pp. 811–818, 1956.
- [33] W. Gerardy, D. Herberg, and H. Kuhn, “Vergleichende Untersuchungen der Lungenfunktion und des Elektroencephalogramms Bie Zwei Patienten Mit Pickwickian-syndrom [in German],” *Z Klin Med*, vol. 156, pp. 362–380, 1960.
- [34] D. Drachman and R. Gumnit, “Periodic Alteration of Consciousness in the ‘Pickwickian’ Syndrome,” *Arch Neurol.*, vol. 6, pp. 63–69, 1962.
- [35] H. Gastaut, C. Tassinari, and B. Duron, “Polygraphic Study of the Episodic Diurnal and Nocturnal (Hypnic and Respiratory) Manifestations of the Pickwickian Syndrome,” *Brain Res.*, vol. 2, pp. 167–186, 1965.
- [36] E. Lugaresi, G. Coccagna, M. Mantovani, and F. Brignani, “Effect of Tracheotomy in Hypersomnia with Periodic Respiration,” *Clin Neurophysiol*, vol. 30, pp. 373–374, 1971.
- [37] M. Kryger, L. Quesney, D. Holder, P. Gloor, and P. MacLeod, “The Sleep Deprivation Syndrome of the Obese Patient: a Problem of Periodic Nocturnal Upper Airway Obstruction,” *Am J Med*, vol. 56, pp. 531–539, 1974.
- [38] C. Guilleminault, F. Eldridge, F. Simmons, and W. Dement, “Sleep Apnea in Eight Children,” *Pediatrics*, vol. 58, pp. 23–30, 1976.
- [39] J. Remmers, W. DeGroot, E. Sauerland, and A. Anch, “Pathogenesis of Upper Airway Occlusion During Sleep,” *J Appl Physiol*, vol. 44, pp. 931–938, 1978.
- [40] C. Sullivan, F. Issa, M. Berthon-Jones, and L. Eves, “Reversal of Obstructive Sleep Apnoea by Continuous Positive Airway Pressure Applied Through the Nares,” *Lancet*, vol. 1, pp. 862–865, 1981.
- [41] S. Fujita, W. Conway, F. Zorick, and T. Roth, “Surgical Correction of Anatomic Abnormalities in Obstructive Sleep Apnea Syndrome: Uvulopalatopharyngoplasty,” *Otolaryngol Head Neck Surg*, vol. 89, pp. 923–934, 1981.
- [42] J. Skatrud and J. Dempsey, “Interaction of Sleep State and Chemical Stimuli in Sustaining

- Rhythmic Ventilation,” *J Appl Physiol*, vol. 55, pp. 813–822, 1983.
- [43] G. Gould *et al.*, “The Sleep Hypopnea Syndrome,” *Am Rev Respir Dis*, vol. 137, pp. 895–898, 1988.
- [44] W. Mezzanotte, D. Tangel, and D. White, “Waking Genioglossal Electromyogram in Sleep Apnea Patients Versus Normal Controls (a Neuromuscular Compensatory Mechanism),” *J Clin Invest*, vol. 89, pp. 1571–1579, 1992.
- [45] E. Fletcher, J. Lesske, W. Qian, C. Miller, and T. Unger, “Repetitive, Episodic Hypoxia Causes Diurnal Elevation of Blood Pressure in Rats,” *Hypertension*, vol. 19, pp. 555–561, 1992.
- [46] S. Redline *et al.*, “The Familial Aggregation of Obstructive Sleep Apnea,” *Am J Respir Crit Care Med*, vol. 151, pp. 682–687, 1995.
- [47] G. Pillar and P. Lavie, “Assessment of the Role of Inheritance in Sleep Apnea Syndrome,” *Am J Respir Crit Care Med*, vol. 151, pp. 688–691, 1995.
- [48] R. Mathur and N. Douglas, “Family Studies in Patients with the Sleep Apnea-Hypopnea Syndrome,” *Ann Intern Med*, vol. 122, pp. 174–178, 1995.
- [49] D. Brooks, R. Horner, L. Kozar, C. Render-Teixeira, and E. Phillipson, “Obstructive Sleep Apnea As a Cause of Systemic Hypertension: Evidence from a Canine Model,” *J Clin Invest*, vol. 99, pp. 106–109, 1997.
- [50] G. D., “Sleep-Disordered Breathing and School Performance in Children,” *Pediatrics*, vol. 102, pp. 616–62, 1998.
- [51] C. Jenkinson, R. Davies, R. Mullins, and J. Stradling, “Comparison of Therapeutic and Subtherapeutic Nasal Continuous Positive Airway Pressure for Obstructive Sleep Apnoea: a Randomised Prospective Parallel Trial,” *Lancet*, vol. 353, pp. 2100–2105, 1999.
- [52] J. Pepperell *et al.*, “Ambulatory Blood Pressure After Therapeutic and Subtherapeutic Nasal Continuous Positive Airway Pressure for Obstructive Sleep Apnoea: a Randomised Parallel Trial,” *Lancet*, vol. 359, pp. 204–210, 2002.
- [53] L. Almazaydeh, K. Elleithy, and M. Faezipour, “Detection of Obstructive Sleep Apnea Through ECG Signal Features,” in *IEEE International Conference on Electro Information Technology*, 2012, pp. 1–6.
- [54] M. J. Sateia, “International Classification of Sleep Disorders-Third Edition: Highlights and Modifications,” *Chest*, vol. 146, no. 5, pp. 1387–1394, Nov. 2014.
- [55] GAllegre, “Patient Equipped for a Sleep Apnea Diagnosis (Polysomnography), Ambulatory Diagnosis (Sleeping at Home),” *wikimedia.org*. [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/5/5d/Polysmonograpy_equipped_patient.jpg. [Accessed: 30-Dec-2019].
- [56] R. Lawton, “A Pediatric Patient Prepared for a Polysomnogram by a Respiratory Therapist, St. Louis Children’s Hospital, St. Louis, Missouri, 2006,” 2006. [Online]. Available: https://en.wikipedia.org/wiki/File:Pediatric_polysomnogram.jpg. [Accessed: 30-Dec-2019].

