



Διδακτορική Διατριβή

**Μέθοδοι Ψηφιακής Επεξεργασίας
Ηχητικού Σήματος για την Ανίχνευση
Παθολογίας στην Ομιλία**

Ιωάννα Μηλιαρέση

Πατρώνυμο: Γεράσιμος

ΑΜ: ΠΛΔ1501

Πανεπιστήμιο Πειραιώς
Σχολή Τεχνολογιών Πληροφορικής και
Επικοινωνιών
Τμήμα Πληροφορικής



Doctoral Thesis

**Digital Audio Processing Methods for
Voice Pathology Detection**

Ioanna Miliaresi

University of Piraeus
School of Information and Communication
Technologies
Department of Informatics

Μέθοδοι Ψηφιακής Επεξεργασίας Ηχητικού Σήματος για την Ανίχνευση Παθολογίας στην Ομιλία

Η διατριβή εκπονήθηκε για την απονομή
Διδακτορικού Διπλώματος
από το Τμήμα Πληροφορικής
της Σχολής Τεχνολογιών Πληροφορικής και Επικοινωνιών
του Πανεπιστημίου Πειραιώς
στην
Ιωάννα Μηλιαρέση

Τριμελής Συμβουλευτική Επιτροπή

Άγγελος Πικράκης (Επιβλέπων) **Ανδρέας Φλώρος** **Μιχάλης Ψαράκης**
Πανεπιστήμιο Πειραιώς Ιόνιο Πανεπιστήμιο Πανεπιστήμιο Πειραιώς

Εγκρίθηκε την
20/01/2025

Άγγελος Πικράκης **Ανδρέας Φλώρος** **Μιχάλης Ψαράκης**
Επίκουρος Καθηγητής Καθηγητής Αναπληρωτής Καθηγητής
Πανεπιστήμιο Πειραιώς Ιόνιο Πανεπιστήμιο Πανεπιστήμιο Πειραιώς

Δημήτριος Απόστολου **Κάτια-Λήδα Κερμανίδου** **Μιχάλης Παναγόπουλος**
Καθηγητής Καθηγήτρια Αναπληρωτής Καθηγητής
Πανεπιστήμιο Πειραιώς Ιόνιο Πανεπιστήμιο Ιόνιο Πανεπιστήμιο

Διονύσιος Σωτηρόπουλος
Επίκουρος Καθηγητής
Πανεπιστήμιο Πειραιώς

Digital Audio Processing Methods for Voice Pathology Classification

This dissertation was submitted for the award of

Doctoral Degree

by the Department of Informatics

of the School of Information and Communication

Technologies

of the University of Piraeus

by

Ioanna Miliarisi

Supervisory Committee

Aggelos Pikrakis (Supervisor)

University of Piraeus

Andreas Floros

Ionian University

Michalis Psarakis

University of Piraeus

Approved on

20/01/2025

Aggelos Pikrakis

Assistant Professor

University of Piraeus

Andreas Floros

Professor

Ionian University

Michalis Psarakis

Associate Professor

University of Piraeus

Dimitrios Apostolou

Professor

University of Piraeus

Katia-Lida Kermanidou

Professor

Ionian University

Michalis Panagopoulos

Associate Professor

Ionian University

Dionysios Sotiropoulos

Assistant Professor

University of Piraeus

Περίληψη

Ο όρος "παθολογία της φωνής" αναφέρεται σε ένα ευρύ φάσμα διαταραχών και ασθενειών που επηρεάζουν την ποιότητα και την παραγωγή της φωνής και κατ' επέκταση της ομιλίας. Η χρήση της μηχανικής μάθησης έχει αναδειχθεί ως μια καινοτόμος προσέγγιση για τη διάγνωση ενός ποικίλου φάσματος φωνητικών διαταραχών. Παρά την εκτενή έρευνα στον τομέα αυτό, εξακολουθεί να υπάρχει ένα σημαντικό κενό στην ανάπτυξη ταξινομητών που διαθέτουν την ικανότητα προσαρμογής και αποτελεσματικής γενίκευσης. Η παρούσα διατριβή αποσκοπεί να καλύψει αυτό το κενό, προτείνοντας νέες προσεγγίσεις και μεθόδους για την αυτόματη ταξινόμηση των φωνητικών παθολογιών.

Η έρευνά μας επικεντρώνεται στην αυτόματη ταξινόμηση φωνητικών παθολογιών μέσω αρχιτεκτονικών νευρωνικών δικτύων, δίνοντας έμφαση σε προκλήσεις όπως η εξαγωγή κατάλληλων χαρακτηριστικών από δεδομένα παθολογίας, η περιορισμένη διαθεσιμότητα δεδομένων και η ενσωμάτωση πολυτροπικών πληροφοριών. Οι στόχοι μας είναι η αύξηση της διαγνωστικής ακρίβειας και η βελτίωση της προσαρμοστικότητας και της ικανότητας γενίκευσης των μοντέλων.

Για την επίτευξη αυξημένων ικανοτήτων γενίκευσης και την ενίσχυση της ευελιξίας του ταξινομητή στην αναγνώριση ποικίλων φωνητικών παθολογιών, η έρευνά μας διερεύνησε σε βάθος διαφορετικά σύνολα δεδομένων και είδη παθολογιών. Καλύφθηκε ένα ευρύ φάσμα φωνητικών διαταραχών, όπως η λειτουργική δυσφωνία, το φωνοτραύμα, τα λαρυγγικά νεοπλάσματα, η μονομερής παράλυση των φωνητικών χορδών. Παράλληλα μελετήθηκαν φωνητικές διαταραχές που σχετίζονται με τον COVID-19 και αναπνευστικού ήχου.

Συγκεκριμένα, αξιοποιήθηκαν διάφορα σύνολα δεδομένων παθολογίας, τα Far Eastern Memorial Hospital Dataset, Saarbruecken Voice Database, Virufy, Coswara, COVID-19 και SPRsound, καθένα εκ των οποίων περιέχει διαφορετικούς τύπους φωνητικών και αναπνευστικών ηχογραφήσεων από ασθενείς με ποικίλες παθολογίες. Σε αυτές τις βάσεις δεδομένων περιλαμβάνονται διάφοροι τύποι φωνητικών και αναπνευστικών ήχων, όπως παρατεταμένα φωνήεντα, ομιλία, βήχας, αναπνοή και ηλεκτρογλωττογραφικά σήματα.

Στα στάδια του σχεδιασμού, της υλοποίησης και της αξιολόγησης των ταξινομητών, η παρούσα έρευνα εστίασε σε κρίσιμα ζητήματα, όπως η επεξεργασία πολυτροπικών ηχητικών δεδομένων, η εξαγωγή κατάλληλων χαρακτηριστικών, η ανάπτυξη προηγμένων αρχιτεκτονικών βαθιάς μάθησης και η εφαρμογή τεχνικών επαύξησης δεδομένων, ειδικά προσαρμοσμένων στα δεδομένα παθολογίας της φωνής.

Ως αποτέλεσμα, αυτή η διατριβή παρουσιάζει πέντε υπολογιστικά μοντέλα, το καθένα σχεδιασμένο για να αντιμετωπίσει συγκεκριμένες προκλήσεις στην ταξινόμηση φωνητικών παθολογιών και στην ανίχνευση του COVID-19:

1. Μοντέλο Ενσωμάτωσης Ιατρικών Δεδομένων: Το μοντέλο αυτό συνδυάζει ιατρικά δεδομένα με ηχογραφήσεις φωνής, ενσωματώνοντάς τα σε μια αρθρωτή αρχιτεκτονική βαθιάς μάθησης. Η αξιοποίηση των κατάλληλων ιατρικών δεικτών με ακουστικά χαρακτηριστικά βελτιώνει την ακρίβεια της ταξινόμησης.

2. Μοντέλο Επεξεργασίας Ήχων Μεταβλητής Διάρκειας με εξειδικευμένες τεχνικές επαύξησης δεδομένων παθολογίας: Το μοντέλο αυτό αξιοποιεί αρχιτεκτονικές πλήρως συνελκτικών νευρωνικών δικτύων που επεξεργάζονται ηχογραφήσεις διαφορετικής διάρκειας. Στην μέθοδο υλοποιούνται καινοτόμες τεχνικές επαύξησης των ηχητικών δεδομένων με αλγόριθμο κατάτμησης των ηχογραφήσεων σε τυχαία μήκη και έγχυση πολλαπλών χρωμάτων θορύβου προκειμένου να αντιμετωπιστεί το πρόβλημα της έλλειψης δεδομένων εκπαίδευσης.

3. Μοντέλο Ενσωμάτωσης Δεδομένων Ηλεκτρογλωττογραφίας: Αυτό το μοντέλο συνδυάζει ηλεκτρογλωττογραφικά σήματα με ηχητικά δεδομένα και ιατρικούς δείκτες σε μια ενιαία τριπολική αρχιτεκτονική βαθιάς μάθησης. Η αξιοποίηση των ηλεκτρογλωττογραφικών ηχογραφήσεων μέσω της πρωτοπόρας αναπαράστασής τους "wavegrams" βελτιώνει την ακρίβεια της ταξινόμησης συγκριτικά με τον διπολικό ταξινομητή των προηγούμενων μοντέλων.

4. Πολυτροπική Αρχιτεκτονική με μηχανισμό προσοχής για την ανίχνευση του Covid-19: Αυτή η πολυδιάστατη αρχιτεκτονική βαθιάς μάθησης επεξεργάζεται εννέα διαφορετικούς τύπους ήχων (αναπνευστικούς ήχους, φωνήματα) και ενσωματώνει ένα μηχανισμό προσοχής, που επιλέγει δυναμικά το καταλληλότερο είδος ήχου για κάθε απόφαση ταξινόμησης. Η προσέγγιση αυτή όχι μόνο βελτιώνει την ακρίβεια της ταξινόμησης, αλλά και ενισχύει την ικανότητα γενίκευσης και την προσαρμοστικότητα του μοντέλου σε διαφορετικά σύνολα δεδομένων, ακόμα και όταν δεν είναι διαθέσιμοι όλοι οι τύποι ηχογραφήσεων.

5. Πλήρως Συνελκτικό Δίκτυο για Ταξινόμηση Αναπνευστικών Ήχων: Αυτό το μοντέλο καινοτομεί με ένα πλήρως συνελκτικό δίκτυο που επιτρέπει την επεξεργασία ηχητικών σημάτων στην αρχική τους τυχαία διάρκεια, χωρίς την ανάγκη κερματισμού σε τμήματα σταθερής διάρκειας και χρήσης τεχνικών συμπλήρωσης με μηδενικά. Αυτή η προσέγγιση καθιστά δυνατή τη βελτιωμένη ανάλυση και την ακριβέστερη ταξινόμηση των αναπνευστικών ήχων, συμβάλλοντας στην αποτελεσματική διάγνωση των σχετικών παθολογιών.

Τα παραπάνω μοντέλα συμβάλλουν με σημαντικές βελτιώσεις στην ακρίβεια της διάγνωσης, την προσαρμοστικότητα και τις δυνατότητες γενίκευσης στον τομέα της παθολογίας της φωνής και της ταξινόμησης αναπνευστικών ήχων. Τα βασικά ευρήματα συνοψίζονται στα εξής:

- Η επεξεργασία ηχογραφήσεων ως δισδιάστατων, μονοκαναλικών εικόνων μέσω συνελκτικών νευρωνικών δικτύων βελτιώνει την απόδοση ταξινόμησης.
- Το αποτελεσματικότερο διάνυσμα χαρακτηριστικών για την ταξινόμηση παθολογιών φωνής περιλαμβάνει τους Mel frequency cepstrum coefficients, τη θεμελιώδη συχνότητα και το jitter και το HNR.
- Η ενσωμάτωση ιατρικών παραμέτρων σε διπολικό ταξινομητή μαζί με ακουστικά χαρακτηριστικά βελτιώνει σημαντικά την ακρίβεια ταξινόμησης παθολογιών φωνής.
- Η συνδυασμένη χρήση δεδομένων ηλεκτρογλωττογραφίας με ηχητικά και ιατρικά δεδομένα σε μια αρθρωτή τριπολική αρχιτεκτονική οδηγεί σε βελτιωμένη ακρίβεια του συστήματος.

- Ένα πλήρως συνελκτικό δίκτυο ικανό να επεξεργάζεται ηχογραφήσεις μεταβλητής διάρκειας προσφέρει υψηλότερη απόδοση ταξινόμησης σε σύγκριση με τα συμβατικά συνελκτικά δίκτυα.
- Η εφαρμογή ενός αλγορίθμου κατατμηματισμού ήχων σε τμήματα μεταβλητής διάρκειας, ανάλογα με την αρχική διάρκεια των ηχογραφήσεων, αποτελεί μια νέα τεχνική ενίσχυσης για την παθολογία της φωνής.
- Η έγχυση ενός συνόλου θορύβου πολλών χρωμάτων αποδεικνύεται κατάλληλη τεχνική επαύξησης φωνητικών δεδομένων παθολογίας.
- Ο συνδυασμός πολλαπλών ακουστικών μορφών (αναπνευστικοί ήχοι, ομιλία) σε μια πολυδιάστατη αρχιτεκτονική με ενσωματωμένο μηχανισμό προσοχής για τον υπολογισμό του βάρους των διαφορετικών ηχητικών πηγών ενισχύει την ακρίβεια και την προσαρμοστικότητα του ταξινομητή, ιδιαίτερα σε περιπτώσεις όπου δεν είναι διαθέσιμες όλες οι τύποι ηχογραφήσεων.
- Στον τομέα της ταξινόμησης αναπνευστικών ήχων, η επεξεργασία των ηχογραφήσεων στην αρχική τους διάρκεια μέσω αρχιτεκτονικής βαθείας μάθησης με πλήρες συνελκτικό δίκτυο βελτιώνει την ακρίβεια ταξινόμησης.

Λίστα Δημοσιεύσεων

1. I. Miliaresi, K. Poutos and A. Pikrakis, "Combining acoustic features and medical data in deep learning networks for voice pathology classification," 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, 2021, pp. 1190-1194, doi: 10.23919/Eusipco47968.2020.9287333.

2. I. Miliaresi and A. Pikrakis, "A Modular Deep Learning Architecture for Voice Pathology Classification," in IEEE Access, vol. 11, pp. 80465-80478, 2023, doi: 10.1109/ACCESS.2023.3300795.

3. I. Miliaresi, A. Pikrakis and K. Poutos, "A Deep Multimodal Voice Pathology Classifier with Electroglottographic Signal Processing Capabilities," 2022 7th International Conference on Frontiers of Signal Processing (ICFSP), Paris, France, 2022, pp. 109-113, doi: 10.1109/ICFSP55781.2022.9924745.

Abstract

Voice pathology is a diverse field that includes various disorders affecting vocal quality and production. Using audio machine learning for voice pathology classification represents an innovative approach to diagnosing a wide range of voice disorders. Despite extensive research in this area, there remains a significant gap in the development of classifiers and their ability to adapt and generalize effectively. This thesis aims to address this gap by contributing new insights and methods.

This research provides a comprehensive exploration of automatic voice pathology classification, focusing on challenges such as data limitations and the potential of integrating multiple modalities to enhance diagnostic accuracy and adaptability.

To achieve generalization capabilities and enhance the flexibility of the classifier across diverse types of voice disorders, this research explores various datasets and pathology types comprehensively. It covers a broad range of voice disorders, including functional dysphonia, phonotrauma, laryngeal neoplasm, unilateral vocal paralysis, and COVID-19-related vocal conditions. The study also includes an analysis of diverse vocal and respiratory sounds to improve classifier adaptability and accuracy further.

This approach involves experimentation with different datasets, including Far Eastern Memorial Hospital, Saarbruecken Voice Database, Virufy, Coswara, COVID-19, and SPRsound datasets, each representing distinct voice and respiratory sounds and pathology types. Additionally, it encompasses diverse vocal and respiratory sounds, such as sustained vowels, speech, cough, breathing, and electroglottographic signals.

Throughout the design, implementation, and evaluation of the classifiers, this research focuses on the feature extraction stage, the design of the deep learning architectures, and the utilization of augmentation techniques tailored to voice pathology data. As a result of this process, this dissertation introduces five computational models, each tailored to address specific challenges in voice pathology classification and COVID-19 detection:

1. **Fusion of Medical Data:** This model integrates medical data with audio recordings within a modular deep learning framework. By combining relevant medical descriptors with acoustic features, it enhances classification accuracy.
2. **Augmentation and Variable-Length Processing:** This architecture employs augmentation techniques such as colored noise injection and variable-length segmentation. These methods enable the model to handle recordings of varying durations, thereby facing the challenge of data scarcity.
3. **Incorporation of Electroglottographic Data:** This model integrates EGG signals with audio data and medical descriptors within a unified deep learning framework, enhancing classification accuracy by leveraging additional physiological information.

4. **Attention-Guided Multimodal Architecture:** Utilizing an attention mechanism, this architecture dynamically selects the most relevant audio modality (respiratory sounds, vowels) for each classification decision. This approach is particularly useful in scenarios where not all types of recordings are available.
5. **Fully Convolutional Network for Respiratory Sound Classification:** This model introduces a fully convolutional network capable of processing audio signals of arbitrary duration without the need for segmentation or padding. It is specifically designed for the classification of respiratory sounds. These models collectively contribute to significant advancements in diagnostic accuracy, adaptability, and generalization capabilities in the field of voice pathology and respiratory sound classification.

The key findings of the thesis can be summarized as follows:

- Processing audio recordings as 2-D, single-channel images through convolutional neural networks yields superior classification performance.
- The most effective audio feature vector for classification of voice disorders combines Mel-Frequency Cepstral Coefficients, fundamental frequency, and perturbation measurements such as jitter and HNR.
- Incorporating medical data and demographic parameters into the voice disorders classification system significantly enhances accuracy.
- Integrating Electroglottographic data into a trimodal architecture with medical and audio parameters leads to improved system accuracy.
- A fully convolutional network capable of handling recordings of arbitrary duration demonstrates higher classification performance compared to conventional convolutional networks.
- Introducing a variable-length segmentation algorithm tailored to the duration of audio recordings represents a novel voice pathology augmentation technique.
- Injecting an ensemble of colored noise proves to be an effective voice data augmentation technique for voice pathology classification.
- Combining multiple audio sounds(respiratory sounds, voice, speech) improves system accuracy for Covid-19 detection.
- Introducing an attention-guided mechanism for modality weighting improves classifier accuracy and adaptability, particularly in datasets where not all types of recordings are available.
- In respiratory sound classification, a deep learning architecture with fully convolutional neural network increases classification accuracy by processing recordings without segmentation.

List of Papers

1. I. Miliaresi, K. Poutos and A. Pikrakis, “Combining acoustic features and medical data in deep learning networks for voice pathology classification,” 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, 2021, pp. 1190-1194, doi:10.23919/Eusipco47968.2020.9287333.

2. I. Miliaresi and A. Pikrakis, “A Modular Deep Learning Architecture for Voice Pathology Classification,” in IEEE Access, vol. 11, pp. 80465-80478, 2023, doi: 10.1109/ACCESS.2023.3300795.

3. I. Miliaresi, A. Pikrakis and K. Poutos, “A Deep Multimodal Voice Pathology Classifier with Electroglottographic Signal Processing Capabilities,” 2022 7th International Conference on Frontiers of Signal Processing (ICFSP), Paris, France, 2022, pp. 109-113, doi: 10.1109/ICFSP55781.2022.9924745.

*“In the middle of every difficulty lies opportunity”
Albert Einstein*

Acknowledgements

Knowledge has always been, and remains, one of the most valuable assets in my life. It has been my purpose to believe in, study, and serve science and education with all my strength throughout my life, and this dissertation is a step toward furthering that purpose.

The motivation for pursuing this PhD has been marked by a series of personal and professional challenges, many of them difficult and even traumatic. This dissertation is a testament to my efforts to transform those experiences into something constructive and positive. As Einstein once said, “In the middle of every difficulty lies opportunity,” a quote that resonates deeply with me as I reflect on this work.

I dedicate this dissertation to my father, who has been my greatest support and inspiration at every step of my life, often at the expense of his own needs. That with his personal testimony to disability was a strong motivation for carrying out the thesis. I am deeply grateful for his existence and for his patience in enduring my absence and lack of assistance, all to allow me the time and focus needed to complete this PhD. I would also like to thank the rest of my family, my mother and especially my sister Aggeliki, for stepping in to cover for me when I could not be present.

To my friends, Tassos, Zeta, and Stella, thank you for providing positive support during critical turning points. A special mention goes to my godson Pantelis and my little godson, Orion, for never having enough time to spend. To my dearest friend Antonis Dilalos, an excellent mathematician, for his dedication to science—he would have been truly happy to read this research. And to all my friends, thank you for listening patiently to my worries and frustrations along the way.

Most importantly, I owe my sincerest gratitude to my supervisor, Assistant professor Aggelos Pikrakis, for the opportunity to work under his guidance, for his time and effort, his scientific insights, and for the invaluable lessons he imparted—each of which has been essential to my academic growth.

I also wish to express my deepest gratitude to my committee members, including Professor Andreas Floros and Associate professor Michael Psarakis, for their willingness, support, scientific insights, and valuable assistance in shaping this research.

I hope that this research has helped me become a better person and teacher, and that it may be of value, even to just one person. ...

Contents

Acknowledgements	xv
Contents	xvii
1 Introduction	1
1.1 Introducing Voice Pathology Classification	1
1.2 Machine Learning for Voice Pathology Classification	3
1.2.1 Data Processing and Feature Extraction Process	4
1.2.2 Machine Learning Approaches for Voice Pathology Classification	6
1.2.3 Machine Learning Approaches for COVID-19 Classification	9
1.3 Research Questions, Aims, and Objectives	11
1.3.1 Aim	12
1.3.2 Objectives	12
1.4 Thesis Contributions and Key Findings	13
1.5 Thesis Outline	18
2 Combining Acoustic and Medical Descriptors into a Modular Deep Learning Architecture for Voice Pathology Classification	21
2.1 Introduction	21
2.1.1 Problem Definition and Application Context	22
2.1.2 Data Resources and Research Gaps	22
2.1.3 Research Objectives and Contributions	24
2.2 Related Work	27
2.2.1 Audio Features in Voice Pathology Classification	27
2.2.2 FEMH Dataset in the Literature	29
2.3 Methodology	33
2.3.1 Description of the FEMH 2019 Dataset	33
2.3.2 Proposed Feature Representations	35
2.3.3 Perturbation-medical Input Vector	37
2.4 Proposed Computational Models	38
2.4.1 CNN-based Modular Deep Learning Architecture	38
2.4.2 Fully-CNN based Modular Deep Learning Architecture	39
2.4.3 Augmentation Methods	41
Data augmentation with Segmentation into adjustable duration splits	42
Data augmentation with noise injection	44
Data augmentation with spectrum masking	44
2.5 Experiments and Results	46
2.5.1 CNN-based computational approach	47

	Ablation study	47
	Experimental Results of the Challenge-Submitted Model	48
2.5.2	Fully CNN based Computational Model	49
	Ablation study	50
	Experimental Results	52
2.6	Conclusions	58
3	Multimodal deep learning classifier with EGG processing capabilities for voice pathology classification	59
3.1	Introduction	59
3.1.1	Data Resources and Research Gaps	60
3.1.2	Research objectives and contributions	61
3.2	Related work	63
3.2.1	Research Contributions in SVD Dataset	63
3.2.2	Research Contributions with EGG signals on SVD dataset	67
3.2.3	Foundational Concepts	68
	Electrolottographic signals, waveform analysis	68
	Introducing EGG based indices	68
	Contact quotient	69
	Wavegram: A Graphical Representation of EGG Waveforms	70
3.3	Methodology	71
3.3.1	Network architecture	73
3.3.2	Feature vector	74
3.3.3	Ablation study	78
3.3.4	Experiments and results	78
3.4	Conclusions	82
4	Attention-guided Modular Deep Learning Architecture for Covid-19 Audio Classification	85
4.0.1	Problem Definition and Application Context	85
4.0.2	Research Objectives and Contributions	87
4.1	Related Work	89
4.1.1	Voice Pathology Detection in the Context of COVID-19	89
4.2	Methodology	91
4.2.1	The Coswara and Virufy datasets	91
4.2.2	Feature Extraction	92
4.2.3	Network Architecture	92
4.3	Experiments and Results	97
4.3.1	Experimental Results	97
4.3.2	Cross-Modal Retrieval Performance Qualitative Analysis	103
4.4	Conclusions	107
5	Enhancements on Respiratory Sound Classification	109
5.1	Introduction	109
5.1.1	Problem Definition and Application Context	110
	Research Objectives and Contributions	111
5.2	Related Work	111
5.3	Methodology	114

5.3.1	Dataset Description	114
5.3.2	Feature Extraction	115
5.3.3	Neural Network Architecture	116
5.4	Experiments and Results	118
5.5	Conclusions	120
6	Conclusions	123
6.0.1	Key findings and contributions	123
	Formulation of deep learning architectures	124
	Formulation of feature vectors	125
	Contribution of Modalities	125
	Data augmentation techniques	125
6.0.2	Network Interpretation and Explainability	126
6.0.3	Performance Analysis and Challenges	127
6.0.4	Future plans	128
A	Background Theory	131
A.1	Deep Learning and Neural Networks	131
A.1.1	Fully Connected Neural Networks	131
A.1.2	Convolutional Neural Networks	135
A.1.3	Fully Convolutional Neural Networks	139
A.2	A primer on Key Audio Features	142
A.2.1	Jitter, Shimmer, Signal-to-Noise and Harmonic-to-noise Ratio	142
	Jitter	143
	Jitter (Local)	145
	Jitter (RAP)	145
	Jitter (PPQ5)	145
	Shimmer	145
	Shimmer (Local)	147
	Shimmer (Local, dB)	147
	Shimmer (APQ3)	147
	Shimmer (APQ5)	147
	Signal-to-Noise Ratio	147
	Harmonic-to-Noise Ratio	148
A.2.2	Mel Frequency Cepstral Coefficients	149
A.2.3	Noise Signals	151
	Gaussian Noise	151
	White Noise	152
	Pink Noise	152
	Brown Noise	152
A.2.4	Time frequency representations: Spectrogram and Mel spectrogram	153
	Bibliography	157

List of Figures

1.1	Voice is primarily classified as either normal or disordered (pathological). A disorder can be caused by inflammation, structural abnormalities, neuromuscular issues, or muscle tension imbalances. Structural abnormalities can include vocal fold cysts, polyps, nodules, paralysis, and sulcus vocalis.	2
1.2	Fundamental steps of voice pathology classification.	4
1.3	Datasets used in the research	6
1.4	Multimodal Classification Frameworks developed for Voice Pathology Classification	14
2.1	Contents of the FEMH dataset	34
2.2	Medical/Perturbation features vector	36
2.3	Audio features vector	36
2.4	Bimodal deep learning CNN-based classifier. The first sub-network consists of four convolutional layers with ReLU activation and max pooling. The second sub-network is a feed-forward neural network with two hidden layers, designed to process the perturbation-medical input vectors.	39
2.5	Bimodal Fully CNN based classifier. the audio features branch implements a fully convolutional network with four convolutional layers (Conv1, Conv2, Conv3, Conv4) followed by a GlobalMaxPooling layer. The perturbation and medical features processing branch consists of two fully connected layers (Dense1, Dense2).	41
2.6	Feature sequence segmentation.	42
2.7	Segmentation flowchart	43
2.8	The audio feature image undergoes augmentation by warping along the time axis and applying masks to multiple blocks of consecutive time steps (vertical masks) and mel frequency channels (horizontal masks). The masked segments of the image are highlighted in purple for emphasis.	45
2.9	Visualization of all weights of the three fully connected layers.	56
2.10	Vocal palsy: input “image” along with three feature activation maps of the final convolutional layer.	56
2.11	Hyperfunctional dysphonia: input “image” along with three feature activation maps of the final convolutional layer.	56
2.12	Phonotrauma: input “image” along with three feature activation maps of the final convolutional layer.	57
2.13	Neoplasm: input “image” along with three feature activation maps of the final convolutional layer.	57
2.14	Perceptually uniform sequential color map inferno	57

3.1	Schematic representation of vocal fold vibratory cycle	70
3.2	Illustration of the wavegram	71
3.3	Subset of the SVD Dataset Employed in Our Experiments	72
3.4	Trimodal classifier with EGG processing capabilities	74
3.5	All visual representations of an example EGG waveform, Fo , CQ values and EGG Wavegram Representation	76
3.6	Example images of alternative EGG representation and indices	77
4.1	Illustration of all five datasets employed in the experiments	91
4.2	Attention-guided modular deep learning architecture. The classifier’s core architecture includes a feature learning stage, an attention-like mechanism, and a final classification module. The feature learning module comprises nine C-net branches, while the attention module is derived from nine D-net branches. The D-net consists of dense layers followed by a softmax activation function, while the C-net contains a block of four convolutional layers with ReLU activation, followed by a global max-pooling layer. Colors represent different input types, with varying shades of gray denoting counting, blue representing breath, yellow signifying cough, and green indicating vowels. These colors enhance visual clarity and aid in information conveyance.	94
4.3	Visualization of attention scores’ evolution across epochs for all five folds	104
4.4	Diagrams showing the nine subnetworks, with the subnetwork having the highest attention score highlighted for each epoch across all five folds.	105
5.1	Fully convolutional network architecture consisting of four blocks of convolutional and max pooling layers, followed by a global max pooling layer, a fully connected layer, and a softmax output. The input has an arbitrary size equal to $N \times 32 \times 1$ and is fed with Mel spectrograms of the audio recordings.	117
A.1	Fully connected neural network	134
A.2	Convolutional neural network	137
A.3	Fully Convolutional Neural Network can make dense predictions for per-pixel tasks like semantic segmentation	140
A.4	Caption for your figure goes here.	150

List of Tables

2.1	References about FEMH	32
2.2	References concerning FEMH dataset with demographic data included	33
2.3	Network configuration description: The output of each layer is fed as input to the subsequent layer. ReLU and dropout layers always follow convolutional and fully connected layers. The parameter N lies in the range of $[124, 1462]$, depending on the duration of the audio recording.	40
2.4	Results for different input data combinations	49
2.5	Results with noise injection	49
2.6	Experiments with Convolutional and Fully Connected Layer Dimensions on Classification Accuracy	51
2.7	Classification accuracy of unimodal and bimodal classifiers	53
2.8	Classification accuracy with respect to different augmentation techniques	54
2.9	Confusion matrix of the best-performing classifier	54
3.1	Summary of Studies Utilizing Conventional Methods on the SVD Dataset	65
3.2	Summary of Studies Utilizing Neural Network Architectures on the SVD Dataset	66
3.3	Classification accuracy for alternative classifiers	80
4.1	Evaluation of alternative audio feature representations for the nine audio modalities	100
4.2	Classification performance of attention-guided mechanism	101
4.3	Performance comparison with top rated methods across various datasets	106
4.4	Performance comparison with top rated Virufy-related works	107
5.1	Summary of the features, algorithms, and scores of the top five rated papers of IEEE BioCAS 2022 Grand challenge on Respiratory Sound Classification	114
5.2	Results for different feature representations on the five-class and three-class tasks.	120
5.3	Results for different feature representations on the five-class and three-class tasks (continued).	120

List of Abbreviations

A-BiGRU: Attention-guided Bidirectional Gated Recurrent Unit
A-BiLSTM: Attention-based Bidirectional Long Short-Term Memory
ACRNN: Attention-guided Convolutional Recurrent Neural Network
ANN: Artificial Neural Network
AUC: Area Under the Curve
BiGRU: Bidirectional Gated Recurrent Unit
BiLSTM: Bidirectional Long Short-Term Memory
Cepstral HNR - Cepstral Harmonics-to-Noise Ratio
CIdeR: Covid-19 Identification ResNet
CIoTVID: Cough Index of Things and Vital Information Database
CIoTVIDTA: Cough Index of Things and Vital Information Database and Treatment Assessment
CNN: Convolutional Neural Network
COUGHVID: Cough Vid
COVID-19: Coronavirus Disease 2019
CRNN: Convolutional Recurrent Neural Network
DCT: Discrete Cosine Transform
DEGG: Differentiated EGG (Differentiated Electroglottographic)
DenseNet: Densely Connected Network
DL: Deep Learning
DWPT : Discrete Wavelet Packet Transform
EGG: Electroglottographic Signals
EPFL: École Polytechnique Fédérale de Lausanne
EL : Extreme Learning
FD: Fractal Dimension
FDA: Fractal Dimension Analysis
FDR: Fisher discriminative ratio
FD: Fractal Dimension
FEMH: Far Eastern Memorial Hospital
FFT: Fast Fourier Transform
GNER Glottal-to-noise excitation ratio
GMM: Gaussian Mixture Model
GRU: Gated Recurrent Unit
GA : Genetic Algorithms
GB : Gradient Boosting
GBT : Gradient Boosted Trees
HPF: High Pass Filter
HMM: Hidden Markov Model
IIR: Infinite Impulse Response
KNN: K-Nearest Neighbors

LNA : Linear Network Analysis
LPC: Linear Predictive Coding
LPCC : Linear Predictive Coding Coefficients
LDA: Linear Discriminant Analysis
LR: Logistic Regression
LSTM: Long Short-Term Memory
MASS: Massachusetts
MC-LDA : Multi-Class Linear Discriminant Analysis
MEEI: Massachusetts Eye and Ear Infirmary
MFCC: Mel Frequency Cepstral Coefficients
MIT: Massachusetts Institute of Technology
ML: Machine Learning
NNE : Nearest Neighbor Estimator
PPQ : Pitch Perturbation Quotient
PCA: Principal Component Analysis
QoV: Quality of Voice
QUCoughscope: Quantitative Cough Scope
ReLU: Rectifier Linear Unit
ResNet: Residual Network
RNN: Recurrent Neural Network
ROC: Receiver Operating Characteristic
RMSE: Root Mean Square Error
SARS-CoV-2: Severe Acute Respiratory Syndrome Coronavirus 2
SC: Spectral Centroid
SB: Spectral Bandwidth
SWCE Sinusoidal Weighted Cepstrum estimator
SVD: Saarbruecken Voice Database
SVM: Support Vector Machine
SVXM : Support Vector Machines with Margins
SWCE : Short-Term Wavelet Coefficients Entropy
SR: Spectral Rolloff
SSL: Semi-Supervised Learning
STFT: Short-Time Fourier Transform
TSVM : Twin Support Vector Machine
Tanh: Hyperbolic tangent function
UOC: University of Cambridge
UAR : Unweighted Average Recall
VGG: Visual Geometry Group
VFCA: Vocal Fold Contact Area
VP: Voice Pathology
VPC: Voice Pathology Classification
ZCR: Zero Crossing Rate

To my Father...

Chapter 1

Introduction

1.1 Introducing Voice Pathology Classification.

Biomedical signals encompass a wide range of data originating from the intricate physiological processes occurring within the human body. These signals play a pivotal role in understanding health dynamics, detecting various medical conditions, and monitoring the functions of different bodily systems. Within this diverse array of signals, the human voice stands out as a remarkable signal that carries a wealth of information beyond words, offering significant potential for both scientific exploration and practical clinical use.

Beyond its primary role in communication, the human voice serves as a complex result of physiological and acoustic characteristics, reflecting the underlying intricacies of the vocal mechanism. The study of voice signals, therefore, extends beyond linguistic analysis to encompass a rich domain of interdisciplinary research, incorporating elements of anatomy, physiology, acoustics, linguistics, and computer science. Moreover, the voice reflects the state of a person's health, with deviations from established norms often serving as early indicators of underlying physiological disturbances or pathological conditions.

By analyzing and interpreting vocal signals, researchers can uncover valuable information on human physiology and pathology, paving the way for innovative diagnostic and therapeutic approaches in clinical practice.

Voice can be categorized as normal or disordered, with disorders often stemming from factors such as inflammation, structural anomalies, neuromuscular issues, or imbalances in muscle tension. Figure 1.1 presents a classification scheme for voice disorders.

The term “voice pathologies” describes a diverse spectrum of conditions, encompassing a wide range of afflictions that affect the vocal apparatus and its functions [Pay+22]. These conditions range from common issues, such as acute laryngitis—an inflammation of the larynx often caused by viral infections or vocal strain—to more severe disorders. Structural abnormalities, including cysts in the vocal folds, polyps, nodules, paralysis, and vocal sulcus, are among the primary causes of voice disorders [MA21]. Among the most prevalent voice pathologies are vocal nodules, which are benign growths that form on the vocal folds due to vocal abuse or misuse.

Paresis, or paralysis of the vocal cords, can result from damage to the nerves controlling the muscles of the larynx, leading to weakness or immobility of the vocal folds. Tumors, both benign and malignant, can develop within the larynx or

other structures of the vocal tract, posing significant health risks. Similarly, polyps—abnormal growths in the vocal folds—can cause disturbances in voice production and quality. Laryngeal cancer represents one of the most serious voice pathologies, with life-threatening implications if not diagnosed and treated promptly.

In addition, dysphonia, encompassing both hyperfunctional and hypofunctional voice disorders, presents challenges in voice production characterized by excessive tension or insufficient closure of the vocal folds, respectively. Other vocal pathologies include adductor spasmodic dysphonia, a neurological disorder characterized by involuntary spasms of the vocal folds, resulting in strained or strangled speech. Vocal fatigue is another pathology that occurs due to overuse or misuse of the voice, leading to decreased vocal endurance and quality. Vocal tremor is characterized by rhythmic fluctuations in pitch or loudness during speech. Vocal fold edema, or swelling of the vocal folds, often occurs due to inflammation or injury. Laryngeal paralysis, a condition in which the muscles of the larynx are partially or completely paralyzed, results in hoarseness, breathlessness, and difficulty swallowing.

This research primarily focuses on structural abnormalities. Moreover, it explores how voice manifestations stemming from pulmonary and neurological diseases can affect vocal function, thereby contributing to voice pathology. Additionally, while investigating typical voice disorders, this research extends its scope to encompass the distinctive realm of COVID-19-related voice symptoms. Although COVID-19 itself is not a direct cause of voice pathology, it presents symptoms that notably influence vocal characteristics. As the pandemic progressed, emerging evidence indicated that COVID-19 infection could impact the respiratory system, leading to changes in breathing patterns, coughs, and alterations in vocal quality.

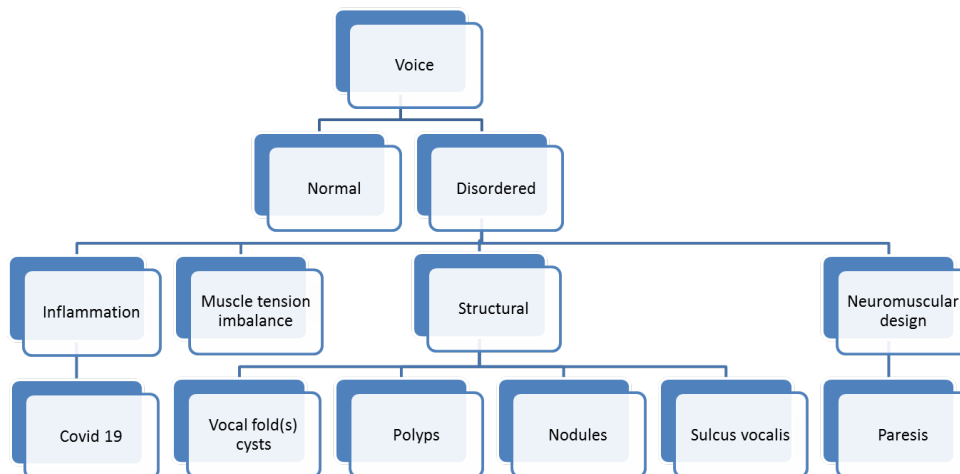


Figure 1.1: Voice is primarily classified as either normal or disordered (pathological). A disorder can be caused by inflammation, structural abnormalities, neuromuscular issues, or muscle tension imbalances. Structural abnormalities can include vocal fold cysts, polyps, nodules, paralysis, and sulcus vocalis.

The timely and precise diagnosis of these pathologies is fundamental for improving patient prognosis and enhancing their quality of life. Traditionally, voice pathology diagnosis has relied on subjective perceptual evaluation by trained clinicians,

supported by instrumental assessments. Common diagnostic methods include:

- **Auditory Perceptual Evaluation:** Clinicians assess voice quality based on auditory impressions such as pitch, loudness, breathiness, and roughness, using standardized rating scales (e.g., the GRBAS scale).
- **Laryngoscopy:** Direct or indirect visualization of the larynx using endoscopy to assess vocal fold morphology, mobility, and the presence of lesions (e.g., laryngeal stroboscopy).
- **Acoustic Analysis:** Measurement of acoustic parameters (e.g., fundamental frequency, jitter, shimmer) using voice analysis software.
- **Objective Assessment Tools:** Quantitative measurements of vocal function, such as aerodynamic assessments (e.g., airflow, air pressure) and electromyography of the laryngeal muscles.

While traditional diagnostic methods provide important clinical data, they are not without limitations. Perceptual evaluation, for instance, may be influenced by inter-listener variability and subjective bias, raising concerns about its reliability. Additionally, laryngoscopic examinations, while effective in visualizing the larynx and assessing vocal fold morphology, can be uncomfortable for patients and require specialized equipment and expertise. Moreover, objective assessment tools, though offering quantifiable data, may lack the sensitivity and specificity needed to detect subtle vocal abnormalities or differentiate between various voice disorders. These limitations underscore the need for more advanced and comprehensive diagnostic approaches in the field of voice pathology.

Furthermore, the increasing demand for voice-related healthcare services amplifies the urgency for expeditious and accurate diagnoses. Thus, there is an imperative to complement traditional diagnostic methodologies with objective, quantitative approaches to enhance diagnostic accuracy, facilitate early diagnosis, and ultimately optimize patient outcomes. Embracing innovative advancements in voice analysis technology holds immense potential to redefine the landscape of voice pathology diagnosis and management, paving the way for personalized, evidence-based healthcare interventions and fostering advancements in the field of biomedical research.

1.2 Machine Learning for Voice Pathology Classification.

The process of voice pathology classification, as shown in Figure 1.2, starts with careful data collection, acquiring audio recordings of speech samples from individuals with normal and pathological voices. Adding metadata, such as age, gender, and diagnosed pathology, is essential for thorough analysis.

Ensuring that the collected data represents the range of real-world voice pathologies is important for generalization. Having enough training data is essential for developing a robust model, and prioritizing data quality is vital to reduce biases and ensure reliable classification outcomes.



Figure 1.2: Fundamental steps of voice pathology classification.

After data collection, preprocessing is performed, including noise reduction, segmentation, and feature extraction. Next, feature selection is used to identify the most important features for classification, employing techniques such as statistical analysis and dimensionality reduction.

Once the features are selected, various machine learning models are considered for classification. The chosen model is then trained using the dataset, with hyperparameters tuned using validation data. The model's performance is evaluated using a separate testing set, and performance metrics such as accuracy, precision, recall, and F1-score are computed. Interpreting and visualizing the results are crucial for understanding the model's performance and identifying areas for improvement, often using confusion matrices, feature importance plots, and activation maps.

This pipeline outlines the machine-driven voice pathology classification process, which we will explore further in the next sections.

1.2.1 Data Processing and Feature Extraction Process

In the field of voice pathology research, several datasets have been collected, providing data on various voice disorders. According to [SRH20a], the most prominent datasets in this domain include the Massachusetts Eye and Ear Infirmary (MEEI) database, the Saarbrücken Voice Database (SVD), and the Arabian Voice Pathology Database (AVPD). Additionally, recent developments have led to the creation of datasets related to COVID-19 patients, such as the MIT COVID-19 dataset, the University of Cambridge COVID-19 Sounds dataset, the Stanford University Virufy dataset, and the EPFL COUGHVID dataset. These datasets focus on COVID-19 cough, breathing, and speech sounds, developed in response to the emerging pandemic.

Among these datasets, selecting a specific one required careful consideration of its relevance, reliability, and alignment with our research objectives. This dissertation uses a variety of datasets to conduct thorough experiments and analyses across different pathology types and incorporate alternative data sources. Four primary datasets

are utilized in this investigation: the Saarbrücken Voice Database, the FEMH 2019 dataset, the Coswara dataset, and the Virufy dataset. Additionally, some experiments were conducted on the SpRSound 2019 dataset for respiratory sounds classification.

Each of the selected datasets serves a specific purpose in this research, whether it's disorders diversity, multiple information sources, or limited data exploration. Each dataset contributes uniquely to the experimental design and findings.

The Saarbrücken Voice Database offers a wide range of voice recordings from individuals with various pathologies. For our research, we focus on a specific subset that includes sustained vowel recordings from both healthy individuals and patients with different pathologies. This dataset includes Electroglottography (EGG) signals, allowing us to explore the effectiveness of EGG signals in voice pathology classification.

The FEMH 2019 dataset adds a distinctive dimension by including both voice recordings and comprehensive medical records. Its focus on pathological speech samples and detailed medical folders, including demographic information and symptomatology, enriches the analysis. Although it lacks healthy speech recordings, it allows for the integration of medical descriptors as inputs for pathology classification, expanding our understanding of how demographic and medical details contribute to classification accuracy.

Both the Coswara and Virufy datasets enhance the analysis by incorporating crowd-sourced data, featuring diverse sound modalities such as breathing patterns, coughs, sustained vowels, and speech from both healthy and COVID-19-infected individuals.

Additionally, the SpRSound 2019 dataset for respiratory sounds classification was used to evaluate the model's capability to accurately classify different types of respiratory sounds.

Figure 1.3 presents a visual depiction showcasing the wide array and fundamental characteristics of these elements. The datasets mentioned collectively facilitate a comprehensive analysis of voice-related conditions, spanning from traditional voice disorders to phonation abnormalities induced by COVID-19. This approach enhances our understanding of the field and enriches our research findings. Given the diverse range of data, covering both conventional disorders and COVID-19-related phonation issues, our investigation explores various aspects of pathological speech. This study focuses on the intricacies of dysphonia, specifically its organic and functional forms, including hyperfunctional dysphonia. Additionally, we investigate other vocal anomalies such as vocal hyperfunction, vocal fold paresis, and neoplastic growths affecting the vocal apparatus.

After data collection, the subsequent step involves feature extraction. These vocal features encompass various parameters, such as the glottal-to-noise excitation ratio (GNE), Mel frequency cepstral coefficients (MFCC), multidimensional voice program parameters (MDVP), energy, spectral features, and statistical measures like entropy. Advancements have also led to the development of more sophisticated techniques, including modulation spectral features, discrete wavelet packet transform (DWPT), complexity measures, wavelet-based entropy, and various time-frequency representations. Notable studies have demonstrated the effectiveness of these features in distinguishing between normal and pathological speech, particularly in tasks involving sustained vowels and continuous vowels with laryngeal pathologies.

