Automatic Detection and Characterization of Obstructive Sleep Apnea Using Computer Vision

Hanif, Umaer Rashid

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Hanif, U. R. (2022). *Automatic Detection and Characterization of Obstructive Sleep Apnea Using Computer Vision*. DTU Health Technology.

DTU Health Tech
Department of Health Technology

Technical University of Denmark

Automatic Detection and Characterization of Obstructive Sleep Apnea Using Computer Vision

Umaer Hanif

PhD Dissertation

2022

Region H Rigshospitalet

Stanford School of Medicine

Automatic Detection and Characterization of Obstructive Sleep Apnea Using Computer Vision

Author
Umaer Rashid Hanif
Ph.D. Student, M.Sc.Eng.
Department of Health Technology
Technical University of Denmark

Main Supervisor
Helge B.D. Sørensen
Associate Professor MSK, Group Leader, Ph.D.
Department of Health Technology
Technical University of Denmark

Co-supervisor
Poul Jennum
Professor, Chief Physician, Ph.D.
Department of Clinical Neurophysiology
Copenhagen University Hospital

Co-supervisor
Emmanuel Mignot
Professor, Division Chief, Ph.D.
Stanford Center for Sleep Sciences and Medicine
Stanford University

Project Period:	March 1st, 2019 - May 31st, 2022
Degree:	Doctor of Philosophy
University:	Technical University of Denmark
Field:	Health Technology
Edition:	1. edition
Class:	Public
Copyrights:	©Umaer Hanif, 2022

umaer hanif
Umaer Hanif

30/05/22
Date

Abstract

Background: Obstructive sleep apnea (OSA) is characterized by recurrent upper airway collapse during sleep and affects up to one billion people worldwide. OSA is associated with increased risk of cardiovascular diseases, stroke, and all-cause mortality. Diagnosis and treatment of OSA are crucial for long term health and a reduced economic burden. However, the gold-standard polysomnography (PSG) is time-consuming, expensive, and requires a great amount of manual labor.

Objective: The objective of this PhD project was to invent fast, cheap, and data-driven screening and scoring systems for OSA based on imaging data. Imaging data can be captured much faster than overnight data collection for detection and characterization of OSA using computer vision.

Methods: Two systems were proposed: 1) an automatic screening system which utilizes 3D craniofacial scans to estimate apnea-hypopnea index (AHI), which measures OSA severity, and 2) an automatic scoring system which utilizes drug-induced sleep endoscopy (DISE) videos to estimate sites of upper airway collapse and obstruction degrees in OSA patients. The main components in both systems were convolutional neural networks, which were trained and evaluated using two different datasets consisting of 1) 1366 3D craniofacial scans collected across 11 different sleep clinics, and 2) 281 DISE videos collected across two sleep clinics and scored by three different surgeons.

Results: For AHI estimation based on 3D craniofacial scans, a mean absolute error of 11.38 events/hour and a Pearson correlation of 0.4 were obtained. Subjects were classified as normal or with OSA with an accuracy of 67%, which was higher than using screening questionnaires. The model's performance was comparable to three sleep specialists, and its performance increased further by adding demographics and questionnaires as features. For automatic scoring of DISE in OSA, a mean F1 score of 70% was obtained across four upper airway sites (velum: 85%, oropharynx: 72%, tongue base: 57%, epiglottis: 65%) with respect to obstruction degrees (0, 1, or 2).

Conclusion: The proposed automatic screening system for detection of OSA based on 3D craniofacial scans has the potential to fulfill the need for a fast and cheap screening method for OSA in clinical practice. The proposed automatic scoring system for DISE videos has the potential to provide consistent scoring without bias, which can aid surgeons in interpretation of DISE and result in improved treatment outcomes for OSA patients.

Resumé

Baggrund: Obstruktiv søvnapnø (OSA) er karakteriseret ved gentagende kollaps i den øvre luftvej under søvn og påvirker op til en milliard mennesker globalt. OSA er forbundet med øget risiko for kardiovaskulære sygdomme, blodpropper og dødelighed af alle årsager. Diagnose og behandling af OSA er afgørende for langtidshelbreddet og en reduceret økonomisk byrde. Guldstandarden, som er polysomnography (PSG), er tidskrævende, dyr og kræver en stor mængde manuelt arbejde.

Formål: Formålet med dette PhD projekt var at opfinde hurtige, billige og datadrevne screenings- og scoringssystemer til OSA baseret på billeddata. Billeddata kan opsamles meget hurtigere end dataopsamling under søvn til detektion og karakterisering af OSA ved hjælp af computer vision.

Metoder: To systemer blev udviklet: 1) et automatisk screeningssystem, der bruger 3D kraniofaciale skanninger til at estimere apnø-hypopnø-indekset (AHI), som måler graden af OSA, og 2) et automatisk scoringssystem, der bruger søvnendoskopi (DISE) videoer til at estimere lokationer af øvre luftvejskollaps og obstruktionsgrader i OSA patienter. Hovedkomponenterne i begge systemer var convolutional neural networks, som blev trænet og evalueret ved at bruge to forskellige datasæt bestående af 1) 1366 3D kraniofaciale skanninger samlet på tværs af 11 forskellige søvnklinikker, og 2) 281 DISE videoer samlet på tværs af to søvnklinikker og scoret af tre forskellige kirurger.

Resultater: AHI estimering baseret på 3D kraniofaciale skanninger gav en mean absolute error på 11.38 events/time og en Pearson korrelation på 0.4. Subjekter blev klassificeret som normale eller OSA med en accuracy på 67%, hvilket var højere end at bruge screeningsspørgeskemaer. Modellens performance var sammenlignelig med tre søvnspecialister og dens performance blev bedre ved at tilføje demografiske data og spørgeskemaer som features. Auomatisk scoring af DISE i OSA gav en gennemsnitlig F1 score på 70% opnået på tværs af de fire øvre luftvejslokationer (velum: 85%, oropharynx: 72%, tunge: 57%, epiglottis: 65%) med hensyn til obstruktionsgrader (0, 1 eller 2).

Konklusion: Det udviklede automatiske screeningssystem til detektion af OSA baseret på 3D kraniofacialle skanninger har potentialet til at opfylde behovet om en hurtig og billig screeningsmetode til OSA i klinisk praksis. Det udviklede automatiske scoringssystem til DISE videoer har potentialet til at levere konsistente scoringer uden bias, som kan hjælpe kirurgerne med fortolkningen af DISE og medvirke i forbedrede behandlingsresultater for OSA patienter.

Preface

This PhD dissertation was prepared at the Section for Digital Health, Department of Health Technology at the Technical University of Denmark (DTU) in fulfilment of the requirements for acquiring a degree of Doctor of Philosophy. The research and results conveyed in this dissertation have been accomplished equally at the Department of Health Technology at DTU, the Stanford Center for Sleep Sciences and Medicine at Stanford University, and the Danish Center for Sleep Medicine at Copenhagen University Hospital. The entire work was conducted within the period March 1st, 2019 until May 31st, 2022 during which four research papers have been prepared.

The dissertation deals with advanced image analysis and deep learning techniques to meet the objectives of the research questions. Although it is not a requirement, a reader with a background similar to the author in signal processing and machine learning will be able to fully comprehend and appreciate the content of this dissertation. Furthermore, a person with a technical background will be able to quickly acquire any missing knowledge about sleep by consulting the clinical background chapter. As a general rule, I always strive to prepare a thesis that is as short and concise as possible, and while this is still a priority, a more detailed dissertation has been developed this time around, allowing for a more pleasant and comprehensive reading experience.

Above all else, the emphasis has been on presenting clear and reproducible research. Consequently, every applied method in this project is described in great detail using text, figures, and tables that support and facilitate the apprehension of each methodology. Furthermore, all the software which has been implemented in Python has been commented properly, which should allow another researcher to grasp and reproduce/continue the work if they desire to do so and the code will be readily available if they request it.

On a personal note, ever since I was in elementary school, I thoroughly enjoyed the art of writing. This passion has continued to grow as I made my way through high school and university. I realized with a heavy heart that this could be my last big report, and I decided to go out with a bang. With this dissertation, I have striven to produce my greatest, most well-written and thorough report to date. I sincerely hope that this is the impression I leave the reader with as well.

Acknowledgements

First and foremost, above all else, I must thank my main supervisor Helge B. D. Sørensen for providing me with this opportunity and having complete confidence in me to achieve a doctorate degree and be a good ambassador for the DTU/Stanford/Copenhagen University Hospital collaboration. At the time of writing, 8 years have passed since I wrote my bachelor's project with him as a supervisor back in 2014. I have come a long way since then on both a professional and personal level, and my lengthy collaboration with Helge has played an important part in me becoming the researcher I am today. I thank Helge, not only for supervising me in this PhD project, but also for supervising me through various other projects throughout my time as a biomedical engineering student and for always believing in me. I hope that I have made him proud with this work.

Next, I wholeheartedly want to thank my co-supervisor Emmanuel Mignot for his guidance throughout the entire time we have worked together. I went to Emmanuel's lab at Stanford University in 2017 to write my Master's thesis, and I immediately knew that I wanted to return and work with him again. Three years later, I was lucky enough to realize that wish through my external stay at Stanford University for 18 months. Emmanuel has one of the greatest minds I've ever known and he has been an extraordinary mentor to me. He never plays it safe and dares to be innovative, while his passion for research and previous achievements have motivated me to aspire for greatness on several occasions. Despite his brilliance, Emmanuel sometimes has a childish sense of humor and never takes himself too seriously, which is very refreshing in the world of academia. I am happy to not only call Emmanuel a supervisor, but also a friend.

I am also tremendously grateful for the help and supervision provided by my co-supervisor Poul Jennum, who, like Helge, first supervised me in my bachelor's project back in 2014. Poul has played a big part in my fascination for sleep science as he has an infectious admiration and passion for the field and a praiseworthy attitude towards helping people with sleep disorders. Poul is a well-recognized figure in the sleep community and has contributed greatly to our understanding of sleep physiology and sleep disorders. He is a true master in the field and he has been a very inspiring figure to me on a professional level.

At DTU, I would like to thank Rasmus R. Paulsen for his valuable input and help, particularly for my paper on 3D craniofacial scans. I am also grateful to my colleagues Matteo Cesari, Alexander N. Zahid, Andreas Brink-Kjær, Mads Olsen, and Søren Møller Rasmussen who all impressed me as great researchers and helped create a fun and social work environment at DTU. At Stanford University, I would like to thank Robson Capasso and Stanley Y. Liu for providing data and input for my paper on drug-induced sleep endoscopy videos. Furthermore, I am grateful to Stephanie Lettieri, Eileen B. Leary, Logan D. Schneider, Hyatt Moore, Valerie Garcia, and Polina Davidenko, all of whom have facilitated my research and made my stay at Stanford University easier.

On a personal level, I firstly want to thank my family for their never-ending love and support for whatever I pursue in life. Sending a son away to a foreign country for a long period of time is not a fun endeavor for any parent, but they have always backed me to succeed and told me to do whatever it takes to fulfill my ambitions. I would also like to thank Andreas Brink-Kjær, Villads H. Jørgensen, Asbjørn W. Helge, and Magnus Ruud Kjær, all of whom I have shared apartment with during different periods of my stay in San Francisco. I also want to thank all my friends, especially in Denmark, since you all give me the energy and motivation to do what I do on a daily basis.

Next, I want to thank the love of my life, Wania Saif. She has been the one to give me perspective when everything seemed pointless and the one to ground me when I got carried away. She has shared with me the inevitable ups and downs of a project like this and there is no one I would rather have had with me on this ride. Thank you for supporting me and believing in me.

Finally, I want to thank my grandmother, who has played a big part in raising me and shaping me into the person I am today. She has always given me unconditional love and support and put me and the rest of the family before herself. Unfortunately, she has been very sick in the past few months, particularly towards the end of my PhD project, during which I have been preparing this dissertation. Seeing her in this condition has been hard, but her undying faith in my success has been a true motivation for me to finish this dissertation and conclude my project in the best possible manner. I dedicate this dissertation to her.

Contents

Abstract ii

Resume iii

Preface iv

Acknowledgements vi

Abbreviations xiii

List of Figures xv

List of Tables xvi

1 Introduction 1

1.1 Research Background 1

1.2 Problem Statement 4

1.3 Dissertation Objective 5

1.4 Dissertation Hypotheses 5

1.5 Dissertation Outline 6

1.6 Research Contributions 7

1.6.1 First-Author Journal Papers 7

1.6.2 First-Author Conference Papers 7

1.6.3 Co-Authored Papers 8

1.6.4 Popular Science Articles 8

2 Clinical Background 9

2.1 Importance of Sleep 9

2.2 Polysomnography 10

2.2.1 Sensors and Signals 10

2.2.2 Scoring of Events 10
2.2.2.1 Sleep Stages and Arousals 11
2.2.2.2 Sleep-Disordered Breathing 13
2.3 Obstructive Sleep Apnea 13
2.3.1 Causes of Obstructive Sleep Apnea 14
2.3.2 Implications of Obstructive Sleep Apnea 14
2.3.3 Scoring of Obstructive Sleep Apnea Events 15
2.3.4 Apnea-Hypopnea Index 15
2.3.5 Screening Questionnaires for Obstructive Sleep Apnea 15
2.3.6 Continuous Positive Airway Pressure 17
3 Detection of Obstructive Sleep Apnea From 3D Craniofacial Scans 18
3.1 Research Background 19
3.2 Research Questions and Objectives 20
3.3 Paper I: Prediction of Patient Demographics Using 3D Craniofacial Scans and Multi-View CNNs 22
3.3.1 Methods 22
3.3.1.1 Data Collection 22
3.3.1.2 Data Description 26
3.3.1.3 Pre-Processing 29
3.3.1.4 Multi-View Convolutional Neural Networks 30
3.3.1.5 Training, Validation, and Testing 31
3.3.1.6 Performance Measures 33
3.3.2 Results and Discussion 34
3.3.2.1 Estimation of Sex 34
3.3.2.2 Estimation of Age 34
3.3.2.3 Estimation of BMI 35
3.3.2.4 Comparison to Similar Work 36
3.3.3 New Analyses After Publication 37
3.3.3.1 Estimation of Sex 38
3.3.3.2 Estimation of Age 38
3.3.3.3 Estimation of BMI 40
3.4 Paper II: Estimation of Apnea-Hypopnea Index Using Deep Learning on 3D Craniofacial Scans 42
3.4.1 Methods 42
3.4.1.1 Data Collection 44
3.4.1.2 Data Description 44

3.4.1.3 Pre-Processing . . . 44
3.4.1.4 Multi-View Convolutional Neural Network . . . 46
3.4.1.5 Training, Validation, and Testing . . . 48
3.4.1.6 Performance Measures . . . 49
3.4.1.7 Sleep Specialists' Ability to Estimate Apnea-Hypopnea Index . . . 49
3.4.1.8 Topographic Display of Important Craniofacial Regions . . . 50
3.4.2 Results and Discussion . . . 50
3.4.2.1 Model Using Craniofacial Scans . . . 51
3.4.2.2 Model Using Craniofacial Scans, Demographics, and Questionnaires 54
3.4.2.3 Comparison to Similar Work . . . 56
3.4.2.4 Comparison to Sleep Medicine Specialists . . . 57
3.4.2.5 Topographic Display of Important Craniofacial Regions . . . 60
3.4.2.6 Limitations . . . 62
3.4.2.7 Future Work . . . 63
3.5 Conclusions . . . 63

4 Automatic Scoring of Drug-Induced Sleep Endoscopy 67
4.1 Research Background . . . 68
4.2 Research Questions and Objectives . . . 71
4.3 Paper III: Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks . . . 73
4.3.1 Methods . . . 73
4.3.1.1 Data Collection . . . 74
4.3.1.2 Data Description . . . 75
4.3.1.3 Pre-Processing . . . 77
4.3.1.4 Convolutional Recurrent Neural Network . . . 78
4.3.1.5 Training, Validation, and Testing . . . 80
4.3.1.6 Performance Measures . . . 81
4.3.2 Results and Discussion . . . 81
4.3.2.1 Overall Performance . . . 81
4.3.2.2 Misclassified Frames . . . 81
4.3.2.3 Performance for Videos . . . 84
4.3.2.4 Limitations . . . 84
4.3.3 New Analyses After Publication . . . 85
4.3.3.1 Overall Performance . . . 85
4.3.3.2 Performance for Videos . . . 85
4.3.3.3 Misclassified Frames . . . 86

4.4 Paper IV: Automatic Scoring of Drug-Induced Sleep Endoscopy for Obstructive Sleep Apnea Using Deep Learning . . . 88
4.4.1 Methods . . . 88
4.4.1.1 Data Collection . . . 90
4.4.1.2 Data Description . . . 90
4.4.1.3 Pre-Processing . . . 92
4.4.1.4 Convolutional Recurrent Neural Network . . . 94
4.4.1.5 Training, Validation, and Testing . . . 94
4.4.1.6 Post-Processing . . . 95
4.4.1.7 Performance Measures . . . 96
4.4.2 Results and Discussion . . . 96
4.4.2.1 Overall Performance . . . 96
4.4.2.2 Sensitivity for Each Obstruction Degree . . . 97
4.4.2.3 Performance for Videos From Each Surgeon . . . 99
4.4.2.4 Performance for 5-Second Clips . . . 99
4.4.2.5 Probability Interpretation . . . 100
4.4.2.6 Comparison to Inter-rater Reliability . . . 101
4.4.2.7 Limitations . . . 102
4.4.2.8 Future Work . . . 103
4.5 Conclusions . . . 103

5 Conclusions 106
5.1 Research Conclusions . . . 106
5.2 Limitations and Future Perspectives . . . 108
5.3 Concluding Remarks . . . 110

References 111

Appendix A STOP-Bang Questionnaire 129

Appendix B Paper I 131

Appendix C Paper II 136

Appendix D Paper III 147

Appendix E Paper IV 152

Abbreviations

AASM - American Academy of Sleep Medicine

ACC - Accuracy

Adam - Adaptive Moment Estimation

AHI - Apnea-Hypopnea Index

AP - Average Pooling

A-P - Antero-Posterior

AUC ROC - Area Under the Receiver Operating Characteristics Curve

Bi-LSTM - Bidirectional Long Short-Term Memory

BMI - Body Mass Index

CNN - Convolutional Neural Network

CPAP - Continuous Positive Airway Pressure

CSA - Central Sleep Apnea

CUH - Copenhagen University Hospital

DCSM - Danish Center for Sleep Medicine

Deep-MVLM - Multi-View Consensus CNN for 3D Facial Landmark Placement

DISE - Drug-Induced Sleep Endoscopy

DTU - Technical University of Denmark

E - Epiglottis

ECG - Electrocardiography

EEG - Electroencephalography

EMG - Electromyography

EOG - Electrooculography

ESS - Epworth Sleepiness Scale

FC - Fully Connected

FN - False Negatives

FP - False Positives

HST - At-Home Sleep Apnea Test

LSTM - Long Short-Term Memory

MAE - Mean Absolute Error

MLP - Multi-Layer Perceptron

MP - Max Pooling

MSE - Mean Squared Error

NREM - Non-Rapid Eye Movement

NSRR - National Sleep Research Resource

O - Oropharynx

ODI - Oxygen Desaturation Index

OSA - Obstructive Sleep Apnea

OTE - Oropharynx, Tongue Base, and Epiglottis Combined

PCC - Pearson Correlation Coefficient

PSG - Polysomnography

ReLU - Rectified Linear Unit

REM - Rapid Eye Movement

RGB - Red Green Blue

RIP - Respiratory Inductance Plethysmography

RNN - Recurrent Neural Network

STAGES - Stanford Technology Analytics and Genomics in Sleep

SUH - Stanford University Hospital

SVM - Support Vector Machine

SVR - Support Vector Regression

T - Tongue base

TN - True Negatives

TORS - Transoral Robotic Surgery

TP - True Positives

V - Velum

VOTE - Velum, Oropharynx, Tongue base, Epiglottis

List of Figures

1.1 Diagnostic accuracy for obstructive sleep apnea vs. time and cost for various methods 3

2.1 A patient undergoing diagnostic polysomnography 11
2.2 30 seconds of a polysomnography recording 12
2.3 A hypnogram 13
2.4 A person experiencing upper airway collapse during sleep 14
2.5 Apneas and hypopneas in a polysomnography recording 16
2.6 Different continuous positive airway pressure masks 17

3.1 A Structure Sensor from Occipital Inc attached to an Ipad Pro 23
3.2 The procedure of capturing a 3D craniofacial scan 24
3.3 A completed 3D surface scan displayed on an iPad Pro 25
3.4 Illustration of vertices, edges, and triangles 26
3.5 A 3D mesh scan and associated triangles displayed from three angles 27
3.6 Distribution of sex, age, and BMI for subjects in the dataset 28
3.7 Detected landmarks on a 3D craniofacial scan 29
3.8 Eight pairs of 2D images and depth maps derived from a 3D craniofacial scan 30
3.9 Network architecture for estimating patient demographics from 3D scans 32
3.10 Confusion matrix for estimating sex on the test set 34
3.11 Bland-Altman plot for estimating age on the test set 35
3.12 Bland-Altman plot for estimating BMI on the test set 36
3.13 Confusion matrix for estimating BMI categories on the test set 37
3.14 Confusion matrix for estimating sex on the entire dataset 38
3.15 Bland-Altman plot for estimating age on the entire dataset 39
3.16 Bland-Altman plot for estimating age on the entire dataset compared to sleep specialists 40
3.17 Bland-Altman plot for estimating BMI on the entire dataset 41
3.18 Confusion matrix for estimating BMI categories on the entire dataset 41
3.19 Block diagram for estimating apnea-hypopnea index from 3D craniofacial scans . . . 43

3.20 Distribution of apnea-hypopnea index for subjects in the dataset 45
3.21 Five pairs of 2D images and depth maps derived from a 3D craniofacial scan 46
3.22 Network architecture for estimating apnea-hypopnea index from 3D scans 47
3.23 Network architecture for estimating apnea-hypopnea index from clinical variables . . 48
3.24 Comparison of Bland-Altman plots between model using 3D scans and clinical variables 52
3.25 Comparison of confusion matrices between model using 3D scans and clinical variables 53
3.26 Scatter plots of apnea-hypopnea index for misclassified normal subjects 54
3.27 Scatter plots of apnea-hypopnea index for misclassified obstructive sleep apnea subjects 55
3.28 Comparison of Bland-Altman plots between proposed model and sleep specialists . . 59
3.29 Comparison of confusion matrices between proposed model and sleep specialists . . . 60
3.30 Topographic display of important craniofacial features 61

4.1 Upper airway sites of collapse in drug-induced sleep endoscopy 69
4.2 Patterns of collapse and obstruction degrees for the velum 70
4.3 Patterns of collapse and obstruction degrees for the oropharynx, tongue, and epiglottis 71
4.4 A drug-induced sleep endoscopy examination being performed 74
4.5 An endoscope used for drug-induced sleep endoscopy 75
4.6 Each class used for upper airway region classification 76
4.7 Network architecture for classifying upper airway regions 79
4.8 Confusion matrix for classifying upper airway regions on the test set 82
4.9 Wrongly classified frames with respect to upper airway regions 83
4.10 Performance distribution for classifying upper airway regions on the entire dataset . 86
4.11 Confusion matrix for classifying upper airway regions on the entire dataset 87
4.12 Block diagram for estimating obstruction degrees for each upper airway site 89
4.13 Distribution of drug-induced sleep endoscopy video durations in the dataset 90
4.14 Distribution of obstruction degrees per site for drug-induced sleep endoscopy videos 91
4.15 Distribution of obstruction degrees per site for 5-second clips 93
4.16 Confusion matrices for estimating obstruction degrees for each site 98
4.17 Performance distribution for estimating obstruction degrees in each video 100
4.18 Predicted probabilities for obstruction degrees for each site 101

List of Tables

3.1 Comparison between models using different inputs to predict apnea-hypopnea index 57
3.2 Comparison between proposed models and performance achieved in the literature . . 58
3.3 Comparison between proposed model and three sleep medicine specialists 59

4.1 The VOTE classification system 69
4.2 Labels used for upper airway region classification 77
4.3 Distribution of labels for upper airway region classification 77
4.4 Performance for classifying upper airway regions on the test set 84
4.5 Performance for classifying upper airway regions on the test set using cross-validation 86
4.6 Annotations provided by surgeons for drug-induced sleep endoscopy videos 91
4.7 Labels for obstruction degrees created for 5-second clips 92
4.8 Performance for estimating obstruction degrees with respect to three different surgeons 99
4.9 Comparison between proposed model and inter-rater reliability between surgeons . . 102

Chapter 1

Introduction

1.1 Research Background

Obstructive sleep apnea (OSA) is characterized by recurrent upper airway narrowing or collapse during sleep, causing reduced airflow or cessation of breathing and frequently displaying symptoms such as heavy snoring, sudden awakenings, and gasping/choking sounds [1]. OSA patients suffer from daytime sleepiness and have increased risk of cardiovascular diseases, hypertension, stroke, type 2 diabetes, and all-cause mortality [2–9]. A recent study estimated that almost half a billion people aged 30-69 have moderate to severe OSA globally [10] and evidence suggests that OSA is a severely underdiagnosed disorder [11–13]. Furthermore, the American Academy of Sleep Medicine (AASM) estimated in 2015 that the economic burden of undiagnosed OSA was $150 billion in the United States, while cost of diagnosis and treatment amounted to only a third of that, emphasizing that diagnosis and treatment of OSA can lead to significant reductions in health care costs [14].

OSA is diagnosed by undergoing a diagnostic sleep test called a polysomnography (PSG) or an at-home sleep apnea test (HST) during which several physiological signals are recorded from the patient to monitor their breathing pattern during sleep [15]. Subsequently, the signals are analyzed manually by a sleep technician, who annotates apneas (cessation of breathing) and hypopneas (reduced airflow) observed in the recording according to criteria defined by the AASM [16]. Diagnosis of OSA is based on the apnea-hypopnea index (AHI), which summarizes the number of apneas and hypopneas per hour of sleep [17, 18]. An AHI of 15 events/hour or greater is the clinical criterion used for defining the presence of OSA (moderate-severe versus no or mild OSA) [17]. PSGs are comprehensive tests, which contain extensive amounts of information about sleep patterns and sleep disorders [15]. However, PSGs are also costly and cumbersome procedures that require a lot of time and resources before, during, and after the procedure [19]. Furthermore, the large number of signals and long recording times put a strenuous workload on sleep technicians who analyze the data, which may explain part of the reason for low inter-rater reliability between technicians who score respiratory events [20–24]. HSTs reduce user burden and resource requirements by allowing overnight monitoring at home using fewer modalities but have been shown to be less accurate than PSGs [25–27]. OSA-related questionnaires, such as the Berlin and STOP-Bang questionnaires, can be used as fast and cheap screening options, but diagnostic accuracy varies a lot and can be extremely poor for some categories of OSA severity [28, 29]. The main disadvantage of using questionnaires in clinical practice is that the screening is based almost solely on subjective information provided by the patient and no objective data as obtained from physiological sensors.

Since PSGs require a lot of resources and are extremely time-consuming, commercial efforts go into developing wearable technology that can detect OSA based on fewer sensors [30, 31], while many research efforts go into developing state-of-the-art machine learning-based algorithms for automatic detection of apneas and hypopneas from PSG signals [32–36]. Such efforts generally compromise on the diagnostic accuracy but lead to faster results by eliminating the need for in-lab visits and/or manual scoring [30, 37]. Although wearables are unable to provide rich and detailed information about sleep behavior in the way PSGs do, they can capture basic sleep metrics for the consumer on a nightly basis without causing any discomfort. Additionally, event detection algorithms eliminate the need for manual analysis of signals by technicians, saving both time and money but at the expense of diagnostic accuracy compared to PSGs and HSTs. The combination of sleep wearables with automatic apnea detection algorithms provide an enormous potential for fast and efficient screening of OSA, which could significantly reduce the current number of undiagnosed subjects suffering from the disorder. However, wearables and apnea detection algorithms still require a subject to sleep an entire night, perhaps even several nights, to continuously collect sufficient and reliable data before analyses can be performed. This consideration highlights the need for a fast and cheap screening system for OSA without the need for overnight data collection but with higher accuracy than screening questionnaires.

In that context, it has been demonstrated that craniofacial features related to the midface, jaw, and neck are indicative of the presence of OSA [38]. Therefore, craniofacial imaging has massive potential as a screening option for OSA. However, research is still at a premature stage and previous studies have only been conducted on a small scale [39–46]. Two major limitations of applications using craniofacial imaging for screening of OSA are the dependence on 1) manually derived measurements from the face and neck, which introduces a burden on clinicians/staff and increases time to diagnosis, and 2) hand-crafted features used to train machine learning models, which introduce human biases. Figure 1.1 shows a trade-off plot between diagnostic accuracy of OSA versus time and cost of using the methods discussed above, while also emphasizing that one of the dissertation objectives is to reduce time and cost of using craniofacial imaging for OSA screening while increasing accuracy in a data-driven manner.

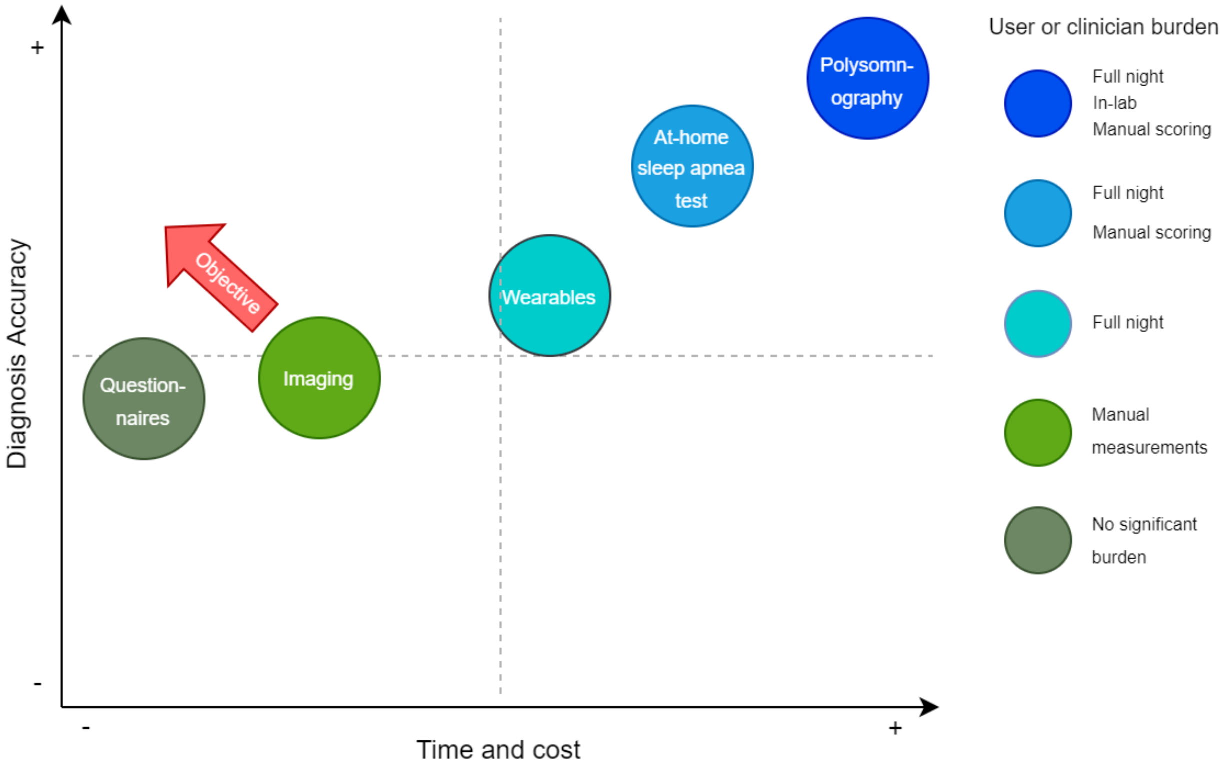


Figure 1.1: Trade-off between diagnostic accuracy of obstructive sleep apnea (OSA) versus time and cost of using different methods. Polysomnography (PSG) is the gold-standard for diagnosis of OSA. It is the most accurate while also being the most time-consuming and costly procedure as it is performed throughout an entire night and requires setting up equipment before the procedure, presence of staff during the procedure, and manual analysis of data after the procedure [15]. At-home sleep apnea test (HST) is less accurate than PSG in diagnosing OSA [25] but it is also less costly because it is performed at home and does not require presence of staff during the procedure [26]. However, HST is time-consuming as it requires data collection during an entire night and manual analysis of data subsequently [27]. Wearables generally yield medium level of diagnostic accuracy [30, 37] while still requiring a full night of sleep from the user. However, they are cheap because they only present a one-time cost and do not require anyone to be present during data collection or to analyze the data due to in-built algorithms capable of detecting events automatically [47, 48]. OSA-related questionnaires, which are used for screening, provide the lowest diagnostic accuracy [28, 29] but they are also the least time-consuming and costly because they can be completed in a few minutes and do not require presence of a specialist [49]. Craniofacial imaging is less accurate for OSA diagnosis than PSG, HST, and wearables [40, 45, 46] because it does not provide nocturnal respiratory information, but it is cheap and much less time-consuming because it does not require the subject to sleep. However, craniofacial imaging for OSA detection typically involves deriving measurements from the face and neck manually [39–43], which increases the time it takes to obtain a diagnosis. The dissertation objective is to move the role of imaging in OSA diagnosis left and upwards in the graph while eliminating the associated manual labor by using a data-driven and unbiased approach.

Why and how frequently the upper airway collapses is due to multiple factors, such as narrow upper airway anatomy, a low arousal threshold, inability to recruit dilator muscles during inspiration, and poor central control of breathing [50]. A strong contributing factor to passive anatomy is obesity, which causes fat deposits around the upper airway that narrow the airway during sleep [51]. Other factors include swollen tonsils or the tongue falling backwards [52]. Continuous positive airway pressure (CPAP) is the gold-standard treatment for OSA and works by providing a constant level of pressure sufficient to keep the upper airway open [53]. Although CPAP is extremely effective, studies show that up to 50% of users give up the device within the first year of therapy for various reasons [54].

For some patients, surgical procedures such as removing excess tissue or advancing the jaw are viable options to prevent collapse or increase upper airway space [55–57]. Prior to surgery, drug-induced sleep endoscopy (DISE) is performed to examine the pattern of upper airway collapse using a fiberoptic endoscope under sedation, which is designed to simulate natural sleep [58]. During a DISE examination, an otolaryngology - head and neck surgeon evaluates where and how the upper airway collapses by determining the upper airway sites of collapse and the degrees of obstruction [59]. According to the VOTE classification system, which is the most used system for scoring DISE, the sites of upper airway collapse are the velum (V), oropharynx (O), tongue base (T), and epiglottis (E), while the obstruction degree for each site is classified as either 0 (no obstruction), 1 (partial obstruction), or 2 (complete obstruction) [59].

The analysis of such examinations is not straightforward due to 1) anatomical variation across subjects, 2) distortion in video quality caused by e.g. mucus or saliva on the camera lens, and 3) several upper airway sites collapsing simultaneously, making it difficult to keep the camera stationary and determine which sites are causing the collapse. Surgeons display poor to moderate inter-rater reliability when scoring DISE videos [60–65], which is clinically important because scoring of DISE determines the treatment strategy for OSA patients and a wrong conclusion may lead to wrong treatment. Requiring a second opinion from a fellow otolaryngology surgeon equals more time and cost until a treatment plan is determined. This consideration highlights the need for a fast and cheap scoring system capable of automatically identifying sites of upper airway collapse and obstruction degrees in OSA patients in an objective and data-driven manner.

1.2 Problem Statement

OSA is a severely underdiagnosed sleep disorder which places an enormous economic burden on society and poses serious long-term health risks for people suffering from the disorder. Diagnosis of OSA is expensive and time-consuming and sleep clinics have a critical need for a fast and cheap screening method that is more accurate than using questionnaires. After OSA diagnosis, the upper airway is examined using DISE if surgery is considered as treatment. These examinations can be very challenging to analyze and inter-rater reliability is low between surgeons scoring DISE. The analysis of DISE can be very subjective despite a well-established classification system and sleep clinics have a critical need for an automatic scoring system capable of identifying sites of upper airway collapse and obstruction degrees to aid surgeons with consistent scoring without bias.

1.3 Dissertation Objective

This PhD project takes a step away from the collection of physiological signals during sleep and focuses on inventing methods that rely on imaging data which can be captured much faster than overnight sleep data collection. The analysis of such imaging data will be conducted in an automatic and data-driven manner by training and evaluating computer vision models based on deep learning architectures. The overall objective of this dissertation is stated as follows:

Dissertation Objective

To invent automatic screening and scoring systems, based on dedicated computer vision models, that rely on imaging data to detect the presence of obstructive sleep apnea and characterize the upper airway collapse pattern in a fast, cheap, and data-driven manner.

1.4 Dissertation Hypotheses

The dissertation objective will be investigated through two main research hypotheses that explore the diagnostic potential of computer vision techniques applied on imaging data. There are two imaging modalities that will be explored in relation to detection and characterization of OSA in this dissertation: 3D craniofacial scans and DISE examination videos. For each imaging modality, a research hypothesis is formulated and stated below:

Dissertation Hypotheses

Hypothesis 1: An automatic screening system can be invented, based on dedicated computer vision models, which utilizes 3D craniofacial scans to estimate presence and severity of obstructive sleep apnea more accurately than current screening questionnaires in a fast, cheap, and data-driven manner.

Hypothesis 2: An automatic scoring system can be invented, based on dedicated computer vision models, which utilizes drug-induced sleep endoscopy examination videos to estimate sites of upper airway collapse and obstruction degrees in obstructive sleep apnea patients with a similar accuracy as otolaryngology - head and neck surgeons.

1.5 Dissertation Outline

The dissertation is composed of two separate research parts concerned with: (1) the estimation of AHI and classification of OSA based on 3D craniofacial scans, and (2) the estimation of upper airway collapse sites and obstruction degrees based on DISE videos. Additional chapters are dedicated to introduce the project, provide the necessary clinical background, and conclude the project, respectively. The dissertation is composed of five chapters and appendix, and each chapter's content is outlined below.

- **Chapter 1** - Contains an introduction to the research background, problem statement, overall objective, and hypotheses of this dissertation. The chapter also provides an outline of the dissertation's content and the papers published during the project period.

- **Chapter 2** - Contains the necessary clinical background to comprehend the purpose and clinical significance of this dissertation for the reader who is not familiar with basic concepts in sleep science and medicine. This includes the importance of sleep and more detailed descriptions of PSG and OSA.

- **Chapter 3** - Contains a research background, a description of the data and methods, and a presentation of results, discussion, and conclusion related to Hypothesis 1 from Section 1.4, which is concerned with inventing an automatic screening system for OSA based on 3D craniofacial scans and computer vision.

- **Chapter 4** - Contains a research background, a description of the data and methods, and a presentation of results, discussion, and conclusion related to Hypothesis 2 from Section 1.4, which is concerned with inventing an automatic scoring system for DISE videos based on computer vision.

- **Chapter 5** - Contains a conclusion of the dissertation as a whole. This includes relating the research findings from Chapters 3 and 4 to Hypotheses 1 and 2 from Section 1.4 and a discussion on limitations associated with the PhD project and future perspectives that could improve it.

- **Appendix** - Contains preprints and published versions of first-author journal papers and conference papers that have been published throughout the PhD project period. These papers are listed in the following section.

1.6 Research Contributions

The papers, which have been published, accepted, submitted, or prepared during the PhD project period are summarized below. They are split into four categories: first-author journal papers, first-author conference papers, co-authored papers, and popular science articles. The scientific content of this dissertation is centered around the first author journal papers and conference papers, and preprints or published versions of these papers are attached in the appendix.

1.6.1 First-Author Journal Papers

- **Umaer Hanif**, Eileen B. Leary, Logan D. Schneider, Rasmus R. Paulsen, Anne Marie Morse, Adam Blackman, Paula K. Schweitzer, Clete A. Kushida, Stanley Y. Liu, Poul Jennum, Helge B. D. Sorensen, and Emmanuel J. M. Mignot, "Estimation of Apnea-Hypopnea Index using Deep Learning on 3D Craniofacial Scans", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 11, pp. 4185-4194, 2021. DOI: 10.1109/JBHI.2021.3078127. (**Published**)

- **Umaer Hanif**, Eva K. Kiær, Robson Capasso, Stanley Y. Liu, Emmanuel J. M. Mignot, Helge B. D. Sorensen, and Poul Jennum, "Automatic Scoring of Drug-Induced Sleep Endoscopy in Obstructive Sleep Apnea Using Deep Learning", *JAMA Otolaryngology - Head and Neck Surgery*, 2022. (**Under Review**)

1.6.2 First-Author Conference Papers

- **Umaer Hanif**, Rasmus R. Paulsen, Eileen B. Leary, Emmanuel Mignot, Poul Jennum, and Helge B. D. Sorensen, "Prediction of Patient Demographics using 3D Craniofacial Scans and Multi-view CNNs", *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1950-1953, 2020.
  DOI: 10.1109/EMBC44109.2020.9176333. (**Published**)

- **Umaer Hanif**, Eric Kezirian, Eva Kirkegaard Kiær, Emmanuel Mignot, Helge B. D. Sorensen, and Poul Jennum, "Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks", *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3957-3960, 2021.
  DOI: 10.1109/EMBC46164.2021.9630098. (**Published**)

1.6.3 Co-Authored Papers

- Villads Hulgaard Joergensen, **Umaer Hanif**, Poul Jennum, Emmanuel Mignot, Asbjoern W. Helge, and Helge B. D. Sorensen, "Automatic Segmentation to Cluster Patterns of Breathing in Sleep Apnea", *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 164-168, 2021.
DOI: 10.1109/EMBC46164.2021.9629624. (**Published**)

- Magnus Ruud Kjær, Andreas Brink-Kjær, **Umaer Hanif**, Emmanuel Mignot, Poul Jennum, and Helge B. D. Sørensen, "Polysomnographic Plethysmography Excursions are Reduced in Obese Elderly Men", *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2396-2399, 2021.
DOI: 10.1109/EMBC46164.2021.9630145. (**Published**)

- Asbjørn Wulff Helge, **Umaer Hanif**, Villads Hulgaard Jørgensen, Poul Jennum, Emmanuel Mignot, and Helge B. D. Sorensen, "Detection of Cheyne-Stokes Breathing Using a Transformer-based Neural Network", *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022. **(Accepted)**

- Magnus Ruud Kjær, **Umaer Hanif**, Andreas Brink-Kjær, Poul Jennum, Helge B. D. Sorensen, and Emmanuel Mignot, "ABED: Automatic Apneic Breathing Event Detector", 2022.
(In Preparation)

- Alexander R. Johansen, **Umaer Hanif**, Anupama Sridhar, Kyu Hur, Helge B. D. Sorensen, Poul Jennum, Emmanuel Mignot, and Michael Snyder, "Wearipedia: A Database for Wearables Measuring Physical Activity, Sleep, and Continuous Glucose", 2022. **(In Preparation)**

1.6.4 Popular Science Articles

- **Umaer Hanif**, "Noninvasiv måling hjælper patienter med søvnapnø", *Medicoteknik*, no. 4, vol. 6, August 2019. (**Published**)

- Majken Lerche Møller, "Artificial intelligence can diagnose sleep apnoea",
www.healthtech.dtu.dk/english/news/nyhed?id=58226F9F-3F5C-4645-AEC3-0AC00075B128
(**Published**)

Chapter 2

Clinical Background

This chapter contains the necessary clinical background to comprehend the purpose and clinical significance of this dissertation for the reader who is not familiar with basic concepts in sleep science and medicine. Section 2.1 highlights the importance of sleep and what the consequences of inadequate sleep are. Section 2.2 provides a description of polysomnography and examples of its use. Section 2.3 describes obstructive sleep apnea, including characteristics, implications, and treatment.

2.1 Importance of Sleep

Sleep is an essential biological process for maintaining normal human function, health, and well-being, and it plays a critical part in restoring energy, removing toxins and waste products from the body, and consolidating memories [66–69]. The National Sleep Foundation recommends 7-9 hours of sleep per night for healthy adults while infants and teenagers require even greater amounts of sleep to grow and develop [70]. The American Academy of Sleep Medicine (AASM) recommends 7 hours or more of sleep to avoid health risks associated with inadequate sleep [71]. These are general guidelines, but an individual's need for sleep depends on factors such as their activity levels and overall health.

Results from the National Health Interview Survey for U.S. non-institutionalized adults aged 18-84 from 2004-2017 (N = 398,382) show that 33% of the U.S. population sleeps less than 6 hours and are thus sleep deprived [72]. Chronic lack of sleep has been linked to altered or decreased cognitive performance [73, 74], cardiovascular diseases [75–77], hypertension [78, 79], stroke [77, 80], altered metabolism [81, 82], and increased inflammation [83, 84], essentially affecting every vital function of the human body. A study examining the economic burden across five different countries (US, Canada, UK, Germany, and Japan) estimated a total annual loss of up to $680 billion due to insufficient sleep ($<$ 6 hours) [85], while another study estimated an economic cost of $45.21 billion for the 2016-2017 financial year in Australia [86]. The huge global impact that sleep deprivation has, both in terms of financial burden and its role in the development of major diseases, underlines the significance of individuals getting adequate sleep on a daily basis.

2.2 Polysomnography

Polysomnography (PSG) is a nocturnal, multi-parametric, diagnostic sleep test which can be used to derive information about sleep patterns and sleep disorders [15].

2.2.1 Sensors and Signals

During a PSG, multiple sensors, as depicted in Fig. 2.1, are utilized to monitor physiologically meaningful signals, all of which are outlined below and illustrated in Fig. 2.2. The sensors and their placement are described in more detail in the AASM manual [16].

- Brain activity measured by electroencephalography (EEG) according to the International 10-20 system [87]
- Eye movements measured by electrooculography (EOG) using two electrodes for each eye placed at the inner and outer canthus [88]
- Airflow measured by a nasal cannula in combination with an oral thermistor [89, 90]
- Submental muscle activity measured by electromyography (EMG) using three electrodes on the chin [91, 92]
- Snoring measured by a microphone, cannula, or piezoelectric sensor [93]
- Blood oxygen saturation measured by a finger pulse oximeter [94]
- Heart activity measured by electrocardiography (ECG) using a single modified Lead II torso electrode placement [95]
- Respiratory effort measured by respiratory inductance plethysmography (RIP) belts on the chest and abdomen [96]
- Leg movements measured by EMG using two electrodes on each leg [91]

2.2.2 Scoring of Events

After the procedure, a sleep technician analyzes the signals manually and annotates sleep-related events such as sleep stages, arousals, sleep-disordered breathing events, blood oxygen desaturations, and periodic limb movements according to rules defined in the AASM manual [16].

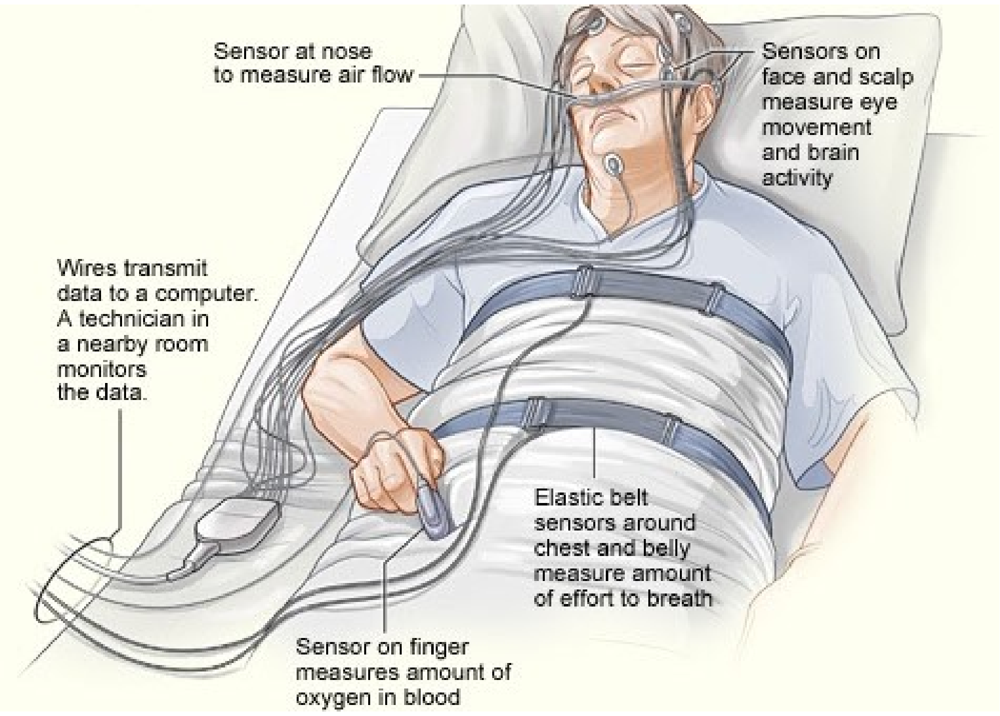


Figure 2.1: An illustration of a patient undergoing a diagnostic polysomnography. Sensors are attached to the patient, which measure physiological signals while he is sleeping to provide insight into sleep behavior and the presence of events related to sleep disorders, such as apneas and periodic limb movements. Source: [97]

2.2.2.1 Sleep Stages and Arousals

Sleep stages are scored in epochs of 30 seconds based on changes in the EEG, EOG, and chin EMG signals [16]. Sleep is categorized into rapid-eye movement (REM) sleep, also known as dream sleep, and non-REM (NREM) sleep, which is further divided into NREM1 (or just N1), NREM2 (N2), and NREM3 (N3) sleep stages, amounting to a total of four distinct sleep stages. NREM sleep and REM sleep alternates cyclically throughout the night and the amount of REM sleep increases per cycle as the night progresses [98]. Healthy subjects typically transition from wakefulness to N1 when they start falling asleep, followed by N2 and N3, and concluding the sleep cycle with REM sleep [99]. A typical sleep cycle is between 90-120 minutes, and it is common to have four to six sleep cycles during a night, where the distribution of sleep stages varies from cycle to cycle [98]. The graphical representation of sleep stages as a function of time is called a hypnogram and is depicted in Fig. 2.3.

An arousal is characterized by a transition from a deeper to a lighter sleep stage (e.g., N3 to N2) or from sleep to wakefulness and is scored using the same signals as for sleep stages (EEG, EOG, and chin EMG) [16]. Arousals can either be spontaneous or occur in association with events such as periodic limb movements or sleep disordered breathing events [100].

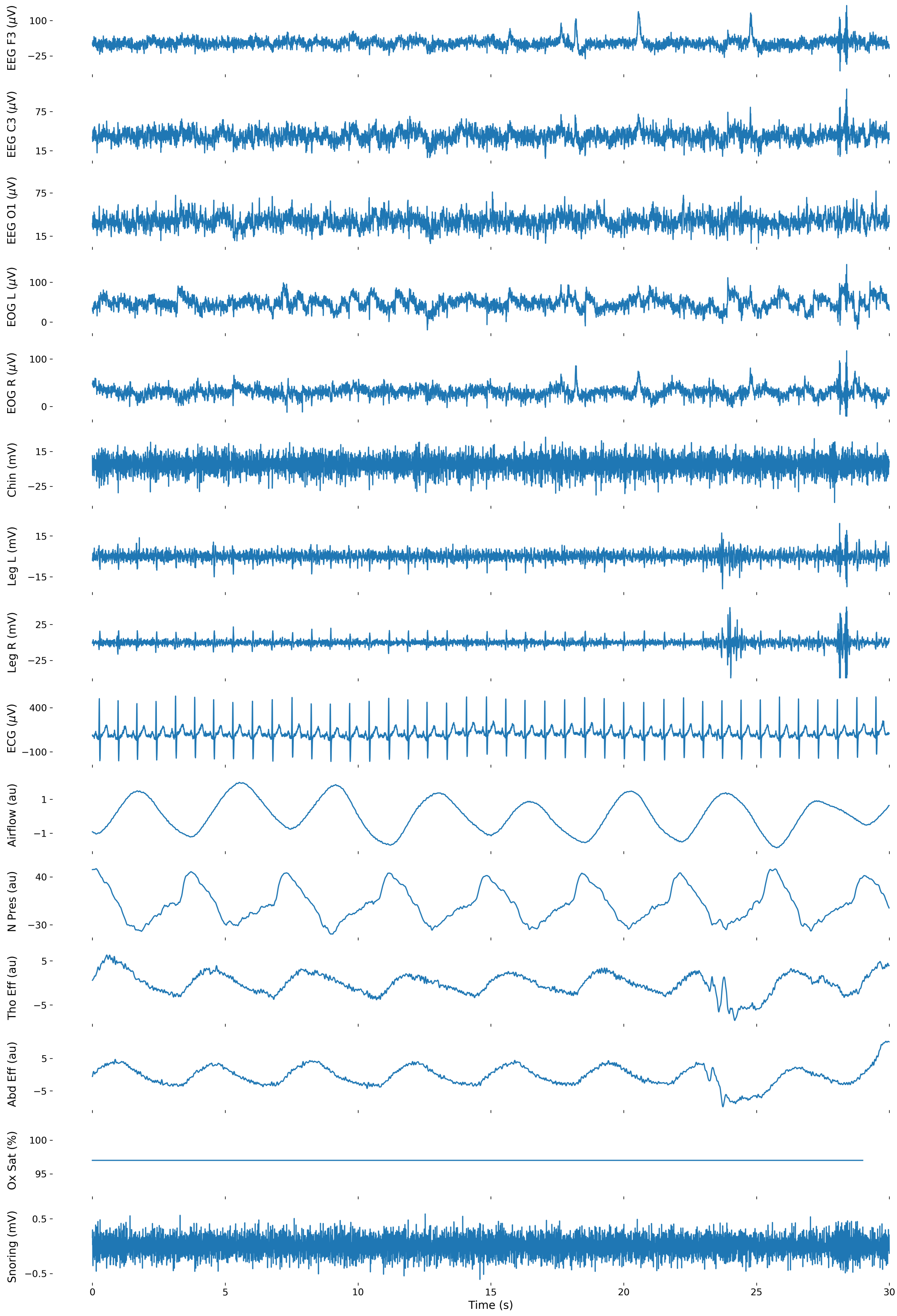


Figure 2.2: 30 seconds of a polysomnography recording. Four EEG channels and one ECG channel have been omitted to save space. F3, C3, and O1 are channels from the frontal, central, and occipital regions of the brain, respectively. L and R are abbreviations for left and right. N Pres - Nasal Pressure, Tho Eff - Thoracic Effort, Abd Eff - Abdominal Effort, Ox Sat - Oxygen Saturation.

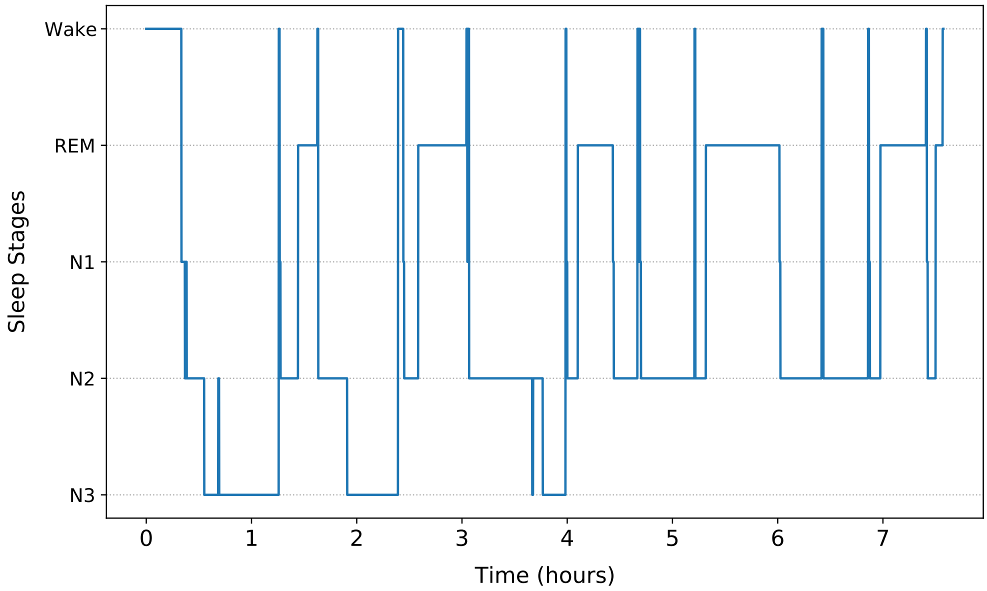


Figure 2.3: An example of a hypnogram, which is a graphical representation of sleep stages as a function of time. Five stages are included in the hypnogram: wake, non-rapid eye movement (NREM) stage 1 (N1), NREM stage 2 (N2), NREM stage 3 (N3), and rapid-eye movement (REM) sleep.

2.2.2.2 Sleep-Disordered Breathing

For sleep disordered breathing conditions, such as sleep apnea, technicians rely on the respiratory signals (airflow, respiratory effort, blood oxygen saturation, and snoring) for detection of respiratory events [16]. Sleep apnea is divided into three types based on the underlying pathology: obstructive sleep apnea (OSA), central sleep apnea (CSA), and mixed or complex sleep apnea. OSA is by far the most common phenotype and is the focus of this dissertation. OSA is characterized by upper airway obstruction, which causes breathing cessation and increased respiratory effort [1]. CSA is much less common and is characterized by the lack of proper signalling between the brain and respiratory muscles, which causes recurrent episodes of breathing cessation [101]. Complex sleep apnea is characterized by having both obstructive and central components [102]. OSA and CSA events are distinguished based on the presence of respiratory effort (OSA) or lack of respiratory effort (CSA) during the event which can be observed in the thoracoabdominal belt signals as presence or lack of excursions [16]. A detailed description of OSA is provided in the following section.

2.3 Obstructive Sleep Apnea

OSA is a sleep disorder characterized by recurrent collapse of the upper airway during sleep, resulting in decreased airflow (hypopnea) or total cessation of breathing (apnea), lasting until the upper airway reopens [1]. Figure 2.4 shows an example of a collapsed soft palate in a person during sleep.

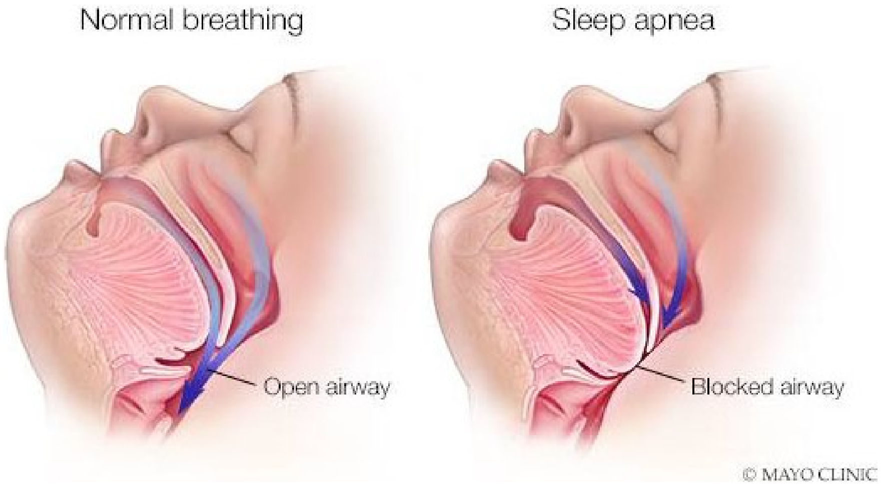


Figure 2.4: An example of a person breathing normally during sleep due to an unobstructed upper airway (left), and a person with a collapsed soft palate during sleep, which causes breathing cessation (right). Source: [103]

2.3.1 Causes of Obstructive Sleep Apnea

Why and how frequently the upper airway collapses is due to multiple factors, such as narrow upper airway anatomy, poor recruitment of dilator muscles during inspiration, central control of breathing (loop gain), and inability/ability to arouse (arousal threshold) [50]. A strong contributing factor to passive anatomy is obesity, which causes fat deposits around the upper airway that narrow the airway during sleep [51]. Other reasons include nasal congestion, enlargement of the tonsils, or the tongue falling backwards into the throat [52]. Supine position is also associated with increased events, and in some cases OSA is positional. A loss of muscle tone in the upper airway, either constitutional, or as the result of age or usage of various sedatives, is also contributing [52]. Finally, research shows that several craniofacial features, mainly related to the midface, jaw, and neck are indicative of the presence of OSA [38]. The relationship between OSA and craniofacial anatomy will be explored thoroughly in Chapter 3.

2.3.2 Implications of Obstructive Sleep Apnea

OSA patients suffer from daytime sleepiness and have increased risk of cardiovascular diseases [2–5], hypertension [6], stroke [7, 8], type 2 diabetes [9], and all-cause mortality [4, 5]. Furthermore, OSA patients have increased risk of being in a motor vehicle crash compared to individuals without the disorder [104]. A recent study estimated that almost half a billion people aged 30-69 have moderate to severe OSA globally [10] and evidence suggests that OSA is a severely underdiagnosed disorder [11–13]. The AASM estimated in 2015 that the economic burden of undiagnosed OSA was $150 billion in the US, while cost of diagnosis and treatment amounted to $50 billion [14], emphasizing that diagnosis and treatment of OSA can lead to significant reductions in health care costs.

2.3.3 Scoring of Obstructive Sleep Apnea Events

OSA is diagnosed by undergoing a PSG or an at-home sleep apnea test (HST) [19]. Subsequently, the signals are annotated manually by a sleep technician with respect to apneas and hypopneas observed in the recording. According to the AASM guidelines, an apnea is defined as a minimum decrease of 90% in airflow amplitude compared to the baseline for a minimum of 10 seconds [16]. A hypopnea is defined as a minimum decrease of 30% in airflow amplitude with either an associated arousal or an oxygen desaturation of at least 3% for a minimum of 10 seconds [16]. Figure 2.5 shows an example of two apneas and one hypopnea as observed in the respiratory PSG signals (oral airflow, nasal pressure, thoracoabdominal effort, oxygen saturation, and snoring). The absence of airflow is most notable in the nasal pressure signal, where the signal flattens out completely during an apnea while the signal amplitude reduces significantly during a hypopnea. Although less apparent than for the nasal pressure signal, the amplitude also reduces in the oral airflow and thoracoabdominal effort signals during these events, while the signal intensity in the snoring signal increases. In the oxygen saturation signal, desaturations are observed after each event, but there is a notable time delay between the decrease in airflow and the associated oxygen desaturations.

2.3.4 Apnea-Hypopnea Index

OSA severity is measured by the apnea-hypopnea index (AHI), representing the number of apneas and hypopneas per hour of sleep [17]. The classification of OSA severity based on AHI is outlined below:

- **Normal:** AHI $<$ 5 events/hour
- **Mild OSA:** 5 $\leq$ AHI $<$ 15 events/hour
- **Moderate OSA:** 15 $\leq$ AHI $<$ 30 events/hour
- **Severe OSA:** AHI $\geq$ 30 events/hour

2.3.5 Screening Questionnaires for Obstructive Sleep Apnea

Questionnaires containing OSA-related questions can be used to screen subjects and determine eligibility for PSG in a simple and cost-effective way [29]. Questionnaires for OSA include the Berlin Questionnaire [105], STOP Questionnaire [106], STOP-Bang Questionnaire [107], and Epworth Sleepiness Scale (ESS) [108]. The first three questionnaires pose questions about snoring, daytime sleepiness, observed episodes of breathing cessation, hypertension, and anthropometic measures such as having a large BMI or neck circumference [29]. If the subject answers yes to more than a certain number of questions (which varies depending on the questionnaire), the subject is classified

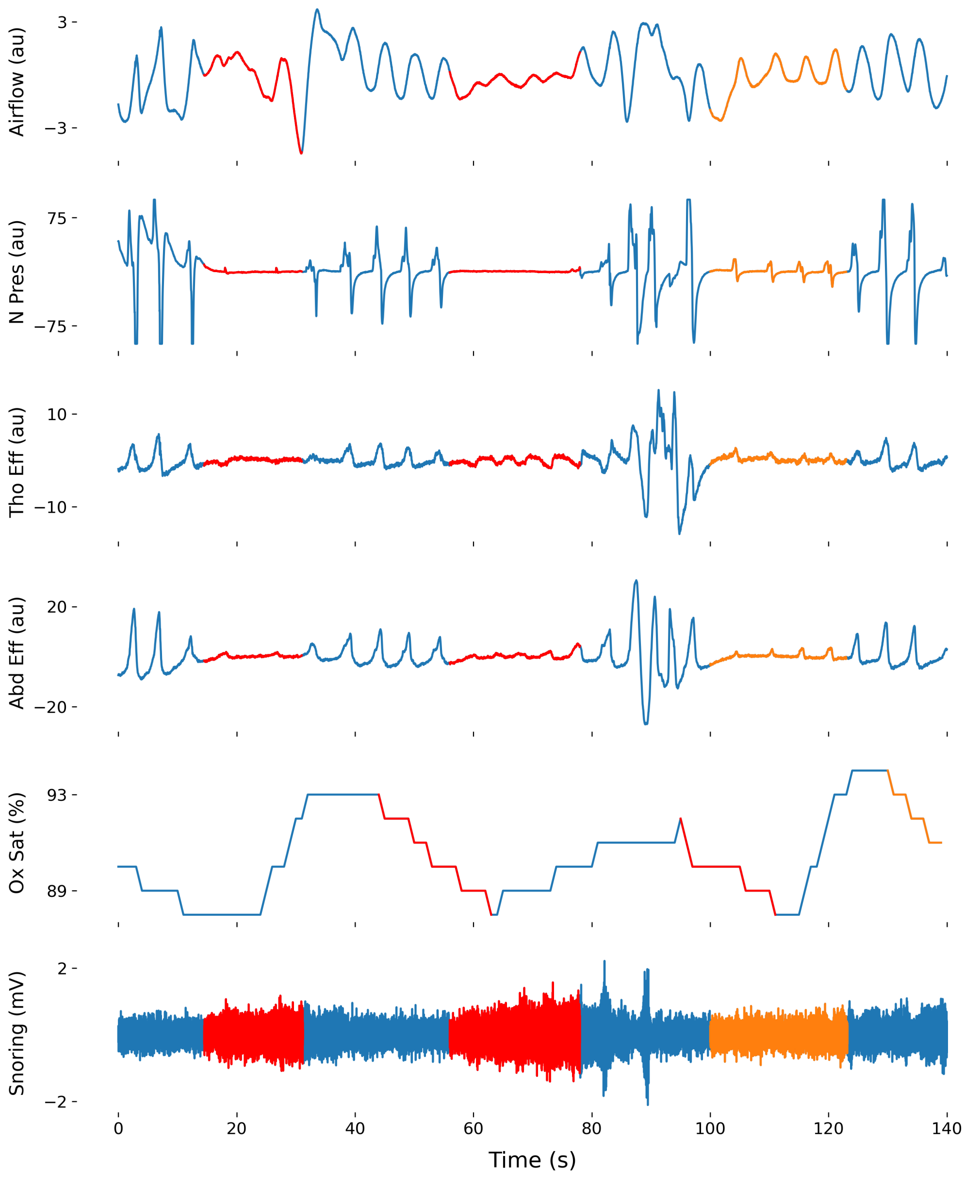


Figure 2.5: Example of two apneas (red) and one hypopnea (orange) in a polysomnography recording segment. Apnea and hypopnea events are immediately reflected in the airflow, nasal pressure (N Pres), thoracic effort (Tho Eff), abdominal effort (Abd Eff), and snoring signals. The events are most apparent in the nasal pressure signal, where the signal flattens out completely during apneas and the signal amplitude reduces dramatically during hypopneas. After the event, an associated oxygen desaturation is observed in the oxygen saturation signal (Ox Sat) with a variable time delay relative to the decreased airflow.

as having high risk of OSA. The ESS is used to assess daytime sleepiness and ranges from 0-24, where a higher score indicates increased daytime sleepiness and OSA risk [108]. The score is obtained by asking the subject to rate, on a scale from 0-3, how likely they are to doze off or fall asleep while engaging in eight different activities and summing the scores for all eight questions.

2.3.6 Continuous Positive Airway Pressure

Continuous positive airway pressure (CPAP) is the gold-standard treatment for OSA and works by providing a constant level of pressure sufficient to keep the upper airway open [53]. Examples of different CPAP masks are provided in Fig. 2.6, showing that some masks only apply pressure intra-nasally, while others (full face masks) provide pressure through nose and mouth e.g., for those who breathe through their mouth during sleep. Although CPAP is extremely effective, studies show that up to 50% of users give up on the device within a year of therapy because of intolerance, noise, discomfort, or a negative impact on intimacy [54]. Oral appliances may reduce upper airway obstructions by advancing the mandible or refraining the tongue and epiglottis from falling back, but, as they are generally less effective than CPAP they are considered the second line of treatment after CPAP [109]. For some OSA patients, surgical procedures such as removing excess tissue or advancing the upper and lower jaw can be viable options to prevent collapse or increase upper airway space [57, 110]. Surgery as a treatment option for OSA and the associated upper airway examination will be explored thoroughly in Chapter 4.