- [57] B. R. Berry, R. Brooks, E. C. Gamaldo, M. S. Harding, C. Marcus, and B. Vaughn, “2012 AASM Manual for the Scoring of Sleep and Associated Events. Rules, Terminology and Technical Specifications (Darien, IL: AASM).” 2012.
- [58] R. Ferber *et al.*, “ASDA Standards of Practice Portable Recording in the Assessment of Obstructive Sleep Apnea,” *Am. Sleep Disord. Assoc. Sleep Res. Soc.*, vol. 17, no. 4, pp. 378–392, 1994.
- [59] O. S. Apnea, “ASDA Standards of Practice Practice Parameters for the Use of Portable Recording in the Assessment of Obstructive Sleep Apnea,” *Am. Sleep Disord. Assoc. Sleep Res. Soc.*, vol. 17, no. 4, pp. 372–377, 1994.
- [60] S. H. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals,” *Circulation*, vol. 101(23), pp. e215–e220, 2000.
- [61] T. Penzel, G. Moody, R. Mark, A. Goldberger, and J. Peter, “The apnea-ECG Database,” in *Computers in Cardiology 2000*, 2000, pp. 255–258.
- [62] “St. Vincent’s University Hospital/University College Dublin Sleep Apnea Database.” [Online]. Available: <https://physionet.org/pn3/ucddb/>.
- [63] C. Iber; S. Ancoli-Israel ; A. L. Chesson ; S. F. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications, 1st edn.* Westchester, IL: American Academy of Sleep Medicine, 2007.
- [64] M. Y. Kiang, “A Comparative Assessment of Classification Methods,” *Decis. Support Syst.*, vol. 35, no. 4, pp. 441–454, 2003.
- [65] I. Ciuca and J. A. Ware, “Layered Neural networks as Universal Approximators,” in *Computational Intelligence Theory and Applications. Fuzzy Days 1997. Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 1997, pp. 411–415.
- [66] M. Gupta, L. Jin, and N. Homma, *Static and Dynamic Neural Networks: from Fundamentals to Advanced Theory*. John Wiley & Sons, 2004.
- [67] C. Cortes, C. Cortes, and V. Vapnik, “Support-Vector Networks,” vol. 20, no. 3, 1995.
- [68] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” in *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*, 1992, pp. 144–152.
- [69] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 2000.
- [70] K. Reza, S. Khatun, M. Jamlos, M. Fakir, and S. Mostafa, “Performance Evaluation of Diversified SVM Kernel Functions for Breast Tumor Early Prognosis,” *ARPN J. Eng. Appl. Sci.*, vol. 9, no. 3, pp. 329–335, 2014.
- [71] T. Hastie, R. Tibshirani, and F. Jerome, *The Elements of Statistical Learning Data Mining*,

- Inference, and Prediction*. Springer-Verlag, 2013.
- [72] M. O. Mendez *et al.*, “Detection of Sleep Apnea from Surface ECG Based on Features Extracted by an Autoregressive Model,” in *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 2007, pp. 6105–6108.
- [73] B. Xie and H. Minn, “Real-Time Sleep Apnea Detection by Classifier Combination,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 469–477, 2012.
- [74] T. Hastie and R. Tibshirani, “Generalized Additive Models,” *Stat. Sci.*, vol. 1, no. 3, pp. 297–318, 1986.
- [75] H. Zhang, “The Optimality of Naive Bayes,” in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, 2004, pp. 562–567.
- [76] H. Zhang, “Exploring Conditions for the Optimality of Naive Bayes,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 19, no. 2, pp. 183–198, 2005.
- [77] L. Kuncheva, “On the Optimality of Naive Bayes with Dependent Binary Features,” *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 830–837, 2006.
- [78] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press: Cambridge, MA, USA, 2016.
- [79] J. S. J. Ren and L. Xu, “On Vectorization of Deep Convolutional Neural Networks for Vision Tasks,” *To Appear AAAI-2015*, no. 2003, pp. 1840–1846, 2015.
- [80] T. Penzel and R. Conradt, “Computer based sleep recording and analysis,” *Sleep Med. Rev.*, vol. 4, no. 2, pp. 131–148, Apr. 2000.
- [81] D. Alvarez-Estevéz and V. Moret-Bonillo, “Computer-Assisted Diagnosis of the Sleep Apnea-Hypopnea Syndrome: A Review,” *Sleep Disord.*, vol. 2015, p. 237878, 2015.
- [82] N. A. Collop *et al.*, “Obstructive Sleep Apnea Devices for Out-Of-Center (OOC) Testing: Technology Evaluation,” *J. Clin. Sleep Med.*, vol. 07, no. 05, pp. 531–548, Oct. 2011.
- [83] S. A. Pullano *et al.*, “Medical Devices for Pediatric Apnea Monitoring and Therapy: Past and New Trends,” *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 199–212, 2017.
- [84] W. W. Flemons *et al.*, “Home Diagnosis of Sleep Apnea: A Systematic Review of the Literature: An Evidence Review Cosponsored by the American Academy of Sleep Medicine, the American College of Chest Physicians, and the American Thoracic Society,” *Chest*, vol. 124, no. 4, pp. 1543–1579, Oct. 2003.
- [85] M. Al-Mardini, F. Aloul, A. Sagahyoon, and L. Al-Husseini, “On the Use of Smartphones for Detecting Obstructive Sleep Apnea,” in *13th IEEE International Conference on BioInformatics and BioEngineering*, 2013, pp. 1–4.
- [86] M. Al-Mardini, F. Aloul, A. Sagahyoon, and L. Al-Husseini, “Classifying Obstructive Sleep Apnea using Smartphones,” *J. Biomed. Inform.*, vol. 52, pp. 251–259, Dec. 2014.
- [87] F. Mendonca, S. S. Mostafa, A. G. Ravelo-Garcia, F. Morgado-Dias, and T. Penzel, “A Review of Obstructive Sleep Apnea Detection Approaches,” *IEEE J. Biomed. Heal. Informatics*, vol.