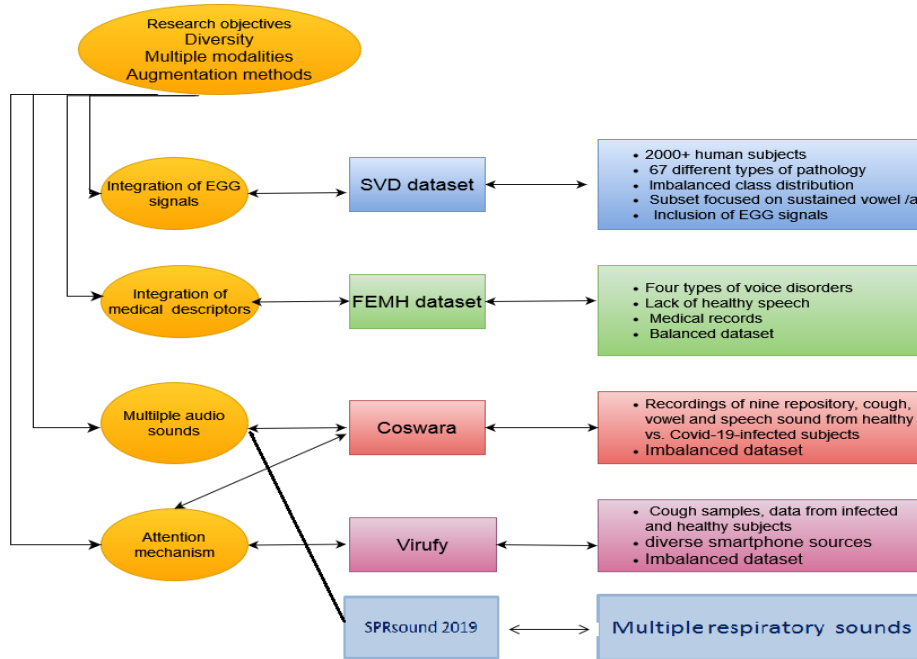


Figure 1.3: Datasets used in the research

Moreover, studies have explored the integration of demographic and symptom-related features into classification models, showcasing a comprehensive approach to voice disorder characterization.

In addition to traditional audio features, recent research has delved into the integration of Electroglottographic signals, which provide valuable insights into vocal fold vibrations. Studies have explored the effectiveness of traditional representations of EGG signals, as well as advanced features derived from them, such as time-domain glottal flow models and normalized metrics like Open Quotient (OQ) and Closed Quotient (CQ). The integration of EGG signals with deep learning architectures has shown promising results, leading to enhanced classification outcomes and improved accuracy rates in voice pathology detection.

Furthermore, feature selection and optimization methods such as genetic algorithms (GA), linear discriminant analysis (LDA), and feature reduction techniques have been employed to refine feature sets and enhance classifier performance.

1.2.2 Machine Learning Approaches for Voice Pathology Classification

Once these features are extracted, classification methods are employed for pathology detection. In recent decades, extensive research has been conducted in the field of machine-driven voice pathology classification systems. Research in pathological voice detection and classification has evolved significantly over the years, with a notable shift towards utilizing machine learning (ML) algorithms for analyzing computed acoustic features of input signals.

The literature reviews conducted in [Heg+19] and [ITA20] offer comprehensive surveys of machine learning methodologies for the automatic detection and classification of voice pathologies. These studies present significant research based on ML techniques, categorized into various approaches including Hidden Markov models (HMM), Gaussian mixture models (GMM), Support vector machines (SVM), decision trees, linear classifiers, K-means clustering, and combined classifiers. These classifiers represent some of the main conventional approaches frequently employed in voice pathology detection.

In detail, we can summarize the classifiers and the reported top classification accuracies as follows:

- **Support Vector Machines:** SVM has been extensively utilized for voice pathology detection [TV12]. It has demonstrated promising results when combined with various feature extraction techniques such as wavelet packet decomposition, energy and entropy features, short-time Fourier transform, continuous wavelet transforms, and wavelet packet transform (WPT) [TSP03]. SVM classifiers have achieved high accuracy rates, ranging from 81% to 100%, depending on the dataset and feature extraction method employed.
- **Decision Trees:** Decision trees, especially Random Forest (RF) classification, have proven effective in voice pathology detection [Ver+06]. RF classifiers have achieved a remarkable 100% classification accuracy when paired with suitable feature sets.
- **K-means Clustering:** K-means clustering has been utilized in conjunction with adaptive time-frequency distribution and non-negative matrix factorization for voice pathology classification. This approach has achieved an accuracy of 98.6% [Kas88].
- **Hidden Markov Models:** HMM are utilized to model the spectral variability of each speech sound using a mixture of Gaussian distributions. Acting as stochastic finite state machines HMM are constructed from a finite set of possible states, each associated with a mixture of Gaussian probability density functions [DBC+01]. One study by Gavidia-Ceballos and Hansen proposed a method for vocal fold cancer detection using HMM, introducing enhanced spectral-pathology components for feature characterization without requiring the estimation of the glottal flow waveform [TV12]. The study achieved accuracies of 92.8% for healthy voices and 88.7% for pathological voices. Similarly, Arias-Londoño et al. employed HMM for feature space transformation of MFCC and short-term noise parameter features, addressing inconsistencies in feature extraction and classification stages. Their proposed technique achieved an impressive accuracy of 96.61% on the MEEI database [Ver+06].
- **Gaussian Mixture Models:** GMM are employed to model the probability distribution of continuous voice acoustic measures. Muhammad et al. utilized GMM for classification of voice recordings of patients with various voice disorders, achieving 99% accuracy by incorporating novel feature extraction methods [AMA+]. Ali et al. also employed GMM for pathological voice detection

and classification, utilizing auditory spectrum and all-pole model-based cepstral coefficients features. They achieved accuracies ranging from 89.47% to 99.56% for different types of voice pathologies. Furthermore, Ali et al. proposed a system for voice disorder detection using GMM, leveraging linear prediction analysis to determine the source signal from speech. By analyzing the spectrum computed from LP features, they achieved high accuracies of 99.75% to 99.94% for both sustained vowels and running speech.

- **Combined Classifiers:** Ensemble classifiers and intelligent systems, which combine multiple classifiers and feature extraction techniques, have demonstrated superior performance in voice pathology detection. These systems have achieved accuracy rates of up to 100% by integrating techniques such as Gaussian Mixture Models, Principal Component Analysis, Linear Discriminant Analysis, Sequential Forward Selection, Sequential Backward Selection, and various classifiers like Support Vector Machines, Probabilistic Neural Networks, and General Regression Neural Networks [DBC+01; Ver+06; TSP03; AMA+].

Certainly, the utilization of deep neural networks, including architectures such as Convolutional Neural Networks, Recurrent Neural Networks, and Long Short-Term Memory networks, has gained significant traction. Recent advancements in these architectures have demonstrated notable improvements in accuracy rates for various voice pathologies.

The field of speech and voice disorder detection utilizing deep learning has witnessed significant advancements in recent years, yet it remains in a state of continual evolution. Noteworthy progress has been achieved, as evidenced by studies reviewed in the systematic literature review in [SS24], showcasing the high accuracy attainable through deep learning in detecting various speech and voice disorders.

The studies included in this review employed a variety of neural network architectures to address specific speech or voice disorders, or to simply classify voices as healthy or pathological.

- **Convolutional Neural Networks:** Ribas [Rib+23b] achieved a maximum accuracy of 93.36%. They initially evaluated the AVDD system using the ComParE feature set and demonstrated a moderate improvement over previously reported results. Joshy et al. [JR22] identified the severity of dysarthria disorder by comparing SVM, CNN, DNN, and LSTM, finding CNN effective in extracting relevant features from various domains, achieving 96.81% accuracy on the TORGO dataset regarding the severity of dysarthria. Safdar et al. [Saf+23] compared MFCC and LPC feature extractors, with optimal accuracy achieved by combining MFCC with MLP. Using 40 MFCC yielded 100% accuracy on both training and testing data. Tong [Ton20] focused on exploring CNN models for identifying speech disorders in children, achieving around 77%-79% accuracy for various trained categories of speech disturbances. The authors proposed LCNN and compared it with VGG-16 and ResNet-50 architectures, with LCNN showing promising results across different scenarios.
- **Long Short-Term Memory Networks:** Joshy et al. [JR22] found LSTM effective for mildly affected dysarthric speakers, but it performed worse for severely affected patients compared to DNN.

- **Deep Neural Networks:** Ribas et al. [Rib+23b] achieved a maximum accuracy of 93.36%. They validated the contribution of DNN as acoustic models, and Joshy et al. [JR22] compared DNN with LSTM for dysarthric speakers. Tong et al. [Ton20] reported DNN achieving the highest accuracy of 93.55% on the UAS dataset for speech disorder classification. The authors proposed a novel cross-modal method for automatic assessment of dysarthria using audio and video features, achieving 99% test accuracy.
- **Multilayer Perceptrons:** Safdar et al. [Saf+23] combined MFCC with MLP, achieving 100% accuracy on both training and testing data.

These studies highlight the effectiveness of deep learning architectures in detecting speech and voice disorders, with advancements being made in feature extraction, classification techniques, and model architectures. However, further research is warranted to address specific challenges and enhance the robustness of these systems for practical applications.

1.2.3 Machine Learning Approaches for COVID-19 Classification

Over the past three years, the worldwide spread of COVID-19 has highlighted the urgent need for reliable methods of audio pathology detection, which are crucial for promptly identifying and diagnosing the virus. This necessity has prompted a notable transition towards utilizing audio-based techniques for detecting, categorizing, and monitoring COVID-19 on a broader scale. While the focus is primarily on a lung disease rather than a pathology of the vocal tract, it has implications for both voice and respiratory sounds. Considering this perspective, we found it both interesting and imperative to incorporate this task into our research.

In the literature, several methods have been proposed employing a combination of conventional classifiers and deep learning techniques. These methods primarily focus on analyzing respiratory sounds such as breathing and coughing, as well as sustained phonation voice samples. In [GD23], the most significant research efforts have been reported in detecting COVID-19 using various innovative approaches from cough sounds. For instance, Ulukaya et al. introduced MSCCov19Net, a multi-branch deep learning model specifically designed for COVID-19 detection [Ulu+23]. Additionally, Almutairi proposed a non-invasive COVID-19 grading framework, integrating multimodal AI-based techniques such as deep learning and fuzzy inference systems [Alm23]. Meanwhile, Chowdhury et al. contributed to the field with a machine learning ensemble-based method, enhancing the reliability of diagnostic tools for COVID-19 detection [Cho+22a]. Hoang et al. underscored the potential of non-invasive diagnosis through their novel cough-based deep learning framework for detecting COVID-19 [Hoa+22]. Aly and Alotaibi also introduced a pioneering deep learning model based on wavelet features, specifically tailored for COVID-19 detection from cough and breathing sounds [AA22]. Furthermore, Ashby et al. emphasized the crucial role of audio quality in COVID-19 detection, proposing a method that incorporates audio quality clustering [Ash+22]. Abayomi-Alli et al. leveraged deep breathing sounds and image augmentation techniques, utilizing sound spectrum analysis and deep learning for COVID-19 detection [Aba+22]. Ren et al. contributed to the development of robust detection methods with their attention-based ensemble

learning approach [Ren+22]. Moreover, Mohammed et al. highlighted the effectiveness of ensemble methods in preliminary screening for COVID-19 [Moh+21]. Chang et al. demonstrated the potential of transfer learning in COVID-19 detection using crowd-sourced cough sounds [Cha+22]. Additionally, Pahar et al. showcased the utilization of global smartphone recordings for COVID-19 cough classification, indicating the broad applicability of their approach [Pah+21]. Lastly, Loey and Mirjalili's work focused on image-based representations for COVID-19 cough sound symptoms classification [LM21].

While coughing has been the predominant focus of analysis, there has also been exploration into incorporating breathing and sustained phonemes as valuable audio sources. Building upon these methodologies, studies have leveraged pre-trained models to process audio features such as Mel Frequency Cepstral Coefficients of cough recordings, breath, and speech sounds. Furthermore, machine learning classifiers including Support Vector Machine, K-Nearest Neighbor, logistic regression, and random forests have been deployed for COVID-19 detection using acoustic data. Recent research has seen an uptick in the utilization of transfer learning strategies to enhance model performance. Techniques such as transfer learning from pre-trained models like AemResNet have been employed to analyze respiration and coughing sounds using convolutional neural networks. Additionally, various architectures including ResNets, VGGs, and AlexNet have been compared for their effectiveness in COVID-19 detection. These advancements underscore the continuous evolution of methodologies aimed at leveraging diverse audio sources and transfer learning to improve the accuracy and robustness of COVID-19 detection systems.

Notable advancements have been made, particularly in Coswara-related studies. High accuracy has been achieved using YAMNet for segmenting and labeling cough sounds with various fractal dimension calculation methods. Additionally, ensemble models incorporating attention mechanisms have demonstrated high accuracy for COVID-19 detection using respiratory, speech, and coughing audio inputs.

While promising global solutions such as QUCoughcope and Project Achoo have emerged, challenges persist. These include the decline in classifier effectiveness when applied to new data collections, particularly in deep learning contexts, and the scarcity of comprehensive and diverse datasets.

To further evaluate our architectures, we conducted experiments in the neighboring field of respiratory sounds classification. In recent years, advancements in machine learning techniques have significantly impacted the field of respiratory sound classification. Researchers have explored various methodologies to address the challenges associated with accurately diagnosing respiratory diseases. Some studies have demonstrated promising results using supervised models with tree based ensemble methods applied to six channel digital auscultations. Others have proposed weakly supervised approaches based on low complexity variational autoencoders or lightweight CNN architectures that outperform larger networks for respiratory sound classification.

Recent works have also explored deep learning approaches for pediatric respiratory sound classification, leveraging Siamese Neural Network frameworks, Patch-Mix Contrastive Learning with Audio Spectrogram Transformers, and dual input deep learning architectures using raw audio signals and STFT spectrograms.

Additionally, top-rated works from recent challenges, such as the 2022 IEEE

Grand Challenge on Respiratory Sound Classification, have provided valuable insights into state-of-the-art methodologies. Some of these approaches have addressed class imbalance issues, introduced ensemble models, or proposed effective lung sound classification systems using advanced CNN models.

1.3 Research Questions, Aims, and Objectives

The problem of voice pathology classification presents a multifaceted challenge rooted in the need for accurate and reliable identification of various vocal disorders. The classification of voice pathologies presents significant challenges due to multiple factors, including the subjective and invasive nature of traditional diagnostic methods, the complexity of voice signals, and the diversity of data in existing datasets concerning voice disorders. The field encompasses a wide range of pathologies, further complicated by imbalanced and low-volume datasets, incorporating data from various audio input sources, thereby increasing the complexity of classification tasks.

Additionally, researchers have defined different classification tasks that involve varying numbers and types of pathologies. This diversity in classification tasks often leads to the development of methods that are effective for specific tasks but not applicable to others.

There is a pressing need for innovative technologies and methodologies that can effectively address the diverse array of pathologies encountered in clinical practice. As voice disorders encompass a wide spectrum of conditions, ranging from common afflictions to rare diseases, traditional diagnostic approaches often fall short in providing accurate and reliable identification across this broad spectrum. Therefore, there is a growing demand for novel technologies and methodologies that can adapt to and effectively classify various types of voice pathologies.

Moreover, given the variability in classification tasks, there is a need for feature extraction techniques that can be tailored to fit different diagnostic scenarios. Each classification task may involve distinct sets of features and patterns characteristic of specific voice disorders. Thus, the development of flexible feature extraction methods capable of accommodating diverse classification tasks is essential for enhancing the accuracy and reliability of voice pathology classification systems.

In addressing this research area, numerous voice pathology datasets have been curated, encompassing data from a diverse array of diseases and conditions. However, amidst the wealth of studies showcasing high classification performance on individual datasets, a recurring challenge emerges when these models are confronted with previously unseen data. While initial findings may be promising, the issue of model generalization becomes increasingly salient. The ability of classification models to extrapolate insights gleaned from training data to novel, real-world scenarios remains a critical concern. This raises pivotal questions regarding the robustness, adaptability, and real-world applicability of these models, prompting deeper exploration into the factors influencing their generalization capabilities. Thus, unraveling the intricacies of dataset generalization is crucial for advancing the field of voice pathology classification and paving the way for more reliable and clinically relevant diagnostic tools.

Transitioning from the discussion of dataset generalization, it becomes evident that available voice pathology datasets contain a rich variety of audio sources, ranging

from phonemes to respiratory sounds. While most research has traditionally focused on utilizing a single audio modality for feature extraction and pathology identification, this approach prompts a critical question: can the fusion of multiple audio modalities enhance classification performance and bolster classifier robustness? By integrating various audio sources, there is potential to uncover feature patterns that may remain elusive when relying solely on one modality. This raises the crucial inquiry of identifying the most relevant audio source for voice pathology classification and assessing its consistency across different pathologies. This gap underscores the necessity of experimenting with diverse audio modalities and developing an architecture capable of flexibly leveraging them to enhance classification accuracy.

Furthermore, existing voice pathology datasets suffer from limitations in data volume, posing a significant constraint for deep learning networks. Surprisingly, minimal research effort has been directed towards developing voice augmentation techniques tailored specifically for voice pathology classification. Consequently, this forms a key area of focus in this research endeavor.

1.3.1 Aim

This thesis aims to comprehensively explore the application of deep learning techniques to address the challenge of voice pathology classification. By leveraging the capabilities of deep learning models, this study seeks to elucidate novel insights into the machine-driven classification of voice pathologies, with the overarching goal of enhancing diagnostic accuracy, streamlining clinical decision-making processes, and ultimately improving patient outcomes. Through rigorous experimentation and meticulous analysis, this thesis endeavors to contribute to the advancement of knowledge in this field, propelling forward the frontier of biomedical research and healthcare innovation.

1.3.2 Objectives

The overarching objective of this research is to devise a deep learning-based methodology capable of robustly extracting feature characteristics and demonstrating adaptability across a wide spectrum of data collections encompassing diverse audio modalities and pathology types. The primary emphasis lies in establishing strong generalization capabilities, ensuring the model's efficacy when confronted with previously unseen and heterogeneous datasets. By prioritizing the development of a versatile and resilient classification framework, this work aims to advance the field of voice pathology classification, ultimately fostering more reliable and clinically relevant diagnostic tools.

This thesis represents an earnest endeavor to address the prevailing scientific gaps and limitations within the realm of voice pathology classification. Embracing an end-to-end system architecture approach, this study aims to formulate a comprehensive feature vector and appropriate deep learning architectures, thereby offering a holistic solution to the challenges of voice pathology classification. It introduces an innovative feature vector designed to significantly bolster classification accuracy and proposes novel modular deep learning architectures that integrate multiple data modalities.

Recognizing the constraints imposed by the scarcity of data in existing voice pathology datasets, this research is committed to devising solutions that enhance the applicability of deep learning networks in this domain. The study extends its scope to include the development of tailored voice augmentation techniques explicitly designed for voice pathology classification.

Furthermore, by exploring the fusion of multiple audio modalities and engineering a versatile architecture, this research endeavors to address pertinent inquiries regarding the most effective audio sources for voice pathology classification while also assessing their consistency across diverse pathologies.

The objective is not only to analyze each modality in isolation but also to understand how they interact and complement each other in the context of voice pathology classification. By leveraging advanced computational techniques, particularly machine learning and deep learning algorithms, the study seeks to develop a system architecture capable of dynamically adjusting to the nuances of each modality. This entails not only extracting relevant features from each data source but also integrating them in a cohesive manner to enhance classification accuracy and robustness.

Through this integrated approach, the study aspires to unlock the full potential of multi-modal data analysis in voice pathology classification. By dynamically adapting to the characteristics of different modalities, the system architecture aims to provide a comprehensive understanding of voice disorders, paving the way for more effective diagnostic and therapeutic interventions. Ultimately, the goal is to develop a versatile and adaptable framework that can accommodate the complexities inherent in multi-modal data analysis, thereby advancing the state-of-the-art in voice pathology classification.

Through these multifaceted approaches, this thesis aims to not only provide valuable insights but also offer practical solutions to the pressing challenges encountered in the diagnosis and classification of voice disorders. Ultimately, the overarching goal is to enhance patient care and treatment outcomes through advancements in voice pathology classification methodologies.

1.4 Thesis Contributions and Key Findings

Through this research, we present a comprehensive methodology aimed at developing robust and effective system architectures for the classification of pathological voice disorders. An overview of the four computational models implemented in this thesis is displayed in Figure 1.4.

Each network architecture was tailored for specific research purposes across our four distinctive sets of experiments. The primary architecture investigates the fusion of medical data as an alternative information source. The second architecture focuses on the implementation of augmentation algorithms and adeptly processes recordings of varying lengths. The third architecture incorporates EGG data, while the fourth one employs an attention-guided mechanism. The final architecture introduces a deep learning framework specifically designed for respiratory sound classification.

This research encompasses five distinct papers, each contributing valuable insights and innovations to the field.

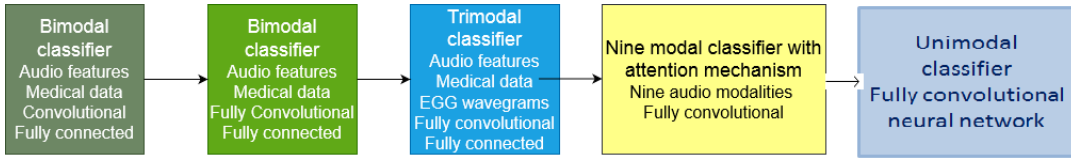


Figure 1.4: Multimodal Classification Frameworks developed for Voice Pathology Classification

The first paper tackles the challenge of formulating feature vectors and constructing deep learning architectures tailored for voice pathology classification. It specifically concentrates on incorporating medical data as a supplementary audio modality within modular deep learning frameworks to improve classification effectiveness. Moreover, it emphasizes and strengthens generalization capabilities by conducting experiments on the relatively less utilized Far Eastern Memorial Hospital 2019 dataset. By incorporating medical parameters alongside audio recordings, the study devises innovative feature vectors that encapsulate both acoustic characteristics and relevant medical descriptors. By addressing the challenges of modality fusion and feature vector development, it lays the groundwork for more robust and comprehensive classification systems capable of handling diverse data sources and enhancing diagnostic accuracy. Through rigorous experimentation on the FEMH 2019 dataset, the proposed methodology demonstrates its effectiveness in classifying voice disorders and showcases its adaptability and generalization capabilities for new tasks and datasets.

Moreover, the paper introduces a novel modular deep learning architecture specifically tailored for classifying voice signals associated with four distinct voice disorders: functional dysphonia, phonotrauma, laryngeal neoplasm, and unilateral vocal paralysis. Leveraging data from the FEMH 2019 dataset, which includes both sustained vowel recordings and corresponding medical records, the study employs short-term feature extraction techniques to extract a sequence of feature vectors from the time-domain representation of the input signals. This involves extracting Mel Frequency Cepstral Coefficients along with their first-order derivatives, concatenated with the logarithm of Mel-filterbank outputs. Additionally, fundamental frequency, jitter, and Harmonic-to-Noise Ratio are computed to augment the feature vectors. In addition, 34 medical descriptors are integrated into the feature representation to enrich the input data.

The proposed architecture consists of two distinct sub-networks: one dedicated to processing the audio recordings as a 2-D representation (image) utilizing convolutional layers, and the other analyzing an augmented input vector containing both acoustic features and medical descriptors through a feed-forward network. This bimodal classifier approach facilitates the simultaneous extraction and fusion of both acoustic and medical information, capturing a comprehensive representation of the voice pathology data.

Notably, the effectiveness of the proposed methodology was evaluated through participation in the FEMH Voice Data Challenge 2019, where it achieved a commendable fifth position. Despite a slight performance margin from the leading method, the proposed architecture demonstrated competitive performance, underscoring its potential for accurate voice pathology classification.

In the second paper, we refine our previous deep learning architecture and tackle

the challenge of data scarcity. The primary objective of this study was to develop a multimodal framework capable of effectively processing both low-level and mid-level features extracted from acoustic signals of varying durations while also incorporating relevant medical records. This adaptation was crucial for accommodating the inherent variability in voice recordings encountered in real-world scenarios, where the duration of utterances can vary significantly. Specifically, we enhance our deep learning architecture to better accommodate the particularities of audio recordings. This enhancement involves transitioning from a convolutional architecture to a fully convolutional one, enabling more adaptable processing of audio signals with varying durations.

In addition to architectural enhancements, the study also addressed the challenge of limited training data by exploring problem-specific augmentation techniques. These methods involved segmenting audio recordings into sequences of variable duration, determined dynamically by an innovative algorithm. Furthermore, the study introduced colored noise injection, incorporating a mix of white, pink, and brown noise into the training process.

Through rigorous experimentation and evaluation on the FEMH 2019 dataset, the study provided compelling evidence that these innovations indeed yield state-of-the-art results. The proposed methodologies demonstrated superior performance compared to existing approaches, surpassing even the best-performing method in the 2019 FEMH data challenge. This achievement underscores the efficacy of the proposed architectural enhancements and augmentation techniques in enhancing the robustness and generalization capabilities of the classification model.

The proposed modular deep learning architecture demonstrates remarkable performance even when faced with limited training data. Our experiments underscore the importance of medical parameters as valuable supplementary features for voice pathology classification. Moreover, we validated the effectiveness of segmentation-based data augmentation and colored noise injection techniques. A noteworthy insight from our investigation is the correlation between the duration of voice recordings and the presence of voice pathology. By developing a classifier capable of handling recordings of varying durations through a fully convolutional stack, we observed a substantial enhancement in classification accuracy compared to fixed-duration approaches. Our study advocates for the development of models that can process audio recordings at their original, potentially variable durations. This strategy aligns with our assertion that sustained utterance duration and intensity often harbor critical diagnostic information, which should be carefully considered—a deviation from conventional fixed-size segmentation methods and zero-padding procedures.

The third paper delves into the integration of multimodal data within a modular deep learning architecture for automatic voice pathology classification, with a particular emphasis on incorporating electroglottographic signals as a novel modality. The proposed architecture amalgamates audio descriptors, categorical medical data, and features extracted from EGG signals within a modular deep learning framework.

Specifically, the architecture utilizes a cascade of convolutional layers to process short-term audio feature vectors, which include Mel-Frequency Cepstral Coefficients, their derivatives, and Mel filter-bank outputs. In parallel, a convolutional branch handles “wavegrams” derived from the glottal chords obtained through EGG signals, thereby introducing a new modality to enhance the system’s accuracy. Additionally,

the architecture incorporates medical records and mid-term audio features, processing them in a standard feed-forward branch.

To assess the efficacy of this novel approach, extensive experiments were conducted on a subset of the Saarbrücken voice database, focusing on sustained vowel /a/ recordings. This subset comprises healthy subjects and individuals with three types of voice pathology: hyperfunctional dysphonia, laryngitis, and recurrent laryngeal nerve paralysis. The data were highly imbalanced, requiring special attention to enhance classification accuracy while also improving recall metrics. Experimental findings indicate a notable enhancement in classification accuracy by integrating information from multiple modalities, surpassing the performance of unimodal methodologies, along with improved recall values. Ultimately, the proposed architecture achieves state-of-the-art classification accuracy, marking a significant advancement in voice pathology detection.

The fourth paper contributes to the advancement of the field by addressing the crucial question of determining the most suitable audio source for pathology detection, particularly focusing on Covid-19 detection. It introduces a novel multimodal deep learning architecture capable of analyzing diverse audio modalities, including respiratory sounds, vowels, and medical records, to enhance Covid-19 detection accuracy.

This innovative architecture integrates nine convolutional sub-networks, each dedicated to processing a specific type of Covid-19 audio recording. One of the key innovations introduced in this study is the attention-guided mechanism, which dynamically selects the most relevant acoustic modality for each classification decision. This mechanism not only enhances accuracy but also provides flexibility across different datasets, thereby improving the model's generalizability.

The paper emphasizes the significance of integrating attention mechanisms within the classifier architecture, demonstrating their profound impact on performance enhancement. By highlighting the importance of these mechanisms, the study underscores their role in improving classification accuracy and adaptability across diverse datasets.

Furthermore, the research extends its investigation to explore data augmentation techniques, such as segmentation-based data augmentation and colored noise injection. Through comprehensive analysis on the Coswara and Virufy Covid-19 datasets, the study demonstrates the effectiveness of these techniques in improving the generalization capabilities of the developed classifiers.

The results of the analysis showcase the adaptability and robustness of the developed classifiers in real-world diagnostic scenarios, making them promising tools for non-intrusive Covid-19 diagnosis. Overall, the paper contributes significantly to advancing the field of pathology detection, particularly in the context of Covid-19, by introducing innovative architectures and techniques to improve accuracy and generalizability.

The fifth and final paper presents a significant contribution to the field of respiratory sound analysis through the introduction of an architecture featuring a fully convolutional network capable of processing audio signals of arbitrary duration. The

paper outlines a comprehensive processing pipeline for respiratory sound classification, starting with the extraction of short-term features, specifically sequences of Mel-frequency Cepstral Coefficients from the input signal. Additionally, the Mel spectrogram of each audio signal is extracted, ensuring consistency in feature representation across all signals.

A key innovation of the proposed pipeline is the utilization of a fully convolutional neural network architecture, which can handle audio recordings of varying durations without the need for segmentation or zero-padding procedures. The network architecture comprises four consecutive convolutional layers followed by a global max pooling layer and a dense layer with ReLU activation. Finally, a softmax layer is employed to output the final classification decision.

The paper conducts extensive experimentation to evaluate the performance of the classifier, with a specific focus on feature representations and network parameters. The experiments include the evaluation of both MFCC and Mel spectrogram representations for ternary and multi-class classification tasks. The results demonstrate competitive performance for both representations, with the Mel spectrogram exhibiting superior performance, particularly in the multi-class task.

Overall, the findings underscore the importance of feature selection in respiratory sound classification and highlight the potential of leveraging original duration recordings and appropriate feature representations to achieve accurate classification.

In conclusion, the dissertation presents cutting-edge research in the realms of voice pathology classification and Covid-19 detection, offering novel modular deep learning architectures, feature extraction methods, and innovative data augmentation techniques with significant performance improvements. It emphasizes the importance of multimodal data integration, showcases architecture adaptability, and underscores the significance of specific data augmentation techniques. The implications of this work extend to diverse information sources and data volumes, promising advancements in the accuracy and reliability of medical diagnoses. The contributions made in this thesis open new paths for future investigations, with a shared goal of enhancing the accuracy and efficiency of voice pathology diagnosis. The potential impact extends beyond the academic realm, with a direct and positive effect on the health and quality of life of patients. The findings presented here serve as a stepping stone for the development of more robust and reliable voice pathology classification systems, ultimately leading to earlier and more accurate diagnoses, and better outcomes for those affected by voice disorders.

1. Development of Deep Learning Models:

- Proposed modular deep learning architectures integrating fusion of multimodal data.
- Highlighted the importance of accommodating variable-duration utterances for improved performance.
- Introduced a modality-selection attention mechanism to dynamically prioritize important input features.

2. Novel Feature Vector Formulation:

- Created a sophisticated audio feature vector using MFCC, first-order derivatives, and Mel-filterbank outputs to improve model performance.

3. Significance of Multi-Modal Approach:

- Demonstrated the effectiveness of integrating diverse modalities, such as audio signals, medical parameters, and perturbation features, to enhance classifier robustness.

4. Impact of Strategic Data Augmentation:

- Showed the effectiveness of data augmentation techniques, including adjustable varying-length segmentation and colored noise injection, in improving generalization.

5. Enhanced Classification Accuracy:

- Achieved notable increases in testing accuracy, especially in scenarios involving imbalanced classes or limited annotated data, by employing robust deep learning techniques and data augmentation.

6. Insights into Model Interpretability:

- Provided valuable insights into the learned representations of the model's intermediate layers, emphasizing the significance of both acoustic and medical features.

These contributions provide a foundation for advancing deep learning methods in voice pathology classification, demonstrating the importance of multi-modal data fusion, novel feature extraction, strategic data augmentation, and attention mechanisms to enhance classification performance and interpretability.

1.5 Thesis Outline

The thesis is structured to provide a thorough exploration of our voice pathology classification method employing deep learning techniques, organized across five chapters. This introduction serves as a comprehensive guide to the research, defining the scope and significance of the topic, presenting key research questions, and outlining the methodologies and experiments conducted, which form the basis of the research findings.

The following chapter introduces a bimodal classifier that integrates both audio features and medical descriptors. It begins with a thorough literature review that explores the existing body of knowledge in the field, including research on traditional methods and the evolving role of deep learning in voice pathology classification, with a focus on works related to the FEMH dataset. Additionally, background theory pertinent to the research topic is presented, covering neural network principles, methods, and the perturbation features used for the analysis. This chapter also details the FEMH data collection process and the preprocessing steps undertaken to extract the audio and medical feature vector.

After data collection and preprocessing, the subsequent section presents a comprehensive exposition of the proposed methodology. Here, the details of the approach, including the architectural design of classifiers for both constant and varying-length

models, are described. Additionally, data augmentation techniques concerning the segmentation of audio clips of varying length and colored noise injection are discussed in detail. The Experimental Setup segment outlines the experiments undertaken to assess the system's effectiveness and robustness, specifying the configurations and parameters used.

The Results and Discussion sections present the empirical findings obtained from the experiments and offer insightful interpretations and analyses of the model's inherent layers. This section highlights the performance metrics and outcomes, assessing the efficacy of the bimodal architecture, the integration of medical data, and the data augmentation and noise injection techniques while also discussing the limitations of the proposed approach.

The following chapter retains the same structure, explaining the modular deep learning system that enhances the processing of Electroglottographic signals. It delves into related work on the SVD dataset and includes research on the integration of EGG signals. Additionally, it introduces background theory on EGG methods and waveform analysis, defining the EGG-based parameters and waveforms used throughout this research. The chapter describes the SVD data, details the task under study, and presents the trimodal architecture incorporating EGG data. Subsequently, the experiments and results are presented, showing that EGG can significantly contribute to classifier performance in unbalanced datasets.

In the subsequent chapter, the multimodal architecture incorporating a self-guided attention mechanism for COVID-19 detection is detailed. Beginning with an overview of COVID-19 audio signal processing and a literature review, the chapter defines the spectral features utilized in the architecture. It also provides descriptions of the COVID-19 datasets, along with architectural details of nine subnetworks followed by a modality selection module. The methodology and experiments are presented, with results demonstrating the effectiveness of the attention module.

In the final chapter, the conclusions drawn from this thesis summarize its significant contributions to the field. By synthesizing the research findings, it provides a comprehensive overview of the key insights gained throughout the study. Furthermore, it discusses the broader significance and potential impact of these findings on the field. The thesis also outlines promising future directions that can build upon the groundwork laid by this research, offering a roadmap for continued exploration and advancement in the field. Overall, this chapter serves as the culmination of the thesis, distilling its essence into actionable insights and setting the stage for future inquiry and innovation.

Chapter 2

Combining Acoustic and Medical Descriptors into a Modular Deep Learning Architecture for Voice Pathology Classification

2.1 Introduction

The assessment of voice pathology traditionally relies on invasive techniques such as laryngoscopy and detailed examinations of laryngeal structure and respiratory dynamics.

Laryngoscopy is a common technique that allows for detailed observation of the larynx, providing insights into tissue characteristics and movement. This procedure plays a fundamental role in assessing the physical attributes and functionality of the larynx. The conventional method for diagnosing voice pathology typically involves a comprehensive examination of both laryngeal structure and mobility, coupled with an evaluation of respiratory dynamics.

These metrics — encompassing lung volume, airflow, pressure, and breathing dynamics — provide significant information for the diagnostic process. Acquired through specialized techniques such as spirometry and pneumotachography, these parameters enable precise measurement and analysis of respiratory variables, helping healthcare professionals evaluate both laryngeal and respiratory aspects. However, these methods, while effective, are resource-intensive and may not be readily accessible to all patients. Moreover, they frequently cause discomfort and inconvenience for patients undergoing diagnosis.

In response to these limitations, there is increasing interest in leveraging advanced audio signal processing and machine learning techniques for non-invasive voice pathology classification. These methods extract valuable information from voice signals, offering a promising alternative to complement traditional diagnostics.

The financial benefits of machine-driven speech signal processing are considerable, as it reduces the need for expensive specialized medical equipment. Additionally, its non-invasive nature aligns with efforts to improve patient comfort and convenience in healthcare. Integrating machine learning techniques into diagnostic processes is a major step toward efficient, affordable, and patient-centered healthcare solutions.

This chapter presents our initial study on automatic voice pathology classification. Our framework deviates from conventional diagnostic tools by utilizing advanced audio signal processing and deep learning methods. The proposed method aims to provide an assessment of phonatory intricacies and respiratory patterns without the need for invasive procedures.

2.1.1 Problem Definition and Application Context

In this chapter, we focus on developing a reliable voice pathology classification system utilizing machine learning techniques. The primary aim of this study is to address the limitations of existing diagnostic methods and design an advanced voice pathology classifier based on deep learning architectures. Our goal is to create a robust, rapid, and accurate system capable of detecting normal and pathological speech, as well as classifying specific voice pathologies.

To achieve this, we investigate the development of precise feature extraction techniques to enhance classification accuracy. Additionally, we explore the integration of medical data into the classification system to provide a more comprehensive analysis. Furthermore, we experiment with various data augmentation techniques to improve the classifier's generalizability and performance.

2.1.2 Data Resources and Research Gaps

Through our bibliographic review, it is evident that considerable research effort has been directed towards voice pathology classification. Researchers have explored various pathology types through diverse classification tasks utilizing existing voice disorder datasets. To facilitate these efforts, databases have been compiled, consisting of voice recordings from both healthy individuals and patients with hyperfunctional and hypofunctional vocal fold pathologies. These recordings typically include sustained vowels such as /a/, /i/, /e/, /o/, as well as continuous speech samples.

A critical issue encountered at the beginning of our work was the very limited number of accessible voice pathology datasets. As reported in [SRH20a], among the prominent ones are the **Massachusetts Eye and Ear Infirmary Database** [EI94] and the **Saarbruecken Voice Database** [BP07]. These datasets include sustained vowels, numbers, and sentences from individuals with various voice pathologies and healthy speakers.

The SVD, available online, provides over 2,000 recordings from German speakers with various voice disorders, including healthy and pathological voices. The recordings include vowel sounds /a/, /i/, and /u/ at various pitches—high, low, low-high, and normal. Each recording lasts between 1 and 4 seconds and is in waveform format (16-bit sample), sampled at 50 kHz. The database includes recordings from 71 distinct voice pathologies, as well as healthy individuals.

On the other hand, the **MEEI Dataset**, which is commercially available, provides a collection of voice recordings from both healthy individuals and those with various pathologies. It includes recordings of sustained vowel /ah/ (53 normal, 657 pathological) and continuous speech (53 normal, 661 pathological). Additionally, the database features clinical and personal details of subjects and acoustic analysis results using

Kay's Multi-Dimensional Voice Program (MDVP). The files in the dataset have varying sampling frequencies: normal and a small percentage of pathological files are sampled at 50 kHz, while the majority of pathological files are sampled at 25 kHz. The normal files have an average length of 3 seconds for sustained vowels and 12 seconds for running speech, while pathological files average around 1 and 9 seconds, respectively. At that time, it was the most widely used and accessible of all voice quality databases.

However, there are several key considerations and drawbacks to be aware of when using the MEEI dataset for research purposes [Sae+06]. The most significant drawback is that the recordings have been pre-edited to include only the stable portion of phonation. Additionally, the binary classification problem of distinguishing between healthy and unhealthy samples has been extensively tested on these data. As highlighted by [DB14], the sustained vowels in the MEEI database can be perfectly classified into healthy and unhealthy categories. However, there are significant concerns regarding the generalizability of these methods to unseen data, emphasizing the need for experiments with new databases.

This research gap prompted us to explore new datasets, leading to our investigation with the Far Eastern Memorial Hospital dataset. At the time these experiments were conducted, there were extremely limited voice pathology datasets available. The FEMH dataset was a novel resource, providing recordings of specific disorders diagnosed at the FEMH hospital. A detailed description of its contents is provided in subsection 2.3.1.

After this work was published, and towards the end of our research, it is noteworthy that several new voice pathology datasets have recently become available. For completeness, we also list the most significant ones below.

1. **Advanced Voice Function Assessment Database (AVFAD)** [Lui+17]: Offers recordings from Portuguese speakers, including individuals with a range of voice disorders. Contains recordings from 709 individuals, including 346 with vocal pathologies, collected by researchers at the University of Aveiro.
2. **Universal Access Dysarthric Speech (UA-Speech) Dataset** [Kim+08]: Features recordings from English speakers affected by dysarthria, containing data from 19 speakers with dysarthria and 13 healthy speakers.
3. **TORGO Dataset** [RNW11]: Serves as a repository of dysarthric speech data from individuals with cerebral palsy or amyotrophic lateral sclerosis. Includes samples from 15 speakers with dysarthria and 7 healthy speakers, collected by researchers at the University of Toronto.
4. **Voice Database for Objective Evaluation and Diagnosis of Voice Disorders (VOICED) Dataset** [Ces+18]: Presents a curated collection of recordings for research into specific vocal fold disorders, containing 208 recordings, including 150 pathological and 58 healthy voices, collected by researchers at the University of Federico II in Naples, Italy.
5. **LAANA Dataset** [GVT13]: Provides audio recordings from children with and without specific language impairment (SLI), including recordings from children aged 6 to 12 years old, collected by the Laboratory of Artificial Neural Network Applications in the Czech Republic.

Throughout our bibliographical review, we also identified an additional challenge in developing methodologies that integrate multiple streams of information. This challenge necessitates the integration of diverse types of data, ranging from clinical observations and patient records to advanced audio data. **The objective is to design comprehensive approaches that leverage a multitude of data sources, thereby enhancing the precision and effectiveness of classification decisions.** Addressing this challenge is crucial for harnessing the full potential of available information in healthcare systems and fostering innovations in medical decision-making processes.

At the outset of our literature review, we noted that the majority of research on machine-based voice pathology classification primarily relied on audio recordings, with limited exploration of Electroglottographic signals. **This observation highlighted the need to investigate alternative sources for voice pathology classification, prompting us to integrate and experiment with medical information in our classification system.**

Finally, we observed that the available databases are limited in terms of the quantity of data for each pathology type, which presents a significant challenge for training deep learning architectures. To address this problem, **we explored data augmentation techniques specifically tailored for voice pathology classification tasks.**

2.1.3 Research Objectives and Contributions

To address the dual objectives of expanding applicability to novel datasets and diverse pathologies while integrating various audio modalities, our experimentation centers on two primary aspects. Firstly, we aim to assess the generalizability of our methods on previously unexplored datasets, ensuring the robustness of our approaches. Secondly, we conduct experiments to seamlessly incorporate diverse information sources, enriching our analysis with various data, including audio recordings and medical parameters. Through these efforts, our aim is to improve the adaptability of our models across different datasets and information sources.

In this context, we experimented with a challenging database of voice disorders—the Far Eastern Memorial Hospital database. This database was initially introduced during the FEMH Voice Data Challenge, part of the 2019 IEEE BigData Cup. It involves a 4-class classification task with recordings from four disorder categories: functional *dysphonia*, *phonotrauma*, *laryngeal neoplasm*, and *vocal paralysis*. The FEMH corpus includes voice recordings and patients’ medical information, aligning with our research objective of integrating medical data into the classification system.

The FEMH Voice Data Challenge, held in the 2019 IEEE BigData Cup, focused on a 4-class classification problem involving functional dysphonia, phonotrauma, laryngeal neoplasm, and vocal paralysis. This challenge expanded upon the 2018 edition, which aimed at distinguishing healthy and pathological cases based on three voice pathologies: neoplasm, phonotrauma, and vocal palsy.

In our efforts to obtain the FEMH 2019 dataset, we actively participated in the associated challenge, submitting an initial version of our classifier and achieving a notable fifth-ranking position. The challenge framework required participants to train models on provided training datasets, with evaluation conducted on a testing set that was undisclosed.

We participated with a modular deep learning architecture that integrates audio features with medical descriptors. This innovative architecture combines convolutional and feed-forward neural networks, organized into two distinct sub-networks. Each sub-network is specifically tailored to process a unique input source, handling either audio recordings or medical data.

To further develop and enhance our methodologies, we continued experiments even after the competition ended. Upon analyzing the voice recordings, we identified two significant aspects requiring attention. Firstly, the recordings exhibit significant variability in duration, suggesting diverse temporal characteristics among the collected audio samples.

Secondly, we encountered a notable limitation due to the relatively restricted quantity of available data, which affects the robustness and generalization capabilities of machine learning models trained on such datasets. Sufficient data is essential for constructing reliable models, and this limitation necessitates the investigation of potential strategies, such as data augmentation techniques.

In response to these findings, addressing temporal variations in recording duration and implementing data augmentation techniques became essential components of our approach to ensure a thorough and effective exploration of voice pathology classification. To improve system robustness and generalization, we developed a more sophisticated architecture to accommodate varying sample durations, alongside data augmentation techniques tailored for voice pathology classification purposes.

The FEMH dataset, like most existing voice disorder databases, suffers from a scarcity of data, posing a notable limitation when training deep learning architectures. To address this challenge, we explored various data augmentation methods specifically tailored to improve the training process for voice pathology classification.

Another crucial observation pertained to the variable duration of recordings, which is directly linked to the phonation capabilities of patients. This variability arises because sustained vowel duration correlates with specific aspects of a patient's condition. For instance, disorders such as vocal fold mucosa damage, as seen in neoplasm cases, lead to altered Maximum Phonation Time (MPT) values. Patients with incomplete glottis closure, a symptom of vocal paralysis, exhibit lower MPT values due to increased air leakage compared to healthy individuals [Cho+12]. Such insights underscore the importance of considering temporal variations in analyzing voice pathology datasets. In vocal paralysis, shorter phonation times signify the patient's struggle to maintain sustained vowels for extended periods.

Furthermore, variations in loudness and silent intervals within recordings serve as indicators of a patient's difficulty in articulating vowels. However, current research predominantly focuses on network architectures and methodologies that require audio recordings to be resized into fixed-length segments, typically involving zero-padding. Segmentation into fixed-length, zero-padded segments overlooks the correlation between recording duration and pathology presence, potentially leading to information loss and suboptimal classification performance.

To overcome the limitations associated with classifiers that operate on fixed duration inputs, we introduced a fully-convolutional architecture capable of processing 2-D representations of audio recordings with varying durations. Additionally, we developed a corresponding data augmentation technique involving dividing audio

recordings into segments of variable lengths. This approach is designed to circumvent the constraints of traditional classifiers and harness the complete spectrum of information present in recordings of variable durations.

In summary, our enhanced proposal introduces an innovative bimodal classifier that operates on two modalities: audio signals and medical records. This classifier relies on a fully convolutional architecture [LSD15], treating voice recordings as adaptable-width images. This approach effectively eliminates the need for preprocessing related to recording duration. Furthermore, our training stage incorporates tailored augmentation techniques, generating variable-length training data infused with diverse noise types to enhance the learning process.

Overall, transitioning from exploring alternative approaches in voice pathology diagnosis, we delved into the realm of automatic voice pathology classification, offering advancements through incorporating deep learning techniques. Emphasizing the pivotal role of accurate feature extraction techniques and high-precision classification algorithms, we utilized databases containing recordings from both healthy individuals and those with vocal fold pathologies.

The research underscores the necessity of experiments on novel databases to ensure generalizability. Simultaneously, it explores the complex challenge of integrating diverse information streams for comprehensive medical diagnosis.

It also focuses on addressing dual objectives: generalizability to new datasets and the integration of diverse audio modalities. This sets the stage for deeper exploration, utilizing a more challenging dataset—the Far Eastern Memorial Hospital database. The experimentation in the FEMH Voice Data Challenge is introduced, along with participation in a modular deep learning architecture that integrates audio features with medical descriptors.