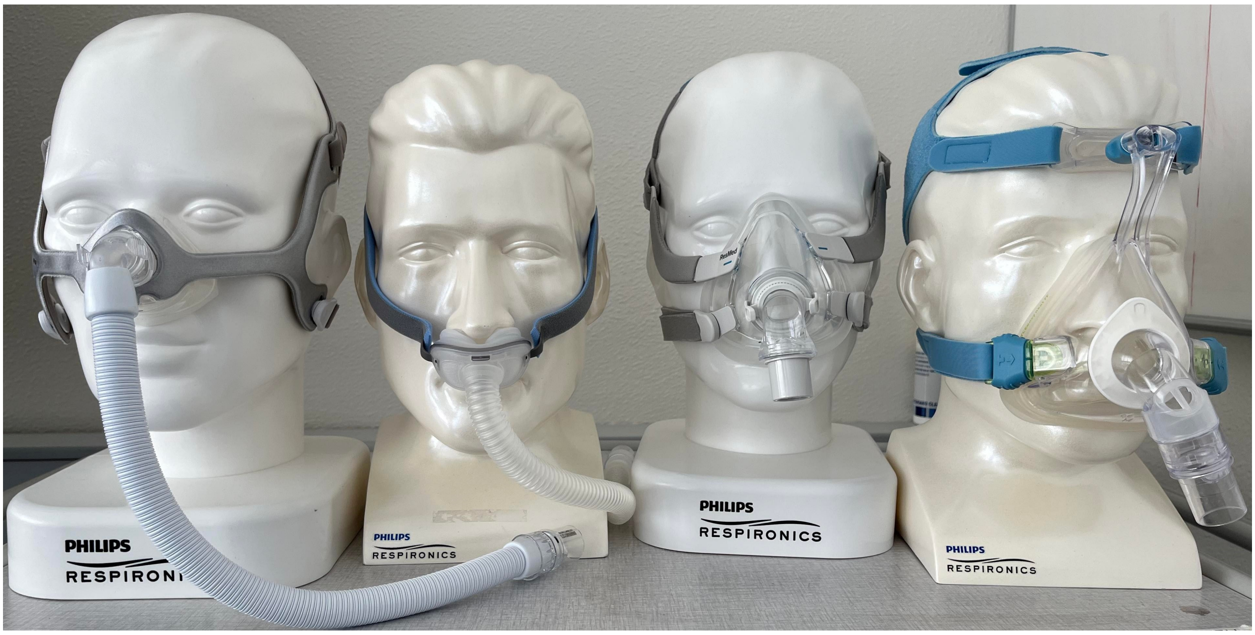


Figure 2.6: Examples of different continuous positive airway pressure masks used to prevent upper airway collapse in obstructive sleep apnea patients. The two masks to the left only apply intra-nasal pressure, while the two masks to the right are full face masks that apply pressure through nose and mouth. Full face masks are required for patients who breathe through their mouth during sleep or patients who tend to get nasal congestion.

Chapter 3

Detection of Obstructive Sleep Apnea From 3D Craniofacial Scans

This chapter explores the diagnostic potential of 3D craniofacial scans in relation to obstructive sleep apnea (OSA), which can be used to develop a clinical screening system for faster and cheaper detection of OSA. The chapter consists of two parts, each based on a published paper:

1) Prediction of Patient Demographics Using 3D Craniofacial Scans and Multi-View CNNs [111]

2) Estimation of Apnea-Hypopnea Index Using Deep Learning on 3D Craniofacial Scans [112]

Part 1) was published as a conference paper (Paper I) and served as a proof of concept for the entire methodology utilized for part 2). The polysomnography (PSG) recordings, which the apnea-hypopnea index (AHI) values are derived from, were not readily available in the initial stages of the PhD project, but the 3D craniofacial scans and corresponding patient demographics were available. Thus, a decision was made to initially attempt estimating sex, age, and body mass index (BMI) from the 3D scans and publish the results in a conference paper. Intuitively, patient demographics should be easier to predict than AHI values from a 3D craniofacial scan, so results from part 1) also served as a baseline for what the maximum expected performance could be for part 2). Part 2) was published as a journal paper (Paper II) and explored how 3D craniofacial imaging can be used to estimate the presence and severity of OSA measured by the AHI. The main content of this chapter is based on that journal paper.

Section 3.1 provides the research background and motivation for this chapter. Section 3.2 states the research questions and objectives of this chapter. Section 3.3 contains a data and methods description, and a presentation of results and discussion for part 1) which is based on Paper I.

Similarly, Section 3.4 contains a data and methods description, and a presentation of results and discussion for part 2) which is based on Paper II. Finally, Section 3.5 concludes the chapter by answering the research questions posed in Section 3.2 and relating the findings to Hypothesis 1, which is stated in Section 1.4 and restated in Section 3.2.

3.1 Research Background

Obesity and craniofacial features strongly contribute to OSA risk; thus, technology and clinical examination tools have been developed to assess these features. In clinical exams, it is frequent for the clinician to examine the size of the jaw, top of the mouth, and position and size of the tongue [113]. Imaging techniques have been proposed, including cephalometry [114–116], computed tomography [116, 117], and magnetic resonance imaging [118–120], showing that typical anatomical features of OSA patients are maxillary deficiency, mandibular retrusion, an abnormal cranial base, and an inferiorly positioned hyoid bone. Furthermore, dynamic collapsibility can be identified with drug-induced sleep endoscopy (DISE) [58]. These methods are, however, rarely used in routine practice, except in case of surgery for OSA.

Since imaging modalities are cumbersome and expensive, recent research has investigated the predictive value of facial imaging [39]. Lee et al. [40] analyzed frontal and profile images of 180 subjects, manually deriving measurements on the face and neck to classify subjects with or without OSA using logistic regression. Others [41–43] used Support Vector Regression (SVR) [121] on similar landmarks to predict the AHI, a procedure subsequently improved by Balaei et al. [44, 45] who used automatic instead of manual placement of landmarks. Islam et al. [46] used 3D scans from 69 subjects, which they converted to 2D depth maps of the frontal face, applied transfer learning on a VGG-16 [122] deep convolutional neural network (CNN), and modified the network to classify subjects into OSA and normal.

Although these studies had some success, all used small sample sizes, and, with the exception of Islam et al. [46], all first derive possibly discriminative facial features from annotated landmarks as detected on 2D frontal and profile images, a process followed by statistical feature selection [123]. Interestingly, these studies typically report different discriminative features and varying number of optimal features using feature selection methods. As attempted by Islam et al. [46] using a small sample size, we believe that feature selection should be unbiased by avoiding manual extraction of features and applying CNNs instead for automatic feature extraction. Facial imaging for OSA diagnosis can be performed in one minute and presents a clear advantage; it does not require an overnight stay at a sleep clinic with several sensors connected to the body and a subsequent manual analysis by sleep technicians, thereby saving both time and resources, while being more comfortable for the patient.

3.2 Research Questions and Objectives

Based on the research background above, Hypothesis 1 from Section 1.4 is restated, followed by research questions derived from Hypothesis 1 that this chapter aims to answer.

Hypothesis 1

An automatic screening system can be invented, based on dedicated computer vision models, which utilizes 3D craniofacial scans to estimate presence and severity of obstructive sleep apnea more accurately than current screening questionnaires in a fast, cheap, and data-driven manner.

Research Questions

Research Question 1: Can a dedicated computer vision model be trained to accurately estimate the sex, age, and BMI of subjects based on their 3D craniofacial scans?

Research Question 2: Can a dedicated computer vision model be trained to accurately estimate the AHI of subjects based on their 3D craniofacial scans?

Research Question 3: Can predicted AHI values from the proposed model be used to accurately detect presence of OSA by classifying subjects as being normal or having OSA?

Research Question 4: Can adding clinically relevant information like demographics and questionnaires increase performance of the proposed model?

Research Question 5: Can the proposed model perform better than current screening questionnaires for OSA?

Research Question 6: Can the proposed model perform at a level similar to that of sleep medicine specialists with years of experience?

Research Question 7: Can regions of the face and neck be identified, which the proposed model focuses on when predicting AHI?

From the research questions posed above, research objectives are formulated below, each one designed to answer a specific research question:

(i) Pre-process 3D craniofacial scans to make them suitable as input for CNNs and extract sex, age, and BMI values for each subject. Then utilize these input and output pairs to train and evaluate dedicated computer vision models to predict the sex, age, and BMI of subjects.

(ii) Pre-process 3D craniofacial scans to make them suitable as input for CNNs and derive AHI values from the corresponding PSGs. Then utilize these input and output pairs to train and evaluate a dedicated computer vision model to predict the AHI of subjects.

(iii) Use the clinically meaningful cut-off of 15 events/hour for the predicted AHI values to detect presence of OSA by classifying subjects as being normal (AHI < 15) or having OSA (AHI ≥ 15).

(iv) Investigate whether adding demographics and questionnaire variables to the proposed model improves performance compared to using only 3D scans.

(v) Compare performance of the proposed model to performance obtained using current screening methods, i.e., screening questionnaires for OSA.

(vi) Recruit three experienced sleep medicine specialists to estimate the AHI values of subjects based on their 3D craniofacial scans and compare the proposed model's performance to the sleep medicine specialists' performance.

(vii) Investigate craniofacial features that the model focuses on when predicting AHI and create a topographic display of these features/regions.

These research objectives will be fulfilled in the following sections.

3.3 Paper I: Prediction of Patient Demographics Using 3D Craniofacial Scans and Multi-View CNNs

Abstract

Purpose: 3D data is becoming increasingly popular and accessible for computer vision tasks. A popular format for 3D data is the mesh format, which can depict a 3D surface accurately and cost-effectively by connecting points in the (x,y,z) plane, known as vertices, into triangles that can be combined to approximate geometrical surfaces. However, mesh objects are not suitable for standard deep learning techniques due to their non-euclidean structure. We present an algorithm which predicts the sex, age, and body mass index of a subject based on a 3D scan of their face and neck.

Methods: This algorithm relies on an automatic pre-processing technique, which renders and captures the 3D scan from eight different angles around the x-axis in the form of 2D images and depth maps. Subsequently, the generated data is used to train three convolutional neural networks, each with a ResNet18 architecture, to learn a mapping between the set of 16 images per subject (eight 2D images and eight depth maps from different angles) and their demographics.

Results: For age and body mass index, we achieved a mean absolute error of 7.77 years and 4.04 kg/m^2 on the respective test sets, while Pearson correlation coefficients of 0.76 and 0.80 were obtained, respectively. The prediction of sex yielded an accuracy of 93%.

Conclusion: The developed framework serves as a proof of concept for prediction of more clinically relevant variables based on 3D craniofacial scans stored in mesh objects.

3.3.1 Methods

This section describes the data collection of 3D craniofacial scans and the subsequent pre-processing and use of these scans to predict patient demographics using multi-view CNNs.

3.3.1.1 Data Collection

Data was collected at 11 different sleep clinic sites as part of the Stanford Technology Analytics and Genomics in Sleep (STAGES) study, which was initiated in 2018 and prematurely terminated in 2020. STAGES was designed to better understand and characterize sleep disorder phenotypes on a large scale by collecting sleep data from 30,000 subjects, but at the end of the study approximately 1800 subjects had participated. For each subject participating in the study, a detailed sleep questionnaire, actigraphy, psychometric testing, a PSG, a 3D craniofacial scan, and blood samples were collected.

The 3D craniofacial scans were performed using a Structure Sensor from Occipital Inc. [124, 125] attached to an iPad Pro from Apple as shown in Fig 3.1. The scanning procedure is as follows: The subject being scanned is seated on a chair and the person capturing the scan walks around the subject, capturing their face and neck from several angles to obtain a complete 3D craniofacial surface scan. This procedure is demonstrated in Fig. 3.2. The reconstruction of the complete surface scan is performed in the app software automatically and the final result as displayed on the iPad is shown in Fig. 3.3. Each scan took approximately one minute to complete and was captured either at night before the PSG or in the morning after the PSG. uGo3D Inc. developed an app for STAGES which was responsible for transferring the scans to the server after they were captured.

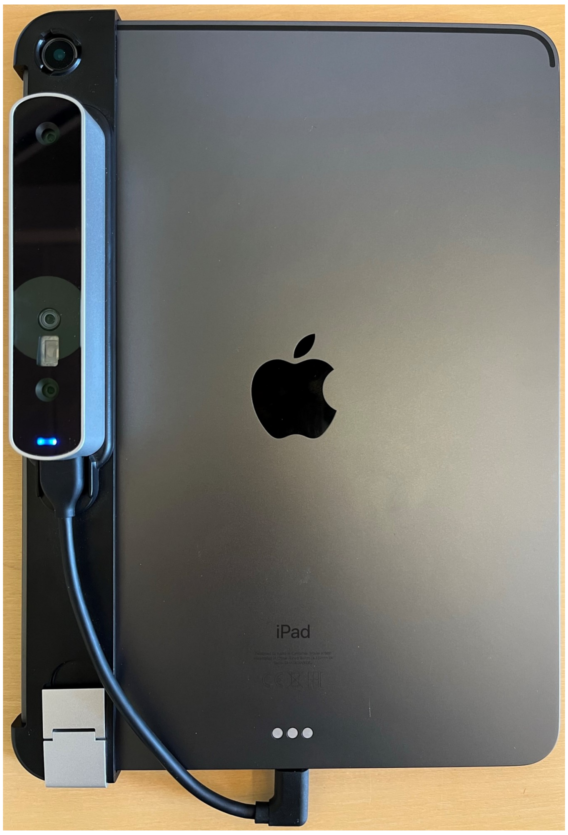


Figure 3.1: A Structure Sensor from Occipital Inc attached to an Ipad Pro from Apple used to capture 3D craniofacial surface scans. Although newer and more expensive versions of the sensor are now available, the cost of the original Structure Sensor used for this study was $379. The Structure Sensor utilizes an infrared sensor to capture objects in 3D and has a depth resolution of 640x480, recommended range of 0.4-3.5 meters, and a field of view of 58x45.

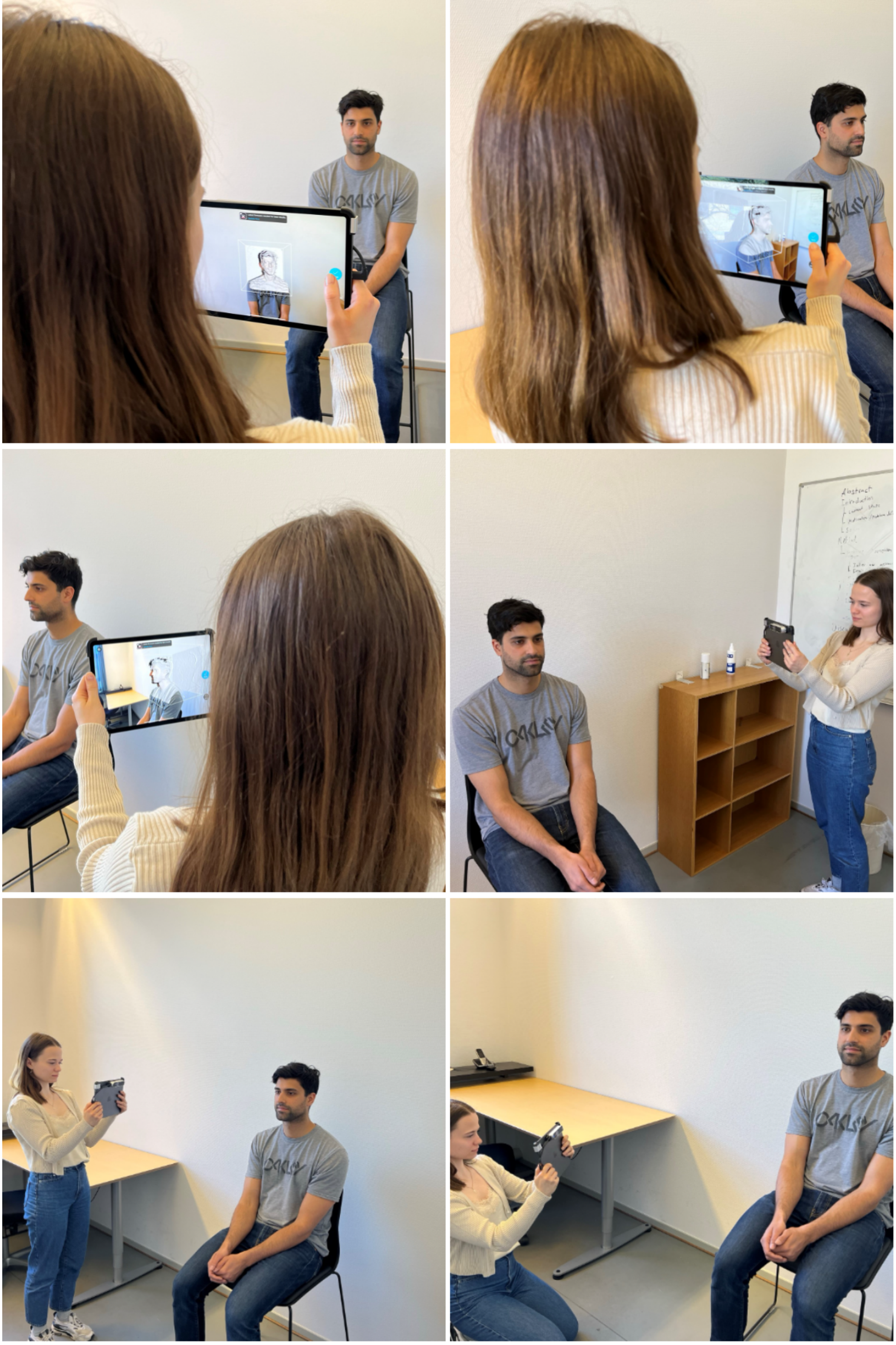


Figure 3.2: The procedure of capturing a 3D craniofacial scan of a subject. The person performing the procedure captures the subject from several angles using an iPad Pro from Apple with a Structure Sensor from Occipital Inc attached to the back. The reconstruction of the surface scan is performed automatically in the app software.

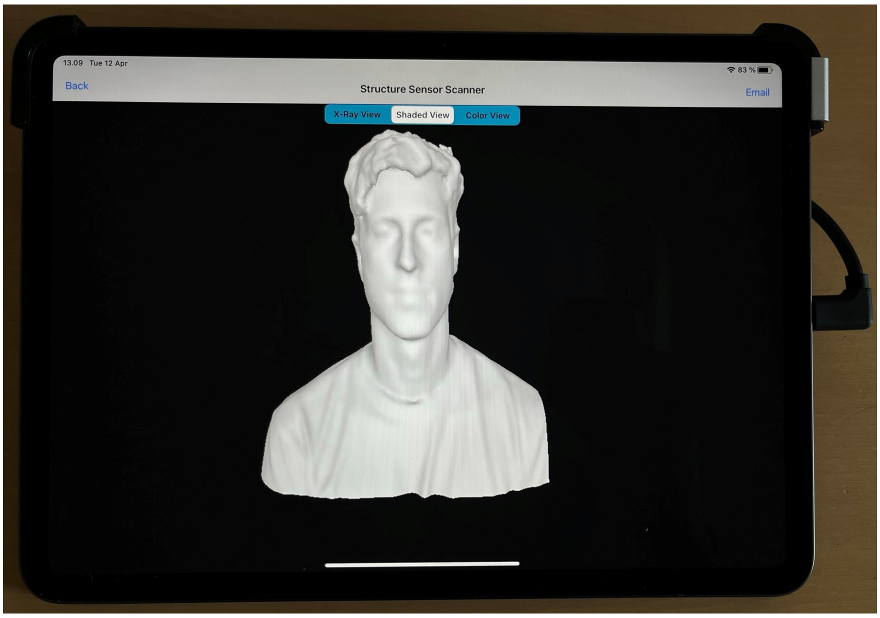


Figure 3.3: A completed surface scan displayed on the iPad which has been used to capture the scan using a Structure Sensor from Occipital Inc attached to the back of the iPad. The surface scan is displayed without associated textures, but a texture file is created with the scan, which can be used to render the scan with textures subsequently on e.g. a computer.

The sleep questionnaire included in this study was a modified STOP-Bang questionnaire [107], where the neck circumference of the subject was excluded. The complete STOP-Bang questionnaire is included in Appendix A of this dissertation. The modified STOP-Bang questionnaire was used as a screening method for OSA, without the need for a specialist, by asking the subject about snoring, tiredness, observed apnea, high blood pressure, sex (is the person male), age (is the person older than 65 years), and BMI (is the BMI greater than 35 kg/m^2).

Each variable gets a value of 1 if the person answers yes, so the total score from the questionnaire ranges from 0 to 7. A score of 0-2 indicates low risk of OSA, 3-4 indicates intermediate risk, and 5-7 indicates high risk. In this study, a score greater than or equal to 3 was used to indicate presence of OSA. The STOP-Bang questionnaire was selected over other OSA screening questionnaires (such as the Berlin Questionnaire or Epworth Sleepiness Scale) due to its superior diagnostic accuracy [28, 29].

Each institution's Ethical Review Board approved all procedures involving human subjects. All participants provided written informed consent to participate in the study. The 3D scans from STAGES are not publicly available, since they count as personal identifiable data, which the Institutional Review Board would not allow to be made public. All other data from STAGES, apart from blood samples, have been made available as part of the National Sleep Research Resource (NSRR) [126].

3.3.1.2 Data Description

Each 3D craniofacial scan was stored as a mesh object in an OBJ format, defined by its set of vertices $V = \{v_1, v_2, ..., v_n\},\ V \in \mathbb{R}^3$, its texture coordinates $V_t = \{v_{t1}, v_{t_2}, ..., v_{t_m}\},\ V_t \in \mathbb{R}^3$, and its triangles, also known as faces, $F = \{(v_1, v_{t1}), (v_2, v_{t2}), ..., (v_k, v_{tk})\},\ F \in \mathbb{R}^3$. A vertex is a point in 3D space defined by its (x, y, z) coordinates, a texture is defined by its (u, v) coordinates, and a triangle consists of three vertex and texture pairs.

Three interconnected vertices form a triangle and several of these triangles are combined to approximate a surface in three dimensions. Textures are bitmap images, that can be laid over the surface scan to make it look more realistic. Figure 3.4 provides a simple example of the basic components of a mesh, i.e., vertices, edges, and triangles. Edges are used to connect two vertices as seen in the illustration. Figure 3.5 shows an example of a craniofacial scan from three different angles, where the triangles that are combined to make the scan are depicted as well.

During data collection for STAGES, a total of 1756 3D scans were captured from enrolled participants. However, subjects were discarded if they had missing sex, age, or BMI, so 1605 subjects were included in this study. Of the 1605 subjects, 855 were female and 750 were male. Mean age ± standard deviation was 45.8 ± 15 years and the BMI was 31.3 ± 8.9 kg/m^2. The distribution of sex for the included subjects is shown in Fig. 3.6 (a), distribution of age is shown in Fig. 3.6 (b), and distribution of BMI is shown in Fig. 3.6 (c).

In the following sections, the pre-processing of 3D scans and their subsequent use to train and evaluate multi-view CNNs for estimation of sex, age, and BMI of subjects is described.

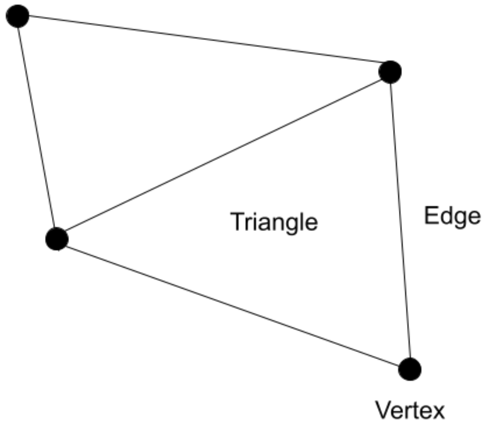


Figure 3.4: An illustration of the basic building blocks of a 3D mesh surface scan. A mesh consists of interconnected vertices, which are points in 3D space (shown in 2D in the illustration for simplicity) characterized by their (x, y, z) coordinates. Vertices are connected using edges and three interconnected vertices form a triangle. A mesh scan is composed of such triangles, which are used to approximate a given surface in 3D. In this example, four vertices have been connected to form two triangles.

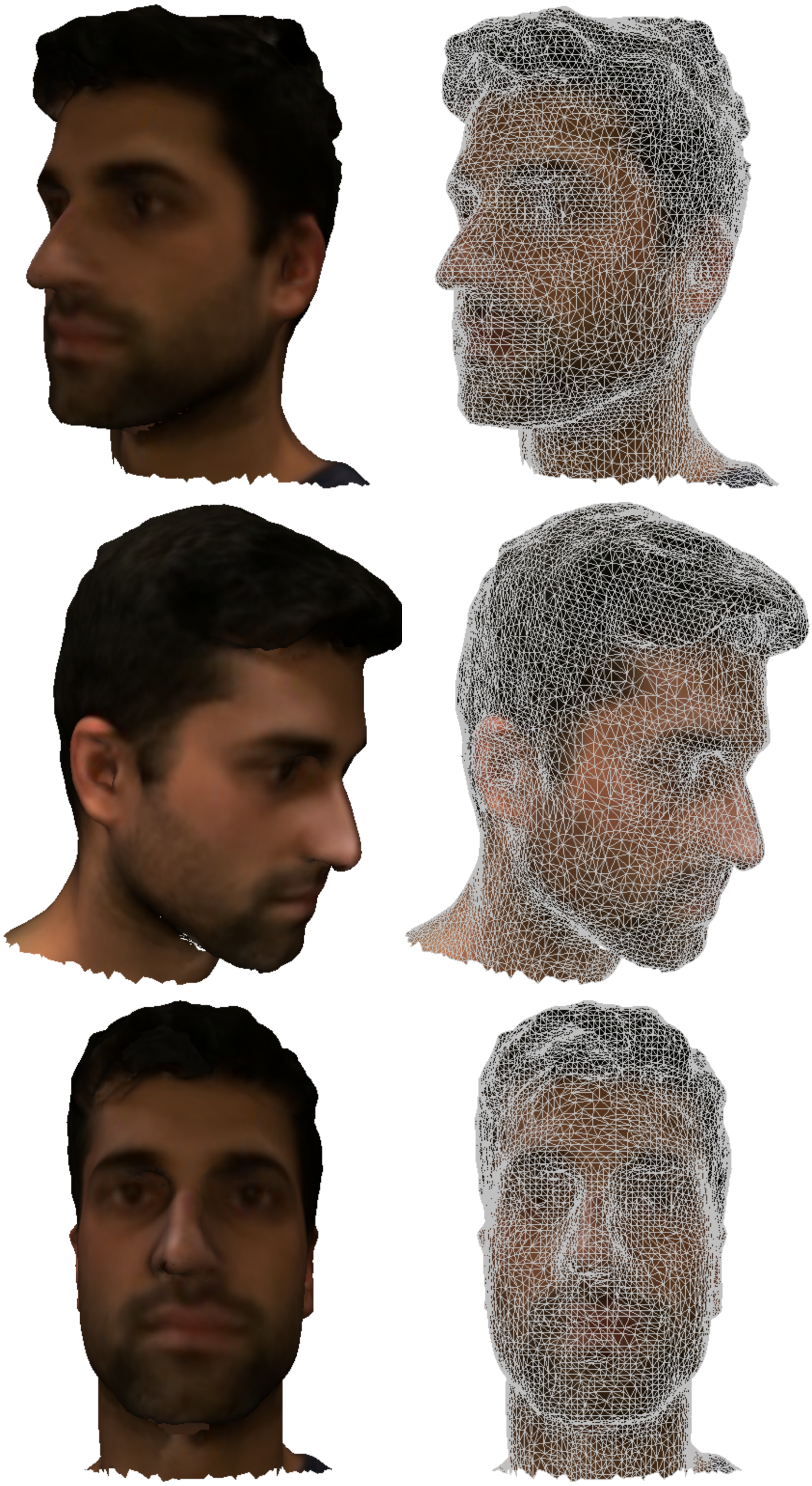


Figure 3.5: Example of a 3D mesh scan rendered from three different angles (left) and the triangles that are used to approximate the surface scan from the same three angles (right). The triangles consist of three interconnected vertices each, where a vertex is a point in 3D space specified by its (x, y, z) coordinates.

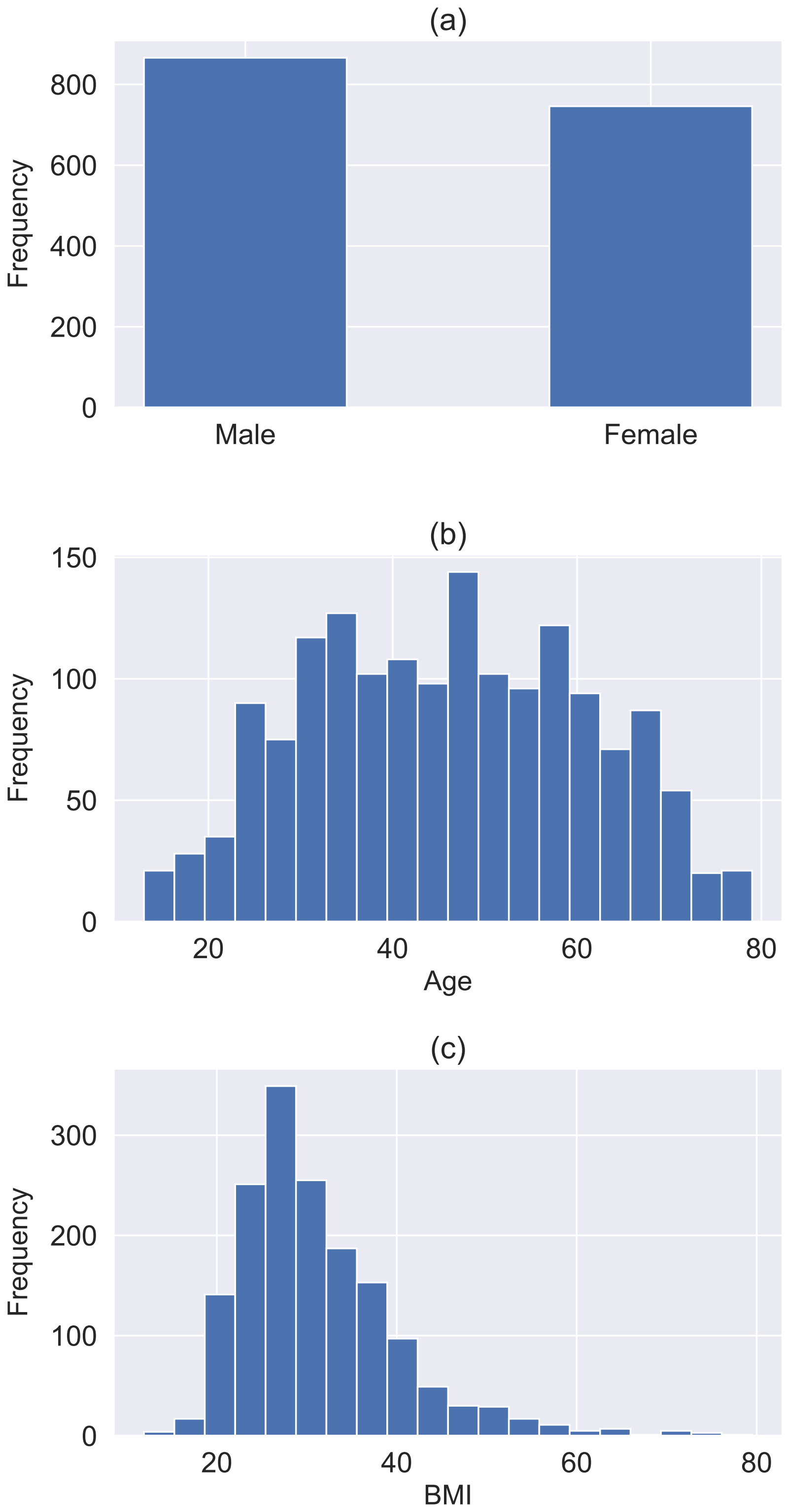


Figure 3.6: Distribution of (a) sex, (b) age, and (c) BMI for 1605 subjects in the dataset used for estimating patient demographics from 3D craniofacial scans based on a proposed computer vision model.

3.3.1.3 Pre-Processing

Since 3D mesh scans are non-Euclidean, meaning not defined in a flat 2D plane, they cannot be used directly as inputs to CNNs. Thus, to make our 3D craniofacial scans suitable for CNNs, the multi-view consensus CNN for 3D facial landmark placement (Deep-MVLM) algorithm [127] was applied to transform each scan into a set of 2D images and depth maps captured from angles around the scan. Deep-MVLM was chosen as it outperforms state of the art algorithms [128, 129] and does not rely on pre-alignment of scans, such that they have the same orientation in 3D space. The choice of metric in Deep-MVLM also makes it more suitable for 3D surfaces than similar methods. Deep-MVLM was only used for alignment of scans and generation of 2D images and depth maps, and not for the subsequent prediction of AHI.

Deep-MVLM works by first applying a pre-trained neural network to automatically detect and place 73 pre-specified landmarks on each mesh object. Figure 3.7 shows an example of these landmarks detected and placed on a subject rendered from three different angles. The landmarks are utilized to align all scans in 3D space, thereby removing the influence of translation and rotation. The alignment of scans is obtained by using a least squares solution based on the detected landmarks and therefore the individual landmark errors are less important for the overall performance [127].

After alignment of 3D scans, each scan is displayed multiple times as flat 2-D images taken from different angles; this is done by rotating the scans 45 degrees consecutively and capturing a 2D image and a depth map at each angle, yielding eight pairs of 2D images and depth maps for each scan. Each 2D image and depth map was normalized to the range [0,1] by dividing each pixel value by 255. This was done to ensure faster convergence during training. The resulting images were stacked in a matrix, representing different views of a subject along the 3$^{\text{rd}}$ dimension of the matrix.

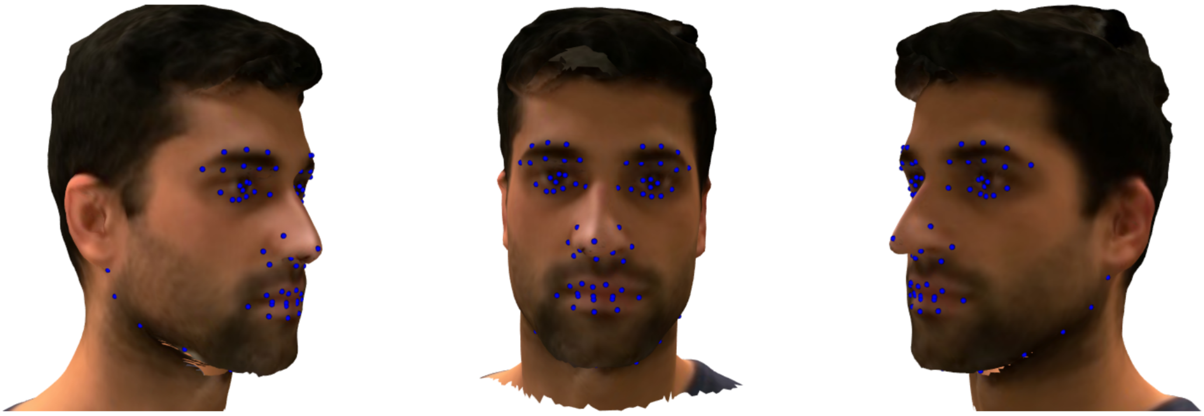


Figure 3.7: Examples of 73 detected landmarks on a 3D mesh scan displayed from three different angles. The landmarks are detected automatically by using the multi-view consensus convolutional neural network for 3D facial landmark placement (Deep-MVLM) algorithm [127].

For this study, all eight pairs of 2D scans and depth maps were included as shown in Fig. 3.8, to retain as much information as possible from the 3D scans. Textures were not utilized as observed in the figure because computational cost was a concern at this time in the PhD project and this study only served as a proof of concept for the next, much larger study. The 2D images and depth maps had one channel each (grayscale) and since there were eight 2D images and eight depth maps, the final stacked matrix per subject contained 8 x 1 + 8 x 1 = 16 channels in total and 224x224 pixels per channel or image.

3.3.1.4 Multi-View Convolutional Neural Networks

The purpose of applying machine learning was to reveal data-driven mapping differences within the multi-view inputs across patient demographics. For this purpose, we implemented multi-view CNNs with a ResNet18 architecture [130]. CNNs were implemented for their ability to automatically extract meaningful features from the input images in relation to the desired outputs. They do so by applying layers of convolutions and non-linear activation functions to the input images, where the weights of the convolution filters are learned during training based on the training data [131]. A ResNet18 architecture was selected because it is state-of-the-art for image recognition compared to other CNN architectures such as AlexNet [132] or VGG-16 [122]. ResNets utilize skip connections between layers to allow deeper architectures (i.e., more layers) without hurting performance while being computationally cost-efficient to train [130]. Figure 3.9 shows the network architecture with all details provided (including skip connections) for estimation of patient demographics using multi-view inputs derived from 3D craniofacial scans.

A separate network was implemented for each demographic (sex, age, and BMI) because the target values were different. The implemented multi-view CNN took 16 channels as input per subject and the input was processed by the standard blocks of a ResNet18 network. One additional fully

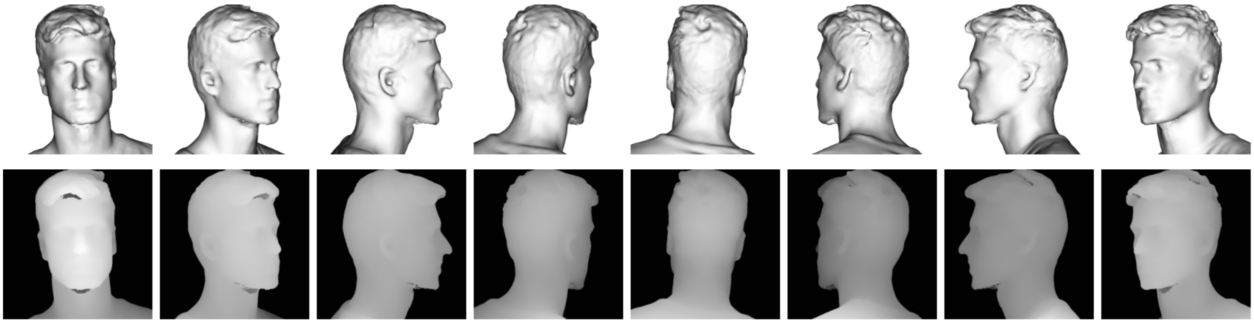


Figure 3.8: Example of eight pairs of 2D images (top row) and depth maps (bottom row) captured at different angles using the multi-view consensus convolutional neural network for 3D facial landmark placement (Deep-MVLM) algorithm [127]. These eight pairs of 2D images and depth maps are used as input for each subject in the proposed multi-view convolutional neural network for predicting patient demographics.

connected layer was added at the end, which was used to reduce 512 features to 256, followed by a ReLU activation function and dropout (probability of 0.4). Dropout was used as a regularizing component to reduce overfitting during training. The dropout probability and number of fully connected layers and neurons were selected based on hyperparameter tuning, which was performed in a grid search-like manner where the hyperparameters were varied and different combinations of these were investigated. The optimal hyperparameters were the ones which yielded the lowest error on the validation set, which is defined in the next section.

The final output layer consisted of one neuron for prediction of age and BMI or two neurons for prediction of sex followed by a sigmoid activation function. For age and BMI, the desired output was a single continuous value, which is why only one output neuron was required. For sex, the desired outputs were two predicted probabilities, one for male and one for female, which is why two output neurons were required. In this case, a sigmoid activation function was used to transform the outputs into probabilities in the range [0,1].

3.3.1.5 Training, Validation, and Testing

Training, validation, and testing was carried out by using a training, validation, and test set split (65% training, 25% validation, and 10% test set). The split was prioritized such that sufficient data was utilized for training and validation of the model to learn an accurate mapping from craniofacial images to patient demographics and hyperparameter optimization, while keeping a sufficient amount of data in the test set for statistical analyses.

For age and BMI estimation, the mean absolute error (MAE) was used as loss function given by:

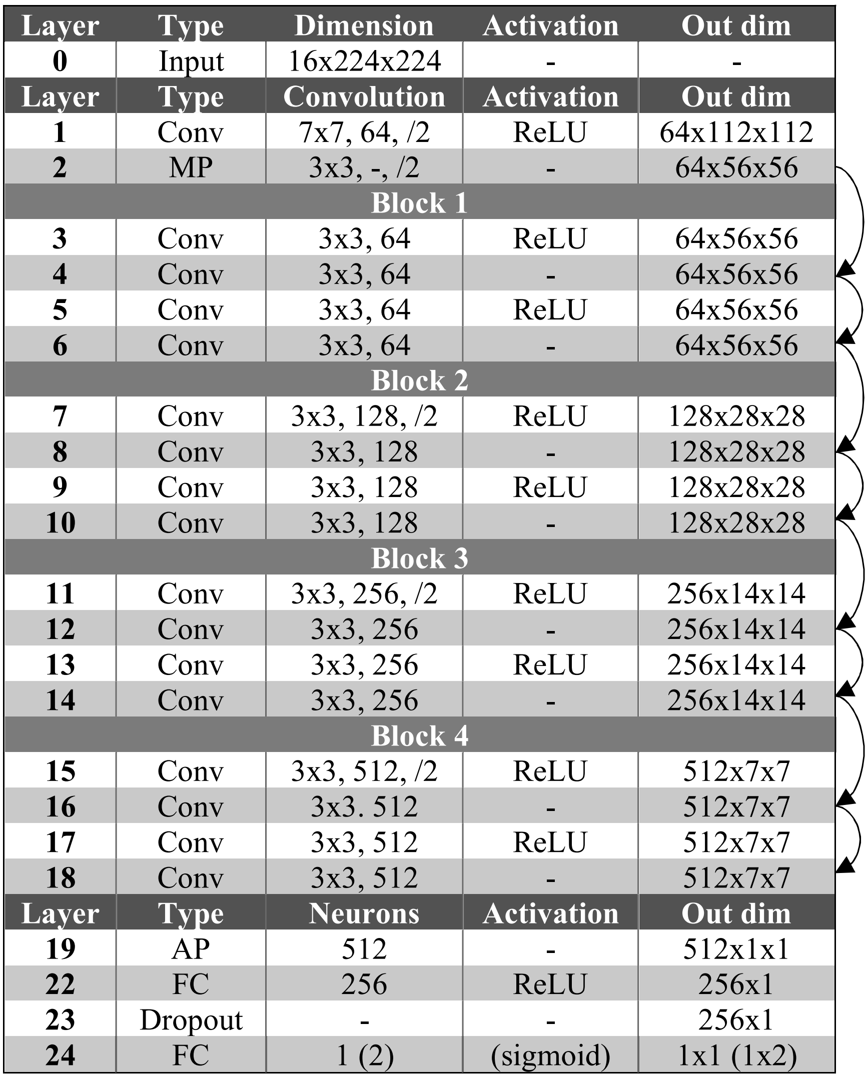
$$L = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{3.1}$$

where y_i is the true age or BMI value, $\hat{y}_i$ is the predicted age or BMI value, and N is the number of samples in the training set. MAE was chosen because it is robust towards outliers, such as the extreme BMI values observed in Fig. 3.6, where values of 60-70 kg/m^2 are observed.

For estimation of sex, binary cross entropy was used as loss function given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log p_i + (1 - y_i) \cdot \log(1 - p_i), \tag{3.2}$$

where $\log p_i$ represents the predicted log probabilities of being male and $\log(1 - p_i)$ represents the predicted log probabilities of being female. Binary cross entropy was chosen as it is the natural choice of loss function for binary classification problems.



Layer	Type	Dimension	Activation	Out dim
0	Input	16x224x224	-	-
Layer	**Type**	**Convolution**	**Activation**	**Out dim**
1	Conv	7x7, 64, /2	ReLU	64x112x112
2	MP	3x3, -, /2	-	64x56x56
		Block 1		
3	Conv	3x3, 64	ReLU	64x56x56
4	Conv	3x3, 64	-	64x56x56
5	Conv	3x3, 64	ReLU	64x56x56
6	Conv	3x3, 64	-	64x56x56
		Block 2		
7	Conv	3x3, 128, /2	ReLU	128x28x28
8	Conv	3x3, 128	-	128x28x28
9	Conv	3x3, 128	ReLU	128x28x28
10	Conv	3x3, 128	-	128x28x28
		Block 3		
11	Conv	3x3, 256, /2	ReLU	256x14x14
12	Conv	3x3, 256	-	256x14x14
13	Conv	3x3, 256	ReLU	256x14x14
14	Conv	3x3, 256	-	256x14x14
		Block 4		
15	Conv	3x3, 512, /2	ReLU	512x7x7
16	Conv	3x3. 512	-	512x7x7
17	Conv	3x3, 512	ReLU	512x7x7
18	Conv	3x3, 512	-	512x7x7
Layer	**Type**	**Neurons**	**Activation**	**Out dim**
19	AP	512	-	512x1x1
22	FC	256	ReLU	256x1
23	Dropout	-	-	256x1
24	FC	1 (2)	(sigmoid)	1x1 (1x2)

Figure 3.9: The applied multi-view convolutional neural network architecture for predicting sex, age, and BMI of a subject based on 16-dimensional input craniofacial images derived from a 3D scan. The input dimensions are given by number of channels x height x width. The convolution layers are specified by filter size (e.g., 3x3), number of channels (e.g., 64), and a stride (e.g., /2). The same applies for the max pooling (MP) layer. The output dimensions of the feature maps are given by number of channels x height x width. The convolution layers are always followed by batch normalization. The dropout layer has a keep probability of 0.4. If the skip connections (arrows) are applied, the feature maps are down sampled instead by applying 1x1 filters with a stride of 2x2. The final output layer consists of 1 neuron for prediction of age and BMI and 2 neurons for prediction of sex followed by a sigmoid layer (indicated by a parentheses). AP - Average pooling, FC - Fully connected.

The learning rate was set to $1 \cdot 10^{-5}$ with a weight decay of $5 \cdot 10^{-4}$ and was chosen using hyperparameter tuning. The optimal choice was a learning rate that is not too high, which can cause the parameter update via gradient descent to diverge from the minima, and not too low, which would slow down training significantly. The weight decay is used to add a penalty term to the loss function during optimization, which shrinks the weights and helps to prevent overfitting.

The batch size was set to 64 for all three networks and was limited by computational resources. The Adam optimizer [133] was used for optimization of the network due to its superior performance compared to other optimization algorithms. Early stopping was applied as a stopping criterion for training when the validation error did not decrease for 5 consecutive epochs (patience of 5) to avoid any overfitting.

Python 3.7.4 and Pytorch 1.3.1 were used for pre-processing and deep learning purposes. Training was carried out on a GeForce GTX 1080 and networks estimating age and BMI took approximately two hours to train, while the network estimating sex took less than an hour.

3.3.1.6 Performance Measures

To evaluate model performance for estimation of age and BMI for subjects in the test set, MAE and Pearson Correlation Coefficient (PCC) were used as performance measures. Calculation of MAE is given by Eq. 3.1, but applied to the test set predictions instead of the training set. The PCC was calculated as:

$$r = \frac{\sum_{i=1}^{M}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{M}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{M}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{3.3}$$

where $\bar{y}$ is the mean of the true age or BMI values, $\bar{\hat{y}}$ is the mean of the predicted age or BMI values, and M is the number of samples in the test set. The PCC expresses the linear correlation between true and predicted age or BMI values and ranges between -1 and 1, where -1 is the maximum negative correlation and 1 is the maximum positive correlation between the two variables.

For estimation of sex, accuracy was used to evaluate model performance, which is calculated by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3.4}$$

where TP are the true positives, TN are the true negatives, FP are the false positives, and FN are the false negatives. A positive was defined as being male, so TP denoted the number of males correctly classified as males and TN denoted the number of females correctly classified as females. FP represented the number of females incorrectly classified as males, and FN was the number of males incorrectly classified as females.

Bland-Altman plots [134] were generated to illustrate the patterns of disagreement between true and predicted age and true and predicted BMI values, respectively, while confusion matrices were displayed for the classification of sex and BMI categories to show the fraction of misclassified subjects in each category as well as the sensitivity of the model in classifying each category.

3.3.2 Results and Discussion

The performance metrics reported on the estimation of sex, age, and BMI were all evaluated on the test set consisting of 160 subjects.

3.3.2.1 Estimation of Sex

For estimation of sex, the network was trained for 13 epochs before early stopping occurred. A mean accuracy of 93% was achieved in classifying the sex of all subjects in the test set. To put this in context, if the network simply predicted all subjects in the test set to be women, an accuracy of 52% would be achieved. Figure 3.10 shows the normalized confusion matrix for the predictions. Approximately 8% of males, corresponding to 6 out of 77 males, were misclassified as females, and 6% of females, corresponding to 5 out of 83 females, were misclassified as males. The estimation of sex from 3D craniofacial scans is a trivial task and the few misclassified subjects are due to the model confusing the sex of subjects based on e.g., short hair for females and slightly feminine features in males.

3.3.2.2 Estimation of Age

For estimation of age, the network was trained for 190 epochs before early stopping occurred. A MAE of 7.77 years and a PCC of 0.76 was obtained between the true and predicted ages in the test set. If the network predicted the age for all subjects in the test set to be the mean value of the distribution, a MAE of 12.9 years would be obtained. Figure 3.11 shows a Bland-Altman plot of the true and predicted ages, i.e. the difference between the true and predicted ages as a function

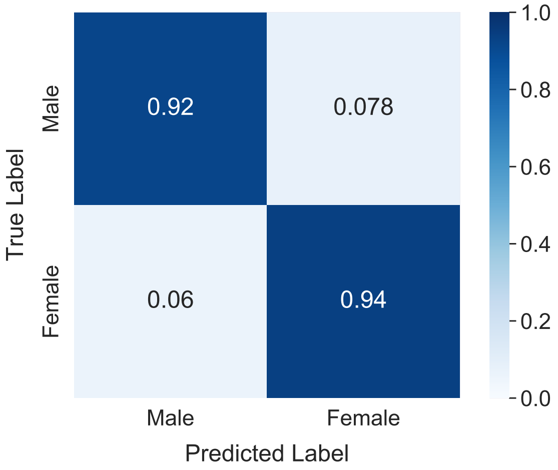


Figure 3.10: A normalized confusion matrix displaying the fraction of correctly and incorrectly classified subjects with respect to their sex for 160 subjects. The classification is performed based on sex of subjects that is predicted using a trained multi-view convolutional neural network, which takes as input eight pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted sex.

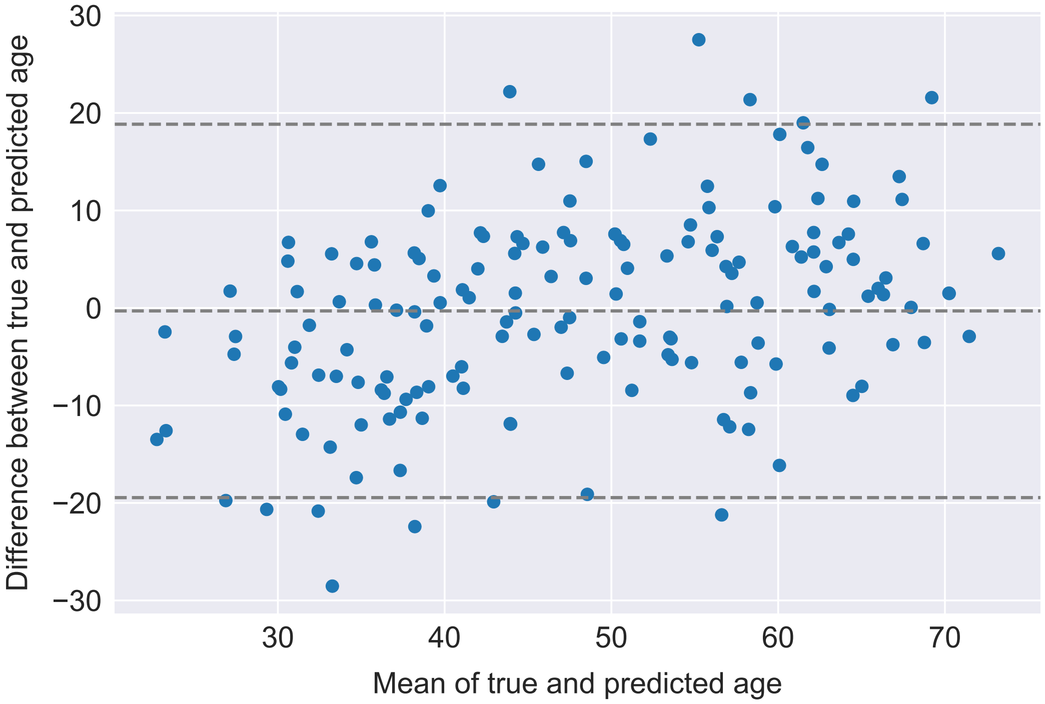


Figure 3.11: Bland-Altman plot for the true and predicted age in the test set consisting of 160 subjects. The predictions are performed using a trained multi-view convolutional neural network, which takes as input eight pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted age. The dashed horizontal lines above and below 0 indicate the limits of the 95% confidence interval.

of the mean of the true and predicted ages. There is a slight trend showing that lower ages are overpredicted and higher ages are underpredicted, and the greatest over and underpredictions are of a similar magnitude (approximately -30 and 30 years).

3.3.2.3 Estimation of BMI

For estimation of BMI, the network was trained for 196 epochs and achieved a MAE of 4.04 kg/m^2 and a PCC of 0.8 with respect to true and predicted BMIs in the test set. If the network predicted the BMI for all subjects in the test set to be the mean value of the distribution, a MAE of 6.9 kg/m^2 would be obtained. Figure 3.12 shows a Bland-Altman plot of the true and predicted BMI values. Small errors around the mean BMI of around 31 kg/m^2 are noted, while the larger BMI values are underpredicted.

The MAE of the age predictor being almost twice that of the BMI predictor makes sense, as in many cases it would be easier to derive someone's BMI based off of their face and neck as compared to age, particularly based on a 3D scan with limited facial details. The main uncertainty associated with prediction of BMI is that the height of the person is unknown, which is used for calculation of BMI, and which cannot be derived from a 3D craniofacial scan alone. Another drawback is that the

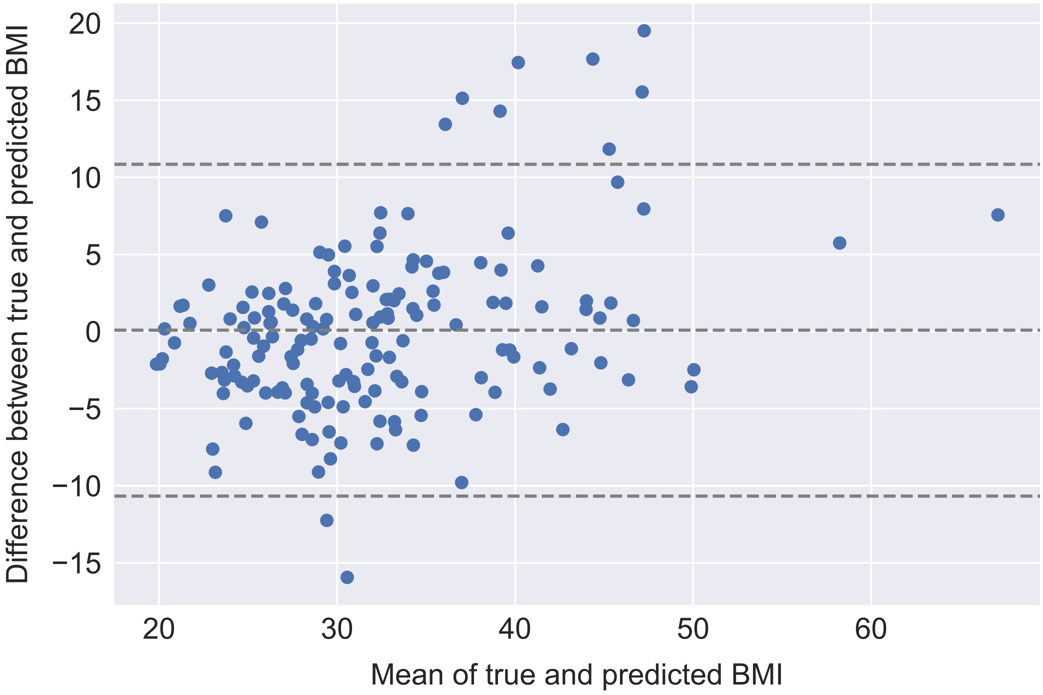


Figure 3.12: Bland-Altman plot for the true and predicted BMI in the test set consisting of 160 subjects. The predictions are performed using a trained multi-view convolutional neural network, which takes as input eight pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted age. The dashed horizontal lines above and below 0 indicate the limits of the 95% confidence interval.

BMI values in the dataset are heavily centered around the mean value of about 31 kg/m^2. Thus, the more extreme cases of either very low or high BMI values are underrepresented in the dataset.

Based on predicted BMI values, subjects were classified into BMI categories of normal, overweight, and obese [135], and Fig. 3.13 shows the resulting confusion matrix. In many cases, the BMI values have been overestimated. Several of the normal subjects are classified as overweight and many of the overweight subjects are classified as obese. However, even a small overestimation could lead to a misclassification, since there are hard cut-offs between each class, and most of the subjects have BMI values centered around the cut-off between overweight and obese. However, the highest sensitivity is obtained for the obese subjects and one could argue that they are the most important to capture, since they are the ones who are medically most at risk. Furthermore, accurate estimation of BMI from 3D craniofacial scans plays a vital role in OSA detection, as increased BMI is heavily associated with development of OSA [136].

3.3.2.4 Comparison to Similar Work

Similar work to a part of this study was presented in [137], where the authors used transfer learning with VGG-Face [138] on social media images of people to predict their BMI. They achieved an

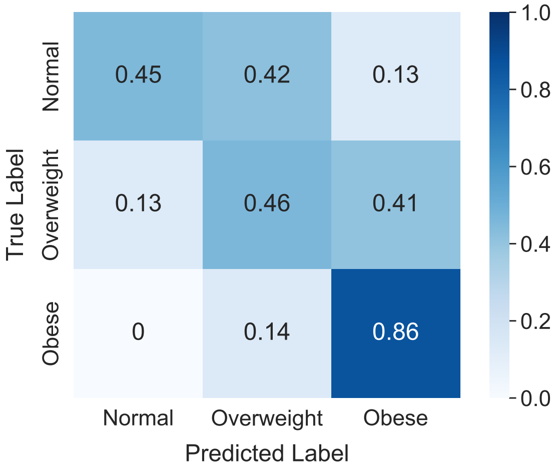


Figure 3.13: A normalized confusion matrix displaying the fraction of correctly and incorrectly classified subjects with respect to their BMI categories for 160 subjects. The classification is performed based on BMI values of subjects that are predicted using a trained multi-view convolutional neural network, which takes as input eight pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted BMI.

overall correlation coefficient of 0.65 compared to our correlation of 0.80. In [139], the authors extracted facial measurements as features instead and used simple regression methods to obtain an overall MAE of 3.14 kg/m^2 compared to our MAE of 4.04 kg/m^2. However, their work relied on hand-engineered features instead of the more data-driven feature extraction presented in this work. The work presented in [140] gives an extensive overview of different age predictors, showing MAEs ranging from 8.84 years to 0.31 years. Compared to our MAE of 7.77, this shows that there is clearly room for improvement with respect to age prediction. However, it must be noted that widely different techniques and datasets are being compared, so one should be careful to place too much emphasis on the comparison. Finally, [141] predicted the sex of subjects based on different deep learning architectures and their best model achieved an accuracy of 93.57%, which is very comparable to the 93% obtained in this work. Although our focus was not to achieve state-of-the-art performance on predicting any demographic, the comparison to other studies still serves as a validation of the proposed framework.

3.3.3 New Analyses After Publication

After publishing Paper I, which the research described above is based on, and working on Paper II, which will be described in Section 3.4, some modifications were made to the methodology. These modifications included 1) using 2D images and depth maps from five angles instead of eight, 2) using textures with the scans, and 3) using 10-fold cross validation instead of a training, validation, and test set split to train and evaluate the proposed models. The reasoning for these modifications are explained in Section 3.4.

The revised approach for Paper II was explored for the problem posed in Paper I, primarily to evaluate performance across the entire dataset instead of on a small test set and to be able to compare results from Paper I and Paper II when trained and evaluated on the same 1366 subjects with the same approach. Consequently, the results have more power and any biases that could be introduced in the test set when using a training, validation, and test set split would be diminished. Furthermore, for the research for Paper II, three sleep medicine specialists were recruited to estimate AHI based on the 3D craniofacial scans as described in Section 3.4.1.7. As part of this experiment, they also guessed the age of each subject. Consequently, a comparison between the age predictions of the proposed model and the sleep medicine specialists will also be included. All results from the new analyses performed after publication are presented below.

3.3.3.1 Estimation of Sex

The accuracy for estimation of sex was 91 ± 3%, which was almost the same as before (93%). Figure 3.14 shows the normalized confusion matrix for the predictions. The sensitivity of 92% for males and 90% for females was again almost the same as before (94% and 92%, respectively), showing that there was no bias with respect to sex in the test set consisting of 160 subjects. However, this was expected since classifying the sex of subjects from their craniofacial scans is a trivial task.

3.3.3.2 Estimation of Age

The MAE for estimation of age was 8.77 ± 0.53 years, which is exactly one year higher on average than before (MAE of 7.77 years). The PCC was 0.68 ± 0.06 which was lower compared to the PCC

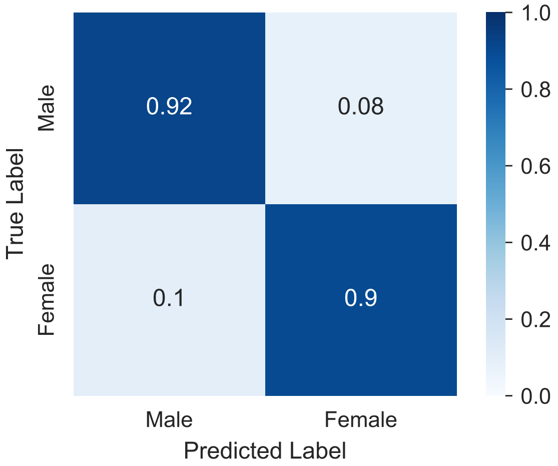


Figure 3.14: A normalized confusion matrix displaying the fraction of correctly and incorrectly classified subjects with respect to their sex for 1366 subjects. The classification is performed based on the sex of subjects that is predicted using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted sex.

of 0.76 obtained before. Figure 3.15 shows a Bland-Altman plot of the true and predicted ages for all 1366 subjects. Comparing Fig. 3.15 to the Bland-Altman plot for 160 subjects given in Fig. 3.11 shows an extremely similar overall trend while the greatest underprediction and overprediction appears to have increased by 5 years in either direction. The results show that the slightly poorer performance for the entire dataset consisting of 1366 subjects provides a better representation of performance for age estimation based on 3D craniofacial scans than the smaller test set consisting of 160 subjects.

The three sleep medicine specialists recruited to estimate AHI from 3D craniofacial scans (described in Section 3.4.1.7) also guessed the age of each subject. They achieved a MAE of 7.34 years and a PCC of 0.82, showing that the model estimating age does not perform as well as humans. Figure 3.16 compares Bland-Altman plots between the model's age predictions and the sleep specialists' age predictions, showing that there is less bias present for the specialists, particularly with respect to overprediction of age, as the model tends to overestimate age more often than the specialists, while the underpredictions appear more similar for both.

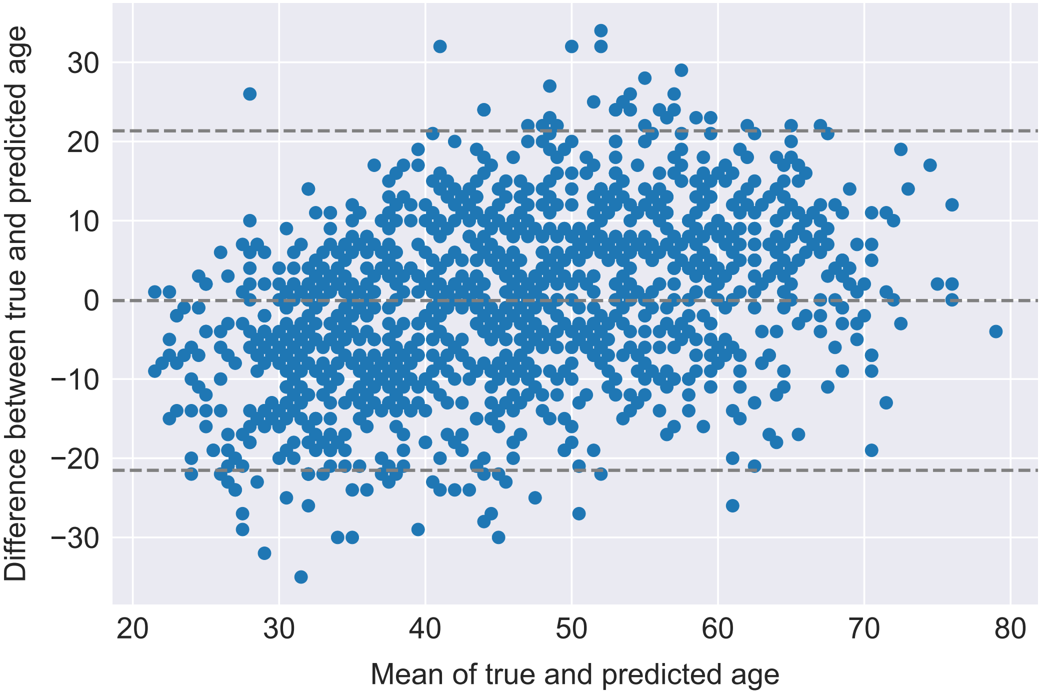


Figure 3.15: Bland-Altman plot for the true and predicted age in the dataset consisting of 1366 subjects. The predictions are performed using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted age. The dashed horizontal lines above and below 0 indicate the limits of the 95% confidence interval.

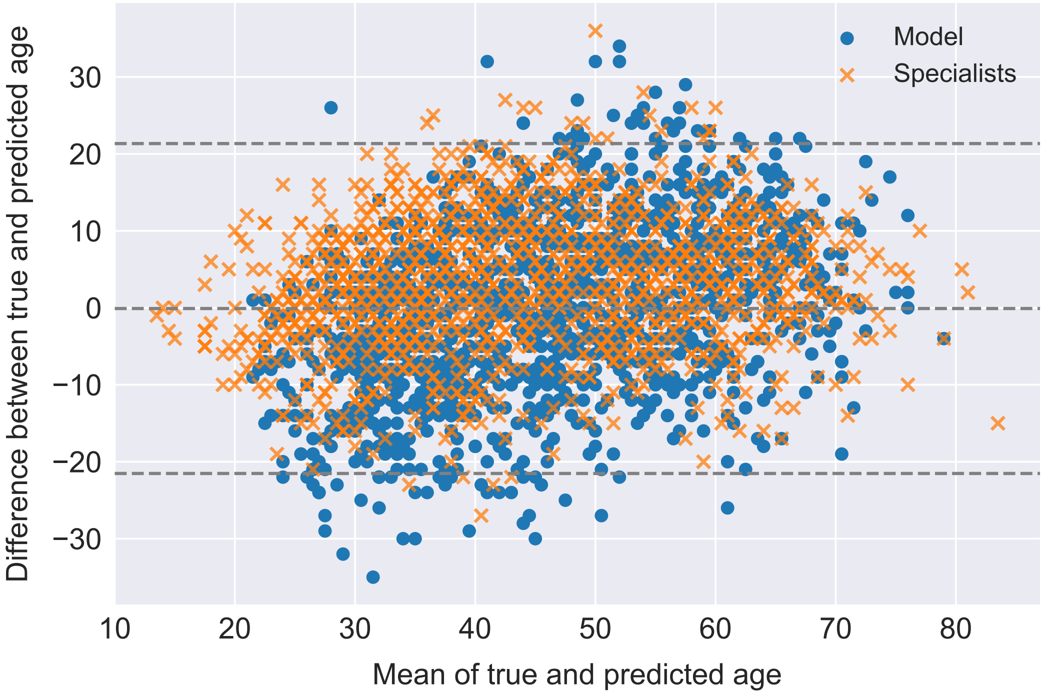


Figure 3.16: Comparison of Bland-Altman plots for age predicted by the proposed model (blue) and by three sleep medicine specialists (orange) evaluated for the entire dataset consisting of 1366 subjects. The age predictions by the proposed model are performed using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted age. The age predictions by the sleep specialists are performed by inspecting 3D craniofacial scans of all subjects and guessing their age. The dashed horizontal lines above and below 0 indicate the limits of the 95% confidence interval.

3.3.3.3 Estimation of BMI

The MAE for estimation of BMI was 4.36 ± 0.29 kg/m^2, which is slightly higher than before (4.04 kg/m^2). The PCC was 0.73 ± 0.05, which was lower compared to the PCC of 0.80 obtained initially. Figure 3.17 shows a Bland-Altman plot of the true and predicted BMIs for all 1366 subjects. The overall trend looks similar to the Bland-Altman plot presented in Fig. 3.12, however the greatest underprediction and overprediction appears to have increased by 10 kg/m^2 in each direction, which is most likely a result of including more subjects with abnormally high BMI to evaluate on.

Figure 3.18 shows the confusion matrix for predicted BMI categories for all 1366 subjects. Comparing this to Fig. 3.13, it is noticed that the sensitivity for normal subjects with respect to BMI decreases slightly (from 45% to 43%), the sensitivity for overweight subjects increases (from 46% to 57%), while the sensitivity for obese subjects decreases (from 86% to 74%). The main difference appears to be that overweight subjects are confused less with obese subjects and that obese subjects are confused more with overweight subjects, again most likely because more subjects with abnormal BMI are included in the analysis.