- 23, no. 2, pp. 825–837, 2019.
- [88] Mostafa, Mendonça, Ravelo-García, and Morgado-Dias, “A Systematic Review of Detecting Sleep Apnea Using Deep Learning,” *Sensors*, vol. 19, no. 22, p. 4934, Nov. 2019.
- [89] S. F. Quan *et al.*, “The Sleep Heart Health Study: Design, Rationale, and Methods,” *Sleep*, vol. 20, no. 12, pp. 1077–85, Dec. 1997.
- [90] I. Y and M. GB, “Development of the Polysomnographic Database on CD-ROM,” *Psychiatry Clin. Neurosci.*, vol. 53, pp. 175–177, 1999.
- [91] D. A. Dean *et al.*, “Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource,” *Sleep*, vol. 39, no. 5, pp. 1151–1164, May 2016.
- [92] D. W. Jung *et al.*, “Real - Time Automatic Apneic Event Detection using Nocturnal Pulse Oximetry,” *IEEE Trans. Biomed. Eng.*, vol. 65, no. 3, pp. 706–712, 2018.
- [93] D. Alvarez, R. Hornero, D. Abásolo, F. del Campo, and C. Zamarrón, “Nonlinear Characteristics of Blood Oxygen Saturation from Nocturnal Oximetry for Obstructive Sleep Apnoea Detection,” *Physiol. Meas.*, vol. 27, no. 4, pp. 399–412, 2006.
- [94] J. V. Marcos, R. Hornero, D. Álvarez, F. Del Campo, and M. Aboy, “Automated Detection of Obstructive Sleep Apnoea Syndrome from Oxygen Saturation Recordings using Linear Discriminant Analysis,” *Med. Biol. Eng. Comput.*, vol. 48, no. 9, pp. 895–902, Sep. 2010.
- [95] J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo, and C. Zamarrón, “Assessment of Four Statistical Pattern Recognition Techniques to Assist in Obstructive Sleep Apnoea Diagnosis from Nocturnal Oximetry,” *Med. Eng. Phys.*, vol. 31, no. 8, pp. 971–978, Oct. 2009.
- [96] J. F. Morales *et al.*, “Sleep Apnea Hypopnea Syndrome Classification in SpO2 Signals using Wavelet Decomposition and Phase Space Reconstruction,” in *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2017, pp. 43–46.
- [97] D. Álvarez, R. Hornero, J. V. Marcos, and F. del Campo, “Feature Selection from Nocturnal Oximetry using Genetic Algorithms to Assist in Obstructive Sleep Apnoea Diagnosis,” *Med. Eng. Phys.*, vol. 34, no. 8, pp. 1049–1057, Oct. 2012.
- [98] D. S. Morillo and N. Gross, “Probabilistic Neural Network Approach for the Detection of SAHS from Overnight Pulse Oximetry,” *Med. Biol. Eng. Comput.*, vol. 51, no. 3, pp. 305–315, Mar. 2013.
- [99] L. Almazaydeh, M. Faezipour, and K. Elleithy, “A Neural Network System for Detection of Obstructive Sleep Apnea Through SpO2 Signal Features,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 5, pp. 7–11, 2012.
- [100] D. Alvarez *et al.*, “Automated Analysis of Unattended Portable Oximetry by Means of Bayesian Neural Networks to Assist in the Diagnosis Of Sleep Apnea,” in *2016 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges (GMEPE/PAHCE)*, 2016, pp. 1–4.
- [101] J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo, C. Zamarrón, and M. López, “Utility of

- Multilayer Perceptron Neural Network Classifiers in The Diagnosis of the Obstructive Sleep Apnoea Syndrome from Nocturnal Oximetry,” *Comput. Methods Programs Biomed.*, vol. 92, no. 1, pp. 79–89, Oct. 2008.
- [102] S. S. Mostafa, J. P. Carvalho, F. Morgado-Dias, and A. Ravelo-García, “Optimization of Sleep Apnea Detection using SpO₂ and ANN,” in *2017 XXVI International Conference on Information, Communication and Automation Technologies (ICAT)*, 2017, pp. 1–6.
- [103] D. Álvarez, R. Hornero, J. V. Marcos, and F. del Campo, “Multivariate Analysis of Blood Oxygen Saturation Recordings in Obstructive Sleep Apnea Diagnosis,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 12, pp. 2816–2824, Dec. 2010.
- [104] J. Lazaro, E. Gil, J. M. Vergara, and P. Laguna, “Pulse Rate Variability Analysis for Discrimination of Sleep-Apnea-Related Decreases in the Amplitude Fluctuations of Pulse Photoplethysmographic Signal in Children,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 1, pp. 240–246, Jan. 2014.
- [105] A. Garde, P. Dehkordi, W. Karlen, D. Wensley, J. M. Ansermino, and G. A. Dumont, “Development of A Screening Tool for Sleep Disordered Breathing in Children using the Phone Oximeter™,” *PLoS One*, vol. 9, no. 11, p. e112959, 2014.
- [106] S. S. Mostafa, F. Mendonça, F. Morgado-Dias, and A. Ravelo-García, “SpO₂ Based Sleep Apnea Detection using Deep Learning,” in *2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*, pp. 91–96.
- [107] R. K. Pathinarupothi *et al.*, “Single Sensor Techniques for Sleep Apnea Diagnosis Using Deep Learning,” in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017, pp. 524–529.
- [108] R. Lin, R.-G. Lee, C.-L. Tseng, H.-K. Zhou, C.-F. Chao, and J.-A. Jiang, “A New Approach for Identifying Sleep Apnea Syndrome using Wavelet Transform and Neural Networks,” *Biomed. Eng. - Appl. Basis Commun.*, vol. 18, no. 3, pp. 138–143, 2006.
- [109] V. P. Rachim, G. Li, and W.-Y. Chung, “Sleep Apnea Classification using ECG-Signal Wavelet-PCA Features,” *Biomed. Mater. Eng.*, vol. 24, no. 6, pp. 2875–2882, 2014.
- [110] A. R. Hassan, “Computer-Aided Obstructive Sleep Apnea Detection using Normal Inverse Gaussian Parameters and Adaptive Boosting,” *Biomed. Signal Process. Control*, vol. 29, pp. 22–30, 2016.
- [111] A. R. Hassan and M. A. Haque, “An Expert System for Automated Identification of Obstructive Sleep Apnea from Single-Lead ECG using Random under Sampling Boosting,” *Neurocomputing*, vol. 235, pp. 122–130, Apr. 2017.
- [112] A. Smruthy, M. Suchetha, and L. Fellow, “Real-Time Classification of Healthy and Apnea Subjects Using ECG Signals With Variational Mode Decomposition,” *3092 IEEE SENSORS J.*, vol. 17, no. 10, pp. 3092–3099, 2017.
- [113] A. R. Hassan, “Automatic Screening of Obstructive Sleep Apnea from Single-Lead