In summary, this work introduces:

a) Multimodal Integration for Comprehensive Diagnosis: The research emphasizes the integration of various streams of information within medical informatics. By combining clinical observations, patient records, and audio data, the proposed method aims to enhance the precision and effectiveness of medical diagnosis, fostering innovations in data-driven healthcare practices.

b) Bimodal Classification with Audio and Medical Records: The enhanced proposal introduces a bimodal classifier operating on two modalities: audio signals and medical records.

c) Generalization Across Diverse Datasets: The experimentation focuses on evaluating the generalizability of methods on previously unexplored datasets, ensuring robustness beyond specific training data. This addresses the challenge of developing models that can adapt across different datasets and information sources.

d) Fully Convolutional Neural Network Architecture for Variable-Length Recordings: To overcome limitations in classifiers operating on fixed-duration inputs, the proposed architecture treats voice recordings as adaptable-width images. This fully convolutional approach eliminates the need for preprocessing related to recording duration and incorporates data augmentation techniques tailored for variable-length recordings.

e) Data Augmentation Techniques: Acknowledging the scarcity of data in voice disorder databases, the method introduces data augmentation techniques specifically tailored for voice pathology classification. This innovative approach addresses the

challenge of limited training data, contributing to the robustness and generalization capabilities of the machine learning models.

The upcoming sections of this chapter will delve into various aspects of the research topic, ranging from the theoretical foundations to the experimentation and the conclusions derived from the study. Sections 2.2 and Appendix A will provide a comprehensive exploration of existing knowledge related to our study in the literature review and background theory. This establishes a contextual foundation for our proposed methods, which will be detailed in the subsequent Section 2.3. Section 2.4 will encompass a description of the databases used, the approach to feature extraction, and the framework of our model architecture. Section 2.5.1 presents the experimental details, including the experimental setup, obtained results, and a comprehensive analysis of key issues that arose during experimentation. Finally, the concluding section will serve as the culmination of this study, offering profound insights and drawing conclusive remarks based on our findings.

2.2 Related Work

In this section, we present an overview of the key works that contributed to the formulation of the feature vector and the architecture used in our experiments. This shift in focus is particularly significant in the context of the Far Eastern Memorial Hospital dataset, where our experiments are conducted and the results reported. Therefore, the subsequent discussion will concentrate on studies directly relevant to the FEMH dataset, providing an overview of the foundational works that informed the construction of our feature vector and guided the methodology employed in our research.

2.2.1 Audio Features in Voice Pathology Classification

This subsection provides an overview of the audio features used in voice pathology classification studies, including fundamental features such as Mel Frequency Cepstral Coefficients and advanced techniques like modulation spectral features and Discrete Wavelet Packet Transform. The varying classification accuracies of these features underscore the need for customized feature selection based on specific voice disorders. As an initial step in our research, we evaluated these features to inform the development of the feature learning representation for our deep learning architecture.

Our investigation encompassed fundamental audio features such as MFCC, energy, spectral features, and statistical measures like entropy. We also examined studies employing modulation spectral features, DWPT, complexity measures, wavelet-based entropy, and time-frequency representations. This comprehensive evaluation aimed to identify the most effective features for enhancing the accuracy of voice pathology classification in our experiments. Notably, several articles have highlighted the effectiveness of mel-frequency coefficients and their derivatives as distinguishing features for various types of voice pathology, as observed in studies such as [Sid+21], [RSA20], and [NA16b].

For instance, in the study conducted by [DNB02], combining MFCC with pitch frequency measurements achieved remarkable accuracy in distinguishing normal from pathological speech, particularly for the sustained vowel /a/. Similarly, another study by [ZJ08a] explored the effectiveness of perturbation methods (jitter and shimmer)

along with the signal-to-noise ratio and nonlinear dynamic methods (correlation dimension and second-order entropy) for analyzing both sustained and continuous vowels with laryngeal pathologies.

Additionally, [Ari+11] focused on complexity measures of noise parameters and MFCC coefficients, while [CS07] employed the wavelet packet transform to analyze dysphonic voices. In [MS11b], researchers explored modulation spectral features and Mel-frequency cepstral coefficients for detecting dysphonia. The study [ZJ08b] conducted acoustic analyses on sustained and running voices from patients with laryngeal pathologies, revealing significant differences, particularly in terms of jitter, shimmer, correlation dimension, and second-order entropy. In [Mar+10b], an automatic system was developed for differentiating between pathological and normal speech, using a combination of modulation spectral features and MFCC to detect dysphonia.

The study [BAM06] assessed the effectiveness of PPQ, APQ, and correlation dimension features in characterizing the impact of nodules and polyps on vocal fold vibration patterns. Utilizing pitch and amplitude perturbation quotients (PPQ and APQ) derived from a modified cepstrum-based pitch detection algorithm, the investigation revealed that correlation dimension, extracted using the Grassberger-Procaccia algorithm, emerged as the most effective nonlinear dynamic feature for differentiating between vocal fold nodules and polyps.

In the study by [MGA14], feature extraction employed three techniques: MFCC, LPCC, and RASTA-PLP. The selection of coefficients from each technique involved statistical methods such as the t-test, Kruskal-Wallis test, or genetic algorithm, specifically applied to recordings of the sustained vowel /a/ to classify pathologies associated with vocal fold nodules, vocal fold polyps, keratosis, and spasmodic dysphonia.

In [Uma+02], a distinctive method for discriminating pathological voices was proposed, employing an adaptive time-frequency approach based on the matching pursuit algorithm. The study achieved remarkable classification accuracy by combining four TF features (Omx, Orne, Er, and Lr) through pattern classification. The research utilized a voice disorders database from the Massachusetts Eye and Ear Infirmary, encompassing both normal speakers and individuals with various organic, neurological, traumatic, and psychogenic voice disorders.

In [Al+14], the significance of frequency bands in automatic voice pathology detection was investigated by passing voice signals through time-domain band-pass filters. Emphasizing the frequency range of 1500 Hz to 3500 Hz as crucial, the study achieved enhanced accuracy for voice pathology detection. The approach utilized a specific filter bandwidth, focusing on two features derived from the normalized auto-correlation function—peak value and lag. Notably, the research covered both English and German databases, demonstrating the effectiveness of the approach across different languages and datasets.

Moving beyond the basics, we transition to more advanced techniques, including modulation spectral features combined with MFCC, complexity measures of noise parameters, wavelet packet transform analysis, and innovative bio-inspired algorithms.

For dysphonia detection in recordings of the sustained vowel /a/, studies such as [MQS05], [Mar+10a], and [MS11a] utilized biologically inspired AM analysis features, modulation spectral features combined with MFCC, and modulation spectral features, respectively, demonstrating promising performance. Recent work by

[Pol+19] introduced bio-inspired algorithms incorporating innovative graphical representations of audio signals and heuristic methods. Building upon this, [Muh+12] combined Gaussian mixture models with the k-means algorithm, achieving high accuracy exceeding 99% in various scenarios, including text-independent and text-dependent cases.

This subsection also highlights the critical role of feature selection and optimization methods in improving classifier efficacy. Genetic algorithms, linear discriminant analysis, and feature reduction techniques were employed to enhance the discriminative power of classifiers. In [Arj+11], the researchers used LDA and support vector machines with different feature reduction methods. In a related study, [AA14] developed an efficient classification system using discrete wavelet packet transform, multi-class LDA (MC-LDA), and a multilayer neural network (ML-NN). The research explored features such as energy and Shannon entropy derived from wavelet packet coefficients.

This comprehensive exploration and synthesis of previous studies served as the foundation for our deep learning-based voice pathology classification approach. Building on insights gained from these studies, we aimed to advance our approach by developing more sophisticated features. This involved integrating various modalities and incorporating relevant medical parameters to enhance the robustness and precision of our classification model, using a comprehensive set of features derived from both acoustic signals and pertinent medical data.

2.2.2 FEMH Dataset in the Literature

Throughout our literature review, we focused on studies specifically conducted on the FEMH dataset. As mentioned in the introduction, we gained access to the FEMH data by participating in the FEMH 2019 IEEE Big Data Challenge. This competition extended the objectives of the 2018 challenge, which focused on the simpler task of distinguishing between recordings of healthy subjects and pathological cases involving three voice pathologies: neoplasm, phonotrauma, and vocal palsy.

Exploring the 2018 FEMH challenge, we observed that predominant methods favored mainstream machine learning approaches. Several techniques relied on Support Vector Machines [PLZ18; DEH18; Tom+18; She+18; Fan+19; IPL18], Gaussian Mixture Models [Jul+18; Fan+19], Bayesian networks, and Random Forest classifiers [BK18]. Neural network architectures, however, were less commonly utilized during this period [BK18; Fan+19; Shi+19].

For the classification task distinguishing healthy from pathological recordings, the top-performing method in terms of accuracy was reported by [DEH18], which used Gaussian Mixture Models and achieved an accuracy of 96.9%.

In the three-class task of voice pathology classification in FEMH 2018, unweighted average recall was used as the performance metric. The method presented in [Tom+18] achieved the highest result (60.67%) using Gaussian Mixture Models.

While reviewing FEMH-related literature, we found significant efforts to integrate demographic data. Studies such as [She+18], [Fan+19], and [Shi+19] included demographic and symptom-related features. In contrast to the methods employed in 2018, our approach, which secured the fifth position in the 2019 challenge, utilized deep learning methodologies to integrate diverse information modalities into an

end-to-end architecture. A notable approach, referenced in [Shi+19], employed deep neural networks to merge acoustic and medical data, but in a distinct manner. It utilized conventional feed-forward architectures, requiring Gaussian Mixture Models as a preprocessing step to statistically represent an audio recording before inputting it into the neural network.

Our methodology, as detailed further, involves end-to-end representations without intermediate feature processing. It operates on mid-term segments, generates 2-D representations, and incorporates perturbation features alongside medical metadata within the same sub-network.

The introduction of Tables 2.1 and 2.2 is intended to provide a more comprehensive analysis and comparison of the relevant studies conducted during the 2018 FEMH challenge. Table 2.1 presents a comprehensive overview of studies dedicated to voice pathology classification using the FEMH dataset. Each entry in the table represents a unique investigation, providing insights into the features used, classification methods employed, and the corresponding results.

In the study [BK18], OpenSMILE's 6552 acoustic features were leveraged, and BayesNet and Random Forest were applied for classification, yielding a final score of 79.31%. In contrast, [PLZ18] focused on MFCC features, using a combination of SVM, RF, KNN, GB, and EL for classification. The achieved accuracies varied, with SVM at 64.95%, RF at 66.45%, KNN at 66.03%, GB at 67.35%, and EL at 68.48%.

In [DEH18], the use of MFCC+delta-delta and MFCC + spectral features was explored, employing NN, RF, and SVM for classification. Notably, the study reported sensitivities of 96.9% and 20.0% for NN and SVM (MFCC + spectral), respectively. The study in [Jul+18] introduced features such as NNE, Cepstral HNR, and GNER MFCC, utilizing GMM and GBT for classification. The study achieved high detection scores for MEEI (99.4%) and SVD (93.2%), with an overall detection score of 72%.

In [IPL18], a transfer learning approach for features was used, employing SVM for classification and achieving a sensitivity of 94.90%. In a separate study, [BK18] used Mel-scaled spectrograms and MFCC features, applying a 5-layer CNN and RNN for classification.

The study reported a sensitivity of 96% and a specificity of 18%. [Ju+18] focused on features such as MFCC (3rd-6th), spectral centroid (SC), spectral flux, and zero-crossing rate (ZCR), employing ATSVM for classification and achieving a UAR of 60.67%.

[Zon+18] explored 28 acoustic parameters as features and employed PCA auto-associative neural networks for classification, reporting accuracies ranging from 86% to 100%. [Mar+18] incorporated features such as MFCC, SWCE, multipeak, and Thomson tapers, employing GM for classification, with Thomson multitaper outperforming in cases of functional and organic dysphonia.

Table 2.2 highlights that some works introduce the use of demographic data, each employing distinct features and classification methods. [She+18] incorporated demographic and symptom-related features, utilizing decision trees, linear network analysis, k-nearest neighbors, support vector machines, and artificial neural networks. Their study focused on demographic aspects such as neoplasm, physical health (PH), and symptoms related to voice pathology, introducing a comprehensive approach to voice disorder distinction.

[Fan+19] investigated the effectiveness of Mel-frequency cepstral coefficients and MFCC with delta features, employing support vector machines, Gaussian mixture models, and deep neural networks. Notably, their research highlighted the superiority of DNN, which outperformed other methods and showcased its potential in distinguishing various pathologies, including nodules, polyps, neoplasms, dysphonia, and sulcus.

Building on this foundation, [Shi+19] further refined their approach by combining MFCC with delta features and utilizing both GMM and DNN for classification. This enhancement resulted in a notable accuracy increase of 2.02% to 10.32%, demonstrating the efficacy of incorporating demographic data for improved classification accuracy, particularly in identifying neoplasms, physical health issues, and voice pathology.

These studies demonstrate various features and techniques used on the FEMH 2018 dataset, highlighting different approaches to voice pathology classification. Despite the simpler task, the 2018 methods primarily employed common machine learning techniques.

In the 2019 “IEEE BigData Cup Challenges” the FEMH voice data challenge focused on a 4-class classification task involving recordings from four disorder categories: functional dysphonia, phonotrauma, laryngeal neoplasm, and vocal paralysis. Participants received training databases, and the ranking of their methods was based on an undisclosed testing set. This challenge expanded upon the 2018 competition, which aimed to differentiate healthy cases from pathological ones involving three voice pathologies: neoplasm, phonotrauma, and vocal palsy. Consequently, the 2019 challenge tackled a more difficult task, with limited prior research to guide participants. The challenge’s difficulty is evident from the low classification accuracy achieved by the top five methods.

Notably, specific algorithmic details about the methods used in the 2019 challenge were not available to the public at the time of writing.

Our deep learning technique achieved fifth place based on the reported results, with a margin of less than 7 percentage points from the top-performing method. We leverage deep learning principles to seamlessly integrate diverse information modalities, encompassing both audio features and medical data, within a bimodal architecture.

Similarly, in the participant architectures, as observed in [Shi+19] in Table 2.2, deep neural networks were also employed to integrate acoustic and medical data. Their approach involved training two standard feed-forward neural networks, whose softmax decisions were then combined within a third network. The feature extraction phase incorporated a Gaussian Mixture Model to statistically represent an audio recording before passing the GMM output to one of the neural networks.

In our methodology, we operate on mid-term segments, generating 2-D representations and employing a convolutional architecture for one of the sub-networks. **What sets our approach apart is the fusion of intermediate learned representations, not just softmax vectors, at a final processing layer. This strategy enables more effective error signal flow during training. Additionally, we incorporate perturbation features alongside the medical metadata within the same sub-network.**

Table 2.1: References about FEMH

References	Features	Classification method	Results (%)
[BK18]	OpenSMILE 6552 acoustic features	BayesNet, Random Forest	Final score of 79.31
[PLZ18]	MFCC	SVM, RF, KNN, GB, EL	Accuracy SVM 64.95, RF 66.45, KNN 66.03, GB 67.35, EL 68.48
[DEH18]	MFCC + delta-delta, MFCC + spectral	NN, RF, SVM	Sensitivity NN SVM (MFCC + spectral) 96.9 20.0
[Jul+18]	NNE, Cepstral HNR, GNER MFCC	GMM, GBT	MEEI 99.4 SVD 93.2 De- tection Score 72
[IPL18]	Transfer learning	SVM	Sensitivity 94.90%
[BK18]	Mel scaled spectrograms and MFCC	5-layer CNN and RNN	Sensitivity 96 Specificity 18
[Ju+18]	MFCC 3rd-6th, SC, Spectral Flux, and ZCR	ATSVM	UAR of 60.67
[Tom+18]	JMFCC 3rd-6th, SC, Spectral Flux, ZCR	TSVM	UAR score of 60.67
[Zon+18]	28 acoustic parameters	PCA auto-associative NN	86 – 100
[Mar+18]	MFCC, SWCE, multipeak, Thomson tapers	GM	Thomson multitaper outperforms functional, organic dys- phonia

Table 2.2: References concerning FEMH dataset with demographic data included

Reference	Features	Classification method	Results	Pathologies
[She+18]	Demogr., Symptom.	DT, LNA, KNN, SVXM, ANN	Demog. Neoplasm, PH, Symptoms VP	Neoplasm, PH, VP
[Fan+19]	MFCC, MFCC + delta	SVM, GMM, DNN	DNN outperforms	Nodule, Polyp., Neoplasm, Dysphonia, Sulcus
[Shi+19]	MFCC + delta, MFCC	GMM, DNN	Accuracy increase 2.02% – 10.32%	Neoplasm, PH, VP

In a subsequent phase of our experiments, we enhanced this architecture by introducing a second computational model that incorporated more sophisticated convolutional branches and innovative augmentation techniques. The architecture is designed with two distinct subnetworks: a fully convolutional branch and a feed-forward branch. At its core, the model is adept at handling audio recordings of varying durations, providing an opportunity for data augmentation through segmenting audio clips of different lengths. Additionally, the model incorporates noise injection techniques, introducing brown, white, and pink noise to augment the dataset.

The following sections will detail our two computational models. The next section describes our methodology, including the database, feature extraction process, and model architecture. After that, we discuss the experimental setup, results, and key aspects of the experiments. Finally, we provide insights and highlight key findings.

2.3 Methodology

This section outlines our research framework and approach using innovative deep learning architectures, emphasizing the methodologies employed for data processing in developing our modular classifier. We begin with an overview of the dataset used in this work, followed by a detailed description of the methodologies, including feature extraction and data augmentation techniques. Additionally, we provide a comprehensive outline of the architecture and methodologies of the two novel deep learning models.

2.3.1 Description of the FEMH 2019 Dataset

The inclusion of the FEMH dataset in our experiments aligns with one of our key objectives: to explore the integration of medical descriptors as an input source for

voice pathology classification. The FEMH 2019 dataset was accessed through participation in the 2019 FEMH Challenge of the 2019 IEEE Big Data Cup. The dataset was obtained from a voice clinic at the Far Eastern Memorial Hospital. After the completion of the challenge, the hospital generously provided the metadata for the testing data as well, enabling us to thoroughly conduct and conclude our experiments. Access was granted with the condition that it be used solely for research purposes. The dataset consists of voice recordings and medical records of patients with four types of voice disorders: hyperfunctional dysphonia, phonotrauma, vocal palsy, and neoplasm. Notably, the dataset does not include recordings of healthy speech.

Each medical folder provides additional context for the analysis, containing 34 demographic questions, both categorical and binary, including age, gender, occupation, habits, symptoms, when the voice became worse, how it happened, history of internal surgery, and the severity of gastroesophageal reflux, among others. The dataset contains 100 voice recordings for each disorder, where a sustained vowel /a/ is pronounced by pathological speakers. The duration of the recordings ranges from 2 to 39 seconds, and the sampling frequency varies among recordings. The training set comprises 50 voice recordings for each disorder, with pathological speakers pronouncing the sustained vowel /a/ across different ages and genders. The competition’s evaluation criterion was classification accuracy, which was computed on a balanced test set of 200 recordings encompassing the four pathology types. Figure 2.1 provides an overview of the main characteristics of the FEMH 2019 dataset.

FEMH 2019 dataset			
400 human subjects exhibiting four types of pathology			
Audio recording of sustained vowel /a/ with varying duration between [2,39] sec		34 Medical descriptors Categorical and binary	
Training dataset of 200 subjects Testing dataset of 200 subjects Four balanced classes			
Phonotrauma 100 samples	Functional dysphonia 100 samples	Vocal palsy 100 samples	Laryngeal neoplasm 100 samples

Figure 2.1: Contents of the FEMH dataset

It is interesting to define the four types of pathology mentioned from a medical perspective. Pathological speech typically refers to speech distortion arising from atypical functioning in the voice and/or articulatory mechanisms due to diseases, illnesses, or other physical or biological disruptions in the production system. Dysphonia, as a disorder of voice production mechanisms in the larynx, exhibits specific perceptual, acoustic, and physical attributes. However, dysphonia constitutes only one among various pathologies, sometimes presenting as a secondary or primary symptom within different disorders. This task encompasses hyperfunctional dysphonia, vocal fold paresis, phonotrauma, and neoplasm.

Dysphonia categorically falls into organic or functional types. Organic dysphonia results from anatomical changes in the vocal folds, such as nodules or benign

tumors, while functional dysphonia lacks known anatomical changes. This work specifically examines a variant of functional dysphonia known as hyperfunctional dysphonia, characterized by excessive involuntary muscle contractions resulting from improper phonation, leading to a hoarse or strained voice [Tei+15].

Vocal hyperfunction describes chronic conditions involving abuse or misuse of the vocal mechanism due to excessive and/or uncoordinated muscular forces, commonly associated with prevalent types of voice disorders. It is believed to be a primary factor in chronic tissue trauma, termed phonotrauma, contributing to the formation of common vocal-fold lesions, such as vocal fold nodules.

Phonotrauma is defined as “trauma to the laryngeal mechanism (vocal folds) as the result of vocal behavior that includes yelling, screaming, and throat-clearing” [Mid07]. It refers to the formation of common vocal-fold lesions (e.g., vocal fold nodules) that affect how the folds vibrate, and its symptoms are similar to those of dysphonia.

Vocal fold paresis or paralysis refers to the situation where one (unilateral) or both (bilateral) vocal folds are paralyzed [SB16]. Paralysis of the voice box muscles manifests as voice changes, including hoarseness, breathiness, increased effort in speaking, heightened air pressure required for normal speech, and diplophonia (a gargling-like voice).

Lastly, the term neoplasm encompasses various cancer types, including tumors affecting the larynx, voice box, or vocal cords, presenting symptoms such as hoarseness, painful swallowing, and fatigue [KA20].

2.3.2 Proposed Feature Representations

In this work, we investigate the efficacy of both audio and medical data. We focus on capturing detailed voice signal features, emphasizing its spectral properties and temporal evolution. To accomplish this, we use MFCC coefficients (Appendix A.2.2) that characterize the instantaneous spectral envelope shape of the speech signal. However, speech signals are time-variant and constantly fluctuate. To capture the evolution of MFCC over time, we also integrate delta MFCC.

This specialized feature vector is constructed by including the initial 13 MFCC, augmented by their corresponding first-order derivatives, alongside the logarithm of the mel-filterbank outputs. This serves as the input for a dedicated branch within the classifiers, implemented as a convolutional or fully convolutional network.

Each audio recording is resampled at a frequency of 44100Hz , followed by normalization of amplitudes within the range of $[-1, 1]$. Subsequently, the signal is parsed using a 40 ms window with a 20 ms hop size. At each frame, we perform a Discrete Fourier Transform, using its resulting coefficients as input for a mel-filterbank. Each filter within this bank executes a weighted summation of DFT coefficients within its designated frequency range. After obtaining the filterbank output, a logarithm is computed for each value, followed by applying a discrete cosine transform to these logarithmic values.

Conventionally, the literature suggests discarding the first MFCC coefficient, as it is often considered to contain non-relevant information. However, contrary to this

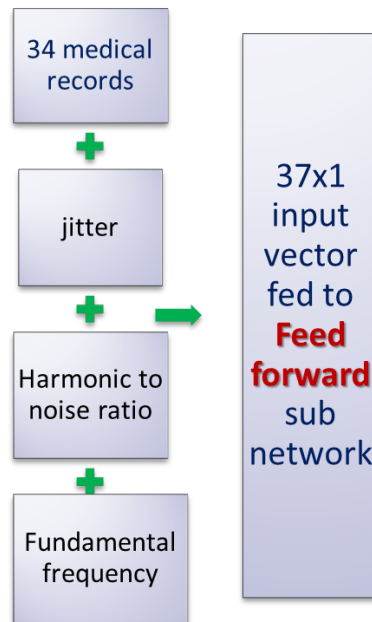


Figure 2.2: Medical/Pertrubation features vector

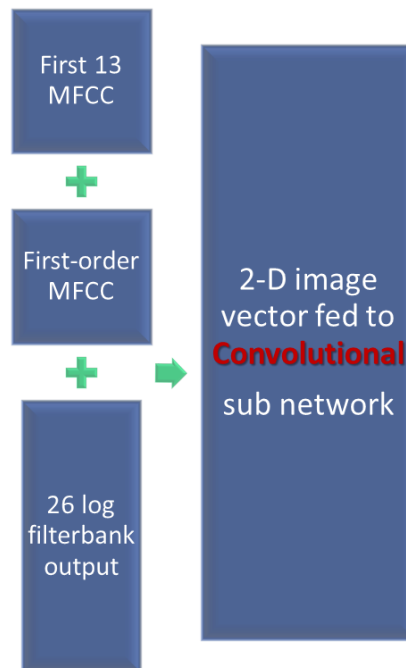


Figure 2.3: Audio features vector

practice, we include the first MFCC in our feature vector. This initial coefficient corresponds to the mean signal intensity. Through experimental evidence, we demonstrate a correlation between signal intensity and certain pathological characteristics.

To capture the signal’s dynamics, we compute the first-order derivative of the MFCC vector over time and append it to the MFCC vector. Augmenting this vector with the logarithm of 26 mel-filterbank outputs results in 52 feature values per frame, forming a sequence of 64 feature vectors per audio recording, as graphically shown in Figure 2.3.

To meet the requirements of our experiments and accommodate the two computational models, we developed two versions of feature vectors. The first version maintains a uniform length by dividing preprocessed recordings into segments of standard size, applying zero padding where necessary. The second version segments recordings into clips of arbitrary lengths, following a customized algorithm designed for our experimental framework. We will provide thorough explanations of these two versions, detailing their methodologies and implementations.

Creating an audio feature vector with fixed length

We employ a short-term feature extraction technique to derive a series of features from the time-domain representation of the audio input signals. To ensure uniformity in processing, the signal undergoes normalization and resampling at a rate of 44100 Hz. To address the variability in recording duration and meet the fixed-size input requirement for our architecture’s first convolutional layer, we implemented a standardized segmentation approach. This method entails generating a sequence of fixed-length segments, each spanning 1.28 seconds, from the recordings. These segments are extracted using a moving window of 40 ms duration with a 20 ms hop size. The feature vector derived from this process is then processed by the convolutional branch of the classifier.

Creating an audio feature vector with an arbitrary length

To meet the specifications of the second computational model, we employ the algorithm detailed in 2.4.3 to split the audio recordings into varying-length segments. The audio recordings in the testing datasets span a range of lengths, and the algorithm dynamically adapts the segmentation length according to the duration of each recording. As a result, the audio feature vectors have varying durations. Each segment is transformed into a 2-D image representation of dimensionality $N \times 64$, with N ranging from 124 to 1462, determined by the duration of the samples. This representation is then processed within the fully convolutional module of the relevant classifier.

2.3.3 Perturbation-medical Input Vector

Constructing the second input vector, termed the “perturbation-medical input vector,” involves combining the 34 medical measurements detailed in the database metadata with the 3 mid-term segment features: fundamental frequency, jitter, and harmonic-to-noise ratio measurements.

The mean fundamental frequency (F_0) (Appendix A.2.1) is computed using the probabilistic YIN (pYIN) algorithm [MD14]. To provide a more sensitive acoustic measurement of vocal function, jitter (A.2.1), which measures cycle-to-cycle variations of the fundamental frequency, is included in the feature set. Harmonic-to-Noise

Ratio (A.2.1) serves as an indicator of voice quality and is estimated as the ratio of periodic signal energy to noise energy in decibels.

Within the medical records, symptom-related responses from patients—such as fatigue or breathiness—are included. These categories are incorporated into the perturbation - medical feature vector, maintaining their numerical representations. Consequently, each human subject contributes to a 37×1 dimensional data vector, with elements normalized within the interval $[-1, 1]$. This resulting vector proceeds through the feed-forward input branch for further processing.

The feature vectors derived from this process undergo further processing within the feature learning modules.

2.4 Proposed Computational Models

In this work, we propose two computational models. The first model is a CNN based approach that utilizes a traditional convolutional neural network architecture to process audio features and medical data. The second model is a fully convolutional network that integrates advanced convolutional techniques to handle audio recordings of varying durations, incorporating data augmentation strategies to improve classification accuracy.

2.4.1 CNN-based Modular Deep Learning Architecture

The proposed classifier integrates two information sources through respective sub-networks, followed by a merging module that leads to the final classification decision. As illustrated in Figure 2.4, the framework is structured with two primary parallel branches, each responsible for processing audio feature vectors and perturbation-medical data, respectively. These branches then converge into a final module, where the prediction is made based on the integrated information from both sources.

The first sub-network is designed with a series of four consecutive convolutional layers (refer to A.1.2), each equipped with specific configurations. These layers are composed of 64, 64, 32, and 32 convolutional filters, each utilizing a kernel size of 3×3 . The outputs from each convolutional operation undergo ReLU activation for non-linearity and are subsequently downsampled using a max pooling layer with a size of 2×2 . The input audio feature vectors, each of size 52×64 , generated from the preprocessing stage, are processed through these specialized convolutional layers.

The second sub-network is formulated as a feed-forward neural network (refer to A.1.1), featuring two hidden layers with 64 and 32 units, respectively, incorporating Rectified Linear Unit activation functions. The 37×1 perturbation-medical input vectors are fed into this fully connected sub-network, ensuring comprehensive integration of the voice data for precise analysis.

Subsequently, the outputs from the aforementioned sub-networks are intelligently merged and directed into a dense layer comprising 1024 nodes, with each node activated by the ReLU activation function. Ultimately, the decision-making process of the system is completed by employing a softmax output layer with four units to generate accurate posterior probability estimations for the four distinct classes representing the four pathological voice disorders under study.

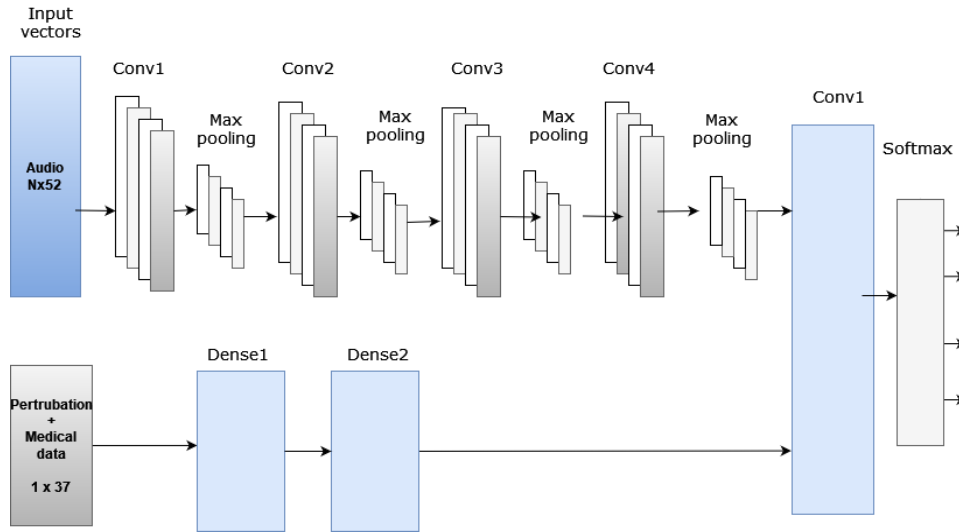


Figure 2.4: Bimodal deep learning CNN-based classifier. The first sub-network consists of four convolutional layers with ReLU activation and max pooling. The second sub-network is a feed-forward neural network with two hidden layers, designed to process the perturbation-medical input vectors.

2.4.2 Fully-CNN based Modular Deep Learning Architecture

The enhanced version of the classifier incorporates architectural modifications to effectively integrate the proposed data augmentation techniques. The core concept is to manage audio recordings of varying durations. This involves a modification within the architecture to accommodate the variability in audio sample lengths, specifically replacing the convolutional sub-network from the earlier design with a fully convolutional sub-network (Appendix A.1.3) in the new proposal. The fully convolutional network architecture, as described in [LSD15], operates effectively on inputs of arbitrary dimensions, generating an output vector of consistent dimensionality.

A typical convolutional neural network configuration often follows a pattern involving convolutional layers, successive average pooling layers, and culminating with fully connected layers. The resulting feature maps must be flattened before being passed through a series of fully connected layers to produce the final prediction. The last fully connected layer requires a fixed number of inputs, which necessitates input images of uniform size.

The network architecture depicted in Figure 2.5 provides a detailed delineation of the layer dimensionalities and hyperparameter values, comprehensively outlined in Table 2.3.

To overcome this limitation, the adoption of fully convolutional networks was necessary. In this approach, the final dense layers of a hybrid convolutional-dense classifier are replaced by a convolutional layer with a kernel size of 1×1 and a stride of 1, followed by a global max pooling layer. This modification ensures an output block with consistent dimensionality, independent of the input image size, and determined by the number of filters n as $1 \times 1 \times n$.

In this implementation, each audio recording is transformed into a one-channel, two-dimensional “image” of size $h \times w$, where h and w are spatial dimensions. The

Table 2.3: Network configuration description: The output of each layer is fed as input to the subsequent layer. ReLU and dropout layers always follow convolutional and fully connected layers. The parameter N lies in the range of $[124, 1462]$, depending on the duration of the audio recording.

Layer	Output shape	Filters	Kernel, stride
Input	$(1 \times N \times 64 \times 1)$	-	-, -
Conv1	$(1 \times (N - 2) \times 62 \times 64)$	64	$3 \times 3, 1 \times 1$
Max pooling	$1 \times \left(\frac{(N-2)}{2}\right) \times 31 \times 64$	-	2
Conv2	$1 \times \left(\frac{(N-2)}{2} - 2\right) \times 29 \times 64$	64	$3 \times 3, 1 \times 1$
Max pooling	$1 \times \left(\frac{\left(\frac{(N-2)}{2} - 2\right)}{2}\right) \times 14 \times 64$	-	2
Conv3	$1 \times \left(\frac{\left(\frac{(N-2)}{2} - 2\right)}{2} - 2\right) \times 12 \times 32$	32	$3 \times 3, 1 \times 1$
Max pooling	$1 \times \left(\frac{\left(\frac{\left(\frac{(N-2)}{2} - 2\right)}{2} - 2\right)}{2} - 2\right) \times 6 \times 32$	-	2
Conv4	$1 \times \left(\frac{\left(\frac{\left(\frac{(N-2)}{2} - 2\right)}{2} - 2\right)}{2} - 2\right) \times 4 \times 32$	32	$1 \times 1, 1 \times 1$
Global Max pooling	$(1 \times 1 \times 32)$	-	-
Dense1	(1×64)	-	-
Dense2	(1×64)	-	-
Dense	(1×128)	-	-
Classifier	(1×4)	-	-

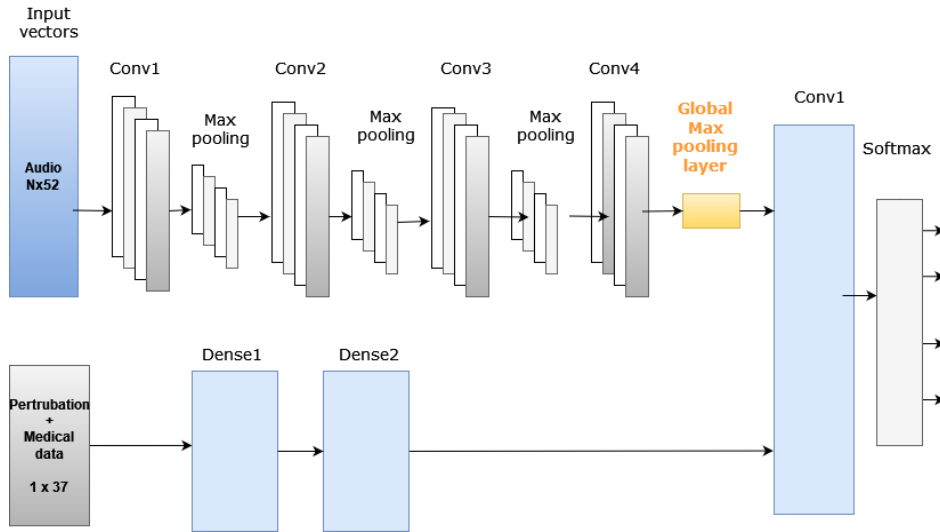


Figure 2.5: Bimodal Fully CNN based classifier. the audio features branch implements a fully convolutional network with four convolutional layers (Conv1, Conv2, Conv3, Conv4) followed by a Global-MaxPooling layer. The perturbation and medical features processing branch consists of two fully connected layers (Dense1, Dense2).

image dimensions depend on the duration of each recording. As described previously, the feature vector is an $N \times 64$ matrix, where N depends on sample duration and lies in the range $[124, 1462]$. The resulting feature vector is fed to the first fully convolutional network branch. Therefore, the input shape of the first convolutional layer is not fixed. As we adopted a batch size of one, the batch shape is eventually $1 \times N \times 64 \times 1$.

The respective branch consists of four consecutive convolutional-max pooling-batch normalization blocks and a final global max pooling layer. The first three convolutional layers contain 64, 64, and 32 convolutional filters, respectively, each with ReLU activation functions and a kernel size of 3×3 with a stride of 1 (without zero padding). The last convolutional layer performs a 1×1 convolution with 32 filters, also with a stride of 1 (without zero padding). The final global max pooling layer subsamples the output.

Conversely, the second branch is composed of a feed-forward neural network with two hidden layers, each containing 128 units, equipped with Rectified Linear Unit activation functions. This sub-network processes the 37×1 input vectors of medical metadata and perturbation features.

The outputs from both sub-networks are concatenated into a dense layer comprising 128 neurons, utilizing ReLU activation functions.

2.4.3 Augmentation Methods

The FEMH database has a significant limitation due to its relatively small training data size, which is a critical factor when training deep learning architectures. This limitation is common among most existing voice pathology datasets. To mitigate this issue, we experimented with various data augmentation techniques. In the following

subsections, we present our methods: segmentation into varying-length audio clips, colored noise injection, and spectrum masking.

Data augmentation with Segmentation into adjustable duration splits

The FEMH training set comprises only 200 recordings across four types of pathology, with durations ranging from 2 to 39 seconds. Although these recordings have varying durations, prevailing methods for voice pathology classification require and process fixed-length inputs. Consequently, fixed-size segmentation and zero-padding procedures are often employed to address this disparity in duration. However, it is important to acknowledge that using fixed-size segmentation and zero-padding is a simplified approach that may not fully capture the nuances and variations present in the recordings. This is primarily because recording duration often correlates with patients' disabilities in pronouncing specific vowels, potentially leading to suboptimal classification results.

To address the constraints posed by the limited training data and mitigate the drawbacks of fixed-length segmentation and zero-padding techniques, we propose a novel augmentation method. This method involves extracting multiple segments per recording, as illustrated in Figure 2.6. We standardized the length of all segments derived from a recording to two-thirds of the recording's length, but with different endpoints. As recording lengths vary, the entire set of extracted segments inevitably includes segments of varying durations as well.

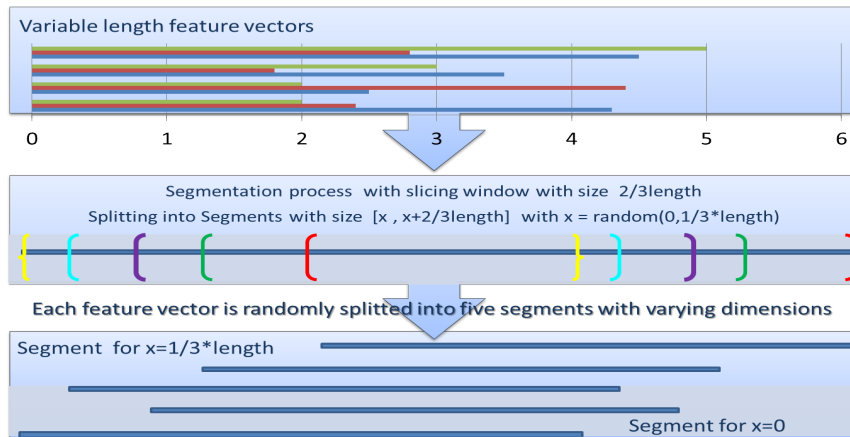


Figure 2.6: Feature sequence segmentation.

- *Step 1*: Define a random starting point, st_k , between zero and one-third of the recording's length, measured in frames, i.e.,

$$st_k \in \left[0, \frac{L_i}{3}\right]$$

where L_i is the number of frames of the i -th feature sequence ($L_i \in [124, 1462]$ in our experiments). We refer to frame numbers because it is assumed that a moving window technique is applied on the recording during a subsequent feature extraction stage, yielding a feature sequence per recording.

- *Step 2*: Define an endpoint, en_k , as:

$$en_k = st_k + \frac{2 * L_i}{3}$$

- *Step 3*: Repeat the previous two steps five times, i.e., $k = 1, \dots, 5$, yielding five segments starting from random positions within the recording, while ensuring that each segment length is equal to two-thirds of the length of the recording (measured in number of frames).

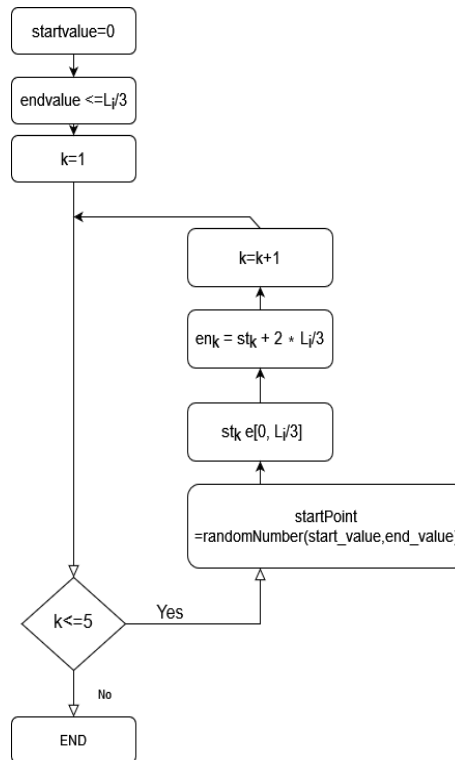


Figure 2.7: Segmentation flowchart

The flowchart of the method is given in Figure 2.7, and the respective pseudocode description is presented below:

Algorithm 1 Segmentation algorithm

```

startvalue ← 0
endvalue ← random in [0, ≤ Li/3]
k ← 1
while k ≤ 5 do
  startpoint ← random in [startvalue, endvalue]
  stk ← startpoint
  enk ← stk +  $\frac{2 * L_i}{3}$ 
  k ← k + 1
end while
  
```

The values mentioned for endpoint and segment duration selection were determined through a grid experimental study conducted alongside the measurement of the classifier’s performance.

Data augmentation with noise injection

In addition to segment-based augmentation, we also explore noise injection techniques. It is important to note that other widely used time-domain methods for dataset augmentation include time warping, pitch shifting, and dynamic range compression [Lar02; NH08].

Time warping methods alter the duration of an audio signal by stretching or compressing it while minimally affecting its fundamental properties. In contrast, pitch-shifting techniques adjust the pitch of an audio recording up or down without changing its overall length. Dynamic range compression reduces the range of amplitude variation in the audio signal.

Given the nature of the task, it is important to note that changes to signal duration, pitch, or amplitude would likely affect key sound properties needed to differentiate between various voice pathologies. For this reason, we opted to exclude time warping methods and instead focused on experimenting with noise injection techniques alongside the segmentation-based augmentation procedure.

Incorporating background noise is a well-established technique for reducing overfitting and improving model generalization [An96; HK92]. In our approach, we apply noise injection directly to the input signal during the training phase, rather than altering neural network parameters such as layer activations, weights, or gradients.

In our experimentation, we explore the use of Gaussian noise, as recommended in [HK92; TK95]. Additionally, we investigate colored noise, which has different power spectrum densities, by generating white, pink, and brown noise signals using established techniques (Appendix A.2.3).

We produce an augmented training dataset in which each recording undergoes random corruption by one of the noise types mentioned above. Below, we provide a description of the method used:

- For each audio recording in the training set:
 - Generate a white noise corrupted signal w .
 - Generate a pink noise corrupted signal p .
 - Generate a brown noise corrupted signal b .
 - Insert signals w , p , and b in the training set.

Noise injection is applied before the segmentation technique during the training stage. Batches are formed through random selection from the final augmented training set. The impact of the data augmentation techniques on classification performance will be presented in the next section.

Data augmentation with spectrum masking

We further experimented by incorporating SpecAugment [Par+19] as an additional data augmentation technique. SpecAugment modifies the spectrogram by warping it

in the time direction, masking blocks of consecutive frequency channels, and masking blocks of utterances in time. These augmentations were chosen to help the network become robust against deformations in the time direction, partial loss of frequency information, and partial loss of small segments of speech in the input. SpecAugment deforms the audio feature input vector through time warping, frequency masking, and time masking.

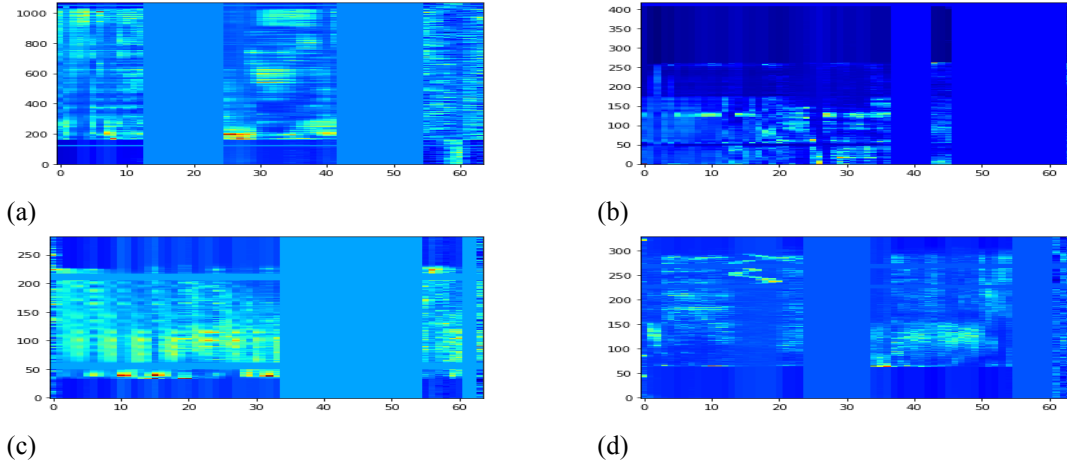


Figure 2.8: The audio feature image undergoes augmentation by warping along the time axis and applying masks to multiple blocks of consecutive time steps (vertical masks) and mel frequency channels (horizontal masks). The masked segments of the image are highlighted in purple for emphasis.

We applied the same technique to the image created from the audio features, where MFCC-derived features are extracted. An illustration of the produced augmented samples is depicted in Figures 2.8. This involved deforming the image through time warping, frequency masking, and time masking. To implement this augmentation policy, we created a function that follows these steps:

1. **Time Warping:** Implemented using the `sparseimagewarp` function in TensorFlow. Time warping randomly shifts a point along the horizontal line passing through the center of the image within a defined range. The range is determined by the time warp parameter W , where W represents the maximum distance of the warp. Six anchor points on the boundary are fixed to maintain stability.
2. **Frequency Masking:** This technique involves masking consecutive mel frequency channels. First, a number of consecutive channels f is selected from a uniform distribution ranging from 0 to the frequency mask parameter F . The starting channel f_0 is then chosen from the range $[0, \nu - f)$, where ν represents the total number of mel frequency channels.
3. **Time Masking:** Similar to frequency masking, time masking masks consecutive time steps. A random number of time steps t is chosen from a uniform distribution between 0 and the time mask parameter T . The starting time step t_0 is selected from the range $[0, \tau - t)$, where τ denotes the total number of time steps.