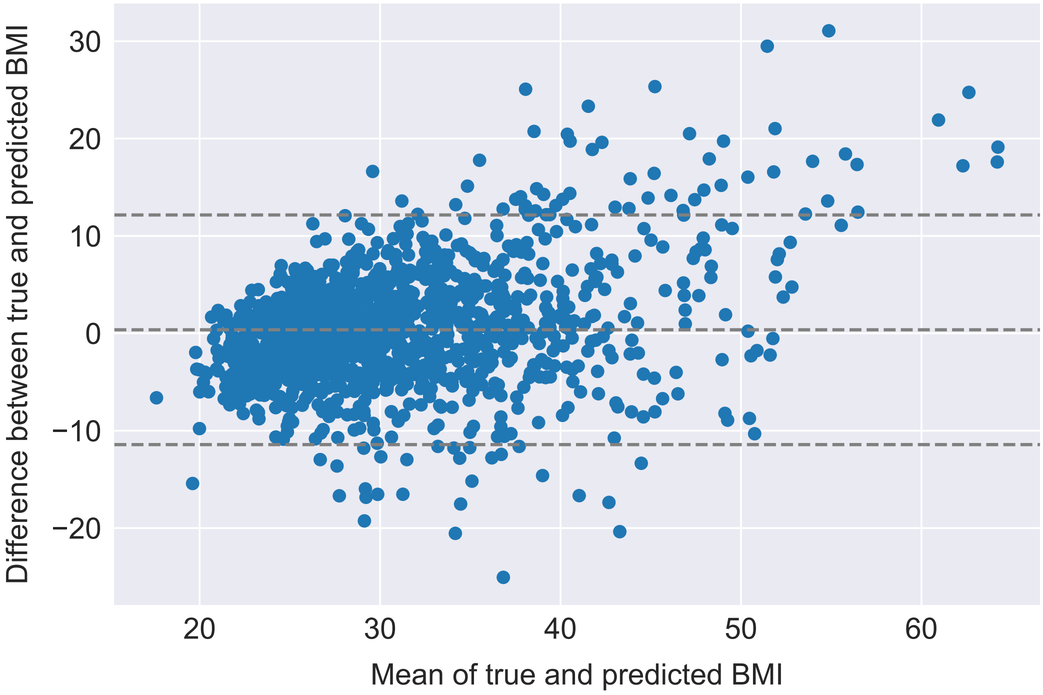


Figure 3.17: Bland-Altman plot for the true and predicted BMI in the dataset consisting of 1366 subjects. The predictions are performed using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted BMI. The dashed horizontal lines above and below 0 indicate the limits of the 95% confidence interval.

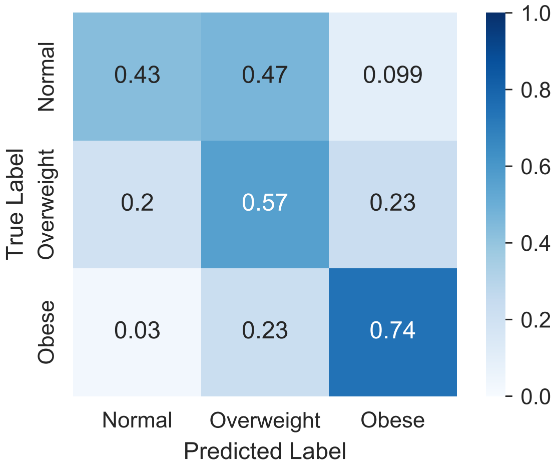


Figure 3.18: A normalized confusion matrix displaying the fraction of correctly and incorrectly classified subjects with respect to their BMI categories for 1366 subjects. The classification is performed based on BMI values of subjects that are predicted using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted BMI.

3.4 Paper II: Estimation of Apnea-Hypopnea Index Using Deep Learning on 3D Craniofacial Scans

Abstract

Purpose: Obstructive sleep apnea (OSA) is characterized by decreased breathing events that occur through the night, with severity reported as the apnea-hypopnea index (AHI), which is associated with certain craniofacial features. In this study, we used data from 1366 patients collected as part of Stanford Technology Analytics and Genomics in Sleep (STAGES) across 11 US and Canadian sleep clinics and analyzed 3D craniofacial scans with the goal of predicting AHI, as measured using gold standard nocturnal polysomnography (PSG).

Methods: First, the algorithm detects pre-specified landmarks on mesh objects and aligns scans in 3D space. Subsequently, 2D images and depth maps are generated by rendering and rotating scans by 45-degree increments. Resulting images were stacked as channels and used as input to multi-view convolutional neural networks, which were trained and validated in a supervised manner to predict AHI values derived from PSGs.

Results: The proposed model achieved a mean absolute error of 11.38 events/hour, a Pearson correlation coefficient of 0.4, and accuracy for predicting OSA of 67% using 10-fold cross-validation. The model improved further by adding patient demographics and variables from questionnaires. We also show that the model performed at the level of three sleep medicine specialists, who used clinical experience to predict AHI based on 3D scan displays. Finally, we created topographic displays of the most important facial features used by the model to predict AHI, showing importance of the neck and chin area.

Conclusion: The proposed algorithm has potential to serve as an inexpensive and efficient screening tool for individuals with suspected OSA.

3.4.1 Methods

This section describes the approach for developing an automatic screening system for OSA based on 3D craniofacial scans. An additional system using 3D scans, demographics, and questionnaire variables to predict AHI was proposed as well. Figure 3.19 shows a block diagram of both systems. The different components of the systems will be described in detail in the following.

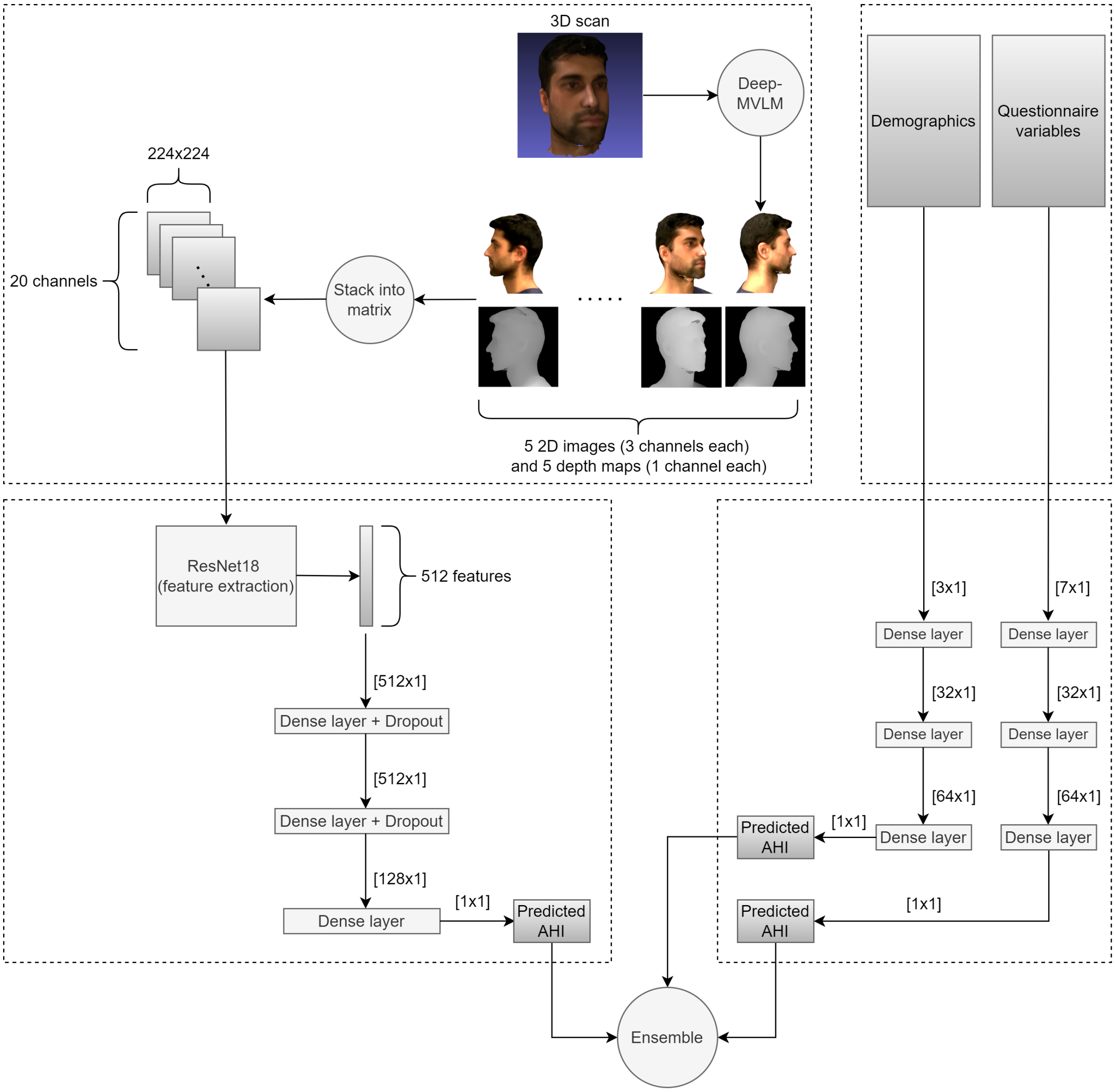


Figure 3.19: A block diagram of the proposed system for estimating apnea-hypopnea index (AHI) from 3D craniofacial scans, demographics, and questionnaires. Top left block: A 3D craniofacial scan is converted to 2D images and depth maps captured from 45-degree increments around the 3D scan using the multi-view consensus convolutional neural network (CNN) for 3D facial landmark placement (Deep-MVLM) algorithm [127]. These images are normalized to the range [0,1] and stacked into a matrix, yielding matrix dimensions of 20x224x224. Bottom left block: The matrix containing craniofacial images is used as input for a CNN with a ResNet18 architecture for feature selection. The output of the CNN is a 512x1 feature vector, which is processed by two dense layers with dropout (one layer preserving the 512 features and the next reducing 512 features to 128). A final dense layer transforms the output into a scalar value for the prediction of AHI. Top right block: Patient demographics (sex, age, and BMI) and questionnaire variables (seven OSA-related questions that the person answers yes or no to) are extracted. Age and BMI values are normalized to the range [0,1] using min-max normalization. Bottom right block: The demographics and questionnaire variables are processed by two separate multi-layer perceptrons (MLPs), which consist of three dense layers each. The first two layers increase the number of features (to 32 and then 64), and the final layer transforms the output into a scalar value for the prediction of AHI. The AHI values predicted using the craniofacial scans, demographics, and questionnaires are ensembled by averaging all three values. The model using only 3D craniofacial scans to predict AHI relies on the top left and bottom left blocks and skips the top right and bottom right blocks entirely.

3.4.1.1 Data Collection

The data collection for this study was the same as for the previous study described in Section 3.3.1.1.

3.4.1.2 Data Description

A detailed description of the 3D craniofacial scans collected for this study, including how the data is structured, was provided in Section 3.3.1.2.

Out of 1756 subjects for whom 3D craniofacial scans were collected, 1366 subjects were included in this study as the rest had missing demographics, questionnaires, and/or PSGs. Of the 1366 subjects, 724 were female and 642 were male. Mean age ± standard deviation was 45.9 ± 14.8 years, BMI was 30.9 ± 8.7 kg/m^2, and AHI was 15.5 ± 19.3 events/hour (median: 9.3, IQR: 17.5). The distributions of sex, age, and BMI were shown in Fig. 3.6 for 1605 subjects in the dataset. The distributions of demographics for the 1366 subjects used for this study are not included as they are approximately the same as in Fig. 3.6. Figure 3.20 shows the distribution of AHI values within the dataset. AHI was derived from each PSG annotation file by using the following formula:

$$AHI = \frac{N_{OSA} + N_{HYP}}{TST} \cdot 60, \tag{3.5}$$

where N_{OSA} are the number of annotated obstructive apneas and N_{HYP} are the number of annotated hypopneas. TST is the total sleep time in minutes, which was calculated as:

$$TST = \frac{(E_{N1} + E_{N2} + E_{N3} + E_{REM}) \cdot 30 \text{ seconds}}{60}, \tag{3.6}$$

where E_{N1}, E_{N2}, E_{N3}, and E_{REM} are the number of annotated epochs (of 30 seconds) for each sleep stage, respectively. TST represents the amount of time in minutes the subject spent asleep during the PSG recording. TST is used instead of the recording time for AHI calculation because apneas and hypopneas only occur during sleep. Central apneas were excluded from the analysis because they have no known relation to craniofacial anatomy. Central hypopneas are very rare so it was assumed that all annotated hypopneas were obstructive.

3.4.1.3 Pre-Processing

The pre-processing steps for this study were the same as detailed in Section 3.3.1.3 for the previous study. However, at this time in the PhD project, it was discovered that some scans were not complete with respect to the back of the head. Since this information does not contribute to AHI

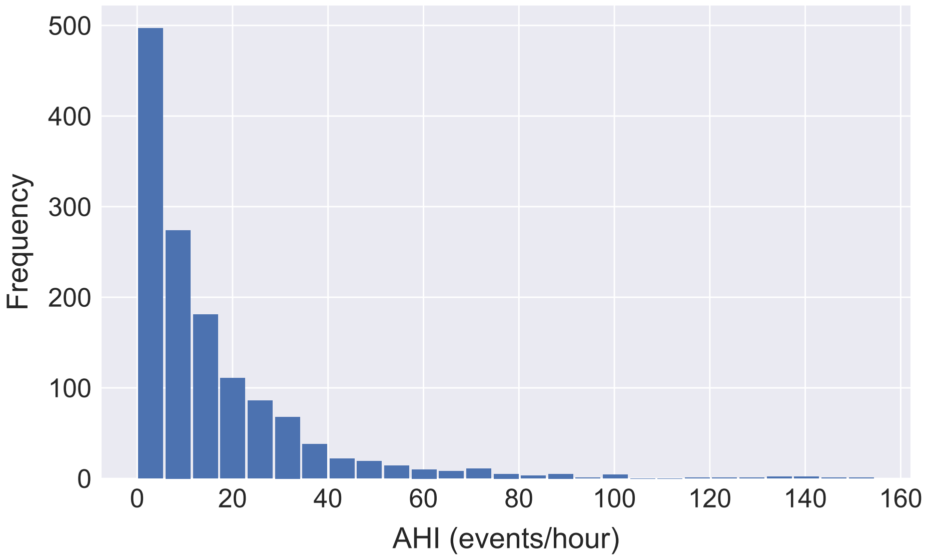


Figure 3.20: Distribution of apnea-hypopnea index (AHI) values for 1366 subjects in the dataset used to train and evaluate a multi-view convolutional neural network for the estimation of AHI from 3D craniofacial scans.

prediction, only five pairs of 2D images and depth maps were used, emphasizing the frontal and profile characteristics of face and neck from several angles as shown in Fig. 3.21. Note that textures for the scans were included in this study as opposed to the previous study since computational cost was not a concern this time around. The 2D images had three channels each (RGB) and the depth maps had one channel each. Since there were five 2D images and five depth maps, the final stacked matrix per subject contained 5×3 + 5×1 = 20 channels in total and 224×224 pixels per channel.

The patient demographics (age and BMI), which were also included as features for one model, were normalized using min-max normalization given by

$$\mathbf{x}_{norm} = \frac{\mathbf{x} - min(\mathbf{x}_{train})}{max(\mathbf{x}_{train}) - min(\mathbf{x}_{train})}, \tag{3.7}$$

where $\mathbf{x}$ is a vector containing one of the demographics for all subjects in the dataset and $\mathbf{x}_{train}$ is a vector containing one of the demographics for all subjects in the training set. This normalization was performed to ensure faster convergence during training of the proposed model. No normalization was necessary for sex because it was binary (0: female, 1: male).

The seven different variables from the modified STOP-Bang questionnaire described in Section 4.5.1 were also included as features for one model and they did not require normalization either because those values were also binary (0: no, 1: yes). The complete STOP-Bang questionnaire is provided in Appendix A. The modified STOP-Bang questionnaire used for this study did not include neck circumference of subjects but was otherwise identical to the one in Appendix A.

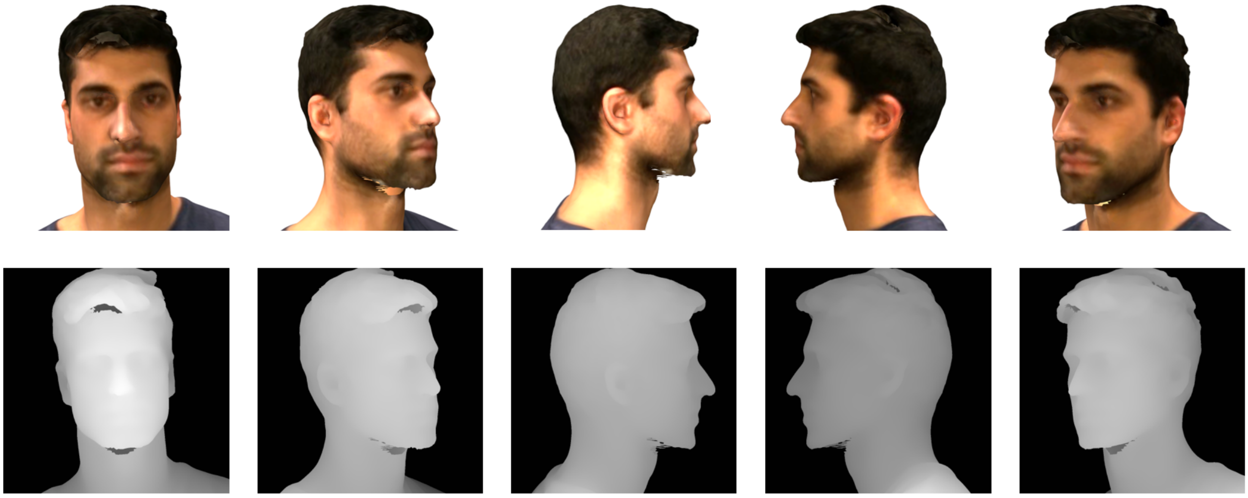


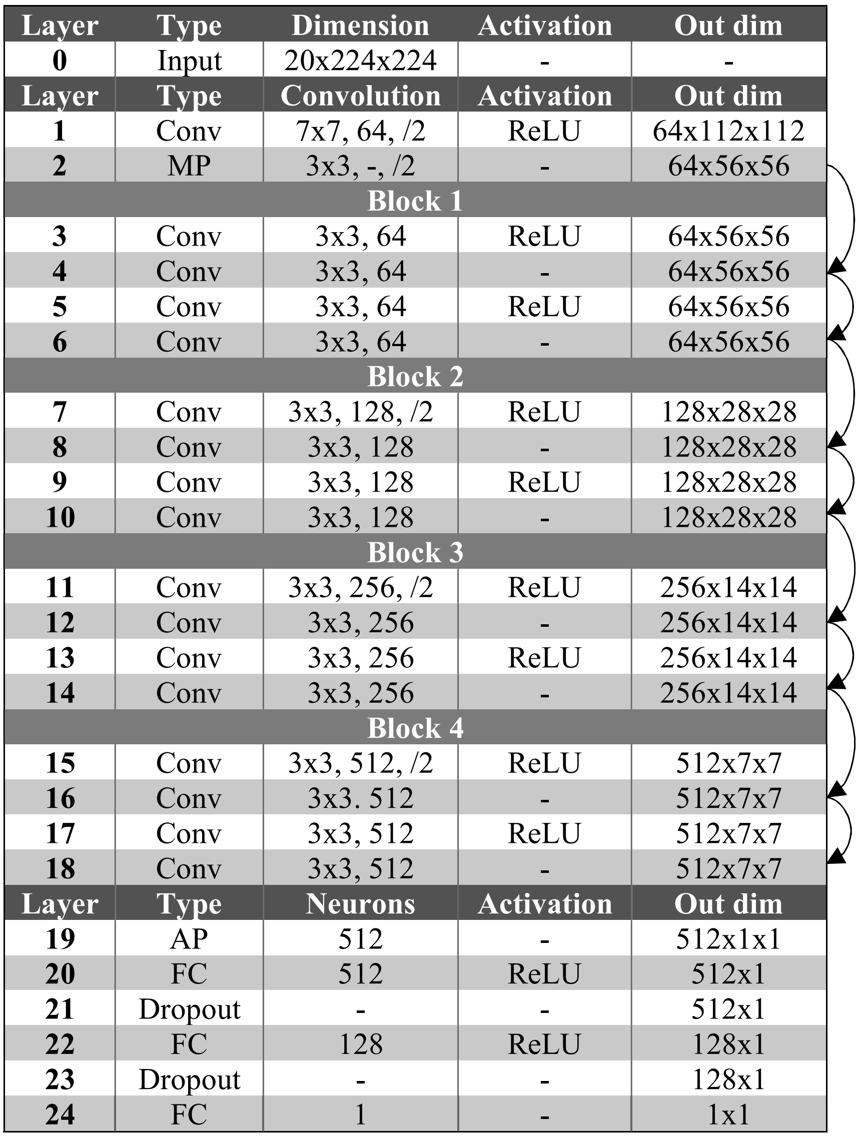
Figure 3.21: Example of five pairs of 2D images (top row) and depth maps (bottom row) captured at different angles from a 3D craniofacial scan and used as input for the multi-view convolutional neural network, which is trained to estimate the apnea-hypopnea index of a subject.

3.4.1.4 Multi-View Convolutional Neural Network

The purpose of applying machine learning was to reveal data-driven mapping differences within the multi-view inputs across AHI values. For this purpose, we implemented a multi-view CNN with a ResNet18 architecture [130]. The network architecture was almost identical to the one presented in Section 3.3.1.4, where reasoning for the choice of network architecture was provided as well, except that an extra fully connected layer followed by dropout was added to this network due to the increased complexity of the task. Figure 3.22 shows the network architecture with all details.

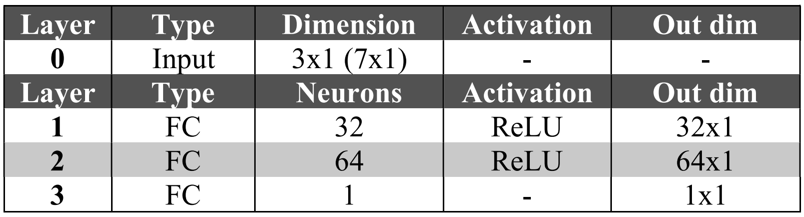
The implemented multi-view CNN took 20 channels as input per subject and the input was processed by the standard blocks of a ResNet18 network. Two additional fully connected layers were added at the end, one preserving the 512 features and the other reducing 512 features to 128. Both fully connected layers were followed by a ReLU activation and dropout (probability of 0.3 and 0.5, respectively). Dropout was added as a regularizing component to reduce overfitting during training. The dropout probability and number of fully connected layers and neurons were selected based on hyperparameter tuning, which was performed in a grid search-like manner where the hyperparameters were varied and different combinations of these were investigated. The optimal hyperparameters were the ones which yielded the lowest error on the validation set. The output layer consisted of a single neuron, since the desired output was a single continuous AHI value.

Two additional networks were developed, one using demographics as input to predict AHI and another using questionnaire variables as input to predict AHI. The first network took three input features (sex, age, and BMI), while the second network took seven input features (seven questionnaire variables). Both networks were multilayer perceptrons (MLPs) with two fully connected layers followed by ReLU activation functions and a final output layer consisting of one neuron for the predicted AHI value. Figure 3.23 shows the network architecture with all details for estimation of AHI based on demographics and questionnaires. MLPs were selected because they are more suitable when the input is a feature vector with independent features as compared to a CNN.



Layer	**Type**	**Dimension**	**Activation**	**Out dim**
0	Input	20x224x224	-	-
Layer	**Type**	**Convolution**	**Activation**	**Out dim**
1	Conv	7x7, 64, /2	ReLU	64x112x112
2	MP	3x3, -, /2	-	64x56x56
		Block 1		
3	Conv	3x3, 64	ReLU	64x56x56
4	Conv	3x3, 64	-	64x56x56
5	Conv	3x3, 64	ReLU	64x56x56
6	Conv	3x3, 64	-	64x56x56
		Block 2		
7	Conv	3x3, 128, /2	ReLU	128x28x28
8	Conv	3x3, 128	-	128x28x28
9	Conv	3x3, 128	ReLU	128x28x28
10	Conv	3x3, 128	-	128x28x28
		Block 3		
11	Conv	3x3, 256, /2	ReLU	256x14x14
12	Conv	3x3, 256	-	256x14x14
13	Conv	3x3, 256	ReLU	256x14x14
14	Conv	3x3, 256	-	256x14x14
		Block 4		
15	Conv	3x3, 512, /2	ReLU	512x7x7
16	Conv	3x3. 512	-	512x7x7
17	Conv	3x3, 512	ReLU	512x7x7
18	Conv	3x3, 512	-	512x7x7
Layer	**Type**	**Neurons**	**Activation**	**Out dim**
19	AP	512	-	512x1x1
20	FC	512	ReLU	512x1
21	Dropout	-	-	512x1
22	FC	128	ReLU	128x1
23	Dropout	-	-	128x1
24	FC	1	-	1x1

Figure 3.22: The applied multi-view convolutional neural network architecture for predicting the apnea-hypopnea index of a subject based on 20-dimensional input craniofacial images derived from a 3D scan. The input dimensions are given by number of channels x height x width. The convolution layers are specified by filter size (e.g., 3x3), number of channels (e.g., 64), and a stride (e.g., /2). The same applies for the max pooling (MP) layer. The output dimensions of the feature maps are given by number of channels x height x width. The convolution layers are always followed by batch normalization in this architecture. The dropout layers have keep probabilities of 0.3 and 0.5, respectively. If the skip connections (arrows) are applied, the feature maps are down sampled instead by applying 1x1 filters with a stride of 2x2. AP - Average pooling, FC - Fully connected.



Layer	Type	Dimension	Activation	Out dim
0	Input	3x1 (7x1)	-	-
Layer	**Type**	**Neurons**	**Activation**	**Out dim**
1	FC	32	ReLU	32x1
2	FC	64	ReLU	64x1
3	FC	1	-	1x1

Figure 3.23: The applied multi-layer perceptron (MLP) architecture for predicting the apnea-hypopnea index of a subject based on the 3-dimensional input demographics (sex, age, and BMI) or the 7-dimensional input questionnaire variables (snoring, tiredness, observed apnea, hypertension, being male, being older than 65 years, and having a BMI greater than 35 kg/m^2), which is indicated by parantheses in the figure. FC - Fully connected.

The AHI predictions of each of the three different networks, using scans, demographics, and questionnaires as inputs, respectively, were averaged using an ensemble approach to form final predictions for AHI values as shown in Fig. 3.19. Thus, two different models were proposed in this study: one predicting AHI based only on 3D craniofacial scans and another predicting AHI based on 3D craniofacial scans, demographics and questionnaire variables combined.

3.4.1.5 Training, Validation, and Testing

Training, validation, and testing was carried out using 10-fold cross-validation. This was done by splitting the dataset into 10 folds of equal size and utilizing 8 folds for training and 1 fold each for validation and testing. Predictions for the test fold samples were computed after training and the process was repeated by assigning new folds to training, validation, and testing. This procedure was repeated 10 times, such that AHI predictions were performed for all subjects in the dataset.

Mean squared error (MSE) was used as loss function given by:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{3.8}$$

where y_i is the true AHI value, $\hat{y}_i$ is the predicted AHI value, and N is the number of samples in the training set. The MSE was chosen to penalize greater errors, since most of the AHI values in the dataset were in the range 0-15 events/hour (Fig. 3.20), and it was important for the model to also learn to estimate AHI values for subjects with severe OSA (AHI $>$ 30 events/hour).

The learning rate was set to $1 \cdot 10^{-5}$ with a weight decay of $5 \cdot 10^{-4}$ for the multi-view CNN, and $1 \cdot 10^{-2}$ (with the same weight decay) for the MLPs. The learning rate was chosen using hyperparameter tuning and the optimal choice was a learning rate that is not too high, which can cause the parameter update via gradient descent to diverge from the minima, and not too low, which would slow down training significantly. The weight decay was used to add a penalty term to the loss function during optimization, which shrinks the weights and helps to prevent overfitting.

The batch size was set to 8 for all three networks and was limited by computational resources. The Adam optimizer [133] was used for optimization of the network due to its superior performance compared to other optimization algorithms. Early stopping was applied when the validation error did not decrease for 3 consecutive epochs (patience of 3) to help prevent overfitting. Python 3.7.4 and Pytorch 1.3.1 were used for pre-processing and deep learning purposes.

Training was carried out on a GeForce RTX 2080 and the entire training, validation, and test setup took approximately one hour to complete for the model using 3D cranofacial scans and approximately fifteen minutes for the models using demographics and questionnaires, respectively.

3.4.1.6 Performance Measures

The measures used to evaluate model performance with respect to true and predicted AHI values in the test set were mean absolute error (MAE) and Pearson Correlation Coefficient (PCC), which are given by Eqs. 3.1 and 3.3.

Furthermore, predicted AHI was used to classify subjects into being normal or having OSA, where an AHI of 15 events/hour or greater was the clinical criterion used for defining the presence of OSA (moderate-severe versus mild or no OSA). Thus, accuracy of classification was another measure used to evaluate the model performance, which was calculated using Eq. 3.4. In this case, the positives were defined as subjects with OSA, so TP denoted the number of subjects with OSA correctly classified as having OSA and TN denoted the number of subjects without OSA correctly classified as being normal. FP represented the number of normal subjects incorrectly classified as having OSA, while FN was the number of subjects with OSA incorrectly classified as normal.

The area under receiver operating characteristic curve (AUC ROC) was used as another performance measure for the classification. The ROC curve plots the true positive rate against the false positive rate for all classification thresholds, and the AUC ROC provides an aggregated metric for model performance in distinguishing between the positive and negative class across all thresholds.

Bland-Altman plots [134] were used to illustrate the patterns of disagreement between true and predicted AHI values, while a confusion matrix was displayed for the classification of normal/OSA to show the fraction of misclassified subjects in both categories as well as the sensitivity of the model in classifying each category.

3.4.1.7 Sleep Specialists' Ability to Estimate Apnea-Hypopnea Index

Three experienced, board certified sleep medicine physicians with in-depth knowledge of OSA were recruited to imitate the task of the proposed model, i.e., estimating AHI based on inspection of the 3D craniofacial scan of each subject. The three physicians scored one third of the dataset each, while also annotating 150 of the same scans to estimate percentage agreement. When annotating a scan, each physician was shown the scan from all desired angles, having the ability to rotate the 3D image for any desired amount of time. Physicians took approximately 30 seconds to score each scan.

Their first thought was to size up the person based on their estimated age, with the awareness that older individuals often have higher AHIs (due to lax musculature, redundant tissue, and an atrophic skeletal scafolding that result in higher risk of airway collapse). Then they would ascertain if the individual looked tired - droopy eyelids (ptosis), drawn face, pallor, circles/bags under the eyes, etc. - that might suggest an underlying sleep disorder. Finally, they would look for some of the high-yield characteristics: looking at the overall head and neck adiposity (fat), the characteristics of the thyromental space, the over/under-bite (to suggest a retrognathic jaw that pushes the tongue into the airway), the craniofacial complex (looking for maxillary or mandibular hypoplasia) and noting whether there was a long/thin face suggestive of life-long nasal congestion ("adenoid facies"), and the cervical lordosis (to see if subjects have their heads thrust forward, suggestive of position modification in order to ease breathing). They would then roughly estimate the AHI based on all these factors and their general knowledge of the known prevalence/proportions of varying severities of OSA.

It is important to note that clinicians do not traditionally estimate AHI, but for this study their estimates served as expert level performance as a comparison for the proposed model.

3.4.1.8 Topographic Display of Important Craniofacial Regions

A topographic display was created by generating saliency maps for the model using craniofacial scans. A saliency map is a visualization technique based on the gradient of the network output with respect to an input image [142]. Consequently, the pixels which contribute most to the prediction of the network can be highlighted. For the topographic displays, we averaged saliency maps for 10 subjects with the highest predicted AHI values per cross-validation test fold and 10 subjects with the lowest predicted AHI values per test fold, yielding an average of 100 saliency maps for the highest and lowest predicted AHI values, respectively.

3.4.2 Results and Discussion

The model using craniofacial scans during cross-validation converged after 6.6 ± 1.6 epochs, the model using demographics converged after 23.1 ± 6.9 epochs, and the model using questionnaires converged after 12.5 ± 6.9 epochs. This was calculated by averaging the epochs it took for each model to converge across all 10 folds. The following results were obtained by evaluating the proposed models for all 1366 subjects in the dataset.

3.4.2.1 Model Using Craniofacial Scans

Without demographic and questionnaire information available, our model achieved a MAE of 11.38 ± 1.36 events/hour and a PCC of 0.40 ± 0.04 using 10-fold cross-validation. In comparison, if AHI for all subjects was predicted as the mean AHI value in the dataset (i.e., 15.5 events/hour), the MAE would be 13.0 events/hour and the PCC would be -0.02. This means that the average absolute deviation from true AHI per subject would be almost 2 events/hour more than the proposed model, although the correlation would be significantly worse.

Figure 3.24 shows the Bland-Altman plot of true and predicted AHI values (blue data points), where underpredictions of great magnitude are observed for subjects with very high AHI (>30 events/hour). In general, subjects with high AHI are hard to estimate because AHI values of 30, 60, or 90 events/hour are not significantly different, as they are all considered abnormal values. There were more than 200 subjects with AHI above 30 events/hour and 12 subjects with AHI more than 100 events/hour. Additionally, 50 subjects had an AHI of 0 events/hour, which were overpredicted on average by 10 events/hour. The subjects with very high AHI and AHI of 0 events/hour contribute to the high standard deviation (19.3) of AHI values in the dataset and increase the MAE as well.

When dividing subjects into normal/OSA using AHI $\geq$ 15 events/hour as a criterion for OSA, an overall accuracy of $67 \pm 4\%$ was obtained based on the predicted AHI. Figure 3.25 (a) shows the resulting confusion matrix of classifying subjects into normal or having OSA based on predicted AHI values, showing that the model is better at classifying normal subjects compared to subjects with OSA. Sensitivity was $59 \pm 8\%$, specificity was $72 \pm 5\%$, and AUC ROC was $65 \pm 4\%$. In comparison, if AHI for all subjects was predicted as the mean AHI value of the dataset, accuracy would be 34%, and AUC ROC would be 50%.

The misclassifications observed in Fig. 3.25 (a) were further explored to reveal any patterns associated with subjects wrongly classified as normal or having OSA. The top row of Fig. 3.26 shows scatter plots of true and predicted AHI of subjects, divided into males and females and color coded according to their BMI category, who were wrongly classified as being normal based on a predicted AHI < 15 events/hour. For the misclassified subjects, it is observed that males have higher predicted AHI than females, which is in accordance with the fact that males have higher OSA risk than females in general. Particularly for males, the predicted AHI for many subjects is around 14 events/hour, which means the model was close to classifying these subjects correctly, which would have increased sensitivity of the model. Importantly though, all predicted AHI values for males and most for females are above 5 events/hour, which means they are predicted to have mild OSA as opposed to having no OSA when using the OSA severity classification described in Section 2.3.4.

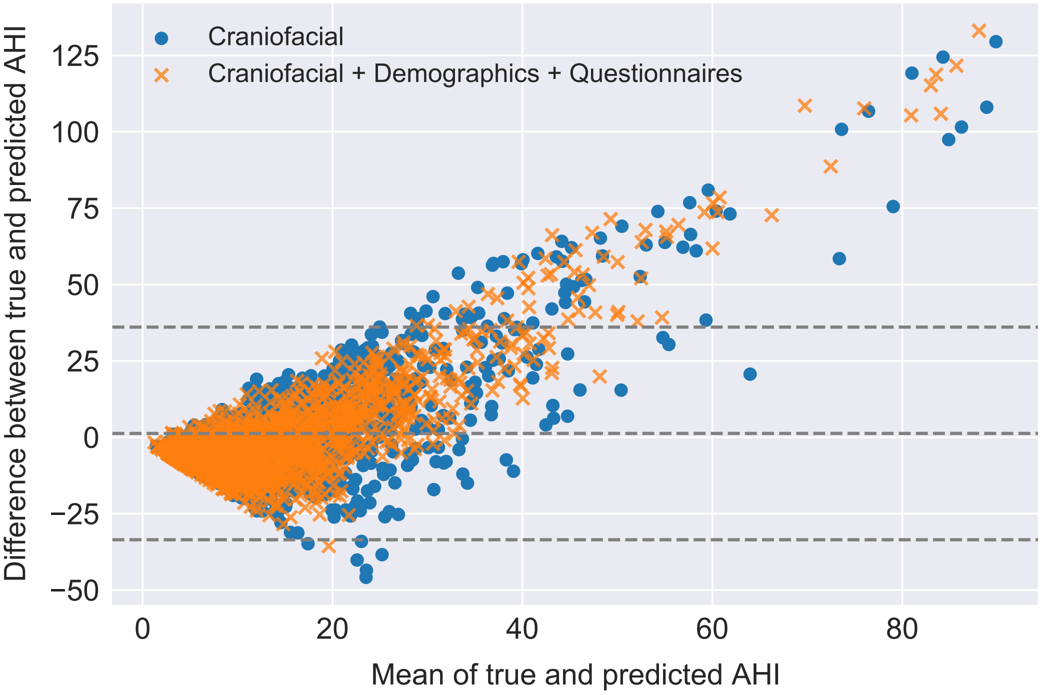


Figure 3.24: Bland-Altman plots of true and predicted apnea-hypopnea index (AHI) in the dataset consisting of 1366 subjects, predicted by the model using only craniofacial images (blue), and predicted by the model combining craniofacial images, demographics, and questionnaires (orange). Predictions for the first model are obtained using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted AHI. Predictions for the second model are obtained by ensembling three different models, using craniofacial images, demographics, and questionnaires as input, respectively, to output predicted AHI. The dashed horizontal lines above and below 0 indicate the 95% confidence interval limits.

The data points in the top left corner represent the subjects who have AHI close to 15 events/hour and for whom an AHI close to 15 events/hour was predicted. Although the AHI estimations were close, a somewhat arbitrary cut-off of 15 events/hour causes misclassifications and reduced sensitivity of the model with respect to OSA detection. Furthermore, the model is not aware of this cut-off for subsequent classification as it has purely been trained to estimate AHI. Although the cut-off for OSA detection can be useful for screening, it does not paint the complete picture. For example, there would not be a big difference between subjects with AHI of 15 and 14.4 events/hour, respectively, yet the cut-off differentiates the two and classifies one as having OSA and the other as being normal. A classification model could be implemented which performs the classification directly, but the benefit of AHI estimation for OSA severity would be lost in that case.

With respect to BMI categories, there is a mixture of normal, overweight, and obese males who are misclassified, with overweight males being slightly more represented. The overweight males are in the middle with respect to BMI categories and perhaps this causes the model to predict AHI values that are somewhat in the middle as well. A likely explanation for subjects with normal BMI

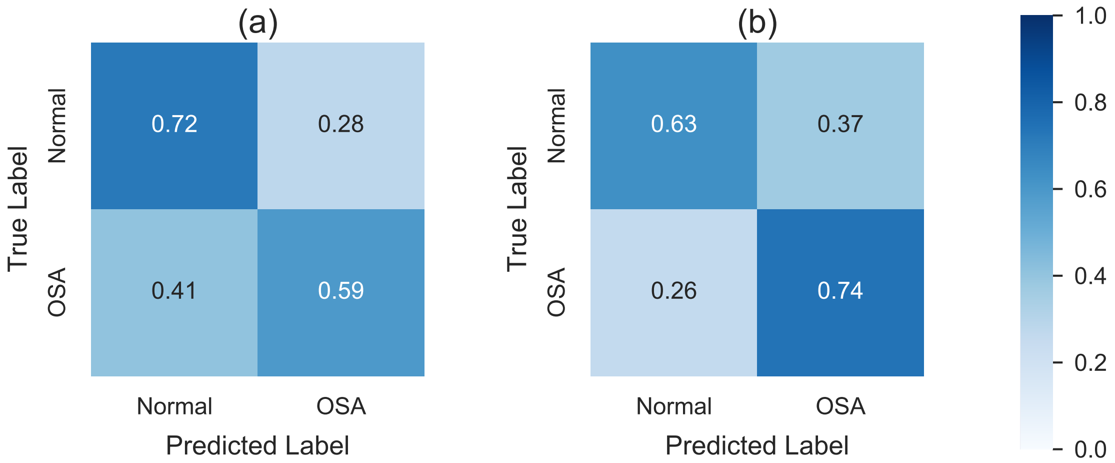


Figure 3.25: (a) Normalized confusion matrix for the model using only craniofacial images. (b) Normalized confusion matrix for the model using a combination of craniofacial images, demographics, and questionnaires. Predictions for the first model are obtained using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted apnea-hypopnea index (AHI). Predictions for the second model are obtained by ensembling three different models, taking craniofacial images, demographics, and questionnaires as input, respectively, to output predicted AHI. The confusion matrices show the results from classifying subjects into being normal or having obstructive sleep apnea (OSA) using AHI $\geq$ 15 as a criterion for OSA based on the predicted AHI values by the proposed models.

not being classified as having OSA is the presence of other factors contributing to OSA, which can not be assessed using craniofacial scans (such as narrow upper airway anatomy). As a result, the model underestimates their AHI based on the fact that they have normal BMI and no apparent craniofacial abnormalities. The males with most extreme AHI values are all overweight or obese, and it is curious that the model predicts low AHI for those. The majority of females who are wrongly classified as being normal are obese, which could be explained by the fact that women have much lower risk of OSA than men despite being obese. Most likely, the model has picked up on this discrimination based on all females in the dataset and thus underpredicts their AHI values to a degree where they are wrongly classified based on the cut-off value.

The top row of Fig. 3.27 shows scatter plots of true and predicted AHI of subjects, divided into males and females and color coded according to their BMI category, who were wrongly classified as having OSA based on a predicted AHI $\geq$ 15 events/hour. Again, the males have higher predicted AHI than females, and for males the overpredicted AHI values are mostly for overweight and obese subjects, while for females, the majority of misclassified subjects are obese. Compared to the top row of Fig. 3.26, much fewer subjects with normal BMI are observed in Fig. 3.27, showing that most overpredicted AHI values are associated with increased BMI. It is also noted that many of the misclassified subjects have AHI values close to 15 events/hour, demonstrating that the model was close to classifying these subjects correctly, which would have increased specificity of the model. This again highlights the shortcomings associated with using such cut-offs for OSA classification.

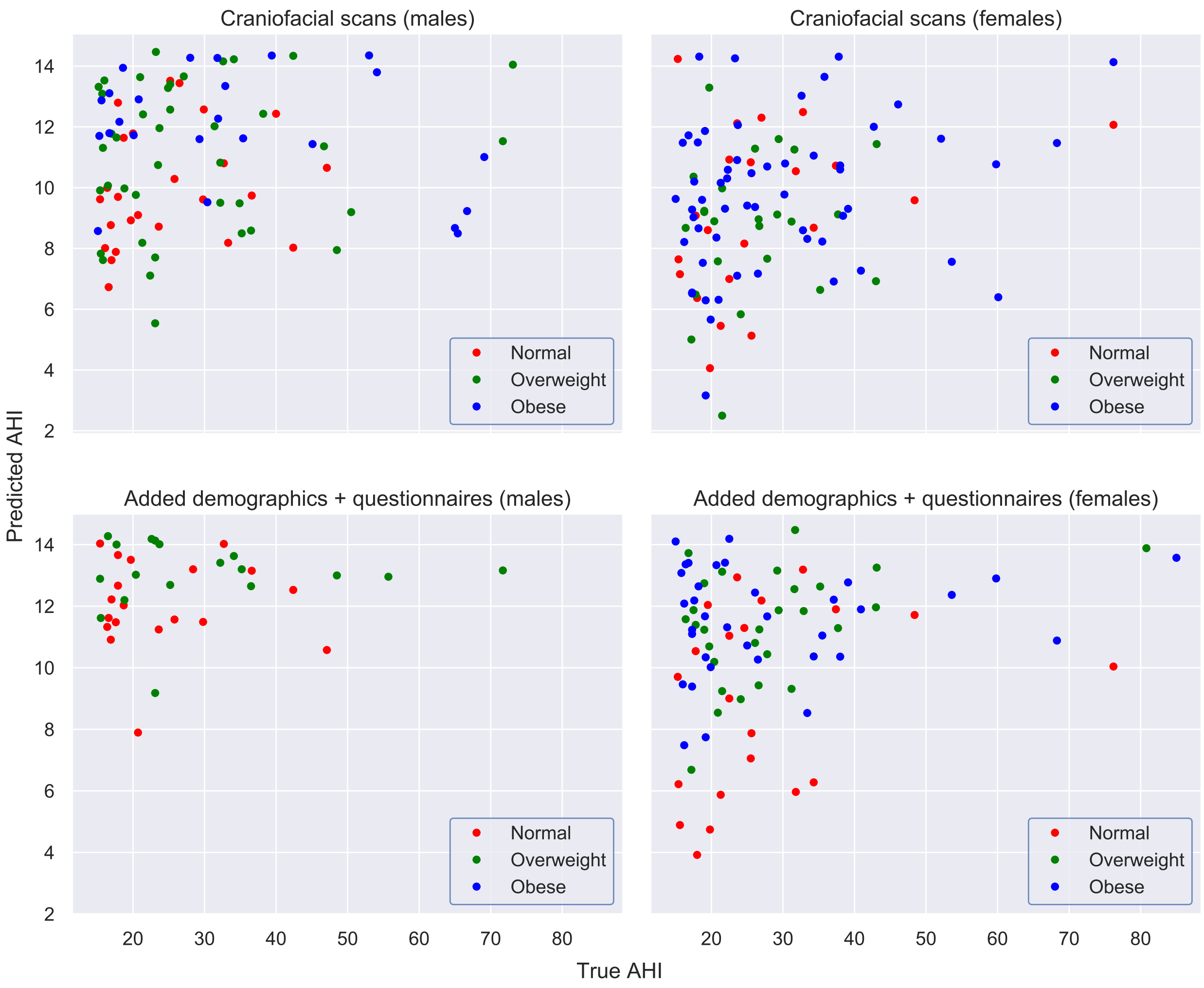


Figure 3.26: Scatter plots showing the true and predicted apnea-hypopnea index (AHI) values for subjects who were wrongly classified as being normal predicted by the model using only craniofacial images (top row) and predicted by the model using a combination of craniofacial images, demographics, and questionnaires (bottom row). The subjects were wrongly classified as being normal because the proposed model predicted an AHI value $<$ 15 events/hour. Separate figures are provided for males and females, and each subject is colored according to their BMI category as obesity is a big risk factor for developing obstructive sleep apnea.

3.4.2.2 Model Using Craniofacial Scans, Demographics, and Questionnaires

Adding the clinically relevant demographics and questionnaire variables improved the model further, yielding a MAE of 11.05$\pm$1.40 events/hour and a PCC of 0.45$\pm$0.04. Figure 3.24 shows the Bland-Altman plot of the true and predicted AHI values (orange data points). Note how the difference between true and predicted AHI is reduced in both directions compared to only using craniofacial images, which gives more accurate AHI estimations as evident from the reduction in MAE from 11.38 to 11.05 events/hour. However, the extreme AHI values are still underpredicted with a similar magnitude, highlighting that the model struggles with such subjects even when adding clinical data.

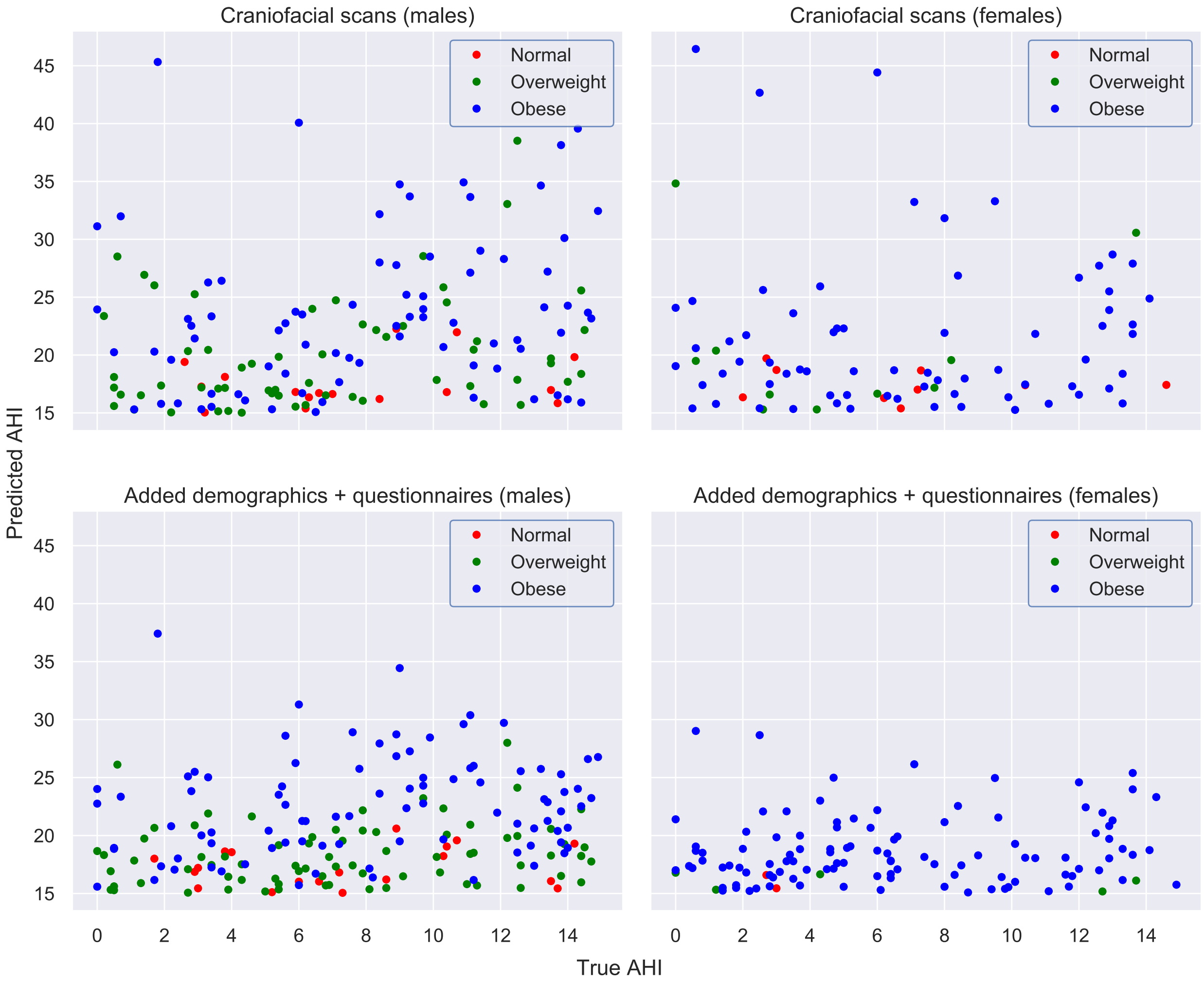


Figure 3.27: Scatter plots showing the true and predicted apnea-hypopnea index (AHI) values for subjects who were wrongly classified as having obstructive sleep apnea (OSA) predicted by the model using only craniofacial images (top row) and predicted by the model using a combination of craniofacial images, demographics, and questionnaires (bottom row). The subjects were wrongly classified as having OSA because the proposed model predicted an AHI value ≥ 15 events/hour. Separate figures are provided for males and females, and each subject is colored according to their BMI category as obesity is a big risk factor for developing OSA.

Using a cut-off of 15 events/hour, subjects were classified as being normal or having OSA based on the predictions of this improved model. Figure 3.25 (b) shows the resulting confusion matrix from the classification. Although the accuracy remained 67 ± 4%, sensitivity increased (from 59% to 74%), while specificity decreased (from 72% to 63%), suggesting that the model became better at predicting subjects with OSA with the added information. Although the overall accuracy did not increase, the AUC ROC improved from 65 ± 4% to 69 ± 3%, which reflects the overall improvements in the classification.

The bottom row of Fig. 3.26 shows scatter plots of subjects, divided into males and females and color coded according to their BMI category, who were wrongly classified as being normal based on a predicted AHI $<$ 15 events/hour. Compared to the results shown in the top row of Fig. 3.26 using only craniofacial images, it is noted that the predicted AHI increases for misclassified subjects after adding demographics and questionnaires, most notably for males. Interestingly, there are still a handful of misclassified males who have predicted AHI around 14 events/hour, reflecting that the model was close to classifying these correctly. Additionally, the number of males with extreme AHI values who are wrongly classified as being normal is reduced compared to before as most data points are concentrated in the upper left corner of the plot. All misclassified males are either normal or overweight with respect to BMI, showing that all obese subjects who were wrongly classified as normal before are correctly classified with the added BMI information and OSA related questions. For women, there is still a mixture of normal, overweight and obese subjects, and the difference in pathophysiology compared to men most likely plays a part in the misclassifications.

The bottom row of Fig. 3.27 shows similar scatter plots for subjects wrongly classified as having OSA based on a predicted AHI $\geq$ 15 events/hour. A decrease in predicted AHI is observed for both males and females. The misclassified males are mostly overweight or obese, while almost all misclassified females are obese, showing that AHI overestimations are associated with increased BMI. Most of the normal and overweight women who were wrongly classified as having OSA based on their craniofacial scans are correctly classified as normal with added clinical information.

Table 3.1 compares performance metrics obtained using different modalities to estimate AHI on the same dataset, e.g., model using only demographics, or deriving a diagnosis from the STOP-Bang questionnaire. Although performance of the model using craniofacial images may appear modest, it achieved a higher accuracy than the 62% obtained in the same dataset using the modified STOP-Bang questionnaire. Given that questionnaires are regularly used as an early screening tool for OSA, it is encouraging to observe that accuracy obtained using craniofacial scans exceeds that of the questionnaires, with added ability to provide an estimate of disease severity (i.e., AHI), which is not possible with a simple screening questionnaire.

3.4.2.3 Comparison to Similar Work

Table 3.2 compares model performance to similar work in the literature in terms of MAE, PCC, accuracy, AUC ROC, and number of subjects used in each study. It is evident that the proposed models are at a similar level to all other studies in terms of every performance measure. However, our models were trained and evaluated on a much larger cohort collected at 11 different sleep clinics and used a very different approach than that of others that predicted AHI using landmark-based measured features [41, 42, 45]. A lot of manual work is needed to derive landmark-based,

Table 3.1: Comparison of models using different inputs to predict the apnea-hypopnea index (AHI) of 1366 subjects. The STOP-Bang questionnaire was applied as in clinics to screen for obstructive sleep apnea (OSA) without machine learning. The subject is classified as having OSA if they answer yes to at least 3 of 7 questions. The models using questionnaire variables and demographics were multi-layer perceptrons and the model using craniofacial scans was a multi-view convolutional neural network. An ensemble approach was used when combining two or more inputs. Mean absolute error (MAE) and Pearson Correlation Coefficient (PCC) were calculated between true and predicted AHI, while accuracy (ACC) and area under the receiver operating characteristics curve (AUC ROC) were calculated by classifying subjects into normal/OSA using AHI $\geq$ 15 for OSA.

Model Input	MAE	PCC	ACC	AUC ROC
STOP-Bang Questionnaire	-	-	$62 \pm 4\%$	$65 \pm 4\%$
Questionnaire variables	11.42 ± 1.27	0.38 ± 0.07	$64 \pm 4\%$	$66 \pm 4\%$
Demographics	11.35 ± 1.26	0.40 ± 0.06	$64 \pm 4\%$	$67 \pm 3\%$
Scans	11.38 ± 1.36	0.40 ± 0.04	$67 \pm 4\%$	$65 \pm 4\%$
Demographics + Questionnaire variables	11.24 ± 1.28	0.41 ± 0.06	$65 \pm 4\%$	$67 \pm 4\%$
Scans + Demographics	11.12 ± 1.36	0.44 ± 0.03	$67 \pm 3\%$	$67 \pm 4\%$
Scans + Questionnaire variables	11.03 ± 1.40	0.45 ± 0.04	$67 \pm 4\%$	$68 \pm 4\%$
All combined	11.05 ± 1.36	0.45 ± 0.04	$67 \pm 4\%$	$69 \pm 3\%$

hand-selected features as opposed to using an entirely data driven approach as we propose. As such, our study is reassuring in that the empirically identified features emphasized by the model, recapitulated clinical expertise without the manual labor or years of clinical training and experience. Only Islam et al. [46] used images directly in a data-driven manner, but these authors only used depth information and only had craniofacial scans for 69 subjects as opposed to our 1366 subjects, which is equivalent to the patient volume seen over the entire course of a clinical sleep medicine training fellowship. We implemented the algorithm proposed by Islam et al. [46] on our dataset as seen in Table 3.2, which decreased the overall accuracy from 67% to 60%. This makes sense because our dataset is much larger and much more diverse, since it was collected at many different sites. Furthermore, it shows that using images from several angles holds an advantage over using only frontal depth maps when predicting OSA, even when a pre-trained network which has been trained on more than two million general facial images is utilized.

3.4.2.4 Comparison to Sleep Medicine Specialists

Table 3.3 compares results from the proposed model using craniofacial scans to those of three sleep medicine specialists predicting AHI values based on 3D craniofacial scan displays. The percentage agreement between the three specialists was 67%, again highlighting the presence of significant interscorer variability within the field of sleep medicine. The comparison shows that the model was at a level similar to all three specialists in terms of each performance measure.

Figure 3.28 compares the Bland-Altman plot of true and predicted AHIs from the model using craniofacial scans with a Bland-Altman plot of true and predicted AHIs by the specialists. As discussed earlier, subjects with high AHI are hard to estimate, which is evident when looking at

Table 3.2: Comparison between the two proposed models in this study (bold font) and similar models presented in the literature for predicting apnea-hypopnea index (AHI) of subjects based on their craniofacial images. MAE - Mean absolute error, PCC - Pearson Correlation Coefficient, ACC - Accuracy, AUC ROC - Area under the receiver operating characteristics curve, N - Total number of subjects, N Test - Number of subjects used for testing.

Predictor	MAE	PCC	ACC	AUC ROC	N	N Test	Method	Validation scheme
Craniofacial	**11.38**	**0.40**	**67%**	**65%**	**1366**	**1366**	**Automatic landmarks with CNN. AHI prediction and OSA classification using 2D images and depth maps from five angles and multi-view CNN.**	**10-fold cross-validation**
Craniofacial + demographics + questionnaires	**11.05**	**0.45**	**67%**	**69%**	**1366**	**1366**	**Same approach as above but with an ensemble of models using scans, demographics, and questionnaires, respectively.**	**10-fold cross-validation**
Espinoza-Cuadros et al. [41]	12.56	0.37	71%	67%	285	285	Automatic landmarks with Active Appearance Model. AHI prediction using measurements and Support Vector Regression.	Leave-one-out cross-validation
Nosrati et al. [42]	13.4	0.52	68%	75%	180	180	Manual landmarks. AHI prediction using measurements and Support Vector Regression.	Leave-one-out cross-validation
Balaei et al. [45]	-	-	69%	-	376	204	Automatic landmarks with Support Vector Machine and cascade regression. OSA classification using measurements and logistic regression.	Training-test-set
Islam et al. [46]	-	-	67%	-	69	14	OSA classification using frontal depth maps and pre-trained VGGFace.	Training-validation-test-set
Islam et al. [46] on our dataset	-	-	60%	64%	1366	1366	Same approach as above.	10-fold cross-validation

the sleep specialists' scorings, as they consistently underpredicted the higher AHI values with a similar magnitude as the proposed model. Figure 3.28 also shows that even though both our model and sleep specialists make large underpredictions for highest AHI values, the model does not make large overpredictions in the same manner as sleep specialists. All overpredicted values by the model are within the confidence interval which most likely stems from a bias in the model towards people with low to moderate AHI values, i.e., 5-30 events/hour.

Table 3.3: Comparison of the main performance measures between the proposed model using craniofacial images to predict apnea-hypopnea index (AHI) and three sleep medicine specialists guessing the AHI of subjects from their 3D craniofacial scans. MAE - Mean absolute error, PCC - Pearson correlation coefficient, ACC - Accuracy, AUC ROC - Area under the receiver operating characteristics curve.

Predictor	MAE	PCC	ACC	AUC ROC
Craniofacial	11.38 ± 1.36	0.40 ± 0.04	$67 \pm 4\%$	$65 \pm 4\%$
Specialists data combined	13.34 ± 1.39	0.35 ± 0.12	$66 \pm 4\%$	$66 \pm 4\%$
Specialist 1	13.39 ± 1.67	0.43 ± 0.11	$68 \pm 5\%$	$66 \pm 4\%$
Specialist 2	14.33 ± 0.69	0.42 ± 0.04	$61 \pm 2\%$	$65 \pm 1\%$
Specialist 3	12.08 ± 1.81	0.53 ± 0.17	$69 \pm 8\%$	$72 \pm 8\%$

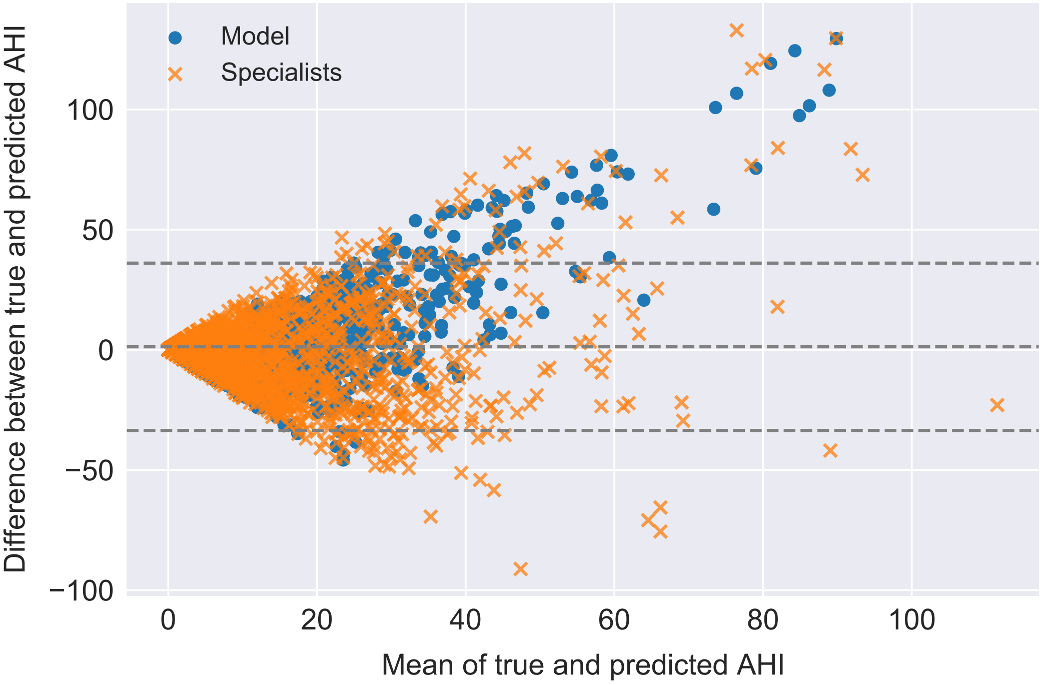


Figure 3.28: Comparison of Bland-Altman plots for apnea-hypopnea index (AHI) predicted by the proposed model (blue) and by three sleep medicine specialists (orange) evaluated for 1366 subjects. Predictions by the proposed model are performed using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted AHI. The predictions by the sleep specialists are performed by inspecting 3D craniofacial scans of all subjects and guessing their AHI.

Figure 3.29 compares confusion matrices of classifying subjects into being normal/OSA based on AHIs predicted from the model using craniofacial scans (Fig. 3.29 (a)) and on AHIs predicted by the sleep specialists (Fig. 3.29 (b)). The sleep medicine specialists achieve a higher sensitivity but a lower specificity than the proposed model. The performance of the specialists is important because it provides context to the performance that can be obtained using craniofacial scans to detect OSA. These specialists represent expert level performance and it is encouraging to see that our model performs to their standards and behaves similarly with respect to AHI predictions.

3.4.2.5 Topographic Display of Important Craniofacial Regions

Figure 3.30 shows saliency maps averaged over the 100 subjects with the highest predicted AHI values and the lowest predicted AHI values, respectively. The topographic display of Fig. 3.30 (a) shows that the network focuses mainly on the neck, jaw, and midface area when predicting high AHI values. These exact same regions have been reported in the literature as being the most important facial features related to OSA [38] and the same regions that sleep specialists focused on when predicting AHI values. Interestingly however, when the network predicts low AHI values, as shown in Fig. 3.30 (b), it seems to focus more selectively on regions of the craniofacial complex that reflect skeletal anatomy (e.g., the maxilla and mandible) for predictions. This may reflect the fact that subjects with milder AHI values have a different pathophysiology where skeletal abnormalities more than body fat may be causing airflow limitations.

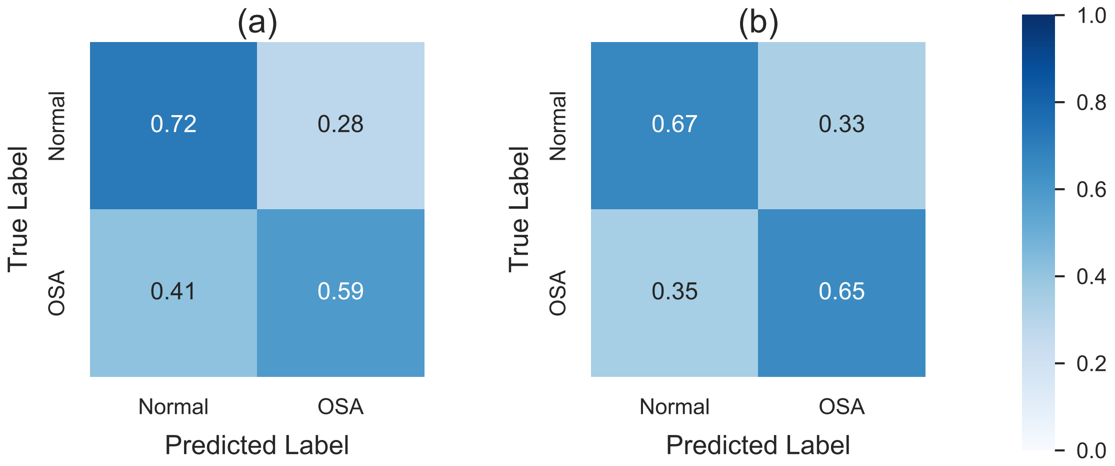


Figure 3.29: (a) Normalized confusion matrix for the model using craniofacial images. (b) Normalized confusion matrix for the three sleep medicine specialists. Predictions by the proposed model are performed using a trained multi-view convolutional neural network, which takes as input five pairs of 2D images and depth maps derived from 3D craniofacial scans and outputs predicted apnea-hypopnea index (AHI). The predictions by the sleep specialists are performed by inspecting 3D craniofacial scans of all subjects and estimating their AHI. The confusion matrices show the results of classifying 1366 subjects into normal/OSA using AHI ≥ 15 as a criterion for OSA based on the predicted AHI values by the proposed model and sleep specialists.

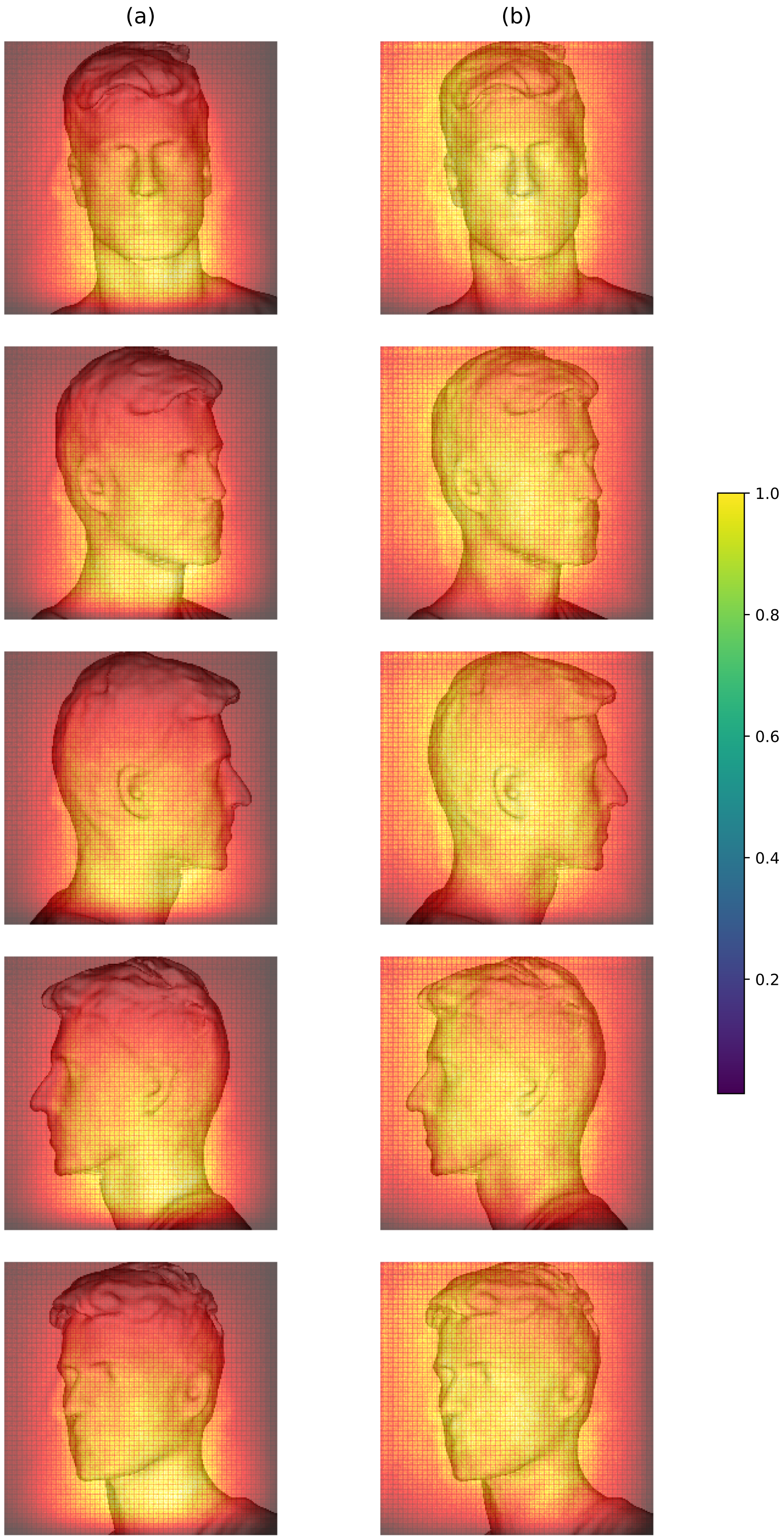


Figure 3.30: (a) Saliency maps averaged over 100 subjects with the highest predicted apnea-hypopnea index (AHI) values by the proposed model. (b) Saliency maps averaged over 100 subjects with the lowest predicted AHI values by the proposed model.