- Electrocardiogram,” in *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 2015, pp. 1–6.
- [114] A. F. F. Quiceno-Manrique, J. B. B. Alonso-Hernandez, C. M. M. Travieso-Gonzalez, M. A. A. Ferrer-Ballester, and G. Castellanos-Dominguez, “Detection of Obstructive Sleep Apnea in ECG Recordings using Time-Frequency Distributions and Dynamic Features,” in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, 2009, pp. 5559–5562.
- [115] J. D. Martinez-Vargas, L. M. Sepulveda-Cano, and G. Castellanos-Dominguez, “On Determining Available Stochastic Features by Spectral Splitting in Obstructive Sleep Apnea Detection,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 6079–6082.
- [116] K. Kesper, S. Canisius, T. Penzel, T. Ploch, and W. Cassel, “ECG Signal Analysis for the Assessment of Sleep-Disordered Breathing and Sleep Pattern,” *Med. Biol. Eng. Comput.*, vol. 50, no. 2, pp. 135–144, Feb. 2012.
- [117] A. G. Ravelo-García *et al.*, “Symbolic Dynamics Marker of Heart Rate Variability Combined with Clinical Variables Enhance Obstructive Sleep Apnea Screening,” *Cit. Chaos*, vol. 24, 2014.
- [118] C. W. Zywietz, V. Von Einem, B. Widiger, and G. Joseph, “ECG Analysis for Sleep Apnea Detection,” *Methods Inf. Med.*, vol. 43, no. 1, pp. 56–9, 2004.
- [119] G. Gutiérrez-Tobal, D. Álvarez, J. Gomez-Pilar, F. del Campo, and R. Hornero, “Assessment of Time and Frequency Domain Entropies to Detect Sleep Apnoea in Heart Rate Variability Recordings from Men and Women,” *Entropy*, vol. 17, no. 1, pp. 123–141, Jan. 2015.
- [120] F. Roche *et al.*, “Predicting Sleep Apnoea Syndrome from Heart Period: a Time-Frequency Wavelet Analysis,” *Eur. Respir. J.*, vol. 22, pp. 937–942, 2003.
- [121] A. G. Ravelo-garcía, J. L. Navarro-mesa, L. Palmas, D. G. Canaria, S. De Neumología, and H. Universitario, “Cepstrum Feature Selection for the Classification of Sleep Apnea-Hypopnea Syndrome based on Heart Rate Variability Cepstrum analysis,” in *2013 Computing in Cardiology Conference*, 2013, pp. 959–962.
- [122] M. O. Mendez, A. M. Bianchi, M. Matteucci, S. Cerutti, and T. Penzel, “Sleep Apnea Screening by Autoregressive Models From a Single ECG Lead,” *IEEE Trans. Biomed. Eng.*, vol. 56, no. 12, pp. 2838–2850, Dec. 2009.
- [123] C. Cheng, C. Kan, and H. Yang, “Heterogeneous Recurrence Analysis of Heartbeat Dynamics for the Identification of Sleep Apnea Events,” *Comput. Biol. Med.*, vol. 75, pp. 10–18, Aug. 2016.
- [124] L. Chen and X. Zhang, “State-Based General Gamma CUSUM for Modeling Heart Rate Variability Using Electrocardiography Signals,” *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 2, pp. 1160–1171, Apr. 2017.

- [125] B. Yılmaz *et al.*, “Sleep Stage and Obstructive Apneic Epoch Classification using Single-Lead ECG,” *Biomed. Eng. Online*, vol. 9, no. 1, p. 39, 2010.
- [126] P. de Chazal, T. Penzel, and C. Heneghan, “Automated Detection of Obstructive Sleep Apnoea at Different Time Scales using the Electrocardiogram,” *Physiol. Meas.*, vol. 25, no. 4, pp. 967–983, 2004.
- [127] A. G. Ravelo *et al.*, “Application of Support Vector Machines and Gaussian Mixture Models for the Detection of Obstructive Sleep Apnoea based on the RR Series,” in *8th WSEAS International Conference on Applied Mathematics*, 2005, pp. 139–143.
- [128] C. M. Travieso, J. B. Alonso, M. del Pozo-Baños, J. R. Ticay-Rivas, and K. Lopez-de-Ipiña, “Automatic Apnea Identification by Transformation of the Cepstral Domain,” *Cognit. Comput.*, vol. 5, no. 4, pp. 558–565, Dec. 2013.
- [129] H. D. Nguyen, B. A. Wilkins, Q. Cheng, and B. A. Benjamin, “An Online Sleep Apnea Detection Method Based on Recurrence Quantification Analysis,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 4, pp. 1285–1293, Jul. 2014.
- [130] L. Chen, X. Zhang, and C. Song, “An Automatic Screening Approach for Obstructive Sleep Apnea Diagnosis Based on Single-Lead Electrocardiogram,” *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 1, pp. 106–115, Jan. 2015.
- [131] M. Cheng, W. J. Sori, F. Jiang, A. Khan, and S. Liu, “Recurrent Neural Network Based Classification of ECG Signal Features for Obstruction of Sleep Apnea Detection,” in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 2017, pp. 199–202.
- [132] R. K. Pathinarupothi, R. Vinaykumar, E. Rangan, E. Gopalakrishnan, and K. P. Soman, “Instantaneous Heart Rate As a Robust Feature for Sleep Apnea Severity Detection using Deep Learning,” in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2017, pp. 293–296.
- [133] T. Penzel *et al.*, “Modulations of Heart Rate, ECG, and Cardio-Respiratory Coupling Observed in Polysomnography,” *Front. Physiol.*, vol. 7, p. 460, 2016.
- [134] C. Song, K. Liu, X. Zhang, L. Chen, and X. Xian, “An Obstructive Sleep Apnea Detection Approach Using a Discriminative Hidden Markov Model from ECG Signals,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1532–1542, Jul. 2016.
- [135] C. Maier, H. Wenz, and H. Dickhaus, “Robust Detection of Sleep Apnea from Holter ECGs,” *Methods Inf. Med.*, vol. 53, no. 04, pp. 303–307, Jan. 2014.
- [136] A. Ravelo-García *et al.*, “Application of the Permutation Entropy over the Heart Rate Variability for the Improvement of Electrocardiogram-based Sleep Breathing Pause Detection,” *Entropy*, vol. 17, no. 3, pp. 914–927, Feb. 2015.
- [137] S. Martín-González, J. L. Navarro-Mesa, G. Juliá-Serdá, J. F. Kraemer, N. Wessel, and A. G. Ravelo-García, “Heart Rate Variability Feature Selection in the Presence of Sleep Apnea: an