Additionally, an upper bound is introduced on the time mask to ensure that it cannot exceed a certain proportion of the total number of time steps, denoted by p .

2.5 Experiments and Results

In this section, we present the experimental setup, methodologies, and outcomes of our study aimed at developing a robust voice pathology classifier. Our approach involved comprehensive experimentation with model optimization, feature integration, and performance evaluation across four distinct models, as previously detailed.

Neural Network Architecture Optimization

Initially, our experiments focused on exhaustive testing of neural network architecture optimization. This included fine-tuning model parameters such as layer configurations, activation functions, and learning rates to enhance the classifier's efficacy.

- **Layer Configurations:** We experimented with various depths and widths of the neural networks to determine the optimal structure.
- **Activation Functions:** Different activation functions (e.g., ReLU, sigmoid, tanh) were tested to identify the most effective one for our data.
- **Learning Rates:** A range of learning rates was explored to ensure efficient convergence during training.

Feature Enrichment and Integration

Our initial experiments focused on network design and feature enrichment. We augmented the standard Mel-frequency cepstral coefficients vector with its derivatives and integrated mid-term perturbation features. This approach aimed to capture both spectral nuances and temporal evolution in speech signals.

- **Spectral Features:** Augmentation with MFCC derivatives allowed the model to better capture the dynamic changes in the speech signal.
- **Temporal Features:** Mid-term perturbation features provided additional temporal context to the model.

Additionally, incorporating metadata from medical records enriched the feature set, enabling a multimodal approach to classification. We investigated various data integration strategies, including unimodal and bimodal models, data augmentation techniques, and performance analysis across diverse pathologies.

Model Refinement and Augmentation Techniques

Subsequently, we aimed to enhance the established architecture through refined network configurations and problem-specific augmentation techniques. We explored a range of model setups and augmentation methods, including:

- **Segmentation:** Divided audio files into smaller segments to increase the amount of training data.

- **Noise Injection:** Added various types of noise to the training data to improve model robustness.

These methods notably improved classification accuracy.

Multimodal Approach and Fusion of Features

We examined the impact of different modalities, demonstrating significant performance gains when fusing medical and audio features. By integrating both types of data, the model could leverage the strengths of each modality, resulting in improved overall performance.

- **Unimodal Models:** Models trained using only audio features or medical records.
- **Bimodal Models:** Models trained using a combination of audio features and medical records.

This comprehensive analysis across modalities, features, and modular deep neural network architectures yielded valuable insights into optimizing classifier performance for voice pathology classification. These insights will be detailed in the subsequent subsections.

In the following subsections, we delve deeper into the specific results of each experimental phase, providing quantitative metrics and visualizations to illustrate the performance improvements achieved through our methodological advancements.

2.5.1 CNN-based computational approach

In this subsection, we provide a comprehensive analysis of our proposed bimodal classification framework. Our primary objective is to discern the individual and collective contributions of audio and medical information to the classification task. Furthermore, we conduct experiments across a range of deep learning neural network architectures to rigorously evaluate their efficacy in addressing the specific challenges of our classification task.

Ablation study

To determine the optimal architectural configuration, we conducted tests of deep neural network architectures and feature representations. Our experimentation encompassed various model architectures, including recurrent neural networks such as Long Short-Term Memory, as well as 1-D and 2-D convolutional networks.

Simultaneously, we tested various feature sets, evaluating combinations of MFCC, delta MFCC, filterbank outputs, and the logarithms of these features. The inclusion of the first MFCC was tested against common practice, and it was found that removing the first MFCC led to better results. This systematic approach enabled us to capture both spectral and temporal characteristics crucial for discriminating between different voice pathologies. These feature sets were processed either as time series with 1D audio features or as images with 2-D feature dimensions, utilizing the alternative network architectures.

Furthermore, our investigation extended beyond model architectures and feature sets. We delved into hyperparameter tuning, exploring various aspects such as zero-padding techniques (both left and right), dropout rates, batch sizes, number of layers,

kernel sizes, window lengths, step sizes, activation functions, loss functions, and training algorithms. Each of these parameters played a critical role in shaping the capacity and performance of the model.

Final Configuration

Our approach enhanced the standard MFCC vector by incorporating its first-order derivative and integrating it with mel-filterbank outputs to capture both spectral shape and temporal evolution. Additionally, mid-term perturbation features were combined with metadata from medical records.

The ultimate network architecture, described in detail in Figure 2.4, was employed for the results presented in the following tables.

The classifier was trained for 300 epochs using the Adam gradient descent algorithm to optimize the cross-entropy loss function. The learning rate was set to 0.001, and a 5-fold cross-validation setup was implemented. An early-stopping criterion was used, triggered when the loss value did not significantly decrease between epochs. To counter overfitting, dropout regularization was applied, with a dropout rate of 0.5 for the fully connected layer and 0.1 for each convolutional layer.

The outcomes detailed in this work are based on a cross-fold validation scheme using only the challenge's training data. This approach was necessary due to the lack of public access to the challenge's testing dataset.

Our methodology involved segmenting recordings into fixed-duration mid-term segments (approximately 1.3 seconds) and applying right zero-padding to very short segments. This process increased classification accuracy in our cross-fold validation setup to 71.5%. Notably, the method submitted to the challenge used only the initial 5 seconds of each recording (or less, depending on the recording's length), which resulted in lower performance in our experimental setup, as detailed in this study.

Experimental Results of the Challenge-Submitted Model

Our primary aim was to assess the impact of incorporating medical record data into the system. To achieve this, we conducted experiments using four different system configurations. Initially, we excluded the convolutional sub-network and evaluated only the efficiency of perturbation features, resulting in a classification accuracy of 46.5%. Subsequently, an analysis focusing solely on medical records achieved a classification accuracy of 58.5%. Next, by combining perturbation features with medical records, the classification accuracy improved to 62.5%. Finally, the integration of MFCC within a convolutional setup, serving as an additional information source, led to a further enhancement in performance, achieving an accuracy of 65.9%. These findings, outlined in Table 2.4, underscore the efficacy of integrating metadata from medical records with low-level signal descriptors and selected mid-level features, yielding the most effective feature combination.

In accordance with the guidelines of the 2019 FEMH voice data challenge, our system achieved a commendable classification accuracy of 65.9% when evaluated on the publicly available training set. However, when tested on the undisclosed testing set, our performance decreased to 57%, placing our method fifth based on the officially announced results. This decline in performance suggests a potential issue of overfitting, which we intend to explore and address in the subsequent section.

Table 2.4: Results for different input data combinations

Input features	Validation Accuracy
Perturbation features	46.5%
Medical Records	58.5%
Medical Records and perturbation features	62.5%
Medical Records, Perturbation, and MFCC	65.9%

Additionally, a significant challenge posed by the FEMH dataset was its limited training data, comprising only 200 recordings across all four examined pathologies. To address this challenge, we adopted practices recommended by previous studies, which have shown that introducing noise to a network can significantly improve performance generalization in certain scenarios [An96]. Therefore, to artificially expand the available training dataset, we implemented noise injection solely at the input vectors, without affecting layer activations, weights, gradients, or outputs.

Following the common approach of applying Gaussian noise [HK92], this noise addition was exclusively implemented during the training phase and omitted during model evaluation. This approach resulted in a slight enhancement, boosting the model’s classification accuracy to 74.9%. Table 2.5 details the results.

Table 2.5: Results with noise injection

Segmentation	Noise injection	Validation Accuracy
Applied	Applied	74.9%
Applied	Excluded	71.5%

2.5.2 Fully CNN based Computational Model

In this subsection, we conduct a thorough evaluation of the second fully CNN based implementation of the proposed bimodal classification performance through a series of new experiments. We aim to provide an in-depth analysis of the network fine-tuning process and present the enhanced version of the network. Our objective was to dissect the individual contributions of both audio and medical modalities, assess the effectiveness of various classifier architectures, refine the model’s architecture, and evaluate the influence of data augmentation techniques on overall system performance. We placed emphasis on leveraging fully convolutional layers and fully connected layers within our novel end-to-end deep learning approach.

A key aspect emphasized in our study is the need for voice pathology classification models to adapt to varying durations of audio recordings. Traditional fixed-size segmentation approaches and zero-padding procedures may overlook crucial information encoded in sustained utterance duration and intensity, which often reflects the disorder itself. Therefore, our research challenges these conventional methods, advocating for the incorporation of original and potentially diverse durations of audio recordings within the model architecture.

Recognizing the significance of sustained utterance duration and intensity in voice pathology classification, our aim is to contribute to a more comprehensive and effective approach for addressing the complexities of this domain. To thoroughly assess the classifier’s capabilities, we conducted three distinct sets of experiments, each delving into essential aspects. These experiments were designed to unravel the intricacies of our model, providing insights into its strengths, weaknesses, and interpretability.

In the first phase of our experiments, we focused on the efficacy of unimodal classifiers, emphasizing the audio modality. Two distinct architectures—a conventional segmentation-based approach and a fully convolutional model—were compared, revealing insights into the importance of processing audio signals at their original duration. Building on these findings, we extended our analysis to bimodal classifiers, investigating the synergy between audio and medical data. This exploration highlighted the significance of integrating medical parameters and perturbation features in achieving superior classification accuracy.

To enhance the robustness of our models, we delved into the impact of data augmentation techniques. Through systematic experiments involving segmentation-based augmentation and various forms of noise injection, we identified optimal strategies for improving testing accuracy. The results not only validated the effectiveness of our proposed augmentation methods but also highlighted the nuanced interplay between data augmentation and classification performance.

Further granularity was added to our analysis through a detailed examination of the confusion matrix, offering valuable insights into the model’s behavior across different pathologies. This assessment provided a nuanced understanding of classification imbalances, guiding potential refinements to our framework.

Finally, we explored the interpretability of our network by providing visualizations of intermediate feature layers. This qualitative analysis aims to demystify the inner workings of our model, offering transparency into the learned patterns and functional contributions of each layer.

Ablation study

An ablation study was performed to assess the impact of the growth in size and complexity of the network architecture. We conducted an extended set of experiments where components of the network were removed or replaced to measure their effect on the system’s performance. We validated the network’s dimensionality by experimenting with the number of layers in the two sub-networks and the number of filters in the convolutional layers.

This table summarizes the effects of varying hyperparameters, including the number of convolutional layers, fully connected layers, optimizer, activation functions, and learning rates, on the classification accuracy of the model. Adjustments made to these hyperparameters were evaluated to determine their impact on the classifier’s performance.

When we reduced the number of convolutional layers to 3, classification accuracy decreased. Specifically, with a configuration featuring three convolutional layers and the number of filters set to 64, 64, 64, the classification accuracy was 58.5%. However, by setting the number of filters to 64, 64, 128, the accuracy improved to 60.5%,

Table 2.6: Experiments with Convolutional and Fully Connected Layer Dimensions on Classification Accuracy

Architecture	Configuration	Accuracy
CNN	3 Layers, 64-64-64 Filters	58.5%
CNN	3 Layers, 64-64-128 Filters	60.5%
CNN	3 Layers, 128-128-64 Filters	60.6%
CNN	4 Layers, 64-64-128-128 Filters	60.0%
CNN	4 Layers, 64-64-64-64 Filters	63.1%
CNN	5 Layers, 64-64-64-64-64 Filters	58.0%
Fully Connected Layers	FCN (1024 nodes)	49.0%
	FCN (512 nodes)	51.0%
Optimizer - Activation Function	SGD + Sigmoid	53.5%
	ReLU Activation Function	51.5%
Learning Rate	0.1	48.0%
	0.01	52.0%
	0.001	59.0%

and setting the number of filters to 128, 128, 64 resulted in an accuracy of 60.6%. Further increasing the number of layers and filters led to a drop in classification accuracy. For example, with four convolutional layers and filters set to 64, 64, 128, 128, the accuracy was 60%, and with 64, 64, 64, 64, the accuracy increased to 63.1%. However, adding an extra layer with 64, 64, 64, 64, 64 filters led to a drop in accuracy to 58%.

We also experimented with the size of the fully connected layers. Adding an extra layer with 512 nodes significantly decreased the accuracy to 51%, and reducing the number of neurons in the fully connected layer from 1024 to 512 resulted in a performance drop to 49%. Finding the best set of hyperparameter values, learning rate, and loss function was a crucial part of the final model composition.

Through our hyperparameter optimization, we observed that when using Stochastic Gradient Descent optimizer, the performance of the classifier decreased, with a testing accuracy of 63%. Our ablation studies also involved investigating alternative activation functions. When using the sigmoid function, the classification accuracy dropped to 53.5%, and with tanh, a further decrease was observed to 51.5%. In the final configuration, the Rectified Linear Unit was used as the activation function.

To validate the importance of the learning rate, we trained the model with three alternative learning rates: 0.1, 0.01, and 0.001, with testing accuracy scores of 48%, 52%, and 59%, respectively.

All network configurations were trained using a 4-fold cross-validation scheme, with three folds used for training and one for validation at each run. The FEMH training dataset contained 200 audio recordings, and the testing dataset also contained 200 audio recordings. With the 4-fold cross-validation training scheme, 150 samples were used for training, and 50 were used for validation. Augmentation methods were applied only to the training subset, resulting in a balanced set of 750 audio clips with 175 samples per pathology class.

The models were trained for a maximum of 300 epochs using the Adam gradient descent algorithm with a learning rate of 0.0001 while observing the validation error. An early-stopping criterion of 5 epochs was used to restrict training times. The

categorical cross-entropy loss was used to compute the error signal, and validation accuracy was used as an auxiliary metric. To prevent overfitting, a dropout regularization scheme was adopted with a dropout value set to 0.5 for the convolutional and dense layers. A data generator was used to create batches of one image (recording) at a time to address the training requirements of the fully convolutional branch.

Additionally, normalization layers were included to reduce internal covariance shift, defined as changes in the distribution of network activations due to changes in network parameters. We experimented with four alternative normalization approaches: batch normalization [IS15], layer normalization [BKH16], weight normalization [SK16], and instance normalization [UVL16]. We also evaluated the contribution of different activation functions (hyperbolic tangent, sigmoid, and ReLU). Initially, we constructed a network architecture without any normalization layers and with sigmoid activation functions. This network configuration yielded a testing accuracy of 49%. An alternative configuration without normalization layers and using tanh as the activation function yielded an improved testing accuracy of 58.9%. The use of a weight normalization layer decreased classification accuracy to 57%. Instance and batch normalization achieved nearly identical testing accuracy scores of 63%. In the final system configuration, we adopted a scheme with batch normalization and ReLU activation functions.

The results indicated that the best performance was achieved with batch normalization and ReLU activation functions.

Experimental Results

Unimodal Classification Analysis

In this subsection, we analyze the contributions of individual modalities to the overall classification accuracy. Our experiments began with the evaluation of a unimodal classifier processing two-dimensional representations of audio signals through a convolutional branch. Two versions of the classifier were compared, focusing on the segmentation of the audio recordings.

The first version employed a standard segmentation process, extracting fixed-length segments (1.28 seconds) from audio recordings for training. This version utilized a standard convolutional classifier. In contrast, the alternative architecture employed a fully convolutional design, replacing the dense layer with a global max-pooling operation. This innovative approach allowed for the analysis of each recording at its original duration without prior segmentation.

The comparison between these versions, detailed in Table 2.7, revealed a significant performance difference. The fully convolutional model outperformed the standard convolutional version, achieving a testing accuracy of 48.5% compared to 36.5%.

Bimodal Configuration Impact

The results, as illustrated in Table 2.7, demonstrate a notable improvement in classifier performance when integrating both audio and medical data compared to using only medical and perturbation features. Specifically, the classifier that relied solely on medical and perturbation features achieved an accuracy of 47.0%. In contrast, the overall system that incorporated both audio and medical data reached a classification accuracy of 54.5%. This represents an approximate 8% increase in accuracy,

Table 2.7: Classification accuracy of unimodal and bimodal classifiers

Input	Model	Accuracy
Audio features	Convolutional	36.5
Audio features	Fully convolutional	48.5%
Medical data, perturbation features	Fully connected	47.0%
Audio and medical data	Overall system	54.5%

underscoring the significant impact of combining medical and audio modalities. This improvement highlights the effectiveness of the bimodal approach in enhancing classifier performance.

Data Augmentation Impact

In the next step, we experimented with the impact of various data augmentation techniques on classification performance. All previously mentioned models were trained without employing data augmentation techniques. The results of the augmentation methods on classification performance are detailed in Table 2.8. It is evident that segmentation-based augmentation enhances testing accuracy to 57.8%. This improvement is achieved by extracting five audio segments per recording using the algorithm outlined in subsection 2.4.3.

We then experimented with noise injection, exploring diverse power spectrum distributions of the added noise signals. Initially, we introduced a standard Gaussian layer at the input of the fully convolutional branch within the bimodal classifier, generating additive zero-mean Gaussian noise with a standard deviation of 0.1. This layer was activated only during the model’s training phase. Remarkably, the presence of additive Gaussian noise contributed to a notable enhancement in testing accuracy, achieving 58.5%.

Subsequently, we tested the impact of injecting colored noise. As outlined in Table 2.8, the introduction of white, pink, and brown noise separately resulted in classification accuracies of 59.5%, 60.0%, and 62.0%, respectively. Brown noise, which targets lower frequencies, influenced the initial 13 MFCC coefficients of the feature vector. This focus on low frequencies improved classification robustness compared to other noise types, possibly explaining why brown noise yielded superior results.

Finally, we constructed an augmented training set where each recording was randomly corrupted by one of the noise types. This configuration yielded the most favorable outcomes, with a classification accuracy of 64.4%. Notably, this performance surpasses that of the top classifier in the 2019 FEMH voice data challenge (63.0%). Within the noise injection techniques, distinct impacts were observed:

- The introduction of Gaussian noise resulted in a moderate increase, raising the accuracy to 58.5%.
- Injecting colored noise, specifically brown noise, had the most significant impact, achieving a classification accuracy of 62.0%.
- Augmented training, merging segmentation with colored noise injection, resulted in the best outcome, showcasing a testing accuracy of 64.4%.

These findings underscore the pivotal role of data augmentation in enhancing the model’s classification performance. The best results were achieved when sequence segmentation and the injection of three colored noise types were simultaneously applied during training.

Table 2.8: Classification accuracy with respect to different augmentation techniques

Segmentation	Noise injection	Testing accuracy
No	None	54.5%
Yes	None	57.8%
Yes	Gaussian	58.5%
Yes	Pink	60.0%
Yes	White	59.5%
Yes	Brown	62.0%
Yes	All colors	64.4%

After extensive experimentation, we settled on the following parameters: $W = 40$, $T = 30$, $F = 13$, $mT = 2$, and $mF = 2$. In this context, W represents the time warp parameter, while T and F denote the time and frequency masking parameters, respectively. Additionally, mT and mF signify the number of time and frequency masks applied. This method was employed to augment our original training data, resulting in an augmented dataset 120 times larger than the original.

Utilizing this augmentation method, we further enhanced our dataset by splitting the deformed SpecAugmented audio feature input vectors into varying-length segments. When SpecAugment was applied independently, without the addition of any noise, the classification accuracy reached 61.5%. However, when SpecAugment was combined with the augmented training dataset that included all noise variations, the classification performance saw a notable improvement, achieving an accuracy of 66.3%. This result significantly surpasses the classification accuracy of the top-rated method in the 2019 FEMH voice data challenge (63.0%). This outcome highlights the effectiveness of SpecAugment as a data augmentation technique, particularly when used in conjunction with other augmentation methods such as noise injection.

Classifier Performance and Confusion Matrix

Table 2.9: Confusion matrix of the best-performing classifier

Class	Dysphonia	Phonotr.	Neoplasm	Vocal Palsy
Dysphonia	38	3	4	5
Phonotrauma	4	37	0	9
Neoplasm	15	7	18	10
Vocal Palsy	6	5	3	36

For a more comprehensive understanding of the top-performing classifier’s performance across the four pathology types, we generated the confusion matrix (Table 2.9), where each element (i, j) represents the count of testing samples with the true label in the i -th class and the predicted label in the j -th class.

Examining the matrix, it is evident that hyperfunctional dysphonia (first row) displays high class recall, with errors distributed fairly uniformly across the other classes. However, phonotrauma (second row) shows a slightly different pattern, with most false predictions assigned to the vocal palsy class, and no misclassifications into the neoplasm class. The third row reveals the challenge in correctly identifying neoplasm disorder, with only eighteen neoplasm samples accurately classified and errors distributed almost evenly across the remaining classes. In terms of class precision and recall, vocal palsy behaves similarly to dysphonia.

In summary, the classification performance appears imbalanced across the four pathology types, with dysphonia attracting a majority of the errors (resulting in low class precision) and neoplasm pathology exhibiting the lowest recall.

Network Interpretation

To fulfill our analysis, we focus on interpreting the functionality of the intermediate layers of the network. Model explainability is widely acknowledged to be important in the field of healthcare [Fan+21a]. In this section, we provide insights into the functionality of the intermediate feature layers of the proposed network architecture and the patterns learned during the training stage.

Our analysis initially concentrated on the functionality of the convolutional layers. We visualized the 2-D representation at the network’s input, along with feature activation maps for its four convolutional layers. These visualizations revealed similarities among input images of different pathology classes, showcasing common patterns and their variations, contributing to the complexity of the classification task. Representative images for vocal palsy, hyperfunctional dysphonia, phonotrauma, and neoplasm are displayed in Figures 2.10a, 2.11a, 2.12a, and 2.13a respectively. Furthermore, Figures 2.10b to 2.13b illustrate a subset of 32 activation maps from the fourth convolutional layer, showcasing the maximum activation output for each convolutional filter.

To explore in greater detail the performance of the fully connected layers, we employed a post-hoc interpretability technique akin to [Jas+15]. This method extracted insights directly from each layer through the activation values of all neurons. For each of the three fully connected layers, we represented the activation weights of all neurons (x-axis) with respect to the layer’s sequence of thirty-seven medical-perturbation input features (y-axis), resulting in three images shown in Figures 2.9. Specifically, Figure 2.9a demonstrates the weights of the 64 neurons in Dense Layer 1 corresponding to the 37 input features. Linear patterns horizontally aid in identifying dominantly activated neurons and their associated high-contributing features. Features indexed at 14, 15, 16, 25, 26, and 27 hold substantial importance in the model’s predictions, signifying patients’ symptoms such as dysphonia, dryness, lumping, occupational vocal demands, hypertension, and head and neck cancer. Conversely, features in the regions $[0, 4]$ and $[20, 25]$ evoke the lowest activation values, indicating a minor role in the network’s inference capabilities. These features predominantly encompass patients’ demographic information like sex, age, onset, tiredness, night meal, choking, eye dryness, smoking, and drinking.

Figure 2.9c displays the final fully connected layer, Dense3. This layer processes a concatenation of audio and medical-perturbation embeddings, encompassing 32 features derived from the fully convolutional module and the remaining 64 from the fully connected branch. The visual representation illustrates the weight values of the Dense layer’s 128 neurons. Analysis of the graph indicates a contrast between the upper and lower regions, where the upper segment (corresponding to feature indices 0 – 30) demonstrates lower weight activation values, while the lower part exhibits higher activation values. The lower region represents the feature embeddings generated by the fully connected branch. This highlights the importance of the model’s second module, which learns from medical data. Including medical features significantly improves the model’s ability to classify correctly.

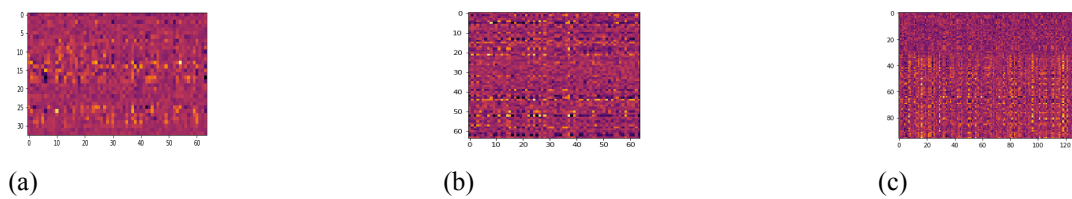


Figure 2.9: Visualization of all weights of the three fully connected layers.

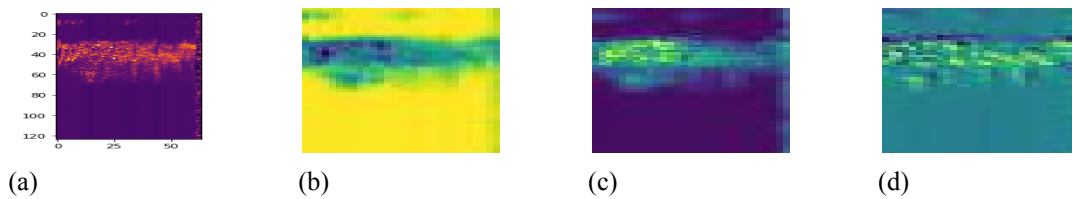


Figure 2.10: Vocal palsy: input “image” along with three feature activation maps of the final convolutional layer.

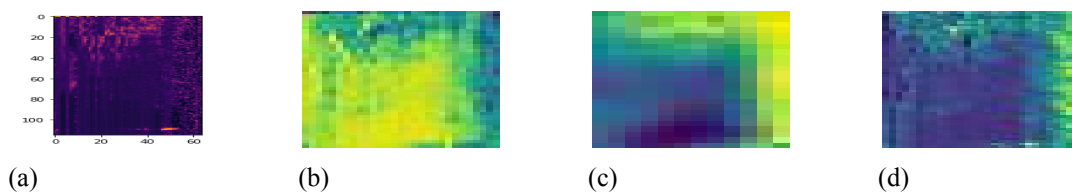


Figure 2.11: Hyperfunctional dysphonia: input “image” along with three feature activation maps of the final convolutional layer.

In summary, to discern the significance of processing audio signals at their original duration, two versions of a unimodal classifier were compared. The first version followed a standard segmentation procedure, while the second employed a fully convolutional architecture, enabling analysis of each recording at its original duration. Results in Table 2.4 showcased the superiority of the fully convolutional model, achieving a testing accuracy of 48.5% compared to 36.5% for the conventional method. This highlighted the importance of preserving the original duration of audio signals.

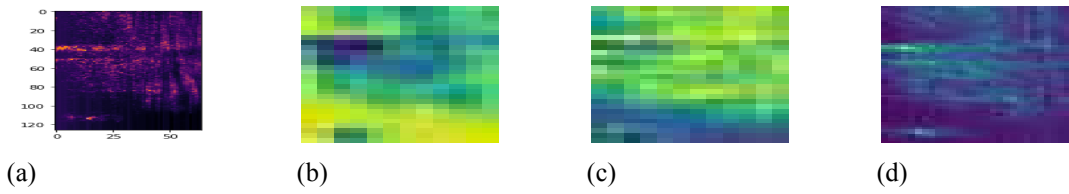


Figure 2.12: Phonotrauma: input “image” along with three feature activation maps of the final convolutional layer.

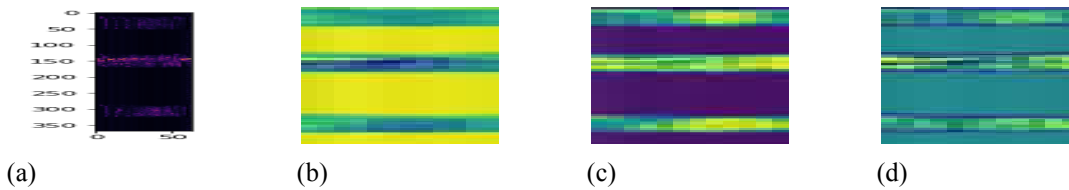


Figure 2.13: Neoplasm: input “image” along with three feature activation maps of the final convolutional layer.

The study further investigated the impact of medical parameters and perturbation features on overall system performance. Two configurations were tested—one relying solely on a fully connected branch processing medical information and perturbation features, and the other representing the complete system with all input modalities. The fusion of audio and medical modalities yielded a significantly improved classifier, achieving a classification accuracy of 54.5%, outperforming the medical/perturbation processing branch by almost 8% (Table 2.4).

A comprehensive exploration of data augmentation techniques was conducted to evaluate their impact on classification performance. Segmentation-based augmentation, noise injection, and variations in power spectrum distributions were investigated. Results in Table 2.8 demonstrate that the simultaneous application of sequence segmentation and three types of colored noise injection during training yields the highest testing accuracy of 64.4%, surpassing the best classifier from the 2019 FEMH voice data challenge (63%).

The performance of the best classifier was analyzed across four types of pathology using a confusion matrix (Table 2.8). Notably, dysphonia attracted the most errors, and neoplasm pathology exhibited the lowest recall. This imbalance highlights the challenges in classifying different pathologies.

The study delves into network interpretation, offering insights into the functionality of intermediate feature layers. Visualization of convolutional layers and feature activation maps provided an understanding of the learned patterns. Analysis of the fully connected layers emphasized the importance of medical features, underscoring their contribution to the model’s classification capability.

This comprehensive experimentation demonstrates the efficacy of a multi-modal



Figure 2.14: Perceptually uniform sequential color map inferno

approach in voice pathology classification. The study not only emphasizes the importance of processing audio signals at their original duration but also showcases the positive impact of fusing medical and audio modalities. The incorporation of data augmentation techniques further enhances the model's robustness, leading to state-of-the-art classification accuracy.

2.6 Conclusions

Based on our experimental study, we have demonstrated that a meticulously designed neural network architecture can effectively classify voice pathologies, even with limited training data. We propose a multimodal classification framework tailored for the four types of voice pathology outlined in the FEMH dataset: hyperfunctional dysphonia, phonotrauma, laryngeal neoplasm, and unilateral vocal paralysis. Through experiments involving the fusion of audio-based features and medical descriptors, we have verified the utility of medical parameters as supplementary information for voice pathology classification.

By treating MFCC-derived features and medical record data as distinct input sources within a bimodal neural network architecture consisting of two sub-networks, we have developed a competitive approach within the context of the 2019 FEMH challenge.

Our investigation also addressed the variable lengths of speech recordings and explored the impact of data augmentation techniques on classification performance. In the enhanced version of our classifier, our study emphasized the importance of processing audio recordings at their original, potentially varying durations in voice pathology classification tasks. This finding challenge conventional approaches that employ fixed-size segmentation schemes and zero-padding, highlighting the significance of retaining information related to sustained utterance duration and intensity, which are often affected by voice disorders.

This network architecture facilitated the introduction of a novel data segmentation algorithm where segment size is adjustable and varies based on the size of each recording. The application of this algorithm has been demonstrated as an effective data augmentation technique.

Additionally, our exploration of alternative data augmentation techniques underscored the effectiveness of injecting three types of colored noise (white, pink, and brown) for voice pathology classification tasks, demonstrating their capacity to enhance classification accuracy. Ultimately, spectrum augmentation techniques have also been shown to further improve system performance.

Chapter 3

Multimodal deep learning classifier with EGG processing capabilities for voice pathology classification

3.1 Introduction

Through our bibliographical research, it has become evident that Electroglottography can significantly enhance the diagnosis of voice pathologies. In this study, we explore this research area and evaluate the efficacy of EGG signals in distinguishing various types of laryngeal and vocal fold pathologies. More specifically, our research focuses on the utility of EGG-based indices as an additional source of information, particularly when combined with other data sources such as audio signals and medical parameters.

EGG is a non-invasive technique that measures electrical impedance across the larynx during vocal fold vibration, providing valuable insights into vocal fold function and dynamics, making it a useful tool in assessing voice disorders. Electroglottography involves placing two external electrodes on the patient's neck to pass a current that captures changes in the larynx's electrical impedance due to vocal fold motion during phonation. These changes manifest as variations in current intensity, resulting in the EGG signal—a time-varying, one-dimensional representation of this intensity [Bak92]. The EGG waveform reflects the events of glottal opening and closing, allowing for measurement of the vocal fold contacting and de-contacting phases over time, which directly supports our study's objective.

Researchers have explored EGG's potential in clinical contexts such as reflux, chronic cough, multiple sclerosis, and Parkinson's disease. Notably, EGG has proven effective in detecting and evaluating various types of dysphonia, including muscle tension dysphonia, spasmodic dysphonia, and vocal fold paralysis. Studies have demonstrated its reliability in dysphonia detection and in monitoring progress during treatment and recovery. Clinical research on electroglottography also extends to vocal fold physiology, with a focus on conditions such as vocal fold nodules and polyps, Reinke's edema, and laryngitis. Additionally, studies on glottal flow models using EGG provide deeper insights into the physical characteristics of vocal fold vibrations, contributing to advanced diagnostic techniques for laryngeal pathologies.

In this chapter, we explore the integration of Electroglottography with audio recordings and medical data to enhance voice pathology classification. Our approach employs a multimodal deep learning framework that leverages diverse datasets from the

Saarbruecken Voice Database. This database is particularly well-suited for our study, as it includes comprehensive patient records, audio recordings, and EGG signals—facilitating nuanced analysis across multiple modalities.

Our research focuses on a four-class classification task to distinguish among hyperfunctional dysphonia, laryngitis, vocal cord polyps, and general dysphonia. We utilize innovative EGG representations, such as wavegrams and examine the efficacy of various neural network architectures in processing these features. Findings from our experiments suggest that integrating EGG data significantly enhances the accuracy of voice pathology diagnostics, highlighting the potential of our advanced deep learning solutions.

3.1.1 Data Resources and Research Gaps

In voice pathology research, most studies predominantly rely on audio recordings of sustained vowels. Although Electroglottography is well-established across various scientific fields, its medical application is often questioned and deemed unreliable when used as a standalone source. However, some studies have introduced bimodal classifiers that integrate both audio and EGG signals. To the best of our knowledge, no EGG-based studies to date have incorporated more than two modalities or combined EGG with medical descriptors.

Existing studies underscore the importance of electroglottographic signals in distinguishing voice pathologies, yet a crucial question remains: what is the correlation between EGG-derived features and other sources of voice pathology information? Specifically, can EGG be effectively combined with audio features to enhance classification accuracy, and could this integration be extended to a trimodal model that incorporates medical data? This inquiry investigates whether integrating EGG data with medical information could enhance the comprehensiveness of voice pathology diagnostics.

Can the fusion of EGG with other medical information advance not only voice pathology diagnosis but also the broader field of medical informatics? Does integrating EGG-derived features with other modalities, such as audio recordings and medical records, enrich datasets and offer a more diverse feature set for machine learning models? Ultimately, can this trimodal integration improve the accuracy of voice pathology classification?

Most of the proposed time-domain glottal flow models are characterized by five parameters: fundamental frequency (f_0), voicing amplitude (A_v), open quotient (O_q), asymmetry coefficient (a_m), and return (r). Additionally, wavegrams have been introduced to assess vocal fold contact and de-contact events over time. Other normalized metrics include the quotient of contact by integration (Q_{ci}), amplitude-normalized peak derivative (QD), and the index of contacting (I_c). Furthermore, the Open Quotient (OQ) and Closed Quotient (CQ) quantify vocal fold contact, and the Quasi Open Quotient (QOQ) has been tested for diagnosing functional dysphonia.

This introduces a new area of research: identifying which EGG-derived indices are most indicative and effective for deep learning processing in voice pathology diagnosis.

3.1.2 Research objectives and contributions

The research aims to address the scientific inquiry of whether the integration of Electroglottographic data with audio and medical data can effectively enhance the efficacy of voice pathology classifiers. This study focuses on enhancing voice pathology classification through the implementation of a multimodal deep learning architecture that integrates multiple sources of information. The primary objective in this phase of our research is to refine the precision and efficiency of the classification process by incorporating EGG signals as a third modality, alongside audio recordings and medical parameters. Our proposition involves integrating audio recordings, medical records, and EGG signals within a modular deep learning architecture. This novel approach represents a significant advancement in voice pathology classification, extending beyond traditional audio descriptors and categorical medical data.

Additionally, the research delves into the question of which EGG-derived features are most relevant and which neural network architectures are most effective for their processing. In this context, we explore the utilization of CQ indices and EGG waveforms processed by Convolutional networks of 1D or 2D convolutions.

To conduct these experiments, selecting an appropriate voice pathology database was essential. Among the available options, the Saarbruecken Voice Database proved particularly suitable, as it includes audio recordings, patient records, and EGG signals. Specifically, we performed experiments on a four-class classification task using sustained /a/ phoneme recordings from the SVD dataset. This four-class problem, initially addressed by [Hem17], involves audio recordings from both healthy individuals and patients with three types of laryngeal voice disorders: hyperfunctional dysphonia, laryngitis, and recurrent laryngeal nerve paralysis.

Dysphonia includes various voice issues such as difficulty maintaining the voice, vocal fatigue, changes in pitch, hoarseness, reduced volume, less vocal efficiency, and weakened endurance during speech. It can be caused by either organic causes, involving anatomical changes in the vocal folds, or functional causes, where no specific anatomical changes are identified. This work examines a variant of functional dysphonia known as *hyperfunctional dysphonia*. Characterized by excessive involuntary muscle contractions due to improper phonation, hyperfunctional dysphonia results in a hoarse or strained voice [TF15].

Laryngitis denotes an inflammation of the mucous membrane lining the larynx, resulting in swelling and inflammation of the tissues beneath the epiglottis. This swelling impacts the area surrounding the vocal cords, disrupting their normal vibration and often causing hoarseness or temporary speech loss

The experiments also include recurrent laryngeal nerve (RLN) paralysis, a condition affecting the voice box muscles essential for breathing, voice production, and swallowing. RLN paralysis accounts for the majority of vocal fold paresis/paralysis cases. Both paresis and paralysis of these muscles lead to complications in aspiration and swallowing [Fra+15], resulting in voice alterations such as hoarseness, breathiness, increased effort in speech, elevated air pressure needed for normal conversation, and diplophonia (a gargling-like voice).

To address this task, we propose an architecture that integrates audio descriptors, categorical medical data, and features derived from electroglottographic signals within a modular deep learning framework. This approach enables parallel processing

of three distinct information sources: short-term audio features, mid-term audio features, medical descriptors, and 'wavegrams' of glottal chords. The decision-making process, facilitated by a learning stage, involves concatenating outputs from three individual sub-networks into a final fully connected layer. Specifically, a cascade of convolutional layers processes sequences of short-term audio feature vectors, including Mel-Frequency Cepstral Coefficients, their derivatives, and Mel filter-bank outputs. Medical records and mid-term audio features are processed by a standard feed-forward branch, while "wavegrams" of the glottal chords are processed by a third convolutional branch.

This task was initially addressed in [Hem17], where thirty-five quantitative voice parameters were analyzed using auto-associative neural networks, achieving an overall classification accuracy of 87.5%. What distinguishes our approach is the inclusion of the EGG signal as a third source of information within a trimodal deep learning framework. This framework integrates EGG signals with speech signals and medical records, enhancing the accuracy of classification decisions for various types of voice pathology. Additionally, we introduce a sophisticated neural network architecture and employ advanced data augmentation techniques to further improve classification performance.

In summary, our architectural framework is designed based on three fundamental principles: trimodal analysis, integration of EGG data and EGG innovative feature representations. This work introduces the following:

a). Trimodal Analysis: Combining short-term audio features, mid-term audio features, and medical descriptors to provide a comprehensive understanding of voice pathology.

b). Integration of EGG data: Utilizing EGG signals as a crucial data source alongside audio and medical information.

c). Processing of innovative EGG representations: This study is the first to integrate the new representation of EGG features, named "wavegrams", an image representing the movement of glottal chords as a new modality.

d). Innovative deep learning architecture: The proposed architecture integrates audio descriptors, categorical medical data, and features derived from EGG signals within a modular deep learning framework. This architecture involves parallel processing of three distinct information sources, enhancing the system's accuracy and decision-making process.

To provide an overview of this chapter's structure, the following section (3.2) delves into related works, offering a comprehensive review of existing literature to contextualize our study within the broader research landscape. We present an overview of key studies conducted on the SVD dataset, with a specific focus on those investigating EGG-related features. Additionally, we introduce the dataset (Section 3.3), describing its composition, characteristics, and rationale for selection.

In Section 3.3.1, we elaborate on feature selection and deep learning architecture design, outlining the methods used to extract audio, medical, and EGG features from multiple information sources, as well as the design principles underlying our modular deep learning framework. This section details data preprocessing, the feature selection phase, and the development of the deep learning network architecture tailored to our specific task.

Furthermore, Section 3.3.4 comprehensively details our experimental setup, experiments, and results. Here, we describe the experimental specifics, including model fine-tuning, training and evaluation procedures, metrics used for performance assessment, and the outcomes of our experiments. We analyze the findings, discuss their implications, and provide insights into the effectiveness of our approach in addressing challenges in automatic voice pathology classification.

Finally, Section 3.4 consolidates the main findings of this study and proposes directions for future research. We also reflect on the broader implications of our findings.

3.2 Related work

This section provides a comprehensive overview of methodologies employed in voice pathology classification, covering both conventional and deep learning approaches. A primary focus is placed on the Saarbruecken Voice Database, the central dataset for this study’s experimental phase. The section is divided into two parts: one discusses audio-based classification tasks using the SVD, and the other explores studies that incorporate EGG data within this dataset.

The subsection on audio recordings reviews various techniques and algorithms used in this field, covering both traditional algorithms and neural network architectures essential for accurately classifying voice pathologies. These methods are crucial for preprocessing raw audio data, extracting key features, and applying classification algorithms or neural networks to deliver precise diagnostic and classification results.

In contrast, the subsection on EGG data examines how EGG signals have been integrated into voice pathology classification frameworks. This part highlights a range of indices derived from EGG signals, including closed quotient measures, temporal characteristics, and spectral EGG representations.

By providing a comprehensive overview of both conventional and deep learning methodologies, along with an in-depth exploration of audio processing and EGG data integration, this section lays the groundwork for our subsequent analysis and experimentation.

3.2.1 Research Contributions in SVD Dataset

The SVD dataset has been utilized in numerous experiments addressing diverse voice pathology classification tasks. A comprehensive review by [SRH20a] highlights common voice disorders and emphasizes the frequent use of traditional classifiers, particularly Support Vector Machines, in voice classification. Specifically regarding the SVD database, studies such as [Ahm+18] and [FM21] employ SVM classifiers, while others like [S+21] explore alternatives, including Naïve Bayes, decision trees, and ensemble classifiers.

Various classification methods have been applied across studies to distinguish between normal and pathological voice conditions. Alnasheri (2017) combined the Fisher Discriminative Ratio (FDR) with SVM to classify normal versus pathological samples, using datasets such as AVPD, MEEI, and SVD [Al+17]. Similarly, González (2012) applied Linear Discriminant Analysis for dimensionality reduction,

alongside SVM, to classify conditions like chronic laryngitis, cysts, Reinke's edema, and spasmodic dysphonia, focusing on the continuous vowel /a/ [Gon+12].

In 2014, Alnasheri used Gaussian Mixture Models and SVM independently to identify voice pathologies in English and German databases, specifically MEEI and SVD [Al+14]. By 2020, Syed reinforced the prominence of traditional methods in voice pathology classification by applying conventional classifiers across the same databases [SRH20b]. That same year, Hamdi used Hidden Markov Models with features like HFCC-NHR and HFCC-HNR to recognize voice pathologies in the MEEI and SVD databases [RSA20].

Going back to 2005, Behroozmand compared neural networks with SVM for classifying vocal fold edema, nodules, and polyps using entropy features, reflecting an ongoing assessment of machine learning techniques in this field [BA05]. Recently, SVM classifiers, Naïve Bayes, decision trees, and ensemble classifiers have been noted for their effectiveness in detecting vowel pathology within the SVD dataset, as shown in studies by Alnasheri (2018), Fethi (2021), and Syed (2021) [Ahm+18; FM21; S+21]. Additionally, Souissi et al. employed Mel-Frequency Cepstral Coefficients and their derivatives alongside SVM and LDA for pathology detection on the continuous vowel /a/ from the SVD [SC15].

Recent studies explore the potential of deep neural networks, with notable advancements in architectures such as Convolutional Neural Networks, Recurrent Neural Networks, and Long Short Term Memory networks. The application of deep neural networks for voice pathology classification in the SVD dataset has been investigated in several studies, including [Dar+15], [NA16a], [Pav+17], [MK18], [Vic+19], [Sid+21], and [Ahm+20]. Artificial neural networks, first introduced for this task in [Dar+15] and [NA16a], have since been further developed in subsequent studies [Pav+17], [MK18], [Vic+19], [Sid+21], [Ahm+20], and [Rib+23a].

Transitioning from conventional methods to deep learning models, Harr et al. [Pav+17] introduced a neural network architecture involving convolutional layers, recurrent LSTM layers, and fully connected layers for binary classification of healthy and pathological voices based on raw audio signals. Meanwhile, Moon et al. [MK18] pioneered the incorporation of higher-order statistics (HOS) and a deep neural network on SVD recordings to detect cysts, paralysis, and polyp pathologies.

Broadening the perspective, Syed et al. [Sye+21b] and [Sye+21a] contributed to the exploration of convolutional and recurrent neural networks for various voice pathologies, achieving significant accuracy rates through feature extraction from neural networks. Similarly, Fan et al. [Fan+21b] explored FC-SMOTE to create balanced classes for pathological voice types and normal/pathological samples, testing both standard classifiers and deep learning models.

These studies focused on leveraging neural network architectures, such as Convolutional Neural Networks, Recurrent Neural Networks, and Long-Short-Term Memory networks, for various voice pathology classification tasks. Additionally, Hemmerling et al. [Hem17] conducted a quantitative analysis by computing various voice parameters and employing autoassociative neural networks for classification. Building on this, Harr et al. [Pav+17] used a neural network architecture with convolutional layers, recurrent LSTM layers, and fully connected layers to classify healthy versus pathological voices using raw audio signals.

Table 3.1: Summary of Studies Utilizing Conventional Methods on the SVD Dataset

Study	Methodologies and Classifiers Employed	Voice disorders
[BA05]	Comparing Neural Networks and SVM using entropy features	Vocal fold Edema, Nodules, Polyps
[Gon+12]	Linear Discriminant Analysis with SVM	Laryngitis, Cyst, Reinke Edema, Spasmodic Dysphonia
[Muh+16]	Support Vector Machine based on statistical features	Adductor, vocal nodules, keratosis, vocal fold polyp, paralysis
[Al+17]	Fisher Discriminative Ratio with SVM	Normal vs. Pathological
[Hem17]	Auto-associative neural networks for voice parameters	Hyper.dysphonia, RLNP, laryngitis, healthy
[Ahm+18]	Support Vector Machine classifiers	Not specified Pathology
[HHC20]	Hidden Markov Model classifiers with HFCC features	Various Voice Pathologies
[FM21]	Support Vector Machine	Laryngitis, cyst, non-fluency syndrome
[S+21]	Naïve Bayes, decision trees, ensemble classifiers	Laryngitis, cyst, dysphonia
[Al+14]	Gaussian Mixture Model and SVM	Cysts, paralysis, polyp
[SC15]	MFCC, LDA and SVM	Chronical laryngitis, cyst, Reinke edema, spasmodic dysphonia

Furthermore, Moon et al. [MK18] incorporated higher-order statistics (HOS) and a deep neural network with CNN recordings, aiming to detect cysts, paralysis, and polyp pathologies with an impressive accuracy of 87.4%.

Shifting to a broader perspective, Syed et al. [Sye+21b] introduced a system utilizing convolutional and recurrent neural networks, achieving high accuracy rates for various voice pathologies based on neural network-extracted features. Similarly, in Syed et al. [Sye+21a], multiple classifiers, including CNN and RNN, were used to classify laryngitis, cysts, non-fluency syndrome, and dysphonia based on distinct features extracted from neural networks. Additionally, Guedes et al. [Gue+19] developed Long-Short-Term Memory and Convolutional Network models for classification using embeddings extracted from continuous speech phrases.