3.4.2.6 Limitations

Even though our proposed model obtains similar performance compared to sleep specialists and similar work in the literature, an average absolute error of more than 11 events/hour is still quite high. First and foremost, it is important to keep in mind that the pathology of OSA is not exclusively attributed to obesity and craniofacial factors. That is also why performance increases only moderately when adding extra information such as demographics and questionnaires. Other variables that could not be assessed from our dataset are internal upper airway anatomy factors, such as size and positioning of the tongue and palate. Other reasons for uncertainty may be physiological changes that occur with age, independent of facial anatomy or obesity such as recruitment of upper airway dilator muscles. Nonetheless, even if the features identified by the model are primarily anatomic in nature, these findings may prove useful to determine phenotypic risk for certain types of OSA. Finally, it should be noted that because our 3D scans were of structural facial features, we intentionally excluded central/mixed apneas from the AHI, in order to focus our model on the anatomical contributions to OSA. However, there are various models of this complex disorder that also account for the physiologic aspects of OSA (e.g., loop gain [143]), which is something that future modeling efforts should take into account.

Another limitation of this study was the quality and quantity of the captured 3D scans. The quality of the scans varied significantly and reflected the fact that they were captured in many different sleep clinics. Some scans had missing parts of the neck, whereas others were affected by poor lighting conditions. Furthermore, we believe that the size of the dataset was too small to truly capture the variation in craniofacial features across humans in relation to OSA in a data-driven manner. Evidently, we observe that similar performance is obtainable in smaller datasets if the features are hand-crafted like landmark-based measurements. The fact that scans were captured either at night before the PSG or in the morning after did not have any effect on diagnostic performance, which was evident when we obtained accuracies of 68% and 66% for scans captured in the morning and at night, respectively, showing that scans can be obtained in both conditions.

Although the focus in this study is on providing a fast, efficient, and cheap screening tool for OSA, efforts to explore alternative screening methods include sleep tests at home [144], usually with very few sensors, such as sound [145, 146], and blood oxygen saturation [147, 148]. Potential of depth and thermal cameras has also been explored in breathing monitoring at an early stage [149, 150]. Even contactless bed sensors have been proposed, although only with moderate success so far [151]. The benefit of using the mentioned approaches is that the person is more comfortable sleeping in their own home wearing few or no wires. However, most studies use a small number of subjects to validate their techniques and still requires a full night's sleep to reach a diagnosis.

3.4.2.7 Future Work

In future work, it would be interesting to explore if, beside AHI, other clinically important variables captured by sleep studies could be better predicted. These could prove to not only make up a more accurate model but could also improve our knowledge of OSA phenotypes and their relation to facial anatomy. For example, Azarbarzin et al. [152] recently suggested that the hypoxic burden is a better measure to use compared to AHI when evaluating sleep apnea severity and resulting cardiovascular risk. Predicting hypoxic burden, oxygen desaturation index (ODI), or the duration of events instead of the AHI, or newer derivatives that better describe sleep disorder breathing heterogenicity could prove more useful and insightful.

3.5 Conclusions

Based on the methods, results, and discussions presented in Sections 3.3 and 3.4, the research questions posed in Section 3.2 are now restated and answered:

Research Question 1: Can a dedicated computer vision model be trained to accurately estimate the sex, age, and BMI of subjects based on their 3D craniofacial scans?

Research Conclusion 1: A framework was presented for predicting patient demographics based on 3D craniofacial scans. This was achieved by converting the 3D images into a series of 2D images and depth maps and implementing a multi-view convolutional neural network for learning. We successfully showed that it is possible to derive variables such as sex (accuracy of 93%), age (MAE of 7.77 years and PCC of 0.76), and BMI (MAE of 4.04 kg/m^2 and PCC of 0.8) of a subject from a surface scan of their face and neck. Subsequently, we implemented a cross-validation approach to obtain performance across all 1366 subjects in the dataset, and results showed that the initial performance was slightly overestimated. The new accuracy for estimation of sex was 91 ± 3%. The new MAE for estimation of age was 8.77 ± 0.53 years, and the new PCC was 0.68 ± 0.06. The new MAE for estimation of BMI was 4.36 ± 0.29 kg/m^2, and the new PCC was 0.73 ± 0.05. This serves as a proof of concept and the applied techniques can be extended to predict more biologically relevant variables such as the AHI of subjects.

Research Question 2: Can a dedicated computer vision model be trained to accurately estimate the AHI of subjects based on their 3D craniofacial scans?

Research Conclusion 2: A deep learning-based computer vision model was presented for predicting AHI automatically from 3D craniofacial scans that can be captured in a minute by a non-specialist. This was achieved by utilizing the framework developed for Research Question 1, but using only five pairs of 2D images and depth maps (focusing on the frontal and profile characteristics) instead of all eight, and including textures with the 3D scans. Furthermore, an additional fully connected layer with dropout was added to the network architecture and training was carried out using 10-fold cross-validation. We successfully showed that it is possible to derive AHI values based on 3D scans using a dedicated computer vision model, which yielded a MAE of 11.38 ± 1.36 events/hour and a PCC of 0.40 ± 0.04 evaluated on 1366 subjects in the dataset.

Research Question 3: Can predicted AHI values from the proposed model be used to accurately detect presence of OSA by classifying subjects as being normal or having OSA?

Research Conclusion 3: We used the clinical meaningful cut-off value for AHI of 15 events/hour to classify subjects as being normal ($\text{AHI} < 15$) or having OSA ($\text{AHI} \geq 15$) and achieved an accuracy of 67 ± 4%, with a sensitivity of 59 ± 8%, a specificity of 72 ± 5%, and an AUC ROC of 65 ± 4%.

Research Question 4: Can adding clinically relevant information like demographics and questionnaires increase performance of the proposed model?

Research Conclusion 4: We added patient demographics (sex, age, and BMI) and seven OSA-related questionnaire variables to the model. This was done by implementing two MLPs, one using demographics and the other using questionnaires as input to predict AHI, and ensembling them with the model using craniofacial scans to predict AHI. This improved model performance by decreasing the MAE from 11.38 ± 1.36 to 11.05 ± 1.40 events/hour and increasing the PCC from 0.40 ± 0.04 to 0.45 ± 0.04. Sensitivity increased from 59% to 74% while specificity decreased from 72% to 63%, but importantly, the overall AUC ROC improved from 65 ± 4% to 69 ± 3%.

Research Question 5: Can the proposed model perform better than current screening questionnaires for OSA?

Research Conclusion 5: All participants had filled out a modified STOP-Bang questionnaire consisting of OSA-related questions. 1366 subjects in the dataset were classified as being healthy or having OSA based on the STOP-Bang questionnaire, which yielded an accuracy of 62% and an AUC ROC of 65%. In comparison, the model based on 3D craniofacial scans yielded an accuracy of 67% and an AUC ROC of 65%, while the model based on a combination of 3D craniofacial scans, patient demographics, and questionnaire variables yielded an accuracy of 67% and an AUC ROC of 69%. These results demonstrate that the proposed models perform better than current screening methods, particularly the model using three different modalities, all of which can be acquired within minutes.

Research Question 6: Can the proposed model perform at a level similar to that of sleep medicine specialists with years of experience?

Research Conclusion 6: Three sleep medicine specialists were recruited to imitate the task of the model, i.e., estimating AHI based on 3D craniofacial scans. The sleep specialist estimated AHI for a third of the dataset each, while also estimating AHI for 150 of the same subjects, which yielded an overall agreement of 67% between the sleep specialists. The model performance was at a similar level to two sleep specialists and better than one, showing that the model is comparable to physicians who have decades of experience working with OSA patients.

Research Question 7: Can regions of the face and neck be identified, which the proposed model focuses on when predicting AHI?

Research Conclusion 7: A topographic display was created based on saliency maps to visualize which craniofacial regions the model focuses on when predicting AHI. The topographic display showed that the model focuses mainly on the neck, jaw, and midface area when predicting high AHI values. These are the same regions which have been reported in the literature as being the most important facial features related to OSA, showing that the model learned the correct features in a data-driven manner.

To summarize the findings and state the overall conclusion, Hypothesis 1 is restated below and answered:

Hypothesis 1

An automatic screening system can be invented, based on dedicated computer vision models, which utilizes 3D craniofacial scans to estimate presence and severity of obstructive sleep apnea more accurately than current screening questionnaires in a fast, cheap, and data-driven manner.

The potential and power of using 3D craniofacial scans to estimate the presence and severity of OSA has been demonstrated. The best performing approach is one using a combination of craniofacial anatomy, patient demographics, and questionnaires, which surpasses performance of current screening questionnaires for OSA. We have shown that our model performs at a level similar to sleep medicine specialists with decades of experience within the field. This is backed by the fact that the model focuses on the exact same craniofacial regions (the neck, jaw, and midface) as the sleep specialists do when they estimate AHI. Furthermore, these regions are associated with development of OSA according to the medical literature. The proposed model has acquired this knowledge by seeing a relatively small sample of 3D scans (<1500) compared to standard current dataset sizes within deep learning and computer vision. A much larger dataset is required for training and evaluation of a screening system such as the one proposed here before clinical applicability can be considered. However, we have demonstrated that such a screening system can be invented and that using it takes a few minutes to derive a diagnosis as opposed to sleeping an entire night and having the data analyzed manually by technicians.

Chapter 4

Automatic Scoring of Drug-Induced Sleep Endoscopy

This chapter explores the potential of automatic scoring of drug-induced sleep endoscopy (DISE) with respect to sites of upper airway collapse and obstruction degrees in patients with obstructive sleep apnea (OSA). The chapter consists of two parts, one based on a published paper and another based on a paper currently under review for publication:

1) Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks [153]

2) Automatic Scoring of Drug-Induced Sleep Endoscopy for Obstructive Sleep Apnea Using Deep Learning

Part 1) was published as a conference paper and served as a proof of concept for the entire methodology utilized for part 2). While waiting to obtain a sufficient number of DISE videos to train and evaluate an automatic deep learning-based scoring system on, we decided to use the small sample of videos available at the time, simplify the problem, and publish the results as a conference paper.

The problem of scoring DISE was simplified to investigate if a dedicated computer vision model could be trained to accurately classify upper airway regions that the endoscope is in throughout a DISE examination. The upper airway regions should be easier to determine than scoring DISE videos due to fewer variables that need to be predicted. Thus, results from part 1) also served as a baseline for what the maximum expected performance could be for part 2). Part 2) was written as a journal paper and explored how a computer vision model that we implemented based on deep learning can be used to automatically score DISE videos with respect to upper airway collapse sites and obstruction degrees. The majority of this chapter's content is based on that journal paper.

Section 4.1 provides the research background and motivation for this chapter. Section 4.2 states the research questions and objectives of this chapter. Section 4.3 contains a data and methods description, and a presentation of results and discussion for part 1), which is based on Paper III. Similarly, Section 4.4 contains a data and methods description, and a presentation of results and discussion for part 2), which is based on Paper IV. Finally, Section 4.5 concludes the chapter by answering the research questions posed in Section 4.2 and relating the findings to Hypothesis 2, which is stated in Section 1.4 and restated in Section 4.2.

4.1 Research Background

Continuous positive airway pressure (CPAP) is the gold-standard treatment for OSA and works by providing a constant level of pressure sufficient to keep the upper airway open [53]. Although CPAP is extremely effective in reducing OSA events, studies show that up to 50% of users give up on the device within a year of therapy because of intolerance, noise, discomfort, or a negative impact on intimacy [54]. Oral appliances may reduce upper airway collapse by advancing the mandible or refraining the tongue and epiglottis from falling back, but, as they are generally less effective than CPAP they are considered the second line of treatment after CPAP [109].

For some patients, surgical procedures can be viable options to prevent collapse or increase upper airway space, with the most common surgery being a modified uvulopalatopharyngoplasty [110], where excess tissue is removed from the soft palate and lateral walls of the pharynx, often combined with tonsillectomy (removal of the palatine tonsils) [56]. Other procedures include TORS (transoral robotic surgery) on the tongue base and epiglottis [154] and maxillomandibular advancement (advancement of the upper and lower jaw) [57].

Prior to surgery, DISE is performed to examine the location and pattern of sleep-related upper airway collapse using a fiberoptic endoscope under sedation, which is designed to simulate natural sleep [58]. The endoscope is introduced through the nasal cavity and is used to examine the upper airway from the nares to the level of the glottis. After a DISE examination, the surgeon evaluates the sites of collapse in the upper airway according to the VOTE (velum, oropharynx, tongue base, epiglottis) classification system, the most commonly used scoring system for DISE [59]. These four sites can collapse, either individually or in combination, causing obstruction in the upper airway. The four different sites are shown in Fig. 4.1 from an endoscope positioned at the top of the soft palate looking downwards into the upper airway.

The VOTE classification system assigns a degree of obstruction to each upper airway site and a pattern of collapse to each site where a collapse occurs. VOTE obstruction degrees are classified either as 0 (no obstruction), 1 (partial obstruction), or 2 (complete obstruction) [59]. Additionally,

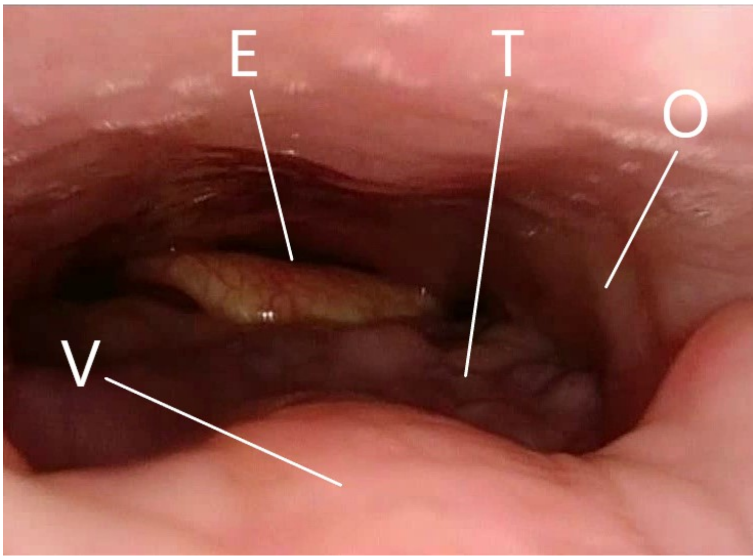


Figure 4.1: A frame taken from a drug-induced sleep endoscopy examination, which is used to characterize the upper airway collapse pattern in patients with obstructive sleep apnea. Collapse can occur at four different upper airway sites: Velum (V), oropharynx (O), tongue base (T), and epiglottis (E). An obstruction degree is assigned to each site and is classified either as 0 (no obstruction), 1 (partial obstruction), and 2 (complete obstruction). In this image, V and O have obstruction degrees of 0, while T and E have obstruction degrees of 1.

there are three possible patterns of collapse: antero-posterior (A-P) collapse, lateral collapse, and concentric collapse [59]. The upper airway can collapse in any of these patterns at V [155], but only lateral collapse can occur at O, only A-P collapse at T, and only lateral and A-P at E as outlined in Table 4.1. Lateral collapse at E is extremely rare in practice [156]. Figure 4.2 shows examples of upper airway collapse patterns and obstruction degrees at V, while Fig. 4.3 shows similar examples for O, T, and E.

DISE suggests location and indication for surgical intervention and its use has been shown to improve OSA surgical outcomes [64]. The analysis however depends on the procedure and the evaluation hereof and as such presents inter-rater variability between surgeons who score DISE.

Table 4.1: Upper airway sites where collapse can occur in obstructive sleep apnea patients according to the VOTE classification system [59], which is used to score drug-induced sleep endoscopy videos. The VOTE classification system assigns an obstruction degree to each site, which is classified as either 0 (no collapse), 1 (partial obstruction), or 2 (complete obstruction). Furthermore, a pattern of collapse is assigned to each site that collapses. Checkmarks indicate the possible patterns of collapse at each site.

Site	**Degree of obstruction**	**Pattern of collapse**		
		Antero-posterior	**Lateral**	**Concentric**
Velum	0, 1, or 2	✓	✓	✓
Oropharynx			✓	
Tongue base		✓		
Epiglottis		✓	✓	

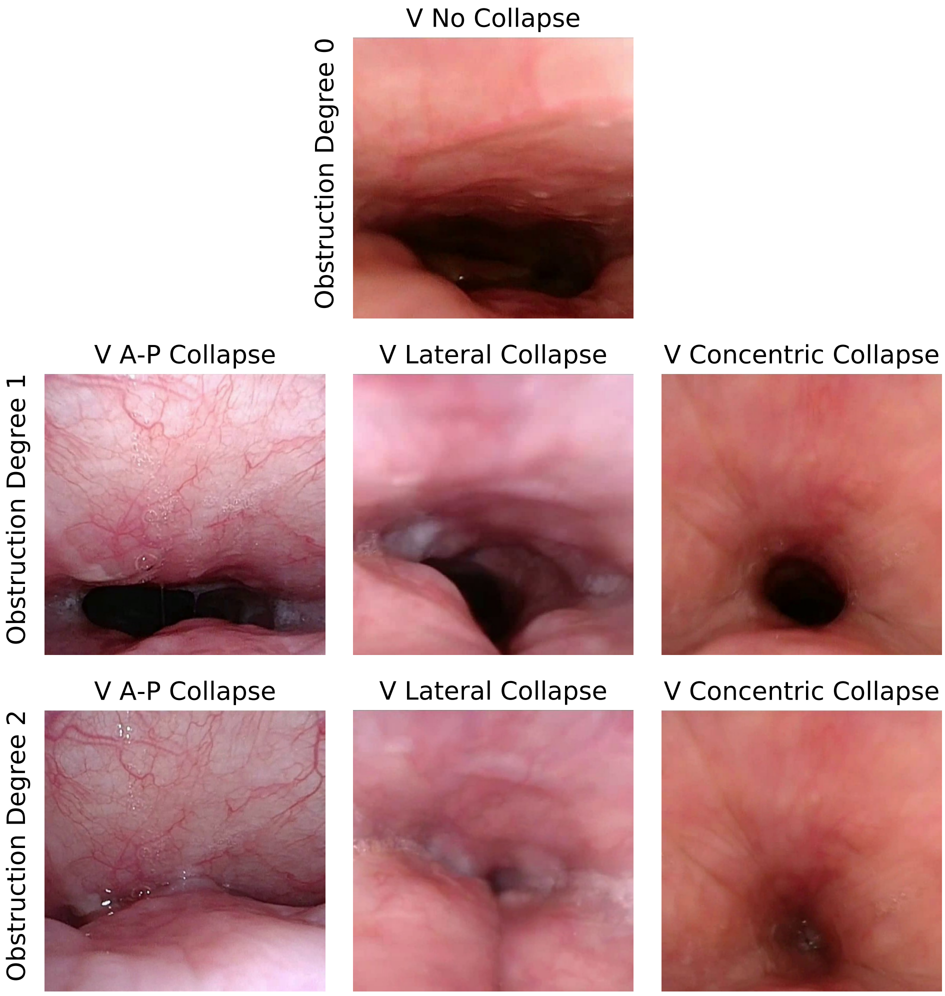


Figure 4.2: Examples of possible ways the velum (V) can collapse in the upper airway in obstructive sleep apnea patients. The obstruction caused by the collapse is classified either as 0 (no collapse), 1 (partial obstruction), or 2 (complete obstruction). The possible patterns of collapse are antero-posterior (A-P), lateral, or concentric.

First, there is an anatomical variation across subjects with respect to the upper airway, and the pattern of collapse may also be affected by the depth of sedation [157–161]. Secondly, DISE videos can appear chaotic due to several sites collapsing simultaneously in patients with severe OSA, essentially pushing the endoscope around and making it difficult to determine the sites of collapse. Mucus or saliva may also cover the endoscope, which can reduce or distort the video quality significantly and at times make it almost impossible to visually inspect the upper airway. Furthermore, studies examining inter-rater reliability between surgeons show poor to moderate agreement [60–65], demonstrating that despite a well-established classification system, interpretation remains subjective. Due to these limitations associated with DISE and the VOTE classification system, surgeons will benefit from a system capable of scoring DISE videos automatically in an unbiased and data-driven manner. Such a system could in turn assist in the planning of surgical treatment by identifying sites of upper airway collapse and associated obstruction degrees in OSA patients.

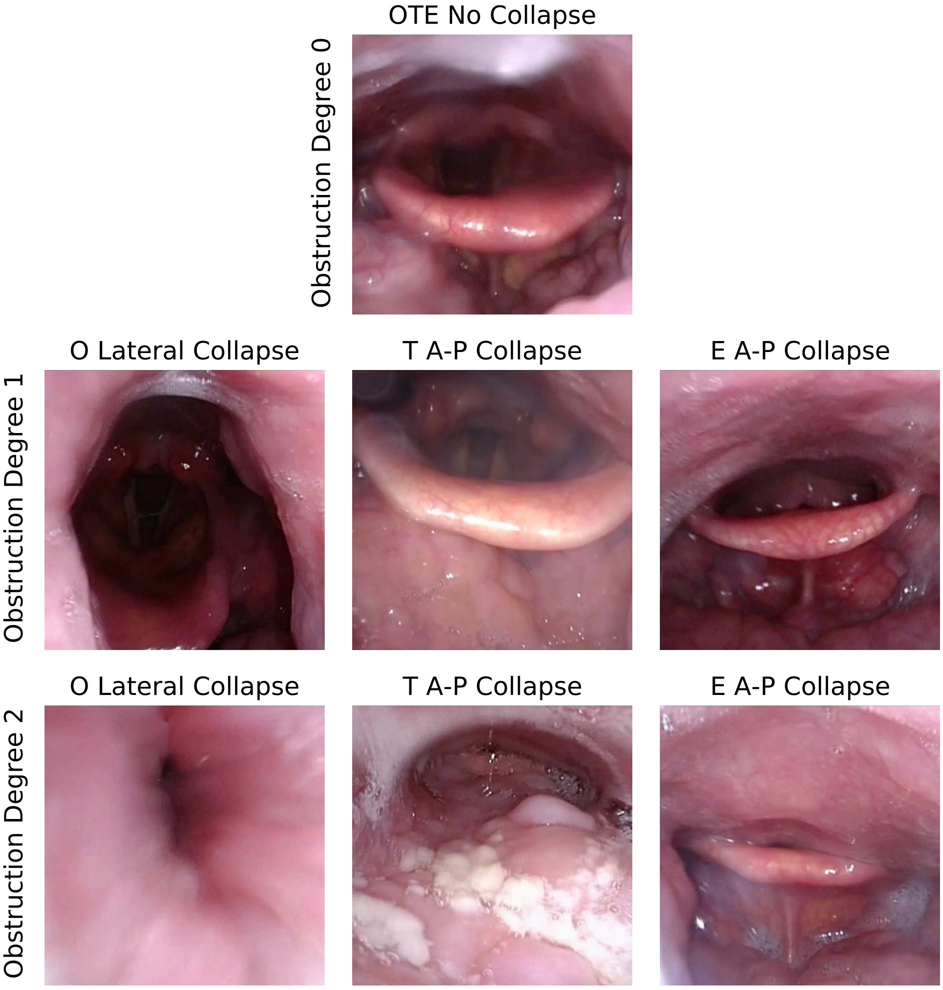


Figure 4.3: Examples of possible ways the oropharynx (O), tongue base (T), and epiglottis (E) can collapse in the upper airway in obstructive sleep apnea patients. The obstruction caused by the collapse is classified either as 0 (no collapse), 1 (partial obstruction), or 2 (complete obstruction). The possible pattern of collapse is lateral for O, antero-posterior (A-P) for T, and A-P for E. Lateral collapse for E has been left out, since it is rare.

4.2 Research Questions and Objectives

Based on the research background above, Hypothesis 2 from Section 1.4 is restated, followed by research questions derived from Hypothesis 2 that this chapter aims to answer.

Hypothesis 2

An automatic scoring system can be invented, based on dedicated computer vision models, which utilizes drug-induced sleep endoscopy examination videos to estimate sites of upper airway collapse and obstruction degrees in obstructive sleep apnea patients with a similar accuracy as otolaryngology - head and neck surgeons.

Research Questions

Research Question 1: Can a dedicated computer vision model be trained to accurately classify upper airway regions in DISE videos?

Research Question 2: Can a dedicated computer vision model be trained to accurately score DISE videos for sites of upper airway collapse and obstruction degrees?

Research Question 3: Can the proposed model generalize well across DISE procedures performed at different sleep centers by different otolaryngology - head and neck surgeons?

Research Question 4: Can the proposed model perform at a level similar to that of otolaryngology - head and neck surgeons with years of experience?

From the research questions posed above, research objectives are formulated below, each one designed to answer a specific research question:

(i) Label each DISE video second-by-second according to upper airway regions, and split the videos into 5-second clips. Then utilize these input and output pairs to train and evaluate a dedicated computer vision model to classify upper airway regions in DISE videos.

(ii) Split DISE videos into 5-second clips and label each clip according to the obstruction degree for each upper airway site. Then utilize these input and output pairs to train and evaluate a dedicated computer vision model for each upper airway site to classify the obstruction degree for each clip.

(iii) Evaluate performance of the proposed model for DISE videos obtained from different sleep centers and performed by different otolaryngology - head and neck surgeons. Investigate if there is any bias towards a specific sleep center or surgeon.

(iv) Compare performance of the proposed model with research papers reporting inter-rater reliability among otolaryngology - head and neck surgeons in scoring obstruction degrees for each upper airway site in DISE videos.

These research objectives will be fulfilled in the following sections.

4.3 Paper III: Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks

Abstract

Purpose: Assessing the upper airway of obstructive sleep apnea patients using drug-induced sleep endoscopy (DISE) before potential surgery is standard practice in clinics to determine the location of upper airway collapse. According to the VOTE classification system, upper airway collapse can occur at the velum (V), oropharynx (O), tongue (T), and/or epiglottis (E). Analyzing DISE videos is not trivial due to anatomical variation, simultaneous upper airway collapse in several locations, and video distortion caused by mucus or saliva. This paper is a proof of concept for classifying upper airway regions using 24 annotated DISE videos.

Methods: The first step towards automated analysis of DISE videos is to determine which upper airway region the endoscope is in at any time throughout the video: V (velum) or OTE (oropharynx, tongue, or epiglottis). An additional class denoted X is introduced for times when the video is distorted to an extent where it is impossible to determine the region. We propose a convolutional recurrent neural network using a ResNet18 architecture combined with a two-layer bidirectional long short-term memory network. The classifications were performed on a sequence of 5 seconds of video at a time.

Results: The network achieved an overall accuracy of 82% and F1-score of 79% for the three-class problem, showing potential for recognition of regions across patients despite anatomical variation.

Conclusion: Results indicate that large-scale training on videos can be used to further predict the location(s), type(s), and degree(s) of upper airway collapse, showing potential for derivation of automatic diagnoses from DISE videos eventually.

4.3.1 Methods

This section describes the data collection of DISE videos and the subsequent pre-processing and use of these videos to train a convolutional recurrent neural network to classify upper airway regions.

4.3.1.1 Data Collection

The DISE examinations were performed in accordance with DISE procedure guidelines described by Kiaer et al. [162] and Lan et al. [163]. Figure 4.4 shows how the DISE procedure is performed, while Fig. 4.5 shows the endoscope that is used for the DISE procedure. Patients who are eligible for a DISE procedure have confirmed OSA and no potential risk of any complications during the procedure, such as being intolerant to propofol, having a nasal obstruction that can prevent the passage of an endoscope, or having an airway that is deemed unsuitable for the procedure [164].

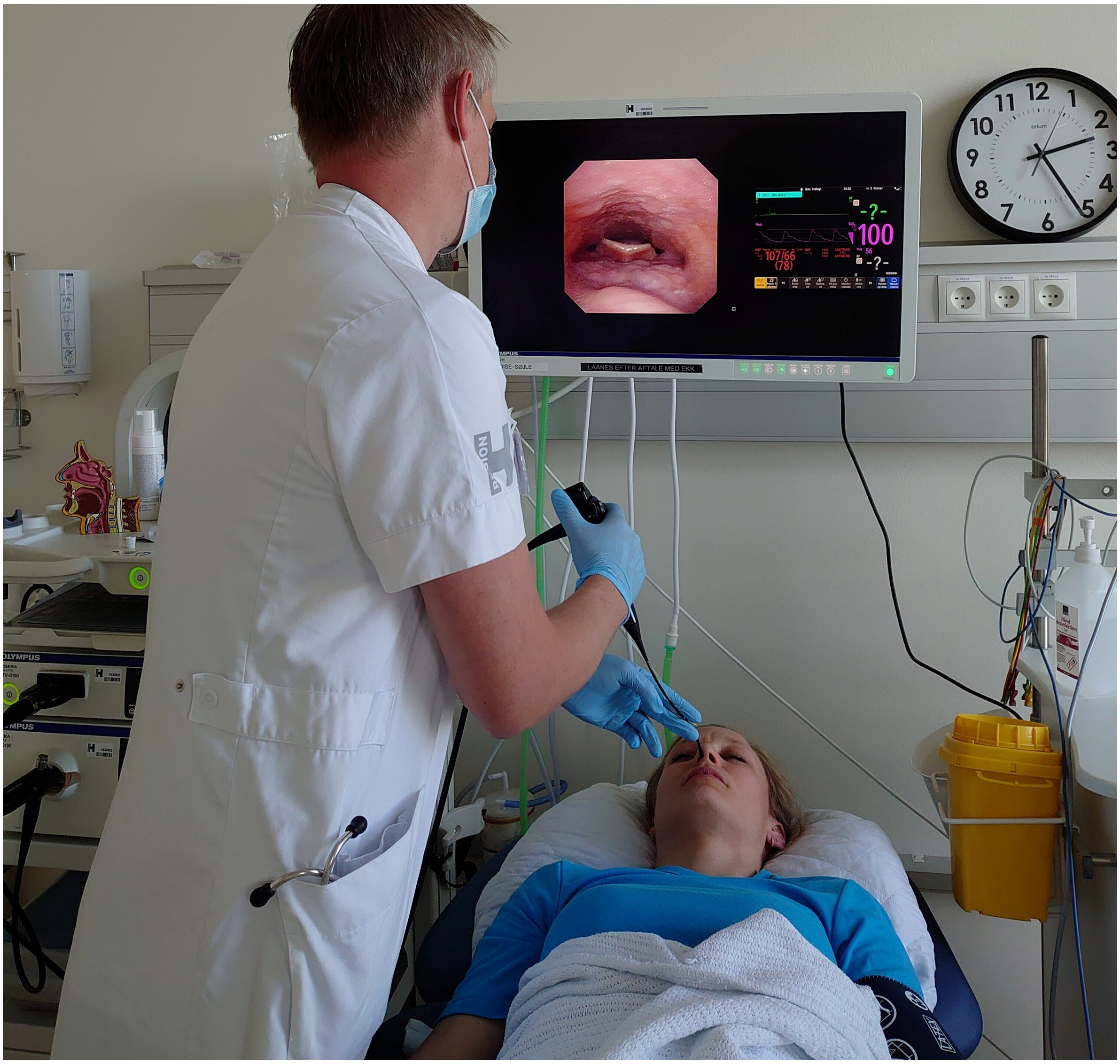


Figure 4.4: A drug-induced sleep endoscopy examination being performed. The patient lying in a supine position is sedated using propofol, which is designed to simulate natural sleep. The otolaryngology surgeon is guiding the long, tube-like endoscope through the nasal cavity into the throat of the patient and examining the upper airway by looking at the screen, which is displaying real-time video from the endoscope. Source: [165]

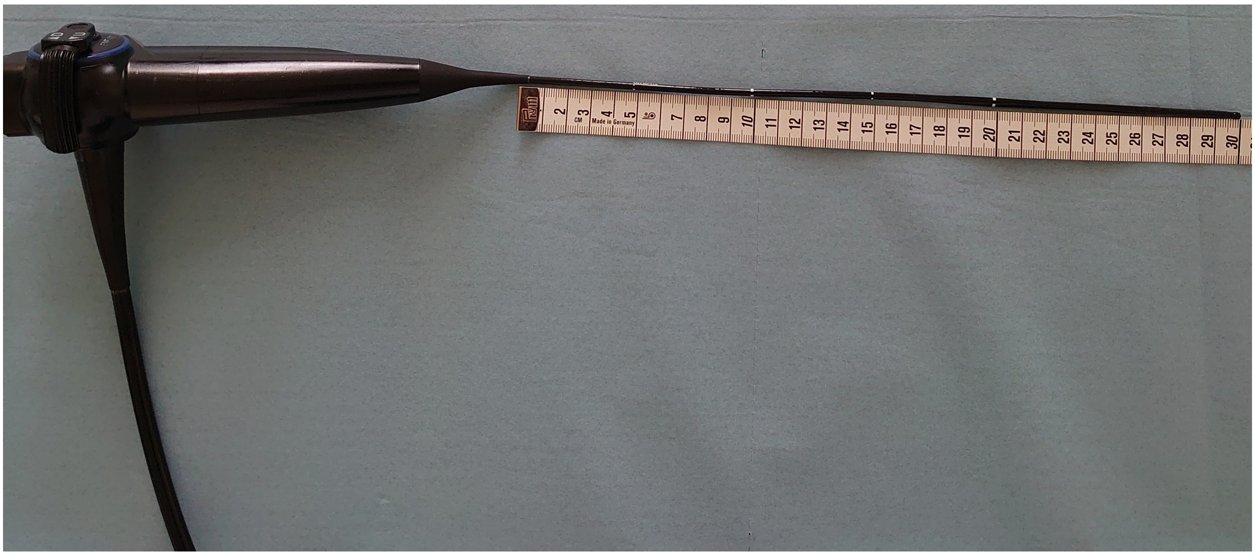

2 CM 3 4 Made in Germany 5 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Figure 4.5: An example of an endoscope used for drug-induced sleep endoscopy. The endoscope is long and tube-like and a ruler is provided in the picture for measurement to show how long the endoscope is. Source: [165]

The procedure is initiated by an anesthesiologist who sedates the patient with propofol, which is designed to simulate natural sleep. After sedation, an otolaryngology - head and neck surgeon introduces an endoscope through one of the nostrils of the patient and guides it through the nasal cavity to the posterior nares. From there, the surgeon examines the upper airway from the posterior nares to the level of the glottis. The surgeon may perform a jaw-thrust maneuver during the examination to investigate the effect of mandibular advancement on upper airway collapse [166]. The video from the endoscope is transmitted in real-time to a screen that the surgeon looks at while performing the procedure. The video is also recorded during the procedure and saved to a video file for subsequent analysis and scoring using the VOTE classification system [59].

4.3.1.2 Data Description

We included a total of 24 DISE videos collected at Copenhagen University Hospital, which were performed in accordance with the DISE procedure guideline described by Kiaer et al. [162]. The Institution's Ethical Review Board approved all experimental procedures involving human subjects. The videos were approximately 2-5 minutes in duration with a frame rate of 25 frames per second and a resolution of 864 × 540 pixels. The combined video duration in the dataset was approximately one hour. All videos were anonymized by removing parts of recordings where the endoscope was not inside the subject.

Each video was initially scored by the surgeon, who collected it, as a single line summary of where, how, and to what degree the upper airway collapsed. However, the purpose of this study was to classify upper airway regions that the endoscope is in at all times throughout a DISE video.

Two upper airway regions were distinguished for this purpose: the region consisting of the velum (V), and the region consisting of the oropharynx, tongue base, and epiglottis combined (OTE). An additional class denoted X was introduced in case the video was distorted (e.g., due to mucus or saliva on the camera) to a degree where it is impossible to recognize the upper airway. Figure 4.6 visualizes different examples of the three classes taken from DISE videos.

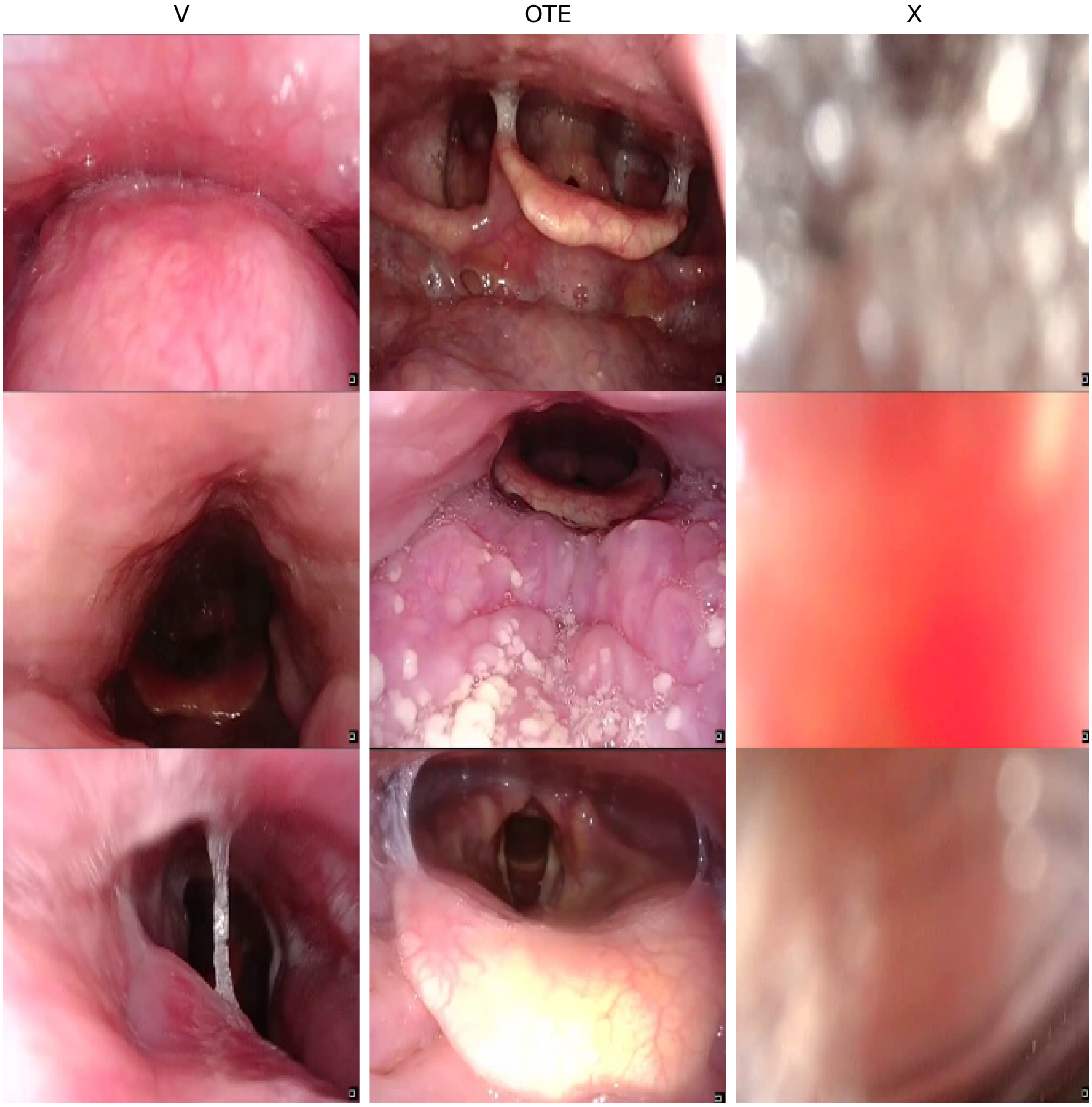


Figure 4.6: Three examples of each class used to classify upper airway regions in drug-induced sleep endoscopy videos for patients with obstructive sleep apnea. The three classes are the velum (V) in the first column, oropharynx, tongue, and epiglottis combined (OTE) in the second column, and distortion in video (X) in the third column.

For machine learning purposes, labels were required that detailed each time the endoscope transitioned either from one region to another (i.e., V to OTE or OTE to V) or from visible video to distorted video and vice versa (i.e., V to X, OTE to X, X to V or X to OTE). Videos were labeled in this manner by consulting with the otolaryngology surgeon who initially labeled the videos [165] and another expert surgeon who introduced the VOTE classification system back in 2011 [59, 167]. Table 4.2 shows an example of the structure of labels created for this study, while Table 4.3 outlines the distribution of the three classes within the dataset, showing that e.g., the class X is massively underrepresented compared to V and OTE.

4.3.1.3 Pre-Processing

Initially, all frames were extracted from each video, yielding 25 frames per second. Subsequently, every 5$^{\text{th}}$ frame was selected, yielding 5 frames per second, because no visual difference was observed between consecutive frames during inspection. Assuming that the model would extract features primarily related to anatomical structures and not color differences, frames were converted to gray scale to reduce computational cost associated with training. All frames were rescaled to 224×224 pixels, which was found to be appropriate for reducing computational cost while still preserving discriminatory information between upper airway regions. Finally, each frame was normalized to the range [0,1], by dividing each pixel value by 255, to ensure faster convergence during training.

The final input output pairs were 5-second clips, consisting of 25 frames, and corresponding labels denoting the upper airway region (V, OTE, or X) for each frame. Note from the structure of the labels (Table 4.2) that is is possible to have different labels within the same 5-second clip if a transition between regions occurs within that clip. This emphasizes the need for labels for each frame instead of utilizing a single label for an entire 5-second clip.

Table 4.2: Example of labels created for drug-induced sleep endoscopy videos using three classes which are used to classify upper airway regions: velum (V), oropharynx, tongue, and epiglottis combined (OTE), and distortion in video (X).

Time (s)	7-15	16-28	29-35	36-40	40-45	45-57	57-60
Region	V	X	OTE	X	V	OTE	V

Table 4.3: Distribution of the three classes in the dataset used for classification of upper airway regions: velum (V), oropharynx, tongue, and epiglottis combined (OTE), and distortion in video (X).

Class	**Total Duration (s)**	**N Frames**
V	1543	7715
OTE	2041	10205
X	376	1880
Total	3960	19800

4.3.1.4 Convolutional Recurrent Neural Network

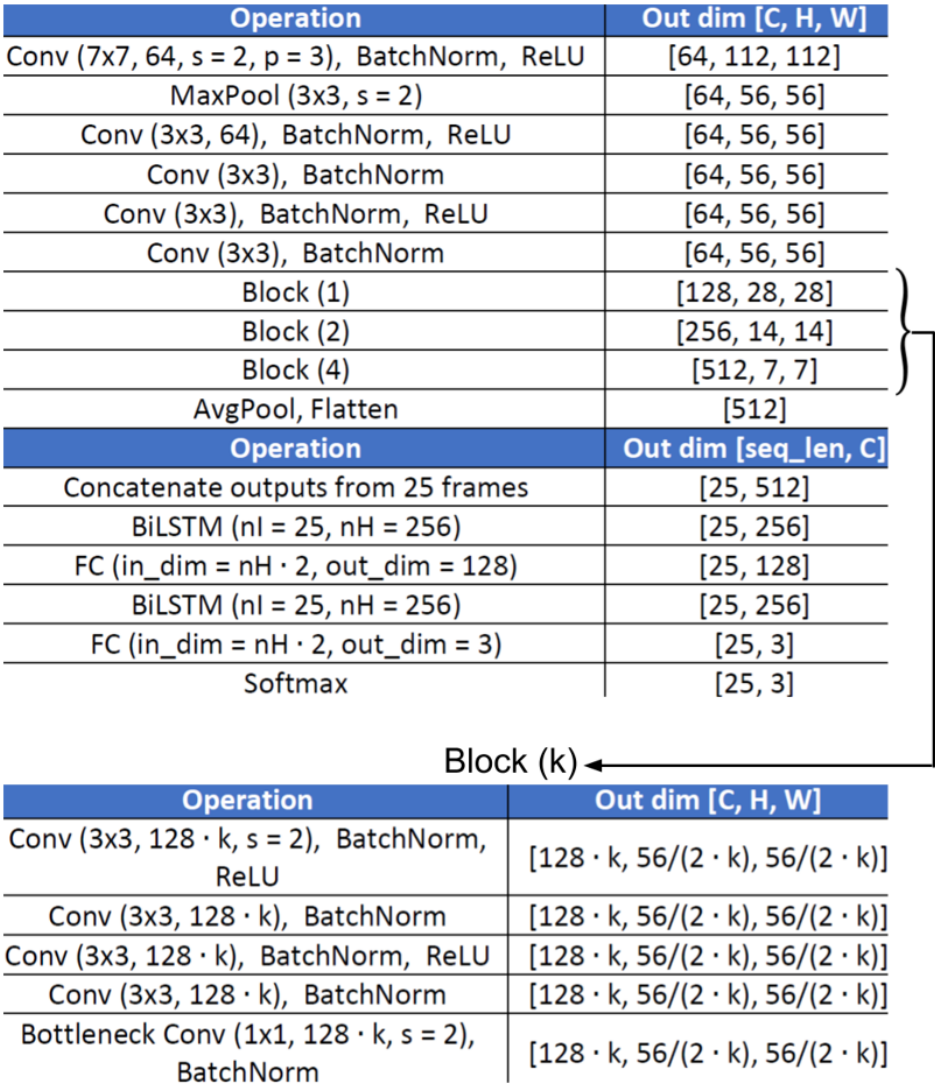
The purpose of applying machine learning was to learn data-driven discriminatory information about V, OTE, and X based on 5-second video clips from DISE examinations. For this purpose, we implemented a convolutional neural network (CNN) with a ResNet18 architecture [130] combined with a two-layer bidirectional long short-term memory (Bi-LSTM) neural network [168]. Figure 4.7 shows the network architecture with all details provided for classification of upper airway regions using 5-second input clips derived from DISE videos.

The CNN was included for its ability to automatically extract meaningful features from the input frames in relation to the desired outputs. They do so by applying layers of convolutions and non-linear activation functions to the input frames, where the weights of the convolution filters are learned during training based on the training data [131]. A ResNet18 architecture was selected because it is state-of-the-art for image recognition compared to other CNN architectures such as AlexNet [132] or VGG-16 [122]. ResNets utilize skip connections between layers to allow deeper architectures (i.e., more layers) without hurting performance while being computationally cost-efficient to train [130].

The Bi-LSTM network was included for its ability to learn context between the input frames in both directions and preserving the temporal information that is present in data such as video [169]. In contrast to regular neural networks, which are typically used when the input features are independent of each other, Bi-LSTMs are suitable for data in which the order of the input features matters, such as for time series data. The Bi-LSTM incorporates information from previous and subsequent outputs when predicting an output at a given time step, which is useful for learning patterns in temporal data [168]. 5-second clips were utilized for learning because the amount of context in both directions was appropriate using this duration. Longer clips would lead to more parameters in the network and were not investigated due to limited computational resources.

The Resnet18 network was implemented such that a 5-second clip (consisting of 25 frames) could be input one frame at a time. The output was a feature map of size 1x512 for each frame. The feature maps for all frames were concatenated to form a 25x512 matrix, where each row is considered a time-step in the original 5-second clip and each column is a feature vector for a particular frame. This matrix was then processed by a Bi-LSTM layer, followed by a dense layer, which reduced the number of features from 512 to 128 while keeping time steps intact, i.e., resulting matrix dimensions of 25x128. A second Bi-LSTM and dense layer reduced the number of features further from 128 to 3, yielding a matrix with dimensions 25x3. Finally, a softmax activation function was applied to yield a probability for each class, i.e., V, OTE, and X, for each frame in the original 5-second clip. The softmax activation function transforms the output vector to the range [0,1], where the probabilities for all three classes sum to 1.

The optimal number of time steps and hidden neurons in the Bi-LSTM layers were found using hyperparameter tuning, which was performed in a grid search-like manner where the hyperparameters were varied and different combinations of these were investigated. The optimal hyperparameters were the ones which yielded the lowest error on the validation set, which is defined in the next section.



Operation	Out dim [C, H, W]
Conv (7x7, 64, s = 2, p = 3), BatchNorm, ReLU	[64, 112, 112]
MaxPool (3x3, s = 2)	[64, 56, 56]
Conv (3x3, 64), BatchNorm, ReLU	[64, 56, 56]
Conv (3x3), BatchNorm	[64, 56, 56]
Conv (3x3), BatchNorm, ReLU	[64, 56, 56]
Conv (3x3), BatchNorm	[64, 56, 56]
Block (1)	[128, 28, 28]
Block (2)	[256, 14, 14]
Block (4)	[512, 7, 7]
AvgPool, Flatten	[512]

Operation	Out dim [seq_len, C]
Concatenate outputs from 25 frames	[25, 512]
BiLSTM (nI = 25, nH = 256)	[25, 256]
FC (in_dim = nH · 2, out_dim = 128)	[25, 128]
BiLSTM (nI = 25, nH = 256)	[25, 256]
FC (in_dim = nH · 2, out_dim = 3)	[25, 3]
Softmax	[25, 3]

Block (k)

Operation	Out dim [C, H, W]
Conv (3x3, 128 · k, s = 2), BatchNorm, ReLU	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm, ReLU	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]
Bottleneck Conv (1x1, 128 · k, s = 2), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]

Figure 4.7: Architecture of the proposed model for classifying upper airway regions in drug-induced sleep endoscopy videos. The input is a 5-second video consisting of 25 frames. The frames are input individually to the convolutional neural network and the outputs are concatenated before the recurrent part of the network. The parameters in the convolution operations (Conv) are kernel size, number of output channels, stride (s), and padding (p), and the output dimensions are specified by number of channels (C), height (H), and width (W). The parameters in the bidirectional long short-term memory (BiLSTM) network layers are number of input features (nI) and number of hidden neurons in each direction (nH). The parameters in the fully connected layers (FC) are input features (in dim) and output features (out dim).

4.3.1.5 Training, Validation, and Testing

Training, validation, and testing was carried out by using a training, validation, and test set split. 18 videos (amounting to 51 minutes) were used for training, 3 videos (amounting to 8 minutes) for validation, and 3 videos (amounting to 7 minutes) for testing. The split was prioritized such that sufficient data was utilized to train the model to learn an accurate discrimination between upper airway regions from DISE videos. The test set consisting of 3 videos was small, but given that the entire dataset only consisted of 24 videos and that this study was solely a proof of concept, the majority of data was assigned to train the proposed model, as it was a concern if this would be enough to generalize across patients with different internal anatomies and videos of varying quality.

Cross-entropy was used as loss function given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_{ij} \log p_{ij}, \tag{4.1}$$

where y_{ij} is the true probability that class j is the label of the i^{th} training sample, and p_{ij} is the predicted probability that class j is the label of the i^{th} training sample. Here N is the number of training samples, and K is the number of classes (i.e., 3). The true probability in each case is 1 for the target class and 0 for the two others. Cross-entropy was used as it is the natural choice for multi-class classification problems.

The learning rate was set to $1 \cdot 10^{-5}$ with a weight decay of $5 \cdot 10^{-4}$ and was chosen using hyperparameter tuning. The optimal choice was a learning rate that is not too high, which can cause the parameter update via gradient descent to diverge from the minima, and not too low, which would slow down training significantly. The weight decay was used to add a penalty term to the loss function during optimization, which shrinks the weights and helps to prevent overfitting. Weights were applied in the loss function for the V and X classes, since the dataset was heavily imbalanced as witnessed in Table 4.3. The weight for V was calculated as the ratio between the majority class (OTE) and V, and the weight for X was calculated as the ratio between OTE and X.

The batch size was set to 2 and was limited by computational resources. The Adam optimizer [133] was used for optimization of the network due to its superior performance compared to other optimization algorithms. Early stopping was applied when the validation error did not decrease for 3 consecutive epochs (patience of 3) to help prevent overfitting.

Python 3.7.4 and Pytorch 1.3.1 were used for pre-processing and deep learning purposes. Training, validation, and testing was carried out on a GeForce RTX 2080 and took approximately one hour to complete.

4.3.1.6 Performance Measures

The classified upper airway region for each frame was compared to the labels. Performance was evaluated using weighted a F1 score [170], which was used instead of accuracy due to a large imbalance between the three classes. The F1 score is calculated as the harmonic mean of precision and recall or alternatively in terms of true positives (TP), false positives (FP), and false negatives (FN):

$$F1\ score = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4.2}$$

Using a frame labeled V as example, TP represents the number of frames that are correctly classified as V, FP represents the number of frames that are incorrectly classified as V, and FN represents the number of frames classified as OTE and X that are actually V. The F1 scores for OTE and X were calculated similarly. The weighted F1 score is calculated by averaging the F1 score of the individual upper airway regions multiplied by their proportion in the dataset. The F1 scores were averaged over all frames in each video, such that performance between each video could be compared. Finally, the mean F1 score across the videos in the test set was calculated to yield the overall performance of the proposed model in the test set.

4.3.2 Results and Discussion

This is the first attempt to use a data-driven approach to identify upper airway regions during the DISE procedure. The performance reported on the classification of upper airway regions in DISE videos was evaluated on the test set consisting of three videos.

4.3.2.1 Overall Performance

The mean weighted F1-score obtained using the proposed model was 79% in the test set. In contrast, if the network had simply predicted all frames to be the majority class in the test set (i.e. OTE), the overall F1-score would be 23%. In this context, the model performs much better than random guessing. Furthermore, mean F1-scores for V, OTE, and X were 74%, 79%, and 68%, respectively.

4.3.2.2 Misclassified Frames

Figure 4.8 shows the confusion matrix for the classification of upper airway regions, while Fig. 4.9 depicts examples of misclassified frames for each class. The X class is intuitively the easiest to recognize since it means that the video is too distorted to derive anything and it would be a trivial task to recognize this class even for a person unfamiliar with DISE videos. This is also reflected

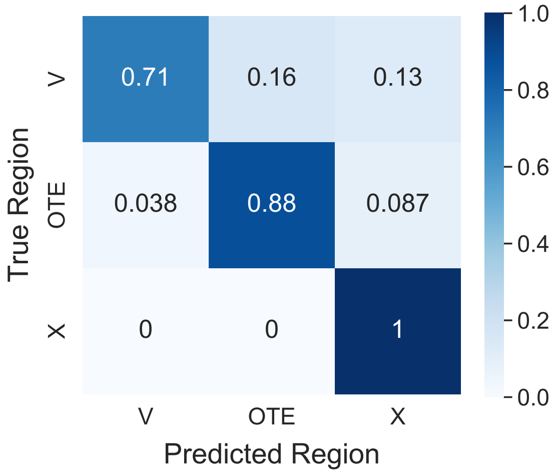


Figure 4.8: Normalized confusion matrix for classifying regions in the upper airway from 2150 frames with three different classes: velum (V), oropharynx, tongue, and epiglottis combined (OTE), and distorted video (X). The frames are classified using a trained convolutional recurrent neural network, which takes as input 25 frames from drug-induced sleep endoscopy videos and outputs the classified upper airway region for each frame.

by the fact that even with the limited number of frames with class X in the dataset, the model was easily able to learn to recognize this class. Looking at Fig. 4.8, it is noted that the sensitivity is 100%, meaning that none of the frames labeled X are misclassified. However, both the V and OTE classes are occasionally misclassified as X, which Fig. 4.9 shows examples of. It is noted that the model classifies a frame as X any time there is mucus or saliva on the endoscope even if some structures are still visible to some degree. When annotating the data, a frame was only labeled as X if there was no way to estimate the region based on the video or context from previous frames, while the model has learned that any mucus or saliva on the endoscope equals a classification of X.

The model also performed well for the OTE class, reflected by a high sensitivity and F1 score. It is noted from Fig. 4.8 that when OTE is misclassified, it is mostly as X, which is again explained by the fact that the model is sensitive to mucus and saliva on the camera, even if it is possible to derive the upper airway region. Scenarios where OTE is misclassified as V is illustrated in Fig. 4.9, where it is observed that this occurs when the endoscope is at the border between the V and OTE regions. Even experts analyzing these frames could have scored them as V instead of OTE, and it appears that the last frame at the bottom has been wrongly labeled as OTE even though the endoscope is in the V region.

For the V class, the model did not perform as well as for OTE and X. Figure 4.8 shows that the misclassifications are almost equally split between OTE and X. The frames misclassified as X are due to the same reason as for OTE. Examples of V being misclassified as OTE are shown in Fig. 4.9. In this case, it appears that the misclassifications do not necessarily occur when the endoscope is close to the OTE region, but rather when the OTE region is visible from the V region so that

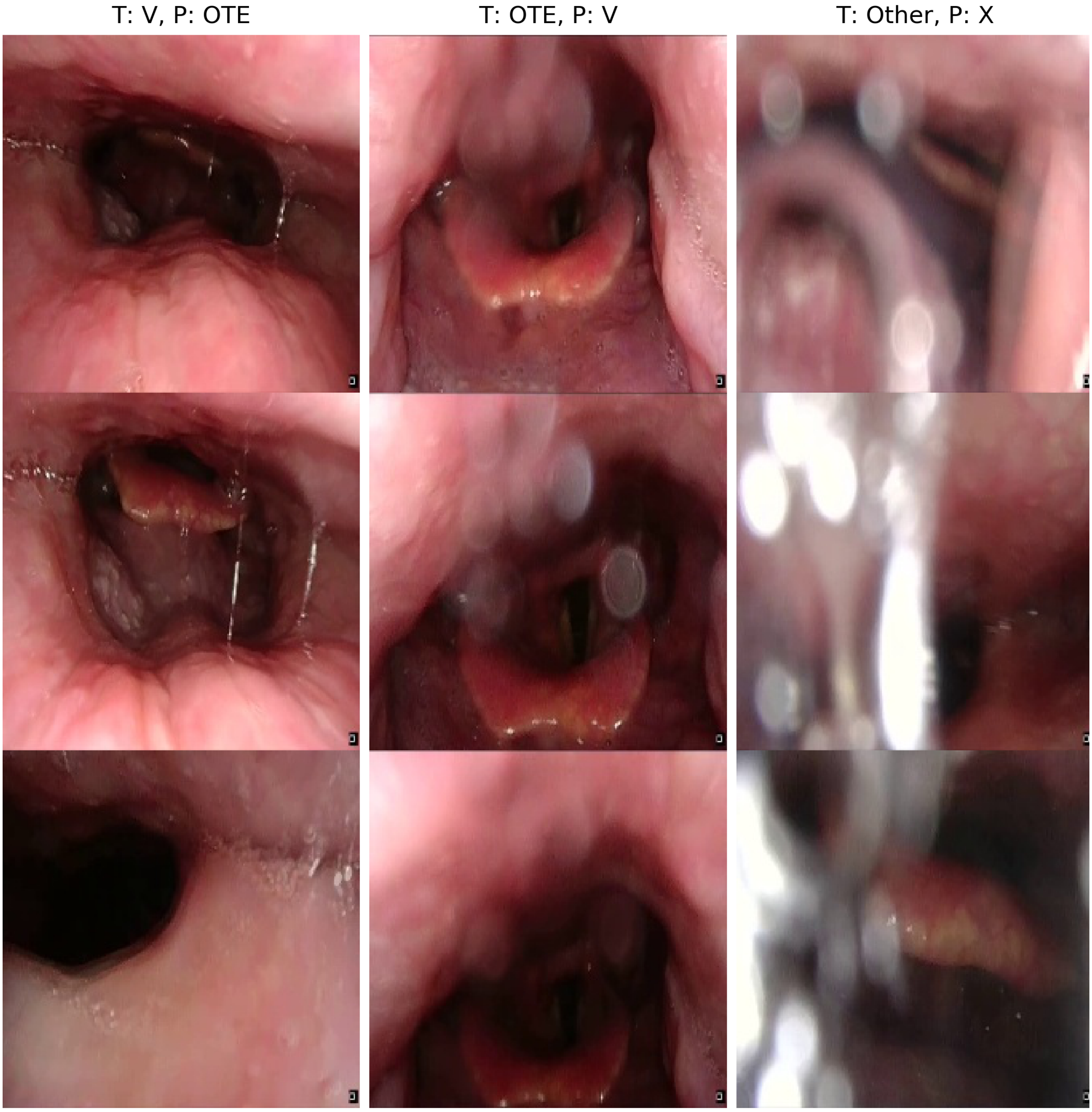


Figure 4.9: Examples of frames wrongly classified with respect to upper airway regions: velum (V), oropharynx, tongue and epiglottis combined (OTE), and distorted video (X). T is the true class and P is the predicted class. The frames are classified using a trained convolutional recurrent neural network, which takes as input 25 frames from drug-induced sleep endoscopy videos and outputs the classified upper airway region for each frame.

the model can recognize structures such as the tongue and epiglottis. It appears that with the limited amount of data the model has seen, it is not able to derive distance-based decisions to estimate the region as well as it recognizes structures associated with a given region. Furthermore, the many noisy frames in the video (approximately 25%) most likely cause noise in the context of the Bi-LSTM, which contributes to the poor performance for video 3.

Table 4.4: Performance for the three videos in the test set obtained by classifying upper airway regions with three different classes: velum (V), oropharynx, tongue, and epiglottis combined (OTE), and distorted video (X). The frames are classified using a trained convolutional recurrent neural network, which takes as input 25 frames from drug-induced sleep endoscopy videos and outputs the classified upper airway region for each frame.

Video	F1	Class	Class F1	N Frames
1	93%	V	94%	386
		OTE	91%	264
		X	-	0
2	75%	V	67%	166
		OTE	93%	741
		X	65%	93
3	62%	V	61%	252
		OTE	54%	123
		X	70%	125

4.3.2.3 Performance for Videos

Table 4.4 summarizes the performance for each of the three individual videos in the test set. It is observed that the best performance is obtained for video 1, which has no frames with distorted video and also few misclassifications for the V and OTE regions. During video 2, the endoscope is by far the most in the OTE region and the F1-score is high for that class. The F1-score for both V and X is modest because V is misclassified as both OTE and X, whereas OTE is misclassified a few times as X as well. During video 3, most time is spent in the V region but the F1 score is modest for all classes. In this case, V is still misclassified as both OTE and X, but OTE is also sometimes misclassified as V and not only X, which is most likely due to wrong labels, similar to the bottom frame in the middle column of Fig. 4.9, where V is labeled as OTE.

4.3.2.4 Limitations

There are two main limitations of this study: the quantity of data is extremely low, and the problem posed is simplistic with respect to applicability in clinical practice. However, the results serve as an important proof of concept, which shows that it is possible to apply deep learning techniques on DISE videos, even though they contain large variations in terms of anatomical structure and angles/positions in the upper airway across videos. Considering this, it is quite impressive that the proposed model achieves high performance on little data and that it manages to learn meaningful mappings between the classes and the series of frames that are used as input.

For a future study, we will obtain a much larger quantity of DISE videos and expand the problem to classify sites of upper airway collapse and obstruction degrees according to the VOTE classification system, which is a clinically more relevant endeavour compared to classifying upper airway regions in DISE.

4.3.3 New Analyses After Publication

After publishing Paper III and working on Paper IV, smaller modifications to the methodology were made. The primary change was to utilize a cross-validation scheme instead of a training, validation, and test set split. This revised approach was explored for the problem posed in Paper III after publication, primarily to evaluate performance across the entire dataset (24 videos) instead of on a small test set (3 videos).

Ideally, this approach would extend to the 281 videos used for Paper IV, however this would require a labeling of all videos with respect to upper airway regions, which would be too time-consuming. Thus, the cross-validation approach was applied for the 24 videos used for Paper III. Consequently, the results have more power and any biases that could be introduced in the test set when using a training, validation, and test split would be reduced. All results from the new analyses performed after publication are presented below.

4.3.3.1 Overall Performance

Performance was evaluated in the same way as shown in Table 4.4, but this time across the full dataset consisting of 24 videos, which yielded a mean F1 score of $66 \pm 20\%$. This was much lower than the performance reported in Paper III across 3 videos (mean F1 score of 79%), showing that the initial test set was too small and not representative of the overall dataset. Similarly, F1 scores for V, OTE, and X were $62 \pm 25\%$, $61 \pm 30\%$, and $53 \pm 25\%$, respectively, which was also lower than the F1 scores obtained for the three videos previously (74%, 79%, and 68%, respectively).

4.3.3.2 Performance for Videos

Table 4.5 outlines performance for the three videos in the previous test set obtained using the cross-validation approach. Interestingly, when comparing Table 4.5 to Table 4.4, it is observed that the F1 score dropped from 93% to 75% for video 1, increased from 75% to 91% for video 2, and remained unchanged for video 3. However, the class F1 scores relative to each other within each video were similar for both approaches.

The change in performance when using cross-validation is expected because splitting the data into five equal folds and using one fold for validation and another for testing yields less data for training, i.e. 60% training data, compared to the training, validation, and test set split approach used previously, which had 75% training data.

Figure 4.10 shows the distribution of F1 scores across the 24 videos in the dataset, which makes it evident that the proposed model's performance varies significantly from video to video, but that the majority of videos achieve F1 scores of 60% or above.

Table 4.5: Performance for the three videos in the previous test set for classifying upper airway regions with three different classes: velum (V), oropharynx, tongue, and epiglottis combined (OTE), and distorted video (X). The frames are classified using a trained convolutional recurrent neural network, which takes as input 25 frames from drug-induced sleep endoscopy videos and outputs the classified upper airway region for each frame.

Video	F1	Class	Class F1	N Frames
1	75%	V	74%	386
		OTE	76%	264
		X	-	0
2	91%	V	78%	166
		OTE	96%	741
		X	68%	93
3	61%	V	65%	252
		OTE	43%	123
		X	67%	125

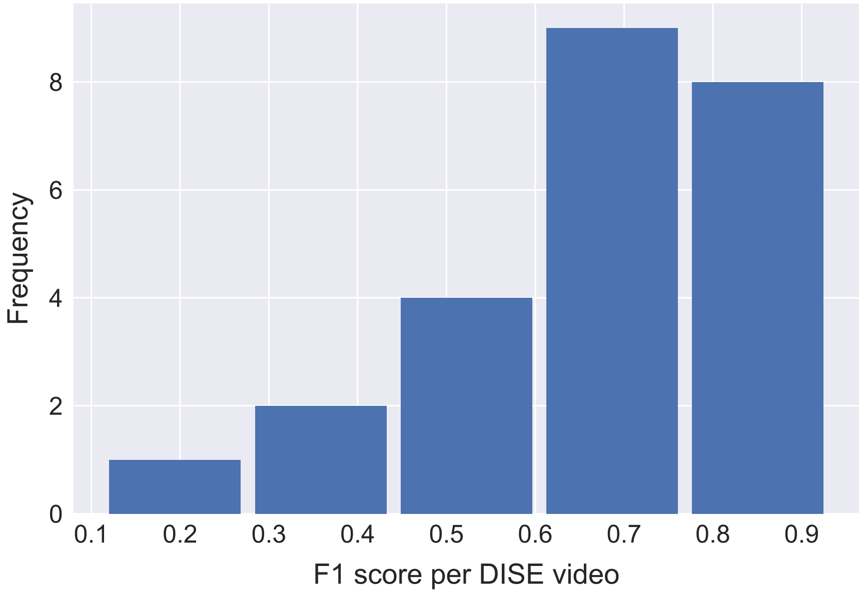


Figure 4.10: Distribution of F1 scores for 24 DISE videos calculated as the average F1 score of all frames making up each DISE examination with respect to three classes: velum (V), oropharynx, tongue, and epiglottis combined (OTE), and distorted video (X). The frames are classified using a trained convolutional recurrent neural network, which takes as input 25 frames from drug-induced sleep endoscopy videos and outputs the classified upper airway region for each frame.

4.3.3.3 Misclassified Frames

Figure 4.11 shows the confusion matrix for the classification of upper airway regions in 24 DISE videos. Comparing this to the confusion matrix shown in Fig. 4.8, it is observed that the sensitivity for V has increased from 71% to 77%, sensitivity for OTE has decreased from 88% to 64%, and sensitivity for X has decreased from 100% to 69%. These results again emphasize the importance of having a sufficiently large test set for evaluating performance of a machine learning algorithm such that the performance is not overestimated. This is particularly true when it comes to applying deep learning on DISE videos, since these videos present a lot of variation from subject to subject, not only with respect to the upper airway anatomy, but also in relation to the video quality and how the examination is filmed.

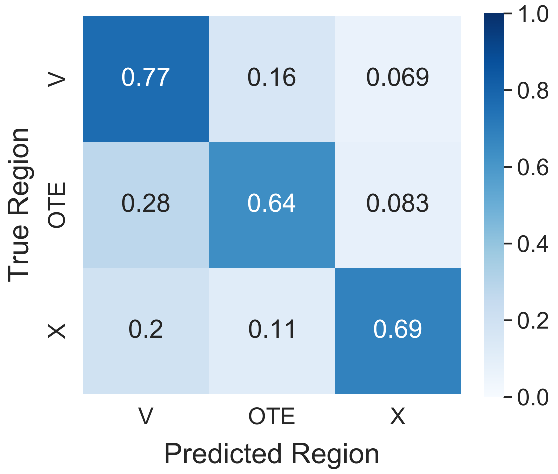


Figure 4.11: Normalized confusion matrix for classifying regions in the upper airway from 19800 frames with three different classes: velum (V), oropharynx, tongue, and epiglottis combined (OTE), and distorted video (X). The frames are classified using a trained convolutional recurrent neural network, which takes as input 25 frames from drug-induced sleep endoscopy videos and outputs the classified upper airway region for each frame.

4.4 Paper IV: Automatic Scoring of Drug-Induced Sleep Endoscopy for Obstructive Sleep Apnea Using Deep Learning

Abstract

Purpose: Treatment of obstructive sleep apnea is crucial for long term health and reduced economic burden. For those considered for surgery, drug-induced sleep endoscopy (DISE) is a method to characterize location and pattern of sleep-related upper airway collapse. According to the VOTE classification system, four upper airway sites of collapse are characterized: velum (V), oropharynx (O), tongue base (T), and epiglottis (E). The degree of obstruction per site is classified as 0 (no obstruction), 1 (partial obstruction), or 2 (complete obstruction). Here we propose a deep learning approach for automatic scoring of VOTE obstruction degrees from DISE videos.