- Expert System for the Characterization and Detection of the Disorder,” *Comput. Biol. Med.*, vol. 91, pp. 47–58, Dec. 2017.
- [138] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, “Support Vector Machines for Automated Recognition of Obstructive Sleep Apnea Syndrome From ECG Recordings,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 1, pp. 37–48, Jan. 2009.
- [139] A. H. Khandoker, C. K. Karmakar, and M. Palaniswami, “Automated Recognition of Patients with Obstructive Sleep Apnoea using Wavelet-Based Features of Electrocardiogram Recordings,” *Comput. Biol. Med.*, vol. 39, pp. 88–96, 2009.
- [140] A. Yildiz, M. Akin, M. Poyraz, M. Akin, and M. Poyraz, “An Expert System for Automated Recognition of Patients with Obstructive Sleep Apnea using Electrocardiogram Recordings,” *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12880–12890, Sep. 2011.
- [141] B. Koley and D. Dey, “Automated Detection of Apnea and Hypopnea Events,” in *2012 Third International Conference on Emerging Applications of Information Technology*, 2012, pp. 85–88.
- [142] B. L. Koley and D. Dey, “Real-Time Adaptive Apnea and Hypopnea Event Detection Methodology for Portable Sleep Apnea Monitoring Devices,” *IEEE Trans. Biomed. Eng.*, vol. 60, no. 12, pp. 3354–3363, Dec. 2013.
- [143] P. Caseiro, R. Fonseca-Pinto, and A. Andrade, “Screening of Obstructive Sleep Apnea using Hilbert–Huang Decomposition of Oronasal Airway Pressure Recordings,” *Med. Eng. Phys.*, vol. 32, no. 6, pp. 561–568, Jul. 2010.
- [144] G. C. Gutiérrez-Tobal, R. Hornero, D. Álvarez, J. V Marcos, and F. del Campo, “Linear and Nonlinear Analysis of Airflow Recordings to Help in Sleep Apnoea–Hypopnoea Syndrome Diagnosis,” *Physiol. Meas.*, vol. 33, no. 7, pp. 1261–1275, Jul. 2012.
- [145] N. Selvaraj and R. Narasimhan, “Detection of Sleep Apnea on a Per-Second Basis using Respiratory Signals,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2013, pp. 2124–2127.
- [146] M. M. Paul and Amithab M, “SAHS Detection Based on ANFIS Using Single Channel Airflow Signal,” *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 5, no. 7, pp. 13053–13061, 2016.
- [147] C. Avcı and A. Akbaş, “Sleep Apnea Classification Based on Respiration Signals by using Ensemble Methods,” *Biomed. Mater. Eng.*, vol. 26, no. s1, pp. S1703–S1710, Aug. 2015.
- [148] G. Ozdemir, H. Nasifoglu, and O. Erogul, “A Time-Series Approach to Predict Obstructive Sleep Apnea (OSA) Episodes,” in *Proceedings of the 2nd World Congress on Electrical Engineering and Computer Systems and Science*, 2016, pp. 1–8.
- [149] R. Haidar, I. Koprinska, and B. Jeffries, “Sleep Apnea Event Detection from Nasal Airflow using Convolutional Neural Networks,” in *Proceedings of the International Conference on Neural Information Processing (ICONIP)*, 2017, pp. 819–827.
- [150] A. Thommandram, J. M. Eklund, and C. McGregor, “Detection of Apnoea from Respiratory

- Time Series Data using Clinically Recognizable Features and KNN Classification,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 5013–5016, 2013.
- [151] Y. Maali and A. Al-Jumaily, “Automated Detecting Sleep Apnea Syndrome: a Novel System Based on Genetic SVM,” in *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, 2011, pp. 590–594.
- [152] T. Rosenwein, E. Dafna, A. Tarasiuk, and Y. Zigel, “Breath-By-Breath Detection of Apneic Events for OSA Severity Estimation using Non-Contact Audio Recordings,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 7688–7691.
- [153] L. Almazaydeh, K. Elleithy, M. Faezipour, and A. Abushakra, “Apnea Detection Based on Respiratory Signal Classification,” *Procedia Comput. Sci.*, vol. 21, pp. 310–316, 2013.
- [154] T. Praydas, B. Wongkittisuksa, and S. Tanthanuch, “Obstructive Sleep Apnea Severity Multiclass Classification Using Analysis of Snoring Sounds,” *Proc. 2nd World Congr. Electr. Eng. Comput. Syst. Sci.*, pp. 16–20, 2016.
- [155] A. K. Ng, T. S. Koh, E. Baey, T. H. Lee, U. R. Abeyratne, and K. Puvanendran, “Could Formant Frequencies of Snore Signals Be an Alternative Means for The Diagnosis of Obstructive Sleep Apnea?,” *Sleep Med.*, vol. 9, no. 8, pp. 894–898, Dec. 2008.
- [156] A. S. Karunajeewa, U. R. Abeyratne, and C. Hukins, “Multi-Feature Snore Sound Analysis in Obstructive Sleep Apnea–Hypopnea Syndrome,” *Physiol. Meas.*, vol. 32, no. 1, pp. 83–97, Jan. 2011.
- [157] E. Goldshtein, A. Tarasiuk, and Y. Zigel, “Automatic Detection of Obstructive Sleep Apnea using Speech Signals,” *IEEE Trans. Biomed. Eng.*, vol. 58, no. 5, pp. 1373–1382, 2011.
- [158] R. Fernández Pozo, J. L. Blanco Murillo, L. Hernández Gómez, E. López Gonzalo, J. Alcázar Ramírez, and D. T. Toledano, “Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker Recognition Techniques,” *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, p. 982531, Dec. 2009.
- [159] A. Montero Benavides, R. Fernández Pozo, D. T. Toledano, J. L. Blanco Murillo, E. López Gonzalo, and L. Hernández Gómez, “Analysis of Voice Features Related to Obstructive Sleep Apnoea and Their Application in Diagnosis Support,” *Comput. Speech Lang.*, vol. 28, no. 2, pp. 434–452, Mar. 2014.
- [160] T. Penzel and A. Sabil, “The Use of Tracheal Sounds for the Diagnosis of Sleep Apnoea,” *Breathe*, vol. 13, no. 2, pp. e37–e45, Jun. 2017.
- [161] C. Kalkbrenner, M. Eichenlaub, S. Rüdiger, C. Kropf-Sanchen, W. Rottbauer, and R. Brucher, “Apnea and Heart Rate Detection from Tracheal Body Sounds for the Diagnosis of Sleep-Related Breathing Disorders,” *Med. Biol. Eng. Comput.*, vol. 56, no. 4, pp. 671–681, Apr. 2018.
- [162] C. Zamarrón, F. Gude, J. Barcala, J. R. Rodriguez, and P. V. Romero, “Utility of Oxygen