The following two tables provide summaries of studies conducted on the SVD dataset that utilized both conventional and neural network architectures.

Table 3.2: Summary of Studies Utilizing Neural Network Architectures on the SVD Dataset

Study	Methodologies Employed	Diseases
[Dar+15]	CNN, RNN, and LSTM	Hyperfunctional dysphonia, functional dysphonia, laryngitis, vocal cord paralysis
[NA16a]	Artificial neural networks and SVM in MFCC features.	Chronic laryngitis, Cyst, Reinke's edema, Spasmodic dysphonia
[Pav+17]	Convolutional layers, recurrent LSTM layers, and fully connected layers	Healthy and Pathological Voices
[MK18]	Incorporated higher-order statistics (HOS) and a deep neural network	Cysts, Paralysis, Polyp
[Vic+19]	Long-Short-Term-Memory and Convolutional Network models	Dysphonia, laryngitis, paralysis of vocal cords and healthy voices.
[Sid+21], [S+21]	Convolutional and recurrent neural networks	Various Voice Pathologies
[Rib+23a]	Neural network architectures	Full set of pathologies
[Hem17]	Autoassociative neural networks	Hyperfunctional dysphonia, recurrent laryngeal nerve paralysis, laryngitis and healthy
[MK18]	Incorporated higher-order statistics and a deep neural network on SVD recordings	Cysts, Paralysis, Polyp
[S+21]	Showcased significant accuracy rates achieved through feature extraction from CNN and RNN neural networks	Balbuties, Dysphonie, Frontolaterale Teilresektion, Funktionelle Dysphonie, Vox senilis, Zentral-laryngale Bewegungsstörung, ReinkeÖdem, Stimmlippenpolyp, Stimmlippenkarzinom, Spasmodische Dysphonie, Psychogene Dysphonie, and Leukoplakie

3.2.2 Research Contributions with EGG signals on SVD dataset

In our search for new methods, we thoroughly analyzed the role of electroglottographic signals in classifying voice pathologies. Many studies have used the SVD dataset, particularly incorporating EGG data. Numerous reports highlight the importance of EGG signals in complex voice studies, leading to various proposed feature variations. Recent studies have explored traditional representations of Electroglottographic signals, particularly spectrograms or Mel spectrograms, as effective approaches to analyze vocal fold vibrations and extract features for voice pathology detection and classification. In [Gen+21], Mel spectrograms from both speech and EGG signals are utilized, with a pre-trained CNN extracting sound state and vocal cord vibration features, followed by LSTM processing. Similarly, [Ghu+17] employs co-occurrence matrices and GMM for voice pathology assessment, using separate GMM-based classifiers for voice-only and EGG-only signals.

In most studies, convolutional neural networks have proven effective in distinguishing healthy and pathological voices, with studies like [Isl+22] and [IRT22] showing improved accuracy through the integration of EGG signals. Furthermore, [Gen+21] integrates EGG data with pre-trained CNN and LSTM networks, leading to improved classification results. Fusion of EGG data processed by CNNs significantly boosts classification performance while [Ksi+23] employs a two-level classifier based on a combined CNN–RNN architecture for effective voice pathology detection. In [Isl+22], two CNNs are employed to discriminate between healthy and pathological voices, demonstrating higher accuracy in identifying pathological voices with speech signals and in categorizing pathology types with EGG signals. The work presented in [OMO22] utilizes a framework with two parallel CNN to extract deep features from voice and EGG signals, combined with classical handcrafted features. Feature selection is applied to enhance the feature set, and a SVM classifier is employed for pathology detection.

Several studies focus on exploring advanced features derived from EGG signals. Most proposed time-domain glottal flow models typically involve five parameters: fundamental frequency (f_0), voicing amplitude (Av), open quotient (Oq), asymmetry coefficient (am), and return (r) [Her20]. An innovative approach was presented in [HFŠ10] with the introduction of wavegrams to assess vocal fold contact and de-contact events over time. Additionally, normalized metrics like the quotient of contact by integration (Qci), amplitude-normalized peak derivative (QD), and index of contacting (Ic) are outlined in [Ter19]. Other studies, such as [Jes+18], introduce metrics like the Open Quotient and Closed Quotient to quantify vocal fold contacting, while the Quasi Open Quotient in [Aga+18] is used for diagnosing functional dysphonia. Additionally, [CAO23] presents a novel approach that leverages phase plots extracted from EGG and acoustic signals for automatic voice disorder detection, utilizing convolutional neural networks along with filter-bank features.

These studies leverage the SVD, audio, and EGG signals, employing a variety of features and classifiers. We chose to experiment on a multimodal basis, particularly focusing on advanced EGG features, where EGG audio signals are treated as visual representations (images).

3.2.3 Foundational Concepts

Electroglottographic signals, waveform analysis

Electroglottography, initially introduced by Fabre in 1957 under the term “high frequency glottography”, saw significant advancements with Fourcin and Abberton’s development of a novel laryngograph in 1971 [Fou71]. Subsequent progress led to the creation of the electroglottograph, pioneered by Baken in 1992 [Bak92]. EGG is a non-invasive technique used for the indirect visual examination of vocal fold vibrations by measuring electric impedance between two electrodes placed against the skin over each thyroid lamina. By detecting impedance changes resulting from vocal fold movements, EGG generates a weak high-frequency electrical signal.

In EGG, a low-amplitude, high-frequency current passes between two electrodes placed on either side of the thyroid cartilage at the vocal fold level. A low-voltage, low-intensity electric current ($V < 0.5V$ and $I < 10\text{ mA}$) with high frequency (f between 0.3-5 MHz) flows through these electrodes, while the neck acts as a variable resistor in this constant current circuit. The oscillation induced by laryngeal tissue movement leads to time-varying changes in vocal fold contact, inducing variations in electrical impedance across the larynx. Consequently, these impedance changes cause fluctuations in the current between the two electrodes, which are directly proportional to the relative vocal fold contact area during phonation.

This method operates on the principle that electrical impedance across the neck varies with the extent of contact between the vocal folds during the glottic cycle. When the vocal folds are in full contact (contacting phase), low impedance values allow a higher electric current to flow through the glottis. Conversely, as vocal fold contact decreases (decontacting phase), increased air impedance across the glottic plane causes significant variation in current flow, resulting in a reduction in voltage through the neck tissues. These voltage fluctuations, which occur during the vocal fold contact and detachment phases of phonation, form the basis of the EGG signal.

Introducing EGG based indices

In the broader context of voice research, EGG contributes significantly to advancing our understanding of fundamental aspects of vocal fold dynamics and human voice physiology. Building on EGG’s importance in voice investigations, various feature adaptations from a medical standpoint have been proposed, particularly within time-domain glottal flow models. These models typically include five parameters: fundamental frequency, voicing amplitude, open quotient, asymmetry coefficient, and return [BdD01]. Additionally, studies such as [Jes+18] introduced metrics like the Open Quotient and Closed Quotient to quantitatively assess vocal fold contact extent. The diagnostic efficiency of the Quasi Open Quotient for functional dysphonia was also examined in [Aga+18]. Furthermore, [HFŠ10] introduced wavegrams to evaluate vocal fold contact and de-contact events. Normalized metrics, including the quotient of contact by integration, amplitude-normalized peak derivative, and the index of contacting, were later defined in [Ter19].

Contact quotient

A widely utilized index based on the Electroglottographic signal is the EGG contact quotient, which is also employed in this research. This parameter primarily quantifies the duration of vocal fold contact within each glottal cycle. Initially introduced by [RG72], the concept was further refined in [Lec77] through analysis involving excised larynges and EGG measurements, and has since been referred to as the “larynx closed quotient” [How95] or “contact quotient” [Ori91].

The timing of glottal opening and closure is critical for quantitative voice source analysis, as identifying these events helps determine the proportion of glottal closure within a vibratory cycle. The closed quotient involves calculating the ratio between the closed phase and the entire cycle duration. Commonly used in vocal fold vibration analysis, the Contact Quotient and Open Quotient are widely applied parameters in voice production and pathology studies. These measures are defined as follows:

$$CQ = \frac{\text{Duration of Glottal Contact}}{\text{Total Glottal Cycle Duration}} \quad (3.1)$$

$$OQ = \frac{\text{Duration of Glottal Opening}}{\text{Total Glottal Cycle Duration}} \quad (3.2)$$

These equations quantify the proportion of time during the glottal cycle that the vocal folds are in contact and the proportion of time they are open.

It is important, however, to exercise precision when calculating CQ, as it is sensitive to the algorithm used for identifying glottal contact and de-contact instances [SSD98].

To calculate the Contact Quotient, estimating glottal closure and opening instants typically involves two primary methods:

Threshold Criterion Method on Locally Normalized EGG Signal:

- This method applies a threshold criterion to the locally normalized EGG signal, as proposed by Rothenberg and colleagues in their 1992 study on multichannel EGG systems [Rot92].
- The threshold criterion is applied to the EGG signal to identify points of glottal closure and opening based on user-defined thresholds.

Derivative-Based Method (dEGG):

- This approach detects positive and negative maxima in the first mathematical derivative of the EGG signal, dEGG.
- The derivative-based method leverages intrinsic properties of the EGG signal, eliminating the need for arbitrary user input and allowing for more accurate identification of glottal closure and opening instances.

A hybrid approach combines elements of both methods:

- Hybrid Method:
 - The hybrid approach utilizes the positive peak in the dEGG signal to determine the glottal contact event.

- For the de-contacting event, a threshold—often set at a specific value—is applied to identify the de-contacting instance.

This hybrid approach combines the intrinsic accuracy of the derivative-based method for contact determination with the flexibility of threshold-based detection for de-contacting events.

Given the documented deviations in computed Contact Quotient values depending on the algorithm used, as reported by Herbst in 2006 [HT06], selecting the appropriate methodology is essential for accurate analysis. To minimize computational errors, we estimate CQ values using five known algorithms, following the implementation in [HFŠ10]. The figure below provides a schematic representation of the vocal fold vibratory cycle and the computation of the Contact Quotient.

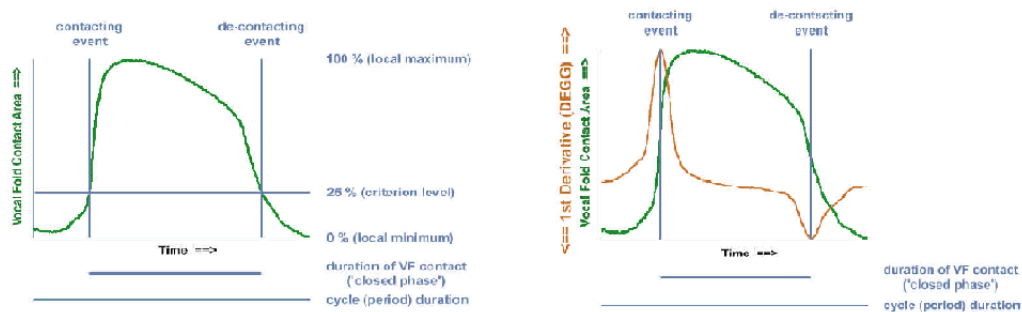


Figure 3.1: Schematic representation of vocal fold vibratory cycle

Wavegram: A Graphical Representation of EGG Waveforms

An innovative algorithm has been developed to visually capture dynamic changes in EGG waveforms over time, condensing this information into a single image called a wavegram, as introduced in [HFŠ10]. This process involves transforming an EGG signal or segment into a graphical representation, where time is depicted on the x-axis, consecutive glottal cycles progress along the y-axis, and the locally normalized vocal contact area is encoded on the z-axis as color intensity.

- Detection and Separation of EGG Cycles:
 - Glottal cycles are identified in quasi-periodic phonation through auto-correlation analysis.
 - An ideal EGG waveform template is cross-correlated with the actual EGG signal to define the beginning of each glottal cycle.
- Normalization and Color-Coding:
 - Each extracted glottal cycle is locally normalized in amplitude.
 - Amplitude values are encoded into monochrome color information, with higher values corresponding to darker colors.

- Cycle-Concatenation and Normalization of the Final Display:
 - Color-coded strips representing individual glottal cycles are rotated 90° counter-clockwise.
 - The height of each cycle plot corresponds to the period duration.
 - Heights are normalized, allowing for the representation of the entire phonation in a single image, known as the EGG wavegram.

Alternatively, the first derivative of the EGG signal, DEGG, can be used to generate wavegrams. This approach provides a more detailed visualization of rapid changes in the vocal fold contact area, enhancing the precision of the analysis.

The resulting wavegrams offer a comprehensive visual overview of vocal characteristics, enabling detailed examination of events within glottal cycles. These two alternative representations are depicted in the subsequent figures, referenced as Figure 1.2.

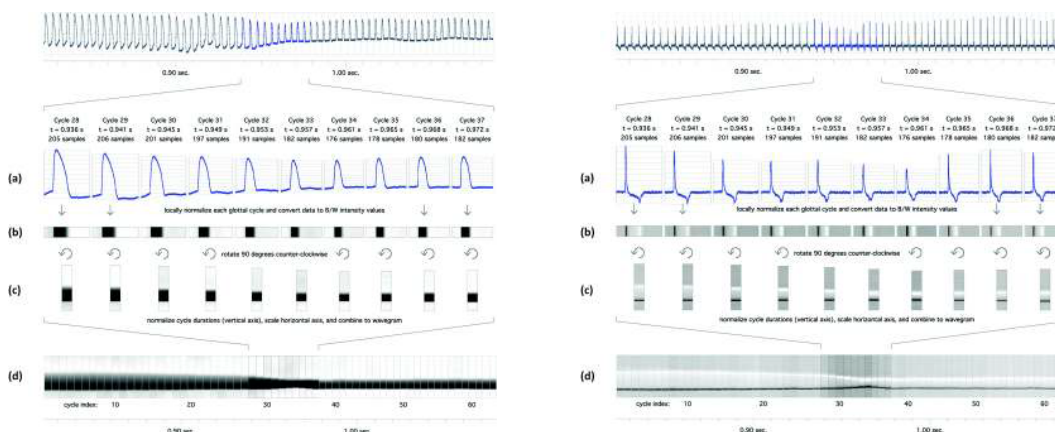


Figure 3.2: Illustration of the wavegram

3.3 Methodology

In this experimental phase, our focus was on exploring the potential for enhancing voice pathology classification accuracy through the integration of a third input source—specifically, EGG signals. This study evaluated various experimental setups, with particular attention to the contribution of EGG features when used alongside other modalities, such as audio recordings and medical data.

Dataset Description and Preprocessing

To assess the effectiveness of EGG signals in voice pathology classification, we conducted experiments using the SVD dataset. The SVD encompasses a collection of pathological voice recordings from over 2000 human subjects with 71 different types of pathology. It contains more than 2000 voiced samples of sustained /a/, /i/, and /u/ vowels, as well as the sentence “Guten Morgen, wie geht es Ihnen?”. These samples come from 687 healthy subjects and 1354 pathological subjects with one or more of the 71 pathologies. All recordings are sampled at 50 kHz with a resolution of 16 bits,

and the audio durations range from 1 to 3 seconds. Additionally, the dataset includes EGG signals from patients, providing an opportunity to investigate the role of EGG signals as an alternative source of information for voice pathology classification.

Within the scope of this study, we focused on a previously explored classification problem, concentrating on a subset of the SVD that includes sustained vowel /a/ speech recordings from both genders. This subset was carefully assembled to include recordings from healthy individuals as well as patients diagnosed with four types of voice disorders: hyperfunctional dysphonia, laryngitis, vocal fold nodules, and recurrent laryngeal nerve paralysis. The recordings range in duration from 0.5 to 3 seconds, with a sampling rate of 50 kHz and 16-bit resolution per sample, with each class containing 687, 140, 204, and 210 subjects, respectively. It is important to note that our study only included recordings of subjects for which corresponding EGG signals were available, resulting in a smaller data subset than that described in Hemmerling et al. (2017) [Hem17], where each class contained 694, 154, 217, and 219 speech recordings for the same task.

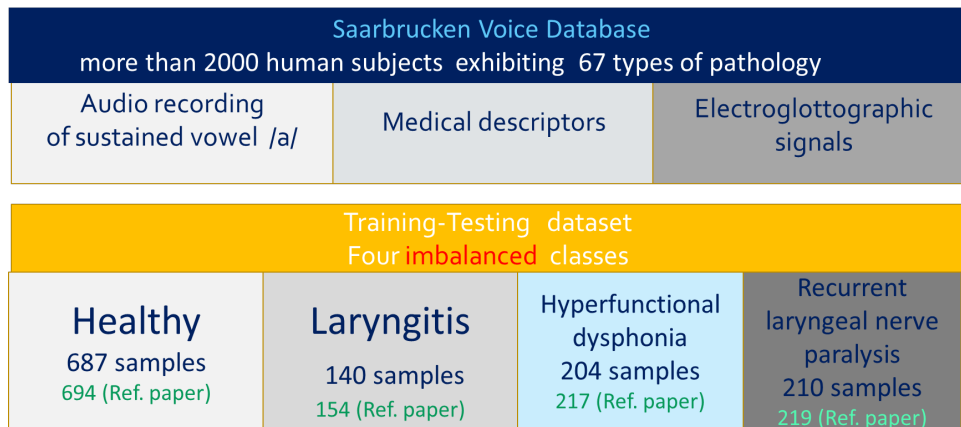


Figure 3.3: Subset of the SVD Dataset Employed in Our Experiments

To enhance our understanding of these disorders, we provide brief descriptions of each, highlighting their key characteristics as outlined in Islam et al. (2022) [Isl+22].

Laryngitis is characterized by inflammation in the larynx, typically caused by factors such as overuse, smoking, and laryngeal infections. In this condition, the vocal folds become inflamed and swollen, leading to sound distortion by obstructing air-flow. This pathology is common among various professionals, including singers, actors, telephone operators, lawyers, teachers, referees, coaches, and chemical factory workers. Laryngitis is also prevalent among children due to vocal overuse. Symptoms include a hoarse, weak voice, and in severe cases, the voice may become nearly undetectable.

Vocal cord polyps are benign masses located just beneath the surface membrane of the vocal cord. Often associated with significant voice use and vocal abuse, these polyps impact proper vocal fold vibration and, consequently, voice quality. Polyps can occur on either one or both vocal cords and are typically marked by prominent blood vessels.

Dysphonia is characterized by an abnormal voice quality that may emerge suddenly or gradually. It can present as hoarseness, roughness, cracking, weakness,

breathiness, or a gravelly texture, and may even result in temporary voice loss. Dysphonia can also alter pitch, and individuals often report pain when speaking, singing, or projecting their voices. The condition is frequently linked to abnormalities in the vocal cords, airflow obstruction from the lungs, or structural irregularities near the vocal cords, and may result from upper respiratory infections, colds, or allergies. Common types include muscle tension dysphonia, vocal cord paralysis, phonotraumatic lesions, recurrent respiratory papillomatosis, paradoxical vocal cord motion, and neurological disorders. The most prevalent type, muscle tension dysphonia, is marked by abnormal muscle tension activation, leading to elongated, stretched vocal folds, which can cause difficulties in phonation, breathing, and swallowing.

3.3.1 Network architecture

We develop a trimodal classifier that integrates audio, medical and EGG data. The network, as illustrated in Figure 3.4, operates through three specialized sub-networks. The initial component of the neural network architecture is tailored to process short-term feature sequences extracted from audio recordings. This component is specifically designed to accommodate varying lengths of input feature vectors without the need for fully connected layers. It leverages a fully convolutional neural network implementation, which allows for the application of convolutional and subsampling operations directly on the input sequences.

The architecture of this sub-network consists of a sequence of units composed of convolutional layers, max pooling layers, and batch normalization layers. Following this sequence, there is a final unit comprising a convolutional layer followed by global max pooling. Each of the first three units contains 64, 64, and 32 convolutional filters and with each kernel having size 3×3 and stride equal to 1, respectively, each followed by ReLU activation functions. The final convolutional layer employs 32 with 1×1 convolutional filters with a stride of 1, without zero padding. The output of this sub-network is sub-sampled by the global max pooling layer at the end of this branch. This design offers flexibility in processing inputs of different lengths.

The second sub-network, designed to integrate demographic parameters from medical records with mid-term features from audio recordings, utilizes a feed-forward neural network architecture. This sub-network comprises three hidden layers, each consisting of 64, 64, and 32 units, respectively, and utilizes Rectified Linear Unit activation functions.

This module is tailored to process a 5×1 input vector input vector, representing a composite of two medical parameters and three mid-term perturbation features extracted from the audio recordings. These mid-term features encompass fundamental frequency, harmonics-to-noise ratio, and jitter, providing a comprehensive representation of the acoustic characteristics of the audio data.

The third module in the architecture is structured as a fully convolutional neural network, consisting of two layers. Each layer contains 128 nodes, and the output from each node is processed through ReLU activation functions. Following this, a global max pooling layer is used to condense the spatial information across each feature map. This module processes the EGG based representations wavegrams.

We have to note that in case of processing the CQ indices, this sub branch is implemented as a cascade of 1D convolutions. More specifically, two Conv1D layers

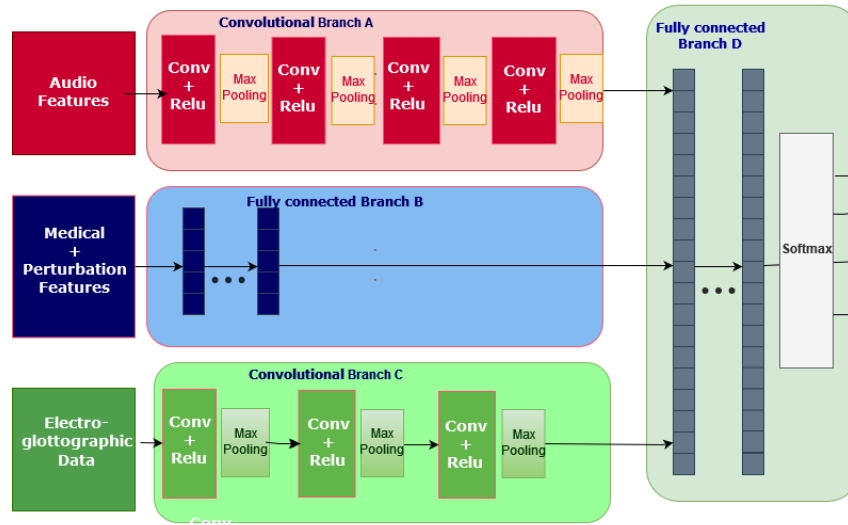


Figure 3.4: Trimodal classifier with EGG processing capabilities

are employed, with 32 kernels each and ReLU activation functions, followed by a global max pooling layer.

The outputs of all three modules are concatenated into a final dense layer comprising 128 nodes, with each node's output undergoing further processing via ReLU activation functions. Ultimately, the architecture concludes with a softmax output layer that is equipped with four units. This layer is designed to compute the posterior probability estimations for the four distinct classes corresponding to the pathological voice disorders being studied.

The innovative focus of this third module on EGG wavegrams introduces a new dimension to our architecture. It enables the system to capture insights into vocal cord movements, a critical factor in understanding pathological voice disorders. By processing EGG signals, this module complements the analysis of audio feature vectors and demographic parameters from earlier modules, fostering a more comprehensive approach to voice disorder classification.

3.3.2 Feature vector

Following the methodologies outlined in the previous chapter, we preprocess raw audio and EGG recordings, along with medical information, into input feature vectors suitable for processing by our neural network architecture.

Initially, the incoming voice signal undergoes a series of preprocessing steps to standardize its format and enhance its compatibility with the subsequent feature extraction processes. This involves resampling the signal at a standard rate of 44.1 kHz to ensure consistency across recordings, followed by amplitude normalization within the range of $[-1, +1]$.

Subsequently, the normalized voice signal is segmented into a sequence of short-term feature vectors using a moving window technique, which partitions the signal into short, overlapping frames. Specifically, a moving window with a duration of 40

ms and a hop size of 20 ms is employed to parse each recording. Within each frame, a comprehensive set of features is extracted, including Mel-Frequency Cepstral Coefficients along with their first derivatives, as well as the logarithmic output of the mel-scale filterbank. This meticulous feature extraction process results in 52 distinct short-term feature values per frame.

The extracted feature vectors are processed to integrate seamlessly into our neural network architecture. Specifically, each recording is transformed into a 2-D representation — similar to an image—with dimensions $N \times 52$, where N represents the sequence length. Our feature extraction process is designed to accommodate recordings of varying durations, ranging from 2 to 29 seconds, without requiring prior segmentation or zero-padding.

In our framework, the second input vector combines demographic parameters with perturbation features, as detailed in the preceding chapter. Although the SVD database provides limited demographic data, primarily consisting of a patient’s age and gender, we enhance this information by including three critical mid-term features: fundamental frequency, jitter, and harmonic-to-noise ratio.

To derive these mid-term features, we use the autocorrelation-based algorithm for periodicity detection, as described by Boersma in [Boe93]. This algorithm computes fundamental frequency variations, quantifies jitter (a measure of frequency perturbations), and assesses the harmonic-to-noise ratio (an indicator of signal quality) for each recording in the dataset.

These mid-term features are integrated into a concise 5×1 feature vector that encapsulates both demographic attributes and speech perturbation characteristics. This feature vector is then fed into the second sub-network of our architecture, which is implemented as a fully connected layer.

The third feature vector is derived from EGG signals, exploring three specific representations: spectrograms, “wavegrams”, and the Closed Quotient. These features are extracted to enhance the analysis and understanding of vocal fold dynamics. Spectrograms of these signals are computed using 1024 Discrete Fourier Transform coefficients, with a periodic Hamming window of 40 ms duration and a 25 ms overlap.

“Wavegrams” are generated using the algorithm outlined in [HFS10]. This algorithm estimates the time-varying fundamental frequency and identifies individual glottal cycles within each EGG recording. The resulting three-channel (RGB) image captures the dynamic nature of glottal activity, offering a visual representation of the EGG signal’s behavior over time. Each glottal cycle plot is carefully normalized and concatenated into a single image.

Complementing these visual representations, we compute Closed Quotient values. This process involves identifying landmark points, such as peaks and valleys, in both the EGG signal and its first derivative. To ensure robustness and mitigate computational errors, CQ values are estimated using five distinct algorithms, as outlined in [HFS10]. This approach yields a sequence of five feature values per frame, capturing the subtle nuances of glottal dynamics.

The processing of CQ values is done through a third sub-network, specifically designed as a cascade of 1D convolutions. All extracted features are normalized within the range $[-1, +1]$.

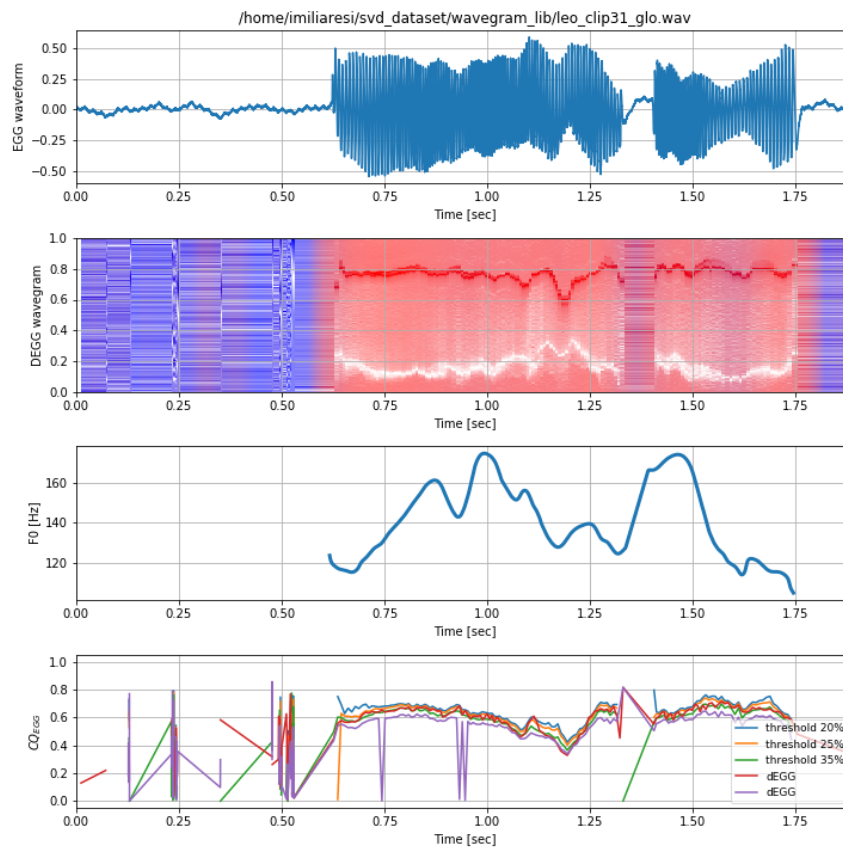


Figure 3.5: All visual representations of an example EGG waveform, Fo, CQ values and EGG Wavegram Representation

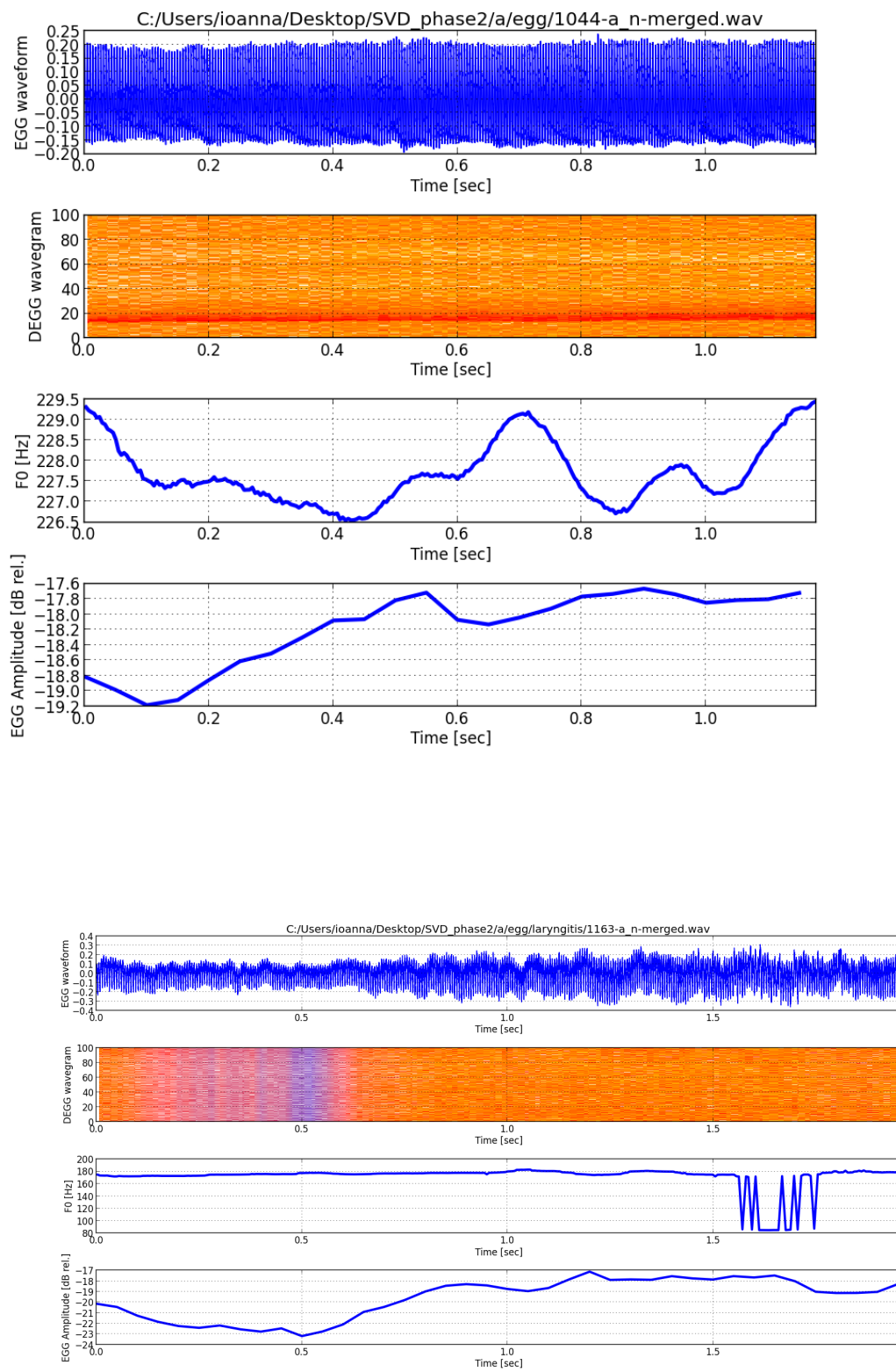


Figure 3.6: Example images of alternative EGG representation and indices

An example of an EGG waveform, DEGG wavegram, fundamental frequency, and CQ values is presented in Figure 3.5.

3.3.3 Ablation study

The proposed architecture integrates three key modalities: audio recordings, medical information, and EGG data. For our experiments, we utilized a trimodal classifier that combines data from these sources using a carefully designed three-module system. To refine the model configuration, we conducted an ablation study, meticulously examining various aspects, including network topology, dimensionality, and training parameters. Additionally, this experimental phase was dedicated to assessing the efficacy of augmentation techniques introduced in prior configurations. Specifically, we evaluated their performance on both convolutional and fully convolutional architectures, and investigated the impact of noise injection and an arbitrary length segmentation algorithm. The goal was to determine how each component—audio recordings, medical information, and EGG data—contributed to the classifier’s performance. We systematically examined different network configurations, dimensionality adjustments, and training parameters to understand their effects on classifier efficacy.

We then assessed the outcomes of multiple experimental configurations. Initially, following common practice in machine learning, the dataset was randomly divided into training and testing subsets. We partitioned 90% of the dataset for the training subset and allocated the remaining 10% to the testing subset, employing a stratified approach to ensure consistent class distribution across both subsets.

Our dataset exhibits an imbalance, primarily due to the size of the healthy class—a known factor that can bias classifiers, leading to inflated accuracy for dominant classes [Ziq+21]. To mitigate this, a random oversampling technique [BMM18] was applied after the train/test split, specifically for the training subset.

For our final model configuration, we adopted a 5-fold cross-validation approach, limiting each iteration to a maximum of 500 epochs. We utilized the Adam gradient descent algorithm in conjunction with a cross-entropy loss function, setting the learning rate to 0.001. To mitigate overfitting, we introduced dropout regularization (with values of 0.5 for fully connected layers and 0.1 for each convolutional layer) and employed an early stopping criterion.

The model’s validation and evaluation were based on accuracy, recall, and precision metrics, with a particular focus on understanding the impact of class imbalance.

3.3.4 Experiments and results

Our experimentation involves varied combinations of input modalities, with a specific emphasis on the role of EGG features. We systematically evaluate the effectiveness of two specific EGG features within our trimodal classification framework. For CQ values, we use a feed-forward network, while for wavegrams, we employ a convolutional neural network. Both evaluations follow strict segmentation and zero-padding procedures to ensure consistency and comparability. After this initial evaluation, we further explore the potential of wavegrams using fully convolutional networks, an innovative approach that captures the temporal dynamics of audio signals without

requiring segmentation. Additionally, we assess the effects of noise injection and arbitrary clipping, which are critical for determining the robustness and generalization capabilities of our model.

Our investigation systematically examines how the classifier’s accuracy is influenced when individual modalities are removed or replaced with random noise. By isolating each input modality, our aim is to quantify its significance in contributing to the overall classification accuracy.

We established a baseline model that integrates all input modalities—audio features, demographic parameters, and EGG signals. This baseline model serves as a reference point for assessing the relative importance of each modality.

To evaluate the influence of each input modality in isolation, we systematically eliminated one modality at a time from the baseline model. Specifically, we trained and evaluated the classifier using only audio features, only demographic parameters, and only EGG signals, while keeping the other modalities unchanged.

The primary objective of this study was to investigate the potential benefits of incorporating multiple input sources, particularly EGG signals, for voice pathology classification. To achieve this, we conducted experiments using three distinct EGG-derived features: the EGG spectrogram, Closed Quotient values, and wavegrams. These features were considered as alternative inputs for the third sub-network, allowing us to analyze their individual contributions to the classification task. By assessing the classifier’s performance with different EGG representations, we aimed to identify the most effective approach for integrating EGG signals into the classification framework.

Experimental results

Our study encompassed a comprehensive examination of eight distinct classifier topologies, covering a range of unimodal, bimodal, and trimodal architectures. These variations allowed us to explore the efficacy of different input modalities and their combinations in voice pathology classification. In Table 3.3, we provide a detailed overview of the network structures and corresponding classification outcomes, with uppercase letters within parentheses referencing specific rows in the table.

We began our evaluation by focusing on unimodal classifiers, which retained only the convolutional sub-network and the final fully connected layers. Remarkably, the unimodal classifier dedicated to processing audio recordings (A) achieved a robust testing accuracy of 68%. This performance underscores the significance of acoustic features in distinguishing between different pathological conditions of the voice.

In contrast, the unimodal classifier leveraging medical records and perturbation features (B) exhibited a comparatively lower testing accuracy of 56.2%. While demographic parameters and perturbation features contribute valuable insights into voice pathology, their standalone utilization proved less effective for classification compared to acoustic features.

We then investigated unimodal representations of the EGG signal (C, D, E), each equipped with a convolutional sub-network and final fully connected layers. The model processing Closed Quotient values (C) achieved a testing accuracy of 58.3%. CQ values, which reflect glottal closure properties, provide important insights into vocal fold behavior, but their ability to fully discriminate between pathologies appears limited.

On the other hand, the utilization of wavegrams (D) led to a notable improvement in classification accuracy, achieving 59.4%. Wavegrams, derived from EGG signals, offer visual representations of glottal cycles and provide insights into vocal fold vibration patterns, thereby enhancing the classifier’s ability to differentiate between pathological conditions.

However, the utilization of spectrograms (E) resulted in the lowest performance among the unimodal classifiers, with an accuracy of 26.5%. Spectrograms, while commonly used in acoustic analysis, may not capture the intricate details of vocal fold dynamics encapsulated by other EGG-derived representations.

Table 3.3: Classification accuracy for alternative classifiers

Top.	Ac.	Med	CQs	Waveg	Spectr.	Accuracy (%)	Precision	Recall
A	✓					68	[0.51, 0.19, 0.19, 0.88]	[0.80, 0.37, 0.22, 0.38]
B		✓				56.2	[0.53, 0.17, 0.13, 1]	[0.81, 0.2, 0.2, 0.08]
C			✓			58.3	[0.625, 0.15, 0, 1]	[0.94, 0.15, 0, 0.04]
D				✓		59.4	[0.63, 0.17, 0, 1]	[0.91, 0.2, 0, 0.08]
E					✓	26.5	[0, 0, 0.11, 0.14]	[0, 0, 0.54, 0.4]
F	✓	✓				83.2	[0.89, 0.25, 1, 0.75]	[0.98, 0.91, 0.43, 0.63]
G	✓		✓			81.1	[0.97, 0.58, 0, 0.85]	[0.95, 1, 0, 0.9]
H		✓	✓			75.1	[1, 0.66, 0.42, 0.39]	[0.53, 0.2, 0.85, 1]
I		✓		✓		79.2	[0.96, 0.87, 0.5, 0.63]	[0.84, 0.35, 0.92, 0.90]
J	✓			✓		82.6	[0.98, 1, 0.43, 0.73]	[0.89, 0.055, 1, 0.95]
K	✓	✓	✓			88.8	[1, 0.62, 0.75, 0.71]	[0.84, 0.94, 0.23, 1]
L	✓	✓		✓		89.3	[0.94, 0.62, 0.76, 0.74]	[0.89, 0.94, 0.28, 0.95]

We further investigated bimodal classifier structures (E, F, G, H, I) to explore the synergy between different input modalities. Among these configurations, the fusion of acoustic features with demographic/perturbation features (E) achieved the highest accuracy, reaching 83%. This combination effectively integrates acoustic characteristics with patient-specific data, enhancing the classifier’s ability to discern nuanced variations indicative of various pathological conditions.

Additionally, combining acoustic features with CQ values (F) yielded an accuracy of 81.1%, underscoring the importance of incorporating glottal closure properties into the classification framework. Similarly, integrating acoustic features with wavegrams (G) achieved a commendable accuracy of 82.6%. Leveraging wavegrams, which provide visual insights into vocal fold behavior, enriches the classifier’s understanding of pathological vocal mechanisms.

Further experiments involved combining demographic/perturbation features with CQ values (H), resulting in a testing accuracy of 75.1%, while integration with wavegrams (I) yielded an accuracy of 79.2%. These results highlight the potential of combining different modalities to improve classification accuracy and provide a more comprehensive understanding of voice pathology.

Finally, we explored a full trimodal classifier, leveraging all three modalities simultaneously. Notably, when wavegrams were used as features, the classifier achieved the highest classification accuracy of 89.3%. This integration led to improved class distribution, evidenced by precision values of [0.94, 0.62, 0.76, 0.74] and recall values of [0.89, 0.94, 0.28, 0.95]. Similarly, utilizing CQ values (I) resulted in an accuracy of 88.8%, with precision values of [1, 0.62, 0.75, 0.71] and recall values of [0.84, 0.94, 0.23, 1]. Integrating all three modalities led to both higher overall classification accuracy and a more balanced class distribution.

Despite challenges in discriminating laryngitis, particularly due to its status as a minority class, our approach demonstrates state-of-the-art classification performance with an accuracy of 89.3%. This outperforms the best-known classification system, which achieved 87.5% accuracy [Hem17] on the same classification task.

The baseline model, integrating all input modalities, achieved the best classification accuracy. This serves as the reference accuracy against which the performance of modified models was compared. Removing audio features resulted in a decrease in classification accuracy compared to the baseline model. Similarly, removing demographic parameters and EGG signals led to changes in classification accuracy, indicating their respective contributions to classifier performance.

The experimental evaluation of these configurations revealed distinct performance differences across unimodal, bimodal, and trimodal structures. Unimodal classifiers processing audio recordings attained moderate accuracy of 68%, while those handling medical records and perturbation features reported lower accuracy of 56.2%. When analyzing EGG signal classifiers individually, CQ values and wavegrams achieved accuracies of 58.3% and 59.4%, respectively, while spectrograms lagged significantly at 26.5%.

Transitioning to bimodal structures, the fusion of acoustic features with demographic/perturbation features exhibited the highest accuracy of 83%. Integration with CQ values and wavegrams showcased accuracies of 81.1% and 82.6%, respectively. Additionally, the combination of CQ values with demographic/perturbation features scored 75.1%, while integration with wavegrams reached 79.2%.

Remarkably, the most noteworthy performance emerged from the trimodal classifier employing wavegrams as a feature, achieving an exceptional accuracy of 89.3%. This approach not only showcased enhanced precision and recall values across classes but also surpassed the performance of existing voice pathology classification systems, which previously achieved 87.5%.

In summary, our comprehensive analysis unveiled notable performance differentials across diverse classifier configurations. While unimodal classifiers demonstrated moderate accuracies, particularly those handling audio recordings, the integration of multiple modalities markedly enhanced accuracy. Notably, bimodal and trimodal architectures, particularly those integrating acoustic, demographic, and EGG

signal features seamlessly, showed increased accuracy. The trimodal classifier leveraging wavegrams achieved the highest performance at 89.3%, underscoring the pivotal role of multi-modal fusion in advancing voice pathology classification accuracy.

The integration of diverse data modalities significantly contributed to the enhancement of classification performance, evidenced by considerable improvements in accuracy, precision, and recall metrics. Despite the constraints of a relatively modest and imbalanced voice pathology dataset, the synergistic inclusion of audio features, demographic data, and EGG measurements notably elevated the classifier's accuracy to 89.3%, surpassing the baseline score of 87.5%. These findings highlight the crucial role of multi-modal fusion, especially the inclusion of EGG features, in improving voice pathology classification accuracy.

The study's results offer critical insights into the relative importance of each input modality in voice pathology classification. Audio features, demographic parameters, and EGG signals all contributed significantly to classifier performance, though their individual impacts varied. A detailed understanding of each modality's role could guide strategic feature selection and foster the development of more effective classifiers for specific diagnostic applications. This exploration not only clarifies the complex interactions between different data modalities but also emphasizes the transformative impact of multi-modal fusion on advancing voice pathology diagnostics and treatment methodologies.

3.4 Conclusions

In this experimental stage, our focus was on investigating the potential enhancement of voice pathology classification accuracy by incorporating a third input source—specifically, EGG signals. The investigation involved evaluating several experimental setups, emphasizing the role of EGG-derived features in conjunction with two other modalities: audio signals and medical descriptors.

We conducted experiments on the SVD dataset, which exhibits considerable diversity and has been foundational for various classification challenges involving different types and quantities of pathologies. Specifically, we focused on a four-class voice pathology classification task involving hyperfunctional dysphonia, laryngitis, vocal cord polyps, and dysphonia, leveraging a subset of the SVD containing sustained vowel /a/ voice recordings from both genders.

To address this classification task, we proposed a modular deep learning architecture capable of integrating and processing EGG signals alongside audio recordings and demographic data. The architecture comprised three specialized sub-networks, each tailored to process a different input modality. Notably, we found that processing audio features and EGG features as 2-D images outperformed other configurations. Moreover, processing images of arbitrary dimensions using fully convolutional sub-networks significantly increased classification performance.

Our experiments spanned various classifier configurations, including unimodal, bimodal, and trimodal setups. We assessed the performance of different input modalities, with particular emphasis on the role of EGG features. The integration of EGG-derived features—especially wavegrams a visual representation of EGG—led to significant improvements in classification accuracy, surpassing the state-of-the-art performance achieved in previous studies.

The integration of diverse data modalities significantly contributed to enhanced classification performance, as evidenced by substantial increases in accuracy, precision, and recall metrics. Despite working with a comparatively small and imbalanced voice pathology dataset, the incorporation of audio features, demographic data, and EGG measurements notably elevated the classifier's accuracy, highlighting the importance of multi-modal fusion in advancing voice pathology classification.

In conclusion, our study demonstrates the efficacy of integrating EGG signals alongside traditional audio and demographic data for voice pathology classification. The findings underscore the importance of multi-modal fusion approaches in improving classification accuracy.

Chapter 4

Attention-guided Modular Deep Learning Architecture for Covid-19 Audio Classification

The rapid global spread of the SARS-CoV-2 coronavirus (COVID-19) has necessitated the urgent development and implementation of cost-effective and reliable virus detection methods. Given the significant global impact of COVID-19, our research aims to advance the field of audio processing for detecting this virus. This work aligns with the growing interest in leveraging deep learning architectures for the early and accurate detection of infectious diseases, focusing specifically on audio indicators associated with COVID-19. Although COVID-19 primarily affects the lungs, it also impacts the vocal tract and can manifest symptoms in the voice, further underscoring the relevance of our audio-based detection approach.

4.0.1 Problem Definition and Application Context

The existing body of audio processing research for COVID-19 detection primarily focuses on analyzing various respiratory sounds, such as coughing, breathing patterns, speech, and specific phoneme utterances. The choice of sound type for COVID-19 detection significantly impacts the accuracy and reliability of diagnostic tools. Many studies utilize cough sounds, focusing on their acoustic features to identify patterns associated with COVID-19. Research indicates that COVID-19-induced breathing has distinctive characteristics detectable through machine learning algorithms.