Methods: We included 281 DISE videos with varying durations (6 seconds – 16 minutes) from two sleep clinics: Copenhagen University Hospital and Stanford University Hospital. Examinations were split into 5-second clips, each receiving annotations of 0, 1, 2, or X (site not visible) for each site (V, O, T, and E), which was used to train a deep learning model. Predicted VOTE obstruction degrees per examination were obtained by taking the highest predicted degree per site across 5-second clips, which were evaluated against VOTE degrees annotated by surgeons.

Results: Mean F1 score of 70% was obtained across all DISE examinations (V: 85%, O: 72%, T: 57%, E: 65%). For each site, sensitivity was highest for degree 2 and lowest for degree 0. No bias in performance was observed between videos from different surgeons/hospitals.

Conclusion: This study demonstrates that automating scoring of DISE examinations show high validity and feasibility in degree of upper airway collapse.

4.4.1 Methods

This section describes the approach for developing an automatic scoring system for DISE videos in OSA by estimating sites of upper airway collapse and obstruction degrees. Figure 4.12 shows a block diagram of the proposed system. The different components of the system will be described in detail in the following.

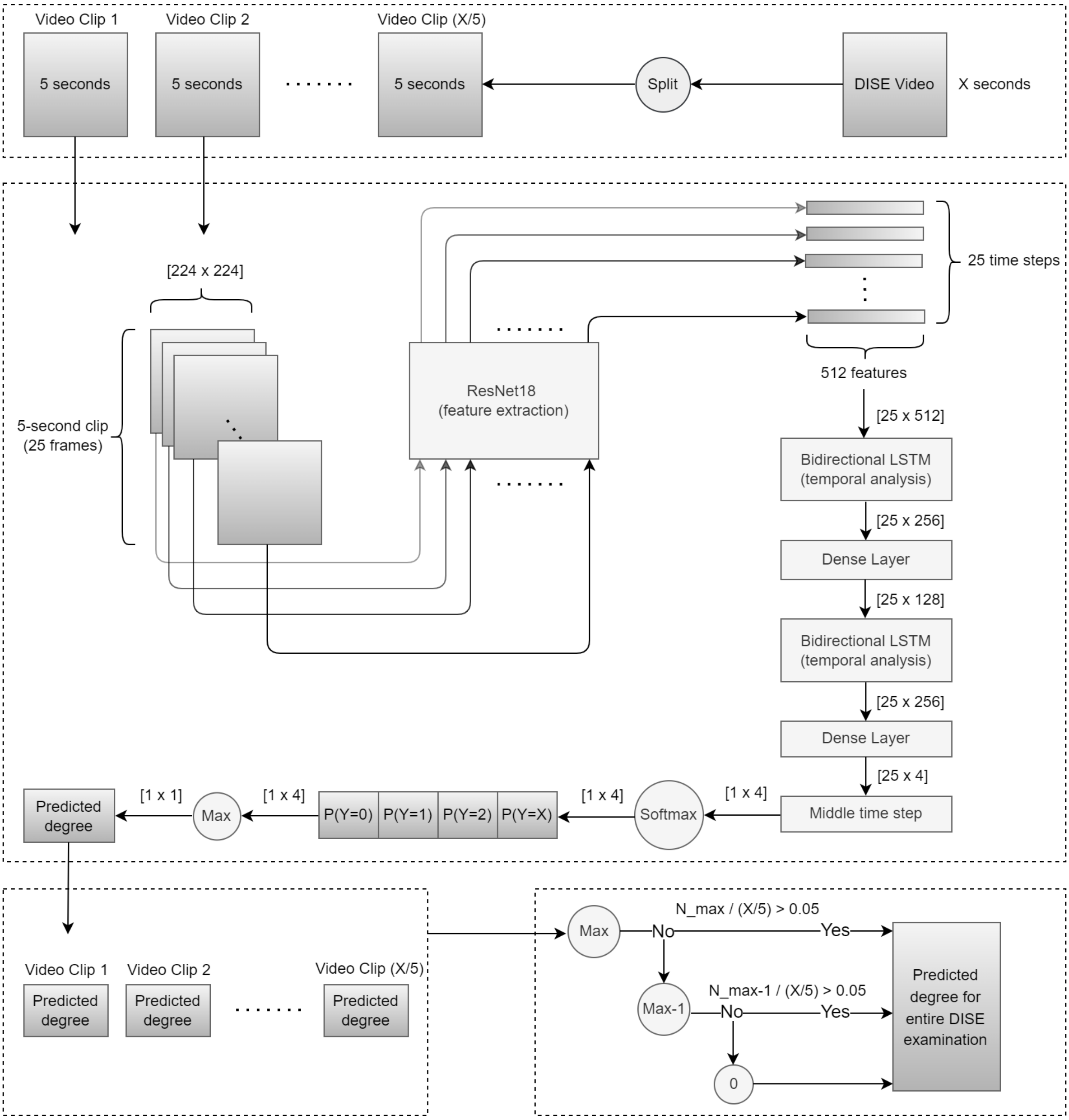


Figure 4.12: A block diagram of the proposed system for predicting obstruction degrees for each upper airway site in subjects using drug-induced sleep endoscopy (DISE) videos. This architecture is repeated for each of the four upper airway sites (velum, oropharynx, tongue base, and epiglottis). Top block: A DISE examination is split into 5-second clips. Middle block: Each individual frame (grayscale) of a 5-second clip is used as input one by one for a convolutional neural network (CNN) with a ResNet18 architecture for feature extraction. All resulting feature vectors (1x512) are concatenated (25x512) and input to a bidirectional long short-term memory network (Bi-LSTM) for temporal analysis, followed by a dense layer to reduce number of features (25x128). Another Bi-LSTM and dense layer reduce the feature vector (1x4), which is run through a softmax activation function. This yields four probabilities, one for each obstruction degree (P(Y=0), P(Y=1), P(Y=2), and P(Y=X), where X means that the site is not visible). The obstruction degree with highest probability is the predicted obstruction degree. Bottom left block: Predictions for all clips within a DISE examination are collected. Bottom right block: The maximum predicted obstruction degree across all 5-second clips that make up a full examination is chosen as the overall degree if the model predicts this degree for at least 5% of all clips. Otherwise, same criterion is checked for the next highest degree and if not fulfilled either, the predicted obstruction degree is 0 by default.

4.4.1.1 Data Collection

The data collection for this study was the same as for the previous study described in Section 4.3.1.1.

4.4.1.2 Data Description

281 DISE videos were obtained in total from three different otolaryngology - head and neck surgeons at two different locations: one surgeon from Copenhagen University Hospital (CUH) (51 videos) and two surgeons at Stanford University Hospital (SUH) (58 and 172 videos, respectively). The Institution's Ethical Review Board approved all experimental procedures involving human subjects.

Each video was anonymized by removing any part where the endoscope was outside of the patient. Median duration of videos after anonymization was 2.1 minutes with an interquartile range of 3.33 minutes (min – max: 6 seconds – 16.4 minutes) and the total amount of video footage was 13.7 hours. Figure 4.13 shows distribution of DISE examination durations, showing that most videos in the dataset are less than 2 minutes long. Videos obtained from CUH had sampling rates of 25 frames per second, while videos from SUH had sampling rates of 30 frames per second.

For each examination, an annotation was obtained containing the VOTE score, i.e., obstruction degree and collapse pattern at each site as shown in Table 4.6. Note that several sites can collapse in the same subject (also in combination) and that a site like V can collapse in more than one way in the same subject. The distribution of obstruction degrees for each site is shown in Fig. 4.14.

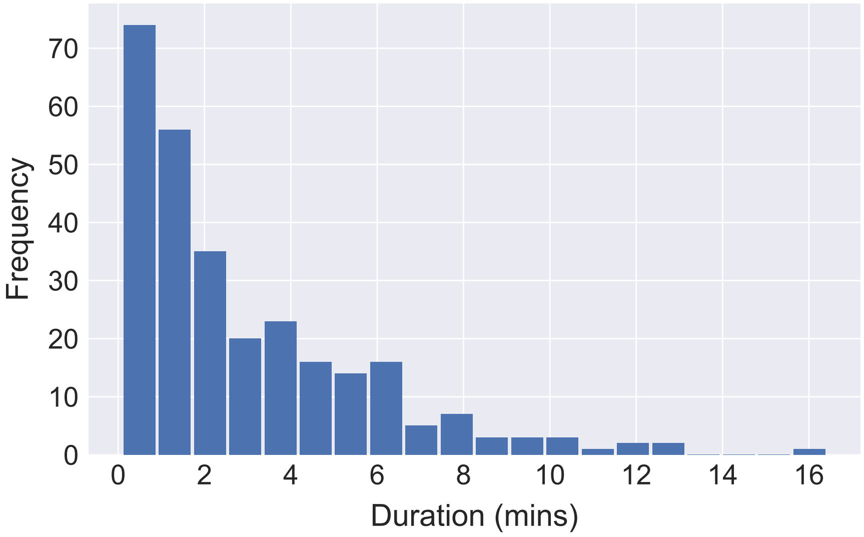


Figure 4.13: Distribution of drug-induced sleep endoscopy video durations for 281 videos in the dataset used to train and evaluate a convolutional recurrent neural network for predicting obstruction degrees for each upper airway site.

Table 4.6: Example of annotations provided by surgeons for drug-induced sleep endoscopy examinations. The obstruction degree for each upper airway site (velum, oropharynx, tongue base, and epiglottis) is indicated by a number (0, 1, or 2) followed by the collapse pattern (A-P, lateral or concentric).

Video	Velum	Oropharynx	Tongue base	Epiglottis
Video 1	2 A-P - Concentric	2 Lateral	1 A-P	2 Lateral
Video 2	2 A-P	2 Lateral	0	0
Video 3	1 A-P	0	2 A-P	2 A-P

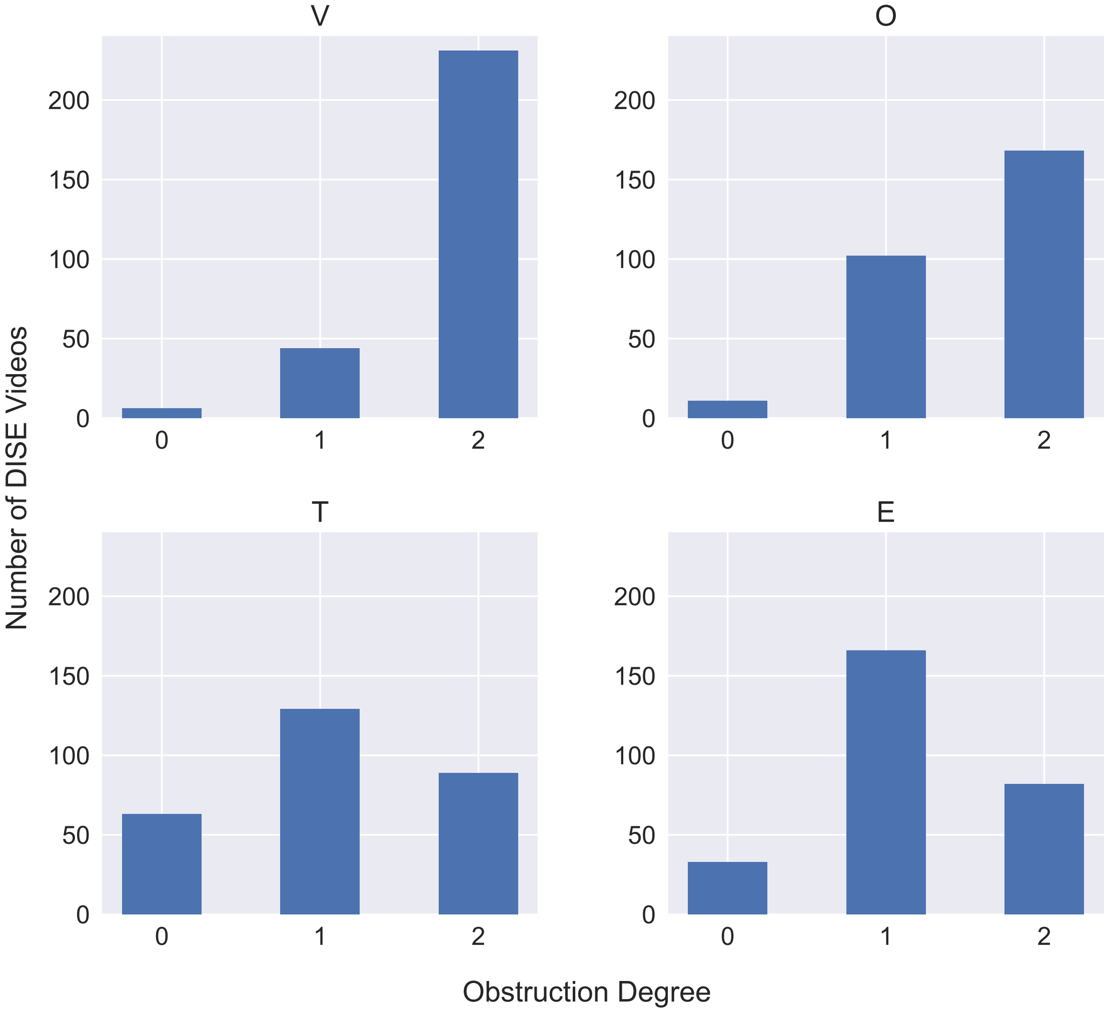


Figure 4.14: Distribution of obstruction degrees (0, 1, and 2) for each of the four upper airway sites (velum (V), oropharynx (O), tongue base (T), and epiglottis (E)) annotated by otolaryngology surgeons for 281 drug-induced sleep endoscopy videos.

4.4.1.3 Pre-Processing

Since DISE examination videos varied greatly with respect to duration (Fig. 4.13) and there was only a one-line annotation per video (Table 4.6), we decided that using data as it was would be unsuitable for deep learning purposes. Consequently, all DISE videos were split into 5-second clips, as shown in the top block of Fig. 4.12, and each clip received a label with respect to each upper airway site. These labels were created by the authors in consultation with a chief surgeon in otolaryngology at CUH [165]. Table 4.7 provides an example of the labels created for 5-second clips. Note that the term *annotations* is used when describing the scored obstruction degrees by the surgeons for an entire DISE video, while the term *labels* is used to describe the scored obstruction degrees by the authors for 5-second clips.

Using 5-second clips, there are many scenarios where one or more upper airway sites are not visible. Thus, another class was introduced for such situations, denoted X, such that there were four classes (0, 1, 2, and X) for each upper airway site (V, O, T, and E) per 5-second clip, essentially amounting to a 16-class classification problem. The idea was to train and evaluate the proposed computer vision model on the 5-second clips with corresponding labels and then summarize all predictions to form a single predicted obstruction degree per upper airway site for each DISE video. This, in turn, can then be evaluated against the surgeons' gold-standard annotations for each DISE video.

Figure 4.15 shows the distribution of labeled 5-second clips for each upper airway site with respect to obstruction degrees for all 9895 5-second clips. Note how the distributions are different from the ones presented in Fig. 4.14 because the longer DISE videos consist of more 5-second clips and thus contribute more to the number of labeled clips with a given degree. Whereas the distributions shown in Fig. 4.14 are based on entire DISE videos, i.e., one annotation for each site per video.

Although sampling rates for videos from RH and SUH were 25 and 30 frames per second, respectively, we used only every 5th and 6th frame, respectively. This was done to reduce computational cost, because no visual difference was observed between consecutive frames during inspection. Consequently, the sampling rate for 5-second clips used in this study was 5 frames per second, yielding a total of 25 frames for a 5-second clip. Instead of using all three color-channels for each clip (R,G,B),

Table 4.7: Example of obstruction degree labels (0, 1, 2, and X) created for 5-second clips of drug-induced sleep endoscopy with respect to the four upper airway sites that can collapse (velum, oropharynx, tongue base, and epiglottis). 0 means no collapse, 1 means partial collapse, 2 means complete collapse, and X means that a given upper airway site is not visible in that 5-second clip.

Video clip	Velum	Oropharynx	Tongue base	Epiglottis
Video 1 Clip 1	2	X	X	X
Video 1 Clip 2	0	1	1	1
Video 1 Clip 3	X	1	2	2

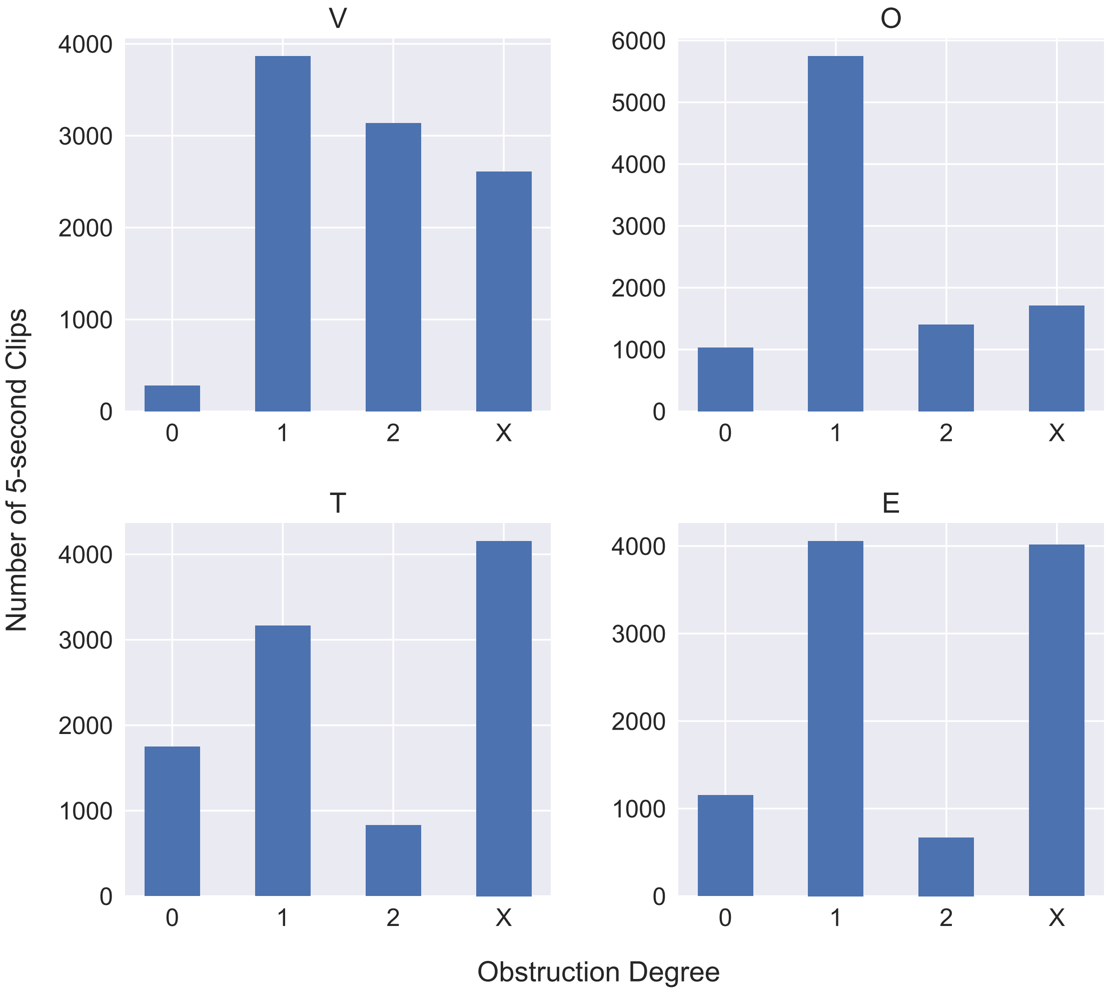


Figure 4.15: Distribution of obstruction degree labels (0, 1, 2, and X) created for 9895 5-second clips of drug-induced sleep endoscopy with respect to the four upper airway sites that can collapse (velum (V), oropharynx (O), tongue base (T), and epiglottis (E)). 0 means no collapse, 1 means partial collapse, 2 means complete collapse, and X means that a given upper airway site is not visible in that 5-second clip.

the videos were converted to grayscale. Both approaches were investigated (with and without color channels) and preserving colors did not make any noticeable difference, most likely because anatomical composition is much more important than small differences in color, so grayscale frames were used to reduce computational cost.

All frames in each 5-second clip were rescaled to 224×224 pixels, which was found to be appropriate for reducing computational cost while still preserving discriminatory information between upper airway sites. Finally, each frame was normalized to the range [0,1] by dividing each pixel value by 255. This was done to ensure faster convergence during training. The final input output pairs were 5-second clips, consisting of 25 frames, and corresponding labeled obstruction degrees (0, 1, 2, or X) for each upper airway site (V, O, T, and E).

4.4.1.4 Convolutional Recurrent Neural Network

The purpose of applying machine learning was to learn data-driven discriminatory information about upper airway collapse at the different sites and whether the collapse is partial or complete. For this purpose, we implemented a CNN with a ResNet18 architecture [130] combined with two Bi-LSTM layers [168] as illustrated in the middle block of Fig. 4.12. The network architecture was almost identical to the one presented in Section 4.3.1.4, where reasoning for the choice of network architecture was provided as well, except for two key differences: 1) here we take the middle time step of the second Bi-LSTM and dense layer output instead of using all time steps, since there is one predicted obstruction degree for each upper airway site per 5-second clip, instead of predictions for each frame, and 2) the number of output probabilities are 4 instead for 3, one for each obstruction degree (0, 1, 2, and X). The architecture was implemented four times, one for each upper airway site.

The CNN was implemented for automatic feature extraction from each video frame, while Bi-LSTM layers were included to include temporal context in both forward and backward directions for each frame. The Resnet18 network was implemented such that a 5-second clip (consisting of 25 frames) could be input one frame at a time. The output of the CNN was then a feature map of size 1x512 for each frame. The feature maps for all frames were concatenated to form a 25x512 matrix, where each row is considered a time-step in the original 5-second clip and each column is a feature vector for a particular frame. This matrix was processed by a Bi-LSTM layer, followed by a dense layer, which reduced the number of features from 512 to 128 while time steps were intact, i.e., resulting matrix dimensions were 25x128. A second bidirectional LSTM and dense layer reduced the number of features further from 128 to 4, yielding a matrix of dimension 25x4. From this matrix, the output at the middle time step, i.e., 13 was taken as it represents the time step where the model has most context in both directions. Finally, a softmax activation function was applied to the resulting 1x4 vector to yield a probability for each class, i.e., 0, 1, 2, and X.

The optimal number of time steps and hidden neurons in the Bi-LSTM layers were found using hyperparameter tuning, which was performed in a grid search-like manner where the hyperparameters were varied and different combinations of these were investigated. The optimal hyperparameters yielded the lowest error on the validation set, which is defined in the next section.

4.4.1.5 Training, Validation, and Testing

The proposed model was trained, validated, and tested using 10-fold cross-validation to get predictions for all DISE videos in the dataset. This was done by splitting the dataset into 10 folds of equal size and utilizing 8 folds for training the network and 1 fold each for validation and testing, respectively. The test fold predictions were stored and the process was repeated by assigning new folds to training, validation, and testing. This procedure was repeated 10 times, such that predicted VOTE obstruction degrees were obtained for all DISE videos in the dataset.

The loss function for each upper airway site was the cross-entropy loss defined in Eq. 4.1 with $K = 4$ classes. The loss functions for all four sites were added together and the combined loss was used to optimize the weights of the model using the Adam optimizer [133]. The combined loss function was used to optimize the model simultaneously with respect to all four upper airway sites. Since the obstruction degrees for each site were imbalanced (Fig. 4.15), penalty weights were introduced in the loss function during training. The weight for a particular class was calculated by dividing the number of samples for the most represented class with the number of samples for a particular class in the training set.

The model was trained using batches of size 8, where computational resources were the limiting factor. The learning rate was set to $1 \cdot 10^{-5}$ with a weight decay of $5 \cdot 10^{-4}$, which was found using hyperparameter tuning. Early stopping was applied with a patience of 3, to help prevent overfitting to the training data.

Python 3.6.10 and Pytorch 1.10.0 were used for implementation of the proposed model. Training of the model was performed using a GeForce RTX 3070 and the entire training, validation, and test setup took approximately 16 hours to run.

4.4.1.6 Post-Processing

Post-processing steps are illustrated in the bottom blocks of Fig. 4.12. For each upper airway site in each 5-second clip, the model predicted probabilities for each of the four different classes (0, 1, 2, and X). The predicted degree for each site was the one which the model predicted the highest probability for. After a prediction was made for each site for each 5-second clip, the overall degree for each site for a particular DISE examination was calculated as the maximum predicted degree across all 5-second clips constituting a single DISE examination.

The maximum degree was selected only if this degree was predicted in at least 5% of the clips which make up a full examination. This is to avoid any coincidences where a degree of e.g., 2 occurs one time by chance or because of other upper airway sites and does not reflect the true behavior of that site in a subject. In case the maximum degree did not satisfy this condition, the next greatest degree was selected if it satisfied the same condition. If this was not satisfied either, the degree was set to 0 by default. This is illustrated in the bottom right block of Fig 4.12.

A voting approach was not applied here, because surgeons annotate DISE examinations according to the highest degree observed for each upper airway site.

4.4.1.7 Performance Measures

The predicted VOTE obstruction degrees were compared to the surgeons' annotations, considered as ground truth, for all DISE videos in the dataset. Performance was evaluated using weighted F1 score [170]. Weighted F1 score was used instead of accuracy due to a large imbalance between the annotated obstruction degrees as observed in Fig. 4.14.

The F1 score was calculated using Eq. 4.2. Using a degree of 0 as example, TP represents the number of 0's that are correctly predicted as 0's, FP represents the number of predicted 0's that are not actually 0's, and FN represents the number of predicted 1's and 2's that are actually 0's. The F1 scores for degrees 1 and 2 were calculated similarly. The weighted F1 score was calculated by averaging the F1 score of the individual degrees multiplied by their proportion in the dataset.

Cohen's kappa [171] was also used to compare model performance with inter-rater reliability reported in the literature with respect to obstruction degrees, either for each site or overall.

4.4.2 Results and Discussion

Here we describe for the first time that deep learning can be used to reliably evaluate DISE videos with the goal of identifying site of collapse and extent of obstruction. Performance was evaluated across all 281 DISE videos in the dataset using 10-fold cross-validation.

4.4.2.1 Overall Performance

Mean F1 score for the 12-class problem, i.e., predicting obstruction degree (0, 1, or 2) for each of the four upper airway sites across all DISE videos, was 70% (V: 85%, O: 72%, T: 57%, E: 65%). If the model had instead predicted all examinations to have the most represented obstruction degree for each site (i.e., 2 for V, O, and 1 for T, E as seen in Fig. 4.14), the average F1 score would be only 48% (V: 74%, O: 45%, T: 29%, 44%).

Mean F1 score for the 16-class problem, i.e., predicting obstruction degrees for each individual 5-second clip (including the class X for when the site is not visible) was $65 \pm 14\%$ (V: $68 \pm 21\%$, O: $64 \pm 22\%$, T: $64 \pm 23\%$, E: $65 \pm 23\%$) and was calculated by averaging all clips that make up a full DISE examination and then averaging the performance over all 281 DISE videos. In contrast, if the model predicted the obstruction degrees for each 5-second clip to be the most represented obstruction degree for each site, respectively, the average F1 score would only be $20 \pm 13\%$ (V: $18 \pm 21\%$, O: $8 \pm 13\%$, T: $24 \pm 28\%$, E: $31 \pm 29\%$). This quantitative analysis supports the fact that the model performs much better than random guessing and emphasizes the large gap that would not be as apparent if a performance metric like accuracy would be used, which does not consider how well the model predicts individual classes.

When distinguishing between collapse/no collapse, i.e., combining obstruction degrees of 1 and 2 as one class, the F1 score increased to 90% (V: 98%, O: 95%, T: 78%, E: 91%). Similarly, for 5-second clips averaged over DISE examinations, where the classes are no collapse, collapse, and X, the F1 score increased to $74 \pm 13\%$ (V: $79 \pm 19\%$, O: $76 \pm 22\%$, T: $70 \pm 22\%$, E: $71 \pm 22\%$). The model predicts especially well whether or not there is a collapse as shown by an F1 score of 90% when combining degrees 1 and 2, showing that in general, the model confuses degrees 1 and 2 more often than 0 and 1 or 0 and 2.

4.4.2.2 Sensitivity for Each Obstruction Degree

Figure 4.16 shows confusion matrices for the model's predicted obstruction degree for each site evaluated against the surgeons' annotations across all DISE videos in the dataset. Figure 4.16 shows that the highest sensitivity for degree 0 is obtained for E (55%), and that most misclassifications occur because the model predicts degree 1. In very few cases (<10%), the model predicts degree 2 and after inspecting the three examinations in question, it occurs for two reasons: 1) E is reflected in saliva causing a mirror image where it looks like E is collapsing, which in fact it is not, and 2) the model confuses an A-P V collapse for an E collapse, particularly when the endoscope is close to the collapse. In these examinations, however, the predicted probability for degree 2 is never higher than 40% and it only occurs in 1-2 clips per examination for the three examinations.

The second highest sensitivity for degree 0 is obtained for T (35%), but it is confused for both degrees 1 and 2. When degree 2 is predicted, it occurs for two reasons: the uvula or lower part of the soft palate resembles the tongue and when it collapses, the model confuses it for the tongue collapsing, and 2) when V or O are collapsing and the endoscope is extremely close to the tissue, it resembles the tissue of the tongue.

The lowest sensitivity for degree 0 is obtained for V (17%), although there are only 6 out of 281 DISE examinations where V has a degree of 0. Presence of collapse at the level of V is extremely common among OSA patients [172] which is also evident in the dataset by the lack of videos where V has a degree of 0. However, it is encouraging that the model only confuses degree 0 with degree 1 and never predicts degree 2 in those cases.

The next lowest sensitivity for degree 0 is obtained for O (27%) and the confusion is equally split between degrees 1 and 2. Again, there are only a small number of DISE examinations where O has degree 0 (11 examinations), but the cases in which they are predicted as 2 are due to three reasons: 1) A lateral collapse at the level of E which is generally not considered part of O, (2) T collapsing completely and the endoscope being very close such that the model mistakes it for a collapse at O, and 3) V collapsing and the model mistakenly predicting a contribution of O as well, which can be difficult to assess even for surgeons [173].

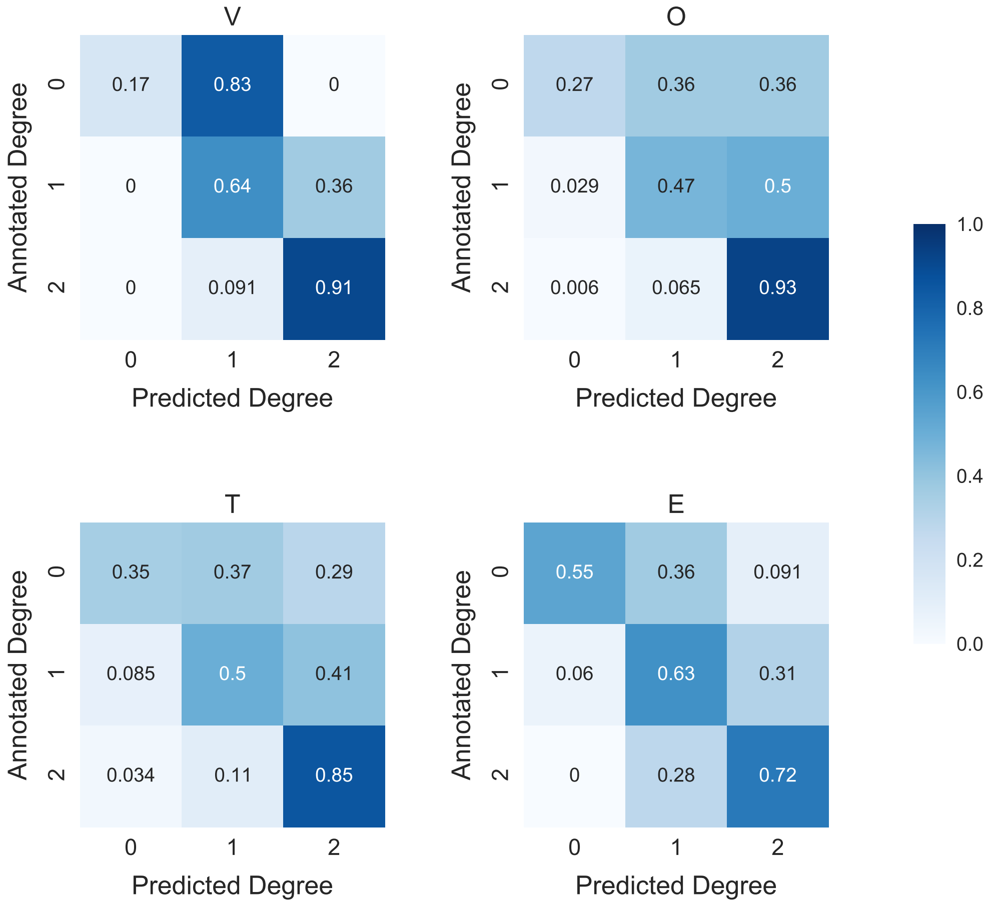


Figure 4.16: Confusion matrices for predicted obstruction degrees evaluated for 281 drug-induced sleep endoscopy (DISE) videos with respect to four different upper airway sites: velum (V), oropharynx (O), tongue base (T), and epiglottis (E). The obstruction degree for each site was predicted using a trained convolutional recurrent neural network, which takes a 5-second DISE clip as input and outputs the obstruction degree for each site. The predicted degrees for an entire DISE video were the maximum predicted degree for each site observed across all 5-second clips which the DISE video consisted of. The gold-standard annotations were provided by otolaryngology surgeons who performed the DISE procedure.

The highest sensitivity for degree 1 is obtained for both V and E (64% and 63%, respectively), while T and O are lower (50% and 47%, respectively). For all four sites, the model primarily confuses degree 1 with degree 2 and very rarely with degree 0 (none for V, <5% for O, <10% for T and E). For degree 2, the sensitivity is very high for V, O, and T (91%, 93%, and 85%, respectively), while the sensitivity for E is lower (72%). For all four sites, the model almost exclusively confuses degree 2 with degree 1 and almost never with degree 0 (none for V and E, <5% for O and T). Although the model confuses degree 1 with degree 2 to some extent for all sites, the model confuses degree 2 with degree 1 only for E, showing that for this site, the model appears to have most difficulty distinguishing between degrees 1 and 2.

4.4.2.3 Performance for Videos From Each Surgeon

Performance was also evaluated with respect to videos obtained from the three different surgeons (one from CUH and two from SUH) to investigate any biases in the proposed model towards videos from a particular surgeon. The results are summarized in Table 4.8, which shows that there is no noticeable difference in overall F1 score between videos obtained from each of the three surgeons, demonstrating that there is no meaningful bias towards any of them and suggesting that the procedures are comparable.

For videos from surgeon 1 (S1) from CUH, the model yields the highest F1 score for V out of all three but also the lowest F1 score for T and E. For videos obtained from surgeon 2 (S2) from SUH, the model has the highest F1 score for T out of all three and the lowest F1 score for O. However, the gap between the highest and lowest F1 score for S2 is much smaller than for S1 and in general the discrepancy in performance between sites is lower for videos from S2. For videos obtained from surgeon 3 (S3) from SUH, the model has the highest F1 scores for both O and E compared to S1 and S2 and the lowest F1 score for V. Again, the gap between highest and lowest F1 score is much smaller than for S1.

4.4.2.4 Performance for 5-Second Clips

Figure 4.17 shows the distribution of F1 scores per DISE examination as the average F1 score with respect to 5-second clips that make up an entire examination. It is noted that most examinations have an F1 score above 50% (86% of all videos), but the lowest F1 score is at 24%. There are three important factors which explain low performance for some examinations: 1) The duration of an examination, since a very short examination consists of only few 5-second clips and even a few misclassified clips reduce performance by a lot, 2) the video quality, since some examinations have very low quality, which makes it difficult to assess the degree for each site, and 3) several sites collapsing simultaneously, which pushes the endoscope around and makes the video appear chaotic.

Table 4.8: F1 scores for predicted obstruction degrees in drug-induced sleep endoscopy with respect to each of the three surgeons who provided the videos. Performance is evaluated against the surgeons' annotations with respect to obstruction degree (0, 1, or 2) for four different upper airway sites: velum (V), oropharynx (O), tongue base (T), and epiglottis (E). The obstruction degree for each site was predicted using a trained convolutional recurrent neural network, which takes a 5-second DISE clip as input and outputs the obstruction degree for each site. The predicted degrees for an entire DISE video were the maximum predicted degree for each site observed across all 5-second clips which the DISE video consisted of.

Surgeon	N Videos	V (F1)	O (F1)	T (F1)	E (F1)	Overall (F1)
S1 (CUH)	51	91%	70%	53%	58%	68%
S2 (SUH)	58	89%	64%	63%	63%	70%
S3 (SUH)	172	82%	74%	56%	67%	70%

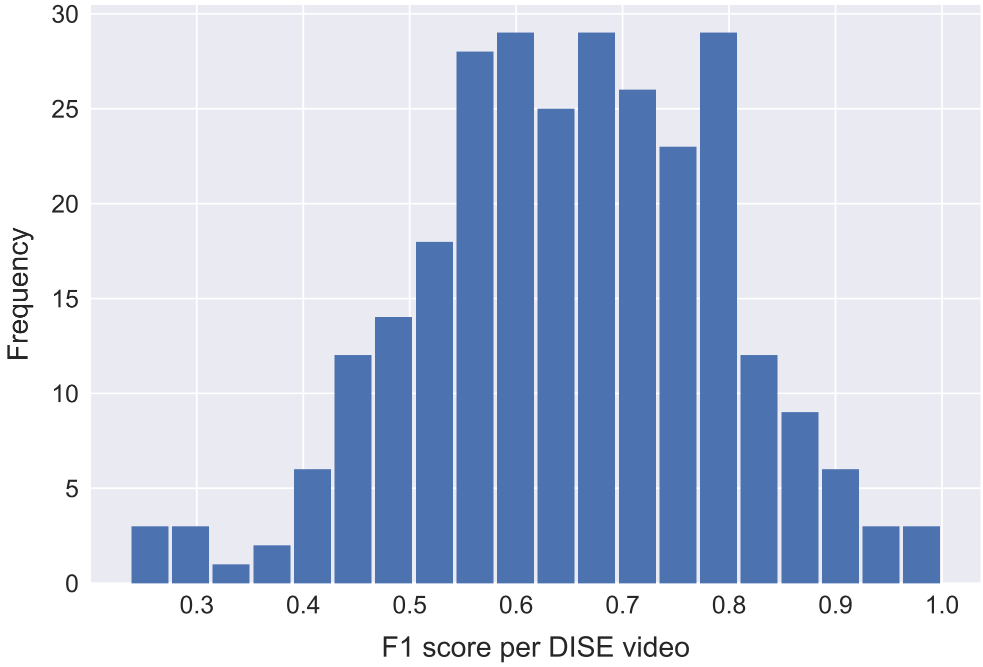


Figure 4.17: Distribution of F1 scores for 281 DISE videos calculated as the average F1 score of all 5-second clips making up each DISE examination with respect to four classes (0, 1, 2, and X) for the upper airway sites (velum, oropharynx, tongue base, and epiglottis). The obstruction degree for each site was predicted using a trained convolutional recurrent neural network, which takes a 5-second DISE clip as input and outputs the obstruction degree for each site.

4.4.2.5 Probability Interpretation

Figure 4.18, which depicts an example of predicted probabilities for each degree (and X) for each site over time, shows that the model regularly produces predictions with high confidence that are easy to interpret due to the way they covary within and across sites. Such a plot provides insight into model behavior and situations where the predictions are made with high or low confidence, which can then be directly compared to the DISE examinations. For example, for time (t) = 0 seconds (s), the model predicts with high probability that V has degree 2 and that OTE are not visible due to the obstruction at V. The probability of O not being visible is lower than for T and E because V reopens in that clip and reveals O for a split second. This observation emphasizes a limitation of the model: predictions are made on 5-second clips, during which several events can occur, causing the model to be less confident in a single obstruction degree. For t = 5 s, V transitions from degree 2 to 1, which enables the model to see OTE and predict degree 0 for all three. However, probabilities for O and E are lower because they are both on the border between having degrees 0 or 1. At t = 10 s, the model is confused whether V is not visible or has degree 1 because the endoscope is being moved down the airway and is at the border between V and O. For O, the model becomes more confident that the degree is 0 as the lateral walls separate further.

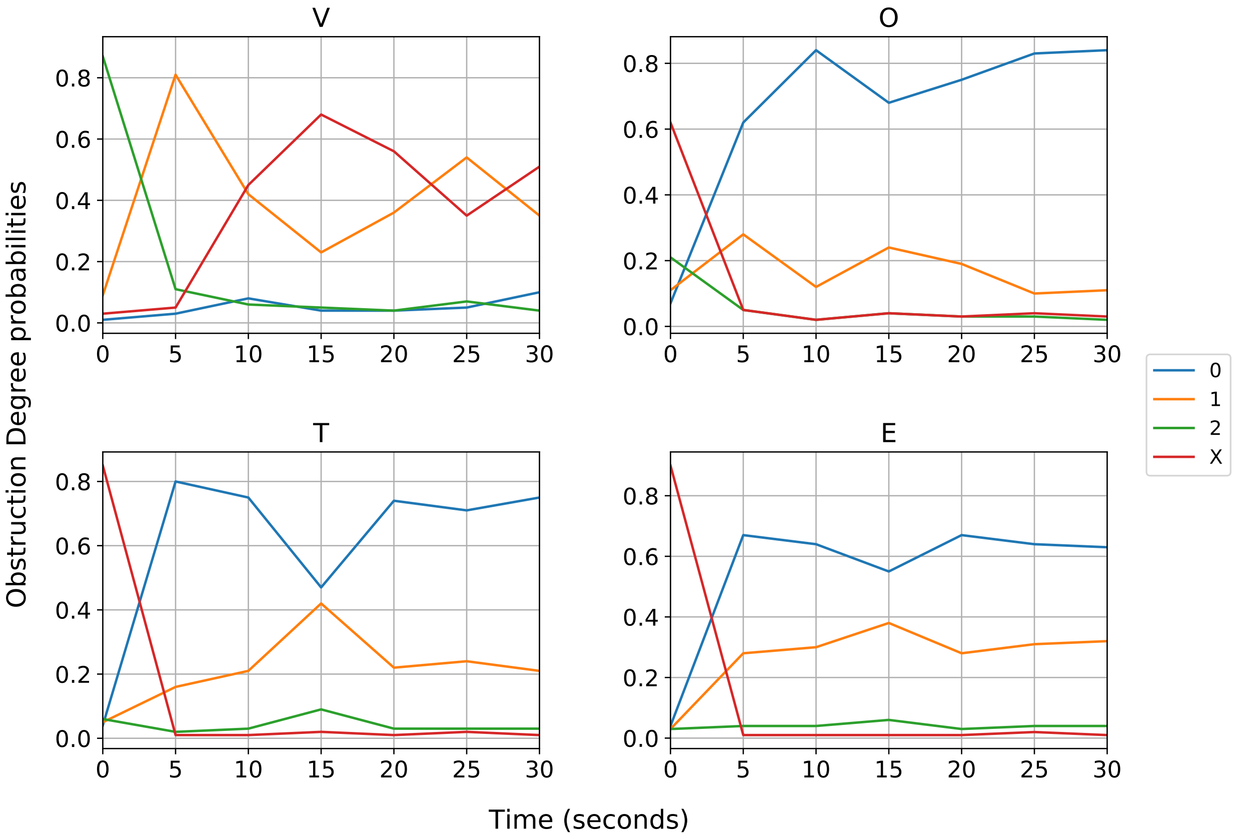


Figure 4.18: Probabilities for obstruction degrees (0, 1, 2) or X (site not visible) predicted by the proposed model for 5-second clips for four different upper airway sites: velum (V), oropharynx (O), tongue base (T), and epiglottis (E). The probability for each obstruction degree per site was predicted using a trained convolutional recurrent neural network, which takes a 5-second drug-induced sleep endoscopy clip as input and outputs the obstruction degree for each site.

At t = 15 s, probabilities for degree 0 decrease and degree 1 increase for OTE because these sites move slightly, but not enough to cause any obstruction. At the same time, the model becomes more confident that V is not visible as the camera is moved further down. For the remainder of the video (t = 20-30 s), the model becomes more confident that OTE have degrees 0 again, but for V, it switches between degree of 1 and X because, although the endoscope is further down in the airway, the uvula occasionally vibrates and becomes visible, causing the model to confuse whether V is visible or not.

4.4.2.6 Comparison to Inter-rater Reliability

Table 4.9 compares performance of the proposed model (in terms of Cohen's kappa) to inter-rater reliabilities reported in the literature between surgeons. The comparisons are made on three levels: for each individual site (V, O, T, and E), region-based (palate and hypopharynx), and overall (all sites combined). For region-based comparison, V was compared to the palate, and OTE were

Table 4.9: Comparison of model performance in terms of Cohen's kappa (κ) to inter-rater reliabilities between surgeons reported in the literature for scoring obstruction degrees in drug-induced sleep endoscopy (DISE). Some studies use palate vs. hypopharynx, where palate corresponds to velum (V) and hypopharynx corresponds to oropharynx (O), tongue base (T), and epiglottis (E) combined. One study reports overall κ across all sites. Our model was a convolutional recurrent neural network, which takes a 5-second DISE clip as input and predicts obstruction degrees for all four sites. The overall obstruction degrees for an entire DISE video were obtained as the maximum predicted degree for each site across all 5-second clips which the DISE video consists of.

Study	**V (κ)**	**O (κ)**	**T (κ)**	**E (κ)**	**N Scorers**	**N DISE**
Our model	**0.55**	**0.45**	**0.38**	**0.44**	**N/A**	**281**
Vroegop et al. [61]	0.30	0.66	0.03	0.61	7	6
Llatas et al. [63]	0.17	0.67	0.35	0.43	2	31
Green et al. [64]	0.40	0.42	0.60	0.55	4	275
Study	**Palate (κ)**	**Hypopharynx (κ)**			**N scorers**	**N DISE**
Our model	**0.55**	**0.43**			**N/A**	**281**
Kezirian et al. [60]	0.60	0.44			2	108
Koo et al. [65]	0.52	0.35			6	100
Study	**Overall (κ)**				**N scorers**	**N DISE**
Our model	**0.46**				**N/A**	**281**
Gillespie et al. [62]	0.27				3	38

combined for hypopharynx comparisons. For V, our model achieves a higher kappa than the inter-rater scores. For O and E, our model has lower kappa than two studies and slightly higher than one. For T, our model has higher kappa than two studies and lower than one, demonstrating that despite our model having low performance for T, surgeons can struggle with it as well which is evident from a huge difference in kappa scores between studies (0.03 - 0.60).

For both the palate and hypopharynx, our model produces higher kappa than one study and lower than another. For the overall evaluation of degrees, our model achieves a much higher kappa than the one study where they conduct such an analysis. This comparison adds context to the model performance and demonstrates that analysis of DISE examinations is not a trivial task, not even for experiences surgeons.

4.4.2.7 Limitations

There are two main limitations of this study: 1) the model is not able to predict the pattern of collapse, which would need to be added for the model to produce complete VOTE annotations as the surgeons do, 2) no healthy controls are used in the study, but the model could benefit from seeing more examples of absence of collapse, particularly for V and O. For the first limitation, the absence of collapse pattern only really affects predictions for V, since O and T only have one possible pattern and lateral obstructions for E are extremely rare [156]. However, the difference in collapse patterns for V (particularly concentric vs. A-P or lateral) can lead to different treatment strategies and is therefore important in clinical practice [174–176]. The second limitation is difficult to compensate for because DISE examinations are performed for people with confirmed OSA as the point of the procedure is to identify sites contributing to upper airway collapse prior to surgery. However, such data could be gathered by using DISE with other medical procedures under sedation.

4.4.2.8 Future Work

In future work, pattern of collapse for V should be added to the output of the model. This could be done by labeling the pattern for V for all 5-second clips, i.e., A-P, lateral, concentric, and X for when there is no collapse. A network identical to the ones predicting VOTE obstruction degrees could be implemented and trained simultaneously with the other networks. Future work would also benefit from using a larger, multi-scored dataset for testing purposes, such that the inter-rater reliability can be compared directly to the model's performance on the same data. Finally, future work should explore the potential of using self-supervised learning on DISE videos, which would require fewer labels and save time and resources compared to the approach followed in this study.

4.5 Conclusions

Based on the described methods, results, and discussions in Sections 4.3 and 4.4, the research questions posed in the beginning of the chapter are now restated and answered.

Research Question 1: Can a dedicated computer vision model be trained to accurately classify upper airway regions in DISE videos?

Research Conclusion 1: A framework was presented for classifying upper airway regions for each frame in DISE videos. This was achieved by labeling 24 DISE videos second by second with respect to where in the upper airway the endoscope is using three classes: velum (V), oropharynx, tongue, or epiglottis combined (OTE), or distortion in video (X). Then, videos were split into 5-second clips, consisting of 25 frames each with a corresponding label, and these input and output pairs were used to train a convolutional recurrent neural network for classifying upper airway regions. We successfully showed that upper airway regions can be classified automatically on a test set consisting of three videos, which yielded an overall F1 score of 79%, while class F1 scores of 74%, 79%, and 68% were obtained for V, OTE, and X, respectively. Subsequently, we implemented a cross-validation approach to obtain performance across all 24 videos in the dataset, and results showed that the initial performance was overestimated. The new F1 score was 66 ± 20%, while class F1 scores were 62 ± 25%, 61 ± 30%, and 53 ± 25% for V, OTE, and X, respectively. This serves as a proof of concept and the applied techniques can be extended to automatically score DISE videos for sites of upper airway collapse and obstruction degrees.

Research Question 2: Can a dedicated computer vision model be trained to accurately score DISE videos for sites of upper airway collapse and obstruction degrees?

Research Conclusion 2: An automatic system was presented for scoring DISE videos with respect to sites of upper airway collapse and obstruction degrees. This was achieved by first splitting 281 DISE videos into 5-second clips. Each clip was labeled according to an obstruction degree (0, 1, or 2) or X, which means that a site is not visible in the clip, for each upper airway site: velum (V), oropharynx (O), tongue base (T), and epiglottis (E). The framework developed for Research Question 1 was modified to yield four outputs and was repeated four times, one for each upper airway site. 5-second clips with corresponding labels were used to train the proposed model to automatically estimate obstruction degrees for each upper airway site. The maximum predicted obstruction degree for each site observed across an entire DISE recording was evaluated against the surgeons' annotations. We succesfully showed that it is possible to automatically score DISE videos using a dedicated computer vision model, which yielded an overall F1 score of 70% (V: 85%, O: 72%, T: 57%, E: 65%).

Research Question 3: Can the proposed model generalize well across DISE procedures performed at different sleep centers by different otolaryngology - head and neck surgeons?

Research Conclusion 3: The 281 DISE videos in the dataset were obtained from two different clinics and were performed by three different surgeons. We investigated performance for videos obtained from each surgeon and compared them to each other. Overall F1 scores of 68%, 70%, and 70% were obtained for each surgeon, respectively, showing no clear bias towards any of them and suggesting that the proposed model generalizes well across DISE procedures performed at different sleep centers and by different surgeons.

Research Question 4: Can the proposed model perform at a level similar to that of otolaryngology - head and neck surgeons with years of experience?

Research Conclusion 4: We compared performance of the proposed model to inter-rater reliabilities reported in the literature between surgeons scoring obstruction degrees for each upper airway site. For comparisons at each individual site, the proposed model in general performed better than surgeons for V and T and worse for O and E. For comparisons at the palate (V) and hypopharynx (OTE), the proposed model in general performed as well as surgeons. For overall comparison across all sites, the model performed notably better than the surgeons, leading to the conclusion that in general, the proposed model performs at a level similar to that of otolaryngology - head and neck surgeons with years of experience.

To summarize the findings and state the overall conclusion, Hypothesis 2 is restated below and answered:

Hypothesis 2

An automatic scoring system can be invented, based on dedicated computer vision models, which utilizes drug-induced sleep endoscopy examination videos to estimate sites of upper airway collapse and obstruction degrees in obstructive sleep apnea patients with a similar accuracy as otolaryngology - head and neck surgeons.

The potential for automatic scoring of DISE videos has been demonstrated by inventing a system based on dedicated computer vision models that is capable of estimating sites of upper airway collapse and obstruction degrees. The proposed system was trained and evaluated on 281 DISE videos and displayed solid performance in estimating obstruction degrees (0, 1, or 2) for each upper airway site (V, O, T, and E). Mean F1 score of 70% was obtained in estimating obstruction degrees across all four sites (V: 85%, O: 72%, T: 57%, E: 65%). Furthermore, we have shown that our model performs at a level similar to otolaryngology - head and neck surgeons with years of experience by comparing the model performance to inter-rater reliabilities reported in the literature. This comparison yielded better performance than some studies and lower performance than others, but in general appeared to be at a similar level as the surgeons. The proposed model has acquired this knowledge by seeing a relatively small sample of DISE videos (< 300) compared to standard current dataset sizes within deep learning and computer vision. A much larger dataset is required for training and evaluation of a scoring system such as the one proposed here before clinical applicability can be considered. However, we have demonstrated that such a scoring system can be invented and that sites of upper airway collapse and obstruction degrees can be obtained in less than a minute to provide surgeons with objective and data-driven conclusions, which can have a positive impact on treatment strategy.

Chapter 5

Conclusions

This chapter concludes the dissertation and discusses future perspectives of the current work. Section 5.1 relates the main research findings to the dissertation objective and research hypotheses presented in Chapter 1. Section 5.2 discusses limitations associated with the project and future perspectives that could improve it. Section 5.3 provides concluding remarks for the dissertation.

5.1 Research Conclusions

The focus of this dissertation was to invent automatic screening and scoring methods for obstructive sleep apnea (OSA) that rely on imaging data. Imaging data can be acquired in a fast and inexpensive manner compared to physiological signals that are traditionally utilized for OSA diagnosis and which require an entire night of sleep. The dissertation objective was stated follows:

Dissertation Objective

To invent automatic screening and scoring systems, based on dedicated computer vision models, that rely on imaging data to detect the presence of obstructive sleep apnea and characterize the upper airway collapse pattern in a fast, cheap, and data-driven manner.

Automatic and data-driven screening and characterization systems for OSA as presented in this dissertation provide several benefits to patients and clinicians if they are implemented and utilized in a clinical setting: 1) fast and cost-effective screening of OSA, based on modalities that can be captured in a few minutes, which performs better than current screening questionnaires, 2) less manual labor for clinicians due to the proposed systems being automatic, and 3) providing clinicians with a second opinion that is objective and data-driven, which has potential of increasing inter-rater reliability and yielding a better diagnosis and treatment plan for the patient.

The two research hypotheses stated in Section 1.4 were answered in Chapters 3 and 4, respectively. In the following, the main findings and conclusions related to each hypothesis are summarized:

Hypothesis 1

An automatic screening system can be invented, based on dedicated computer vision models, which utilizes 3D craniofacial scans to estimate presence and severity of obstructive sleep apnea more accurately than current screening questionnaires in a fast, cheap, and data-driven manner.

Chapter 3 presented an automatic system for screening of OSA based on dedicated computer vision models used on 3D craniofacial scans. In comparison to current screening methods, such as questionnaires consisting of OSA-related questions, the proposed system using 3D craniofacial scans performed better with an accuracy of 67% compared to 62% (obtained using a modified STOP-Bang questionnaire) in classifying subjects as being normal or having OSA, although both methods had the same area under receiver operating characteristic curve (AUC ROC) of 65%. However, the best performing system used a combination of three modalities (3D scans, demographics, and questionnaires), which yielded an accuracy of 67% but an improved AUC ROC of 69%. This approach is extremely fast and cost-efficient as any of the three modalities can be acquired in a few minutes and none of them require the presence of a physician. In this project, we have taken one step towards clinical applicability by showing that 1) the proposed system estimates apnea-hypopnea index (AHI) and OSA in a similar way as sleep medicine specialists with decades of experience when inspecting craniofacial anatomy, 2) the proposed system correctly identifies craniofacial regions that contribute to OSA development and that correspond perfectly with what is described in the medical literature, and 3) the proposed system generalizes well across data from different sleep clinics, as the dataset utilized to train and evaluate the proposed system comprises 3D scans and polysomnographies (PSGs) acquired from 11 different sites in the US and Canada.

Hypothesis 2

An automatic scoring system can be invented, based on dedicated computer vision models, which utilizes drug-induced sleep endoscopy examination videos to estimate sites of upper airway collapse and obstruction degrees in obstructive sleep apnea patients with a similar accuracy as otolaryngology - head and neck surgeons.

Chapter 4 presented an automatic scoring system for identifying sites of upper airway collapse and obstruction degrees in OSA patients based on dedicated computer vision models used on drug-induced sleep endoscopy (DISE) videos. To our knowledge, it is the first time that automated scoring of DISE has been attempted. The proposed system displayed solid performance with a mean weighted F1 score of 70% in estimating obstruction degrees across all four upper airway sites (velum (V): 85%, oropharynx (O): 72%, tongue base (T): 57%, epiglottis (E): 65%). Performance of the proposed system was compared to performance of otolaryngology - head and neck surgeons in scoring obstruction degrees for different upper airway sites. This was done by calculating Cohen's Kappa coefficients for the system and comparing to the inter-rater reliabilities reported in the literature. Results showed that the proposed system performs at a level similar to these surgeons. The proposed system is fast and cost-efficient, but requires a DISE video as input, which is more resource dependent than performing a 3D scan because it requires the presence of a surgeon and time to sedate the patient. Consequently, the benefit of this system in a clinical setting is that it provides surgeons with objective and data-driven scoring of DISE videos without the presence of subjective human biases that reduce the inter-rater reliability among surgeons. Once the surgeon has recorded a DISE video, it would take less than a minute to use the proposed system to obtain sites of upper airway collapse and obstruction degrees.

5.2 Limitations and Future Perspectives

Although the two systems presented in Chapters 3 and 4 performed well compared to medical experts within the fields of OSA and otolaryngology, both systems have limitations that should be improved in the future for better clinical applicability. The quantity of data used for training and evaluation of both systems should be increased, because they rely on anatomical data in which there is a large variation present across subjects. For computer vision models to see a sufficient number of examples and learn representative mappings for an entire population of OSA patients, we expect that both systems would benefit from a five-to-ten-fold increase in their respective dataset sizes, i.e. from 1366 3D scans to at least 10,000 scans, and from 281 DISE videos to at least 1000 videos.

Another important consideration that applies to both systems is the quality of data used for training and evaluation. The quality of 3D scans varied significantly and reflected the fact that they were captured in different sleep clinics and under different conditions. Some scans had missing parts of the neck, whereas others were affected by poor lighting conditions. Similarly, there was a large variation in the quality of DISE videos, where some videos were affected significantly by mucus or saliva on the camera lens while other videos had poor choice of lighting, making it difficult to see some upper airway sites. If data collection was dedicated towards developing systems such as the ones presented in this dissertation, more care would have been taken towards collecting good quality data and ensuring that all the data would have been collected under similar conditions.

For the system described in Chapter 3, the most inherent limitation of the approach is that the pathophysiology of OSA is not exclusively attributed to obesity and craniofacial features [1]. Upper airway anatomy factors, such as the size and positioning of the tongue and palate cannot be assessed with the proposed system as it only models the outer anatomy of a subject. A procedure like DISE, which is used to examine the internal anatomy, is performed after confirmed OSA and is therefore not used to screen for OSA [164]. A solution in future work could be to take a picture inside the subject's mouth or a picture of the tongue extending outwards to include additional information about potential factors contributing to OSA in the screening system. Other factors that the system may fail to capture are physiological changes that occur with age, such as recruitment of upper airway dilator muscles, which is independent of facial anatomy or obesity [50].

Although the proposed system detects OSA based exclusively on structural features, there are models that account for physiological factors of OSA, which should be considered in future efforts to develop a more complete screening system for OSA [177, 178]. Furthermore, it would be interesting to explore if, besides AHI, other clinically important variables captured by sleep studies could be better predicted, such as the hypoxic burden described by Azarbarzin et al. [152]. These could prove to not only make up a more accurate model but could also improve our knowledge of OSA phenotypes and their relation to facial anatomy. For example, the AHI receives criticism by some members of the sleep medicine community for being a frequency-based metric that does not adequately capture OSA severity and associated outcomes. New metrics that arise to complement the limitations of AHI can naturally be considered as target variables in future efforts investigating the diagnostic potential of 3D craniofacial scans in relation to OSA.

For the system described in Chapter 4, the biggest limitation is that it does not estimate pattern of collapse. The VOTE classification system categorizes three different variables: site of upper airway collapse, obstruction degree, and pattern of collapse [59]. The latter was purposefully excluded from the proposed system because the learning problem was growing complex enough as it was by estimating sites of collapse and obstruction degrees, which amounts to a 16-class classification problem when including the class X, which denotes when a particular site is not visible during a 5-second clip. However, extending the current framework to include pattern of collapse would not be difficult, as the pattern of collapse only varies at the level of the velum (antero-posterior, lateral, or concentric) and epiglottis (antero-posterior or lateral). Since lateral collapse is extremely rare for the epiglottis [156], we would only have to include pattern of collapse at the velum. This could be done by including another network identical to the four networks that were implemented for each site to predict obstruction degree, and labelling each 5-second video according to the pattern of collapse (antero-posterior, lateral, concentric, or X, which means there is no collapse). With inclusion of pattern of collapse in the proposed system, it would be able to estimate VOTE scores exactly as the surgeons do, which would increase the clinical applicability of the system even further.

For the system described in Chapter 4, a big limitation is the lack of a multi-scored DISE cohort annotated by several surgeons. Although performance of the proposed system was compared to inter-rater reliabilities between surgeons reported in the literature, this is only the second-best solution. Deriving inter-rater reliability from the same dataset that the system was trained and evaluated on would have enabled us to compare the system performance directly to the surgeons in a legitimate manner. The comparisons in Table 4.9 show that inter-rater reliabilities are estimated from as little as six DISE videos in one case, which gives less power to the results and takes away from the comparison to inter-rater reliabilities.

Finally, a limitation of the proposed system described in Chapter 4 was that it required a lot of time and resources for labeling of the entire dataset. All 281 DISE videos in the dataset were split into 5-second clips, which amounted to approximately 10,000 clips. The labeling of these clips took almost three months to complete, which is valuable time that could have been spent on other research activities. However, supervised deep learning requires an enormous quantity of labeled data, particularly as the task grows more complex. In future work, semi-supervised [179] or self-supervised [180] learning techniques should be explored within the problem of scoring DISE videos automatically. Although the computational requirements to train such models would increase dramatically and there would be no guarantee that it would work well, it is worth investigating the feasibility of a self-supervised approach, which would not require labeling of data, or a semi-supervised approach, which would require much fewer labels than the supervised approach. If additional videos could be obtained for future work, a semi-supervised approach could be utilized, where the existing labels would be included and there would be no need to create additional labels for the new videos. This would be a good compromise between supervised and self-supervised learning and would most likely lead to more successful results as the resulting model would benefit from using more data without the need for additional labels.

5.3 Concluding Remarks

In this dissertation, we showed the potential role of imaging in the detection and characterization of OSA. The advantage of using imaging data is that it can be captured in a few minutes compared to overnight sleep tracking. We invented two systems based on computer vision and demonstrated their clinical applicability. The advantages of both systems are that they present fast, cheap, and data-driven methods for screening OSA patients and characterizing upper airway collapse automatically. Future efforts should attempt to increase the dataset sizes, which the proposed systems are trained and evaluated on, to take one step further towards the implementation of such systems in a clinical setting.

Bibliography

[1] Patrick Lévy, Malcolm Kohler, Walter T McNicholas, Ferran Barbé, R Doug McEvoy, Virend K Somers, Lena Lavie, and Jean-Louis Pépin. Obstructive Sleep Apnoea Syndrome. *Nature Reviews Disease Primers*, 1(1):1–21, 2015.

[2] Abu S M Shamsuzzaman, Bernard J Gersh, and Virend K Somers. Obstructive Sleep Apnea Implications for Cardiac and Vascular Disease. *Jama*, 290(14):1906–1914, 2003.

[3] Yüksel Peker, Jan Hedner, Jeanette Norum, Holger Kraiczi, and Jan Carlson. Increased Incidence of Cardiovascular Disease in Middle-aged Men with Obstructive Sleep Apnea: A 7-Year Follow-up. *American Journal of Respiratory and Critical Care Medicine*, 166(2):159–165, 2002.

[4] Chengjuan Xie, Ruolin Zhu, Yanghua Tian, and Kai Wang. Association of obstructive sleep apnoea with the risk of vascular outcomes and all-cause mortality: a meta-analysis. *BMJ Open*, 7(12):e013983, 2017.

[5] Wojciech Trzepizur, Margaux Blanchard, Timothée Ganem, Frédéric Balusson, Mathieu Feuilloy, Jean Marc Girault, Nicole Meslier, Emmanuel Oger, Audrey Paris, Thierry Pigeanne, Jean Louis Racineux, Abdel Kebir Sabil, Chloé Gervès-Pinquié, and Frédéric Gagnadoux. Sleep Apnea–Specific Hypoxic Burden, Symptom Subtypes, and Risk of Cardiovascular Events and All-Cause Mortality. *American Journal of Respiratory and Critical Care Medicine*, 205(1):108–117, 2022.

[6] C. Gonzaga, A. Bertolami, M. Bertolami, C. Amodeo, and D. Calhoun. Obstructive sleep apnea, hypertension and cardiovascular diseases. *Journal of Human Hypertension*, 29(12):705–712, 2015.

[7] H. Klar Yaggi, John Concato, Walter N. Kernan, Judith H. Lichtman, Lawrence M. Brass, and Vahid Mohsenin. Obstructive Sleep Apnea as a Risk Factor for Stroke and Death. *New England Journal of Medicine*, 353(19):2034–2041, 2005.

[8] Susan Redline, Gayane Yenokyan, Daniel J. Gottlieb, Eyal Shahar, George T. O'Connor,

Helaine E. Resnick, Marie Diener-West, Mark H. Sanders, Philip A. Wolf, Estella M. Geraghty, Tauqeer Ali, Michael Lebowitz, and Naresh M. Punjabi. Obstructive Sleep Apnea–Hypopnea and Incident Stroke: The Sleep Heart Health Study. *American Journal of Respiratory and Critical Care Medicine*, 182(2):269–277, 2010.

[9] Sirimon Reutrakul and Babak Mokhlesi. Obstructive Sleep Apnea and Diabetes: A State of the Art Review. *Chest*, 152(5):1070–1086, 2017.

[10] Adam V Benjafield, Najib T Ayas, Peter R Eastwood, Raphael Heinzer, Mary S M Ip, Mary J Morrell, Carlos M Nunez, Sanjay R Patel, Thomas Penzel, Jean-Louis Pépin, Paul E Peppard, Sanjeev Sinha, Sergio Tufik, Kate Valentine, and Atul Malhotra. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet Respiratory Medicine*, 7(8):687–698, 2019.

[11] Vishesh Kapur, Kingman P. Strohl, Susan Redline, Conrad Iber, George O'Connor, and Javier Nieto. Underdiagnosis of Sleep Apnea Syndrome in U.S. Communities. *Sleep and Breathing*, 6(02):049–054, 2002.

[12] Claire Fuhrman, Bernard Fleury, Xuân Lan Nguyên, and Marie Christine Delmas. Symptoms of sleep apnea syndrome: High prevalence and underdiagnosis in the French population. *Sleep Medicine*, 13(7):852–858, 2012.

[13] Dayna A Johnson, Na Guo, Michael Rueschman, Rui Wang, James G Wilson, and Susan Redline. Prevalence and correlates of obstructive sleep apnea among African Americans: the Jackson Heart Sleep Study. *Sleep*, 41(10), 2018.

[14] Nathaniel F. Watson. Health Care Savings: The Economic Value of Diagnostic and Therapeutic Care for Obstructive Sleep Apnea. *Journal of Clinical Sleep Medicine*, 12(8):1075–1077, 2016.

[15] Jessica Vensel Rundo and Ralph Downey. Polysomnography. *Handbook of Clinical Neurology*, 160:381–392, 2019.

[16] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, Robin M Lloyd, Carole L Marcus, and Bradley V Vaughn. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Technical Specifications Version 2.2. *American Academy of Sleep Medicine*, 2015.

[17] Dirk A. Pevernagie, Barbara Gnidovec-Strazisar, Ludger Grote, Raphael Heinzer, Walter T. McNicholas, Thomas Penzel, Winfried Randerath, Sophia Schiza, Johan Verbraecken, and Erna S. Arnardottir. On the rise and fall of the apnea-hypopnea index: A historical review and critical appraisal. *Journal of Sleep Research*, 29(4):e13066, 2020.

[18] Atul Malhotra, Indu Ayappa, Najib Ayas, Nancy Collop, Douglas Kirsch, Nigel Mcardle, Reena Mehra, Allan I Pack, Naresh Punjabi, David P White, and Daniel J Gottlieb. Metrics of sleep apnea severity: beyond the apnea-hypopnea index. *Sleep*, 44(7), 2021.

[19] Nicholas Scalzitti, Shana Hansen, Stephen Maturo, Joshua Lospinoso, and Peter O'Connor. Comparison of home sleep apnea testing versus laboratory polysomnography for the diagnosis of obstructive sleep apnea in children. *International Journal of Pediatric Otorhinolaryngology*, 100:44–51, 2017.

[20] Nancy Collop. Scoring Variability between Polysomnography Technologists in Different Sleep Laboratories. *Sleep Medicine*, 3(1):43–47, 2002.

[21] Ulysses J. Magalang, Ning-Hung Chen, Peter A. Cistulli, Annette C. Fedson, Thorarinn Gíslason, David Hillman, Thomas Penzel, Renaud Tamisier, Sergio Tufik, Gary Phillips, Allan I. Pack, and for the SAGIC Investigators. Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers. *Sleep*, 36(4):591–596, 2013.