- Saturation and Heart Rate Spectral Analysis Obtained From Pulse Oximetric Recordings in the Diagnosis of Sleep Apnea Syndrome,” *Chest*, vol. 123, no. 5, pp. 1567–1576, May 2003.
- [163] D. Álvarez, R. Hornero, J. Víctor Marcos, F. Del Campo, and M. López, “Spectral Analysis of Electroencephalogram and Oximetric Signals in Obstructive Sleep Apnea Diagnosis,” in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*, 2009, pp. 400–403.
- [164] L. Poupard, C. Philippe, M. D. Goldman, R. Sartène, and M. Mathieu, “Novel Mathematical Processing Method of Nocturnal Oximetry for Screening Patients with Suspected Sleep Apnoea Syndrome,” *Sleep Breath.*, vol. 16, no. 2, pp. 419–425, Jun. 2012.
- [165] G. Memis and M. Sert, “Multimodal Classification of Obstructive Sleep Apnea using Feature Level Fusion,” in *2017 IEEE 11th International Conference on Semantic Computing*, 2017, pp. 85–88.
- [166] K. V. Madhav, E. H. Krishna, and K. A. Reddy, “Detection of Sleep Apnea from Multiparameter Monitor Signals using Empirical Mode Decomposition,” in *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 2017, pp. 1–6.
- [167] J. Y. Tian and J. Q. Liu, “Apnea Detection Based on Time Delay Neural Network,” in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 2005, pp. 2571–2574.
- [168] A. Yadollahi, E. Giannouli, and Z. Moussavi, “Sleep Apnea Monitoring and Diagnosis Based on Pulse Oximetry and Tracheal Sound Signals,” *Med. Biol. Eng. Comput.*, vol. 48, no. 11, pp. 1087–1097, Nov. 2010.
- [169] H. M. Al-Angari and A. V. Sahakian, “Automated Recognition of Obstructive Sleep Apnea Syndrome Using Support Vector Machine Classifier,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 3, pp. 463–468, May 2012.
- [170] D. Sommermeyer, D. Zou, L. Grote, and J. Hedner, “Detection of Sleep Disordered Breathing and Its Central/Obstructive Character Using Nasal Cannula and Finger Pulse Oximeter,” *J. Clin. Sleep Med.*, vol. 8, no. 5, pp. 527–33, Oct. 2012.
- [171] K. Polat, Ş. Yosunkaya, and S. Güneş, “Comparison of Different Classifier Algorithms on the Automated Detection of Obstructive Sleep Apnea Syndrome,” *J. Med. Syst.*, vol. 32, no. 3, pp. 243–250, Jun. 2008.
- [172] F. Espinoza-Cuadros, R. Fernández-Pozo, D. T. Toledano, J. D. Alcázar-Ramírez, E. López-Gonzalo, and L. A. Hernández-Gómez, “Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment,” *Comput. Math. Methods Med.*, vol. 2015, pp. 1–13, Nov. 2015.
- [173] R. Jayaraj, J. Mohan, and A. Kanagasabai, “A Review on Detection and Treatment Methods of Sleep Apnea,” *J. Clin. Diagn. Res.*, vol. 11, no. 3, pp. VE01–VE03, Mar. 2017.
- [174] K. Li, W. Pan, Y. Li, Q. Jiang, and G. Liu, “A Method to Detect Sleep Apnea Based on Deep Neural Network and Hidden Markov Model using Single-Lead ECG Signal,” *Neurocomputing*,

- vol. 294, pp. 94–101, Jun. 2018.
- [175] J. Pan and W. J. Tompkins, “A Real-Time QRS Detection Algorithm,” *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.
- [176] T. Kim, J.-W. Kim, and K. Lee, “Detection of Sleep Disordered Breathing Severity using Acoustic Biomarker and Machine Learning Techniques,” *Biomed. Eng. Online*, vol. 17, no. 1, p. 16, Dec. 2018.
- [177] P. Lakhan, A. Dittthapron, N. Banluesombatkul, and T. Wilaiprasitporn, “Deep Neural Networks with Weighted Averaged Overnight Airflow Features for Sleep Apnea-Hypopnea Severity Classification,” in *arXiv*, 2018, pp. 1–5.
- [178] I. De Falco *et al.*, “Deep Neural Network Hyper-Parameter Setting for Classification of Obstructive Sleep Apnea Episodes,” in *2018 IEEE Symposium on Computers and Communications (ISCC)*, 2018, pp. 01187–01192.
- [179] E. Urtnasan, J.-U. Park, and K.-J. Lee, “Multiclass Classification of Obstructive Sleep Apnea/Hypopnea Based on A Convolutional Neural Network from a Single-Lead Electrocardiogram,” *Physiol. Meas.*, vol. 39, no. 6, p. 065003, Jun. 2018.
- [180] E. Urtnasan, J. Park, E. Joo, and K. Lee, “Automated Detection of Obstructive Sleep Apnea Events from a Single-Lead Electrocardiogram Using a Convolutional Neural Network,” *J. Med. Syst.*, vol. 42, no. 6, p. 104, Jun. 2018.
- [181] D. Dey, S. Chaudhuri, and S. Munshi, “Obstructive Sleep Apnoea Detection using Convolutional Neural Network Based Deep Learning Framework,” *Biomed. Eng. Lett.*, vol. 8, no. 1, pp. 95–100, Feb. 2018.
- [182] S. H. Choi *et al.*, “Real-Time Apnea-Hypopnea Event Detection During Sleep by Convolutional Neural Networks,” *Comput. Biol. Med.*, vol. 100, pp. 123–131, Sep. 2018.
- [183] R. Haidar, S. McCloskey, I. Koprinska, and B. Jeffries, “Convolutional Neural Networks on Multiple Respiratory Channels to Detect Hypopnea and Obstructive Apnea Events,” in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–7.
- [184] S. McCloskey, R. Haidar, I. Koprinska, and B. Jeffries, “Detecting Hypopnea and Obstructive Apnea Events using Convolutional Neural Networks on Wavelet Spectrograms of Nasal Airflow,” in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2018, pp. 361–372.
- [185] L. Cen, Z. L. Yu, T. Kluge, and W. Ser, “Automatic System for Obstructive Sleep Apnea Events Detection Using Convolutional Neural Network,” in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 3975–3978.
- [186] T. Van Steenkiste, W. Groenendaal, D. Deschrijver, and T. Dhaene, “Automated Sleep Apnea Detection in Raw Respiratory Signals using Long Short-Term Memory Neural Networks,” *IEEE J. Biomed. Heal. Informatics*, vol. PP, no. c, p. 1, 2018.

- [187] S. Hochreiter and J. Jürgen Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [188] D. Novak, K. Mucha, and T. Al-Ani, “Long Short-Term Memory for apnea detection based on Heart Rate Variability,” in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp. 5234–5237.
- [189] E. Urtnasan, J. U. Park, and K. J. Lee, “Automatic Detection of Sleep-Disordered Breathing Events using Recurrent Neural Networks from an Electrocardiogram Signal,” *Neural Comput. Appl.*, vol. 2, 2018.
- [190] S. Biswal, H. Sun, B. Goparaju, M. B. Westover, J. Sun, and M. T. Bianchi, “Expert-Level Sleep Scoring with Deep Neural Networks,” *J. Am. Med. Informatics Assoc.*, vol. 25, no. 12, pp. 1643–1650, 2018.
- [191] N. Banluesombatkul, T. Rakthanmanon, and T. Wilaiprasitporn, “Single Channel ECG for Obstructive Sleep Apnea Severity Detection using a Deep Learning Approach,” in *arxiv*, 2018.
- [192] J. B. Blank *et al.*, “Overview of Recruitment for the Osteoporotic Fractures In Men Study (MrOS),” *Contemp. Clin. Trials*, vol. 26, no. 5, pp. 557–68, Oct. 2005.
- [193] “Sleep Heart Health Study.” [Online]. Available: <https://sleepdata.org/datasets/shhs>. [Accessed: 11-Jan-2019].
- [194] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, “Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity.,” *Emerg. (Tehran, Iran)*, vol. 3, no. 2, pp. 48–9, 2015.
- [195] S. S. Mostafa, F. Morgado-Dias, and A. G. Ravelo-García, “Comparison of SFS and mRMR for Oximetry Feature Selection in Obstructive Sleep Apnea Detection,” *Neural Comput. Appl.*, pp. 1–21, Apr. 2018.
- [196] S. S. Mostafa, F. Mendonça, G. Juliá-Serdá, F. Morgado-Dias, and A. G. Ravelo-García, “SC3: Self-Configuring Classifier Combination for Obstructive Sleep Apnea,” *Neural Comput. Appl.*, 2019.
- [197] A. G. . Ravelo-García *et al.*, “Application of RR Series and Oximetry to a Statistical Classifier for the Detection of Sleep Apnoea/Hipopnoea,” in *Computers in Cardiology, 2004*, 2004, pp. 305–308.
- [198] A. R. Warley, J. H. Mitchell, and J. R. Stradling, “Evaluation of the Ohmeda 3700 Pulse Oximeter,” *Thorax*, vol. 42, no. 11, pp. 892–896, 1987.
- [199] L. G. Olson, A. Ambrogetti, and S. G. Gyulay, “Prediction of Sleep-Disordered Breathing by Unattended Overnight Oximetry,” *J. Sleep Res.*, vol. 8, no. 1, pp. 51–55, 1999.
- [200] S. Gyulay, L. G. Olson, M. J. Hensley, M. T. King, K. M. Allen, and N. A. Saunders, “A Comparison of Clinical Assessment and Home Oximetry in the Diagnosis of Obstructive Sleep Apnea,” *Am. Rev. Respir. Dis.*, vol. 147, no. 1, pp. 50–53, Jan. 1993.
- [201] P. de Chazal, C. Heneghan, and W. T. McNicholas, “Multimodal Detection of Sleep Apnoea