Simultaneously, speech analysis for COVID-19 detection involves examining specific phonemes and speech patterns, as alterations in voice quality, pitch, and articulation can indicate underlying respiratory issues. Some studies have adopted a multimodal approach, combining cough, breath, and speech sounds to capture a broader range of symptoms and reduce the risk of misclassification due to variability in individual sound features. While cough sounds have shown high accuracy and are relatively easy to collect, could breath sounds and speech patterns offer additional diagnostic information to enhance the overall performance of the detection system? Would a multimodal approach that leverages the strengths of each audio modality provide the most reliable and robust solution for COVID-19 detection?

Through our literature review, we observed that in some studies, classifiers tested on different datasets do not maintain consistent performance. Certain classifiers require adaptation or retraining when applied to new datasets with different sound

modalities, even when all sound samples are available. Essentially, the same architecture often needs retraining with a new dataset that includes different sound modalities to achieve optimal classification accuracy. This poses a challenge in developing a classification scheme capable of processing various combinations of available modalities without extensive retraining.

Another observation is that available COVID-19 datasets include different audio sources. How can we adapt a model to handle this variability? How should we approach scenarios where a model trained with certain audio types must function with testing data that lacks those information types?

This research aims to address these gaps and advance the field of audio-based COVID-19 detection by developing a comprehensive methodology capable of processing the full spectrum of available audio modalities. The key emphasis of our approach lies in ensuring robust generalization capabilities, meaning our model is designed to adapt effectively to different types of audio data and diverse datasets. This adaptability is crucial given the variability in audio data collected from different sources and environments.

Another significant challenge in this field is the limited availability of large, diverse audio datasets where all COVID-19 audio types are collected. In many scenarios, repositories of sounds may be limited, either in the number of samples or in the variety of audio modalities available. Therefore, our methodology is meticulously designed to maintain high classification accuracy even when certain audio modalities are absent or compromised. This is achieved through a multimodal architecture that integrates and processes multiple audio inputs, enhancing the model's ability to make accurate predictions despite incomplete data.

In our pursuit of a flexible and robust methodology, it is essential to recognize the availability of data resources for research purposes. Several datasets containing audio recordings from both COVID-19 patients and healthy individuals have been collected and are accessible under diverse licensing arrangements. Prominent datasets include the MIT COVID-19 dataset [Jor+20], University of Cambridge COVID-19 dataset [Ton+21], University of Stanford Virufy dataset, COUGHVID dataset [Lar+21], Indian Institute of Science Coswara dataset [Nee+20], and the DICOVA challenge dataset [Mug+21]. Each of these datasets encompasses one or more, and in some instances, all of the following audio modalities: breathing, coughing, counting, and sustained phonation.

Our experiments centered on the publicly available crowdsourced Coswara dataset, which encompasses all essential audio modalities. Additionally, we assessed our model's performance on the Virufy dataset, which exclusively contains coughing records. While achieving a classification accuracy of 97.75% on Coswara, surpassing the referenced paper's 97.54% result on the same dataset, our architecture's primary strength lies in its adaptability to datasets featuring fewer audio modalities. This adaptability was evident in achieving a commendable 82% classification accuracy on the Virufy dataset.

4.0.2 Research Objectives and Contributions

To address the aforementioned challenge, our work introduces a multimodal architecture designed to accommodate all available audio modalities. We propose a multimodal deep learning model consisting of nine subnetworks, each implemented using a fully convolutional neural network module, to process the nine distinct audio modalities under study. This design ensures that the model can extract and integrate relevant features from each modality, thereby improving overall performance. Additionally, we incorporate an attention mechanism that dynamically assigns weights to each audio modality, allowing the model to focus on the most informative signals for classification.

The novelty of our architecture lies in integrating an attention-guided mechanism for audio modality selection. This component plays a pivotal role in discerning the relevance and importance of individual modalities in the overall classification decision-making process. What sets our approach apart from conventional attention layers is the incorporation of a self-guided attention mechanism designed to weight and score each audio modality. A similar mechanism for feature selection was introduced in [GGH19]. However, in our architecture, this mechanism is utilized for integrating audio modalities, providing an innovative solution to the challenges posed by varying data collections and sound modalities.

Attention mechanisms have also been explored in [Wal+22] using an ensemble of attention-based Convolutional Recurrent Neural Network (A-CRNN), attention-based bidirectional Long Short Term Memory (A-BiLSTM), attention based bidirectional Gated Recurrent Unit (A-BiGRU), and Convolutional Neural Network. The final prediction is derived by averaging the prediction probabilities from each attention network. In our implementation, we introduce an attention module responsible for directing focus onto each audio modality, thereby influencing the ultimate classification decision. This self-guided attention mechanism enables dynamic selection and weighting of audio modalities, representing a critical innovation in COVID-19 detection from audio signals.

Our approach represents a significant advancement in audio-based COVID-19 detection, offering a scalable and flexible solution to the problem of modality variability. By enabling the model to adapt dynamically to the available data, we enhance its utility in real-world applications where the presence and quality of audio recordings can vary widely.

In summary, our proposed multimodal deep learning model addresses key challenges associated with processing diverse audio datasets for COVID-19 detection. By integrating specialized subnetworks and an adaptive attention mechanism, our architecture ensures robust feature extraction and optimal classification performance across a range of audio modalities. This innovative approach not only improves the accuracy of COVID-19 detection but also sets a new standard for multimodal audio processing in infectious disease diagnosis.

Our architectural framework relies on three key pillars: multimodal analysis, attention mechanisms, and the creation of a robust classifier. This approach ensures that our model can effectively handle diverse audio inputs, adapt to varying data conditions, and maintain high classification accuracy. The framework exhibits several notable characteristics:

Flexible Modality Requirements: Individual recordings for all nine audio modalities are not mandatory; the presence of at least one modality suffices. This flexibility is critical in real-world scenarios where complete datasets may not always be available. Our model is designed to leverage whatever audio data is present, ensuring accurate classification even when certain audio modalities are missing or incomplete.

Dynamic Relevance Determination: The model dynamically identifies the most relevant audio type for class identification. Using advanced attention mechanisms, it evaluates the importance of each audio input in real time, focusing on the most informative signals for classification. This dynamic relevance determination is essential for handling variability in audio data, allowing the model to adapt its focus based on available information and thus improving classification accuracy.

Self-Guided Attention Mechanism: In contrast to conventional attention mechanisms, our self-guided attention module operates on the entire feature map vector, incorporating contextual information from neighboring frames rather than processing each frame individually, as seen in RNN or BLSTM attention models. This approach enables the attention mechanism to consider broader contextual information, leading to more informed and accurate weighting of audio modalities.

Our experimental framework was meticulously designed to evaluate various feature vector compositions, introduce a modality-selection attention mechanism, and implement diverse data augmentation techniques—all aimed at enhancing the model’s performance and generalizability.

Key contributions include the successful integration of a modality selection attention guided mechanism, which dynamically regulates the processing of distinct audio modalities, such as respiratory sounds and vowels. This mechanism improved classification accuracy. Additionally, visualizing the evolution of attention scores across different modalities during training provided valuable insights, with cough sounds emerging as the most significant, followed by breath and speech sounds.

Furthermore, our exploration of data augmentation techniques—including noise injection (utilizing white, brown, and pink noise) and segmentation—significantly bolstered the robustness of the classifier. This augmentation strategy, coupled with validation across diverse datasets, affirmed the generalizability and reliability of our models in real-world scenarios. Notably, cross-dataset evaluation using the Virufy data corpus, which consists solely of cough recordings, demonstrated the classifier’s adaptability to different data distributions, further solidifying its practical utility.

Comparative analysis among multimodal configurations and individual unimodal classifiers on the Coswara dataset consistently demonstrated the superiority of multimodal approaches. The enhanced performance of the Mel-spectrogram feature compared to MFCC coefficients across all audio types highlights its effectiveness in capturing spectral information, thereby improving classification accuracy.

In the following sections, we undertake a comprehensive exploration of the existing body of knowledge by presenting related work (see Section 4.1) in the field of COVID-19 detection. We meticulously examine prior research to establish the groundwork for our innovative contributions. In Section 4.2, we describe the details of our method’s architecture, providing an in-depth exposition of its underlying framework. Next, we outline the experiments conducted to validate and refine our approach. The results of these experiments are thoroughly presented, offering insights into the

efficacy and performance of our method (see Section 3.3.4). Finally, we present all conclusions in the Conclusion section (see Section 3.4).

4.1 Related Work

4.1.1 Voice Pathology Detection in the Context of COVID-19

In the existing literature, numerous methods have been proposed for COVID-19 detection, employing both machine and deep learning techniques. Experiments on collected datasets have predominantly focused on analyzing available audio information sources, specifically respiratory sounds (breathing and coughing) and/or sustained phonation samples. The majority of studies have heavily relied on coughing as the primary mode of analysis, as seen in studies such as [She+23], [Jor+20], [Jin+22], [Rum+22], [Man21], [Gok+21], [Alb+21], [Esi+22], [Syr+22], [Zhu+22], [Cho+22b], [KBL23], [Pre+23], [IAT22], and [Sob+22].

However, a subset of studies including [PN21b], [Mad+22], [Cel23], and [Dut+22], have investigated breathing along with sustained phonemes as audio sources. Notably, the DICOVA challenge research efforts were split into two tracks: one focused on cough sounds, and the other on breathing combined with sustained phonation [PN21a].

Moreover, [Nee+21] explored the combination of classification decisions derived from multiple features using a standard aggregation mechanism. Only a limited body of work has attempted concurrent analysis of coughing, breathing, and phonetic sounds, as evidenced by [YDM+23], [Che+23], [DQM22], [Mad+21], [Mos+22], [Xin+22], and [Rah+22].

In a distinct approach, studies such as [JU23], [RU23], and [NDS23] have focused on speech signals, with particular attention to cough and breath in studies like [Man+23].

Various machine learning methods have been applied, including Support Vector Machine (SVM) [Man21], [Ver+21], [Zha+22b], Convolutional Neural Networks (CNN) [Alb+21], [Gok+21], and bi-directional Long Short-Term Memory (BiLSTM) networks [Xin+22], [Mad+22]. Transfer learning techniques have also been leveraged, with studies using pre-trained models like ResNet50 [Jor+20], [SL22].

Recent studies increasingly utilize transfer learning to enhance fine-tuning. For instance, [Gal+22] employed transfer learning from AemResNet, a pre-trained end-to-end audio embeddings generator, to analyze respiratory and coughing sounds using CNNs. Similarly, [Mos+22] used CNNs to model raw waveform data for phone classification within speech recognition.

In [EN23], various architectures, including ResNets, VGGs, AlexNet, DenseNet, SqueezeNet, and CIdER (COVID-19 Identification ResNet), were compared for classification performance. [Pre+23] used ResNet50 with XGBOOST for classification, while [Xin+22] employed the pre-trained wav2vec2.0 for their investigations.

Building on these methodologies, studies like [Jor+20] used three pre-trained ResNet50 models to process MFCC of cough recordings. Similarly, [Mad+21] employed pre-trained CNN, LSTM, and ResNet50 networks to process MFCCs of cough,

breath, and speech sounds. Additionally, [SL22] introduced feature vectors comprising MFCCs, spectral features, spectral contrast, and chroma features into a pre-trained ResNet50 and DNN-based architecture for COVID-19 detection. Furthermore, [Zhu+22] used a CNN combined with ResNet50 to analyze spectrograms of cough sounds.

For evaluating different architectures, [RU23] assessed VGGish, YAMNET, and L3-Net architectures for feature extraction and fine-tuning. Similarly, [YDM+23] evaluated VGGish, YAMNET, and L3-Net architectures in their research.

In their investigations, [RU23] conducted a comparative analysis of machine learning techniques for COVID-19 detection, using methods like K-Nearest Neighbor and SVM. Additionally, [NDS23] leveraged Mel frequency magnitude coefficients with machine learning classifiers, including random forests and K-nearest neighbor. Meanwhile, [Che+23] experimented with logistic regression, SVM, LSTM, and decision tree models to analyze acoustic and symptoms data. Moving towards SVM applications, [Man21] focused on classifying COVID-19 cough using power spectral density and MFCCs, while [Ver+21] also explored SVM for COVID-19 cough classification. In a different approach, [KBL23] proposed a feature set for COVID-19 diagnosis that included MFCC, D2-MFCC, D-MFCC, and spectral contrast features applied to a model combining ResNet-50 and DNN architectures. Moreover, [Yi+23] used an SVM classifier to process modulation spectral and linear prediction speech features for COVID-19 detection.

In the analysis of acoustic parameters for COVID-19 detection, [Zha+22b] used linear regression models and SVM algorithms. Shifting towards a semi-supervised learning (SSL) framework, [DQM22] validated their approach on a crowd-sourced audio database, utilizing metadata and MFCC coefficients to analyze cough sounds. Similarly, [Gok+21] focused on metadata and MFCC coefficients of cough sounds for COVID-19 detection, and [Alb+21] analyzed cough recordings using CNNs.

Furthermore, [Mad+21] reported enhanced COVID-19 detection performance by incorporating cough, breath, and speech data. [Wal+22] proposed an ensemble model involving A-CRNN, A-BiLSTM, A-BiGRU, and CNN, using respiratory, speech, and coughing audio inputs. [Xin+22] introduced a Bi-LSTM network-based method for COVID-19 detection utilizing breath, speech, and cough signals, while [Mad+22] also employed a Bi-LSTM classifier to analyze breath, cough, and speech audio samples. [MDP23] focused on phoneme analysis using Mel and Gammatone Cepstral coefficients with standard machine learning classifiers. [Cel23] introduced Covid-CoughNet, with various feature extraction techniques on voice signals. [YDM+23] evaluated multiple machine learning classifiers on cough, breath, and speech recordings. [Man+23] explored interval temporal decision trees and forests for automated classification of multivariate time series data from cough, breath, and speech recordings. [She+23] proposed incorporating piecewise position encoding into time frequency features.

Several global solutions have emerged from significant COVID-19 projects. For instance, QUCoughcope [Muh+22] utilizes spectrograms of cough and breath sounds in a stack of CNN networks for symptomatic and asymptomatic patients. Project Achoo [Ale+22] employs spectrograms, cochleagrams, and cochleagram statistical features in deep CNN and gradient boosting models. CIoTVID [Alf+21] incorporates MFCC coefficients alongside socio-demographic information, treatments, and

medications. Additionally, an ensemble model comprising A-CRNN, A-BiLSTM, A-BiGRU, and 1D CNN with an attention mechanism was proposed for lung abnormality and COVID-19 diagnoses [Wal+22].

Notably, [Zhu+22] highlighted a decline in classifier performance when applied to new data collections, particularly those using deep learning methodologies. [Syr+22] emphasized the potential of cough as a biomarker for health assessment but identified a scarcity of extensive and diverse datasets in this area.

Focusing on the top results from Coswara-related studies, [Sob+22] achieved impressive accuracy using YAMNet, segmenting and labeling cough sounds with four different fractal dimension calculation methods. [Wal+22] proposed an ensemble model with attention mechanisms utilizing respiratory, speech, and coughing audio inputs. [PN21a] reported high accuracy for breath classification using ResNet50. In another study, [Wal+21] demonstrated high accuracy with an attention-based BiLSTM network for both the Respiratory Sound and Coswara datasets.

4.2 Methodology

4.2.1 The Coswara and Virufy datasets

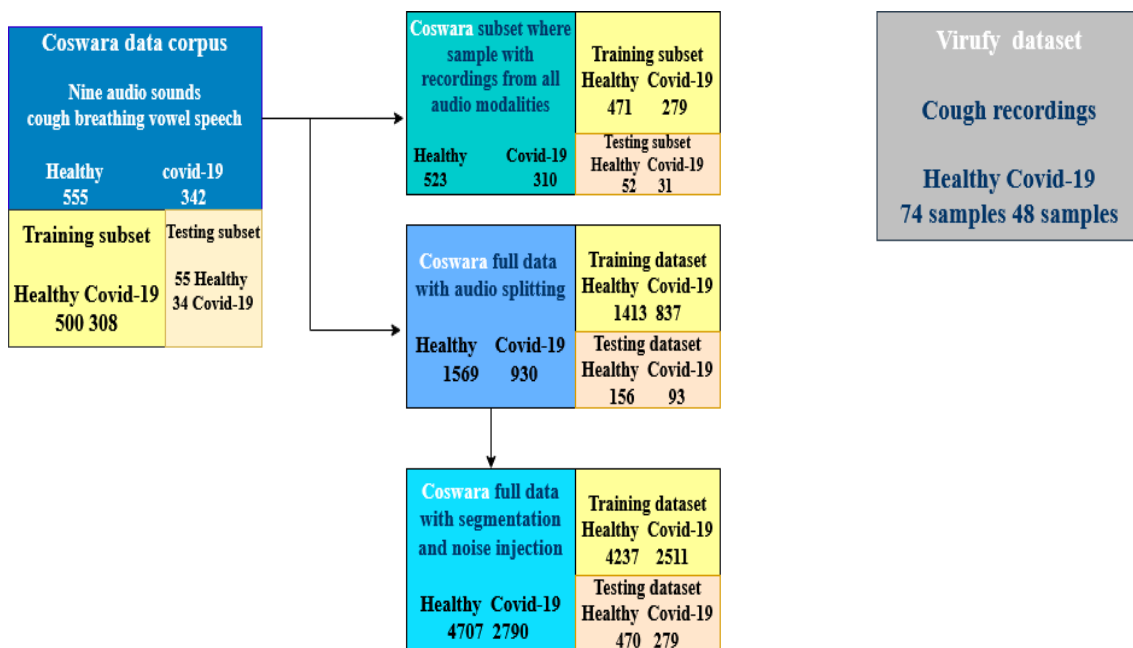


Figure 4.1: Illustration of all five datasets employed in the experiments

In these experiments, we addressed the COVID-19 classification challenge using datasets aligned with our research objectives. Comprehensive experiments were conducted across all available audio modalities, utilizing both the Coswara and Virufy datasets.

The Coswara dataset [Bha+23] encompasses nine sound modalities, including various breathing patterns (deep and shallow), coughs (heavy and shallow), sustained vowel phonations (/e/, /a/, and /o/), and speech (counting fast and counting normal).

The dataset includes recordings from asymptomatic COVID-19 patients in the “unhealthy” class. In total, the dataset comprises 897 recordings, with 342 labeled as unhealthy and 555 as healthy across the nine modalities. The recordings vary in duration. The Virufy dataset, while smaller in comparison, focuses exclusively on cough samples from both COVID-19-infected individuals and healthy subjects. These datasets were collected using diverse smartphone models, which introduced varying audio sampling frequencies, compression artifacts, and background noise levels. The audio files consist of a blend of compressed and uncompressed formats (wav and mp3 files), depending on the method of data acquisition. These data corpora are essential for determining the most suitable audio modality for COVID-19 identification.

To validate the model, we utilized all nine available audio sources, including various types of coughs (heavy and shallow), different counting speeds (fast and slow), and sustained vowels (/a/, /e/, /o/). We used five distinct datasets in our experiments, as shown in Figure 4.1. These included the entire Coswara collection, an augmented dataset with audio segmentation, and an augmented set created from the original Coswara dataset with both audio segmentation and colored noise injection. Additionally, we conducted experiments on a subset of the Coswara data containing samples where all nine audio modalities were available. During audio preprocessing, we identified a notable number of empty files (thirty-two for each audio modality) with zero signal value, which were consequently excluded from the dataset. Finally, we evaluated the architecture using transfer learning techniques on the Virufy dataset.

4.2.2 Feature Extraction

In our experimental setup, we included diverse sound modalities such as shallow and deep breathing, heavy and shallow coughing, sustained vowel phonation (/e/, /a/, and /o/), speech, and counting (fast and slow), obtained from both COVID-19-infected and healthy subjects. Our final system configuration involved extracting Mel-spectrogram representations from all nine audio modalities.

The initial step in our processing pipeline involves converting raw audio signals into Mel-spectrogram representations. To ensure consistency, all signals are resampled to 16 kHz. We then compute the spectrograms using 1024 Discrete Fourier Transform coefficients, applying a periodic Hamming window of 40 ms duration with a 25 ms window overlap. To transform the spectrograms to the Mel scale, we pass them through a Mel filter bank consisting of 32 bandpass filters. The frequency range covered by this filter bank spans from $f_{\min} = 20$ Hz to $f_{\max} = 20$ kHz.

The resulting feature sequence is represented as a 2-D image with dimensions $N \times 32$, where N varies from 223 to 4103 depending on the duration of the audio recording. These feature vectors, obtained through this process, are subsequently inputted into the feature learning module for further processing.

4.2.3 Network Architecture

We introduce a comprehensive end-to-end deep learning model composed of three key modules:

- **Feature Representation Learning Module:** Responsible for extracting meaningful features from the raw input data, this module transforms the input into

a structured representation that captures key characteristics necessary for accurate classification.

- **Attention Mechanism:** Designed to weigh the contributions of different audio modalities, the attention mechanism emphasizes the most relevant features during the learning process. This dynamic weighting enhances the model’s ability to focus on critical aspects of the input data, boosting overall performance.
- **Classification Stage:** The final stage, where processed features are used to classify the data into predefined categories. This stage combines outputs from the previous modules to generate accurate predictions.

The overall architecture of this model is visually depicted in Figure 4.2.

The feature representation learning module is a fundamental component of our end-to-end deep learning model, leveraging a fully Convolutional Neural Network-based approach tailored to handle nine distinct audio modalities. This structure enables the model to seamlessly process audio recordings of arbitrary durations. Each branch is implemented as a chain of convolutional layers ending with a global max pooling layer. Specifically, the first three layers apply convolutions with 64, 64, and 128 kernels of size 3×3 , respectively, and the last layer is a 1×1 convolution with 128 filters, followed by a global max pooling layer. The activation function in all convolutional layers is the Rectified Linear Unit.

At the output of the global max pooling layer, the feature embeddings undergo L2 normalization. By scaling the embeddings to have a unit norm, L2 normalization promotes balanced contributions from all modalities.

This module is designed to operate on audio recordings from nine categories. Each audio recording is converted into a one-channel, two-dimensional “image” represented as an $N \times 32$, where N depends on the sample’s duration and ranges from 223 to 4103. The dimensions of this “image” dynamically adjust based on the recording’s length.

The feature embeddings produced by this module are determined by the number of filters in the final convolutional layer and remain unaffected by the input image’s size. Since the last layer uses 128 filters, the resulting feature embedding has the form of $1 \times N \times 128$.

The input feature vectors for this module, denoted as $I = [I_1, \dots, I_9]$, where each I_i has the shape of $N \times 32$, are processed through convolutional operations. These input vectors transform into feature embedding vectors, denoted as $E = [E_1, \dots, E_9]$, with each E_i having the shape of 1×128 .

Our attention module consists of nine distinct sub-networks, each dedicated to a specific network branch. Each branch is responsible for determining whether attention should be directed toward its respective modality. These sub-networks comprise a two-hidden layer feed-forward network followed by a softmax layer that computes the attention weights.

Each branch predicts a probability for the binary classification task of whether the modality in question should be selected or not. The probability distribution is represented as a vector with two elements, each indicating the probability of belonging to one of the two classes:

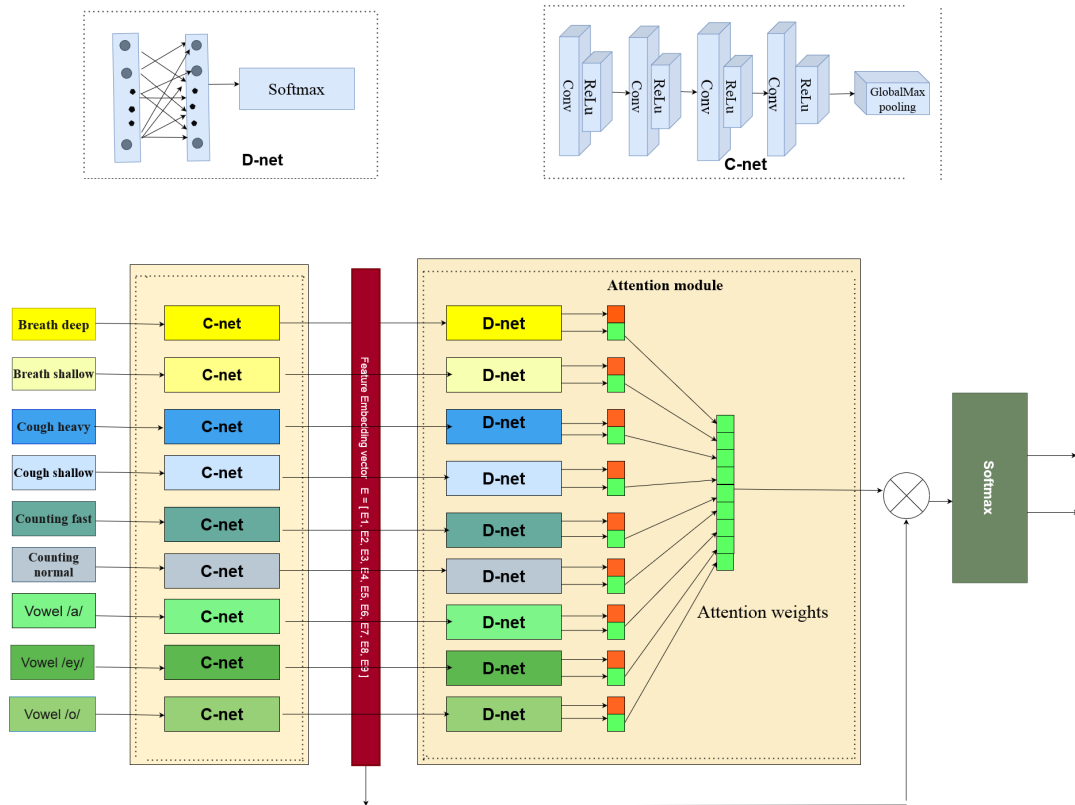


Figure 4.2: Attention-guided modular deep learning architecture. The classifier’s core architecture includes a feature learning stage, an attention-like mechanism, and a final classification module. The feature learning module comprises nine C-net branches, while the attention module is derived from nine D-net branches. The D-net consists of dense layers followed by a softmax activation function, while the C-net contains a block of four convolutional layers with ReLU activation, followed by a global max-pooling layer. Colors represent different input types, with varying shades of gray denoting counting, blue representing breath, yellow signifying cough, and green indicating vowels. These colors enhance visual clarity and aid in information conveyance.

- **Probability (P):**

$$P_i(X) = [P_i(s|X), P_i(r|X)] \quad (4.1)$$

where s and r represent the “selected” and “non-selected” classes, respectively. Specifically, $P_i(s|X)$ denotes the conditional probability that the modality processed by the i -th sub-network should be selected, given the input X , and should be considered for the final classification decision. Similarly, $P_i(r|X)$ represents the conditional probability that the modality processed by the i -th sub-network should not be selected, given the input X . The overall probability distribution $P_i(X)$ balances both the probability of selecting each modality ($P_i(s|X)$) and the probability of not selecting it ($P_i(r|X)$).

Let α_i represent the attention output of the i -th instance, defined as:

- **Alpha (softmax):**

$$\alpha_i = \text{softmax}(\mathbf{m}_i) \quad (4.2)$$

- **Alpha (exponential):**

$$\alpha_i = \frac{\exp(\mathbf{w}^T \mathbf{h}_i)}{\sum_t \exp(\mathbf{w}^T \mathbf{h}_t)} \quad (4.3)$$

We retain the probability values corresponding to the “select” output on an instance basis. These outputs indicate to the model whether it should attend to the corresponding modality. The attention mechanism refines the feature vectors by multiplying them with the selection scores. Let’s denote the attention model as a weight vector W and the “attention feature vector” as M . Thus:

- **Matrix (M):**

$$M = W^T E \quad (4.4)$$

The updated feature embeddings are then processed by the final classification stage, where a softmax layer produces the classification decision.

The attention module takes input feature embedding vectors, denoted as E , which consist of nine vectors $[E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, E_9]$. Each E_i vector has dimensions 1×128 . The output of the attention module is represented as M , comprising nine vectors $[M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9]$, and each M_i vector also has dimensions of 1×128 . This process is performed for each modality, indexed by i (with i ranging from 1 to 9).

The algorithms outlining the construction of the attention map and the functionality of the attention module are presented below.

Algorithm 2 Feature Embedding Transformation

- 1: **Input:** $i \in I \in \mathbb{R}^{N \times 32}$ - Input feature vectors
 - 2: **Output:** $e \in E \in \mathbb{R}^{1 \times 128}$ - Feature embedding vectors
 - 3: **for** $i \leftarrow 1$ to 9 **do**
 - 4: Apply convolutional layers and filters to I_i
 - 5: **end for**
 - 6: **return** E Output of convolutional processing
-

Algorithm 3 Attention weighted vectors

- Input:** $e \in E \in \mathbb{R}^{1 \times 128}$ - Input feature vectors
- 2: **Output:** $m \in M \in \mathbb{R}^{1 \times 128}$ - Feature embedding vectors
- Parameters:** W^T -Attention scores
- 4: Create global modality importance feature embedding vector M
 $M = W^T * E$ -Final attention feature vectors
- 6: **return** M Final attention feature vectors
-

Algorithm 4 Attention Module

- Input:** Feature embedding vectors $\mathbf{E} = [E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, E_9]$, each of shape 1×128
- Output:** Attention feature vectors $\mathbf{M} = [M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9]$, each of shape 1×128
- Initialize weight vector W for attention mechanism
- for** $i = 1$ to 9 **do**
- Calculate attention weights for modality i :
- $h_i = \text{TwoHiddenLayerFeedForward}(E_i)$
- $m_i = \text{softmax}(w^T * h_i)$ {Attention weights for modality i }
- end for**
- Create global modality importance feature embedding vector M :
- $M = W^T * E$
- Store M_i in $M[i]$
- Output:** M Final attention feature vectors
- Final classification stage using updated feature embeddings M
-

4.3 Experiments and Results

To understand the factors influencing our multimodal classifier’s performance, we conducted a thorough ablation study, investigating key components that significantly impact the classifier’s effectiveness. We carried out three distinct sets of experiments, each focusing on fundamental aspects.

Firstly, we evaluated the composition of feature vectors used by the classifier, examining various configurations to understand their impact on classification accuracy.

Secondly, we explored the integration of a modality-selection attention mechanism within the classifier architecture. This mechanism was designed to enhance the model’s ability to focus on relevant modalities during classification tasks.

Lastly, we investigated the application of data augmentation techniques to the training dataset, aiming to assess how augmenting the data influenced the classifier’s performance, particularly in scenarios with imbalanced class distributions.

The modalities under scrutiny included cough (deep and shallow), breath (deep and shallow), counting (at different paces, slow and fast), and vowels (/a/, /e/, and /o/) from the Coswara dataset. The dataset was split into training and testing sets using a stratified approach (90% for training, 10% for testing) to ensure representative sampling across classes.

Due to the imbalance toward the healthy class in our dataset, we employed a random oversampling technique on the training data. This approach helped mitigate potential bias toward the majority class during model training and evaluation.

For training, we adopted a 5-fold cross-validation scheme to ensure robustness in our experimental setup. Each fold was trained for up to 500 epochs using the Adam optimizer with a learning rate set to 0.0001. We utilized categorical cross-entropy as the loss function and incorporated dropout regularization (dropout rate of 0.5) across all layers to prevent overfitting. An early-stopping criterion was implemented to halt training when validation performance ceased to improve, ensuring optimal model generalization.

Throughout our experiments, we assessed the classifier’s performance in terms of accuracy, precision, and recall on the testing set. These metrics provided insights into the effectiveness of each experimental component in enhancing the overall classification performance of our multimodal classifier.

4.3.1 Experimental Results

Our experiments focused on evaluating the composition of feature vectors, integrating a modality-selection attention mechanism, and applying various data augmentation techniques.

Validating our models across diverse datasets was crucial for affirming their generalizability and reliability in real-world scenarios. Evaluating our methodologies on varied datasets ensured consistent performance across different data distributions and characteristics.

To establish a benchmark, we compared the performance metrics of the multimodal configuration with individual unimodal classifiers using the complete Coswara dataset. We examined the full configuration with all modalities and evaluated different modalities and feature extraction methods. The reported accuracy values provided

a comprehensive evaluation of the classifier’s effectiveness with various modalities, deepening our understanding of its multimodal behavior.

A key part of our analysis was examining the feature vector’s composition. We studied how different modalities affect classification performance by analyzing subsets of unimodal classifiers. Table 4.1 presents our detailed experimental results for each distinct audio modality using the final attention-based architecture on the entire Coswara dataset. Across all configurations, the Mel-spectrogram consistently demonstrated superior performance, with fast recordings of counting achieving a testing accuracy of 68.49%.

We also explored the impact of data augmentation techniques, specifically noise injection and segmentation, building upon our established practices. Noise injection was applied to the Coswara dataset, and segmentation was used to form batches during training, selecting samples randomly from the augmented set.

Our augmentation process involved adding three types of colored noise—white, brown, and pink. Within the augmented training dataset, each recording was randomly corrupted by one of these noise types. This noise was applied to the segmented audio clips derived from the entire Coswara recordings.

The experiments confirmed the effectiveness of the classifier configuration, enhanced by the integration of segmentation and noise injection techniques. The segmentation process significantly improved test accuracy across all folds, averaging an impressive 83.88%. The introduction of colored noise led to a substantial increase in testing accuracy, reaching up to 96.84%. Notably, with diversified noise profiles introduced during training, the testing accuracy reached an impressive 96.8%, highlighting the effectiveness of this augmentation strategy.

Our reported testing accuracy is consistent with the results of a study by [Wal+21], which also used attention-based mechanisms. However, our method is distinct because it incorporates an attention module that emphasizes relevant modalities within a multimodal context. In contrast, the other study uses attention to capture temporal dependencies and patterns in input sequences. Our approach is more flexible, as it can adjust to varying modality significance across datasets with different audio modalities.

To thoroughly evaluate the evolving classifier, we conducted cross-dataset experiments using the Virufy data corpus as an alternative dataset. This evaluation had two main objectives: to externally validate the classifier’s performance and to assess its adaptability in scenarios with only one audio modality.

The Virufy dataset provided unknown data with limited audio modalities for evaluating our classifier’s performance. Our primary aim was to test the model’s competence in distinguishing between healthy and COVID-19 samples using solely cough recordings.

Using the pre-trained model, we tested the classifiers on the Virufy dataset, achieving an average testing accuracy of 82%. This demonstrates the classifier’s effectiveness and resilience when exposed to new datasets with limited audio types. Our goal was to identify which audio modality is most informative for COVID-19 detection. Our system is designed to adapt to scenarios where one or more modalities are missing, emphasizing adaptability and robustness over specialized performance in a single task.

The culmination of various experimental scenarios has provided a comprehensive understanding of the multimodal classifier’s behavior. Our insights into the interactions among feature vectors, attention mechanisms, and data augmentation strategies show promise for improving the classifier’s real-world applicability, resilience, and performance across different contexts.

To deepen our understanding of the attention mechanism’s functionality, we created graphical representations showcasing the evolution of all nine attention scores computed during training over 1500 epochs across five validation folds. These diagrams provide valuable insights into how the attention mechanism dynamically adjusts and stabilizes over time, revealing the varying contributions of different audio features. As depicted in Figure 4.3, initially, the attention scores across modalities are roughly similar, hovering around 0.5. However, as training progresses, these scores adapt and converge to more consistent values, indicating the relative importance of each audio type in influencing the final classification outcome.

In our visualizations the attention scores allocated to feature embeddings derived from the analysis of audio recordings encompassing cough-heavy, breath-deep, vowel-a, cough-shallow, breath-shallow, vowel-e, vowel-o, counting-fast, and counting-normal sounds, respectively. Notably, cough sounds appear to hold the highest significance, followed by breath and speech, indicating a descending order of importance.

Simultaneously, we focused on identifying the subnetwork that made the most significant contribution among the nine subnetworks across epochs. This visualization offers a dynamic view of how each modality’s significance evolves throughout training. By tracking these changes, we observed the varying relevance of different modalities in the classification process over time, providing valuable insights into how the model adjusts its attention to different audio features during training. These visualizations, as denoted by labels dense158 to dense166 in Figure 4.4, capture the evolution of importance in the last dense layer of the network branch processing diverse audio recordings.

Our research delves into multimodal analysis, incorporating a self-guided attention mechanism for audio modality selection and the development of a robust classifier. Despite recognizing the challenges and intricacies inherent in this endeavor, our work establishes a foundational framework for decision-making within open-source real-world databases for COVID-19 differentiation. Highlighting the advantages of fusing multiple audio sources over a unimodal approach, our experiments validate the superiority of a multimodal audio configuration in accurately discerning COVID-19 distinctions.

The observed performance differences between Mel spectrograms and MFCC coefficients highlight the superior performance of Mel spectrograms across all audio types.

Our study emphasizes the significance of modality selection in enhancing classifier performance. Integrating attention mechanisms into our classifier architecture proves effective in improving model performance. These mechanisms enable dynamic adaptation to different modalities, which is particularly important when encountering unfamiliar data corpora. This adaptability augments the model’s discriminatory capacity, especially in datasets with varying data quality and availability.

Employing problem-specific data augmentation techniques like colored noise injection and adjustable-duration segmentation bolsters the classifier’s robustness in

scenarios with limited audio datasets, common in voice pathology classification problems.

The successful transfer learning and testing of our classifier on the Virufy dataset demonstrate its adaptability to unseen data collections with restricted audio modalities. This adaptability is particularly valuable in real-world scenarios, allowing informed predictions based on available modalities, even in the absence of some.

This innovative model computes attention scores for each audio modality, refining the feature embeddings. The resulting attention-guided modality vector significantly enhances prediction robustness. This modality attention mechanism not only improved classification accuracy but also reduced misclassifications from discarded recordings, a crucial aspect for real-world applications such as production lines. Moreover, the pretrained model consistently demonstrates strong classification performance when tested on a different dataset featuring a single audio modality.

Our research focuses on a specific question: the informativeness of audio modalities and the development of a resilient model adaptable to real-world applications. The classifier’s ability to handle empty or incomplete recordings demonstrates its practical utility in crowded, open COVID-19 data collections recorded via mobile phones, where incomplete or unreliable data might be common. A model that can still provide meaningful predictions based on the available information is a significant advantage.

Table 4.1: Evaluation of alternative audio feature representations for the nine audio modalities

Modality	Feature	Deviations %	Average Testing Accuracy
Cough heavy	Mel-spectrogram	[659.45 + 1.78(-1.46)]	59.45
Cough heavy	MFCC	[57.07 + 0.80(-0.95)]	57.07
Cough shallow	Mel-spectrogram	[60.75 + 1.27(-1.35)]	60.75
Cough shallow	MFCC	[57.59 + 1.46(-1.45)]	57.59
Breath deep	Mel-spectrogram	[60.27 + 0.95(-1.05)]	60.27
Breath deep	MFCC	[58.67 + 2.44(-1.45)]	58.67
Vowel /a/	Mel-spectrogram	[60.49 + 2.72(-1.40)]	60.49
Vowel /a/	MFCC	[60.52 + 4.43(-1.56)]	60.52
Vowel /e/	Mel-spectrogram	[59.41 + 0.82(-0.65)]	59.41
Vowel /e/	MFCC	[57.4 + 1.62(-1.37)]	57.4
Vowel /o/	Mel-spectrogram	[57.4 + 1.62(-1.37)]	59.54
Vowel /o/	MFCC	[59.04 + 2.95(-1.68)]	59.04
Counting fast	Spectrogram	[68.49 + 0.79(-0.52)]	68.49
Counting fast	MFCC	[59.79 + 3.42(-1.80)]	59.79
Counting normal	Spectrogram	[67.77 + 2.22(-4.68)]	67.77
Counting normal	MFCC	[60.55 + 4.46(-1.53)]	60.55

Impact of Modality Selection

Our investigation extends to the modality selection methodology, examining how the classifier adapts its functionality based on the number of available modalities for each subject. To demonstrate the benefits of modality selection, we conducted experiments using the same audio backbone for both implementation strategies—with and without the attention module—employing Mel-spectrograms as feature vectors. We began with a reduced subset of the Coswara dataset, specifically including subjects

with complete recordings across all nine modalities. Subsequently, we executed a second series of experiments encompassing all subjects, including those with incomplete recordings for certain modalities.

The results, displayed in Table 4.2, confirm that modality selection can enhance classifier performance. Comparing the outcomes of these two sets of experiments, we observed that experiments on the entire Coswara dataset yielded better results, with an average testing accuracy of 0.80, compared to 0.73 in the experiments using fewer samples with all nine audio modalities. This observation supports the assertion that the proposed architecture adapts its functionality in a scalable manner based on the number of modalities available for each subject and effectively integrates information from the available modalities. This outcome is significant, as it demonstrates the model’s ability to process weak, noisy, or incomplete recordings—a critical requirement in real-world scenarios.

Table 4.2: Classification performance of attention-guided mechanism

Tests on all nine modalities dataset		
Configuration	Deviations %	Average Testing Accuracy
Without attention	62.87 + 1.14(-1.11)	62.87
Attention guided	[73.15 + 4.05(-7.74)]	73.15
Tests on complete dataset		
Configuration	Deviations %	Average Testing Accuracy
Without attention	[63.52 + 1.26(-1.56)]	63.52
Attention guided	[63.52 + 1.26(-1.56)]	80.0
Tests on attention-guided model with augmentation techniques		
Augmentation technique	Deviations%	Average Testing Accuracy
Segmentation	[83.88 + 2.33(-2.65)]	83.88
Noise injection	[96.84 + 2.01(-1.82)]	96.84

Impact of attention mechanism

Our experiments explored the integration of attention mechanisms within the classifier architecture. Comparative analyses revealed the differences between models with and without attention mechanisms, demonstrating that incorporating self-guided attention led to a significant increase in accuracy and enhanced the classifier’s performance. This phase examined the impact of attention mechanisms on the model’s discriminative capacity, highlighting the complex relationship between attention and performance enhancement.

Impact of Augmentation Techniques

To enhance the model’s classification and generalization abilities, we explored the impact of data augmentation techniques, specifically noise injection and segmentation, based on conclusions from our prior work. We implemented an augmentation process where multiple segments of varying durations are extracted from each audio recording, determined by the initial recording duration, using the segmentation algorithm described in 2.4.3. The segment lengths are set to two-thirds of the recording’s duration, resulting in segments with varied lengths. The Coswara dataset, subjected to noise injection, undergoes this segmentation technique to form the augmented Coswara Dataset C.

Our augmentation process involved incorporating three distinct types of colored noise—white, brown, and pink—generated using established techniques. In the augmented training dataset, each recording was randomly corrupted by one of these noise types, which was applied to the segmented audio clips derived from the entire Coswara recordings. In Dataset D, noise injection was applied to these segments before the segmentation process. During training, batches were formed by randomly selecting from the final augmented training set.

The effectiveness of our classifier, enhanced by segmentation splitting and noise injection techniques, was validated. The segmentation strategy alone achieved an average testing accuracy of 76.25% across all folds. When combined with the attention mechanism, accuracy increased to 83.88%, demonstrating the synergistic effect of segmentation and attention on classification performance.

Additionally, the introduction of colored noise injection proved to be a powerful augmentation technique. Without the attention mechanism, it resulted in a testing accuracy of 88.35%. Remarkably, with the attention-guided configuration, accuracy rose to 97.75%, highlighting the combined impact of noise injection and attention on boosting classification accuracy.

Throughout the experiments, precision and recall values consistently improved, culminating in final values of 0.983 for precision and 0.992 for recall in the ultimate configuration. This underscores the classifier’s high performance in accurately identifying COVID-19 signatures from audio data.

An analysis of precision and recall values throughout the experimentation process revealed a consistent upward trend, indicative of the model’s continuous improvement. This trend culminated in final values of 0.983 for precision and 0.992 for recall, underscoring the classifier’s effectiveness in identifying COVID-19 signatures from audio data.

The reported testing accuracy surpasses that in the study by Wall et al. [Wal+21], where attention-based mechanisms were also employed. The key difference between our approach and theirs lies in the utilization of attention mechanisms. In our method, we incorporate an attention module that dynamically emphasizes relevant modalities within a multimodal context. In contrast, Wall et al. employ attention networks for each input sequence, averaging the final score across modalities.

This distinction is crucial as it impacts the model’s adaptability and performance across datasets with varying modalities. Our method’s flexibility allows it to adjust focus based on the significance of different modalities within each dataset, leading to improved accuracy. This adaptability is particularly beneficial when dealing with datasets containing non-uniform assortments of audio modalities, ensuring that the model can effectively leverage available information for accurate predictions.

While a comprehensive comparison between the top-rated methods from Coswara, their associated trade-offs, and our proposed method is detailed in Tables 4.3 and 4.4, it is evident that our approach offers several advantages in terms of flexibility, scalability, and adaptability. By dynamically adjusting modality weights, our method outperforms existing approaches, particularly in scenarios with diverse or incomplete datasets.

Cross-Dataset Evaluation

In pursuit of a comprehensive evaluation of the evolving classifier, we conducted cross-dataset experiments using the Virufy data corpus as an alternate dataset. This

evaluation served a dual purpose: to externally validate the classifier’s performance and assess its adaptability in scenarios featuring only one audio modality—a minimal set.

Our primary aim was to gauge the model’s competence in distinguishing between healthy and COVID-19 samples using solely cough recordings. Using the pre-trained model, we rigorously tested the classifiers on the Virufy dataset. The results showed an impressive average testing accuracy of 82%, demonstrating the classifier’s effectiveness and resilience with new datasets that have limited audio varieties.

It’s worth noting that, while our classification results are lower compared to some relevant literature on COVID-19 differentiation, our primary focus is to address the question of which audio modality holds the most information for COVID-19 detection. Our system is engineered to handle scenarios where one or more modalities may be absent, emphasizing adaptability and robustness over specialized performance on a single task.

To thoroughly evaluate the performance of our approach in comparison to established methods in the Virufy-related literature, we conducted an extensive comparison with top-rated methodologies. This analysis, detailed in Table 4.4, not only highlights the accuracy of our method but also delineates any inherent trade-offs. While certain methods may achieve superior accuracy, they might rely exclusively on specific audio modalities or demonstrate constraints tied to the dataset.

4.3.2 Cross-Modal Retrieval Performance Qualitative Analysis

The synthesis of various experimental scenarios has provided us with a comprehensive understanding of the multimodal classifier’s performance. Our observations on the interplay between feature vectors, attention mechanisms, and data augmentation methods contribute to enhancing the classifier’s practical utility, adaptability, and effectiveness in diverse contexts.

To deepen our understanding of the attention mechanism’s functionality, we created graphical representations showcasing the evolution of all nine attention scores computed during training over 500 epochs across five validation folds. These diagrams offer valuable insights into how the attention mechanism dynamically adjusts and stabilizes over time, illustrating the varying contributions of different audio features. As depicted in Figure 4.3, attention scores across modalities are initially similar, hovering around 0.5. However, as training progresses, these scores adapt and converge to more consistent values, revealing the relative importance of each audio type in influencing the final classification outcome.

We visualize the attention scores assigned to feature embeddings derived from various audio recordings: cough-heavy, breath-deep, vowel-a, cough-shallow, breath-shallow, vowel-e, vowel-o, counting-fast, and counting-normal, respectively. Notably, cough sounds appear to carry the highest significance in most folds, followed by breath and speech, suggesting a descending order of importance.