[22] Samuel T. Kuna, Ruth Benca, Clete A. Kushida, James Walsh, Magdy Younes, Bethany Staley, Alexandra Hanlon, Allan I. Pack, Grace W. Pien, and Atul Malhotra. Agreement in Computer-Assisted Manual Scoring of Polysomnograms across Sleep Centers. *Sleep*, 36(4):583–589, 2013.

[23] Atul Malhotra, Magdy Younes, Samuel T. Kuna, Ruth Benca, Clete A. Kushida, James Walsh, Alexandra Hanlon, Bethany Staley, Allan I. Pack, and Grace W. Pien. Performance of an Automated Polysomnography Scoring System Versus Computer-Assisted Manual Scoring. *Sleep*, 36(4):573–582, 2013.

[24] Valentin Thorey, Albert Bou Hernandez, Pierrick J. Arnal, and Emmanuel H. During. AI vs Humans for the diagnosis of sleep apnea. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 1596–1600, 2019.

[25] Jan B. Pietzsch, Abigail Garner, Lauren E. Cipriano, and John H. Linehan. An integrated health-economic analysis of diagnostic and therapeutic strategies in the treatment of moderate-to-severe obstructive sleep apnea. *Sleep*, 34(6):695–709, 2011.

[26] Richard D. Kim, Vishesh K. Kapur, Julie Redline-Bruch, Michael Rueschman, Dennis H. Auckley, Ruth M. Benca, Nancy R. Foldvary-Schafer, Conrad Iber, Phyllis C. Zee, Carol L. Rosen, Susan Redline, and Scott D. Ramsey. An Economic Evaluation of Home Versus Laboratory-Based Diagnosis of Obstructive Sleep Apnea. *Sleep*, 38(7):1027–1037, 2015.

[27] Russell Rosenberg, Max Hirshkowitz, David M. Rapoport, and Meir Kryger. The role of home

sleep testing for evaluation of patients with excessive daytime sleepiness: focus on obstructive sleep apnea and narcolepsy. *Sleep Medicine*, 56:80–89, 2019.

[28] Hsiao Yean Chiu, Pin Yuan Chen, Li Pang Chuang, Ning Hung Chen, Yu Kang Tu, Yu Jung Hsieh, Yu Chi Wang, and Christian Guilleminault. Diagnostic accuracy of the Berlin questionnaire, STOP-BANG, STOP, and Epworth sleepiness scale in detecting obstructive sleep apnea: A bivariate meta-analysis. *Sleep Medicine Reviews*, 36:57–70, 2017.

[29] Babak Amra, Behzad Rahmati, Forogh Soltaninejad, and Awat Feizi. Screening Questionnaires for Obstructive Sleep Apnea: An Updated Systematic Review. *Oman Medical Journal*, 33(3):184, 2018.

[30] Gabriele B. Papini, Pedro Fonseca, Merel M. van Gilst, Jan W. M. Bergmans, Rik Vullings, and Sebastiaan Overeem. Wearable monitoring of sleep-disordered breathing: estimation of the apnea–hypopnea index using wrist-worn reflective photoplethysmography. *Scientific Reports*, 10(1):1–15, 2020.

[31] Yibing Chen, Weifang Wang, Yutao Guo, Hui Zhang, Yundai Chen, and Lixin Xie. A Single-Center Validation of the Accuracy of a Photoplethysmography-Based Smartwatch for Screening Obstructive Sleep Apnea. *Nature and Science of Sleep*, 13:1533, 2021.

[32] Hisham Elmoaqet, Mohammad Eid, Martin Glos, Mutaz Ryalat, and Thomas Penzel. Deep Recurrent Neural Networks for Automatic Detection of Sleep Apnea from Single Channel Respiration Signals. *Sensors*, 20(18):5037, 2020.

[33] T S Dimitrov, M He, M J Prerau, D Yao, L Chieng, R Chiang, J Thybo, A N Olesen, M Olsen, E Leary, P Arnal, H B Sørensen, P Jennum, and E Mignot. 0451 Fully Automatic Detection of Sleep Disordered Breathing Events. *Sleep*, 43:A172–A173, 2020.

[34] Junming Zhang, Zhen Tang, Jinfeng Gao, Li Lin, Zhiliang Liu, Haitao Wu, Fang Liu, and Ruxian Yao. Automatic detection of obstructive sleep apnea events using a deep CNN-LSTM model. *Computational Intelligence and Neuroscience*, 2021.

[35] S. M. Isuru Niroshana, Xin Zhu, Keijiro Nakamura, and Wenxi Chen. A fused-image-based approach to detect obstructive sleep apnea using a single-lead ECG and a 2D convolutional neural network. *PLOS ONE*, 16(4):e0250618, 2021.

[36] Manish Sharma, Divyash Kumbhani, Jainendra Tiwari, T. Sudheer Kumar, and U. Rajendra Acharya. Automated detection of obstructive sleep apnea in more than 8000 subjects using frequency optimized orthogonal wavelet filter bank with respiratory and oximetry signals. *Computers in Biology and Medicine*, 144:105364, 2022.

[37] Yanxia Xu, Qiong Ou, Yilu Cheng, Miaochan Lao, and Guo Pei. Comparative study of a wearable intelligent sleep monitor and polysomnography monitor for the diagnosis of obstructive sleep apnea. *Sleep and Breathing*, 1:1–8, 2022.

[38] Richard W W Lee, Kate Sutherland, and Peter A Cistulli. Craniofacial Morphology in Obstructive Sleep Apnea: A Review. *Clinical Pulmonary Medicine*, 17(4):189–195, 2010.

[39] Richard W W Lee, Andrew S L Chan, Ronald R Grunstein, and Peter A Cistulli. Craniofacial Phenotyping in Obstructive Sleep Apnea—A Novel Quantitative Photographic Approach. *Sleep*, 32(1):37–45, 2009.

[40] Richard W W Lee, Peter Petocz, Tania Prvan, Andrew S L Chan, Ronald R Grunstein, and Peter A Cistulli. Prediction of Obstructive Sleep Apnea with Craniofacial Photographic Analysis. *Sleep*, 32(1):46–52, 2009.

[41] Fernando Espinoza-Cuadros, Rubén Fernández-Pozo, Doroteo T Toledano, José D Alcázar-Ramírez, Eduardo López-Gonzalo, and Luis A Hernández-Gómez. Speech Signal and Facial Image Processing for Obstructive Sleep Apnea Assessment. *Computational and Mathematical Methods in Medicine*, 2015.

[42] Hadis Nosrati, Nadi Sadr, and Philip de Chazal. Apnoea-Hypopnoea Index Estimation using Craniofacial Photographic Measurements. *CinC 2016, IEEE*, pages 1033–1036, 2016.

[43] Philip De Chazal, Asghar Tabatabaei Balaei, and Hadis Nosrati. Screening Patients for Risk of Sleep Apnea using Facial Photographs. *EMBC 2017, IEEE*, pages 2006–2009, 2017.

[44] Asghar Tabatabaei Balaei, Kate Sutherland, Peter A Cistulli, and Philip de Chazal. Automatic Detection of Obstructive Sleep Apnea using Facial Images. *ISBI 2017, IEEE*, pages 215–218, 2017.

[45] Asghar Tabatabaei Balaei, Kate Sutherland, Peter Cistulli, and Philip de Chazal. Prediction of Obstructive Sleep Apnea using Facial Landmarks. *Physiological Measurement*, 39:094004, 2018.

[46] Syed M S Islam, Hassan Mahmood, Adel Ali Al-Jumaily, and Scott Claxton. Deep Learning of Facial Depth Maps for Obstructive Sleep Apnea Prediction. *iCMLDE 2018, IEEE*, pages 154–157, 2018.

[47] Ignacio Perez-Pozuelo, Bing Zhai, Joao Palotti, Raghvendra Mall, Michaël Aupetit, Juan M. Garcia-Gomez, Shahrad Taheri, Yu Guan, and Luis Fernandez-Luque. The future of sleep health: a data-driven revolution in sleep science and medicine. *npj Digital Medicine*, 3(1):1–15, 2020.

[48] Massimiliano De Zambotti, Nicola Cellini, Aimée Goldstone, Ian M. Colrain, and Fiona C. Baker. Wearable Sleep Technology in Clinical and Research Settings. *Medicine and Science in Sports and Exercise*, 51(7):1538, 2019.

[49] Amir Abrishami, Ali Khajehdehi, and Frances Chung. A systematic review of screening questionnaires for obstructive sleep apnea. *Canadian Journal of Anesthesia*, 57(5):423–438, 2010.

[50] Amal M Osman, Sophie G Carter, Jayne C Carberry, and Danny J Eckert. Obstructive sleep apnea: current perspectives. *Nature and Science of Sleep*, 10:21–34, 2018.

[51] Alan R. Schwartz, Susheel P. Patil, Alison M. Laffan, Vsevolod Polotsky, Hartmut Schneider, and Philip L. Smith. Obesity and Obstructive Sleep Apnea: Pathogenic Mechanisms and Therapeutic Approaches. *Proceedings of the American Thoracic Society*, 5(2):185–192, 2008.

[52] Jerome A Dempsey, Sigrid C Veasey, Barbara J Morgan, and Christopher P O'donnell. Pathophysiology of Sleep Apnea. *Physiological Reviews*, 90(1):47–112, 2010.

[53] TL Giles, TJ Lasserson, BJ Smith, J White, J Wright, and CJ Cates. Continuous positive airways pressure for obstructive sleep apnoea in adults. *Cochrane Database of Systematic Reviews*, (1), 2006.

[54] Kim J, Tran K, Seal K, Fernanda A, Glenda R, Messier R, Tsoi B, Garland S, Rader T, Duthie K, Bond K, Mann J, and Kaunelis D. Interventions for the Treatment of Obstructive Sleep Apnea in Adults: A Health Technology Assessment. *Canadian Agency for Drugs and Technologies in Health*, 2019.

[55] David Sheen and Saif Abdulateef. Uvulopalatopharyngoplasty. *Oral and Maxillofacial Surgery Clinics*, 33(2):295–303, 2021.

[56] Thorbjörn Holmlund, Karl A. Franklin, Eva Levring Jäghagen, Marie Lindkvist, Torbjörn Larsson, Carin Sahlin, and Diana Berggren. Tonsillectomy in adults with obstructive sleep apnea. *The Laryngoscope*, 126(12):2859–2862, 2016.

[57] Soroush Zaghi, Jon Erik C. Holty, Victor Certal, Jose Abdullatif, Christian Guilleminault, Nelson B. Powell, Robert W. Riley, and Macario Camacho. Maxillomandibular Advancement for Treatment of Obstructive Sleep Apnea: A Meta-analysis. *JAMA Otolaryngology–Head & Neck Surgery*, 142(1):58–66, 2016.

[58] W. Hohenhorst, M. J.L. Ravesloot, E. J. Kezirian, and N. De Vries. Drug-induced sleep endoscopy in adults with sleep-disordered breathing: Technique and the VOTE Classification system. *Operative Techniques in Otolaryngology-Head and Neck Surgery*, 23(1):11–18, 2012.

[59] Eric J Kezirian, Winfried Hohenhorst, and Nico de Vries. Drug-Induced Sleep Endoscopy: the VOTE Classification. *European Archives of Oto-Rhino-Laryngology*, 268(8):1233–1236, 2011.

[60] Eric J. Kezirian, David P. White, Atul Malhotra, Wendy Ma, Charles E. McCulloch, and Andrew N. Goldberg. Interrater Reliability of Drug-Induced Sleep Endoscopy. *Archives of Otolaryngology–Head & Neck Surgery*, 136(4):393–397, 2010.

[61] Anneclaire V.M.T. Vroegop, Olivier M. Vanderveken, Kristien Wouters, Evert Hamans, Marijke Dieltjens, Nele R. Michels, Winfried Hohenhorst, Eric J. Kezirian, Bhik T. Kotecha, Nico De Vries, Marc J. Braem, and Paul H. Van De Heyning. Observer Variation in Drug-Induced Sleep Endoscopy: Experienced Versus Nonexperienced Ear, Nose, and Throat Surgeons. *Sleep*, 36(6):947–953, 2013.

[62] M. Boyd Gillespie, Ryan P. Reddy, David R. White, Christopher M. Discolo, Frank J. Overdyk, and Shaun A. Nguyen. A trial of drug-induced sleep endoscopy in the surgical management of sleep-disordered breathing. *The Laryngoscope*, 123(1):277–282, 2013.

[63] Marina Carrasco-Llatas, Vanessa Zerpa-Zerpa, and José Dalmau-Galofre. Reliability of drug-induced sedation endoscopy: interobserver agreement. *Sleep and Breathing*, 21(1):173–179, 2017.

[64] Katherine K. Green, David T. Kent, Mark A. D'Agostino, Paul T. Hoff, Ho Sheng Lin, Ryan J. Soose, M. Boyd Gillespie, Kathleen L. Yaremchuk, Marina Carrasco-Llatas, B. Tucker Woodson, Ofer Jacobowitz, Erica R. Thaler, José E. Barrera, Robson Capasso, Stanley Yung Liu, Jennifer Hsia, Daljit Mann, Taha S. Meraj, Jonathan A. Waxman, and Eric J. Kezirian. Drug-Induced Sleep Endoscopy and Surgical Outcomes: A Multicenter Cohort Study. *The Laryngoscope*, 129(3):761–770, 2019.

[65] Soo Kweon Koo, Sang Hoon Lee, Tae Kyung Koh, Young Jun Kim, Ji Seung Moon, Ho Byung Lee, and Geun Hyung Park. Inter-rater reliability between experienced and inexperienced otolaryngologists using Koo's drug-induced sleep endoscopy classification system. *European Archives of Oto-Rhino-Laryngology*, 276(5):1525–1531, 2019.

[66] Kannan Ramar, Raman K Malhotra, Kelly A Carden, Jennifer L Martin, Fariha Abbasi-Feinberg, R Nisha Aurora, Vishesh K Kapur, Eric J Olson, Carol L Rosen, James A Rowley, Anita V Shelgikar, and Lynn Marie Trotti. Sleep is essential to health: an American Academy of Sleep Medicine position statement. *Journal of Clinical Sleep Medicine*, 17(10):2115–2119, 2021.

[67] Emmanuel Mignot. Why we sleep: The temporal organization of recovery. *PLoS Biology*, 6(4):661–669, 2008.

[68] Oxana Semyachkina-Glushkovskaya, Dmitry Postnov, Thomas Penzel, and Jürgen Kurths. Sleep as a Novel Biomarker and a Promising Therapeutic Target for Cerebral Small Vessel Disease: A Review Focusing on Alzheimer's Disease and the Blood-Brain Barrier. *International Journal of Molecular Sciences*, 21(17):6293, 2020.

[69] Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.

[70] Max Hirshkowitz, Kaitlyn Whiton, Steven M. Albert, Cathy Alessi, Oliviero Bruni, Lydia DonCarlos, Nancy Hazen, John Herman, Eliot S. Katz, Leila Kheirandish-Gozal, David N. Neubauer, Anne E. O'Donnell, Maurice Ohayon, John Peever, Robert Rawding, Ramesh C. Sachdeva, Belinda Setters, Michael V. Vitiello, J. Catesby Ware, and Paula J. Adams Hillard. National Sleep Foundation's sleep time duration recommendations: methodology and results summary. *Sleep Health*, 1(1):40–43, 2015.

[71] Nathaniel F. Watson, M. Safwan Badr, Gregory Belenky, Donald L. Bliwise, Orfeu M. Buxton, Daniel Buysse, David F. Dinges, James Gangwisch, Michael A. Grandner, Clete Kushida, Raman K. Malhotra, Jennifer L. Martin, Sanjay R. Patel, Stuart F. Quan, and Esra Tasali. Recommended Amount of Sleep for a Healthy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society. *Journal of Clinical Sleep Medicine*, 11(06):591–592, 2015.

[72] Connor M. Sheehan, Stephen E. Frochen, Katrina M. Walsemann, and Jennifer A. Ailshire. Are U.S. adults reporting less sleep?: Findings from sleep duration trends in the National Health Interview Survey, 2004-2017. *Sleep*, 42(2), 2019.

[73] Jeffrey S. Durmer and David F. Dinges. Neurocognitive consequences of sleep deprivation. *Seminars in Neurology*, 25(1):117–129, 2005.

[74] Seung Schik Yoo, Peter T. Hu, Ninad Gujar, Ferenc A. Jolesz, and Matthew P. Walker. A deficit in the ability to form new human memories without sleep. *Nature Neuroscience*, 10(3):385–392, 2007.

[75] Janet M. Mullington, Monika Haack, Maria Toth, Jorge M. Serrador, and Hans K. Meier-Ewert. Cardiovascular, Inflammatory, and Metabolic Consequences of Sleep Deprivation. *Progress in Cardiovascular Diseases*, 51(4):294–302, 2009.

[76] M. A. Miller and F. P. Cappuccio. Biomarkers of cardiovascular risk in sleep-deprived people. *Journal of Human Hypertension*, 27(10):583–588, 2013.

[77] Kazuomi Kario, Satoshi Hoshide, Michiaki Nagai, Yukie Okawara, and Hiroshi Kanegae. Sleep and cardiovascular outcomes in relation to nocturnal hypertension: the J-HOP Nocturnal Blood Pressure Study. *Hypertension Research*, 44(12):1589–1596, 2021.

[78] Daniel J. Gottlieb, Susan Redline, F. Javier Nieto, Carol M. Baldwin, Anne B. Newman, Helaine E. Resnick, and Naresh M. Punjabi. Association of Usual Sleep Duration With Hypertension: The Sleep Heart Health Study. *Sleep*, 29(8):1009–1014, 2006.

[79] David A. Calhoun and Susan M. Harding. Sleep and Hypertension. *Chest*, 138(2):434–443, 2010.

[80] Dae Lim Koo, Hyunwoo Nam, Robert J. Thomas, and Chang Ho Yun. Sleep Disturbances as a Risk Factor for Stroke. *Journal of Stroke*, 20(1):12, 2018.

[81] Kristen L. Knutson, Karine Spiegel, Plamen Penev, and Eve Van Cauter. The metabolic consequences of sleep deprivation. *Sleep Medicine Reviews*, 11(3):163–178, 2007.

[82] Sarah K. Davies, Joo Ern Ang, Victoria L. Revell, Ben Holmes, Anuska Mann, Francesca P. Robertson, Nanyi Cui, Benita Middleton, Katrin Ackermann, Manfred Kayser, Alfred E. Thumser, Florence I. Raynaud, and Debra J. Skene. Effect of sleep deprivation on the human metabolome. *Proceedings of the National Academy of Sciences of the United States of America*, 111(29):10761–10766, 2014.

[83] Michael R. Irwin, Minge Wang, Capella O. Campomayor, Alicia Collado-Hidalgo, and Steve Cole. Sleep Deprivation and Activation of Morning Levels of Cellular and Genomic Markers of Inflammation. *Archives of Internal Medicine*, 166(16):1756–1762, 2006.

[84] Michael R. Irwin, Richard Olmstead, and Judith E. Carroll. Sleep Disturbance, Sleep Duration, and Inflammation: A Systematic Review and Meta-Analysis of Cohort Studies and Experimental Sleep Deprivation. *Biological Psychiatry*, 80(1):40–52, 2016.

[85] Marco Hafner, Martin Stepanek, Jirka Taylor, Wendy M. Troxel, and Christian van Stolk. Why Sleep Matters—The Economic Costs of Insufficient Sleep: A Cross-Country Comparative Analysis. *Rand Health Quarterly*, 6(4), 2017.

[86] David Hillman, Scott Mitchell, Jared Streatfeild, Chloe Burns, Dorothy Bruck, and Lynne Pezzullo. The economic cost of inadequate sleep. *Sleep*, 41(8):1–13, 2018.

[87] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalography and clinical neurophysiology. Supplement*, 52:3–6, 1999.

[88] Donnell J. Creel. The electrooculogram. *Handbook of Clinical Neurology*, 160:495–499, 2019.

[89] Robert G. Norman, Muhammed M. Ahmed, Joyce A. Walsleben, and David M. Rapoport. Detection of Respiratory Events During NPSG: Nasal Cannula/Pressure Sensor Versus Thermistor. *Sleep*, 20(12):1175–1184, 1997.

[90] R. Farré, J. M. Montserrat, and D. Navajas. Noninvasive monitoring of respiratory mechanics during sleep. *European Respiratory Journal*, 24(6):1052–1060, 2004.

[91] M. A. C. Garcia and T. M. M. Vieira. Surface electromyography: Why, when and how to use it. *Rev Andal Med Deporte*, 4(1):17–28, 2011.

[92] Muhammad Najjar. The Utility of Recording Submental Electrical Activity in Polysomnography. *Cureus*, 13(8), 2021.

[93] Erna S. Arnardottir, Bardur Isleifsson, Jon S. Agustsson, Gunnar A. Sigurdsson, Magdalena O. Sigurgunnarsdottir, Gudjon T. Sigurdarson, Gudmundur Saevarsson, Atli T. Sveinbjarnarson, Sveinbjorn Hoskuldsson, and Thorarinn Gislason. How to measure snoring? A comparison of the microphone, cannula and piezoelectric sensor. *Journal of Sleep Research*, 25(2):158–168, 2016.

[94] Rafael Ortega, Christopher J Hansen, Kelly Elterman, and Albert Woo. Videos in clinical medicine: Pulse oximetry. *New England Journal of Medicine*, 364:33, 2011.

[95] Daniel E. Becker. Fundamentals of Electrocardiography Interpretation. *Anesthesia Progress*, 53(2):53, 2006.

[96] Dmitriy Kogan, Arad Jain, Shawn Kimbro, Guillermo Gutierrez, and Vivek Jain. Respiratory inductance plethysmography improved diagnostic sensitivity and specificity of obstructive sleep apnea. *Respiratory Care*, 61(8):1033–1037, 2016.

[97] https://www.sleep-apnea-guide.com/images/polysomnogram.jpg. Polysomnogram - Sleep Apnea Guide.

[98] MA Carskadon and WC Dement. Normal human sleep: an overview. *Principles and Practice of Sleep Medicine*, 4(1):13–23, 2005.

[99] Benjamin D. Yetton, Elizabeth A. McDevitt, Nicola Cellini, Christian Shelton, and Sara C. Mednick. Quantifying sleep architecture dynamics and individual differences using big data and Bayesian networks. *PLoS ONE*, 13(4), 2018.

[100] Péter Halász, Mario Terzano, Liborio Parrino, and Róbert Bódizs. The nature of arousal in sleep. *Journal of Sleep Research*, 13(1):1–23, 2004.

[101] Danny J. Eckert, Amy S. Jordan, Pankaj Merchia, and Atul Malhotra. Central Sleep Apnea: Pathophysiology and Treatment. *Chest*, 131(2):595, 2007.

[102] Shahrokh Javaheri, Jason Smith, and Eugene Chung. The Prevalence and Natural History of Complex Sleep Apnea. *Journal of Clinical Sleep Medicine*, 5(3):205–211, 2009.

[103] https://wholisticdentistry.com.au/obstructive-sleep-apnea/. Obstructive Sleep Apnea – Wholistic Dentistry.

[104] Stephen Tregear, James Reston, Karen Schoelles, and Barbara Phillips. Obstructive Sleep Apnea and Risk of Motor Vehicle Crash: Systematic Review and Meta-Analysis. *Journal of Clinical Sleep Medicine*, 5(6):573–581, 2009.

[105] Nikolaus C. Netzer, Riccardo A. Stoohs, Cordula M. Netzer, Kathryn Clark, and Kingman P. Strohl. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. *Annals of Internal Medicine*, 131(7):485–491, 1999.

[106] Frances Chung, Balaji Yegneswaran, Pu Liao, Sharon A. Chung, Santhira Vairavanathan, Sazzadul Islam, Ali Khajehdehi, and Colin M. Shapiro. STOP Questionnaire: A Tool to Screen Patients for Obstructive Sleep Apnea. *Anesthesiology*, 108(5):812–821, 2008.

[107] Frances Chung, Hairil R Abdullah, and Pu Liao. STOP-Bang Questionnaire: A Practical Approach to Screen for Obstructive Sleep Apnea. *Chest*, 149(3):631–638, 2016.

[108] M. W. Johns. A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale. *Sleep*, 14(6):540–545, 1991.

[109] Marijke Dieltjens and Olivier M. Vanderveken. Oral Appliances in Obstructive Sleep Apnea. *Healthcare*, 7(4):141, 2019.

[110] BA Stuck, T Eschenhagen, and U Sommer. Uvulopalatopharyngoplasty with or without tonsillectomy in the treatment of adult obstructive sleep apnea–A systematic review. *Elsevier*, 50:152–165, 2018.

[111] Umaer Hanif, Rasmus R Paulsen, Eileen B Leary, Emmanuel Mignot, Poul Jennum, and Helge B D Sorensen. Prediction of Patient Demographics using 3D Craniofacial Scans and Multi-view CNNs. In *EMBC 2020, IEEE*, pages 1950–1953. IEEE, 2020.

[112] Umaer Hanif, Eileen Leary, Logan Schneider, Rasmus Paulsen, Anne Marie Morse, Adam Blackman, Paula Schweitzer, Clete A. Kushida, Stanley Liu, Poul Jennum, Helge Sorensen, and Emmanuel Mignot. Estimation of Apnea-Hypopnea Index Using Deep Learning on 3-D Craniofacial Scans. *IEEE Journal of Biomedical and Health Informatics*, 25(11):4185–4194, 2021.

[113] Richard J Schwab, Sarah E Leinwand, Cary B Bearn, Greg Maislin, Ramya Bhat Rao,

Adithya Nagaraja, Stephen Wang, and Brendan T Keenan. Digital Morphometrics: a New Upper Airway Phenotyping Paradigm in OSA. *Chest*, 152(2):330–342, 2017.

[114] Christian Guilleminault, Robert Riley, and Nelson Powell. Obstructive Sleep Apnea and Abnormal Cephalometric Measurements: Implications for Treatment. *Chest*, 86(5):793–794, 1984.

[115] Andrew Jamieson, Christian Guilleminault, Markku Partinen, and Maria Antonia Quera-Salva. Obstructive Sleep Apneic Patients Have Craniomandibular Abnormalities. *Sleep*, 9(4):469–477, 1986.

[116] Alan A Lowe, John A Fleetham, Satoshi Adachi, and C.Francis Ryan. Cephalometric and Computed Tomographic Predictors of Obstructive Sleep Apnea Severity. *American Journal of Orthodontics and Dentofacial Orthopedics*, 107(6):589–595, 1995.

[117] Takumi Ogawa, Reyes Enciso, Werner H Shintaku, and Glenn T Clark. Evaluation of Cross-Section Airway Configuration of Obstructive Sleep Apnea. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, 103(1):102–108, 2007.

[118] Richard J Schwab, Michael Pasirstein, Robert Pierson, Adonna Mackley, Robert Hachadoorian, Raanan Arens, Greg Maislin, and Allan I Pack. Identification of Upper Airway Anatomic Risk Factors for Obstructive Sleep Apnea with Volumetric Magnetic Resonance Imaging. *American Journal of Respiratory and Critical Care Medicine*, 168(5):522–530, 2003.

[119] Mau Okubo, Masaaki Suzuki, Atsushi Horiuchi, Shinichi Okabe, Katsuhisa Ikeda, Shuichi Higano, Hideo Mitani, Wataru Hida, Toshimitsu Kobayashi, and Junji Sugawara. Morphologic Analyses of Mandible and Upper Airway Soft Tissue by MRI of Patients with Obstructive Sleep Apnea Hypopnea Syndrome. *Sleep*, 29(7):909–915, 2006.

[120] Richard W W Lee, Kate Sutherland, Andrew S L Chan, Biao Zeng, Ronald R Grunstein, M Ali Darendeliler, Richard J Schwab, and Peter A Cistulli. Relationship between Surface Facial Dimensions and Upper Airway Structures in Obstructive Sleep Apnea. *Sleep*, 33(9):1249–1254, 2010.

[121] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[122] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2014.

[123] B. Venkatesh and J. Anuradha. A review of Feature Selection and its methods. *Cybernetics and Information Technologies*, 19(1):3–26, 2019.

[124] M. Kalantari and M. Nechifor. Accuracy and utility of the Structure Sensor for collecting 3D indoor information. *Geo-spatial Information Science*, 19(3):202–209, 2016.

[125] Paul G.M. Knoops, Caroline A.A. Beaumont, Alessandro Borghi, Naiara Rodriguez-Florez, Richard W.F. Breakey, William Rodgers, Freida Angullia, N. U. Owase Jeelani, Silvia Schievano, and David J. Dunaway. Comparison of three-dimensional scanner systems for craniomaxillofacial imaging. *Journal of Plastic, Reconstructive and Aesthetic Surgery*, 70(4):441–449, 2017.

[126] Guo Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 25(10):1351–1358, 2018.

[127] Rasmus R Paulsen, Kristine Aavild Juhl, Thilde Marie Haspang, Thomas Hansen, Melanie Ganz, and Gudmundur Einarsson. Multi-View Consensus CNN for 3D Facial Landmark Placement. *Asian Conference on Computer Vision*, pages 706–719, 2018.

[128] Syed Zulqarnain Gilani, Faisal Shafait, and Ajmal Mian. Shape-Based Automatic Detection of a Large Number of 3D Facial Landmarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4639–4648, 2015.

[129] Carl Martin Grewe and Stefan Zachow. Fully Automated and Highly Accurate Dense Correspondence for Facial Surfaces. In *European Conference on Computer Vision*, pages 552–568. Springer, 2016.

[130] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CVPR 2016*, pages 770–778, 2016.

[131] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):1–74, 2021.

[132] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 2012.

[133] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980 [cs.LG]*, 2014.

[134] Douglas G Altman and J Martin Bland. Measurement in Medicine: the Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32:307–317, 1983.

[135] Frank Q. Nuttall. Body mass index: obesity, BMI, and health: a critical review. *Nutrion Today*, 50(3):117, 2015.

[136] A Romero-Corral, SM Caples, F Lopez-Jimenez, and VK Somers. Interactions between obesity and obstructive sleep apnea: implications for treatment. *Chest*, 137(3):711–719, 2010.

[137] E. Kocabey, M. Camurcu, F. Ofli, Y. Aytar, J. Marin, A. Torralba, and I. Weber. Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media | Proceedings of the International AAAI Conference on Web and Social Media, 2017.

[138] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. 2015.

[139] Lingyun Wen and Guodong Guo. A computational approach to body mass index prediction from face images. *Image and Vision Computing*, 31(5):392–400, 2013.

[140] Raphael Angulu, Jules R. Tapamo, and Aderemi O. Adewumi. Age estimation via face images: a survey. *Eurasip Journal on Image and Video Processing*, (1), 2018.

[141] K Ito, H Kawai, T Okano, and T Aoki. Age and gender prediction from face images using convolutional neural network. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 7–11, 2018.

[142] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034 [cs.CV]*, 2013.

[143] Danny J. Eckert, David P. White, Amy S. Jordan, Atul Malhotra, and Andrew Wellman. Defining phenotypic causes of obstructive sleep apnea: Identification of novel therapeutic targets. *American Journal of Respiratory and Critical Care Medicine*, 188(8):996–1004, 2013.

[144] Fábio Mendonça, Sheikh Shanawaz Mostafa, Antonio G. Ravelo-García, Fernando Morgado-Dias, and Thomas Penzel. Devices for home detection of obstructive sleep apnea: A review. *Sleep Medicine Reviews*, 41:149–160, 2018.

[145] Hiroshi Nakano, Tomokazu Furukawa, and Takeshi Tanigawa. Tracheal sound analysis using a deep neural network to detect sleep apnea. *Journal of Clinical Sleep Medicine*, 15(8):1125–1133, 2019.

[146] Shota Hayashi, Meiyo Tamaoka, Tomoya Tateishi, Yuki Murota, Ibuki Handa, and Yasunari Miyazaki. A New Feature with the Potential to Detect the Severity of Obstructive Sleep

Apnoea via Snoring Sound Analysis. *International Journal of Environmental Research and Public Health*, 17(8):2951, 2020.

[147] Sami Nikkonen, Isaac O. Afara, Timo Leppänen, and Juha Töyräs. Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea. *Scientific Reports*, 9(1):1–9, 2019.

[148] Fábio Mendonça, Shanawaz Mostafa, Fernando Morgado-Dias, and Antonio G Ravelo-García. An Oximetry Based Wireless Device for Sleep Apnea Detection. *Sensors*, 20(3):888, 2020.

[149] Ali Al-Naji, Ali J. Al-Askery, Sadik Kamel Gharghan, and Javaan Chahl. A System for Monitoring Breathing Activity Using an Ultrasonic Radar Detection with Low Power Consumption. *Journal of Sensor and Actuator Networks*, 8(2):32, 2019.

[150] A Procházka, M Schätz, O Ťupa, M Yadollahi, O Vyšata, and M Walls. The MS kinect image and depth sensors use for gait features detection. *2014 IEEE International Conference on Image Processing*, pages 2271–2274, 2014.

[151] Ibrahim Sadek, Terry Tan Soon Heng, Edwin Seet, and Bessam Abdulrazak. A new approach for detecting sleep apnea using a contactless bed sensor: Comparison study. *Journal of Medical Internet Research*, 22(9):e18297, 2020.

[152] Ali Azarbarzin, Scott A Sands, Katie L Stone, Luigi Taranto-Montemurro, Ludovico Messineo, Philip I Terrill, Sonia Ancoli-Israel, Kristine Ensrud, Shaun Purcell, and David P White. The Hypoxic Burden of Sleep Apnoea Predicts Cardiovascular Disease-Related Mortality: the Osteoporotic Fractures in Men Study and the Sleep Heart Health Study. *European Heart Journal*, 40:1149–1157, 2019.

[153] Umaer Hanif, Eric Kezirian, Eva Kirkegaard Kiar, Emmanuel Mignot, Helge B.D. Sorensen, and Poul Jennum. Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 3957–3960, 2021.

[154] T Vauterin, G Garas, and A Arora. Transoral robotic surgery for obstructive sleep apnoea-hypopnoea syndrome. *ORL*, 80(3-4):134–147, 2018.

[155] E Van de Perck, C Heiser, and O M Vanderveken. Concentric versus anteroposterior-laterolateral collapse of the soft palate in patients with obstructive sleep apnea. *ERJ Open Research*, 166(4):782–785, 2022.

[156] Carlos Torre, Macario Camacho, Stanley Yung Chuan Liu, Leh Kiong Huon, and Robson

Capasso. Epiglottis collapse in adult obstructive sleep apnea: A systematic review. *The Laryngoscope*, 126(2):515–523, 2016.

[157] Katelyn J. Kotlarek, Abigail E. Haenssler, Kori E. Hildebrand, and Jamie L. Perry. Morphological variation of the velum in children and adults using magnetic resonance imaging. *Imaging Science in Dentistry*, 49(2):153, 2019.

[158] F. Gao, Y. R. Li, W. Xu, Y. S. An, H. J. Wang, J. F. Xian, and D. M. Han. Upper airway morphological changes in obstructive sleep apnoea: effect of age on pharyngeal anatomy. *The Journal of Laryngology & Otology*, 134(4):354–361, 2020.

[159] Rohan Diwakar, Anuraj Singh Kochhar, Harshita Gupta, Harneet Kaur, Maninder Singh Sidhu, Helen Skountrianos, Gurkeerat Singh, and Michele Tepedino. Effect of Craniofacial Morphology on Pharyngeal Airway Volume Measured Using Cone-Beam Computed Tomography (CBCT)—A Retrospective Pilot Study. *International Journal of Environmental Research and Public Health*, 18(9):5040, 2021.

[160] Melinda A. Ma, Rajesh Kumar, Paul M. Macey, Frisca L. Yan-Go, and Ronald M. Harper. Epiglottis cross-sectional area and oropharyngeal airway length in male and female obstructive sleep apnea patients. *Nature and Science of Sleep*, 8:297, 2016.

[161] Ning Zhou, Jean Pierre T.F. Ho, Cornelis Klop, Ruud Schreurs, Ludo F.M. Beenen, Ghizlane Aarab, and Jan de Lange. Intra-individual variation of upper airway measurements based on computed tomography. *PLOS ONE*, 16(11):e0259739, 2021.

[162] Eva Kirkegaard Kiær, Philip Tønnesen, Henrik Bredahl Sørensen, Niclas Rubek, Anne Hammering, Christine Møller, Anne Marie Hildebrandt, Poul Jørgen Jennum, and Christian Von Buchwald. Propofol sedation in Drug Induced Sedation Endoscopy without an anaesthesiologist - a study of safety and feasibility. *Rhinology*, 57(2):125–131, 2019.

[163] Ming Chin Lan, Stanley Y.C. Liu, Ming Ying Lan, Rahul Modi, and Robson Capasso. Lateral pharyngeal wall collapse associated with hypoxemia in obstructive sleep apnea. *The Laryngoscope*, 125(10):2408–2412, 2015.

[164] Karlien Van den Bossche, Eli Van de Perck, Andrew Wellman, Elahe Kazemeini, Marc Willemen, Johan Verbraecken, Olivier M. Vanderveken, Daniel Vena, and Sara Op de Beeck. Comparison of Drug-Induced Sleep Endoscopy and Natural Sleep Endoscopy in the Assessment of Upper Airway Pathophysiology During Sleep: Protocol and Study Design. *Frontiers in Neurology*, 12:2284, 2021.

[165] Eva Kirkegaard Kiær. Chief Surgeon in otorhinolaryngology at Copenhagen University Hospital.

[166] Donghwi Park, Jung Soo Kim, and Sung Jae Heo. The Effect of the Modified Jaw-Thrust Maneuver on the Depth of Sedation During Drug-Induced Sleep Endoscopy. *Journal of Clinical Sleep Medicine*, 15(10):1503–1508, 2019.

[167] Eric Kezirian. Professor and Chief Surgeon in otorhinolaryngology at University of Southern California.

[168] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[169] Y Bin, Y Yang, F Shen, X Xu, and HT Shen. Bidirectional long-short term memory for video description. *Proceedings of the 24th ACM international conference on Multimedia*, pages 436–440, 2016.

[170] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2020.

[171] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276, 2012.

[172] Anneclaire V. Vroegop, Olivier M. Vanderveken, An N. Boudewyns, Joost Scholman, Vera Saldien, Kristien Wouters, Marc J. Braem, Paul H. Van De Heyning, and Evert Hamans. Drug-induced sleep endoscopy in sleep-disordered breathing: Report on 1,249 cases. *The Laryngoscope*, 124(3):797–802, 2014.

[173] Danny Soares, Hadeer Sinawe, Adam J. Folbe, George Yoo, Safwan Badr, James A. Rowley, and Ho Sheng Lin. Lateral Oropharyngeal Wall and Supraglottic Airway Collapse Associated With Failure in Sleep Apnea Surgery. *The Laryngoscope*, 122(2):473–479, 2012.

[174] Anneclaire V. Vroegop, Olivier M. Vanderveken, and Johan A. Verbraecken. Drug-Induced Sleep Endoscopy: Evaluation of a Selection Tool for Treatment Modalities for Obstructive Sleep Apnea. *Respiration*, 99(5):451–457, 2020.

[175] Sebastian Susan K, Sharma Ankur, Chawla Omprakash, and Garg Payal. Management Concentric Collapse of Velopharynx in Obstructive Sleep Apnoea Using a Modified Barbed Palato-Pharyngoplasty Technique. *Journal of Sleep Disorders and Management*, 6(1), 2020.

[176] Stanley Yung Chuan Liu, Michael J. Hutz, Sasikarn Poomkonsarn, Corissa P. Chang, Michael Awad, and Robson Capasso. Palatopharyngoplasty Resolves Concentric Collapse in Patients Ineligible for Upper Airway Stimulation. *The Laryngoscope*, 130(12):E958–E962, 2020.

[177] Yanru Li, Jingying Ye, Demin Han, Xin Cao, Xiu Ding, Yuhuan Zhang, Wen Xu, Jeremy Orr, Rachel Jen, Scott Sands, Atul Malhotra, and Robert Owens. Physiology-based modeling

may predict surgical treatment outcome for obstructive sleep apnea. *Journal of Clinical Sleep Medicine*, 13(9):1029–1037, 2017.

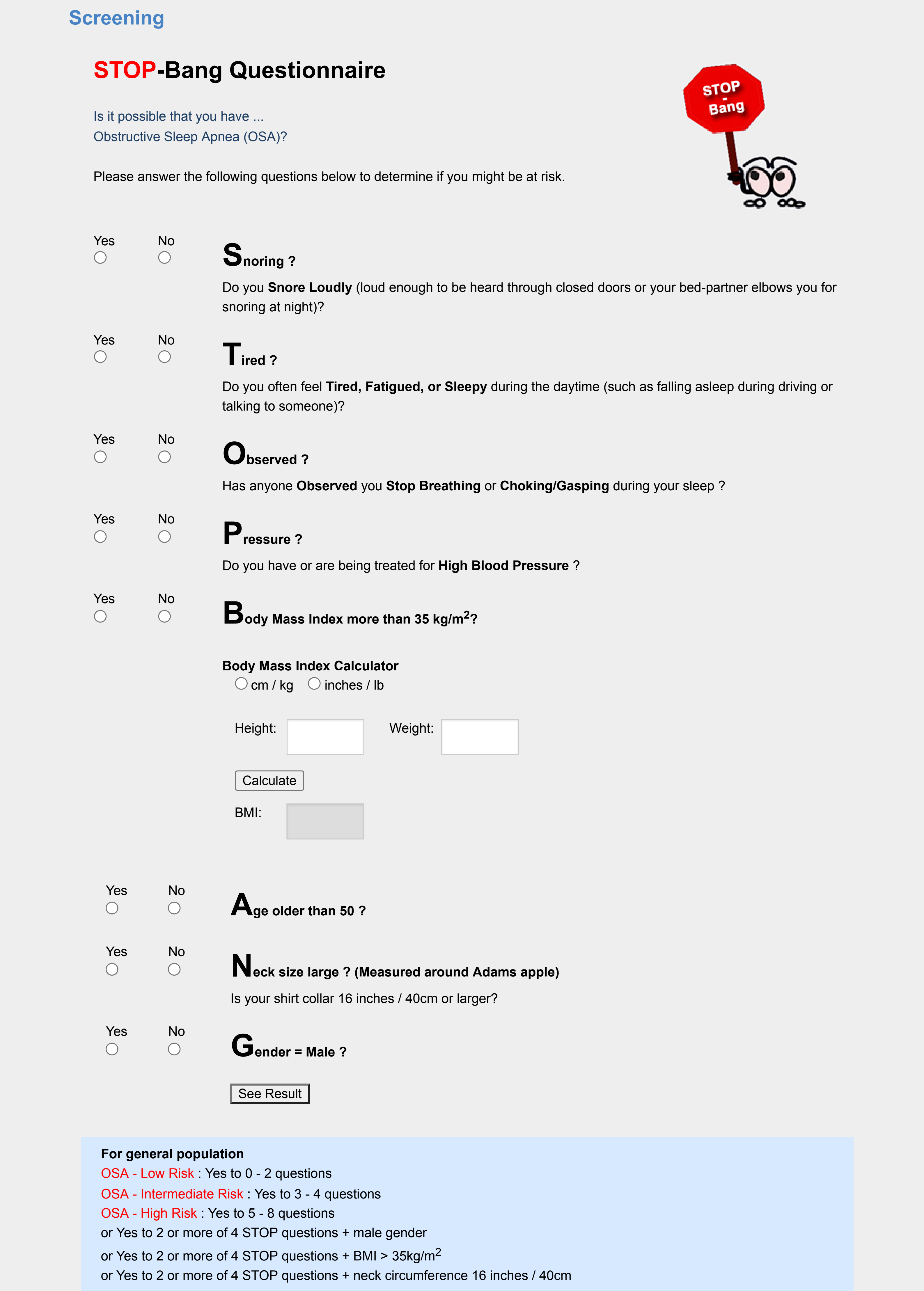
[178] Andrey V Zinchuk, Sangchoon Jeon, Brian B Koo, Xiting Yan, Dawn M Bravata, Li Qin, Bernardo J Selim, Kingman P Strohl, Nancy S Redeker, John Concato, and Henry K Yaggi. Polysomnographic phenotypes and their cardiovascular implications in obstructive sleep apnoea. *Thorax*, 73(5):472–480, 2018.

[179] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A Survey on Deep Semi-supervised Learning. *arXiv preprint arXiv:2103.00550*, 2021.

[180] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-Supervised Representation Learning: Introduction, Advances and Challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2021.

Appendix A

STOP-Bang Questionnaire



Screening

STOP-Bang Questionnaire

Is it possible that you have ...
Obstructive Sleep Apnea (OSA)?

Please answer the following questions below to determine if you might be at risk.

Yes ○ No ○

Snoring ?

Do you **Snore Loudly** (loud enough to be heard through closed doors or your bed-partner elbows you for snoring at night)?

Yes ○ No ○

T ired ?

Do you often feel **Tired, Fatigued, or Sleepy** during the daytime (such as falling asleep during driving or talking to someone)?

Yes ○ No ○

Observed ?

Has anyone **Observed** you **Stop Breathing** or **Choking/Gasping** during your sleep ?

Yes ○ No ○

Pressure ?

Do you have or are being treated for **High Blood Pressure** ?

Yes ○ No ○

Body Mass Index more than 35 kg/m^2?

Body Mass Index Calculator

○ cm / kg ○ inches / lb

Height: ☐ Weight: ☐

Calculate

BMI: ☐

Yes ○ No ○

Age older than 50 ?

Yes ○ No ○

Neck size large ? (Measured around Adams apple)

Is your shirt collar 16 inches / 40cm or larger?

Yes ○ No ○

Gender = Male ?

See Result

For general population

OSA - Low Risk : Yes to 0 - 2 questions

OSA - Intermediate Risk : Yes to 3 - 4 questions

OSA - High Risk : Yes to 5 - 8 questions

or Yes to 2 or more of 4 STOP questions + male gender

or Yes to 2 or more of 4 STOP questions + BMI > 35kg/m^2

or Yes to 2 or more of 4 STOP questions + neck circumference 16 inches / 40cm

Appendix B

Paper I

Title: Prediction of Patient Demographics using 3D Craniofacial Scans and Multi-view CNNs

Authors: Umaer Hanif, Rasmus R. Paulsen, Eileen B. Leary, Emmanuel Mignot, Poul Jennum, and Helge B. D. Sorensen

Conference: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)

Status: Published

DOI: 10.1109/EMBC44109.2020.9176333

Full citation: U. Hanif, R. R. Paulsen, E. B. Leary, E. Mignot, P. Jennum and H. B. D. Sorensen, "Prediction of Patient Demographics using 3D Craniofacial Scans and Multi-view CNNs", *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1950-1953, 2020. DOI: 10.1109/EMBC44109.2020.9176333.

Prediction of Patient Demographics using 3D Craniofacial Scans and Multi-view CNNs*

Umaer Hanif[1,3,4], *Member, IEEE*, Rasmus R. Paulsen[2], Eileen B. Leary[3], Emmanuel Mignot[3,5], Poul Jennum[4,5], and Helge B. D. Sorensen[1,5], *Senior Member, IEEE*

Abstract— **3D data is becoming increasingly popular and accessible for computer vision tasks. A popular format for 3D data is the mesh format, which can depict a 3D surface accurately and cost-effectively by connecting points in the (x,y,z) plane, known as vertices, into triangles that can be combined to approximate geometrical surfaces. However, mesh objects are not suitable for standard deep learning techniques due to their non-euclidean structure. We present an algorithm which predicts the sex, age, and body mass index of a subject based on a 3D scan of their face and neck. This algorithm relies on an automatic pre-processing technique, which renders and captures the 3D scan from eight different angles around the x-axis in the form of 2D images and depth maps. Subsequently, the generated data is used to train three convolutional neural networks, each with a ResNet18 architecture, to learn a mapping between the set of 16 images per subject (eight 2D images and eight depth maps from different angles) and their demographics. For age and body mass index, we achieved a mean absolute error of 7.77 years and 4.04 kg/m^2 on the respective test sets, while Pearson correlation coefficients of 0.76 and 0.80 were obtained, respectively. The prediction of sex yielded an accuracy of 93%. The developed framework serves as a proof of concept for prediction of more clinically relevant variables based on 3D craniofacial scans stored in mesh objects.**

I. INTRODUCTION

As the field of deep learning continues to revolutionize computer vision, increased emphasis is being placed on the 3D domain and its potential to push boundaries within computer vision even further. However, as 3D data becomes more accessible and easier to utilize with the emergence of datasets such as ModelNet [1], ShapeNet [2], and SHREC'16 [3], limitations and problems related to this domain also become more evident [4].

A popular format for storing 3D data is a mesh object, which can depict 3D geometries accurately and in a computationally cost-effective manner considering the level of detail that can be achieved [5]. Mesh objects are made up of vertices, which are points in space, described by their (x,y,z) coordinates. Three interconnected vertices form a triangle, also known as a face, and several faces are combined to construct a surface in 3D. Faces are the basic building blocks of a mesh object, where finer details can be approximated by using very small faces and cruder details can be depicted by using a few large faces.

Even though mesh objects provide a cheap and convenient way of storing 3D data, they are unsuitable for convolutional neural networks (CNNs), which are traditionally applied in a wide variety of computer vision tasks. This is due to the non-euclidean structure of mesh objects [6]. In this work, we attempt to convert mesh data to make it suitable as an input for a CNN while still preserving as much of the 3D information as possible. Specifically, we predict the sex, age, and body mass index (BMI) of a subject by utilizing and transforming a 3D scan of their face and neck into 2D images and depth maps from several angles around the subject. These images can then be applied as inputs to multi-view CNNs to learn the mapping between a subject's craniofacial scan and their demographics. The motivation is to develop a framework which can be extended to predict several clinically significant variables based on 3D craniofacial scans stored in mesh objects.

II. MATERIALS

A. Experimental Protocol

The 3D craniofacial images were captured as part of an extensive study on sleep disorders known as the Stanford Technology Analytics and Genomics in Sleep (STAGES) program. Before the initiation of a diagnostic sleep test on each patient, a 3D surface scan was performed of the patient's face and neck. For the scanning, a Structure Sensor from Occipital Inc. was attached to an Apple iPad Pro 10.5-inch w/ 64 GB Storage, while the software used was the STAGES 3D app developed by uGo3D. A scan was performed by going around the subject and capturing their face and neck from all angles, while the reconstruction of the complete surface scan was done using the app software. The Institution's Ethical Review Board approved all procedures involving human subjects.

B. Data Description

The dataset used for this study contained craniofacial scans from 1605 patients, where 855 were women and 750 were men. The mean age ± standard deviation was 45.8 ± 15 years and the BMI was 31.3 ± 8.9 kg/m^2. Each scan was stored in a mesh object, consisting of vertices, textures, and vertex normals. Fig. 1 depicts an example of a 3D craniofacial scan, where it is possible to see all the faces that are combined to make up the surface scan.

*Research has been supported by The Klarman Family Foundation, Stanford University, Technical University of Denmark, and Rigshospitalet

[1] Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[2] Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

[3] Stanford Center for Sleep Sciences and Medicine, Stanford University, 94304 Palo Alto, USA

[4] Danish Center for Sleep Medicine, Rigshospitalet, 2600 Glostrup

[5] Shared last authors

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

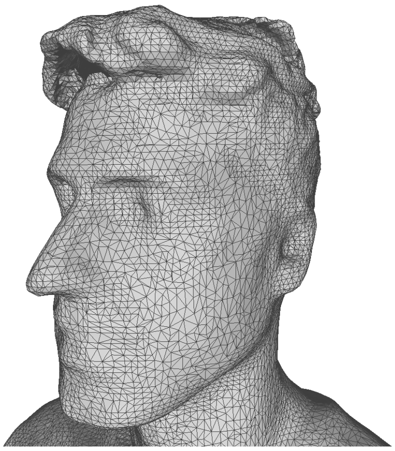


Fig. 1: A 3D craniofacial scan depicted with the faces (triangles) which are combined to form the surface.

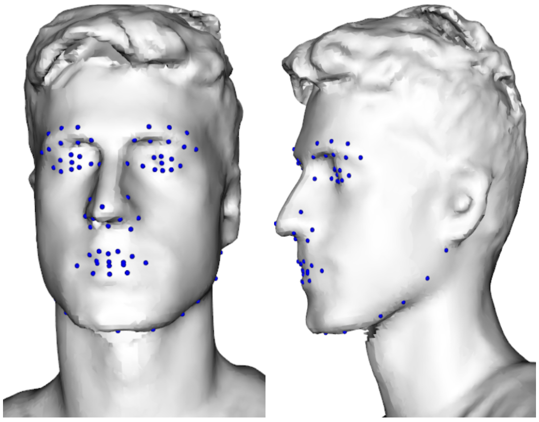


Fig. 2: Examples of detected landmarks on a subject from two different angles.

III. Methods

A. Preprocessing

In order to capture 2D images and depth maps from different angles around each subject, the Deep-MVLM algorithm was utilized [7]. Using a pre-trained neural network, 73 pre-specified landmarks were detected automatically on each of the face scans. Fig. 2 illustrates two examples of these landmarks detected on a subject rendered from two different angles. Once detected, these landmarks were used to align all scans to remove the influence of rotation and translation. Subsequently, all scans were rotated 180 degrees around the x-axis. Images were captured each time the scan had been turned 22.5 degrees, yielding eight different 2D images per subject. Additionally, each of these images were converted to depth maps as well, resulting in a total of 16 input images for each patient, which were stacked as a single matrix with 16 channels. Each image was then normalized to the range [0, 1] by dividing all pixel values by 255. Fig. 3 visualizes a few examples of different angles at which 2D images and depth maps have been captured for a subject. The dataset was split into a training set (65%), a validation set (25%), and a test set (10%).

B. Neural Network

To learn the mapping between the input images and demographic values, a ResNet18 architecture was utilized for training [8]. The input layer was modified to take a 16-channel input instead of RGB images, which have 3 channels. The final layer originally consisted of a fully connected layer taking a feature map of 512 features and predicting 1000 classes. This was modified to have a fully connected layer which reduces the feature map from 512 to 256 features, followed by a ReLU activation and dropout [9] with a probability of 0.4, and finally adding a single neuron for prediction. For prediction of age and BMI, this neuron predicted the values directly due to the regression setting, but two separate networks were trained as the target values were different. For prediction of sex, a sigmoid activation

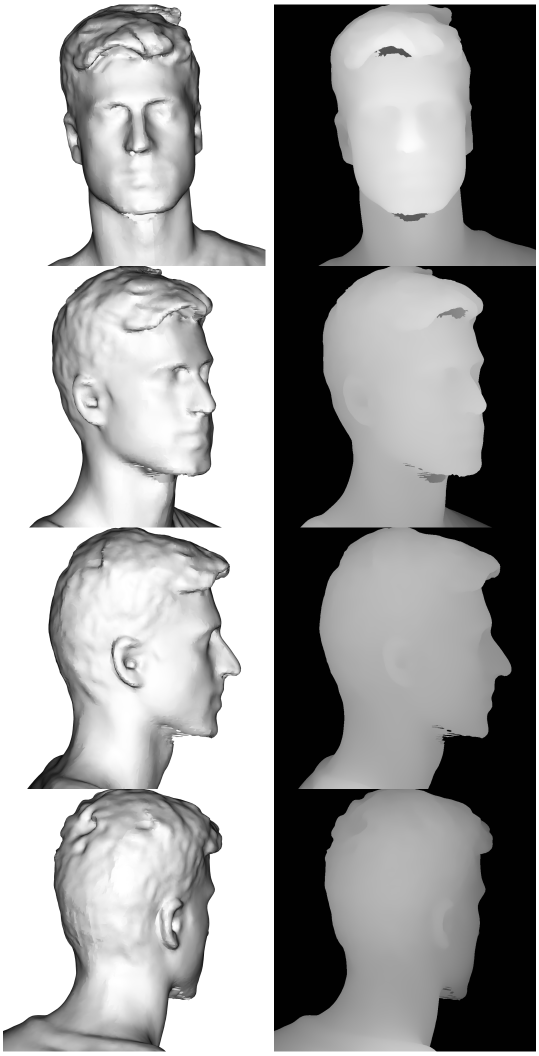


Fig. 3: Four examples of different angles at which 2D images and depth maps have been captured.

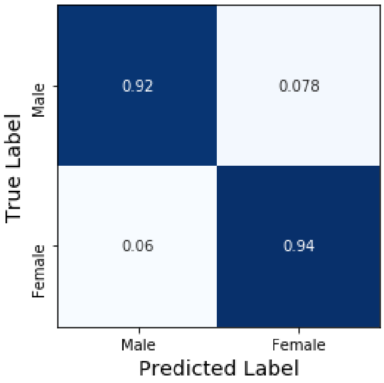


Fig. 4: A confusion matrix displaying the fraction of correctly and incorrectly classified patients with respect to their sex.

was added at the end as this was a two-class classification problem. The learning rate was set to $1 \cdot 10^5$ with a weight decay of $5 \cdot 10^{-4}$. For age and BMI prediction, the mean absolute error (MAE) was used as loss function, while binary cross entropy loss was used for prediction of sex. Finally, Adam [10] was used as an optimizer during training of the network and the validation loss was used as an early stopping criterion.

IV. Results

The performance metrics reported on the prediction of sex, age, and BMI were all evaluated on the test set consisting of 161 patients.

A. Sex

The network was trained for 13 epochs before early stopping occurred. An accuracy of 93% was achieved in classifying the sex of each patient in the test set. Fig. 4 shows the normalized confusion matrix for the predictions.

B. Age

The network was trained for 190 epochs before early stopping occurred. A MAE of 7.77 and a correlation coefficient of 0.76 was obtained between the true and predicted ages in the test set. Fig. 5 shows a Bland-Altman plot [11] of the true and predicted ages, i.e. the difference between the labels and predicted ages as a function of the average of the two.

C. BMI

The network was trained for 196 epochs and achieved a MAE of 4.04 and a correlation coefficient of 0.8 with respect to true and predicted BMIs in the test set. Fig. 6 shows a Bland-Altman plot, illustrating the differences in the true and predicted BMI values of the test set. Based on the predicted BMI values, a classification was also attempted into the standard BMI categories of normal, overweight, and obese [12]. Fig. 7 depicts the confusion matrix displaying the results from the classification. Several of the normal patients are classified as overweight and many of the overweight patients are classified as obese.

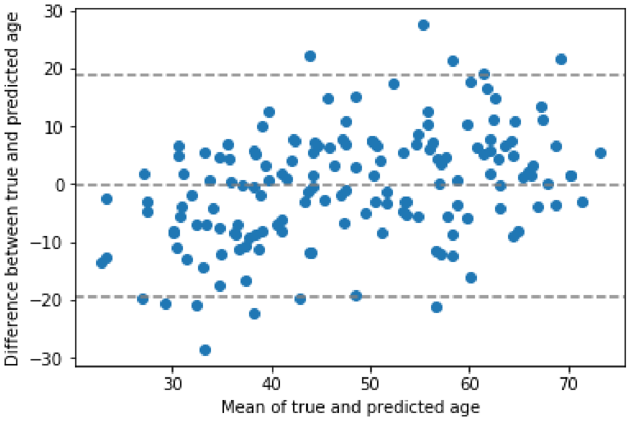


Fig. 5: Bland-Altman plot for the labels and predicted age in the test set. The dashed horizontal lines above and below 0 indicate the limits of the 5% confidence interval.

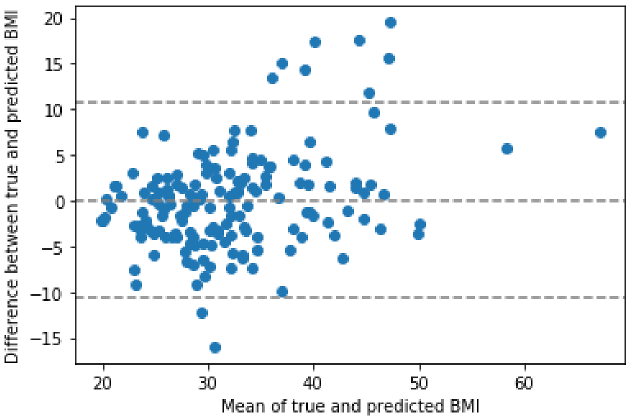


Fig. 6: Bland-Altman plot for the labels and predicted BMI in the test set. The dashed horizontal lines above and below 0 indicate the limits of the 5% confidence interval.

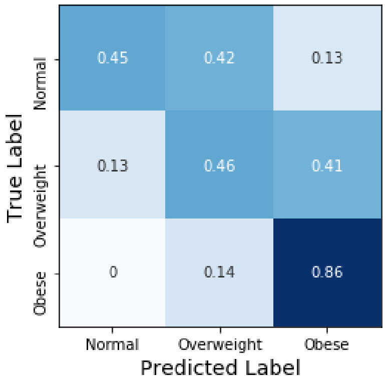


Fig. 7: A confusion matrix displaying the fraction of correctly and incorrectly classified patients with respect to their BMI categories.

V. Discussion

The prediction of a subject's sex yielded an accuracy of 93%. To put this in context, if the network simply predicted all patients in the test set to be women, an accuracy of 52% would be achieved. Similarly, if the network predicted the age and BMI for all patients to be the mean values in their respective distributions, a MAE of 12.9 years and 6.9 kg/m^2 would be obtained, as compared to the actual MAEs of 7.77 years and 4.04 kg/m^2, respectively. The MAE of the age predictor being almost twice that of the BMI predictor also makes sense, as in many cases it would be easier to derive someone's BMI based off of their face as compared to age.

Looking at the Bland-Altman plot for age prediction in Fig. 5, it is noted that there is a slight trend showing that lower ages are over predicted and higher ages are under predicted. The Bland-Altman plot for BMI prediction in Fig. 6 shows small errors around the mean BMI of around 31 kg/m^2, while the larger BMI values are under predicted. The main uncertainty associated with prediction of BMI is that the height of the person is unknown. Another drawback is that the BMI values in the dataset are heavily centered around the mean value of about 31 kg/m^2. Thus, the more extreme cases of either very low or high BMI values are underrepresented in the dataset.

Using the predicted BMI values to classify the patients into normal, overweight and obese shows that in many cases, the BMI values have been overestimated. However, even a small overestimation could lead to a misclassification, since the boundary between each class is so subtle, and most of the patients have BMI values centered around the cut-off between overweight and obese. However, the highest accuracy is obtained for the obese class and one could argue that these patients are the most important to capture, since they are the ones who are medically most at risk.

Similar work to a part of this study was presented in [13], where they used transfer learning with VGG-Face [14] on social media images of people to predict their BMI. They achieved an overall correlation coefficient of 0.65 compared to our correlation of 0.80. In [15], the authors extracted facial measurements as features instead and used simple regression methods to obtain an overall MAE of 3.14 compared to our MAE of 4.04. However, their work relied on hand-engineered features instead of the more data-driven feature extraction presented in this work. [16] gives an extensive overview of different age predictors, showing MAEs ranging from 8.84 to 0.31. Compared to our MAE of 7.77, this shows that there is clearly room for improvement with respect to age prediction. However, it must be noted that widely different techniques and datasets are being compared, so one should be careful to place too much emphasis on it. Finally, [17] predicted the sex of subjects based on different deep learning architectures and their best model achieved an accuracy of 93.57%, which is very comparable to the 93% obtained in this work. Although our focus wasn't to achieve state-of-the-art performance on predicting any demographic, the comparison to other studies still serves as a validation of the proposed framework.

VI. Conclusion

This work presents a framework for predicting patient demographics based on 3D craniofacial scans and deep learning. We have successfully showed that it is possible to derive variables such as sex, age, and BMI of a patient from a surface scan of their face and neck. This serves as a proof of concept and the applied techniques can be extended to predict more biologically significant variables instead.

References

[1] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, 3D ShapeNets: A Deep Representation for Volumetric Shapes, The IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015, pp. 1912-1920.

[2] L. Yi, V. G. Kim, D, Ceylan, I. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. Guibas, A Scalable Active Framework for Region Annotation in 3D Shape Collections, ACM Transactions on Graphics, vol. 35, no. 210, pp. 4503-4514, Nov. 2016.

[3] L. Cosmo, E. Rodola, M. M. Bronstein, A. Torsello, D. Cremers, and Y. Sahillioglu, SHREC'16: Partial Matching of Deformable Shapes, 9th Eurographics Workshop on 3D Object Retrieval, pp. 61-67, 2016.

[4] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, Geometric Deep Learning: Going Beyond Euclidean Data, IEEE Signal Processing Magazine, vol. 34, pp. 18-42, 2017.

[5] M. Botsch, L. Kobbelt, M. Pauly, P. Alliez, and B. Lévy, Polygon Mesh Processing, AK Peters, 2010.

[6] E. Ahmed, A. Saint, A. Shabayek, K. Cherenkova, R. Das, G. Gusev, D. Aouada, A Survey on Deep Learning Advances on Different 3D Data Representations, arXiv:1808.01462 [cs.CV].

[7] R. R. Paulsen, K. A. Juhl, T. M. Haspang, T. Hansen, M. Ganz, and G. Einarsson, Multiview Consensus CNN for 3D Facial Landmark Placement, Lecture Notes in Computer Science, vol. 11361, pp. 706-719, 2019.

[8] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, Deep Residual Learning for Image Recognition, The IEEE Conference on Computer Vision and Pattern Recognition, Nevada, 2016, pp. 770-778.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.

[10] D. P. Kingma and J. L. Ba, Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, San Diego, 2015.

[11] D. Giavarina, Understanding Bland Altman Analysis, Biochemia medica, vol. 25, no. 2, pp. 141-151, 2015.

[12] F. Q. Nuttall, Body Mass Index: Obesity, BMI, and Health: A Critical Review, vol. 50, no. 3, pp. 117-128, 2015.

[13] E. Kocabey, M. Camurcu, F. Ofli, Y. Aytar, J. Marin, A. Torralba, and I. Weber, Face-to-BMI: Using Computer Vision to Infer Body Mass Index on Social Media, 11th International Conference on Web and Social Media, Montreal, 2017, pp. 572-575.

[14] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, University of Oxford, 2015.

[15] L. Wen, and G. Guo, A Computational Approach to Body Mass Index Prediction from Face Images, Image and Vision Computing, vol. 31, pp. 392-400, 2013.

[16] R. Angulu, J. R. Tapamo, and A. O. Adewumi, Age Estimation via Face Images: A Survey, EURASIP Journal on Image and Video Processing, 2018.

[17] K. Ito, H. Kawai, T. Okano, and T. Aoki, Age and Gender Prediction from Face Images Using Convolutional Neural Network, APSIPA Annual Summit and Conference, 2018.

Appendix C

Paper II

Title: Estimation of Apnea-Hypopnea Index Using Deep Learning On 3-D Craniofacial Scans

Authors: Umaer Hanif, Eileen B. Leary, Logan D. Schneider, Rasmus R. Paulsen, Anne Marie Morse, Adam Blackman, Paula K. Schweitzer, Clete A. Kushida, Stanley Y. Liu, Poul Jennum, Helge B. D. Sorensen, and Emmanuel J. M. Mignot

Journal: IEEE Journal of Biomedical and Health Informatics

Status: Published

DOI: 10.1109/JBHI.2021.3078127

Full citation: U. Hanif, E. B. Leary, L. D. Schneider, R. R. Paulsen, A. M. Morse, A. Blackman, P. K. Schweitzer, C. A. Kushida, S. Y. Liu, P. Jennum, H. B. D. Sorensen, and E. J. M. Mignot, "Estimation of Apnea-Hypopnea Index Using Deep Learning On 3-D Craniofacial Scans", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 11, pp. 4185-4194, 2021. DOI: 10.1109/JBHI.2021.3078127.

Copyright information: ©2021 IEEE

Estimation of Apnea-Hypopnea Index Using Deep Learning On 3-D Craniofacial Scans

Umaer Hanif, *Member, IEEE*, Eileen B. Leary, Logan D. Schneider, Rasmus R. Paulsen, Anne Marie Morse, Adam Blackman, Paula K. Schweitzer, Clete A. Kushida, Stanley Y. Liu, Poul Jennum, Helge B. D. Sorensen, *Senior Member, IEEE*, and Emmanuel J. M. Mignot

***Abstract*—Obstructive sleep apnea (OSA) is characterized by decreased breathing events that occur through the night, with severity reported as the apnea-hypopnea index (AHI), which is associated with certain craniofacial features. In this study, we used data from 1366 patients collected as part of Stanford Technology Analytics and Genomics in Sleep (STAGES) across 11 US and Canadian sleep clinics and analyzed 3D craniofacial scans with the goal of predicting AHI, as measured using gold standard nocturnal polysomnography (PSG). First, the algorithm detects prespecified landmarks on mesh objects and aligns scans in 3D space. Subsequently, 2D images and depth maps are generated by rendering and rotating scans by 45-degree increments. Resulting images were stacked as channels and used as input to multi-view convolutional neural networks, which were trained and validated in a supervised manner to predict AHI values derived from PSGs. The proposed model achieved a mean absolute error of 11.38 events/hour, a Pearson correlation coefficient of 0.4, and accuracy for predicting OSA of 67% using 10-fold cross-validation. The model improved further by adding patient demographics and variables from questionnaires. We also show that the model performed at the level of three sleep medicine specialists, who used clinical experience to predict AHI based on 3D scan displays. Finally, we created topographic displays of the most important facial features used by the model to predict AHI, showing importance of the neck and chin area. The proposed algorithm has potential to serve as an inexpensive and efficient screening tool for individuals with suspected OSA.**

***Index Terms*—Apnea, craniofacial scans, deep learning, mesh, multi-view.**

I. INTRODUCTION

OBSTRUCTIVE sleep apnea (OSA) is a sleep disorder characterized by recurrent collapses of the upper airway (UA) during sleep, resulting in decreased airflow (hypopnea) or total cessation of breathing (apnea), lasting until the UA reopens [1] and causing daytime sleepiness and increased cardiovascular risk [2], [3]. OSA severity is measured by the apnea-hypopnea index (AHI), representing the number of apneas and hypopneas per hour of sleep. The presence of sleep disordered breathing (SDB) is extremely common; in a recent study, the prevalence of moderate-to-severe SDB ($\geq$15 events/hour) was 23.4% in women and 49.7% in men older than 40 [4]. Why and how frequently the UA collapses is due to multiple factors, such as narrow UA anatomy, poor recruitment of dilator muscles during inspiration, central control of breathing (loop gain), and inability/ability to arouse (arousal threshold) [5]. A strong contributing factor to passive anatomy is obesity, which causes fat deposits around the UA that narrow the airway during sleep [6]. Other reasons include loss of muscle tone in the UA or the tongue falling backwards into the throat [7]. Finally, research shows that several craniofacial features, mainly related to the midface, jaw, and neck are indicative of the presence of OSA [8].