- using Electrocardiogram and Oximetry Signals,” *Philos. Trans. R. Soc. London A Math. Phys. Eng. Sci.*, vol. 367, no. 1887, pp. 369–89, 2009.
- [202] C. M. Nunes, A. de S. Britto Jr., C. A. A. Kaestner, and R. Sabourin, “Feature Subset Selection Using an Optimized Hill Climbing Algorithm for Handwritten Character Recognition,” in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 2004, pp. 1018–1025.
- [203] M. Lovay, G. Peretti, and E. Romero, “Application of Genetic Algorithms in The Design of Robust Active Filters,” in *2015 Argentine School of Micro-Nanoelectronics, Technology and Applications (EAMTA)*, 2015, pp. 1–6.
- [204] X. Yu and M. Gen, *Introduction to Evolutionary Algorithms*, 1st ed. London: Springer-Verlag London, 2010.
- [205] “PhysioNet.” [Online]. Available: www.physionet.org.
- [206] M. Schrader *et al.*, “Detection of Sleep Apnea in Single Channel ECGs from the PhysioNet Data Base,” in *Computers in Cardiology 2000. Vol.27 (Cat. 00CH37163)*, 2000, pp. 263–266.
- [207] L. I. Kuncheva, *Combining pattern classifiers: Methods and algorithms*. John Wiley & Sons, 2004.
- [208] G. Fumera and F. Roli, “A Theoretical and Experimental Analysis of Linear Combiners for Multiple Classifier Systems,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 942–956, 2005.
- [209] A. B. Santos, A. de A. Araújo, and D. Menotti, “Combiner of Classifiers Using Genetic Algorithm for Classification of Remote Sensed Hyperspectral Images,” in *2012 IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 4146–4149.
- [210] M. Mohandes, M. Deriche, and S. Aliyu, “Classifiers Combination Techniques: A Comprehensive Review,” *IEEE Access*, vol. 6, pp. 19626–19639, 2018.
- [211] S. S. Mostafa, F. Mendonça, A. Ravelo-García, and F. Morgado-Dias, “Combination of Deep and Shallow Networks for Cyclic Alternating Patterns Detection,” in *13th APCA International Conference on Automatic Control and Soft Computing (CONTROLO)*, 2018, pp. 98–103.
- [212] J. Hartmanis and J. Van Leeuwen, *Multiple Classifier Systems: First International Workshop Proceedings*. 2000.
- [213] G. G. Berntson, J. T. Cacioppo, and K. S. Quigley, “Respiratory Sinus Arrhythmia: Autonomic Origins, Physiological Mechanisms, and Psychophysiological Implications,” *Psychophysiology*, vol. 30, 1993.
- [214] O. Yildirim, U. B. Baloglu, R. S. Tan, E. J. Ciaccio, and U. R. Acharya, “Heart rate variability Standards Standards of Measurement, Physiological Interpretation, and Clinical Use,” *Eur. Hear. J.*, vol. 17, pp. 354–381, 1996.
- [215] A. Otero, S. F. Dapena, P. Felix, J. Presedo, and M. Tarasco, “A Low Cost Screening Test for Obstructive Sleep Apnea That Can Be Performed at the Patient’s Home,” in *2009 IEEE*

- International Symposium on Intelligent Signal Processing*, 2009, pp. 199–204.
- [216] O. Salem, Yaning Liu, and A. Mehaoua, “Pervasive Detection of Sleep Apnea using Medical Wireless Sensor Networks,” in *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2014, pp. 435–440.
- [217] F. Shaffer and J. P. Ginsberg, “An Overview of Heart Rate Variability Metrics and Norms,” *Front. Public Heal.*, vol. 5, no. September, pp. 1–17, 2017.
- [218] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, “Deep Learning for Visual Understanding: a Review,” *Neurocomputing*, vol. 187, pp. 27–48, Apr. 2016.
- [219] S. Kiranyaz, T. Ince, and M. Gabbouj, “Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, 2016.
- [220] S. Albelwi and A. Mahmood, “Framework for Designing The Architectures of Deep Convolutional Neural Networks,” *Entropy*, vol. 19, no. 6, p. 242, 2017.
- [221] Y. Zhining and P. Yunming, “The Genetic Convolutional Neural Network Model Based on Random Sample,” *Int. J. u- e- Serv. Sci. Technol.*, vol. 8, no. 11, pp. 317–326, 2015.
- [222] E. Real *et al.*, “Large-Scale Evolution of Image Classifiers,” in *arXiv preprint*, 2017.
- [223] L. Xie and A. Yuille, “Genetic CNN,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1388–1397.
- [224] S. R. Young *et al.*, “Evolving Deep Networks Using HPC,” in *MLHPC’17 Proceedings of the Machine Learning on HPC Environments*, 2017.
- [225] F. Assuncao, N. Lourenco, P. Machado, and B. Ribeiro, “Evolving the Topology of Large Scale Deep Neural Networks,” in *EuroGP 2018: Proceedings of the 21st European Conference on Genetic Programming*, 2018, pp. 19–34.
- [226] A. Baldominos, Y. Saez, and P. Isasi, “Evolutionary Convolutional Neural Networks: an Application to Handwriting Recognition,” *Neurocomputing*, vol. 283, pp. 38–52, Mar. 2018.
- [227] A. Baldominos, Y. Saez, and P. Isasi, “Evolutionary Design of Convolutional Neural Networks for Human Activity Recognition in Sensor-Rich Environments,” *Sensors*, vol. 18, p. 1288, 2018.
- [228] A. Martín, R. Lara-Cabrera, F. Fuentes-Hurtado, V. Naranjo, and D. Camacho, “EvoDeep: a New Evolutionary Approach for Automatic Deep Neural Networks Parametrisation,” *J. Parallel Distrib. Comput.*, vol. 117, pp. 180–191, 2018.
- [229] H. Tian, S. Pouyanfar, J. Chen, S.-C. Chen, and S. S. Iyengar, “Automatic Convolutional Neural Network Selection for Image Classification Using Genetic Algorithms,” in *The 19th IEEE International Conference on Information Reuse and Integration for Data Science (IEEE IRI 2018)*, 2018, pp. 444–451.
- [230] K. Deb, S. Pratab, S. Agarwal, and T. Meyarivan, “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II,” *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.