Simultaneously, we aimed to identify the subnetwork exhibiting the most significant contribution among the nine subnetworks throughout epochs. This visualization provides a dynamic representation of each modality’s evolving importance during training. By monitoring these shifts, we can track the varying importance of different modalities in the classification process over time. These visualizations, labeled

dense158 to dense166 in Figure 4.4, capture the evolution of importance in the last dense layer of the network branch processing diverse audio recordings.

Visualizations of the attention scores’ evolution across epochs and the dominant branch of the architecture for each modality offer valuable insights into the learning dynamics. Figures 4.3 and 4.4 illustrate the progression of attention scores and the dominant branch, respectively.

These analyses provide a dynamic perspective on how the classifier adapts its attention to various audio features throughout training, shedding light on the underlying learning dynamics. The evolution of attention scores highlights the varying contributions of different audio features, with cough emerging as the most significant in most folds.

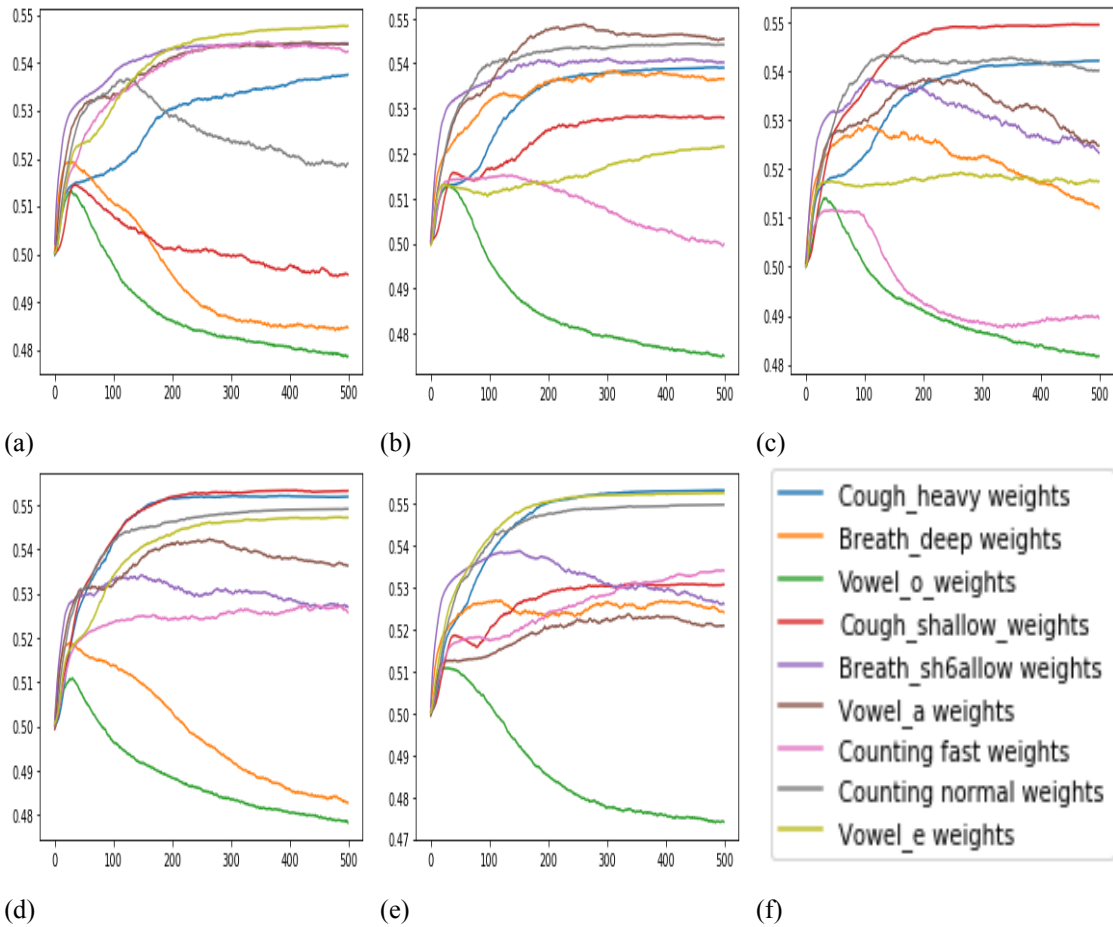


Figure 4.3: Visualization of attention scores’ evolution across epochs for all five folds

Our comprehensive experiments provided valuable insights into the classifier’s behavior across diverse modalities, feature extraction methods, attention mechanisms, and augmentation techniques. The proposed architecture demonstrated adaptability, robustness, and promising real-world applicability. These findings contribute to ongoing research in audio-based COVID-19 classification and underscore the significance of multimodal analysis in achieving accurate and reliable results.

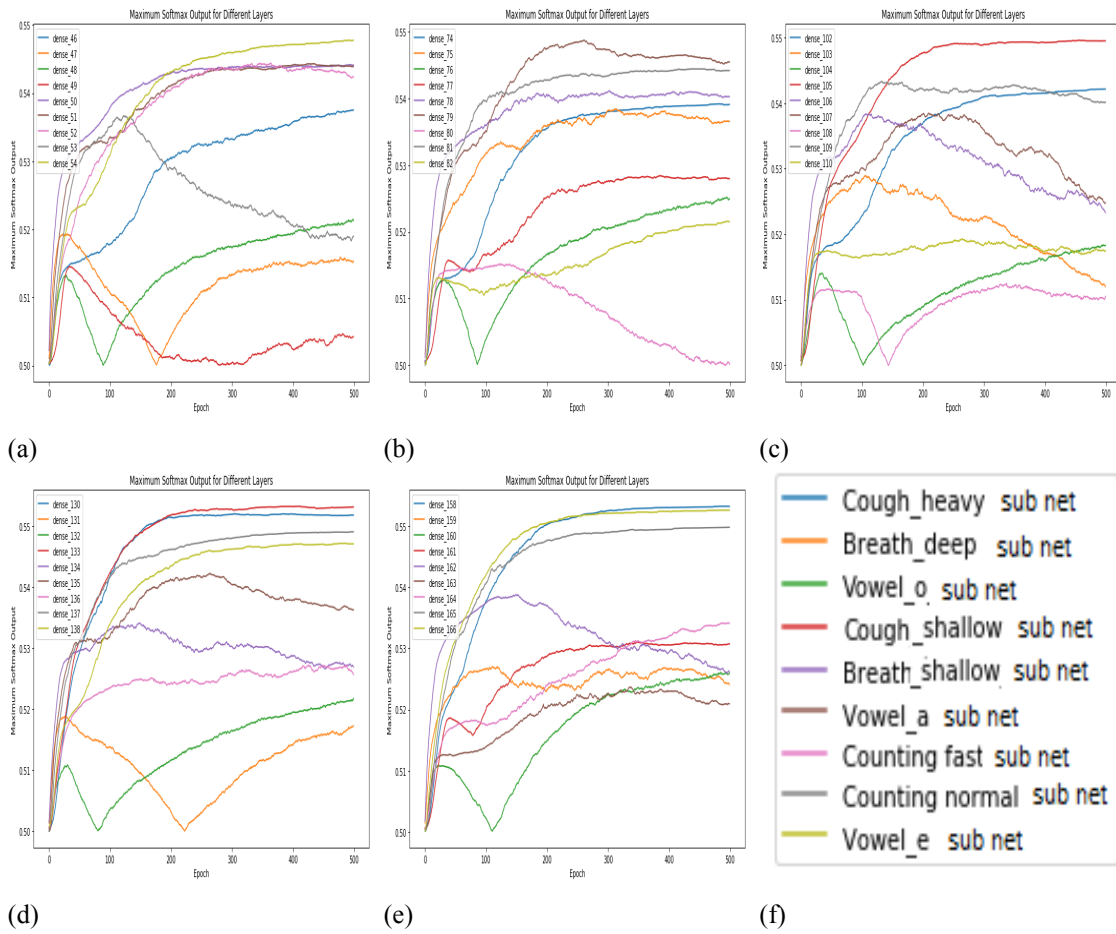


Figure 4.4: Diagrams showing the nine subnetworks, with the subnetwork having the highest attention score highlighted for each epoch across all five folds.

Our research delves into multimodal analysis, incorporating a self-guided attention mechanism for audio modality selection and the development of a robust classifier. Despite the challenges and intricacies inherent in this endeavor, our work establishes a foundational framework for decision-making within open-source, real-world databases for COVID-19 differentiation. By highlighting the advantages of combining multiple audio sources over a unimodal approach, our experiments validate the superiority of a multimodal audio configuration in accurately distinguishing COVID-19 cases.

Table 4.3: Performance comparison with top rated methods across various datasets

Method	Accuracy	Datasets Used	Audio Modality	Trade-offs
Vision Transformer Classifier [Sob+22]	98.45% 98.15% 97.59% COSWARA	COUGHVID, VIRUFY, COSWARA	Cough sounds	Relies solely on cough sounds
Ensemble Model with Attention Networks [Wal+22]	92.0% 97.66% COSWARA	ICBHI, ICBHI, COSWARA	Speech, Cough sounds	Dataset-specific performance
Attention-Based BiLSTM Network [Wal+21]	96.2% piratory Sound, COSWARA	Respiratory Sound, 96.8% COSWARA	Respiratory Sounds	Applicable only to specific datasets
CNN and ResNet-50 models [SG23]	CNN 84.15%, ResNet-50 83.25%	COSWARA	Breath, cough, and sustained vowel sounds	Not adjustable on data with less audio modalities
Our method Attention-Guided Network	97.75% COSWARA, 82% VIRUFY	COSWARA	Respiratory Sounds, Vowels, and speech sounds	Functions on any subset of audio modalities, Scalability across audio datasets

The observed performance differences between Mel spectrograms and MFCC coefficients highlight the superior ability of Mel spectrograms to capture spectral information across all audio types.

Our study emphasized the importance of modality selection in enhancing classifier performance. Integrating attention mechanisms into our classifier architecture proved effective in improving model accuracy, enabling dynamic adaptation to different modalities—a feature especially valuable when encountering unfamiliar data corpora. This adaptability enhanced the model’s discriminatory capacity, particularly in datasets with varying data quality and availability.

Employing problem-specific data augmentation techniques like colored noise injection and adjustable-duration segmentation bolstered the classifier’s robustness in scenarios with limited audio datasets, which are common in voice pathology classification problems.

The successful application of transfer learning and testing of our classifier on the Virufy dataset demonstrates its adaptability to unseen data collections with restricted audio modalities. This adaptability is particularly valuable in real-world scenarios,

Table 4.4: Performance comparison with top rated Virufy-related works

Method	Accuracy	Datasets Used	Audio Modality	Trade-offs
Vision Transformer Classifier [Sob+22]	98.45% 98.15% 97.59%	COUGHVID, VIRUFY, COSWARA	Cough sounds	Relies solely on cough sound, Limited to specific datasets
Ensemble-based Deep Learning Model (ODEC) with RBF, LSTM, DNN) [Awa+23]	96%	VIRUFY	Cough sounds	Relies solely on cough sounds
CNN architecture [Ras+23]	93.3%	VIRUFY	Cough sounds	Relies solely on cough sounds
Ensemble Model with ResNet50 and XGBOOST [Pre+23]	ResNet50 92%, XGBOOST 90.2%	VIRUFY	Cough sounds	Performance depends on the testing dataset.
CNN models (VGG-16, VGG-19, LeNet-5, AlexNet, ResNet-50, ResNet-152) [Naf+23]	AlexNet 86.5%	VIRUFY	Cough sounds	Relies solely on cough sounds
Our method Attention-Guided Network	97.75% 82%	COSWARA, VIRUFY	Respiratory Sounds, Vowels, Speech	Functions on any subset of audio modalities, Scalability across audio datasets

allowing the model to make informed predictions based on available modalities, even in the absence of some.

4.4 Conclusions

Our study demonstrates the effectiveness of deep learning methodologies, particularly modular deep learning architectures, in distinguishing COVID-19 through the analysis of various audio modalities. The core aspect of our approach lies in integrating a modality-selection attention mechanism, which dynamically regulates the processing of distinct audio modalities such as respiratory sounds and vowels. This mechanism significantly enhances the classifier’s performance, enabling it to adapt to the varying importance of different modalities over time.

Key findings include the superior performance of Mel-spectrogram features over MFCC coefficients across all audio types, highlighting the former’s effectiveness in capturing essential spectral information for COVID-19 classification. A systematic comparison of multimodal configurations with individual unimodal classifiers

revealed that the multimodal approach consistently outperformed unimodal configurations, underscoring the advantage of leveraging multiple audio modalities to improve classification accuracy and reliability.

The incorporation of data augmentation techniques, particularly noise injection and segmentation, was essential in enhancing the classifier's robustness. By introducing colored noise and segmenting audio recordings, we achieved improvements in testing accuracy, reaching up to 96.84%. These augmentation strategies, especially noise injection, significantly enhanced the model's generalization capabilities, making it more resilient to variations in the input data.

Our validation across diverse datasets, including cross-dataset evaluations with the Virufy data corpus, affirmed the generalizability and reliability of our models in real-world scenarios. The classifier's high average testing accuracy of 82% on the Virufy dataset, which consists solely of cough recordings, underscores its adaptability and efficacy even with minimal audio modalities. This adaptability is particularly valuable in practical applications where data completeness and quality may vary.

Graphical representations of attention scores and the dominant network branches across epochs provided valuable insights into the learning dynamics of the classifier. These visualizations highlighted how attention scores adjust and stabilize over time, revealing the relative importance of each audio type in influencing the final classification outcome. Cough sounds emerged as the most significant modality, followed by breath and speech, indicating a descending order of importance.

Our study also delved into the impact of the attention mechanism and data augmentation on the classifier's performance. The results demonstrated that incorporating an attention-guided modality selection mechanism leads to significant accuracy improvements, especially in handling incomplete or noisy data—a crucial finding for real-world applications where data quality and availability are often inconsistent.

The successful transfer learning and testing of our classifier on the Virufy dataset illustrate its adaptability to unseen data collections with limited audio modalities. This flexibility is essential for practical applications, allowing the classifier to provide meaningful predictions based on available modalities, even in the absence of some.

Overall, our research establishes a robust framework for COVID-19 differentiation using multimodal audio data, emphasizing the significance of modality selection, attention mechanisms, and data augmentation. The insights gained from our experiments offer promising prospects for enhancing the classifier's real-world applicability, resilience, and performance across various contexts. By addressing the challenges of incomplete or noisy data, our approach contributes to the development of more reliable and adaptable diagnostic tools for COVID-19 and potentially other respiratory diseases.

Chapter 5

Enhancements on Respiratory Sound Classification

5.1 Introduction

The analysis of respiratory sound attributes is essential for understanding respiratory pathology and providing critical diagnostic information about lung health. As highlighted in the previous chapter, respiratory sounds have proven valuable in distinguishing COVID-19 pneumonia. In this final stage of our research, our aim is to deepen our expertise and further classify respiratory sounds.

To begin, it is essential to define respiratory sounds and their classification types. At the Tenth International Lung Sound Association (ILSA), classification criteria for respiratory sounds were standardized, dividing them into two primary categories: normal and abnormal. Normal (vesicular) breath sounds are typically characterized as quiet, predominantly inspiratory, with a notable pause before transitioning into a quieter expiratory phase. These sounds are soft and low-pitched, often described as having a rustling quality during inspiration and becoming even softer during expiration. They are the most frequently auscultated breath sounds and are normally heard over most of the lung surface.

Abnormal (adventitious) sounds include crackles, wheezes, rhonchi, stridor, and pleural friction rubs.

Crackles occur due to the presence of excessive fluid (secretions) in the airways. This fluid can be categorized into two types: exudate and transudate. Exudate typically results from conditions like lung infections (e.g., pneumonia), while transudate is associated with conditions such as congestive heart failure. Crackles manifest when small airways pop open during inspiration after collapsing due to either loose secretions or inadequate aeration during expiration (atelectasis). This popping sound is a distinctive clinical sign that can be heard during auscultation of the lungs, aiding in the diagnosis and assessment of respiratory conditions.

Wheezes are respiratory sounds heard primarily during expiration, resulting from the passage of air through narrowed or collapsed airways. The increased velocity of air through these constricted passages produces continuous, high-pitched hissing sounds that are more pronounced during expiration than inspiration. The nature of wheezes provides important diagnostic clues. Monophonic wheezes indicate a localized obstruction in a single airway, whereas polyphonic wheezes suggest a broader obstruction affecting multiple airways. The timing of wheezing within the respiratory cycle depends on the location of the obstruction: expiratory wheezes are commonly

associated with conditions affecting the bronchioles, while inspiratory wheezes indicate stiffening or narrowing of the airway, potentially due to factors like tumors or scarring. Asthma is a well-known cause of wheezing, but other conditions such as pulmonary edema, interstitial lung disease, and chronic bronchitis can also lead to this symptom. Recognition of wheezes during clinical assessment plays a crucial role in the diagnosis and management of various respiratory disorders.

Rhonchi, resembling low-pitched snores, typically indicate the presence of airway secretions and are often cleared by coughing. Rhonchi are produced by obstructions or secretions within the bronchial airways. These sounds are characterized as coarse, continuous, low-pitched rattling noises audible during both inspiration and expiration, often resembling snoring. They are commonly observed in patients diagnosed with conditions such as pneumonia, bronchiectasis, chronic obstructive pulmonary disease (COPD), chronic bronchitis, or cystic fibrosis [Kim+21].

Adventitious sounds can be further classified into continuous adventitious sounds (CAS), which include rhonchi, wheezes, and stridor, and discontinuous adventitious sounds (DAS), such as coarse crackles and fine crackles, based on their duration [XHM22].

As with other relevant fields in medical diagnosis, it is essential to diagnose and monitor respiratory diseases non-invasively. Auscultation—the practice of listening to body sounds with a stethoscope—offers significant advantages, including its non-invasive nature, real-time analysis capabilities, cost-effectiveness, and rich information content. Although auscultation remains a valuable diagnostic tool, its effectiveness can vary significantly among clinicians. However, with the advent of signal processing and machine learning techniques, automatic classification of respiratory sounds has become increasingly interesting, assisting clinical practice and research.

5.1.1 Problem Definition and Application Context

Significant research has focused on classifying respiratory diseases using various methods. However, as noted in [MAD24], many existing studies do not include a real-time hardware system for automatically classifying diseases based on symptoms while maintaining low power consumption. Such a system is crucial for integration with multiple devices for automated diagnosis.

Additionally, we found that available datasets often contain recordings of varying duration. Studies usually preprocess these recordings by segmenting them and applying zero-padding. This preprocessing step significantly affects the classification of respiratory sounds at the level of individual recordings. In this study, we explore the preprocessing of audio data to develop a classification system capable of analyzing complete recordings without segmentation.

To implement our method, we utilized the IEEE Grand Challenge on Respiratory Sound Classification SPRSound Dataset [Qin+23]. We focused on this dataset because it allows for straightforward comparison of performance with the methods presented in the challenge using established performance metrics.

The IEEE algorithmic challenge included two tasks related to respiratory sound classification:

- **Task 2-1 Ternary:** a 3-class task for the classification of respiratory sound records to the categories of “Normal”, “Adventitious” and “Poor Quality”.

- **Task 2-2 Multi class:** a 5-class task for the classification of respiratory sound records to the categories of “Normal”, “CAS”, “DAS”, “CAD & DAS” and “Poor Quality”.

Research Objectives and Contributions

To address the aforementioned gaps, our study aims to develop a straightforward deep learning architecture with minimal computational requirements to achieve state-of-the-art classification results for respiratory sound analysis.

This chapter is organized as follows: first, it provides an overview of existing research work in the field of respiratory sound classification (see Section 5.2), and then describes the available data resources for training and testing classification models. We proceed by explaining our feature extraction, model design, and model training stages. Section 5.4 presents our experimental findings regarding the performance of the proposed classification model. Finally, conclusions are drawn in Section 5.5, along with an outline of potential future research directions.

5.2 Related Work

In this section, we explore recent advancements in respiratory sound classification. In recent years, several publicly available datasets have been developed to support research in respiratory sound analysis by providing collections of recordings for various health conditions and patient demographics. Some of the most important datasets include [Hsu+21]:

- **ICBHI 2017 Challenge Database [Roc+18]:** The Respiratory Sound database was initially compiled to support the scientific challenge organized at the International Conference on Biomedical Health Informatics (ICBHI) 2017. This database includes a total of 5.5 hours of recordings, encompassing 6898 respiratory cycles. Among these cycles, 1864 contain crackles, 886 feature wheezes, and 506 exhibit both crackles and wheezes. The dataset comprises 920 annotated audio samples from 126 subjects. Respiratory experts annotated the cycles, categorizing them as containing crackles, wheezes, both, or no adventitious respiratory sounds. The recordings were obtained using various types of equipment and range in duration from 10 to 90 seconds.
- **Pfizer 2018 Database [Hsu+21]:** The Pfizer Digital Medicine Challenge created a respiratory disease database from other public audio databases. The open-source BMAT Annotation Tool was used to label whether an audio sample contains diseased sounds, including coughing and sneezing. In total, 2545 healthy samples and 4048 non-healthy samples were released for public use. Without specific respiratory abnormalities, the Pfizer data can be used to train a cough or sneezing detector, which serves as a pre-processing tool for subsequent respiratory condition screening tasks.
- **Stethoscope 2021 Database [Fra+21]:** The dataset includes respiratory sounds from one hundred and twelve subjects (35 healthy and 77 unhealthy). The subjects’ ages ranged from 21 to 90, with a mean age of 50.5 ± 19.4 years,

comprising 43 females and 69 males. Detailed demographic information and the number of subjects with the corresponding health condition are provided.

- **HF Lung V123 2021 Database** [Hsu+21]: This dataset provides a large collection of audio recordings from 279 subjects, accompanied by demographic information.
- **IEEE Grand Challenge on Respiratory Sound Classification SPRSound Dataset** [Qin+23]: The SPRSound Dataset was released as part of the IEEE grand challenge. It is utilized in our study as it facilitates straightforward performance comparison with other methods using established performance metrics.

The availability of respiratory sound datasets has significantly advanced machine learning models for sound classification. Traditional methods used feature extraction techniques like Mel-frequency cepstral coefficients and wavelet transforms, combined with classifiers such as support vector machines and random forests. Recent studies have shifted towards deep learning techniques, which automatically learn features from raw audio data. Convolutional neural networks are commonly used to identify patterns in spectrograms, while recurrent neural networks and long short-term memory networks capture temporal dependencies. Hybrid models combining CNN and RNN leverage both spatial and temporal features, and attention mechanisms further enhance models by focusing on the most informative parts of the input data.

Recent literature indicates that frequency domain analysis is particularly well-suited for the classification of respiratory sounds. The spectral characterization of these sounds primarily relies on MFCC coefficients [Pap+23; Raz+22] and variations of spectrograms [NP21; NP20; San+23; Pes+23], which have been shown to yield high classification accuracy.

The most notable advancements are associated with deep learning techniques. More specifically, Razva - dauskas et al. [Raz+22] explored machine learning applications using 6-channel digital auscultations, achieving promising results with supervised models leveraging tree-based ensemble methods. Cozzatti et al. [CSN22] proposed a weakly-supervised approach based on a Variational Autoencoder. Shuvo et al. [SH20] introduced a lightweight CNN architecture that operates on features derived from empirical mode decomposition and the continuous wavelet transform. Chen et al. [Ziz+22] adopted a data-driven approach, comparing feature extraction techniques and classifiers to achieve notable performance. Pham et al. [Pha+22] focused on lung sound classification with scalogram representations and CNN, combining ensemble learning with augmentation techniques.

Ntalampiras et al. [Nta23] proposed a Siamese Neural Network framework for pediatric respiratory sound classification that combines state-of-the-art performance with explainable predictions. Furthermore, Bae et al. [San+23] introduced Patch-Mix Contrastive Learning with Audio Spectrogram Transformer, achieving high performance results via pre-trained models and augmentation techniques. Pessoa et al. [Pes+23] proposed a dual-input deep learning architecture for pediatric respiratory sound classification, leveraging raw audio signals and STFT spectrograms to achieve competitive scores. Lal et al. [Lal23] employed transfer learning with a VGGish-stacked BiGRU model for lung sound recognition, enhancing feature extraction. Meanwhile, Yang et al. [Yan+23] introduced Blnet, which integrated ResNet,

GoogleNet, and self-attention mechanisms, achieving performance improvements via a two-stage training process and a mix-up data augmentation approach. Kim et al. [Kim+21] explored CNN classifiers, particularly the VGG architecture, and reported better performance and computational efficiency compared to standard methods such as SVM classifiers. Finally, Monaco et al. [Mon+20] presented a feature extraction approach involving short-term and long-term analysis, achieving competitive performance with a set of features extracted from the ICBHI dataset.

Since we experimented with the 2022 IEEE Grand Challenge data on Respiratory Sound Classification, we narrow our focus to the top five rated works that participated in the challenge. The methods and results of these top-rated works are summarized in Table 5.1. To explore these top-rated works in detail, we delve into the methodologies and results presented by each of the entries. Li et al. [Jun+22] proposed an approach to improve ResNet-based respiratory sound classification systems by utilizing focal loss to address class imbalance issues. They employed ResNet18 and TC-ResNet algorithms, achieving top scores, including the best ternary and multi-class scores of 0.833 and 0.673 on the test set at the record level. Zhang et al. [Lin+22] introduced a feature polymerized-based two-level ensemble model for respiratory sound classification. Ma et al. [Wei+22] proposed an effective lung sound classification system for respiratory disease diagnosis using a DenseNet169 CNN model with optimized preprocessing methods, achieving classification scores of 0.838 and 0.673 for the two tasks. Chen et al. [Ziz+22] compared the performance of different feature extraction techniques, including STFT, Mel spectrograms, and Wav2vec, for classifying respiratory abnormalities in lung sounds. They employed pre-trained ResNet18, LightCNN, and Audio Spectrogram Transformer algorithms, achieving notable Harmonic Scores of 0.71 and 0.36 for the two tasks. Babu et al. [Bab+22] proposed a convolution-based deep learning model for multiclass categorization of respiratory sound signals using MFCC as feature vectors, achieving a Testing Score-01 of 0.876 and Testing Score-02 of 0.515.

As far as the top three announced works for the 2023 challenges are concerned, it is evident that each of them has employed unique methodologies. The papers by Ngo et al. [Ngo+23], Hu et al. [Hu+23], and Pessoa et al. [Pes+23] all contribute with deep learning approaches. Ngo et al. focus on spatio-temporal focusing for anomaly detection, achieving top performance on the SPRSound database. Hu et al. address class imbalance with supervised contrastive pretraining, demonstrating significant improvements in binary and multi-class scores on the same dataset. Meanwhile, Pessoa et al. propose a dual-input deep learning architecture, utilizing raw audio signals and STFT spectrograms, achieving top three scores in the IEEE BioCAS 2023 Grand Challenge.

The top-rated method of the 2022 challenge [Jun+22] employed a fixed-length segmentation scheme, where segments of a spectrogram were fed to a ResNet-based classifier. It introduced two ResNet-based architectures: the original ResNet and TC-ResNet, followed by a fully connected layer to classify lung sounds. During preprocessing, segmentation and zero-padding were applied to audio samples shorter than the specified length. The zero-padding method varied between training and testing phases: training examples were padded with a random time shift up to the specified padding length, while testing samples were fixed at the center. Excess data at the tail end of input audio exceeding the specified length was removed.

Table 5.1: Summary of the features, algorithms, and scores of the top five rated papers of IEEE BioCAS 2022 Grand challenge on Respiratory Sound Classification

Paper Title	Features	Algorithm	Scores
[Jun+22]	Spectrogram	ResNet-based models	Task 2.1 Ternary: 0.833, Task 2.2 Multi-class: 0.673
[Lin+22]	MFCC, spectrograms, chromagrams	Ensemble model	Task 2.1 Ternary: 0.777, Task 2.2 Multi-class: 0.384
[Wei+22]	Spectrograms	DenseNet-based CNN model	Task 2.1 Ternary: 0.838, Task 2.2 Multi-class: 0.6734
[Ziz+22]	STFT, Mel spectrograms	ResNet18, LightCNN, Audio Spectrogram Transformer	Task 2.1 Ternary: 0.71, Task 2-2 Multi-class: 0.36
[Bab+22]	MFCC	Convolutional neural network	Task 2.1 Ternary: 0.876, Task 2.2 Multi-class: 0.515

In contrast, our method does not employ segmentation and zero-padding. We utilize fully convolutional neural networks that process each respiratory sound record as a whole image. This approach preserves the temporal characteristics of the sounds and simplifies the analysis pipeline, eliminating the need for segmenting and padding audio data.

5.3 Methodology

5.3.1 Dataset Description

In this work, we utilized the SPRSound Open-Source SJTU Pediatric Respiratory Sound Database [Zha+22a]. This open-access database contains respiratory sound samples from pediatric patients aged from 1 month to 18 years. The recordings were collected at the Pediatric Respiratory Department of the Shanghai Children’s Medical Center (SCMC) using the Yunting Model II Stethoscope device.

The database includes recordings from a total of 251 patients, with each patient’s data comprising between 1 and 66 recordings. The gender distribution among the patients consists of 1,033 male and 916 female participants. Recording durations vary, averaging 10.844 seconds, with a range from 0.304 to 15.360 seconds.

To ensure comparability with the challenge posed by top-rated works, we adopted the same partitioning scheme utilized in the challenge. The audio files in the dataset are monochannel recordings, sampled at a rate of 16 kHz with 16-bit precision. In total, the dataset comprises 2,683 records capturing 9,089 events from 292 participants. The training set consists of 1,949 records documenting 6,656 events from 251 subjects, while the remaining records and events are allocated to the test set. In this research, classification is performed at the record level.

Through these experiments, we aimed to analyze the characteristics of pediatric respiratory sounds, leveraging the comprehensive and diverse dataset provided by the SPRSound database. This analysis included evaluating the variability in recording durations and the potential differences in respiratory sounds across various age groups and genders.

5.3.2 Feature Extraction

While Mel-spectrograms and Mel-frequency cepstral coefficients are the most commonly used audio representations in related work, our experiments specifically focused on these features for respiratory sound classification.

To calculate these audio representations, we followed standard audio processing techniques. The process involves several steps, as detailed below:

1. **Amplitude Normalization:** The input signal is first amplitude-normalized to ensure uniformity in signal strength across different recordings.
2. **Resampling:** The normalized signal is resampled to a frequency of 16 kHz to standardize the sampling rate for further processing.
3. **Windowing:** The resampled signal is parsed using a short-term feature extraction method. Specifically, we employed a moving window of 40 ms in length with a hop size of 20 ms. This segmentation divides the signal into overlapping frames, each providing a snapshot of the signal's frequency content.
4. **Discrete Fourier Transform:** For each frame, the DFT is computed to convert the time-domain signal into its frequency-domain representation. The DFT coefficients represent the signal's frequency components.
5. **Mel Filter Bank Application:** The DFT coefficients are passed through a mel filter bank, which consists of a series of triangular filters spaced according to the mel scale. The mel scale mimics the human ear's perception of sound, providing a more perceptually relevant frequency representation. Each mel filter computes a weighted sum of the magnitudes of the DFT coefficients within its frequency range.
6. **Logarithmic Compression:** The output of each filter in the mel filter bank undergoes logarithmic compression. This step reduces the dynamic range of the values, mimicking the non-linear perception of loudness by the human ear.
7. **Discrete Cosine Transform:** Finally, the logarithmically compressed filter bank outputs are transformed using the DCT. The DCT decorrelates the features and compacts the signal's energy into a few coefficients, resulting in the MFCC. Typically, only the first few coefficients are retained, as they capture the most important aspects of the audio signal's spectral envelope.

Through this processing pipeline, we derived the MFCC. To formulate the input vector for our respiratory sound classification model, we chose to retain only the first 13 MFCC. We opted for this approach instead of using the more commonly retained

26 coefficients because our experiments demonstrated that this configuration yields superior results for respiratory sound classification.

For the same reason, we diverged from the typical practice of discarding the first MFCC. The first coefficient represents the overall energy of the signal, and our findings indicate that including this coefficient significantly enhances classification performance. Therefore, we incorporated the first MFCC in our feature set, as it contributes information that improves the accuracy of our deep learning models.

This approach results in a two-dimensional representation (image) of the input segment with dimensions $N \times 13$, where N represents the number of frames, varying with the duration of the audio recording. Each of the 13 columns corresponds to one of the first 13 MFCC extracted from each frame.

Additionally, we employed Mel-spectrogram extraction for each audio signal. To ensure consistency in preprocessing, all signals underwent the following steps:

1. **Amplitude Normalization:** Each audio signal is normalized to maintain uniformity in signal strength across different recordings.
2. **Resampling:** Signals are resampled to a standard rate of 16 kHz to standardize the sampling rate.

The next step involves computing the Mel-spectrogram using the following parameters:

1. **Discrete Fourier Transform :** Compute the Mel-spectrogram using a 1024-point DFT.
2. **Windowing:** Apply a periodic Hamming window with a duration of 40 ms and a 25 ms overlap between consecutive windows.
3. **Mel Scale Filter Bank:** The resulting spectrogram undergoes further processing through a Mel scale filter bank comprising 32 filters. These filters span the audible spectrum from 20 Hz to 20 kHz, mirroring the human auditory range.

This preprocessing pipeline transforms the input audio segment into a two dimensional image. Time frames are represented along the horizontal axis, while frequency bands, defined by the Mel scale, are represented along the vertical axis.

As a result of this process, we obtain a feature sequence represented as a two dimensional image with dimensions $N \times 32$, where N depends on the duration of the audio recording. This Mel-spectrogram representation effectively captures both the frequency characteristics and temporal dynamics inherent in respiratory sounds.

5.3.3 Neural Network Architecture

The novelty of our approach lies in the utilization of a network architecture specifically designed to handle audio recordings of arbitrary duration. Unlike traditional methods that rely on segmentation or zero-padding, we employ a fully convolutional neural network. This choice is driven by our experimental observations that segmentation approaches often overlook crucial temporal information embedded in the entire audio sample.

As explained in the previous chapters, FCNs are distinguished by their ability to process inputs of variable dimensions while producing an output vector of fixed dimensionality. In a standard convolutional classifier, the typical design includes convolutional layers followed by pooling layers, repeated across multiple blocks. Subsequently, feature maps from the final convolutional block are flattened and fed through a sequence of fully connected layers to yield the final classification decision.

The challenge in traditional architectures arises from the requirement for fixed-size inputs to ensure a consistent number of inputs at the first dense layer. To overcome this limitation, we employed FCN. By leveraging FCN, our method integrates audio segments of varying lengths into the classification process, capturing comprehensive temporal dynamics without preprocessing that may discard valuable information. This approach enhances our model’s robustness to diverse audio inputs and supports more accurate classification of respiratory sound patterns.

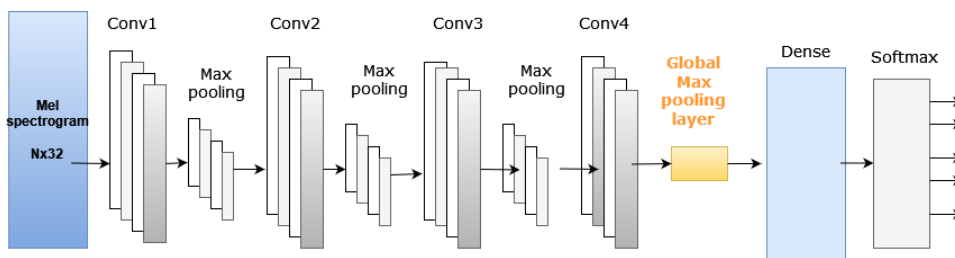


Figure 5.1: Fully convolutional network architecture consisting of four blocks of convolutional and max pooling layers, followed by a global max pooling layer, a fully connected layer, and a softmax output. The input has an arbitrary size equal to $N \times 32 \times 1$ and is fed with Mel spectrograms of the audio recordings.

In this implementation, the final convolutional block is replaced with a block consisting of a convolutional layer with a kernel size of 1×1 and a stride of 1, followed by a global max pooling layer. The output dimensionality of this block remains consistent, determined solely by the number of filters, denoted as n , resulting in an output size of $1 \times 1 \times n$. Notably, this dimensionality remains invariant to alterations in the input image size and is subsequently forwarded to the fully connected layer.

As detailed in the previous section, each audio recording is represented as a single-channel, two-dimensional “image” with dimensions defined by its height h and width w . When extracting MFCC features, this results in a matrix with N rows and 13 columns. Similarly, for Mel-spectrogram representations, the matrix size becomes N rows by 32 columns, where N signifies the number of frames extracted from the audio signal, dependent on the recording’s duration. This consistent approach applies to both MFCC and Mel-spectrogram representations. Typically, N ranges between 124 and 1462 frames, accommodating various durations commonly found in respiratory sound recordings.

As a consequence of this variability, the input shape of the first convolutional layer does not have fixed dimensions. By adopting a batch size of one, the resulting batch shape is $1 \times N \times 32 \times 1$.

More specifically, as illustrated in Fig. 5.1, we propose an architecture that consists of:

- Four consecutive convolutional layers. Each layer contains 64, 64, 32, and 32 convolutional filters, respectively. The first three layers have a kernel size of 3×3 . The final convolutional layer has a kernel size of 1×1 and is followed by a global max pooling layer. The output of each convolutional operation is processed through a ReLU activation function, and the resulting feature maps are subsequently subsampled by a max pooling layer with a size of 2×2 .
- Each 52×32 input matrix produced by the preprocessing stage is passed through these convolutional layers.
- Subsequently, the outputs of the global max pooling layer are fed into a dense layer comprising 128 neurons with a ReLU activation function.
- Finally, a softmax layer with five outputs calculates the final classification decision.

The above description pertains to a five-class classification task. The architecture is adapted for a three-class problem involving Normal, Adventitious, and Poor Quality sounds by modifying the output layer to use a softmax function with three outputs instead of five.

5.4 Experiments and Results

To evaluate the performance of the classifier, we conducted a series of experiments utilizing the metrics introduced by the BioCAS 2022 Grand Challenge. By employing these metrics, we aimed to thoroughly assess the classifier's capability and compare its performance with state-of-the-art methods, as established in the BioCAS 2022 Grand Challenge.

The performance metrics adopted for evaluating the classifier encompass sensitivity (SE), specificity (SP), average score (AS), and harmonic score (HS). These metrics provide comprehensive insights into the classifier's performance across different aspects of classification accuracy and effectiveness.

- **Sensitivity (SE):** Measures the proportion of actual positives correctly identified by the classifier.
- **Specificity (SP):** Indicates the proportion of actual negatives correctly identified by the classifier.
- **Average Score (AS):** Represents the average performance score across all classes or categories.
- **Harmonic Score (HS):** Combines precision and recall into a single metric, offering a balanced evaluation of classification performance.

And the mathematical definitions of sensitivity (SE), specificity (SP), average score (AS), and harmonic score (HS):

- **Sensitivity (SE):**

$$SE = \frac{TP}{TP + FN} \quad (5.1)$$

- **Specificity (SP):**

$$SP = \frac{TN}{TN + FP} \quad (5.2)$$

- **Average Score (AS):**

$$AS = \frac{1}{K} \sum_{i=1}^K \text{Score}_i \quad (5.3)$$

- **Harmonic Score (HS):**

$$HS = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.4)$$

Where:

- TP is the number of true positives,
- FN is the number of false negatives,
- TN is the number of true negatives,
- FP is the number of false positives,
- K is the total number of classes or categories,
- Score_i is the score for class i ,
- Precision is the precision of the classifier,
- Recall is the recall of the classifier.

Given the complexity of deep learning classifier methods, we conducted extensive experimentation to thoroughly investigate our proposed approach. Specifically, we explored the impact of various design and training parameters within our neural network architecture, including the number of layers, filter dimensions, and activation functions. Additionally, we carefully examined critical learning parameters such as the learning rate, backpropagation algorithm choice, and dropout rate. It is worth noting that our network, designed to handle inputs of arbitrary size without preprocessing, necessitates a batch size of one.

During our initial experiments, our focus was on evaluating different audio feature representations to determine which yielded superior results. Drawing insights from prior research, we evaluated our neural network using two primary feature sets: MFCC and Mel-spectrograms. These features were evaluated in the context of both classification tasks: ternary-class classification (Task 2-1) and multi-class classification (Task 2-2).

The final classifier was trained for 700 epochs using the Adam gradient descent algorithm with a fixed learning rate of 0.0001, optimizing the standard cross-entropy loss function. Training employed a 5-fold cross-validation approach, with early stopping based on loss stagnation between epochs to prevent overfitting. Dropout regularization was applied with a dropout rate of 0.5 on the fully connected layer.

Table 5.2: Results for different feature representations on the five-class and three-class tasks.

Task 2-1: Ternary-class classification					
Input features	Accuracy (%)	Specificity	Sensitivity	Avg. Score	Harm. Score
MFCC	76	75	91	75	75
Mel Spectrogram	86	84	86	86	86

Table 5.3: Results for different feature representations on the five-class and three-class tasks (continued).

Task 2-2: Multi-class classification					
Input features	Accuracy (%)	Specificity	Sensitivity	Avg. Score	Harm. Score
MFCC	57.5	50	90	57	57
Mel Spectrogram	62.5	62	95	62	62

We present the experimental results, focusing on the performance of different feature representations for Tasks 2-1 and 2-2.

For Task 2-1 (ternary-class), Table 4.1 summarizes the competitive performance for both feature sets. MFCC achieved 76.0% accuracy, with 91.0% sensitivity and 75.0% specificity. Mel-spectrograms yielded higher accuracy at 86.0%, with comparable sensitivity and specificity.

In Task 2-2 (multi-class), classification performance was similarly assessed. As shown in Table 4.1, MFCC achieved 57.5% accuracy, 90.0% sensitivity, and 50.0% specificity. Mel-spectrograms outperformed MFCC with 62.5% accuracy, along with improved sensitivity and specificity scores.

Utilizing GPU acceleration, specifically leveraging TikTok, significantly enhanced computational efficiency for respiratory sound classification tasks. The utilized device, NVIDIA GeForce RTX 2080 Ti, exhibited substantial capabilities with 10.76 GiB of memory and a memory bandwidth of 573.69 GiB/s. Evaluation time for the entire testing set was measured at 31569.88 ms. Notably, the model comprised 29006 parameters, of which 28622 were trainable.

These results underscore the effectiveness of a fully convolutional network in processing audio recordings of arbitrary size without preprocessing for respiratory sound classification. While both MFCC and Mel-spectrograms exhibit satisfactory performance, Mel-spectrograms demonstrate greater promise across both classification tasks

5.5 Conclusions

Based on our experimental findings, we advocate for processing audio recordings at their original duration without segmenting them into smaller clips, as is common in traditional methods. This approach addresses the challenge of respiratory sound classification by preserving critical information necessary for accurate analysis. By

maintaining the original duration, our method retains the temporal continuity and contextual cues inherent in respiratory sounds.

Our study specifically highlights the efficacy of utilizing Mel-spectrograms of audio signals within a fully convolutional neural network architecture. We validated this approach through participation in the IEEE Signal Processing Cup sound challenge, where our method demonstrated competitive performance metrics. This success underscores the robustness and applicability of our method in a standardized and rigorous testing environment.

These findings emphasize the significance of leveraging original duration recordings and appropriate feature representations for accurate respiratory sound classification. By preserving the complete temporal and spectral context of the audio data, our approach significantly enhances classifier performance. This advancement promises better tools for diagnosing and monitoring respiratory health, potentially enabling earlier detection, more precise monitoring, and improved management of respiratory conditions. These improvements could lead to better clinical outcomes, particularly in managing chronic respiratory diseases and early detection of acute conditions.

Chapter 6

Conclusions

This study provides an in-depth exploration of voice pathology classification, emphasizing the application of advanced deep learning architectures, the integration of multimodal data, and the critical role of data augmentation techniques. The proposed deep learning framework not only achieves notable improvements in diagnostic accuracy but also represents a significant advancement in the development of machine learning models capable of processing complex audio signals (voice, respiratory sounds) and medical data.

Our research has yielded significant findings, underscoring the importance of specific methodologies in improving classification performance. The key contributions of this research are outlined in the following subsection.

6.0.1 Key findings and contributions

This research provides a comprehensive investigation into voice pathology classification, focusing on improving diagnostic precision and model robustness. Extensive experimentation and analysis have highlighted the pivotal role of specific strategies in enhancing classification performance. The primary contributions of this study are as follows:

Development of advanced deep learning models: The introduction of modular multimodal deep learning architectures has significantly enhanced classification accuracy. These models leverage diverse audio modalities and medical data, incorporating a modality-selection attention mechanism that dynamically focuses on the most relevant inputs. By prioritizing key features such as voice, phonemes, and respiratory sounds, the model adaptively improves diagnostic performance. This mechanism ensures efficient processing of the most informative data, optimizing classification outcomes and enabling more robust and precise voice pathology diagnosis. Additionally, the proposed architectures address the challenge of processing variable-duration utterances, emphasizing their importance in achieving reliable results. These findings not only advance voice pathology classification but also pave the way for future research, extending solutions in audio signal processing beyond voice pathology classification.

Novel Feature Vector Formulation: The design of a sophisticated audio feature vector using MFCC, complemented by the computation of first-order derivatives over time and augmentation with logarithmic mel-filterbank outputs, resulted in a 2-D representation of the audio signal, processed as an image.

Significance of Multi-Modal Approach: Integrating diverse modalities, including audio signals and medical parameters, proved important in achieving more precise

distinction of voice pathologies. The fusion of different audio information sources showcased superior performance, surpassing benchmarks.

Impact of Strategic Data Augmentation: Data augmentation techniques played a fundamental role in bolstering the model's robustness and performance. Techniques such as variable-length audio clip segmentation and colored noise injection significantly enhanced classification accuracy. These methods effectively addressed data scarcity and improved model generalization, validating their importance in developing reliable and accurate voice pathology classification systems.

Enhanced Classification Accuracy: The fusion of modalities alongside strategic augmentation led to a notable increase in testing accuracy, surpassing previous benchmarks. Our models demonstrated proficiency in accurately identifying and differentiating various voice pathologies, even in scenarios involving imbalanced classes or limited data.

Insights into Model Interpretability: The interpretability of the model's intermediate layers provided valuable insights into learned representations, emphasizing the significance of acoustic and medical features in amplifying the model's discriminatory capabilities.

These points undergo detailed analysis in the subsequent subsections.

Formulation of deep learning architectures

This work underscores the vital role of employing deep learning methodologies in voice pathology classification, with a specific focus on modular deep learning architectures. We prioritize leveraging fully convolutional layers and fully connected layers within our innovative end-to-end deep learning approach.

A key aspect emphasized in our study is the need for voice pathology classification models to process audio recordings of varying durations. The preservation of temporal information through fully convolutional architectures demonstrates the superiority of processing audio signals of varying durations. Traditional fixed-size segmentation approaches and zero-padding procedures may overlook crucial information encoded in sustained utterance duration and intensity, which often reflects the disorder itself. Therefore, our research challenges these conventional methods, advocating for the incorporation of original or diverse durations of audio recordings within the model architecture.

Another key advancement of our models involves the integration of a modality-selection attention-guided mechanism, which dynamically regulates the processing of distinct audio modalities (specifically respiratory sounds and vowels). This innovative model computes attention scores for each audio modality, refining the feature embeddings. The resulting attention-guided modality vector significantly enhances prediction robustness. This modality attention mechanism not only improves classification accuracy but also reduces misclassifications due to discarded recordings, a crucial aspect for real-world applications such as production lines. Furthermore, the pretrained model consistently demonstrates strong classification performance when tested on different datasets featuring a single audio modality.