Nocturnal polysomnography (PSG) is the accepted gold-standard for diagnosing OSA [9]. PSGs are performed in a

Manuscript received November 30, 2020; revised March 30, 2021 and April 26, 2021; accepted April 30, 2021. Date of publication May 7, 2021; date of current version November 5, 2021. This work was supported in part by a grant from the Klarman Family Foundation and in part by the Technical University of Denmark, and the Danish Center for Sleep Medicine. The work of Umaer Hanif at Stanford University was supported in part by Danmark-Amerika Fondet, Vera og Carl Johan Michaelsens Legat, Reinholdt W. Jorck og Hustrus Fond, Torben og Alice Frimodts Fond, Christian og Ottilia Brorsons Rejselegat, Marie og M.B. Richters Fond, Oberstløjtnant Max Nørgaard og hustru Magda Nørgaards Legat, William Demant Fonden, Augustinus Fonden, Rudolph Als Fondet, Knud Højgaards Fond, Otto Mønsteds Fond, Julie von Müllens Fond, and Direktør Einar Hansen og hustru fru Vera Hansens Fond. *(Corresponding author: Emmanuel J. M. Mignot.)*

Umaer Hanif, Eileen B. Leary, Logan D. Schneider, Clete A. Kushida, and Emmanuel J. M. Mignot are with the Stanford University Center for Sleep Sciences and Medicine, Stanford University, CA 94304 USA (e-mail: umaerhanif@hotmail.com; eileen@eileenleary.com; logands@gmail.com; clete@stanford.edu; mignot@stanford.edu).

Helge B. D. Sorensen is with the Biomedical Signal Processing & AI Research Group, Department of Health Technology, Technical University of Denmark, 2800 Kongens Lyngby, Denmark (e-mail: hbds@dtu.dk).

Poul Jennum is with the Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, 2600 Glostrup, Denmark (e-mail: poul.joergen.jennum@regionh.dk).

Rasmus R. Paulsen is with the Department for Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, 2800 Kongens Lyngby, Denmark (e-mail: rapa@dtu.dk).

Anne Marie Morse is with the Geisinger Commonwealth School of Medicine, Geisinger Medical Center, PA 18510 USA (e-mail: amorse@geisinger.edu).

Adam Blackman is with MedSleep, Toronto, ON M4P 1P2, Canada (e-mail: adamb@medsleep.com).

Paula K. Schweitzer is with Sleep Medicine & Research Center, St. Luke's Hospital, MO 63017 USA (e-mail: paula.schweitzer@stlukes-stl.com).

Stanley Y. Liu is with Otolaryngology/Head & Neck Surgery, Stanford University School of Medicine, CA 94133 USA (e-mail: ycliu@stanford.edu).

Digital Object Identifier 10.1109/JBHI.2021.3078127

sleep clinic or laboratory where the patient can sleep for a full night. While the individual is sleeping, sensors measure airflow, respiratory effort, snoring sounds, blood oxygen levels, eye movements, leg movements, and electrical activity of the heart (ECG) and brain (EEG). These signals are manually annotated by sleep technicians, who follow standard definitions for apneas, hypopneas, periodic leg movements, and sleep stages [10]. A PSG, though gold-standard for OSA diagnosis, is expensive and impractical; it requires a great number of modalities and a trained technician to perform, then manually view and evaluate data. Furthermore, scoring of PSG data is prone to errors, day-to-day variance, and high interscorer variability [11], [12]. Finally, a PSG is supposed to accurately portray sleep behavior, but this is often difficult as only one night is recorded, the individual is wearing equipment, and is connected to several wires, thereby reducing their comfort. Due to these limitations, research efforts have gone into finding alternative, data-driven approaches for the diagnosis of sleep disorders, including use of deep neural networks to score events [13]–[17].

Obesity and craniofacial features strongly contribute to OSA risk; thus, technology and clinical examination tools have been developed to assess these features. In clinical exams, it is frequent for the clinician to examine the size of the jaw, top of the mouth, position, and size of the tongue [18]. Imaging techniques have been proposed, including cephalometry [19]–[21], computed tomography [21], [22], and magnetic resonance imaging [23]–[25]. Furthermore, dynamic collapsibility can be identified with drug-induced sleep endoscopy (DISE) [26]. These methods are, however, rarely used in routine practice, except in case of surgery for OSA.

Since imaging modalities are cumbersome and expensive, recent research has investigated the predictive value of facial imaging [27]. Lee et al. [28] analyzed frontal and profile images of 180 patients, manually deriving measurements on the face and neck to classify subjects with or without OSA using logistic regression. Others [29]–[31] used Support Vector Regression (SVR) on similar landmarks to predict the AHI, a procedure subsequently improved by Balaei et al. [32] who used automatic instead of manual placement of landmarks. Islam et al. [33] used 3D scans from 69 subjects, which they converted to 2D depth maps of the frontal face, applied transfer learning on a VGG-16 deep convolutional neural network (CNN), and modified the network to classify subjects into OSA and non- OSA. Although these studies had some success, all used small sample sizes, and, with the exception of Islam et al. [33], all first derive possibly discriminative facial features from annotated landmarks as detected on 2D frontal and profile images, a process followed by statistical feature selection. As attempted by Islam et al. [33] using a small sample size, we believe that feature selection should be unbiased by avoiding manual extraction of features and applying CNNs instead.

In this study, taking advantage of the rapid development of depth imaging in most hand held devices, we aimed to explore how state of the art deep learning techniques as applied to a dataset of more than 1300 3D images could be used to develop a fully automatic system for the prediction of OSA severity and AHI values. This study aims to (1) investigate how accurately such a system could predict the AHI, (2) determine how accurately we can classify OSA/non-OSA based on an AHI cutoff of 15 events/hour, (3) estimate the predictive value of adding clinically relevant information like demographics and questionnaire variables to the model, and (4) use a data-driven approach to identify which regions of the face and neck the CNN found most useful in predicting OSA severity. From a clinical standpoint, the study further aims to (5) compare our model performance to that of three sleep medicine specialists asked to predict OSA severity after inspecting the scans.

The novelty of this study is four-fold: the dataset consists of 3D surface scans instead of frontal and profile images; we compare our model predictions to those of three sleep medicine specialists imitating the task of the model; we use a purely data-driven approach to identify and reveal craniofacial features related to OSA; and our dataset is four times greater than any other study predicting OSA from craniofacial images. Facial imaging for OSA diagnosis can be performed in one minute with the current setup and presents a clear advantage; it does not require an overnight stay at a sleep clinic with several sensors connected to the body and a subsequent manual analysis by sleep technicians, thereby saving both time and resources, while being more comfortable for the patient.

II. Materials

A. Data Collection

Data was collected at 11 different sleep clinic sites as part of the Stanford Technology Analytics and Genomics in Sleep (STAGES) study, which was initiated in 2018 and prematurely terminated in 2020. STAGES was designed to better understand and characterize sleep disorder phenotypes on a large scale. For each subject participating in the study, a detailed sleep questionnaire, actigraphy, psychometric testing, a PSG, a 3D craniofacial scan, and blood samples were collected. The 3D craniofacial scans were collected according to a procedure described in Hanif et al. [34]. The scans were performed using a Structure Sensor from Occipital Inc. [35], [36] attached to an iPad Pro from Apple, which was used around the subject to get a complete surface scan of face and neck. Each scan took approximately one minute to complete and was captured either at night before the PSG or in the morning after the PSG. uGo3D Inc. developed an app for STAGES which was responsible for transferring the scans to the server after they were captured. The sleep questionnaire was a modified STOP-Bang questionnaire [37] without neck circumference of the subject. The modified STOP-Bang questionnaire is used as a screening tool for OSA without the need for a specialist by asking the subject about snoring, tiredness, observed apnea, high blood pressure, sex (is the person male), age (is the person older than 65), and BMI (is the BMI greater than 35 kg/m^2). Each variable gets a value of 1 if the person answers yes, so the total score from the questionnaire ranges from 0 to 7, where a score of 3 or above indicates presence of OSA. Each institution's Ethical Review Board approved all procedures involving human subjects. All participants provided written informed consent to participate in the study. The 3D scans from STAGES are not publicly available, since they count

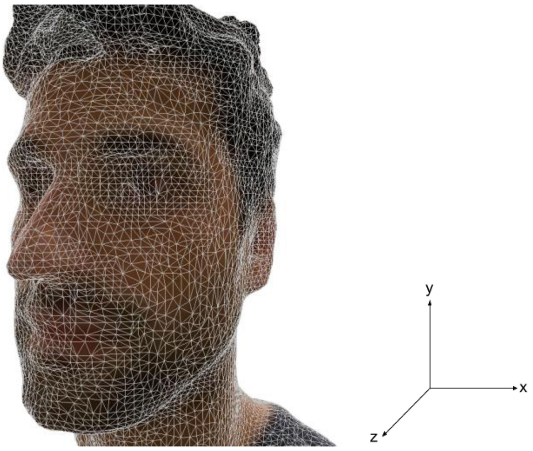


Fig. 1. An example of a 3D craniofacial scan (first author). Each triangle is formed by connecting three vertices, which are points in 3D space. Crude details are approximated using large triangles, whereas finer details require smaller triangles. The axes to the right specify the orientation of the coordinates.

as personal identifiable data, which the Institutional Review Board would not allow to be made public. All other data from STAGES, apart from blood samples, will be made available as part of the National Sleep Research Resource (NSRR).

B. Data Description

During data collection for STAGES, a total of 1756 scans were captured from enrolled participants. However, subjects were discarded for this study if they had missing PSGs, demographics, or questionnaires, so 1366 subjects were used; 724 females and 642 males. Mean age ± standard deviation was 45.9 ± 14.8 years, body mass index (BMI) was 30.9 ± 8.7 kg/m^2, and AHI was 15.5 ± 19.3 events/hour (median: 9.3, IQR: 17.5). Fig. A.1. (appendix) shows the distribution of AHI values within the dataset. AHI was derived from each PSG by summing the number of annotated obstructive apneas and hypopneas and dividing this number by the total sleep time in hours. Central apneas were excluded from the analysis because they have no known relation to craniofacial anatomy. Each scan was stored as a mesh object, defined by its vertices $V = \{v_1, v_2, \ldots, v_n\}$, $V \in \mathbb{R}^3$, its triangles F, given as three interconnected vertices, and its texture coordinates $T = \{t_1, t_2, \ldots, t_m\}$, $T \in \mathbb{R}^3$. A vertex is a point in 3D space described by its (*x*,*y*,*z*) coordinates. Three interconnected vertices form a triangle and several of these triangles are combined to approximate a surface in three dimensions. Furthermore, each scan also contains associated textures. Fig. 1 shows a typical example of a craniofacial scan where the triangles that make up the scan are also depicted. Fig. A.2. (appendix) provides a simple example of the basic components of a mesh, i.e., vertices, edges, and triangles.

III. Methods

A. Preprocessing

Since 3D mesh scans are non-Euclidean, they cannot be used directly as inputs to CNNs. Thus, to make craniofacial scans suitable for CNNs, the multi-view consensus CNN for 3D facial landmark placement (Deep-MVLM) algorithm [38] was applied to transform each scan into a set of 2D images and depth maps captured from angles around the scan. Deep-MVLM was chosen as it outperforms state of the art algorithms [39], [40] and does not rely on pre-alignment of scans, such that they have the same orientation in 3D space. The choice of metric in Deep-MVLM also makes it more suitable for 3D surfaces than similar methods. Deep-MVLM was only used for alignment of scans and not for the subsequent prediction of AHI.

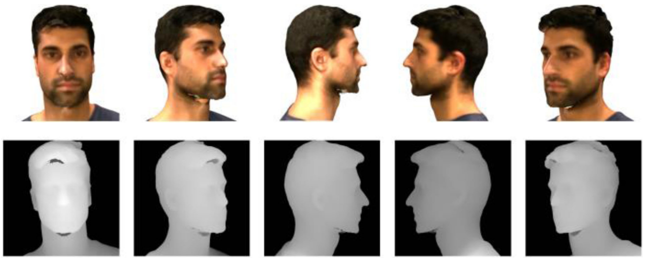


Fig. 2. Example of the five pairs of 2D images (top row) and depth maps (bottom row) captured at different angles and used as input for each subject in the multi-view convolutional neural network for predicting AHI.

Deep-MVLM works by first applying a pre-trained neural network to automatically detect and place 73 pre-specified landmarks on each mesh object. The alignment of scans is obtained by using a least squares solution based on the detected landmarks and therefore the individual landmark errors are less important for the overall performance. Subsequently, these landmarks are utilized to align all scans in 3D space, thereby removing the influence of translation and rotation. Finally, each scan is displayed multiple times as flat 2-D images taken from different angles; this is done by rotating the scans 45 degrees around the *y*-axis consecutively and capturing a 2D image and a depth map at each angle. This results in eight pairs of 2D images and depth maps for each subject. Since some scans were not complete with respect to the back of the head and since this information does not contribute to AHI prediction, only five pairs of 2D images and depth maps were used, emphasizing the frontal and profile characteristics of face and neck from several angles as shown in Fig. 2. For each subject, these images were stacked into a matrix. The 2D images had three channels each (RGB) and the depth maps had one channel each. Since there were five 2D images and five depth maps, the final input matrix per subject contained $5 \times 3 + 5 \times 1 = 20$-channels in total and 224 × 224 pixels.

Each image was normalized to the range [0,1] by dividing each pixel value by 255. This was done to ensure faster convergence during training. Similarly, the patient demographics (age and bmi) were normalized using min-max normalization given by

$$\mathbf{x}_{\text{norm}} = \frac{\mathbf{x} - \min(\mathbf{x}_{\text{train}})}{\max(\mathbf{x}_{\text{train}}) - \min(\mathbf{x}_{\text{train}})}$$

where $\mathbf{x}$ and $\mathbf{x}_{\text{train}}$ are vectors containing one of the demographics for all patients in the entire dataset and training set, respectively. No normalization was necessary for sex or the seven different variables from the modified STOP-Bang questionnaires because these values were binary.

TABLE I
THE APPLIED CNN ARCHITECTURE FOR PREDICTING AHI VALUES BASED ON THE 20-DIMENSIONAL INPUT CRANIOFACIAL IMAGES. THE INPUT DIMENSIONS ARE GIVEN BY NUMBER OF CHANNELS×HEIGHT×WIDTH. THE CONVOLUTION LAYERS ARE SPECIFIED BY FILTER SIZE (E.G., 3×3), NUMBER OF CHANNELS (E.G., 64), AND A STRIDE (E.G., /2). THE SAME APPLIES FOR THE MAX POOLING (MP) LAYER. THE OUTPUT DIMENSIONS OF THE FEATURE MAPS ARE GIVEN BY NUMBER OF CHANNELS×HEIGHT×WIDTH. THE CONVOLUTION LAYERS ARE ALWAYS FOLLOWED BY BATCH NORMALIZATION. THE DROPOUT LAYERS HAVE KEEP PROBABILITIES OF 0.3 AND 0.5, RESPECTIVELY. IF THE SKIP CONNECTIONS (ARROWS) ARE APPLIED, THE FEATURE MAPS ARE DOWN SAMPLED INSTEAD BY APPLYING 1X1 FILTERS WITH A STRIDE OF 2X2. AP – AVERAGE POOLING, FC – FULLY CONNECTED

Layer	Type	Dimension	Activation	Out dim
0	Input	20×224×224	-	-
Layer	**Type**	**Convolution**	**Activation**	**Out dim**
1	Conv	7×7, 64, /2	ReLU	64×112×112
2	MP	3×3, -, /2	-	64×56×56
Block 1				
3	Conv	3×3, 64	ReLU	64×56×56
4	Conv	3×3, 64	-	64×56×56
5	Conv	3×3, 64	ReLU	64×56×56
6	Conv	3×3, 64	-	64×56×56
Block 2				
7	Conv	3×3, 128, /2	ReLU	128×28×28
8	Conv	3×3, 128	-	128×28×28
9	Conv	3×3, 128	ReLU	128×28×28
10	Conv	3×3, 128	-	128×28×28
Block 3				
11	Conv	3×3, 256, /2	ReLU	256×14×14
12	Conv	3×3, 256	-	256×14×14
13	Conv	3×3, 256	ReLU	256×14×14
14	Conv	3×3, 256	-	256×14×14
Block 4				
15	Conv	3×3, 512, /2	ReLU	512×7×7
16	Conv	3×3. 512	-	512×7×7
17	Conv	3×3, 512	ReLU	512×7×7
18	Conv	3×3, 512	-	512×7×7
Layer	**Type**	**Neurons**	**Activation**	**Out dim**
19	AP	512	-	512×1×1
20	FC	512	ReLU	512×1
21	Dropout	-	-	512×1
22	FC	128	ReLU	128×1
23	Dropout	-	-	128×1
24	FC	1	-	1×1

TABLE II
THE APPLIED MLP ARCHITECTURE FOR PREDICTING AHI VALUES BASED ON THE 3-DIMENSIONAL INPUT DEMOGRAPHICS. THE MLP USING QUESTIONNAIRES AS INPUT IS IDENTICAL EXCEPT FOR THE INPUT DIMENSIONS, WHICH ARE 7×1. FC – FULLY CONNECTED

Layer	Type	Dimension	Activation	Out dim
0	Input	3×1	-	-
Layer	**Type**	**Neurons**	**Activation**	**Out dim**
1	FC	32	ReLU	32×1
2	FC	64	ReLU	64×1
3	FC	1	-	1×1

B. Multi-View CNN

Our purpose for applying machine learning was to reveal data-driven mapping differences within the multi-view inputs across AHI values. For this purpose, we utilized a CNN with a ResNet18 architecture [41], which was modified to take 20 channels as input per subject. Two additional fully connected layers were added at the end, reducing 512 features to 128, followed by a ReLU activation and dropout (probability of 0.3 and 0.5, respectively), and finally a single neuron for prediction of the continuous AHI value. This architecture is illustrated in Table I. The optimal number of neurons in the fully connected layers were found by hyperparameter tuning. Additional networks were developed for the demographics and questionnaires, respectively. Both networks were multilayer perceptrons (MLPs) with three layers and varying number of inputs (three for demographics and seven for questionnaires). This architecture is shown in Table II. The AHI predictions of each of the three networks, using scans, demographics, and questionnaires, were averaged using an ensemble approach to form the final prediction.

Training, validation, and testing was carried out using 10-fold cross-validation. Mean squared error (MSE) was used as loss function to train the networks. This was done to penalize greater errors, since most of the AHI values in the dataset were in the range 0-15 events/hour (Fig. A.1.). Learning rate was set to $1 \cdot 10^{-5}$ with a weight decay of $5 \cdot 10^{-4}$ for the multi-view CNN, and $1 \cdot 10^{-2}$ (with the same weight decay) for the MLPs, while batch size was set to 8 for all three networks. The Adam optimizer [42] was used for training, and early stopping was applied when the validation error did not decrease for 3 consecutive epochs (patience of 3). Python 3.7.4 and Pytorch 1.3.1 were used for preprocessing and deep learning purposes. Training was carried out on a GeForce RTX 2080 and each model took one hour to train.

The measures used to evaluate model performance on the test set were mean absolute error (MAE) and Pearson Correlation Coefficient (PCC). Furthermore, predicted AHI was used to classify subjects into having OSA or not, where an AHI of 15 events/hour or greater was the clinical criterion used for defining the presence of OSA (moderate-severe versus mild or no sleep apnea). Thus, accuracy of classification was another measure used to evaluate the model performance. Bland-Altman plots [43] were used to illustrate the patterns of disagreement between true and predicted AHI values.

C. Sleep Specialists' Ability to Guess AHI

Three experienced, board certified sleep medicine physicians with in-depth knowledge of OSA were recruited to imitate the task of the proposed algorithm, estimating AHI based on inspection of the 3D scan of each subject (face and neck). The three physicians scored one third of the dataset each, while also annotating 150 of the same scans to estimate percentage agreement. When annotating a scan, each physician was shown the scan from all desired angles, having the ability to rotate the 3D image for any desired amount of time. Physicians took approximately 30 seconds to score each scan. Their first thought was to size up the person based on their estimated age, with the awareness that older individualsoften have higher AHIs (due to lax musculature, redundant tissue, and an atrophic skeletal scafolding that result in higher risk of airway collapse). Then they would ascertain if the individual looked tired - droopy eyelids (ptosis), drawn face, pallor, circles/bags under the eyes, etc. - that might suggest an underlying sleep disorder. Finally, they would look for some of the high-yield characteristics: looking at the overall head and neck adiposity (fat), the characteristics of the thyromental space, the over/under-bite (to suggest a retrognathic jaw that pushes the tongue into the airway), the craniofacial complex (looking for maxillary or mandibular hypoplasia) and noting whether there was a long/thin face suggestive of life-long nasal congestion ("adenoid facies"), and the cervical lordosis (to see if subjects have their heads thrust forward, suggestive of position modification in order to ease breathing). They would then roughly estimate the AHI based on all these factors and their general knowledge of the known prevalence/proportions of varying severities of OSA. It is important to note that clinicians do not traditionally estimate AHI, but for this study their estimates served as expert level performance as a comparison for the proposed model.

D. Topographic Display of Important Features

A topographic display was created by generating saliency maps for the model using craniofacial scans. A saliency map is a visualization technique based on the gradient of the network output with respect to an input image [44]. Consequently, the pixels which contribute most to the prediction of the network can be highlighted. For the topographic displays, we averaged saliency maps for 10 subjects with the highest predicted AHI values per fold and 10 subjects with the lowest predicted AHI values per fold, yielding an average of 100 saliency maps for the highest and lowest predicted AHI values, respectively.

IV. Results

A. Performance of Multi-view CNNs

The model using craniofacial scans during cross-validation converged after 6.6 ± 1.6 epochs, the model using demographics converged after 23.1 ± 6.9 epochs, and the model using questionnaires converged after 12.5 ± 6.9 epochs.

Without demographic and questionnaire information available, our model achieved a MAE of 11.38 ± 1.36 events/hour and a PCC of 0.40 ± 0.04 using 10-fold cross-validation. Fig. 3 shows the Bland-Altman plot of true and predicted AHI values. When dividing subjects from the test set into non-OSA/OSA using AHI $\geq$ 15 events/hour as a criterion for OSA, an overall accuracy of 67 ± 4% was obtained. Sensitivity was 59 ± 8%, specificity was 72 ± 5%, and area under the receiver operating characteristic (AUC ROC) was 65 ± 4%. Fig. 4(a) shows the resulting confusion matrix of the classification task. Cohen's kappa coefficient was 0.29 for the classification.

Adding clinically relevant demographics and questionnaire scores to the model yielded a MAE of 11.05 ± 1.40 and a PCC of 0.45 ± 0.04. Fig. 3 shows the Bland-Altman plot of the true and predicted AHI values. An accuracy of 67 ± 4% was achieved, with sensitivity of 74 ± 7%, specificity of 63 ± 7%, and AUC ROC of 69 ± 3%. Fig. 4(b) shows the resulting confusion matrix of the classification task. Cohen's kappa coefficient was 0.34 for the classification.

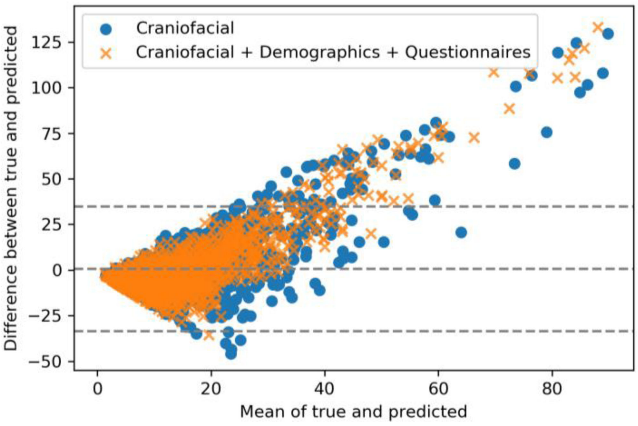


Fig. 3. Bland-Altman plot for the model using only craniofacial images and for the model using a combination of craniofacial images, demographics, and questionnaire scores. Means of true and predicted AHI values are displayed on the abscissa axis and difference between true and predicted AHI on the ordinate axis. The dashed horizontal lines above and below 0 indicate the borders of the 95% confidence interval.

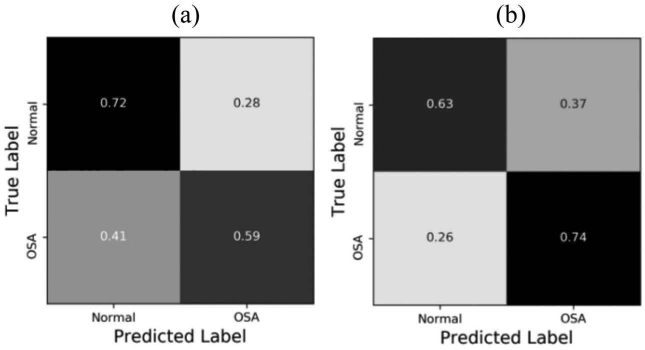


Fig. 4. (a) Confusion matrix for the model using only craniofacial images. (b) Confusion matrix for the model using a combination of craniofacial images, demographics, and questionnaire scores. The confusion matrix shows the results from classifying subjects into non-OSA/OSA using AHI $\geq$ 15 as a criterion for OSA based on the predicted AHI values by the proposed models.

Table III compares model performance to similar work in the literature. Performance measures for comparisons are MAE, PCC, accuracy, AUC ROC, and number of subjects used in each study. As an additional comparison, Table IV compares accuracies obtained using different methods on the same dataset, i.e., model with only demographics, or deriving a diagnosis from the modified STOP-Bang questionnaire, or a combination of both.

B. Performance of Sleep Medicine Specialists

Table V summarizes and compares results from the CNN to those of three sleep medicine specialists predicting AHI values based on 3D craniofacial scan displays from the test set. The percentage agreement between the three specialists was 67%. Fig. 5 compares the Bland-Altman plot of true and predicted AHIs from the model using craniofacial scans with a Bland-Altman plot of true and sleep specialists' AHIs.

TABLE III

COMPARISON BETWEEN THE TWO MODEL PERFORMANCES AND PERFORMANCE ACHIEVED IN THE LITERATURE. MAE – MEAN ABSOLUTE ERROR, PCC – PEARSON CORRELATION COEFFICIENT, AUC ROC – AREA UNDER THE RECEIVER OPERATING CHARACTERISTICS CURVE, N – TOTAL NUMBER OF SUBJECTS IN THE STUDY, N TEST – NUMBER OF SUBJECTS USED FOR TESTING, AMM - ACTIVE APPEARANCE MODEL, SVR – SUPPORT VECTOR REGRESSION, SVM – SUPPORT VECTOR MACHINE

Predictor	MAE	PCC	Accuracy	AUC ROC	N	N Test	Method	Validation scheme
Craniofacial	11.38	0.40	67%	65%	1366	1366	Automatic landmarks with CNN. AHI prediction and OSA classification using 2D images and depth maps from five angles and multi-view CNN.	10-fold cross-validation
Craniofacial + demographics + questionnaires	11.05	0.45	67%	69%	1366	1366	Same approach as above but with an ensemble of models using scans, demographics, and questionnaires, respectively.	10-fold cross-validation
Espinoza-Cuadros et al. [29]	12.56	0.37	71%	67%	285	285	Automatic landmarks with AMM. AHI prediction using measurements and SVR.	Leave-one-out cross-validation
Nosrati et al. [30]	13.4	0.52	68%	75%	180	180	Manual landmarks. AHI prediction using measurements and SVR.	Leave-one-out cross-validation
Balaei et al. [45]	-	-	69%	-	376	204	Automatic landmarks with SVM and cascade regression. OSA classification using measurements and logistic regression.	Training-test-set
Islam et al. [33]	-	-	67%	-	69	14	OSA classification using frontal depth maps and pre-trained VGGFace.	Training-validation-test-set
Islam et al. [33] on our dataset	-	-	60%	64%	1366	1366	Same approach as above.	10-fold cross-validation

TABLE IV

COMPARISON OF PERFORMANCE MEASURES OBTAINED USING DIFFERENT VARIABLES TO TRAIN AND TEST NEURAL NETWORKS USING AN ENSEMBLE APPROACH IN CASE OF TWO OR MORE MODALITIES. THE FIRST ROW IS OBTAINED BY USING THE MODIFIED STOP-BANG QUESTIONNAIRE AS IT IS UTILIZED CLINICALLY TO SCREEN FOR OSA, I.E., WITHOUT MACHINE LEARNING MAE – MEAN ABSOLUTE ERROR, PCC – PEARSON CORRELATION COEFFICIENT, AUC ROC – AREA UNDER THE RECEIVER OPERATING CHARACTERISTICS CURVE

Predictor	MAE	PCC	Accuracy	AUC ROC
Questionnaires	-	-	62±4%	65±4%
Questionnaire variables	11.42±1.27	0.38±0.07	64±4%	66±4%
Demographics	11.35±1.26	0.40±0.06	64±4%	67±3%
Scans	11.38±1.36	0.40±0.04	67±4%	65±4%
Demographics + Questionnaire variables	11.24±1.28	0.41±0.06	65±4%	67±4%
Scans + Demographics	11.12±1.36	0.44±0.03	67±3%	67±4%
Scans + Questionnaire variables	11.03±1.40	0.45±0.04	67±4%	68±4%
All Combined	11.05±1.36	0.45±0.04	67±4%	69±3%

TABLE V

COMPARISON OF THE MAIN PERFORMANCE MEASURES BETWEEN THE PROPOSED MODELS AND THREE SLEEP MEDICINE SPECIALISTS. MAE – MEAN ABSOLUTE ERROR, PCC – PEARSON CORRELATION COEFFICIENT, AUC ROC – AREA UNDER THE RECEIVER OPERATING CHARACTERISTICS CURVE

Predictor	MAE	PCC	Accuracy	AUC ROC
Craniofacial	11.38±1.36	0.40±0.04	67±4%	65±4%
Specialists data combined	13.34±1.39	0.35±0.12	66±4%	66±4%
Specialist 1	13.39±1.67	0.43±0.11	68±5%	66±4%
Specialist 2	14.33±0.69	0.42±0.04	61±2%	65±1%
Specialist 3	12.08±1.81	0.53±0.17	69±8%	72±8%

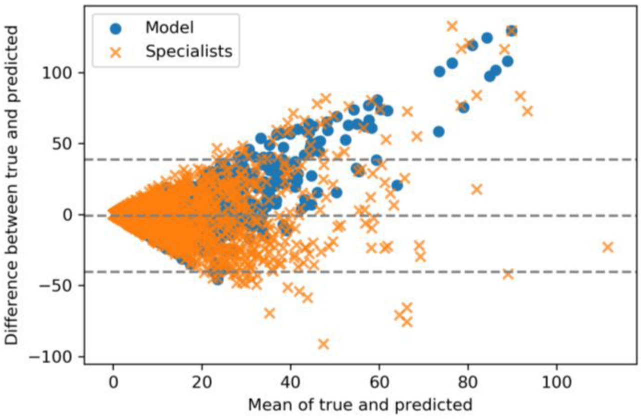


Fig. 5. Bland-Altman plot of the craniofacial model predictions compared to the sleep medicine specialists' predictions.

C. Topographic Display of Important Features

Fig. 6 (a) shows saliency maps averaged over the 100 subjects with the highest predicted AHI values, whereas Fig. 6 (b) shows saliency maps averaged over the 100 subjects with the lowest predicted AHI values.

V. Discussion

Our ML model based on craniofacial images alone achieved a MAE of 11.38 events/hour, a PCC of 0.40, and an overall accuracy of 67%. In comparison, if AHI for all subjects was predicted as the mean AHI value of the dataset (i.e., 15.5 events/hour), the MAE would be 13.0 events/hour, the PCC would be -0.02, accuracy would be 34%, and AUC ROC would be 50%. This means that the average absolute deviation from true AHI per subject would be almost 2 events/hour more than the proposed model, though both correlation and accuracy would be significantly worse.

Adding the clinically relevant demographics and questionnaire scores improved the model further, yielding a MAE of 11.05, a PCC of 0.45, and an accuracy of 67%. Importantly, however, sensitivity increased (from 59% to 74%), while specificity decreased (from 72% to 63%), suggesting that the model became better at predicting subjects with OSA with the added information. Although the overall accuracy did not increase, the AUC ROC improved after introducing demographics and questionnaire scores into the model. Although this performance may appear modest, both models achieved a higher accuracy than the 62% obtained in the same dataset, respectively, using the modified STOP-Bang questionnaire. Given that questionnaires are regularly used as an early screening tool for OSA, it is encouraging to observe that accuracy yielded using craniofacial scans exceeds that of the questionnaires, with added ability to provide an estimate of disease severity (i.e., AHI), which is not possible with a simple screening questionnaire.

The Bland-Altman plots shown in Fig. 3 show significant underpredictions for subjects with very high AHI (>30 events/hour). In general, subjects with high AHI are hard to estimate, which is evident when looking at the sleep specialists' scorings in Fig. 4, as they consistently underpredicted the higher AHI values as well. In OSA context, AHI values of 30, 60, or 90 events/hour are not significantly different, since they are all considered abnormal values. There were more than 200 subjects with AHI above 30 events/hour and 12 subjects with AHI more than 100 events/hour. Additionally, 50 subjects had an AHI of 0 events/hour, which were overpredicted on average by 10 events/hour. The subjects with very high AHI and AHI of 0 events/hour contribute to the high standard deviation (19.3) of AHI values in the dataset and increase the MAE as well.

Compared to similar work, the proposed models are at a similar level to all other studies in terms of overall accuracy (Table III). However, our models were trained and validated on a much larger cohort collected at 11 different sleep clinics and used a very different approach than that of others that predicted AHI using landmark-based measured features [29], [30], [45]. A lot of manual work is needed to derive landmark-based, hand-selected features as opposed to using an entirely data-driven approach as we propose. As such, our study is reassuring in that the empirically identified features emphasized by the model, recapitulated clinical expertise without the manual labor or years of clinical training and experience. Only Islam et al. [33] used images directly in a data-driven manner, but these authors only used depth information and only had craniofacial scans for 69 subjects as opposed to our 1366 subjects, which is equivalent to the patient volume seen over the entire course of a clinical sleep medicine training fellowship. We implemented the algorithm proposed by Islam et al. [33] on our dataset as seen in Table III, which decreased the overall accuracy from 67% to 60%. This makes sense because our dataset is much larger and much more diverse, since it was collected at many different sites. Furthermore, it shows that using images from several angles holds an advantage over using only frontal depth maps when predicting OSA, even when a pretrained network which has been trained on more than two million general facial images is utilized.

When comparing model performance to that of three sleep specialists guessing the AHI (Table V), it was observed that the model was almost at the same level as two specialists and better than one in terms of overall accuracy. Fig. 5 also shows that even though both our model and sleep specialists make large underpredictions for highest AHI values, the model does not make large overpredictions in the same manner as sleep specialists. All overpredicted values by the model are within the confidence interval which most likely stems from a bias in the model towards people with low to moderate AHI values, i.e., 5-30 events/hour. Of note, however, the most underpredicted values are underpredicted by sleep specialists by a similar magnitude. The percentage agreement between the specialists was 67%, again highlighting the presence of significant interscorer variability within the field of sleep medicine.

The topographic display of Fig. 6(a) shows that the network focuses mainly on the neck, jaw, and midface area when predicting high AHI values. These exact same regions have been reported in the literature as being the most important facial features related to OSA [8] and also the same regions that sleep specialists focused on when predicting AHI values. Interestingly however, when the network predicts low AHI values, as shown

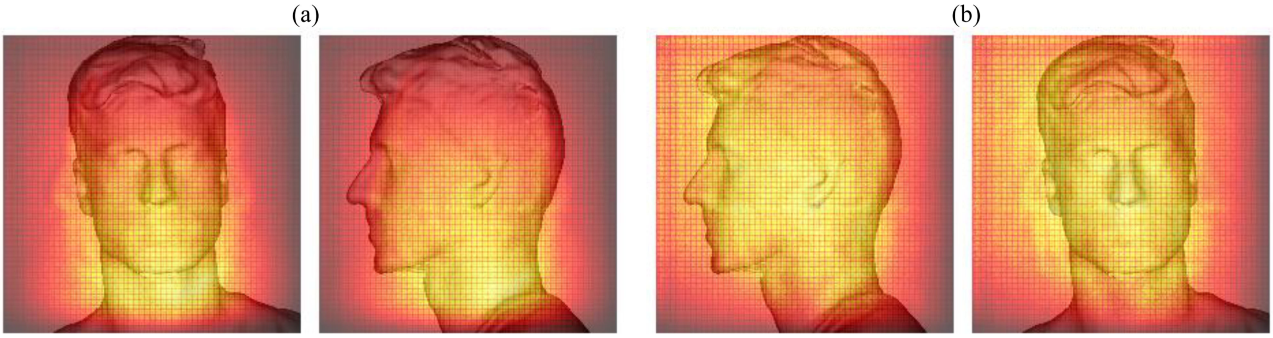


Fig. 6. (a) Saliency maps averaged over 100 subjects with the highest predicted AHI values. (b) Saliency maps averaged over 100 subjects with the lowest predicted AHI values.

in Fig. 6 (b), it seems to focus more selectively on regions of the craniofacial complex that reflect skeletal anatomy (e.g., the maxilla and mandible) for predictions. This may reflect the fact that subjects with milder AHI values have a different pathophysiology where skeletal abnormalities more than body fat may be causing airflow limitations. Even though our proposed model obtains similar performance compared to sleep specialists and similar work in the literature, an average absolute error of more than 11 events/hour is still quite high. First and foremost, it is important to keep in mind that the pathology of OSA is not exclusively attributed to obesity and craniofacial factors. That is also why performance increases only moderately when adding extra information such as demographics and questionnaires. Other variables that could not be assessed from our dataset are internal upper airway anatomy factors, such as size and positioning of the tongue and palate. Other reasons for uncertainty may be physiological changes that occur with age, independent of facial anatomy or obesity such as recruitment of upper airway dilator muscles. Nonetheless, even if the features identified by the model are primarily anatomic in nature, these findings may prove useful to determine phenotypic risk for certain types of OSA. Finally, it should be noted that because our 3D scans were of structural facial features, we intentionally excluded central/mixed apneas from the AHI, in order to focus our model on the anatomical contributions to OSA. However, there are various models of this complex disorder that also account for the physiologic aspects of OSA (e.g., loop gain [46]), which is something that future modeling efforts should take into account.

Another limitation of this study was the quality and quantity of the captured 3D scans. The quality of the scans varied significantly and reflected the fact that they were captured in many different sleep clinics. Some scans had missing parts of the neck, whereas others were affected by poor lighting conditions. Furthermore, we believe that the size of the dataset was too small to truly capture the variation in craniofacial features across humans in relation to OSA in a data-driven manner. Evidently, we observe that similar performance is obtainable in smaller datasets if the features are hand-crafted like landmark-based measurements. The fact that scans were captured either at night before the PSG or in the morning after did not have any effect on diagnostic performance, which was evident when we obtained accuracies of 68% and 66% for scans captured in the morning and at night, respectively, showing that scans can be obtained in both conditions.

Although the focus in this study is on providing a fast, efficient, and cheap screening tool for OSA, efforts to explore alternative screening methods include sleep tests at home [47], usually with very few sensors, such as sound [48], [49] and blood oxygen saturation [50], [51]. Potential of depth and thermal cameras has also been explored in breathing monitoring at an early stage [52], [53]. Even contactless bed sensors have been proposed, although only with moderate success so far [54]. The benefit of using the mentioned approaches is that the person is more comfortable sleeping in their own home wearing few or no wires. However, most studies use a small number of patients to validate their techniques and still requires a full night's sleep to reach a diagnosis.

In future work, it would be interesting to explore if, beside AHI, other clinically important variables captured by sleep studies could be better predicted. These could prove to not only make up a more accurate model but could also improve our knowledge of OSA phenotypes and their relation to facial anatomy. For example, Azarbarzin et al. [55] recently suggested that the hypoxic burden is a better measure to use compared to AHI when evaluating sleep apnea severity and resulting cardiovascular risk. Predicting hypoxic burden, oxygen desaturation index (ODI), or the duration of events instead of the AHI, or newer derivatives that better describe sleep disorder breathing heterogenicity could prove more useful and insightful.

VI. Conclusion

The main purpose of this study was to develop an automatic algorithm predicting AHI based on 3D craniofacial images that can be captured in a minute by a non-specialist. This was achieved by converting the 3D images into a series of 2D images and depth maps and utilizing a multi-view CNN for learning. Two models were implemented, one based exclusively on craniofacial images and one using a combination of craniofacial images, demographics, and questionnaire variables. Using these models, a MAE of 11.38 events/hour was obtained with an accuracy of 67%, which was at a similar level to performance achieved in similar work and higher than the current screening method, i.e., questionnaires. A topographic display was created, highlighting

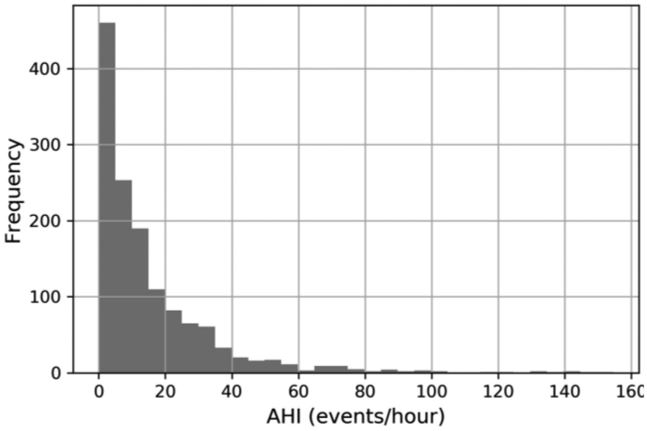


Fig. A.1. The distribution of apnea-hypopnea index (AHI) values in the dataset.

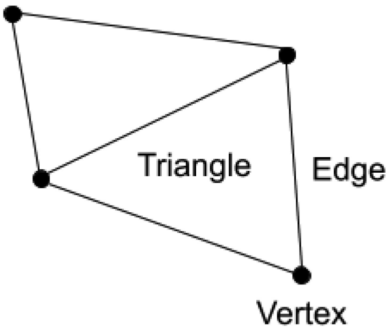


Fig. A.2. Example of four vertices that have been connected using edges, forming two triangles, which are the basic blocks of a 3D mesh surface.

the most important regions of the face when predicting OSA in a data-driven manner. These regions corresponded well with what is reported in the medical literature. The obtained results were at a similar level to two sleep specialists imitating the task and better than one. These sleep specialists had an overall agreement of 67% when scoring the scans. With this work, we have shown that it is possible to derive AHI values based on 3D scans and deep learning techniques for some OSA phenotypes and on a level similar to or higher than vastly experienced physicians. The proposed model has the potential to serve as a clinical screening tool for suspected OSA patients before they undergo a PSG.

Appendix

Fig. A.1. shows the distribution of AHI values within the dataset, whereas Fig. A.2. example of the basic components of a mesh, i.e., vertices, edges, and triangles.

Acknowledgment

We would like to thank all clinic research coordinators who helped collect data as part of STAGES from the following sites: Geisinger Medical Center, Mayo Clinic, MedSleep, SleepMed, St. Luke's Hospital, and Stanford University. We would also like to thank Peter Yoo (), Sungwook Su (), and Terry Izumi () from uGo3D for developing software for the collection of 3D scans.

Dr. Leary is a full-time employee of Jazz Pharmaceuticals, who, during this employment, has received stock options exercisable for, and other stock awards of, ordinary shares of Jazz Pharmaceuticals, plc.

Dr. Schneider reports other support from Sierra Pacific (VISN21) VA Mental Illness Research Education Clinical, Centers of Excellence, during the conduct of the study; personal fees from Alphabet, Inc, personal fees from Jazz Pharmaceuticals, outside the submitted work.

Dr. Morse reports personal fees from Jazz Pharmaceuticals and Rhythm Pharmaceuticals unrelated to this work. She has received research or clinical trial funding from NIH/NIMH and Jazz Pharmaceuticals unrelated to this work.

Dr. Mignot has received research or clinical trial funding unrelated to this work from Jazz Pharmaceuticals, Merck, Takeda, Sunovion and Apple. He owns stocks from Alerion, Dreem, Inoxia and Orexia, whose work is unrelated to this study, and is on an advisory board for Phillips.

The remaining authors have no conflicts of interest to declare.

References

[1] P. Lévy *et al.*, "Obstructive sleep apnoea syndrome," *Nature Rev. Dis. Primers*, vol. 1, no. 1, pp. 1–21, 2015, doi: 10.1038/nrdp.2015.43.

[2] L. F. Drager, S. M. Togeiro, V. Y. Polotsky, and G. Lorenzi-Filho, "Obstructive sleep apnea," *J. Amer. Coll. Cardiol.*, vol. 62, no. 7, pp. 569–576, 2013, doi: 10.1016/j.jacc.2013.05.045.

[3] A. S. M. Shamsuzzaman, B. J. Gersh, and V. K. Somers, "Obstructive sleep apnea implications for cardiac and vascular disease," *JAMA*, vol. 290, no. 14, pp. 1906–1914, 2003, doi: 10.1001/jama.290.14.1906.

[4] R. Heinzer *et al.*, "Prevalence of sleep-disordered breathing in the general population: The hypnolaus study," *Lancet Respir. Med.*, vol. 3, no. 4, pp. 310–318, 2015, doi: 10.1016/S2213-2600(15)00043-0.

[5] A. M. Osman, S. G. Carter, J. C. Carberry, and D. J. Eckert, "Obstructive sleep apnea: Current perspectives," *Nature Sci. Sleep*, vol. 10, pp. 21–34, 2018, doi: 10.2147/NSS.S124657.

[6] A. R. Schwartz, S. P. Patil, A. M. Laffan, V. Polotsky, H. Schneider, and P. L. Smith, "Obesity and obstructive sleep apnea: Pathogenic mechanisms and therapeutic approaches," *Proc. Amer. Thorac. Soc.*, Feb. 2008, vol. 5, no. 2, pp. 185–192, doi: 10.1513/pats.200708-137MG.

[7] J. A. Dempsey, S. C. Veasey, B. J. Morgan, and C. P. O'donnell, "Pathophysiology of sleep apnea," *Physiol. Rev.*, vol. 90, no. 1, pp. 47–112, 2010, doi: 10.1152/physrev.00043.2008.

[8] R. W. W. Lee, K. Sutherland, and P. A. Cistulli, "Craniofacial morphology in obstructive sleep apnea: A review," *Clin. Pulm. Med.*, vol. 17, no. 4, pp. 189–195, 2010, doi: 10.1097/CPM.0b013e3181e4bea7.

[9] C. A. Kushida *et al.*, "Practice parameters for the indications for polysomnography and related procedures: An update for 2005," *Sleep*, vol. 28, no. 4, pp. 499–523, 2005, doi: 10.1093/sleep/28.4.499.

[10] R. B. Berry *et al.*, "The AASM manual for the scoring of sleep and associated events: Rules, terminology, and technical specifications version 2.2," *Amer. Acad. Sleep Med.*, 2015. [Online]. Available: www.aasmnet.org

[11] N. Collop, "Scoring variability between polysomnography technologists in different sleep laboratories," *Sleep Med.*, vol. 3, no. 1, pp. 43–47, 2002, doi: 10.1016/S1389-9457(01)00115-0.

[12] M. Younes *et al.*, "Reliability of the american academy of sleep medicine rules for assessing sleep depth in clinical practice," *J. Clin. Sleep Med.*, vol. 14, no. 2, pp. 205–213, 2018, doi: 10.5664/jcsm.6934.

[13] J. B. Stephansen *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Commun.*, vol. 9, no. 1, pp. 1–15, Dec. 2018, doi: 10.1038/s41467-018-07229-3.

[14] U. Hanif *et al.*, "Non-invasive machine learning estimation of effort differentiates sleep-disordered breathing pathology," *Physiol. Meas.*, vol. 40, no. 2, Feb. 2019, Art. no. 025008, doi: 10.1088/1361-6579/ab0559.

[15] A. Brink-Kjaer *et al.*, “Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness,” *Clin. Neurophysiol.*, vol. 131, no. 6, pp. 1187–1203, Jun. 2020, doi: 10.1016/j.clinph.2020.02.027.

[16] M. Piriyajitakonkij *et al.*, “SleepPoseNet: Multi-view learning for sleep postural transition recognition using UWB,” *IEEE J. Biomed. Heal. Inform.*, vol. 25, no. 4, pp. 1305–1314, Apr. 2021.

[17] N. Banluesombatkul *et al.*, “MetaSleepLearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning,” *IEEE J. Biomed. Heal. Inform.*, pp. 1–1, to be published, doi: 10.1109/JBHI.2020.3037693.

[18] R. J. Schwab *et al.*, “Digital morphometrics: A new upper airway phenotyping paradigm in OSA,” *Chest*, vol. 152, no. 2, pp. 330–342, 2017, doi: 10.1016/j.chest.2017.05.005.

[19] C. Guilleminault, R. Riley, and N. Powell, “Obstructive sleep apnea and abnormal cephalometric measurements: Implications for treatment,” *Chest*, vol. 86, no. 5, pp. 793–794, 1984, doi: 10.1378/chest.86.5.793.

[20] A. Jamieson, C. Guilleminault, M. Partinen, and M. A. Quera-Salva, “Obstructive sleep apneic patients have craniomandibular abnormalities,” *Sleep*, vol. 9, no. 4, pp. 469–477, 1986, doi: 10.1093/sleep/9.4.469.

[21] A. A. Lowe, J. A. Fleetham, S. Adachi, and C. F. Ryan, “Cephalometric and computed tomographic predictors of obstructive sleep apnea severity,” *Amer. J. Orthod. Dentofac. Orthop.*, vol. 107, no. 6, pp. 589–595, 1995, doi: 10.1016/s0889-5406(95)70101-x.

[22] T. Ogawa, R. Enciso, W. H. Shintaku, and G. T. Clark, “Evaluation of cross-section airway configuration of obstructive sleep apnea,” *Oral Surgery Oral Med. Oral Pathol. Oral Radiol. Endodontol.*, vol. 103, no. 1, pp. 102–108, 2007, doi: 10.1016/j.tripleo.2006.06.008.

[23] R. J. Schwab *et al.*, “Identification of upper airway anatomic risk factors for obstructive sleep apnea with volumetric magnetic resonance imaging,” *Amer. J. Respir. Crit. Care Med.*, vol. 168, no. 5, pp. 522–530, 2003, doi: 10.1164/rccm.200208-866OC.

[24] M. Okubo *et al.*, “Morphologic analyses of mandible and upper airway soft tissue by MRI of patients with obstructive sleep apnea hypopnea syndrome,” *Sleep*, vol. 29, no. 7, pp. 909–915, 2006, doi: 10.1093/sleep/29.7.909.

[25] R. W. W. Lee *et al.*, “Relationship between surface facial dimensions and upper airway structures in obstructive sleep apnea,” *Sleep*, vol. 33, no. 9, pp. 1249–1254, 2010, doi: 10.1093/sleep/33.9.1249.

[26] E. J. Kezirian, W. Hohenhorst, and N. de Vries, “Drug-induced sleep endoscopy: The VOTE classification,” *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 268, no. 8, pp. 1233–1236, 2011

[27] R. W. W. Lee, A. S. L. Chan, R. R. Grunstein, and P. A. Cistulli, “Craniofacial phenotyping in obstructive sleep apnea—A novel quantitative photographic approach,” *Sleep*, vol. 32, no. 1, pp. 37–45, 2009, doi: 10.5665/sleep/32.1.37.

[28] R. W. W. Lee, P. Petocz, T. Prvan, A. S. L. Chan, R. R. Grunstein, and P. A. Cistulli, “Prediction of obstructive sleep apnea with craniofacial photographic analysis,” *Sleep*, vol. 32, no. 1, pp. 46–52, 2009, doi: 10.5665/sleep/32.1.46.

[29] 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, 2015, Art. no. 489761, doi: 10.1155/2015/489761.

[30] H. Nosrati, N. Sadr, and P. de Chazal, “Apnoea-hypopnoea index estimation using craniofacial photographic measurements,” in *Proc. Comput. Cardiol. Conf.*, 2016, pp. 1033–1036.

[31] P. De Chazal, A. T. Balaei, and H. Nosrati, “Screening patients for risk of sleep apnea using facial photographs,” in *Proc. 39th Annu. Int. Conf. IEEE Eng. Medicine Biol. Soc.*, 2017, pp. 2006–2009.

[32] A. T. Balaei, K. Sutherland, P. A. Cistulli, and P. de Chazal, “Automatic detection of obstructive sleep apnea using facial images,” in *Proc. IEEE 14th Int. Symp. Biomed. Imag.*, 2017, pp. 215–218.

[33] S. M. S. Islam, H. Mahmood, A. A. Al-Jumaily, and S. Claxton, “Deep learning of facial depth maps for obstructive sleep apnea prediction,” in *Proc. IEEE Int. Conf. Mach. Learn. Data Eng.*, 2018, pp. 154–157.

[34] U. Hanif, R. R. Paulsen, E. B. Leary, E. Mignot, P. Jennum, and H. B. D. Sorensen, “Prediction of patient demographics using 3D craniofacial scans and Multi-view CNNs,” in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2020, pp. 1950–1953.

[35] M. Kalantari, and M. Nechifor, “Accuracy and utility of the structure sensor for collecting 3D indoor information,” *Geo-Spatial Inf. Sci.*, vol. 19, no. 3, pp. 202–209, Jul. 2016, doi: 10.1080/10095020.2016.1235817.

[36] P. G. M. Knoops *et al.*, “Comparison of three-dimensional scanner systems for craniomaxillofacial imaging,” *J. Plast. Reconstr. Aesthetic Surg.*, vol. 70, no. 4, pp. 441–449, Apr. 2017, doi: 10.1016/j.bjps.2016.12.015.

[37] 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, doi: 10.1378/chest.15-0903.

[38] R. R. Paulsen, K. A. Juhl, T. M. Haspang, T. Hansen, M. Ganz, and G. Einarsson, “Multi-view consensus CNN for 3D facial landmark placement,” in *Proc. Asian Conf. Comput. Vis.*, 2018, pp. 706–719, doi: 10.1007/978-3-030-20887-5_44.

[39] S. Zulqarnain Gilani, F. Shafait, and A. Mian, “Shape-based automatic detection of a large number of 3D facial landmarks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4639–4648.

[40] C. M. Grewe and S. Zachow, “Fully automated and highly accurate dense correspondence for facial surfaces,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 552–568.

[41] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014. *arXiv1412.6980 [cs.LG]*.

[43] D. G. Altman and J. M. Bland, “Measurement in medicine: The analysis of method comparison studies,” *J. R. Statist. Soc. Ser. D (Statist.)*, vol. 32, pp. 307–317, 1983, doi: 10.2307/2987937.

[44] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” 2013. *arXiv1312.6034 [cs.CV]*.

[45] A. T. Balaei, K. Sutherland, P. Cistulli, and P. de Chazal, “Prediction of obstructive sleep apnea using facial landmarks,” *Physiol. Meas.*, vol. 39, 2018, Art. no. 094004, doi: 10.1088%2F1361-6579%2Faadb35

[46] D. J. Eckert, D. P. White, A. S. Jordan, A. Malhotra, and A. Wellman, “Defining phenotypic causes of obstructive sleep apnea: Identification of novel therapeutic targets,” *Amer. J. Respir. Crit. Care Med.*, vol. 188, no. 8, pp. 996–1004, Oct. 2013, doi: 10.1164/rccm.201303-0448OC.

[47] 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, Oct. Jan., 2018, doi: 10.1016/j.smrv.2018.02.004.

[48] H. Nakano, T. Furukawa, and T. Tanigawa, “Tracheal sound analysis using a deep neural network to detect sleep apnea,” *J. Clin. Sleep Med.*, vol. 15, no. 8., pp. 1125–1133, Aug. 15, 2019, doi: 10.5664/jcsm.7804.

[49] S. Hayashi, M. Tamaoka, T. Tateishi, Y. Murota, I. Handa, and Y. Miyazaki, “A new feature with the potential to detect the severity of obstructive sleep apnoea via snoring sound analysis,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 8, pp. 2951, 2020, doi: 10.3390/ijerph17082951.

[50] S. Nikkonen, I. O. Afara, T. Leppänen, and J. Töyräs, “Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Dec. 2019, doi: 10.1038/s41598-019-49330-7.

[51] F. Mendonça, S. Mostafa, F. Morgado-Dias, and A. G. Ravelo-García, “An oximetry based wireless device for sleep apnea detection,” *Sensors*, vol. 20, no. 3, pp. 888, 2020, doi: 10.3390/s20030888.

[52] A. Al-Naji, A. J. Al-Askery, S. K. Gharghan, and J. Chahl, “A system for monitoring breathing activity using an ultrasonic radar detection with low power consumption,” *J. Sens. Actuator Netw.*, vol. 8, no. 2, p. 32, May 2019, doi: 10.3390/jsan8020032.

[53] A. Procházka, and M. Schätz, O. Ťupa, M. Yadollahi, O. Vyšata, and M. Walls, “The MS kinect image and depth sensors use for gait features detection,” *ICIP 2014*, pp. 7025460, 2014, doi: 10.1109/ICIP.2014.7025460.

[54] I. Sadek, T. Tan Soon Heng, E. Seet, and B. Abdulrazak, “A new approach for detecting sleep apnea using a contactless bed sensor: Comparison study,” *J. Med. Internet Res.*, vol. 22, no. 9, Sep. 2020, Art. no. e18297, doi: 10.2196/18297.

[55] A. Azarbarzin *et al.*, “The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: The osteoporotic fractures in men study and the sleep heart health study,” *Eur. Heart J.*, vol. 40, pp. 1149–1157, 2019, doi: 10.1093/eurheartj/ehy624.

Appendix D

Paper III

Title: Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks

Authors: Umaer Hanif, Eric Kezirian, Eva K. Kiær, Emmanuel Mignot, Helge B. D. Sorensen, and Poul Jennum

Journal: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)

Status: Published

DOI: 10.1109/EMBC46164.2021.9630098

Full citation: U. Hanif, E. Kezirian, E. K. Kiær, E. Mignot, H. B. D. Sorensen, and P. Jennum, "Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks", *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3957-3960, 2021. DOI: 10.1109/EMBC46164.2021.9630098.

2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks*

Umaer Hanif[1,3,4], *Member, IEEE*, Eric Kezirian[2,5], Eva Kirkegaard Kiær[3,5], Emmanuel Mignot[4,5],
Helge B. D. Sorensen[1,5], *Senior Member, IEEE*, and Poul Jennum[3,5]

Abstract— **Assessing the upper airway (UA) of obstructive sleep apnea patients using drug-induced sleep endoscopy (DISE) before potential surgery is standard practice in clinics to determine the location of UA collapse. According to the VOTE classification system, UA collapse can occur at the velum (V), oropharynx (O), tongue (T), and/or epiglottis (E). Analyzing DISE videos is not trivial due to anatomical variation, simultaneous UA collapse in several locations, and video distortion caused by mucus or saliva. The first step towards automated analysis of DISE videos is to determine which UA region the endoscope is in at any time throughout the video: V (velum) or OTE (oropharynx, tongue, or epiglottis). An additional class denoted X is introduced for times when the video is distorted to an extent where it is impossible to determine the region. This paper is a proof of concept for classifying UA regions using 24 annotated DISE videos. We propose a convolutional recurrent neural network using a ResNet18 architecture combined with a two-layer bidirectional long short-term memory network. The classifications were performed on a sequence of 5 seconds of video at a time. The network achieved an overall accuracy of 82% and F1-score of 79% for the three-class problem, showing potential for recognition of regions across patients despite anatomical variation. Results indicate that large-scale training on videos can be used to further predict the location(s), type(s), and degree(s) of UA collapse, showing potential for derivation of automatic diagnoses from DISE videos eventually.**

I. INTRODUCTION

Obstructive sleep apnea (OSA) is a sleep disorder during which the upper airway (UA) collapses throughout the night, causing events with partial or complete cessation of breathing during sleep [1]. The development of OSA can be physiologically caused (loop gain, arousal threshold, poor recruitment of dilator muscles) [2] or anatomically caused (craniofacial abnormalities, obesity, narrow UA) [3] and treatment varies depending on the underlying cause. If the pathology of OSA has an anatomical component, surgery may be necessary for treatment [4]. Prior to a potential surgical procedure, it is critical to examine the location(s) of collapse in the UA, which according to the VOTE classification system [5] can occur on four different levels: velum, oropharynx, tongue, and/or epiglottis. The examination is commonly performed using drug-induced sleep endoscopy (DISE) during which the surgeon navigates the endoscope from the velum to the epiglottis to determine the location(s), type(s), and degree(s) of collapse occuring in the UA during OSA events [6].

Analyzing DISE videos to determine the appropriate type of surgery is not a trivial task. First, there is a huge anatomical variation in the UA across subjects. Additionally, movements in the UA stemming from several structures collapsing simultaneously push the endoscope back and forth, while mucus or saliva covering the endoscope distorts the video and reduces quality significantly. These challenges are reflected in a relatively high interscorer variability when different surgeons analyze DISE videos [7]. Due to these limitations, surgeons will benefit from an algorithm capable of analyzing DISE videos automatically to assist in determining the locations(s), type(s), and degree(s) of collapse.

The first step towards such a goal is being able to estimate which region of the UA the endoscope is in at any given time. The clinically meaningful distinction is between the velum (V) and anything below the velum (OTE). Thus, the aim of this study is to classify whether the endoscope is in the V or OTE region at any given time in a DISE video. Furthermore, we introduce a third class (X) for any time the video is so distorted that it is impossible to determine where the endoscope is. For this problem, we propose a convolutional recurrent neural network (CRNN) which is trained, validated and tested on a small dataset of annotated DISE videos. This study is the first attempt to apply a data-driven approach to identify regions in the UA during a DISE procedure.

II. DATA DESCRIPTION

We included a total of 24 DISE videos collected at Copenhagen University Hospital, which were performed in accordance with the DISE procedure guideline described by Kiaer et al. [8]. The Institution's Ethical Review Board approved all experimental procedures involving human subjects.

The videos were approximately 2-5 minutes in duration with a frame rate of 25 frames per second and a resolution of 864×540 pixels. All videos were anonymized by removing parts of recordings where the endoscope was not inside the

*Research has been supported by the Klarman Family Foundation, Stanford University, Technical University of Denmark, and Rigshospitalet with supporting grants from Danmark-Amerika Fondet, Vera og Carl Johan Michaelsens Legat, Reinholdt W. Jorck og Hustrus Fond, Torben og Alice Frimodts Fond, Christian og Ottilia Brorsons Rejselegat, Marie og M.B. Richters Fond, Oberstløjtnant Max Nørgaard og hustru Magda Nørgaards Legat, William Demant Fonden, Augustinus Fonden, Rudolph Als Fondet, Knud Højgaards Fond, Otto Mønsteds Fond, Julie von Müllens Fond, and Direktør Einar Hansen og hustru fru Vera Hansens Fond

[1] Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, umaerhanif@hotmail.com

[2] USC Caruso Department of Otolaryngology - Head & Neck Surgery, Keck School of Medicine of USC, Los Angeles, CA 90033, USA

[3] Danish Center for Sleep Medicine, Rigshospitalet, 2600 Glostrup, Denmark

[4] Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA 94304, USA

[5] Shared last authors

subject. Each video was initially labeled by the surgeon who collected them as a single line summary of where, how, and to what degree the UA collapsed. However, for machine learning purposes, labels were required that detailed each time the endoscope transitioned either from one region to another (i.e. V to OTE or OTE to V) or from visible video to distorted video or vice versa (i.e. V to X, OTE to X, X to V or X to OTE). Videos were labeled in this manner by consulting with the surgeon who initially labeled the videos and another expert surgeon who introduced the VOTE classification in 2011 [5]. Fig. 1 visualizes different examples of the three classes, while Table I shows an example of the structure of labels created for this study. Finally, Table II outlines the distribution of the three classes within the dataset.

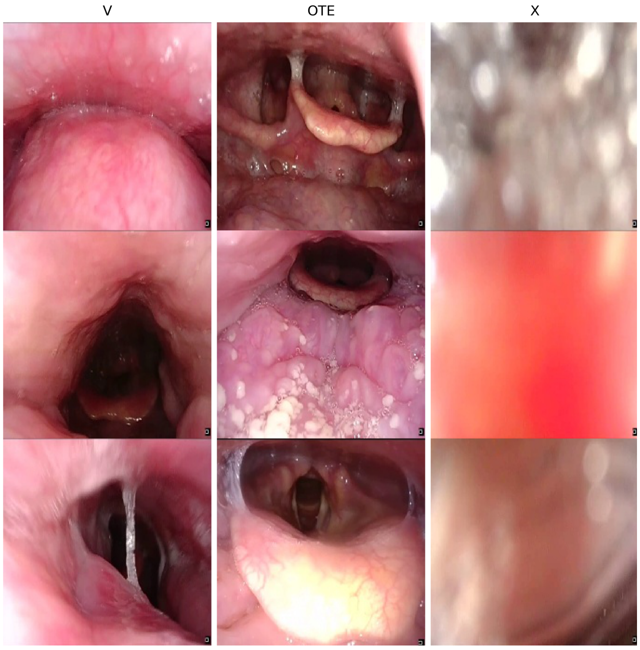


Fig. 1. Three examples of each class representing a region in the upper airway, i.e. velum (V) in the first column, oropharynx, tongue or epiglottis (OTE) in the second column, and distortion in video (X) in the third column.

TABLE I

EXAMPLE OF LABELS CREATED FOR PART OF A DISE VIDEO USING THE THREE CLASSES, I.E. VELUM (V), OROPHARYNX, TONGUE OR EPIGLOTTIS (OTE), AND DISTORTION IN VIDEO (X).

Time (s)	7-15	16-28	29-35	36-40	40-45
Region	V	X	OTE	X	V

TABLE II

DISTRIBUTION OF THE THREE CLASSES IN THE DATASET: VELUM (V), OROPHARYNX, TONGUE OR EPIGLOTTIS (OTE), AND DISTORTION IN VIDEO (X).

Class	Total duration (s)	N Frames
V	1,543	7,715
OTE	2,041	10,205
X	376	1,880
Total	3,960	19,800

III. METHODS

A. Preprocessing

Initially, all frames were extracted from each video, yielding 25 frames per second. Subsequently, every 5th frame was selected, yielding 5 frames per second, because no visual difference was observed between consecutive frames during inspection. Assuming the network would extract features primarily related to anatomical structures and not color differences, all frames were converted to gray scale to reduce computational cost of training the subsequent network. All frames were rescaled to 224×224 pixels, which was found to be appropriate for reducing computational cost while still preserving discriminatory information between UA structures. Finally, the dataset was split into a training set (18 videos amounting to 15,275 frames), a validation set (3 videos amounting to 2,375 frames), and a test set (3 videos amounting to 2,150 frames).

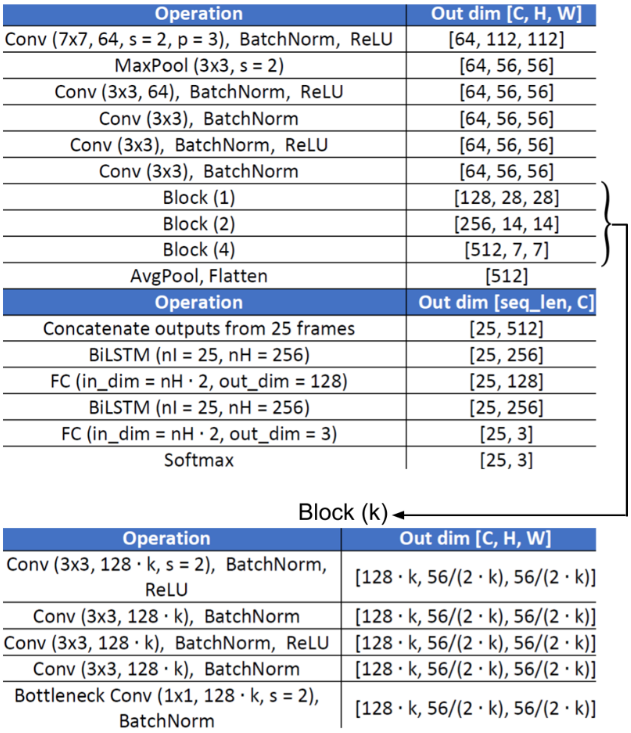
B. Convolutional Recurrent Neural Network

The proposed network architecture for learning was a combination of a ResNet18 [9] convolutional neural network (CNN) and a two-layer bidirectional long short-term memory (LSTM) neural network [10] as shown in Fig. 2. The input layer of the ResNet18 model was modified to take a 1-channel input instead of RGB images with 3 channels, since the frames were grayscale. The input consisted of 25 frames amounting to 5 seconds at 5 frames per second. Each frame was individually input to the CNN and resulting outputs were subsequently concatenated, forming a 25×512 dimensional feature matrix, i.e. 25-time steps each with 512 features. This sequence of features was then input to the bidirectional LSTM to learn context in both forward and backward directions. Both LSTM layers had 128 hidden neurons in both directions followed by a softmax activation function with three outputs such that each class had an output probability. The optimal number of time-steps and hidden neurons were found using hyperparameter tuning.