- [231] S. S. Mostafa, N. Horta, A. G. Ravelo-García, and F. Morgado-Dias, “Analog Active Filter Design using a Multi Objective Genetic Algorithm,” *AEU - Int. J. Electron. Commun.*, vol. 93, pp. 83–94, Sep. 2018.
- [232] M. Li, S. Liu, L. Zhang, H. Wang, F. Meng, and L. Bai, “Non-Dominated Sorting Genetic Algorithms-Ibased on Multi-Objective Optimization Model in the Water Distribution System,” *Procedia Eng.*, vol. 37, pp. 309–313, Jan. 2012.
- [233] A. Seshadri, “Multi-Objective Optimization using Evolutionary Algorithms,” *MathWorks*, 2009. [Online]. Available: <https://www.mathworks.com/matlabcentral/fileexchange/10351-multi-objective-optimizaion-using-evolutionary-algorithm>. [Accessed: 20-Sep-2017].
- [234] S. M. R. Rafiei, M. H. Kordi, G. Griva, and H. Yassami, “Multi-Objective Optimization based Optimal Compensation Strategies Study for Power Quality Enhancement Under Distorted Voltages,” in *2010 IEEE International Symposium on Industrial Electronics*, 2010, pp. 3284–3291.
- [235] D. P. Kingma and J. L. Ba, “ADAM: a Method for Stochastic Optimization,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015, pp. 1–15.
- [236] K. Deb and R. B. Agrawal, “Simulated Binary Crossover for Continuous Search Space,” *Complex Syst.*, vol. 9, pp. 115–148, 1995.
- [237] H. Liu, K. Simonyan, O. Vinyals, C. Fernando, and K. Kavukcuoglu, “Hierarchical Representations for Efficient Architecture Search,” Nov. 2017.
- [238] T. Cetto, J. Byrne, X. Xu, and D. Moloney, “Size/Accuracy Trade-Off in Convolutional Neural Networks: An Evolutionary Approach,” in *Oneto L., Navarin N., Sperduti A., Anguita D. (eds) Recent Advances in Big Data and Deep Learning. INNSBDDL 2019. Proceedings of the International Neural Networks Society*, Springer, Cham, 2020, pp. 17–26.
- [239] E. Real, A. Aggarwal, Y. Huang, and Q. V Le, “Regularized Evolution for Image Classifier Architecture Search,” Feb. 2018.
- [240] J. Snoek *et al.*, “Scalable Bayesian Optimization Using Deep Neural Networks,” Feb. 2015.
- [241] C. Liu *et al.*, “Progressive Neural Architecture Search,” Dec. 2017.
- [242] R. Negrinho and G. Gordon, “DeepArchitect: Automatically Designing and Training Deep Architectures,” Apr. 2017.
- [243] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *ICML’15 Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, 2015, pp. 448–456.
- [244] J. Hua and X. Gong, “A Normalized Convolutional Neural Network for Guided Sparse Depth Upsampling,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2018.
- [245] J. Nagi and F. Ducatelle, “Max-Pooling Convolutional Neural Networks for Vision-Based Hand Gesture Recognition,” *2011 IEEE Int. Conf. Signal Image Process. Appl.*, pp. 342–347,

- 2011.
- [246] F. Vaquerizo-Villar *et al.*, “Convolutional Neural Networks to Detect Pediatric Apnea-Hypopnea Events from Oximetry,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, pp. 3555–3558.
- [247] R. B. Berry *et al.*, “Rules for Scoring Respiratory Events in Sleep: Update of The 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of The Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine,” *J. Clin. Sleep Med.*, vol. 8, no. 5, pp. 597–619, Oct. 2012.
- [248] G. Lu, F. Yang, J. A. Taylor, and J. F. Stein, “A Comparison of Photoplethysmography and ECG Recording to Analyse Heart Rate Variability in Healthy Subjects,” *J. Med. Eng. Technol.*, vol. 33, no. 8, pp. 634–641, 2009.
- [249] “Top 50 Countries/Markets by Smartphone Users and Penetration,” *Newzoo*. [Online]. Available: <https://newzoo.com/insights/rankings/top-50-countries-by-smartphone-penetration-and-users/>. [Accessed: 20-Aug-2019].
- [250] “325,000 mobile health apps available in 2017 – Android now the leading mHealth platform,” *research2guidance*. [Online]. Available: <https://research2guidance.com/325000-mobile-health-apps-available-in-2017/>. [Accessed: 20-Aug-2019].
- [251] A. A. Ong and M. B. Gillespie, “Overview of Smartphone Applications for Sleep Analysis,” *World J. Otorhinolaryngol. Neck Surg.*, vol. 2, no. 1, pp. 45–49, 2016.
- [252] V. P. Cornet and R. J. Holden, “Systematic Review of Smartphone-Based Passive Sensing for Health and Wellbeing,” *J. Biomed. Inform.*, vol. 77, no. July 2017, pp. 120–132, Jan. 2018.
- [253] B. L. Koley and D. Dey, “On-Line Detection of Apnea/Hypopnea Events using Spo2 Signal: a Rule-Based Approach Employing Binary Classifier Models,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 1, pp. 231–239, 2014.
- [254] F. Mendonça, S. S. Mostafa, F. Morgado-Dias, J. L. Navarro-Mesa, G. Juliá-Serdá, and A. G. Ravelo-García, “A Portable Wireless Device Based on Oximetry for Sleep Apnea Detection,” *Computing*, vol. 100, no. 11, pp. 1203–1219, 2018.
- [255] N. Oliver and F. Flores-Mangas, “HealthGear: Automatic Sleep Apnea Detection and Monitoring with a Mobile Phone,” *J. Commun.*, vol. 2, no. 2, pp. 1–9, 2007.
- [256] F. Chung, H. R. Abdullah, and P. Liao, “STOP-Bang Questionnaire a Practical Approach to Screen for Obstructive Sleep Apnea,” *Chest*, vol. 149, no. 3, pp. 631–638, 2016.