Optimization of the network was performed using a batch size of 2 with cross entropy as loss function and Adam [11] as optimizer. Weights were applied in the loss function for the V and X classes, since the dataset was heavily imbalanced as witnessed in Table II. The weights were calculated as the ratio between the majority class (OTE) and a given other class. The learning rate was set to $1 \cdot 10^{-5}$ with a weight decay of $5 \cdot 10^{-4}$. Early stopping was applied when the validation loss did not decrease for 3 consecutive epochs. The network was implemented in Pytorch and all experiments were carried out on a GeForce RTX 2080 graphics card. The model took approximately one hour to train on this dataset.

C. Performance

Model performance was evaluated on the three videos in the test set. Accuracy, F1-score, and the confusion matrix were computed by summing correct classifications on a frame-by-frame basis and averaged over individual videos as well as over the entire test set, respectively.



Operation	Out dim [C, H, W]
Conv (7x7, 64, s = 2, p = 3), BatchNorm, ReLU	[64, 112, 112]
MaxPool (3x3, s = 2)	[64, 56, 56]
Conv (3x3, 64), BatchNorm, ReLU	[64, 56, 56]
Conv (3x3), BatchNorm	[64, 56, 56]
Conv (3x3), BatchNorm, ReLU	[64, 56, 56]
Conv (3x3), BatchNorm	[64, 56, 56]
Block (1)	[128, 28, 28]
Block (2)	[256, 14, 14]
Block (4)	[512, 7, 7]
AvgPool, Flatten	[512]

Operation	Out dim [seq_len, C]
Concatenate outputs from 25 frames	[25, 512]
BiLSTM (nI = 25, nH = 256)	[25, 256]
FC (in_dim = nH · 2, out_dim = 128)	[25, 128]
BiLSTM (nI = 25, nH = 256)	[25, 256]
FC (in_dim = nH · 2, out_dim = 3)	[25, 3]
Softmax	[25, 3]

Block (k)

Operation	Out dim [C, H, W]
Conv (3x3, 128 · k, s = 2), BatchNorm, ReLU	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm, ReLU	[128 · k, 56/(2 · k), 56/(2 · k)]
Conv (3x3, 128 · k), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]
Bottleneck Conv (1x1, 128 · k, s = 2), BatchNorm	[128 · k, 56/(2 · k), 56/(2 · k)]

Fig. 2. Architecture for the proposed network for classifying UA regions. The input is a 5-second video consisting of 25 frames. The frames are input individually to the CNN and the outputs are concatenated before the recurrent part of the network. The parameters in the convolution operations (Conv) are kernel size, number of output channels, stride (s), and padding (p), and the output dimensions are specified by number of channels (C), height (H), and width (W). The parameters in the bidirectional LSTM (BiLSTM) are number of input features (nI) and number of hidden neurons in each direction (nH). The parameters in the fully connected layers (FC) are input features (in_dim) and output features (out_dim).

IV. RESULTS

The best performing model converged after 2 epochs of training. An overall accuracy of 82% and F1-score of 79% was obtained over the entire test set. Furthermore, F1-scores for V, OTE, and X were 68%, 80%, and 88%, respectively. Fig. 3 shows the confusion matrix for the classification, while Fig. 4 depicts examples of misclassified frames for each class. Table III summarizes the performance for each of the 3 individual videos in the test set, respectively.

V. DISCUSSION

This is the first attempt to use a data-driven approach to identify UA regions during the DISE procedure and the overall accuracy and F1-score obtained using the proposed model was 82% and 79%, respectively. In contrast, if the network had simply predicted all frames to be the majority class in the test set (i.e. OTE), the overall accuracy and F1-score would be 52% and 23%, respectively. In this context, the model performs much better than random guessing. In terms of class F1 scores, the model performed best for the X class, then OTE, and finally V.

The X class is intuitively the easiest to recognize since it means that the video is too distorted to derive anything and

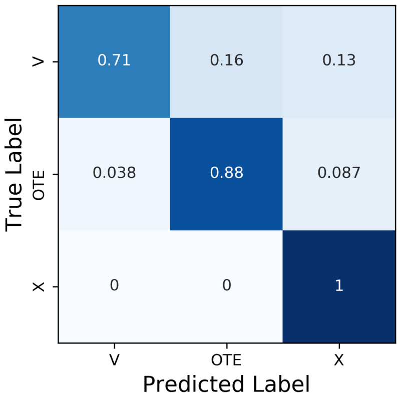


Fig. 3. Confusion matrix for classifying regions in the upper airway with three different classes: velum (V), oropharynx, tongue or epiglottis (OTE), and distorted video (X).

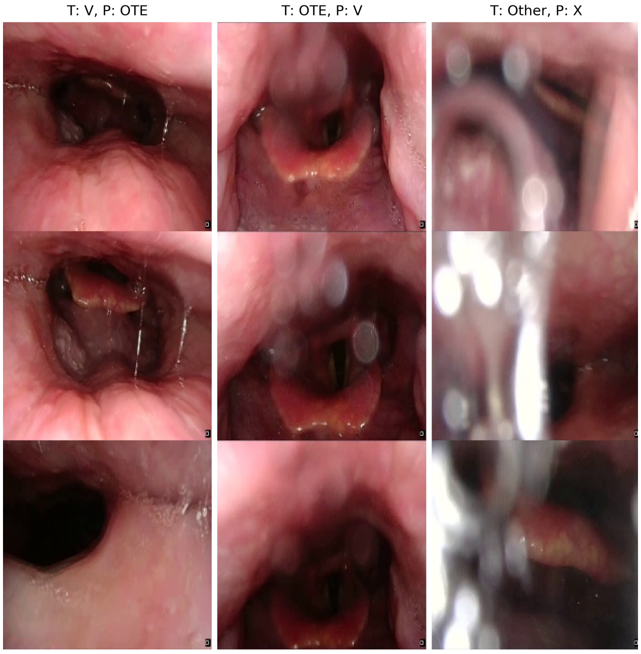


Fig. 4. Examples of misclassifications for each class, i.e. velum (V), oropharynx, tongue or epiglottis (OTE), and distorted video (X), where T is the true class and P is the predicted class.

TABLE III

PERFORMANCE FOR THE THREE VIDEOS IN THE TEST SET FOR CLASSIFYING REGIONS IN THE UPPER AIRWAY WITH THREE DIFFERENT CLASSES: VELUM (V), OROPHARYNX, TONGUE OR EPIGLOTTIS (OTE), AND DISTORTED VIDEO (X).

Video	Accuracy	F1	Class	Class F1	N Frames
1	93%	93%	V	94%	386
			OTE	91%	264
			X	-	0
2	86%	75%	V	67%	166
			OTE	93%	741
			X	65%	93
3	63%	62%	V	61%	252
			OTE	54%	123
			X	70%	125

it would be a trivial task to recognize this class even for a person unfamiliar with DISE videos. This is also reflected by the fact that even with the limited number of frames with class X in the dataset (Table II), the model was easily able to learn to recognize this class. Looking at Fig. 3, it is noted that the sensitivity is 100%, meaning that none of the frames labeled X are misclassified. However, both the V and OTE classes are occasionally misclassified as X, which Fig. 4 shows examples of. It is noted that the model classifies a frame as X any time there is mucus or saliva on the endoscope even if some structures are still visible to some degree. When annotating the data, a frame was only labeled as X if there was no way to estimate the region based on the video or context from previous frames, while the model has learned the relation that any mucus or saliva on the endoscope equals a classification of X.

The model also performed well for the OTE class, reflected by a high sensitivity and F1 score. It is noted from Fig. 3 that when OTE is misclassified, it is mostly as X, which is again explained by the fact that the model is sensitive to mucus and saliva on the camera, even if it is possible to derive the UA region. Scenarios where OTE is misclassified as V is illustrated in Fig. 4, where it is observed that this occurs when the endoscope is at the border between the V and OTE regions. Even experts analyzing these frames could have scored them as V instead of OTE, and it appears that the last frame at the bottom has been wrongly annotated as OTE even though the endoscope is in the V region.

For the V class, the model did not perform as well as for the two other classes. Fig. 3 shows that the misclassifications are almost equally split between OTE and X. The frames misclassified as X are due to the same reason as for OTE. Examples of V being misclassified as OTE are shown in Fig. 4. In this case it appears that the misclassifications do not necessarily occur when the endoscope is close to the OTE region, but rather when the OTE region is visible from the V region so that the model can recognize structures such as the tongue and epiglottis. It makes sense that with the limited amount of data the model has seen, it is not able to derive distance-based decisions to estimate the region as well as it recognizes structures associated with a given region. Furthermore, the large amount of noisy frames in the video (approximately 25%) most likely causes noise in the context of the bidirectional LSTM, which contributes to the poor performance for video 3.

Table III outlines performance for each individual video in the test set. It is observed that the best performance is obtained for video 1, which has no frames with distorted video and also few misclassifications for the V and OTE regions. During video 2, the endoscope is by far the most in the OTE region and the F1-score is high for that class. The F1-score for both V and X is modest because V is misclassified as both OTE and X, whereas OTE is misclassified a few times as X as well. During video 3, most time is spent in the V region but the F1 score is modest for all classes. In this case, V is still misclassified as both OTE and X, but OTE is also sometimes misclassified as V and not only X, which is most likely due to wrong annotations, similar to the bottom frame in the middle column of Fig. 4, where V is labeled as OTE.

There are two main limitations of this study: the quantity of data is extremely low, and the problem posed is simplistic with respect to utilizing this in clinical practice. However, the results serve as an important proof of concept, which shows that it is possible to apply deep learning techniques on DISE videos, even though they depict large variations in terms of both anatomical structure and angles/positions in the UA across videos. Considering this, it is quite impressive that the proposed model obtains such a high performance on so little data and that it actually manages to learn meaningful mappings between the classes and the series of frames that are used as input. For a future study, we will obtain a much larger quantity of data (1000 videos) and expand the problem for classification of where the UA collapses, how it collapses, and what the degree of collapse is.

VI. CONCLUSION

This study shows potential for large-scale learning on DISE videos in order to automatically recognize regions in the UA and thereby derive where the collapse occurs during OSA events, which is critical before any potential surgery to treat OSA. The study was performed on a very limited dataset and serves as a proof of concept for a future study, where a larger quantity of data will be utilized and several variables will be predicted. The presented method has potential application for use in clinical medicine to identify UA collapse.

REFERENCES

[1] P. Lévy, M. Kohler, W. T. McNicholas, F. Barbé, R. D. McEvoy, V. K. Somers, L. Lavie, and J. Pépin, Obstructive sleep apnoea syndrome, Nature Reviews Disease Primers, vol. 1, no. 1, pp. 1-21, 2015.

[2] A. M. Osman, S. G. Carter, J. C. Carberry, and D. J. Eckert, Obstructive sleep apnea: current perspectives, Nature and Science of Sleep, vol. 10, pp. 21–34, 2018.

[3] R. W. W. Lee, K. Sutherland, and P. A. Cistulli, Craniofacial Morphology in Obstructive Sleep Apnea: A Review, Clinical Pulmonary Medicine, vol. 17, no. 4, pp. 189–195, 2010.

[4] K. K. Green et al., Drug-Induced Sleep Endoscopy and Surgical Outcomes: A Multicenter Cohort Study, Laryngoscope, vol 129, pp. 761–770, 2019.

[5] E. J. Kezirian, W. Hohenhorst, and N. de Vries, Drug-induced sleep endoscopy: the VOTE classification, European Archives of Oto-Rhino-Laryngology, vol. 268, pp. 1233–1236, 2011.

[6] W. Hohenhorst, M. J. L. Ravesloot, E. J. Kezirian, and N. de Vries, Operative Techniques in Otolaryngology, vol. 23, no. 1, pp. 3-10, 2012.

[7] E. J. Kezirian, D. P. White, A. Malhotra, W. Ma, C. E. McCulloch, and A. N. Goldberg, Interrater Reliability of Drug-Induced Sleep Endoscopy, Archives of Otolaryngology - Head & Neck Surgery, vol. 136, no. 4, pp. 393-397, 2010.

[8] E. K. Kiaer, P. Tonnesen, H. B. Sorensen, N. Rubek, A. Hammering, C. Moller, A.M. Hildebrandt - P.J. Jennum - C. von Buchwald, Propofol sedation in Drug Induced Sedation Endoscopy without an anaesthesiologist – a study of safety and feasibility, Rhinology, vol. 57, no. 2, pp. 125-131, 2019.

[9] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, Deep Residual Learning for Image Recognition, The IEEE Conference on Computer Vision and Pattern Recognition, Nevada, 2016, pp. 770-778.

[10] M. Schuster and K. K. Paliwal, Bidirectional Recurrent Neural Networks, IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673-2681, 1997.

[11] D. P. Kingma and J. L. Ba, Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, San Diego, 2015.

Appendix E

Paper IV

Title: Automatic Scoring of Drug-Induced Sleep Endoscopy for Obstructive Sleep Apnea Using Deep Learning

Authors: Umaer Hanif, Eva K. Kiær, Robson Capasso, Stanley Y. Liu, Emmanuel J. M. Mignot, Helge B. D. Sorensen, and Poul Jennum

Journal: JAMA Otolaryngology - Head and Neck Surgery

Status: Under review

Title: Automatic Scoring of Drug-Induced Sleep Endoscopy for Obstructive Sleep Apnea Using Deep Learning

Umaer Hanif, Msc.Eng.[a,d,e]

Eva Kirkegaard Kiaer, Ph.D.[b]

Robson Capasso, Associate Professor[c]

Stanley Y. Liu, Associate Professor[c]

Emmanuel J. M. Mignot, Professor[d] *

Helge B. D. Sorensen, Associate Professor[a] *

Poul Jennum, Professor[e] *

[a]Biomedical Signal Processing & AI Research Group, Department of Health Technology, Technical University of Denmark, Oersteds Plads 345B, 2800 Kongens Lyngby, Denmark (work was performed here)

Email: umaerhanif@hotmail.com, hbds@dtu.dk

[b]Danish Center for Sleep Surgery, Department of Otorhinolaryngology, Head and Neck Surgery and Audiology, Copenhagen University Hospital (Rigshospitalet), Inge Lehmanns Vej 8, 2100 Copenhagen, Denmark

Email: eva.kirkegaard.kiaer.01@regionh.dk

[c]Department of Otolaryngology/Head & Neck Surgery, Stanford University School of Medicine, 801 Welch Road, Palo Alto, CA, 94304

Email: ycliu@stanford.edu, rcapasso@stanford.edu

[d]Stanford University Center for Sleep and Circadian Sciences, Stanford University, 3165 Porter Dr., CA 94304, Palo Alto, U.S.A. (work was performed here)

Email: mignot@stanford.edu

[e]Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, University of Copenhagen, Nordre Ringvej 57, 2600 Glostrup, Denmark

Email: poul.joergen.jennum@regionh.dk

* Shared last authors

Name and address of corresponding author:

Poul Jennum

Danish Center for Sleep Medicine, Department of Clinical Neurophysiology

Nordre Ringvej 57, 2600 Glostrup, Denmark

poul.joergen.jennum@regionh.dk

Text word count: 3000

Abstract word count: 346

Figure count: 3

Table count: 2

Reference count: 53

Supplementary Figure Count: 3

Supplementary Table Count: 2

Key points:

Question: Can drug-induced sleep endoscopy examinations in obstructive sleep apnea patients be scored automatically based on deep learning with respect to site of upper airway collapse and obstruction degree?

Findings: Mean F1 score across all upper airway sites with respect to obstruction degree was 70% for 281 drug-induced sleep endoscopy videos and the proposed model generalized well across videos obtained from different clinicians and hospitals.

Meaning: Otolaryngology surgeons can benefit from an automatic scoring tool that provides objective and data-driven estimations which may result in better surgical treatment for obstructive sleep apnea patients.

Abstract

Importance: Scoring drug-induced sleep endoscopy in obstructive sleep apnea patients for site of upper airway collapse can be difficult and presents low to moderate interrater reliability among otolaryngology surgeons. Development of an automatic scoring tool can provide surgeons with objective and data-driven estimations of site of collapse which may reduce uncertainty and result in better surgical treatment.

Objective: To develop a deep learning-based prediction model for estimating site of upper airway collapse and obstruction degree automatically from drug-induced sleep endoscopy videos.

Design: A diagnostic/prognostic study was conducted and included drug-induced sleep endoscopy videos with varying durations (6 seconds – 16 minutes). Videos were split into 5-second clips, each receiving annotations for obstruction degree (0, 1, 2, or X if site is not visible) for each site (velum, oropharynx, tongue, and epiglottis), which was used to train a deep learning model. Predicted obstruction degrees per examination was obtained by taking the highest predicted degree per site across 5-second clips, which was evaluated against obstruction degrees annotated by surgeons.

Setting: Drug-induced sleep endoscopy videos were obtained from two sleep clinics: Copenhagen University Hospital and Stanford University Hospital. These videos were collected by three different otolaryngology surgeons, one at Copenhagen University Hospital and two at Stanford University Hospital.

Participants: 281 drug-induced sleep endoscopy videos were included and all examinations were performed for people with confirmed obstructive sleep apnea to evaluate where and how the upper airway collapses prior to potential surgical intervention.

Main Outcomes and Measures: Ability to estimate site of upper airway collapse and obstruction degree assessed through F1 score, Cohen’s Kappa, sensitivity, and confusion matrices.

Results: Mean F1 score of 70% was obtained across all videos (velum: 85%, oropharynx: 72%, tongue: 57%, epiglottis: 65%). For each site, sensitivity was highest for degree 2 and lowest for degree 0. No bias in performance was observed between videos from different clinicians/hospitals.

Conclusions and Relevance: This study demonstrates that automating scoring of drug-induced sleep endoscopy videos show high validity and feasibility in site and degree of upper airway collapse. Surgeons can benefit from an automatic scoring system in clinical practice but additional study is required to validate the proposed model in a multi-scored dataset.

1. Introduction

Obstructive sleep apnea (OSA) is characterized by partial or complete obstruction of the upper airway during sleep, causing events with reduced airflow (hypopneas) or cessation of breathing (apneas) [1]. Presence of repeated apneas and hypopneas cause disturbances in sleep leading to daytime sleepiness [2–4], increased risk of cardiovascular diseases [5–7], motor vehicle accidents [8], and elevated mortality rates [9]. Prevalence of OSA is high; almost half a billion adults worldwide aged 30-69 years suffer from moderate to severe OSA [10]. The economic burden of undiagnosed or untreated OSA is $150 billion in the United States alone [11].

Key contributors to OSA pathogenesis are narrow upper airway anatomy, low arousal threshold, inability to recruit dilator muscles during inspiration, and poor central control of breathing [12]. Obesity is the most frequent cause of upper airway narrowing due to presence of excess fat tissue in the tongue and around the neck area [13]. Other factors include enlarged tonsils, excess tissue in the soft palate or tongue base, the tongue falling backwards, and an underdeveloped or protracted jaw [14,15]. Continuous positive airway pressure (CPAP) [16] is the gold-standard treatment for OSA and works by providing a constant level of pressure sufficient to keep the upper airway open. Although CPAP is extremely effective, studies show that up to 50% of users give up on the device within a year of therapy because of intolerance, noise, discomfort, or a negative impact on intimacy [17]. Oral appliances may reduce upper airway obstructions by advancing the mandible or refraining the tongue and epiglottis from falling back, but, as they are generally less effective than CPAP, they are considered the second line of treatment after CPAP [18].

For some patients, surgical procedures can be viable options to increase upper airway space, with the most common surgery being a modified uvulopalatopharyngoplasty, where excess tissue is removed from the soft palate and lateral walls of the pharynx, often combined with tonsillectomy (removal of the palatine tonsils) [19,20]. Other procedures include TORS (transoral robotic surgery) on the tongue base and epiglottis [21] and maxillomandibular advancement (advancement of the upper and lower jaw) [22]. Prior to surgery, drug-induced sleep endoscopy (DISE) is often performed to examine the location and pattern of sleep-related upper airway collapse using a fiberoptic endoscope under sedation, which is designed to simulate natural sleep [23]. The endoscope is introduced through the nasal cavity and examines the upper airway from the nares to the level of the glottis. After a DISE examination, the surgeon evaluates the sites of collapse in the upper airway according to the VOTE (velum, oropharynx lateral walls, tongue base, epiglottis) classification system, the most used scoring system [24]. Collapse can occur at these four sites, either individually or in combination, causing obstruction in the upper airway. The VOTE classification system assigns a degree of obstruction and pattern of collapse to each site where a collapse occurs. VOTE obstruction degrees are classified either as 0 (no obstruction), 1 (partial <50% obstruction), or 2 (complete >50% obstruction) [24]. Additionally, there are three patterns of collapse: antero-posterior (A-P) collapse, lateral collapse, and concentric collapse [24]. The upper airway can collapse in any of these patterns at V [25], but only lateral collapse can occur at O, only A-P collapse at T, and only lateral and A-P at E as outlined in Table 1. Lateral collapse at E is extremely rare in practice [26]. Figure 1 shows examples of V and Figure 2 shows examples of O, T, and E with respect to obstruction degrees and collapse patterns.

DISE suggests location and indication for surgical intervention and its use has been shown to improve OSA surgical outcomes [27]. The analysis however depends on the procedure and the evaluation hereof and as such presents interrater variability. First, there is an anatomical variation across subjects with respect to the upper airway and the pattern of collapse may also be affected by the depth of sedation [28–32]. Secondly, DISE videos can appear chaotic due to several sites collapsing simultaneously in patients with severe OSA, essentially pushing the endoscope around and making it difficult to determine the sites where collapses are occurring. Mucus or saliva may also cover the endoscope, which can reduce or distort the video quality significantly and at times making it almost impossible to visually inspect the upper airway. Furthermore, studies examining interscorer reliability between surgeons show poor to moderate agreement [27,33–37], demonstrating that despite a well-established classification system, interpretation remains subjective.

In this study, we hypothesize that a deep learning-based model predicting VOTE obstruction degrees from DISE videos automatically could be trained and would generalize across subjects and centers when validated on a large amount of DISE videos (+10 hours of footage). Such a model could aid surgeons in the scoring of DISE videos and consequentially in the planning of surgical treatment. Deep learning techniques are chosen over other machine learning models because they allow for automatic and data-driven feature extraction from video frames through convolutional neural networks and context-based predictions through long short-term memory networks. In a former smaller study, we evaluated the use of deep learning on DISE examinations for estimation of upper airway regions with promising results [38]. In this current study, we collect a much larger number of videos and design a model capable of predicting obstruction degree at each of the four different sites (VOTE), which is evaluated against gold-standard annotations provided by surgeons. To simplify the problem, we leave out pattern of collapse, which can affect treatment strategy at the level of the velum and lateral pharyngeal wall [39–42].

2. Methods

A. Data Description

281 DISE videos were obtained in total from three different surgeons at two different locations: one surgeon from Copenhagen University Hospital (CUH) in Denmark (51 videos) and two surgeons at Stanford University Hospital (SUH) in California, USA (58 and 172 videos, respectively). The DISE examinations were performed for subjects with confirmed OSA in accordance with DISE procedure guidelines described by Kiaer et al. [43] and Lan et al. [44].

Median duration of videos after anonymization was 2.1 minutes with an interquartile range of 3.33 minutes (min – max: 6 seconds – 16.4 minutes) and the total amount of video footage was 13.7 hours. eFigure 1 in the Supplement shows distribution of DISE examination durations in the dataset. For each examination, an annotation was obtained containing the VOTE score, i.e., obstruction degree and collapse pattern at each site as shown in eTable 1 in the Supplement. The distribution of obstruction degrees for each site is shown in eFigure 2 in the Supplement. The institutions IRB approved the study under IRB-64418.

B. Pre-processing

Since DISE examinations varied greatly with respect to duration (eFigure 1 in the Supplement) and there was only a one-line annotation per video (eTable 1 in the Supplement), we decided that using data as it was would be unsuitable for deep learning purposes. Consequently, all video examinations were split into 5-second clips, as shown in the top block of eFigure 3, and each clip received an annotation with respect to each site. These annotations were created in consultation with a chief surgeon in otorhinolaryngology at CUH (EKK).

Using 5-second clips, there are many scenarios where one or more sites are not visible. Thus, another class was introduced for such situations, denoted X, such that there were four classes (0, 1, 2, X) for each site (VOTE) per 5-second clip, essentially amounting to a 16-class classification problem. The idea was to train and validate the proposed deep learning model on the 5-second clips and then summarize all predictions to form a single VOTE obstruction degree prediction for each DISE examination, which could then be evaluated against the surgeons' annotations.

C. Deep Learning Architecture

The proposed architecture is a combination of a convolutional neural network (CNN) with Resnet18 architecture [45] and a bidirectional long short-term memory (Bi-LSTM) network [46] as illustrated in the middle block of eFigure 3 in the Supplement. The CNN is implemented for automatic feature extraction from each video frame, while Bi-LSTM layers are included to include temporal context in both forward and backward directions for each frame. This architecture is identical to the one presented in earlier work [38] except for two differences: 1) the middle time step of the second Bi-LSTM output is taken instead of using all time steps, and 2) there are 4 output probabilities instead for 3. The proposed model was trained, validated, and tested using 10-fold cross-validation to get predictions for all DISE videos in the dataset.

D. Post-processing

Post-processing steps are illustrated in the bottom blocks of eFigure 3. For each site in each 5-second clip, the model predicts probabilities for each of the four different classes (0, 1, 2, and X). The predicted degree for each site is the one which the model predicted the highest probability for. After a prediction is made for each site for each 5-second clip, the overall degree for each site for a particular DISE examination is calculated as the maximum predicted degree across all 5-second clips constituting a single DISE examination.

E. Performance Evaluation

The predicted obstruction degree for each site for each DISE examination was compared to the surgeons' annotations, considered as ground truth. Performance is evaluated using weighted F1 score [47]. Weighted F1 score is used instead of accuracy due to a large imbalance between the obstruction degrees. The F1 score is calculated as the harmonic mean of precision and recall or alternatively in terms of true positives (TP), false positives (FP), and false negatives (FN):

$$F1\ score = \frac{TP}{TP + \frac{1}{2}\ (FP + FN)}$$

Cohen's kappa [48] is also used to compare model performance with interrater agreement reported in the literature with respect to degrees, either for each site or overall.

3. Results

Overall F1 score for the 12-class problem, i.e., predicting obstruction degree (0, 1, or 2) for each of the four upper airway sites across 281 DISE videos was 70% (V: 85%, O: 72%, T: 57%, E: 65%). Figure 3 shows confusion matrices for the model's predicted degree for each site evaluated against the surgeons' annotations across all DISE videos in the dataset.

Performance is also evaluated with respect to videos obtained from the three different surgeons (one from RH and two from SUH) to investigate any biases in the model towards videos from a particular surgeon. The results are summarized in eTable 2 in the Supplement.

When distinguishing between no obstruction/obstruction, i.e., combining obstruction degrees of 1 and 2 as one class, the F1 score increased to 90% (V: 98%, O: 95%, T: 78%, E: 91%).

Finally, Table 2 compares performance of the model (in terms of Cohen's kappa) to interrater reliabilities reported in the literature between surgeons. The comparisons are made on three levels: for each site (VOTE), region-based (palate and hypopharynx), and overall. For region-based comparison, V was compared to the palate, and OTE were combined for hypopharynx comparisons.

4. Discussion

Here we describe for the first time that DL can be used to reliably evaluate DISE videos with the goal of identifying site and extent of obstruction.

B. Sensitivity for each obstruction degree

The model predicts especially well whether or not there is obstruction as shown by an F1 score reaching 90% (V: 98%, O: 95%, T: 78%, E: 91%) when combining degrees 1 and 2, showing that in general, the model confuses degrees 1 and 2 more often than 0 and 1 or 0 and 2. Figure 3 shows that the highest sensitivity for degree 0 is obtained for E (55%), and that most misclassifications occur because the model predicts degree 1. In very few cases (<10%), the model predicts 2 and after inspecting the three examinations in question, it occurs for two reasons: 1) E is reflected in saliva causing a mirror image where it looks like E is collapsing, which in fact it is not, and 2) the model confuses an A-P V collapse for an E collapse, particularly when the endoscope is close to the collapse. In these examinations, however, the predicted probability for degree 2 is never higher than 40% and it only occurs in 1-2 clips per examination.

The second highest sensitivity for degree 0 is obtained for T (35%), but it is confused for both degrees 1 and 2. When degree 2 is predicted, it occurs for two reasons: the uvula or lower part of the soft palate resembles the tongue and when it collapses, the model confuses it for the tongue collapsing, and 2) when V or O are collapsing and the endoscope is extremely close to the tissue, it resembles the tissue of the tongue.

The lowest sensitivity for degree 0 is obtained for V (17%), although there are only 6 out of 281 DISE examinations where V has a degree of 0. Presence of collapse at the level of V is extremely common among OSA patients [49] which is also evident in the dataset by the lack of videos where V has a degree of 0. However, it is encouraging that the model only confuses degree 0 with degree 1 and never predicts degree 2 in those cases.

The next lowest sensitivity for degree 0 is obtained for O (27%) and the confusion is equally split between degrees 1 and 2. Again, there are only a small number of DISE examinations where O has degree 0 (11 examinations), but the cases in which they are predicted as 2 are due to three reasons: 1) A lateral collapse at the level of E which is generally not considered part of O, (2) T collapsing completely and the endoscope being very close such that the model mistakes it for a collapse at O, and 3) V collapsing and the model mistakenly predicting a contribution of O as well, which can be difficult to assess even for surgeons [50].

The highest sensitivity for degree 1 is obtained for both V and E (64% and 63%, respectively), while T and O are lower (50% and 47%, respectively). For all four sites, the model primarily confuses it with degree 2 and very rarely with degree 0 (none for V, <5% for O, <10% for T and E). For degree 2, the sensitivity is very high for V, O, and T (91%, 93%, and 85%, respectively), while the sensitivity for E is lower (72%). For all four sites, the model almost exclusively confuses degree 2 with degree 1 and almost never with degree 0 (none for V and E, <5% for O and T). Although the model confuses degree 1 with degree 2 to some extent for all sites, the model confuses degree 2 with degree 1 only for E, showing that for this site, the model appears to have most difficulty distinguishing between degrees 1 and 2.

C. Performance for videos from each surgeon eTable 2 shows that there is no noticeable difference in overall F1 score between videos obtained from each of the three surgeons, demonstrating that there is no meaningful bias towards any of them and suggesting that the procedures are comparable. For videos from surgeon 1 (S1) from RH, the model yields the highest F1 score for V out of all three but also the lowest F1 score for T and E. For videos obtained from surgeon 2 (S2) from SUH, the model has the highest F1 score for T out of all three and the lowest F1 score for O. However, the gap between the highest and lowest F1 score for S2 is much smaller than for S1 and in general the discrepancy in performance between sites is lower for videos from S2. For videos obtained from surgeon 3 (S3) from SUH, the model has the highest F1 scores for both O and E compared to S1 and S2 and the lowest F1 score for V. Again, the gap between highest and lowest F1 score is much smaller than for S1.

G. Limitations

There are two main limitations of this study: 1) the model is not able to predict the pattern of collapse, which would need to be added for the model to produce complete VOTE annotations as the surgeons do, 2) no healthy controls are used in the study, but the model could benefit from seeing more examples of absence of collapse, particularly for V and O. For the first limitation, the absence of collapse pattern only really affects predictions for V, since O and T only have one possible collapse pattern and lateral obstructions for E are extremely rare [26]. However, the difference in collapse patterns for V (particularly concentric vs A-P or lateral) can lead to different treatment strategies at the level of the velum and lateral pharyngeal wall and is therefore important in clinical practice [51–53]. The second limitation is difficult to compensate for because DISE examinations are only performed for people with confirmed OSA as the whole point of the procedure is to identify sites contributing to upper airway collapse prior to surgery. However, such data could be gathered by utilizing DISE under sedation associated with other medical procedures. There are three important factors which explain very low performance for some examinations: 1) The duration of an examination, since a very short examination consists of only few 5-second clips and even a few misclassified clips reduce performance by a lot, 2) the video quality, since some examinations have extremely low quality, which makes it difficult to assess the degree for each site, and 3) several sites collapsing simultaneously, causing the endoscope to be pushed back and forth and making the video chaotic to interpret.

5. Conclusion

This study presents the first ever model for predicting sites and obstruction degrees of upper airway collapse from DISE examinations using a dedicated deep learning model. The model produces solid performance with an overall F1 score of 70% and predicts obstruction degrees for the velum and oropharynx well but displays moderate performance for the tongue base and epiglottis. The main limitation is that it does not predict the pattern of collapse, which can affect treatment strategy at the level of the velum and lateral pharyngeal wall. The proposed model has potential to aid surgeons in interpreting DISE examinations in an automated manner but needs further validation on a multi-scored dataset and the added ability to predict collapse pattern.

References

1. Lévy P, Kohler M, McNicholas WT, et al. Obstructive Sleep Apnoea Syndrome. *Nature Reviews Disease Primers*. 2015;1(1):1-21. doi:https://doi.org/10.1038/nrdp.2015.43
2. Johns MW. A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale. *Sleep*. 1991;14(6):540-545. doi:10.1093/SLEEP/14.6.540
3. Gabryelska A, Białasiewicz P. Association between excessive daytime sleepiness, REM phenotype and severity of obstructive sleep apnea. *Scientific Reports 2020 10:1*. 2020;10(1):1-6. doi:10.1038/s41598-019-56478-9
4. Léger D, Stepnowsky C. The economic and societal burden of excessive daytime sleepiness in patients with obstructive sleep apnea. *Sleep Medicine Reviews*. 2020;51. doi:10.1016/j.smrv.2020.101275
5. Baguet JP, Barone-Rochette G, Tamisier R, Levy P, Pépin JL. Mechanisms of cardiac dysfunction in obstructive sleep apnea. *Nature Reviews Cardiology 2012 9:12*. 2012;9(12):679-688. doi:10.1038/nrcardio.2012.141
6. Gonzaga C, Bertolami A, Bertolami M, Amodeo C, Calhoun D. Obstructive sleep apnea, hypertension and cardiovascular diseases. *Journal of Human Hypertension*. 2015;29(12):705-712. doi:10.1038/jhh.2015.15
7. McEvoy RD, Antic NA, Heeley E, et al. CPAP for Prevention of Cardiovascular Events in Obstructive Sleep Apnea. *New England Journal of Medicine*. 2016;375(10):919-931. doi:10.1056/NEJMOA1606599/SUPPL_FILE/NEJMOA1606599_DISCLOSURES.PDF
8. Tregear S, Reston J, Schoelles K, Phillips B. Obstructive Sleep Apnea and Risk of Motor Vehicle Crash: Systematic Review and Meta-Analysis. *Journal of Clinical Sleep Medicine*. 2009;5(6):573-581. doi:10.5664/JCSM.27662
9. Xie C, Zhu R, Tian Y, Wang K. Association of obstructive sleep apnoea with the risk of vascular outcomes and all-cause mortality: a meta-analysis. *BMJ Open*. 2017;7(12):e013983. doi:10.1136/BMJOPEN-2016-013983
10. Benjafield A v., Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *The Lancet Respiratory Medicine*. 2019;7(8):687-698. doi:10.1016/S2213-2600(19)30198-5
11. Watson NF. Health Care Savings: The Economic Value of Diagnostic and Therapeutic Care for Obstructive Sleep Apnea. *Journal of Clinical Sleep Medicine*. 2016;12(8):1075-1077. doi:10.5664/JCSM.6034
12. Osman AM, Carter SG, Carberry JC, Eckert DJ. Obstructive sleep apnea: current perspectives. *Nature and Science of Sleep*. 2018;10:21-34. doi:https://doi.org/10.2147/NSS.S124657
13. Schwartz AR, Patil SP, Laffan AM, Polotsky V, Schneider H, Smith PL. Obesity and Obstructive Sleep Apnea: Pathogenic Mechanisms and Therapeutic Approaches. *Proc Am Thorac Soc*. 2008;5(2):185-192. doi:https://doi.org/10.1513/pats.200708-137MG
14. Dempsey JA, Veasey SC, Morgan BJ, O’donnell CP. Pathophysiology of Sleep Apnea. *Physiological Reviews*. 2010;90(1):47-112. doi:https://doi.org/10.1152/physrev.00043.2008

15. Lee RWW, Sutherland K, Cistulli PA. Craniofacial Morphology in Obstructive Sleep Apnea: A Review. *Clinical Pulmonary Medicine*. 2010;17(4):189-195. doi:https://doi.org/10.1097/CPM.0b013e3181e4bea7
16. Giles T, Lasserson T, Smith B, White J, Wright J, Cates C. Continuous positive airways pressure for obstructive sleep apnoea in adults. *Cochrane Database of Systematic Reviews*. 2006;(1). doi:10.1002/14651858.CD001106.PUB2/MEDIA/CDSR/CD001106/REL0002/CD001106/IMAGE_N/NCD001106-CMP-005-08.PNG
17. J K, K T, K S, et al. Interventions for the Treatment of Obstructive Sleep Apnea in Adults: A Health Technology Assessment. *Canadian Agency for Drugs and Technologies in Health*. Published online 2019. Accessed March 27, 2022. http://europepmc.org/books/NBK535532
18. Dieltjens M, Vanderveken OM. Oral Appliances in Obstructive Sleep Apnea. *Healthcare*. 2019;7(4):141. doi:10.3390/HEALTHCARE7040141
19. Stuck B, Eschenhagen T, Sommer U. Uvulopalatopharyngoplasty with or without tonsillectomy in the treatment of adult obstructive sleep apnea–A systematic review. *Elsevier*. 2018;50:152-165. Accessed April 10, 2022. https://www.sciencedirect.com/science/article/pii/S1389945718301813?casa_token=rzBW1gJ4k3YAAAAA:GfbVVDFDX07YEBJxM1syF5d1EeAV89mswR1I-PhyAiixUHHvSyU8IcqM3SNz5BUh3v0xTLDGgQ
20. Holmlund T, Franklin KA, Levring Jäghagen E, et al. Tonsillectomy in adults with obstructive sleep apnea. *Laryngoscope*. 2016;126(12):2859-2862. doi:10.1002/LARY.26038
21. Vauterin T, Garas G, Arora A. Transoral robotic surgery for obstructive sleep apnoea-hypopnoea syndrome. *ORL*. 2018;80(3-4):134-147. doi:10.1159/000489465
22. Zaghi S, Holty JEC, Certal V, et al. Maxillomandibular Advancement for Treatment of Obstructive Sleep Apnea: A Meta-analysis. *JAMA Otolaryngology–Head & Neck Surgery*. 2016;142(1):58-66. doi:10.1001/JAMAOTO.2015.2678
23. Hohenhorst W, Ravesloot MJL, Kezirian EJ, de Vries N. Drug-induced sleep endoscopy in adults with sleep-disordered breathing: Technique and the VOTE Classification system. *Operative Techniques in Otolaryngology-Head and Neck Surgery*. 2012;23(1):11-18. doi:10.1016/J.OTOT.2011.06.001
24. Kezirian EJ, Hohenhorst W, de Vries N. Drug-induced sleep endoscopy: The VOTE classification. *European Archives of Oto-Rhino-Laryngology*. 2011;268(8):1233-1236. doi:10.1007/S00405-011-1633-8
25. Perck E van de, Heiser C, Vanderveken OM. Concentric versus anteroposterior-laterolateral collapse of the soft palate in patients with obstructive sleep apnea. *ERJ Open Research*. 2022;166(4):782-785. doi:10.1183/23120541.SLEEPANDBREATHING-2021.71
26. Torre C, Camacho M, Liu SYC, Huon LK, Capasso R. Epiglottis collapse in adult obstructive sleep apnea: A systematic review. *Laryngoscope*. 2016;126(2):515-523. doi:10.1002/LARY.25589

27. Green KK, Kent DT, D'Agostino MA, et al. Drug-Induced Sleep Endoscopy and Surgical Outcomes: A Multicenter Cohort Study. *Laryngoscope*. 2019;129(3):761-770. doi:10.1002/LARY.27655

28. Kotlarek KJ, Haenssler AE, Hildebrand KE, Perry JL. Morphological variation of the velum in children and adults using magnetic resonance imaging. *Imaging Science in Dentistry*. 2019;49(2):153. doi:10.5624/ISD.2019.49.2.153

29. Gao F, Li YR, Xu W, et al. Upper airway morphological changes in obstructive sleep apnoea: effect of age on pharyngeal anatomy. *The Journal of Laryngology & Otology*. 2020;134(4):354-361. doi:10.1017/S0022215120000766

30. Diwakar R, Singh Kochhar A, Gupta H, et al. Effect of Craniofacial Morphology on Pharyngeal Airway Volume Measured Using Cone-Beam Computed Tomography (CBCT)—A Retrospective Pilot Study. *International Journal of Environmental Research and Public Health*. 2021;18(9):5040. doi:10.3390/ijerph18095040

31. Ma MA, Kumar R, Macey PM, Yan-Go FL, Harper RM. Epiglottis cross-sectional area and oropharyngeal airway length in male and female obstructive sleep apnea patients. *Nature and Science of Sleep*. 2016;8:297. doi:10.2147/NSS.S113709

32. Zhou N, Ho JPTF, Klop C, et al. Intra-individual variation of upper airway measurements based on computed tomography. *PLOS ONE*. 2021;16(11):e0259739. doi:10.1371/JOURNAL.PONE.0259739

33. Vroegop AVMT, Vanderveken OM, Wouters K, et al. Observer Variation in Drug-Induced Sleep Endoscopy: Experienced Versus Nonexperienced Ear, Nose, and Throat Surgeons. *Sleep*. 2013;36(6):947-953. doi:10.5665/SLEEP.2732

34. Carrasco-Llatas M, Zerpa-Zerpa V, Dalmau-Galofre J. Reliability of drug-induced sedation endoscopy: interobserver agreement. *Sleep and Breathing*. 2017;21(1):173-179. doi:10.1007/S11325-016-1426-9

35. Kezirian EJ, White DP, Malhotra A, Ma W, McCulloch CE, Goldberg AN. Interrater Reliability of Drug-Induced Sleep Endoscopy. *Archives of Otolaryngology–Head & Neck Surgery*. 2010;136(4):393-397. doi:10.1001/ARCHOTO.2010.26

36. Koo SK, Lee SH, Koh TK, et al. Inter-rater reliability between experienced and inexperienced otolaryngologists using Koo's drug-induced sleep endoscopy classification system. *European Archives of Oto-Rhino-Laryngology*. 2019;276(5):1525-1531. doi:10.1007/S00405-019-05386-9

37. Gillespie MB, Reddy RP, White DR, Discolo CM, Overdyk FJ, Nguyen SA. A trial of drug-induced sleep endoscopy in the surgical management of sleep-disordered breathing. *Laryngoscope*. 2013;123(1):277-282. doi:10.1002/LARY.23506

38. Hanif U, Kezirian E, Kiar EK, Mignot E, Sorensen HBD, Jennum P. Upper Airway Classification in Sleep Endoscopy Examinations using Convolutional Recurrent Neural Networks. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Published online 2021:3957-3960. doi:10.1109/EMBC46164.2021.9630098

39. Liu SYC, Huon LK, Iwasaki T, et al. Efficacy of Maxillomandibular Advancement Examined with Drug-Induced Sleep Endoscopy and Computational Fluid Dynamics Airflow Modeling. *Otolaryngology - Head and Neck Surgery (United States)*. 2016;154(1):189-195. doi:10.1177/0194599815611603

40. Liu S, Huon L, Powell N, Riley R, … HCJ of O and, 2015 undefined. Lateral pharyngeal wall tension after maxillomandibular advancement for obstructive sleep apnea is a marker for surgical success: observations from drug. *Elsevier*. Accessed April 17, 2022. https://www.sciencedirect.com/science/article/pii/S0278239115001007?casa_token=WThWJ_-CNEAAAAAA:s3X_Leng7K1_NbWVL1nM3ZUkyoCeYLIwwNT32KTDIc0TkO0BYKRoxuGWxQpS5rdHs9Jk8APk

41. Liu S, Awad M, Riley R, clinics RCS medicine, 2019 undefined. The role of the revised stanford protocol in today's precision medicine. *sleep.theclinics.com*. Accessed April 17, 2022. https://www.sleep.theclinics.com/article/S1556-407X(18)30093-6/abstract

42. Liu S, Riley R, … AP… S, 2019 undefined. Sleep surgery in the era of precision medicine. *oralmaxsurgeryatlas.theclinics.com*. Accessed April 17, 2022. https://www.oralmaxsurgeryatlas.theclinics.com/article/S1061-3315(18)30207-5/abstract

43. Kiær EK, Tønnesen P, Sørensen HB, et al. Propofol sedation in Drug Induced Sedation Endoscopy without an anaesthesiologist - a study of safety and feasibility. *Rhinology*. 2019;57(2):125-131. doi:10.4193/RHIN18.066

44. Lan MC, Liu SYC, Lan MY, Modi R, Capasso R. Lateral pharyngeal wall collapse associated with hypoxemia in obstructive sleep apnea. *Laryngoscope*. 2015;125(10):2408-2412. doi:10.1002/LARY.25126

45. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *CVPR 2016*. Published online 2016:770-778. doi:https://doi.org/10.1109/CVPR.2016.90

46. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*. 1997;45(11):2673-2681. doi:10.1109/78.650093

47. Tharwat A. Classification assessment methods. *Applied Computing and Informatics*. Published online 2020. doi:10.1016/j.aci.2018.08.003

48. McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica*. 2012;22(3):276. doi:10.11613/bm.2012.031

49. Vroegop A v., Vanderveken OM, Boudewyns AN, et al. Drug-induced sleep endoscopy in sleep-disordered breathing: Report on 1,249 cases. *Laryngoscope*. 2014;124(3):797-802. doi:10.1002/LARY.24479

50. Soares D, Sinawe H, Folbe AJ, et al. Lateral Oropharyngeal Wall and Supraglottic Airway Collapse Associated With Failure in Sleep Apnea Surgery. *Laryngoscope*. 2012;122(2):473-479. doi:10.1002/LARY.22474

51. Vroegop A v., Vanderveken OM, Verbraecken JA. Drug-Induced Sleep Endoscopy: Evaluation of a Selection Tool for Treatment Modalities for Obstructive Sleep Apnea. *Respiration*. 2020;99(5):451-457. doi:10.1159/000505584

52. Susan K S, Ankur S, Omprakash C, Payal G. Management Concentric Collapse of Velopharynx in Obstructive Sleep Apnoea Using a Modified Barbed Palato-Pharyngoplasty Technique. *Journal of Sleep Disorders and Management*. 2020;6(1). doi:10.23937/2572-4053.1510028

53. Liu SYC, Hutz MJ, Poomkonsarn S, Chang CP, Awad M, Capasso R. Palatopharyngoplasty Resolves Concentric Collapse in Patients Ineligible for Upper Airway Stimulation. *Laryngoscope*. 2020;130(12):E958-E962. doi:10.1002/LARY.28595

Figures

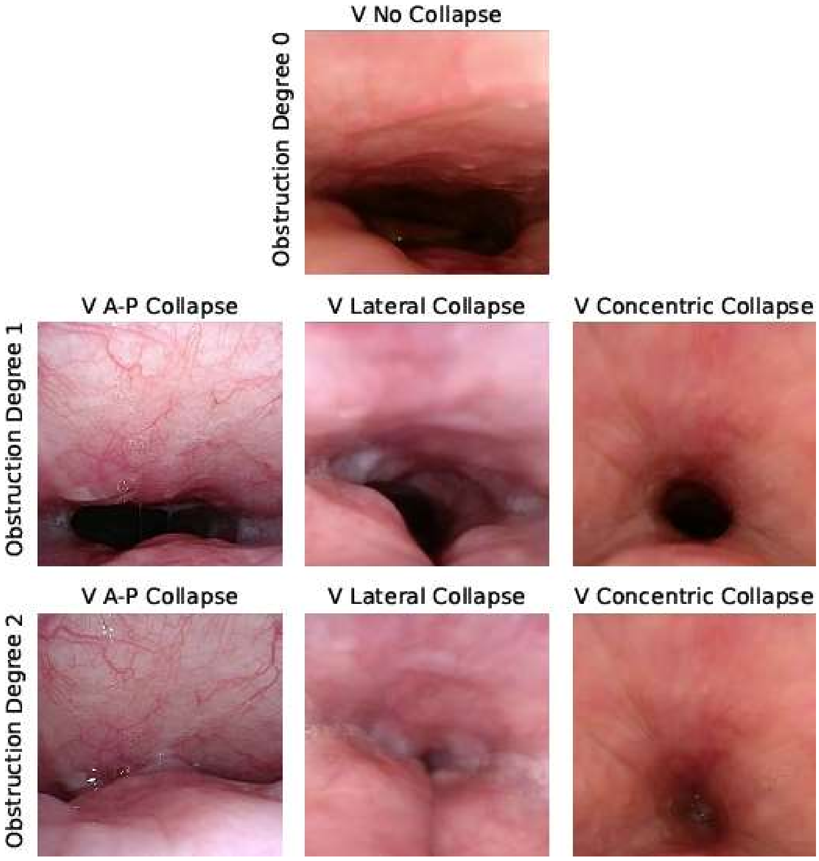


Figure 1. Examples of possible ways the velum (V) can collapse in the upper airway. The collapse can be either partial (obstruction degree 1) or complete (obstruction degree 2). The patterns of collapse are antero-posterior (A-P), lateral, or concentric.

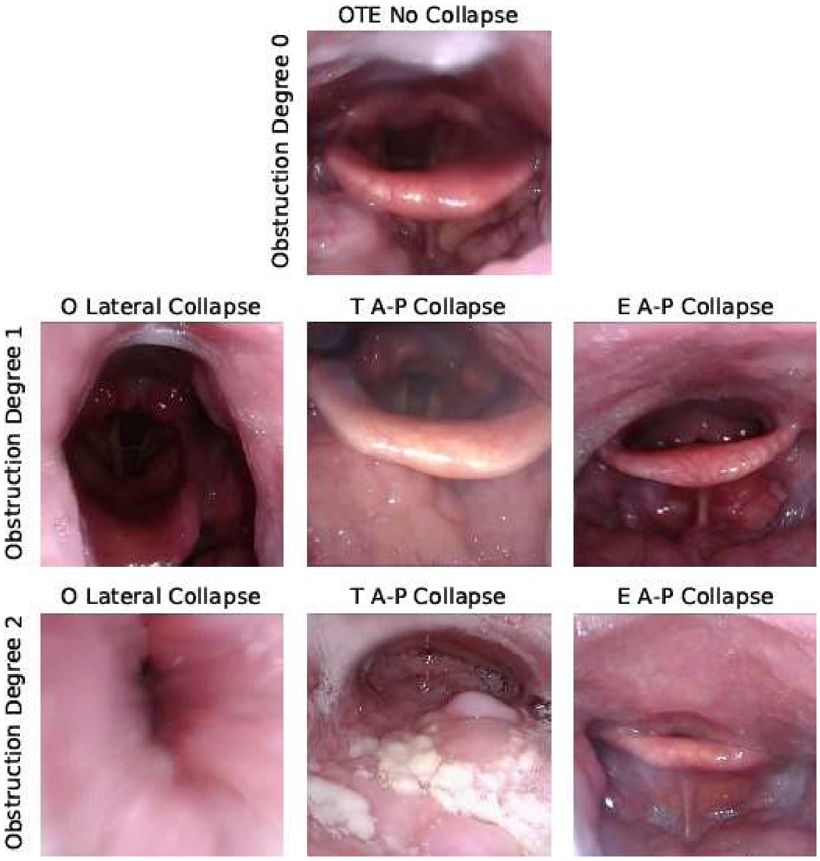


Figure 2. Examples of possible ways the oropharynx (O), tongue base (T), and epiglottis (E) can collapse in the upper airway. The collapse can be either partial (obstruction degree 1) or complete (obstruction degree 2). The pattern of collapse is lateral for O, antero-posterior (A-P) for T, and A-P and lateral for E. Lateral collapse for E has been left out, since it is extremely rare.

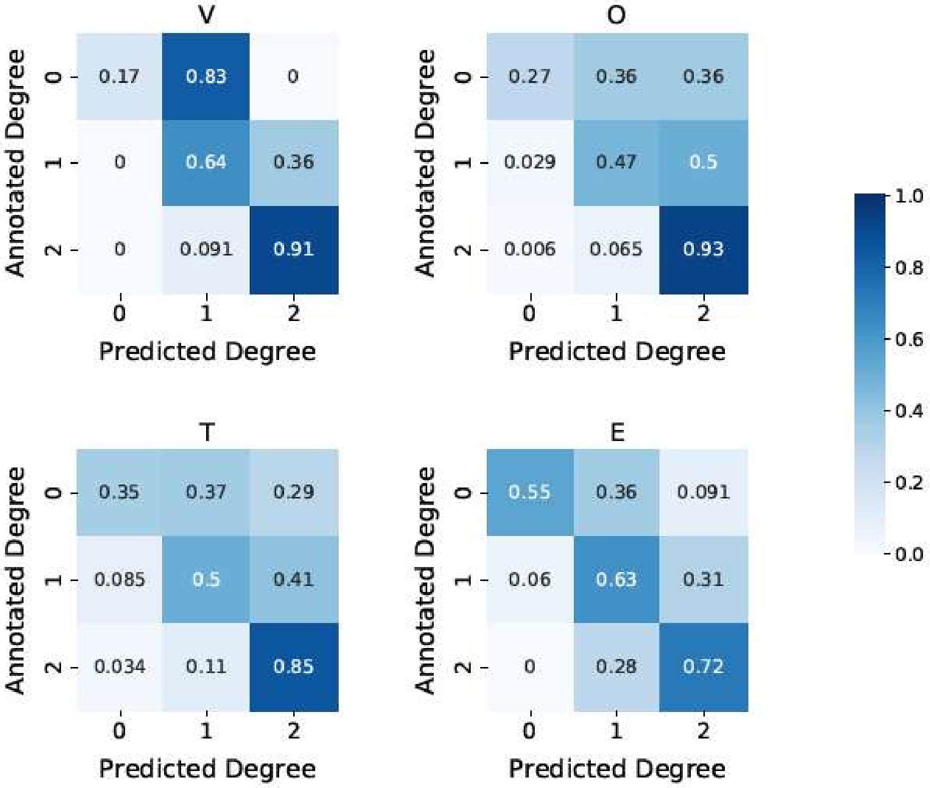


Figure 3. Confusion matrices for the predicted obstruction degrees. Evaluation is done for 281 drug-induced sleep endoscopy videos with respect to the four different upper airway sites. V – velum, O – Oropharynx lateral walls, T – Tongue base, E – Epiglottis.

Tables

Table 1. Upper airway sites where collapse can occur during sleep according to the VOTE classification system. The degree of obstruction caused by collapse is either 0 (no collapse), 1 (partial <50% obstruction), or 2 (complete >50% obstruction). Checkmarks indicate the possible pattern of collapse at each site.

Site	Degree of obstruction	Pattern of collapse		
		Antero-posterior	**Lateral**	**Concentric**
Velum	0, 1, or 2	✓	✓	✓
Oropharynx			✓	
Tongue base		✓		
Epiglottis		✓	✓	

Table 2. Comparison of model performance in terms of Cohen's kappa (κ) to interscorer reliabilities calculated in studies from the literature. Some studies use palate vs hypopharynx, where palate corresponds to V and hypopharynx corresponds to O, T, and E combined. One study reports κ overall across all sites.

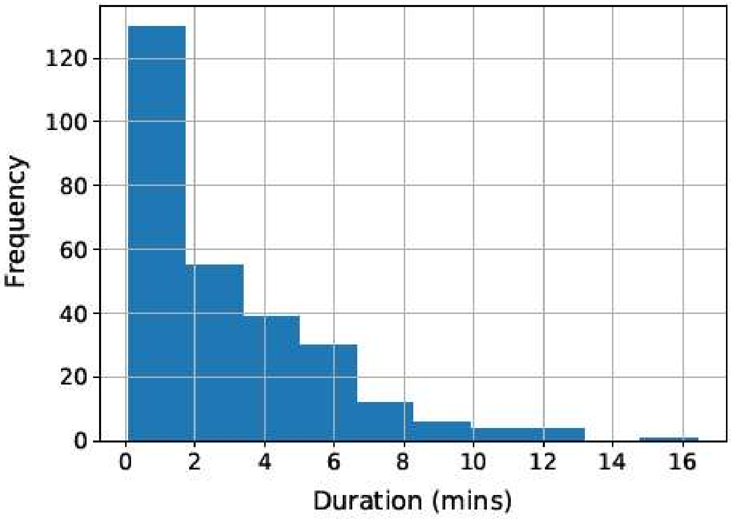
Study	V (κ)	O (κ)	T (κ)	E (κ)	N scorers	N DISE
Our model	***0.55***	***0.45***	***0.38***	***0.44***	***N/A***	***281***
Vroegop et al. [33]	0.30	0.66	0.03	0.61	7	6
Llatas et al. [34]	0.17	0.67	0.35	0.43	2	31
Green et al. [27]	0.40	0.42	0.60	0.55	4	275
Study	**Palate (κ)**	**Hypopharynx (κ)**			**N scorers**	**N DISE**
Our model	***0.55***	***0.43***			***N/A***	***281***
Kezirian et al. 10 [35]	0.60	0.44			2	108
Koo et al. [36]	0.52	0.35			6	100
Study	**Overall (κ)**				**N scorers**	**N DISE**
Our model	***0.46***				***N/A***	***281***
Gillespie et al. [37]	0.27				3	38

Supplementary Material

2. Methods

A. Data Description

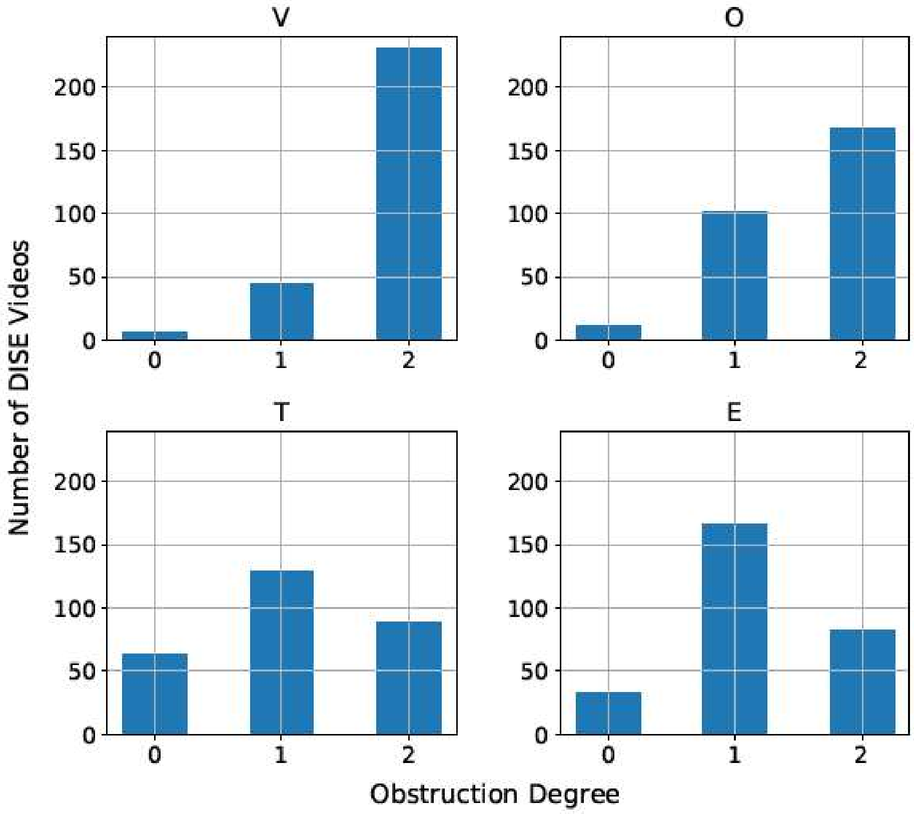
Each video was anonymized by removing any part where the endoscope was outside of the patient and renaming the video file. eFigure 1 shows distribution of DISE examination durations in the dataset after anonymization, showing that most videos in the dataset are less than 2 minutes long. Videos obtained from CUH had sampling rates of 25 frames per second, while videos from SUH had sampling rates of 30 frames per second. For each examination, an annotation was obtained containing the VOTE score, i.e., obstruction degree and collapse pattern at each site as shown in eTable 1. Note that several sites can collapse in the same subject (sometimes even in combination) and that a site like V can collapse in more than one way in the same subject. The distribution of obstruction degrees for each site is shown in eFigure 2.



eFigure 1. Distribution of durations of drug-induced sleep endoscopy videos in the dataset.

eTable 1. Examples of annotations provided by surgeons for drug-induced sleep endoscopy examinations. The number (0, 1, or 2) for each upper airway site (VOTE) indicates the obstruction degree followed by the collapse pattern (A-P, lateral or concentric). V – velum, O – oropharyngeal lateral wall, T – tongue base, E – epiglottis, A-P – antero-posterior.

Video	V	O	T	E
Video 1	2 A-P - Concentric	2 Lateral	1 A-P	2 Lateral
Video 2	2 A-P	2 Lateral	0	0
Video 3	1 A-P	0	2 A-P	2 A-P



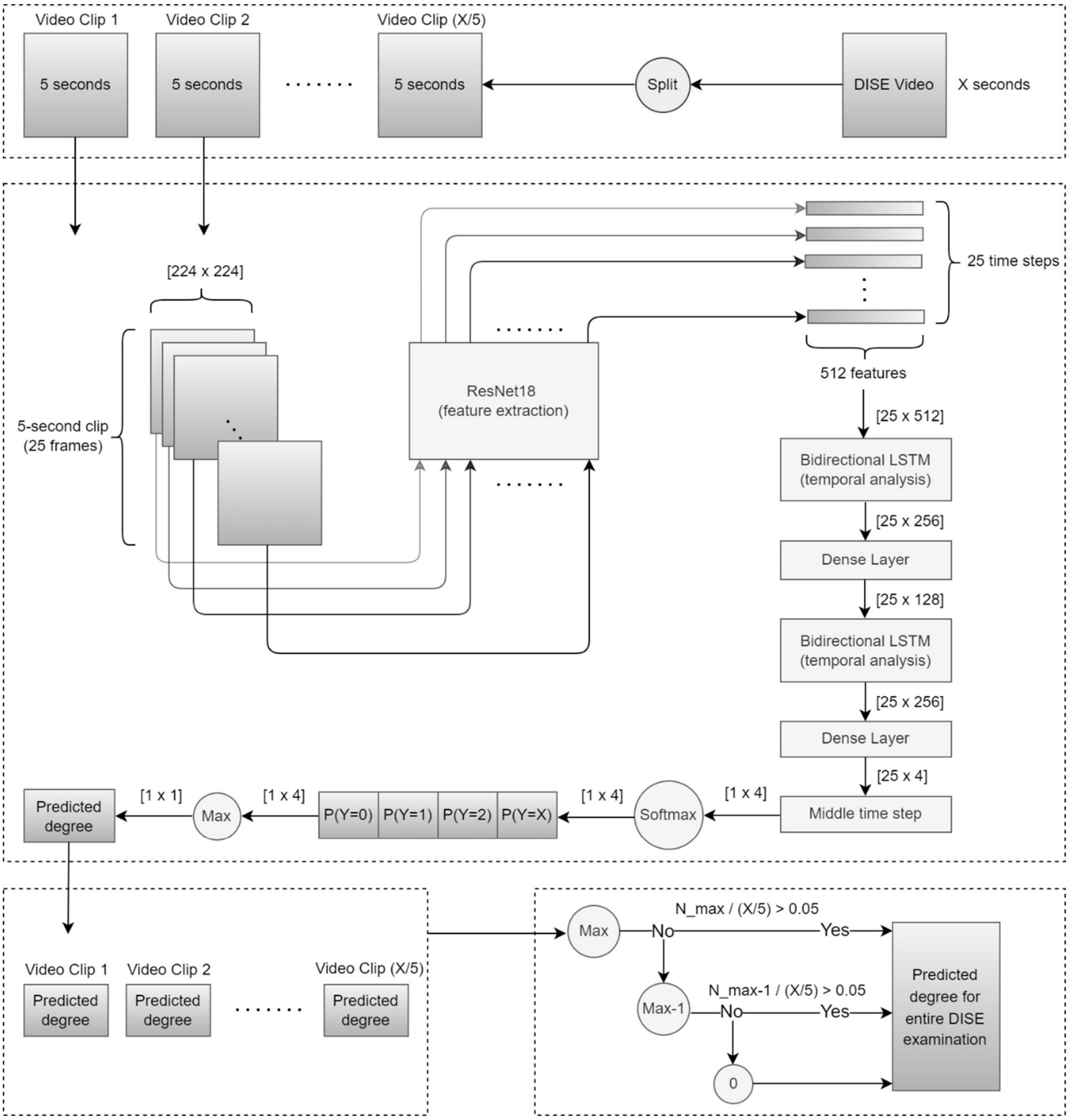
eFigure 2. Distribution of obstruction degrees (0, 1, and 2) for each of the four upper airway sites (VOTE). DISE – drug-induced sleep endoscopy, V – velum, O – oropharynx lateral walls, T – tongue base, E – epiglottis.

B. Pre-processing

Although sampling rates for videos from RH and SUH were 25 and 30 frames per second, respectively, we used only every 5th and 6th frame, respectively, to reduce computational cost. Consequently, the sampling rates for 5-second clips used in the study were 5 frames per second, yielding a total of 25 frames for a 5-second clip. Instead of using all three color-channels for each clip (R,G,B), the videos were converted to grayscale. Both approaches were investigated (with and without color channels) and preserving colors did not make any noticeable difference, most likely because anatomical composition is much more important than small differences in color, so grayscale frames were used to reduce computational cost.

C. Deep Learning Architecture

The Resnet18 network is implemented such that a 5-second clip (consisting of 25 frames) could be input one frame at a time as shown in eFigure 3. The output is then a feature map of size 1x512 for each frame. The feature maps for all frames are concatenated to form a 25x512 matrix, where each row is considered a time-step in the original 5-second clip and each column is a feature vector for a particular frame. This matrix is processed by a bidirectional LSTM layer, followed by a dense layer, which reduces the number of features from 512 to 128 while time steps are intact, i.e., resulting matrix dimensions are 25x128. A second bidirectional LSTM and dense layer reduces the number of features further from 128 to 4, yielding a matrix of dimension 25x4. From this matrix, the output at the middle time step, i.e., 13 is taken as it represents the time step where the model has most context in both directions. Finally, a softmax activation function is applied to the resulting 1x4 vector to yield a probability for each class, i.e., 0, 1, 2, and X. This architecture is repeated four times, one for each site with their own loss function.



eFigure 3. Architecture of proposed model for predicting obstruction degree based on 5-second clips from drug-induced sleep endoscopy (DISE) examinations. This architecture is repeated for each of the four upper airway sites (velum, oropharynx lateral walls, tongue base, epiglottis). Top block: A DISE examination is split into 5-second clips. Middle block: Each individual frame (grayscale) of a 5-second clip is used as input one by one for a convolutional neural network (CNN) with a ResNet18 architecture for feature extraction. All resulting feature vectors (1x512) are concatenated (25x512) and input to a bidirectional long short-term memory network (Bi-LSTM) for temporal analysis, followed by a dense layer to reduce number of

features (25x128). Another Bi-LSTM and dense layer reduce the feature vector (1x4), which is run through a softmax activation function. This yields four probabilities, one for each obstruction degree (P(Y=0), P(Y=1), P(Y=2), and P(Y=X), where X means that the site is not visible). The obstruction degree with highest probability is the predicted obstruction degree. Bottom left block: Predictions for all clips within a DISE examination are collected. Bottom right block: The maximum predicted obstruction degree across all 5-second clips that make up a full examination is chosen as the overall degree if the model predicts this degree for at least 5% of all clips. Otherwise, same criterion is checked for the next highest degree and if not fulfilled either, the predicted obstruction degree is 0 by default.

D. Training, Validation, and Testing

The proposed model was trained, validated, and tested using 10-fold cross-validation to get predictions for all DISE videos in the dataset. The loss function for each site is the cross-entropy loss, which is the loss function of choice in multi-classification settings. The loss functions for all four sites are added together and the combined loss function was used for optimizing the weights of the model using the Adam optimizer. The combined loss function is used to optimize the model simultaneously with respect to all four sites. Since the degrees for each site are imbalanced, penalty weights were introduced during training. The weight for a particular class is calculated by dividing the number of samples for the most represented class with the number of samples for a particular class in the training set. The model was trained using batch sizes of 8 and the learning rate was set to $1 \cdot 10^{-5}$ with a weight decay of $5 \cdot 10^{-4}$. Early stopping was applied when the validation error stopped decreasing for 3 consecutive epochs, i.e., a patience of 3, to avoid overfitting to the training data. Python 3.6.10 and Pytorch 1.10.0 were used for implementation of the proposed model. Training of the model was performed using a GeForce RTX 3070 and the entire training/validation/test setup took approximately 16 hours to run.

E. Post-processing

The maximum degree is selected only if this degree is predicted in at least 5% of the clips which make up a full examination. This is to avoid any coincidences where a degree of e.g., 2 occurs one time by chance or because of other sites and does not reflect the true behavior of that site in a subject. In case the maximum degree does not satisfy this condition, the next greatest degree is selected if it satisfies the same condition. If this is not satisfied either, the degree is 0 by default. This is illustrated in the bottom right box of eFigure 3. A voting approach is not applied here, because surgeons annotate DISE examinations according to the highest degree observed.

3. Results

Performance is also evaluated with respect to videos obtained from the three different surgeons (one from RH and two from SUH) to investigate any biases in the model towards videos from a particular surgeon. The results are summarized in eTable 2.

eTable 2. F1 score for drug-induced sleep endoscopy examinations with respect to each of the three surgeons who provided the videos. Performance is evaluated against the surgeons' annotations with respect to obstruction degree (0, 1, or 2) for four different upper airway sites. V – velum, O – oropharynx lateral walls, T – tongue base, E – epiglottis, CUH – Copenhagen University Hospital, SUH – Stanford University Hospital.

Surgeon	N Videos	V (F1)	O (F1)	T (F1)	E (F1)	Overall (F1)
S1 (CUH)	51	91%	70%	53%	58%	68%
S2 (SUH)	58	89%	64%	63%	63%	70%
S3 (SUH)	172	82%	74%	56%	67%	70%