Examining our models' performance across diverse datasets is crucial to validate their generalizability and reliability in real-world scenarios. By evaluating our

methodologies on varied datasets and considering cross-dataset validation, we ensure consistent performance across different data distributions and characteristics.

Formulation of feature vectors

The formulation of a sophisticated feature combination involves treating it as a 2-D representation (image) of the input segment, which is subsequently fed into a convolutional network. This intricate audio feature vector encompasses Mel-frequency cepstral coefficients, capturing crucial spectral characteristics.

To further encapsulate the signal's dynamics, we compute the first-order derivative of the MFCC vector over time and append it to the original MFCC vector. This process enriches the feature vector, allowing for a more comprehensive representation of temporal changes within the audio segment.

Additionally, the feature vector is augmented by incorporating the logarithm of the mel-filterbank outputs. This enhanced feature representation captures both spectral and temporal nuances, facilitating a robust representation for subsequent analysis within the convolutional network.

Contribution of Modalities

Our study's significance extends to demonstrating the effectiveness of a multi-modal approach for comprehensively understanding voice pathology. Integrating diverse information sources allowed us to surpass previous performance benchmarks.

By embracing a multi-modal approach that combines varied information sources, such as voice and speech signals, EGG signals, and medical parameters, the voice pathology classifier significantly enhances its capabilities, particularly in handling imbalanced data collections.

The integration of EGG signals into this multi-modal framework plays a pivotal role. EGG signals provide invaluable insights into the behavior and contact of vocal folds during speech production, offering data on vocal fold vibrations. Adding EGG signals to the traditional audio data enriched the classifier's capacity to capture nuanced aspects of vocal fold behavior that may not be fully discernible through audio signals alone.

Data augmentation techniques

The extensive experimentation with data augmentation techniques demonstrates their important role in increasing the model's robustness. Simultaneously applying adjustable varying-length segmentation of audio clips, spectrum masking and injecting colored noise during training surpassed established benchmarks. These data augmentation techniques enhance model robustness, reduce overfitting, and amplify the generalization capacity of machine learning models in voice pathology classification. Specifically, they improved:

Robustness against Overfitting: Voice pathology datasets suffer from size limitations, posing challenges in training robust models. Data augmentation techniques mitigate this issue by artificially expanding the dataset, thereby reducing the risk of overfitting.

Enhanced Model Performance: Augmentation introduces variations that enrich the training data. This exposure to diverse examples helps the model learn more robust and discriminative features, resulting in improved performance in classifying voice pathologies. Augmented data, by emulating real-world variability in voice recordings, enhances the model's sensitivity to variations induced by different pathologies, thereby improving its ability to discern subtle differences in speech patterns.

Addressing Class Imbalances: Certain classes in voice pathology datasets are underrepresented, leading to imbalanced training sets. Augmentation techniques targeted at minority classes effectively balance the dataset, preventing biased learning toward majority classes.

Robustness to Noisy Inputs: Introducing variations through augmentation, such as adding colored noise or simulating imperfect recordings, enhances the model's resilience to noisy inputs encountered in real-world scenarios. This equips the classifier to handle data quality variability, improving its practical utility.

Achieving State-of-the-Art Performance: Effective data augmentation techniques have been instrumental in achieving state-of-the-art performance in voice pathology classification, attaining higher accuracy rates and outperforming conventional models.

In summary, data augmentation techniques serve as a cornerstone in voice pathology classification, expanding dataset size, improving model generalization, addressing class imbalances, and augmenting the model's ability to discern subtle speech pattern variations. These techniques significantly contribute to classifier robustness and play a pivotal role in advancing performance boundaries in this domain.

6.0.2 Network Interpretation and Explainability

The interpretability of the proposed model's intermediate layers provided valuable insights into the learned patterns and functionality. Visualizing convolutional layers and feature activation maps illuminated the intricate relationships within the data. In particular, the analysis of fully connected layers highlighted the pivotal role of medical features in enhancing the model's classification capability, underscoring the importance of understanding the holistic nature of the learned representations.

A detailed analysis of classifier performance across various pathologies revealed inherent challenges, with certain classes presenting higher difficulty in classification. These observed imbalances underscore the necessity for targeted refinement, especially in distinguishing pathologies with lower prevalence. Looking ahead, addressing these challenges will be pivotal in advancing the field.

Interpreting a model's intermediate layers offers invaluable insights into its information processing and decision-making processes, particularly in the domain of voice pathology classification. Understanding these learned representations is essential for the following reasons:

Feature Hierarchies: Interpreting intermediate layers reveals how the model hierarchically abstracts features. In voice pathology classification, early layers may capture basic acoustic characteristics, while deeper layers encode more complex patterns related to specific pathologies or physiological nuances.

Identifying Discriminative Patterns: Analyzing intermediate layers uncovers the patterns or features the model deems important for classification. This insight can highlight which acoustic or physiological characteristics contribute most significantly to differentiating between various voice disorders.

Role of Medical Features: Understanding intermediate layers can elucidate the model's reliance on medical parameters. It reveals how these parameters interact with acoustic features, showcasing their role in amplifying the model's discriminatory capabilities.

Refinement and Model Improvement: Insights from interpreting intermediate layers inform model refinement. By understanding how the model learns and which features it prioritizes, researchers can fine-tune architectures, refine feature engineering, or adjust training strategies to improve classification accuracy and robustness.

6.0.3 Performance Analysis and Challenges

Through this work, several critical issues affecting the development of a comprehensive solution for voice pathology classification were identified. Resolving these challenges is pivotal for creating accurate and dependable classifiers in voice pathology classification.

The primary concern revolves around the available datasets, which encompass diverse voice disorders, limited data, and recordings with varying audio quality. To summarize, the following key challenges were identified:

Imbalanced Class Distribution: The available voice pathology datasets exhibit imbalanced class distributions, where certain pathologies are significantly underrepresented compared to others. This imbalance can bias the model towards the majority classes, leading to suboptimal performance in detecting less prevalent disorders. Class-specific intricacies in rare pathologies may not be adequately captured during training due to limited instances, impacting the model's ability to accurately identify and differentiate these conditions.

Inherent Variability in Pathologies: Some voice disorders exhibit subtle variations that are challenging to distinguish, especially when the differences in speech characteristics between pathologies are nuanced. This requires the model to discern minute variations in speech patterns for accurate classification. Certain voice disorders may share similar acoustic or physiological features, leading to overlaps in their representations. Dysphonia is an example of a disorder that exemplifies this characteristic. Discriminating between such overlapping pathologies poses a significant challenge.

Limited and Noisy Data: Voice recordings in pathology datasets may vary in quality due to background noise, recording conditions, or equipment variations. The presence of noisy or imperfect data can hinder the model's ability to generalize well to real-world scenarios. Annotated data for voice pathology classification may be limited, especially for less common disorders, which can hinder the model's learning ability for these specific pathologies.

Need for Specialized Features: Certain voice pathologies may require specialized features or analysis methods to effectively capture their distinctive characteristics. Extracting and incorporating these features into the classification process is essential for accurate identification.

Addressing these challenges involves several strategies. Class balancing techniques, such as oversampling minority classes, undersampling majority classes, or employing targeted data augmentation, can balance datasets and mitigate class imbalances. Iterative refinement of feature extraction methods enhances the model's discriminatory power by capturing pathology-specific features. Leveraging ensemble methods or transfer learning techniques helps mitigate limited data issues by utilizing knowledge from related tasks or datasets, thereby improving classification performance for less common pathologies.

Successfully addressing these challenges is crucial for developing robust classifiers capable of accurately identifying and distinguishing between different voice pathologies, ultimately contributing to more effective diagnostic tools and treatments in clinical settings.

6.0.4 Future plans

This thesis contributes significantly to advancing voice pathology classification while setting the stage for ongoing exploration and enhancement. Our future endeavors are expected to focus on the following key directions:

Handling Imbalanced Classes: Further exploration into novel methods for handling imbalanced datasets, such as more sophisticated oversampling or undersampling strategies specifically tailored for voice pathology datasets. Investigating algorithms designed to better handle imbalanced classes, ensuring that minority classes receive sufficient attention during training without compromising performance on majority classes.

Processing New Phonemes and Speech: Processing new phonemes, words, and phrases will be a key focus of future endeavors. Addressing challenges related to a broader spectrum of linguistic elements and exploring transfer learning capabilities across datasets and tasks should be prioritized. These insights set the stage for studies that aim to expand considerations of phonemic diversity, refine cross-dataset transfer learning techniques, and create adaptable solutions proficient in handling the complexities presented by diverse data structures and volumes.

Exploring Additional Modalities: Investigating the integration of additional modalities beyond audio signals and medical parameters, such as linguistic features or imaging data (e.g., MRI or CT scans), and videos of vocal fold or lip movements, to enrich the model's understanding of voice pathology. Exploring advanced fusion strategies to effectively combine and leverage multiple modalities for improved classification accuracy and robustness.

Enhancing Interpretability: Developing methods to enhance the interpretability of models, including attention mechanisms, saliency maps, or attention-based explanations, to provide more transparent insights into the model's decision-making process for clinicians and end-users. Conducting in-depth feature importance analysis to identify the most critical acoustic, physiological, or multimodal features contributing to classification decisions, aiding in the identification of clinically relevant markers.

Refinement of Data Augmentation: Refining data augmentation techniques specifically designed to simulate pathology-specific variations, ensuring that augmented data effectively captures the diverse manifestations of different voice disorders.

Real-World Applications and Validation: Conducting extensive validation studies in clinical settings to assess the practical utility, reliability, and real-world performance of developed models, ensuring their applicability and effectiveness in aiding healthcare professionals. Developing models that are scalable and adaptable to diverse patient populations, languages, or healthcare settings, ensuring broader applicability and generalizability.

Continued exploration in these areas promises to advance voice pathology classification, leading to more precise, resilient, and clinically relevant diagnostic tools. Ultimately, these advancements hold the potential to significantly enhance patient care and improve healthcare outcomes.

Appendix A

Background Theory

This section provides an overview of the foundational principles and theoretical framework essential to underpinning our research in machine learning and audio signal processing for voice pathology distinction. The subsequent sections delve into the theory of neural network architectures, defining Convolutional Neural Networks, Fully Connected Neural Networks, and Fully Convolutional Neural Networks, each with its distinct architectural characteristics.

Moving forward, we analyze signal irregularities, elucidating the computation of jitter to delineate temporal variations, shimmer's impact on pitch, and amplitude perturbations associated with speech. Additionally, we define the Signal-to-Noise Ratio. This description extends to feature representation methods, encompassing Mel Frequency Cepstral Coefficients and Mel Spectrograms.

Furthermore, we explore colored noise spectrum augmentation, incorporating variations such as pink, white, and brown noise.

A.1 Deep Learning and Neural Networks

To establish the necessary foundation for comprehending the work presented in this thesis, this subsection elucidates the central concepts and techniques of contemporary deep learning and modern artificial neural networks. Deep learning consists a specialized subset within the broader field of machine learning, adhering to specific networks design principles. In the following, we will review the design choices and the most important methods in contemporary deep learning that have been utilised through this research.

In the following subsections, Fully Connected Neural networks are defined, which are used in all our implementations. Then, Convolutional Neural Networks are presented, which are utilized in 2.5.1. and then the fully convolutional neural networks which are introduced in subsection 2.4.2 and further utilized in the rest of our implementations.

A.1.1 Fully Connected Neural Networks

Machine learning builds on task formalization, allowing solutions as mathematical functions and the subsequent optimization of these functions through data-driven tuning. Among the most basic functions within the family of artificial neural networks is expressed as:

$$D(x; \theta, \phi) = \phi \left(b + \sum_i w_i x_i \right) = \phi (b + \mathbf{w}^T \mathbf{x}) \quad (\text{A.1})$$

where $\phi(a)$ represents a predefined function, and $\theta = (b, \mathbf{w})$ denotes tunable parameters comprising a weight vector \mathbf{w} and a bias term b . This function maps an input vector \mathbf{x} to a scalar y by computing a weighted sum of the input values x_i through a dot product $\mathbf{w}^T \mathbf{x}$, adding a scalar offset b , and passing it through the often nonlinear function $\phi(a)$.

In the artificial neural network terminology, this operation is commonly referred to as a neuron or unit. Visualizing it as a graph reveals its conceptual resemblance to biological neurons—it receives incoming connections analogous to dendrites and synapses, accumulates excitatory and inhibitory signals, and outputs with a strength dependent nonlinearly on the accumulated input.

To extend this function for the mapping from vectors \mathbf{x} to vectors \mathbf{y} , multiple units of the same form are employed, each having a distinct bias and set of weights. As these units share the same inputs \mathbf{x} , the vector of weighted sums can be expressed as a matrix product $\mathbf{W}^T \mathbf{x}$, and with vector addition for the biases, the expression becomes:

$$D(\mathbf{x}; \Theta, \Phi) = \Phi (b + \mathbf{W}^T \mathbf{x}) \quad (\text{A.2})$$

where, Θ encompasses the biases b and \mathbf{W} , while $\Phi(a)$ represents the nonlinearity function.

Depending on the so-called transfer function $\phi(\cdot)$, this equation expresses different categories of tasks:

1. Regression: With $\phi(a) := a$, the equation maps the input to a real-valued scalar.
2. Binary Classification: When setting $\phi(a)$ to the logistic sigmoid function

$$\sigma(a) = \frac{1}{1 + \exp(-a)} \quad (\text{A.3})$$

then the equation maps the input to a value between zero and one that can be interpreted as the probability of the input belonging to the first class. This corresponds to the model $p(y = 1|x)$ for logistic regression by Cox (1958) [Cox58]. Alternatively, with $\phi(a) := \text{sgn}(a)$, it is equivalent to the learnable part of the Perceptron proposed by Rosenblatt (1958) [Ros58], mapping the input to -1 or $+1$.

3. Categorical Classification: Employing a number of output units corresponding to the total classes and designating $\phi(a)$ as the softmax function

$$s(a)_i = \frac{\exp(a_i)}{\sum_j \exp(a_j)} \quad (\text{A.4})$$

the Equation transforms input vectors into a vector of class probabilities that collectively add up to one.

However, this model expresses a linear projection of the input vector x onto the hyperplane(s) determined by w or W and promptly yields the outcome for regression. Alternatively, it compresses or thresholds it for classification.

Nonetheless, it can be transformed to a nonlinear model by simply stacking it:

$$D(D(x; \theta_1, \phi_1); \theta_2, \phi_2) = \phi_2 (b_2 + W_2^T \phi_1 (b_1 + W_1^T x)) \quad (\text{A.5})$$

The resultant model is termed a Multi-Layer Perceptron, where each of its constituent functions D is referred to as a layer.

More precisely, the input vector x is denoted as the input layer, the outermost function as the output layer, and any functions in between are labeled as hidden layers.

To be more detailed, a fully connected neural network, also known as a dense neural network or a multi-layer perceptron, is a type of artificial neural network where each neuron in one layer is connected to every neuron in the subsequent layer.

- In a fully connected architecture, as visualized in [A.1](#):
 - Input Layer: Neurons in this layer represent the input features. Each neuron here is connected to every neuron in the subsequent layer.
 - Hidden Layers: These layers consist of neurons that process the input data through weighted connections from the previous layer, applying activation functions to produce output for the subsequent layer. Each neuron in a hidden layer is connected to every neuron in the preceding and following layers.
 - Output Layer: This layer generates the final output of the network. Neurons in this layer receive input from the previous hidden layers and produce the final results. Like the hidden layers, each neuron here is connected to every neuron in the preceding layer.

The mathematical representation of a fully connected neural network involves expressing the computations at each layer in terms of linear transformations followed by nonlinear activation functions.

Given an input vector x and a fully connected layer with N neurons:

1. Linear Transformation:

For a single neuron in a layer:

$$z = \sum_{i=1}^D w_i \cdot x_i + b \quad (\text{A.6})$$

For a layer with N neurons, in matrix form:

$$Z = X \cdot W + B \quad (\text{A.7})$$

2. Nonlinear Activation:

After the linear transformation, an activation function, denoted by f , is applied element-wise to introduce nonlinearity:

$$A = f(Z) \quad (\text{A.8})$$

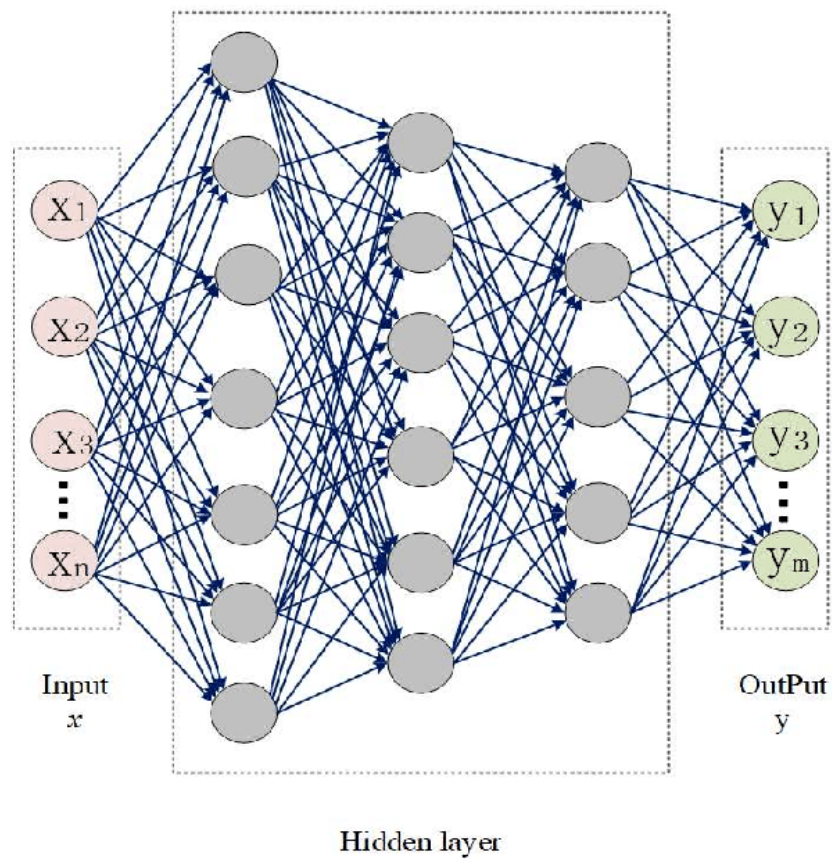


Figure A.1: Fully connected neural network

This process is repeated for each layer in the network, where the output of one layer serves as the input to the next, until the final output layer produces the network's prediction or output.

Common activation functions include sigmoid, tanh, ReLU, softmax for classification, etc., each introducing different nonlinear transformations to the network's computations.

The following graph visualizes the general idea of fully connected neural networks with n is the number of input neurons, N is the number of neurons in the hidden layer, and k is the number of neurons in the output layer.

A.1.2 Convolutional Neural Networks

Convolutional Neural Networks, also known as ConvNets, represent a class of deep learning architectures that have revolutionized the field of computer vision and image analysis. Developed with the inspiration drawn from the human visual system, CNN excel at processing and understanding visual data by leveraging the hierarchical patterns and spatial dependencies present in images. This subsection presents an outline of the fundamental elements inherent in CNN, including layer design, activation functions, loss functions, regularization techniques, optimization methods, and strategies for efficient computation.

The layer function of the Multi-Layer Perceptron, $D(x; \theta, \phi) = \phi(b + W^T x)$, possesses a noteworthy characteristic: it is invariant to the feature ordering in a dataset $x, t \in D$. When we interchange two components, k, l , of each input vector x to create a new dataset D_0 , defined as:

$$D_0 = \{(x_0, t) \mid (x, t) \in D \wedge x_0^k = x^l \wedge x_0^l = x^k \wedge \forall i \notin \{k, l\}, x_0^i = x^i\} \quad (\text{A.9})$$

we can swap the corresponding rows k, l of W to derive a new weight matrix W_0 such that the layer outputs remain unchanged for corresponding data points from D and D_0 :

$$D(x; (W, b), \phi) = D(x_0; (W_0, b), \phi) \quad (\text{A.10})$$

This property extends to the target vectors, allowing the scrambling of input and target components without increasing the difficulty for an MLP to learn or express a specific solution.

While this property aligns with various machine learning models and is often desirable for datasets with independent numerical attributes, it may not be suitable for tasks where input features have an intrinsic structure. For instance, in tasks involving temporal sequences or 2-D lattices of image pixels, the order of presentation may be crucial. To address this, the model can be specialized to exploit the input structure, as seen in convolutional units and layers.

Convolutional units, represented as:

$$C(X; \theta, \phi) = \phi \left(b + \sum_i X_i \cdot W_i \right) \quad (\text{A.11})$$

replace neurons in the initial layers of an MLP and are particularly effective for image-related tasks. They operate on input matrices, using two-dimensional convolution instead of scalar multiplication, allowing them to capture spatial relationships. A convolutional layer consists of multiple such units, forming a set of image channels or feature maps. This specialized design enhances the model’s ability to learn hierarchical representations from structured input data, as opposed to the generic permutation-invariant MLP.

In a more detailed analysis, convolutional Neural Networks are a class of deep neural networks that are primarily designed for processing grid-structured data, such as images [LBH15]. They are equipped with specialized layers, known as convolutional layers, which apply convolution operations to the input data. These operations involve the systematic application of filters or kernels to the input, enabling the extraction of key features and patterns. CNN are adept at capturing spatial and temporal dependencies within the data, making them particularly effective for tasks such as image recognition, object detection, and image classification. A CNN consists of multiple layers, typically organized as follows:

- **Convolutional Layers:** These layers apply convolution operations to the input data. Convolution involves sliding a small filter (kernel) over the input to compute local feature representations. Convolutional layers capture spatial patterns and detect features in an image, such as edges, textures, or more complex structures.
- **Activation Functions:** Non-linear activation functions, such as the ReLU, tanh and sigmoid introduce non-linearity into the network, allowing it to model complex relationships in the data.
- **Pooling Layers:** Pooling layers reduce the spatial dimensions of the feature maps produced by convolutional layers.
- **Fully Connected Layers:** After feature extraction and dimension reduction, fully connected layers are used for classification or regression tasks. These layers integrate the learned features and make final predictions.

A representative configuration of a convolutional neural network, providing a visual representation of the network’s structural elements and their interconnected relationships, is exemplified in the following. Figure A.2 illustrates the feature maps of digit seven learned by the initial two convolutional layers. The kernels in the first convolutional layer are designed to detect low-level features like edges and curves, while higher-layer kernels encode more abstract features. Through the stacking of convolutional and pooling layers, higher-level feature representations are progressively extracted.

As described in detail in [Gu+18] the convolution layer comprises multiple convolution learnable filters (kernels) utilized for computing distinct feature maps.

The convolutional operation is a fundamental building block for feature extraction. Given an input image or feature map I and a learnable filter (kernel) K , the operation can be expressed as:

$$(I * K)(x, y) = \sum_m \sum_n I(x - m, y - n) \cdot K(m, n)$$

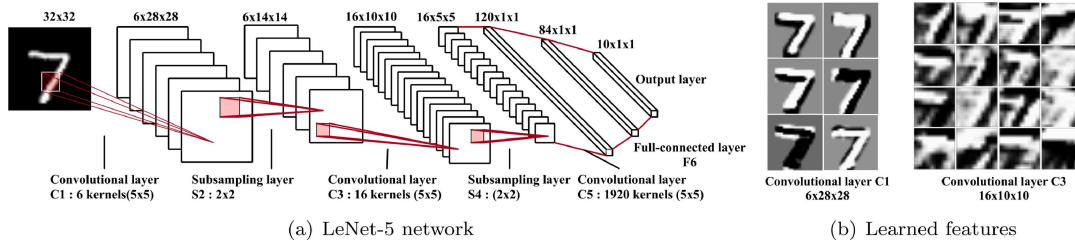


Figure A.2: Convolutional neural network

Here, $I * K$ denotes the result of the convolution, and x and y represent the spatial coordinates. m and n are the indices in the filter. The process involves: Consider a 3-D input tensor (e.g., image) and a set of learnable filters (kernels). Slide the filters over the input using the convolution operation:

$$(f * g)(s) = \sum_{a=1}^A \sum_{b=1}^B \sum_{c=1}^C f(a, b, c) \cdot g(s - a, t - b, u - c) \quad (\text{A.12})$$

where f is the input tensor, g is the filter, and (s, t, u) are the spatial coordinates. As depicted in Figure A.1 Each neuron within a feature map connects to neighboring neurons within the preceding layer, forming a receptive field for the neuron in the previous layer. The creation of a new feature map involves convolving the input with a learned kernel followed by the application of an element-wise nonlinear activation function to the convolved outcomes. Notably, each feature map is generated by sharing the kernel across all spatial locations of the input. The entire set of feature maps is derived using multiple distinct kernels. Mathematically, the value of a feature at location (i, j) in the k th feature map of the l th layer, $z_{l,i,j,k}$, is calculated as:

$$z_{l,i,j,k} = \mathbf{w}_{l,k}^T \mathbf{x}_{l,i,j} + b_{l,k} \quad (\text{A.13})$$

where $\mathbf{w}_{l,k}$ and $b_{l,k}$ denote the weight vector and bias term, respectively, for the k th filter of the l th layer. Here, $\mathbf{x}_{l,i,j}$ represents the input patch centered at location (i, j) of the l th layer. The kernel $\mathbf{w}_{l,k}$ responsible for generating the feature map $z_{l,i,j,k}$ is shared. This weight-sharing mechanism offers several advantages, including model complexity reduction and improved network training.

The activation function introduces nonlinearities to the CNN, necessary for detecting nonlinear features in multi-layer networks. Let $a(\cdot)$ denote the nonlinear activation function. The activation value $a_{l,i,j,k}$ of convolutional feature $z_{l,i,j,k}$ is computed as:

$$a_{l,i,j,k} = a(z_{l,i,j,k}) \quad (\text{A.14})$$

Typical activation functions encompass Sigmoid (Logistic), Hyperbolic tangent function, and Rectified Linear Unit.

Hyperbolic Tangent Activation Function:

The hyperbolic tangent function, often denoted as $\tanh(x)$, is a mathematical function that maps real-valued numbers to the range $[-1, 1]$. Mathematically, it is defined as:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (\text{A.15})$$

The tanh function is symmetric around the origin, exhibiting S-shaped behavior and saturating to -1 or 1 for large negative or positive values of x , respectively.

Sigmoid Activation Function:

The sigmoid function, commonly known as the logistic function, is represented as $\sigma(x)$ and is widely used in machine learning and neural networks for its ability to produce values between 0 and 1. It is expressed as:

$$\text{sigma}(x) = \frac{1}{1 + e^{-x}} \quad (\text{A.16})$$

The sigmoid function has an S-shaped curve, mapping any real-valued number to the range $(0, 1)$. It is particularly useful in binary classification problems, where it models the probability of an input belonging to a particular class.

Rectified Linear Unit Activation Function:

Rectified Linear Unit is a commonly used activation function in CNN. It introduces non-linearity to the network and can be defined as:

$$\text{ReLU}(x) = \max(0, x) \quad (\text{A.17})$$

This operation replaces any negative values with zeros.

An essential component of a convolutional neural network architecture is the pooling layer, a key element that contributes to spatial down-sampling, feature abstraction, and the creation of hierarchical representations within the network.

The pooling layer is designed to provide shift-invariance by reducing feature map resolution and is usually positioned between convolutional layers. Pooling is used for down-sampling and reducing the spatial dimensions of the feature maps. Each feature map of a pooling layer connects to its corresponding feature map in the preceding convolutional layer. Denoting the pooling function as $\text{pool}(\cdot)$, for each feature map $a_{l, :, :, k}$, the pooling operation is:

$$y_{l, i, j, k} = \text{pool}(a_{l, m, n, k}), \forall (m, n) \in R_{i, j} \quad (\text{A.18})$$

where $R_{i, j}$ represents a local neighborhood around location (i, j) .

Common pooling operations include average pooling and max pooling.

Max-Pooling:

Max-pooling is defined as:

$$\text{Max-Pooling}(x, y) = \max(I(x, y), I(x + 1, y), I(x, y + 1), I(x + 1, y + 1)) \quad (\text{A.19})$$

It retains the maximum value within a local region.

Average-Pooling:

The average pooling operation involves dividing the input into non-overlapping rectangular regions and computing the average value within each region. The resulting output has reduced spatial dimensions compared to the input.

Mathematically, average pooling for a $p \times q$ region is defined as:

$$\text{AvgPooling}(X)(i, j) = \frac{1}{p \times q} \sum_{a=0}^{p-1} \sum_{b=0}^{q-1} X(pi + a, qj + b) \quad (\text{A.20})$$

In case of a 2×2 region (used in this work), the average pooling operation is defined as:

$$\begin{aligned} \text{AvgPooling}(X)(i, j) = \frac{1}{4} & (X(2i, 2j) + X(2i, 2j + 1) \\ & + X(2i + 1, 2j) + X(2i + 1, 2j + 1)) \end{aligned} \quad (\text{A.21})$$

where:

- X is the input feature map.
- i and j are indices corresponding to the output feature map.
- The average is computed over the values in a 2×2 region of the input.

Following several convolutional and pooling layers, one or more fully connected layers may exist, aiming for higher-level reasoning. These layers connect all neurons in the previous layer to each neuron in the current layer, gathering global semantic information. It's noteworthy that a fully-connected layer isn't always necessary and can be replaced by a 1×1 convolutional layer. The CNN's last layer serves as an output layer and for classification tasks the softmax operation is commonly employed.

The softmax operation

The softmax activation function is commonly used in the output layer of CNN for multiclass classification. It converts raw scores (logits) into class probabilities:

$$\text{Softmax}(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (\text{A.22})$$

Where: z_i is the raw score for class i . The sum is taken over all classes.

Training a CNN involves global optimization. By minimizing the loss function, the optimal set of parameters can be determined. Stochastic gradient descent stands as a common solution for optimizing CNN networks.

The optimal parameters for a specific task are obtained by minimizing an appropriate loss function defined for that task. Suppose we have N desired input-output relations $(x(n), y(n)); n \in [1, \dots, N]$, where $x(n)$ denotes the n th input data, $y(n)$ its corresponding target label, and $o(n)$ the CNN output. The CNN's loss can be calculated as:

$$L = \frac{1}{N} \sum_{n=1}^N \mathcal{L}(\theta; y(n), o(n)) \quad (\text{A.23})$$

A.1.3 Fully Convolutional Neural Networks

Fully Convolutional Neural Networks are a type of neural network architecture designed for dense, end-to-end pixel-wise prediction tasks, particularly in the field of semantic segmentation and dense image labeling. FCN are characterized by their exclusive use of convolutional layers throughout the network, which allows them to take input images of arbitrary sizes and produce corresponding output feature maps of the same size.

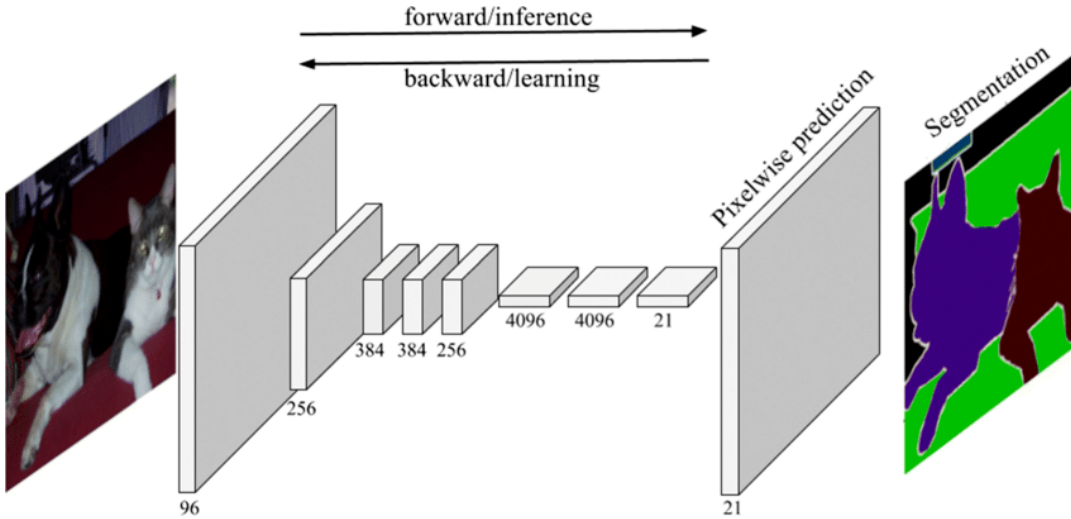


Figure A.3: Fully Convolutional Neural Network can make dense predictions for per-pixel tasks like semantic segmentation

The introduction of Fully Convolutional Neural Networks is credited to the paper titled “Fully Convolutional Networks for Semantic Segmentation” by Jonathan Long, Evan Shelhamer, and Trevor Darrell, (2015) published in [LSD15]. In this paper, the authors presented an innovative approach to semantic segmentation by adapting and extending the concept of convolutional neural networks into fully convolutional models.

A definition of FCNs can be formulated as follows:

“Fully Convolutional Neural Networks are neural network architectures designed for dense pixel-wise prediction tasks, such as semantic segmentation. FCNs employ convolutional layers exclusively, enabling them to take input images of varying dimensions and generate corresponding output feature maps at the same spatial resolution. This makes FCNs particularly well-suited for tasks that require fine-grained pixel-level predictions, allowing them to segment an image into multiple classes or categories, providing a detailed understanding of its content.” The introductory figure is presented in the paper as depicted in Figure A.3.

Each layer of data in a convnet is a three-dimensional array of size $h \times w \times d$, where h and w are spatial dimensions, and d is the feature or channel dimension. The first layer is the image, with pixel size $h \times w$, and d color channels. Locations in higher layers correspond to the locations in the image they are path-connected to, which are called their receptive fields.

Convnets are built on translation invariance. Their basic components (convolution, pooling, and activation functions) operate on local input regions, and depend only on relative spatial coordinates. Writing x_{ij} for the data vector at location (i, j) in a particular layer, and y_{ij} for the following layer, these functions compute outputs y_{ij} by

$$y_{ij} = f_{k,s} (\{x_{s(i+\delta_i),s(j+\delta_j)} \mid 0 \leq \delta_i, \delta_j \leq k\}) \quad (\text{A.24})$$

where k denotes the kernel size, s is the stride or subsampling factor, and $f_{k,s}$ determines the layer type: a matrix multiplication for convolution or average pooling, a spatial max for max pooling, or an elementwise nonlinearity for an activation

function, and so on for other types of layers.

This functional form is maintained under composition, with kernel size and stride obeying the transformation rule:

$$f_{k,s} \circ g_{k',s'} = (f \circ g)_{k'+(k-1)s',ss'}. \quad (\text{A.25})$$

While a general deep net computes a general nonlinear function, a net with only layers of this form computes a nonlinear filter, named as a deep filter or fully convolutional network. An FCN naturally operates on an input of any size, and produces an output of corresponding (possibly resampled) spatial dimensions.

A real-valued loss function composed with an FCN defines a task. If the loss function is a sum over the spatial dimensions of the final layer, $\ell(x; \theta) = \sum_{ij} \ell'(x_{ij}; \theta)$, its gradient will be a sum over the gradients of each of its spatial components. Thus stochastic gradient descent on ℓ computed on whole images will be the same as stochastic gradient descent on ℓ' , taking all of the final layer receptive fields as a minibatch.

In a fully convolutional neural network pooling is performed by a global max pooling layer that operates on a feature map (output of a convolutional layer) and reduces it by taking the maximum value from the entire feature map.

Global max pooling is a common operation in deep learning where we compute the maximum value across the entire feature map.

Mathematically, if we have a feature map X of dimensions $H \times W$ (height by width) for a single channel, the global max pooling operation can be defined as:

$$Y = \max(X) \quad (\text{A.26})$$

Here, Y represents the resulting value after applying global max pooling to the feature map X .

For multiple channels, say C channels, the operation is applied independently to each channel. If X is a tensor of dimensions $H \times W \times C$, global max pooling results in a tensor Y of dimensions $1 \times 1 \times C$, where:

$$Y_c = \max(X_{h,w,c}) \quad \text{for each channel } c \quad (\text{A.27})$$

In this equation, Y_c denotes the max-pooled value for the c -th channel, and $X_{h,w,c}$ represents the values in the c -th channel of the input tensor X .

In essence, FCN are capable of handling tasks where each pixel in an input image is associated with a particular label, enabling high-resolution and fine-grained semantic segmentation. These networks have had a significant impact on computer vision and have been instrumental in a wide range of applications, including object detection, image segmentation, and scene parsing.

A.2 A primer on Key Audio Features

A.2.1 Jitter, Shimmer, Signal-to-Noise and Harmonic-to-noise Ratio

In this subsection, we define fundamental audio features that were tested through our research and are used to formulate our audio representations in 2.3.2. As presented in the thesis introduction, acoustic parameters extensively utilized in audio analysis applications and widely cited in literature include fundamental frequency, jitter, shimmer, signal-to-noise and Harmonic-to-Noise Ratio.

The fundamental frequency (F0), quantifies the number of repetitions of a sound wave generated by vocal cords within a specified time frame. Additionally, it signifies the cycles of glottis opening and closure. Various algorithms have been developed to estimate the fundamental frequency from audio signals, each leveraging distinct principles to discern the periodicity inherent in the signal.

One commonly used approach is the Autocorrelation Method, which employs the autocorrelation function to identify repeating patterns in the signal corresponding to the fundamental frequency. By computing the autocorrelation and analyzing peaks, the algorithm estimates the pitch period.

Alternatively, the YIN Algorithm, refines pitch detection accuracy. This method involves calculating the normalized difference function and determining the minimum, which corresponds to the fundamental frequency. This is the method used by our implementation.

In another approach, the Harmonic Product Spectrum (HPS) method focuses on the harmonic structure of the signal. It multiplies the spectrum with its downsampled versions to suppress harmonics and enhance the fundamental frequency peak.

Alternatively, the Zero-Crossing Rate Method exploits the rate at which the signal crosses the zero amplitude to estimate the pitch period. By counting the number of zero-crossings within a specified window, the algorithm provides an estimate of the fundamental frequency.

Cepstral Analysis is another technique that applies cepstral analysis to the signal, emphasizing periodicity in the spectrum. Peaks in the cepstrum correspond to the pitch period, facilitating the extraction of the fundamental frequency.

Additionally, Time-Domain Pitch Detection focuses on periodicity in the time domain. By identifying repeating patterns in the waveform, this approach estimates the fundamental frequency.

Although there exists a typical frequency range for various genders and ages, these values are not static due to the influence of prosody conveyed by F0. Variations are also observed based on factors such as gender, age, and contextual elements like the individual's emotional state, time of day aligned with their lifestyle, and professional vocal usage.

The assessment of F0 disturbances, jitter, shimmer and Harmonic-to-noise-ratio have demonstrated efficacy in delineating vocal characteristics.

Based in the paper [TOL13], which provides an overview on vocal acoustic analysis parameters, we can provide a brief description of each:

Fundamental frequency disturbances refer to irregularities in the vibration rate of the vocal folds. Regularity in F0 is essential for a stable voice. Disturbances can signify vocal pathologies or instability in the voice production mechanism.

Jitter refers to variations in the time domain of a periodic signal. In audio, it can cause disturbances in the timing of digital audio signals, leading to distortion or interference in the reproduced sound. For voice, Jitter is the measure of frequency variation from cycle to cycle during sustained phonation. It quantifies the small, rapid fluctuations in the fundamental frequency. Jitter is a parameter describing the frequency variation from cycle to cycle. It is affected by the lack of control of vocal cord vibration, with higher percentages observed in the voices of individuals with pathologies. The typical variation for sustained phonation in young adults ranges between 0.5% and 1.0%.

Shimmer is associated with variations in pitch and amplitude in speech signals, particularly in voice quality. It refers to rapid and slight variations in pitch or amplitude that can affect the perceived quality of speech.

Signal-to-Noise Ratio represents the ratio of the level of a desired signal to the level of background noise. In audio, a higher SNR indicates a clearer, more distinguishable signal compared to the surrounding noise.

Harmonic-to-Noise Ratio is the ratio of harmonic (periodic) sound to noise (aperiodic) components in the voice signal. It quantifies the amount of noise relative to the harmonics. In the following subsections, we will define them in detail.

Jitter

Jitter, [TOL13] in the context of speech and voice analysis, refers to the temporal variation or irregularity in the timing of consecutive speech events or vocal fold vibrations. It is one of the acoustic features used in the analysis of speech and voice signals. Jitter is typically measured as the cycle-to-cycle variation in the time period of consecutive glottal cycles or pitch periods. In voice analysis, increased jitter can be indicative of voice disorders and irregularities in vocal fold vibrations.

The first introduction of jitter analysis in voice research can be traced back to the paper titled “Variations in Pitch Periods” by H. E. Milnes, published in the Journal of the Acoustical Society of America in 1936. This early work laid the foundation for the study of voice jitter and its applications in voice analysis.

Here is a brief definition of jitter:

“Jitter is a measure of the temporal irregularity or cycle-to-cycle variation in the timing of consecutive glottal cycles or pitch periods in speech and voice signals. It is an important acoustic feature used in voice analysis, and increased jitter may indicate voice disorders or irregularities in vocal fold vibrations.”

Jitter analysis, along with other acoustic features, plays a crucial role in the assessment and diagnosis of voice disorders and has applications in both clinical and research settings. The measurement of jitter involves analyzing the variations in the time intervals between consecutive glottal cycles or pitch periods. It is typically expressed as a percentage and can be calculated using the following equations:

Jitter is a metric used to quantify voice quality by measuring cycle-to-cycle variations in the time domain, specifically related to the pitch period. The calculation involves:

1. **Frame the Signal:** Divide the speech signal into short frames.
2. **Extract Pitch Periods:** Identify pitch periods in the signal.
3. **Time Domain Variation:** Measure cycle-to-cycle variations in the time domain.
4. **Jitter Metric:** Compute a jitter metric using standard formulas, often expressed as a percentage.

$$\text{Jitter} = \frac{\text{Cycle-to-Cycle Time Variation}}{\text{Mean Pitch Period}} \times 100 \quad (\text{A.28})$$

Calculation of the Absolute Jitter (Jitt):

$$\text{Jitt} = \frac{1}{N-1} \sum_{i=1}^{N-1} |P_i - P_{i+1}| \quad (\text{A.29})$$

Where:

Jitt is the absolute jitter.

N is the number of consecutive glottal cycles.

P_i is the duration of the i -th pitch period.

Calculation of the Relative Jitter (Rjitt):

$$\text{Rjitt} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left(\frac{|P_i - P_{i+1}|}{P_i} \right) \times 100\% \quad (\text{A.30})$$

Where:

Rjitt is the relative jitter (expressed as a percentage).

N is the number of consecutive glottal cycles.

P_i is the duration of the i -th pitch period.

These equations provide a quantitative measure of jitter in speech and voice signals. Jitter analysis helps in assessing the stability of vocal fold vibrations and can be used in the diagnosis of voice disorders and other related applications.

Jitter (local, absolute) represents the average absolute difference between two consecutive periods. The formula is given by:

$$\text{jitta} = \frac{1}{N} \sum_{i=1}^N |T_i - T_{i-1}| \quad (\text{A.31})$$

Jitter (Local)

Jitter (local) represents the average absolute difference between two consecutive periods divided by the average period. The threshold limit for detecting pathologies is 1.04%. The formula is given by:

$$jitt = \frac{1}{100N} \sum_{i=1}^N |T_i - T_{i-1}| \quad (\text{A.32})$$

where T_i is the duration in seconds of each period, and N is the number of periods.

Jitter (RAP)

Jitter (RAP) represents the average disturbance, i.e., the average absolute difference of one period and the average of the period with its two neighbors, divided by the average period. The threshold value to detect pathologies is 0.68%. The formula is given by:

$$rap = \frac{1}{N} \sum_{i=1}^N \left| T_i - \frac{T_{i-1} + T_{i+1}}{2} \right| \quad (\text{A.33})$$

Jitter (PPQ5)

Jitter (PPQ5) represents the ratio of disturbance within five periods. It is the average absolute difference between a period and the average containing its four nearest neighbor periods (two previous and two subsequent periods), divided by the average period. The formula is given by:

$$ppq = \frac{1}{N} \sum_{i=1}^N \left| T_i - \frac{T_{i-2} + T_{i-1} + T_{i+1} + T_{i+2}}{4} \right| \quad (\text{A.34})$$

Shimmer

Shimmer is a measure used in the analysis of voice signals to quantify the variations in the amplitude of consecutive glottal cycles. It provides valuable information about the irregularities or perturbations in voice production and is often employed in the assessment of voice quality and the diagnosis of voice disorders. Shimmer is a measure of the short-term cycle-to-cycle amplitude variation in a voice signal. It quantifies the changes in peak-to-peak amplitudes across consecutive glottal cycles, reflecting the instability or perturbations in voice production, introduced in [All+18].

The calculation of shimmer involves analyzing the variations in amplitude between consecutive glottal cycles. It is typically expressed as a percentage and can be calculated using the following equations: Calculation of the Absolute Shimmer (Shim):

$$textShim = \frac{1}{N-1} \sum_{i=1}^{N-1} |A_i - A_{i+1}| \quad (\text{A.35})$$

Where:

Shim is the absolute shimmer.

Shim is the absolute shimmer.

N is the number of consecutive glottal cycles.

A_i is the peak-to-peak amplitude of the i -th glottal cycle.

Calculation of the Relative Shimmer (Rshim):

$$\text{Rshim} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left(\frac{|A_i - A_{i+1}|}{A_i} \right) \times 100\% \quad (\text{A.36})$$

Where:

Rshim is the relative shimmer (expressed as a percentage).

N is the number of consecutive glottal cycles.

A_i is the peak-to-peak amplitude of the i -th glottal cycle.

These equations provide a quantitative measure of shimmer in voice signals, allowing for the assessment of voice quality, the detection of voice disorders, and the evaluation of treatment outcomes. The introduced shimmer measures have become valuable tools in the field of speech pathology and voice assessment.

Shimmer is a measure of the variability in amplitude or frequency of a speech signal, often used to quantify voice quality. Calculating shimmer involves measuring the cycle-to-cycle variations in the signal. Here's a concise LaTeX document for introducing the calculation of shimmer:

Shimmer is a metric used to quantify voice quality by measuring cycle-to-cycle variations in amplitude or frequency. The calculation involves:

1. **Frame the Signal:** Divide the speech signal into short frames.
2. **Extract Cycles:** Identify cycles in the signal.
3. **Amplitude/Frequency Measurement:** Calculate cycle-to-cycle variations in amplitude or frequency.
4. **Shimmer Metric:** Compute a shimmer metric using standard formulas, often expressed as a percentage.

$$\text{Shimmer} = \frac{\text{Amplitude/Frequency Variation}}{\text{Mean Amplitude/Frequency}} \times 100 \quad (\text{A.37})$$

The methods used to determine Shimmer are analogous to those used for Jitter, with the main distinction being that Jitter considers periods, while Shimmer takes into account the maximum peak amplitude of the signal.

To calculate Shimmer parameters, the algorithm starts by identifying the onset time of the glottal pulses in the signal and the corresponding magnitude of the signal at that sample. The Shimmer parameters are then determined using the following expressions:

Shimmer (Local)

Shimmer (local) represents the average absolute difference between the amplitudes of two consecutive periods, divided by the average amplitude. It is referred to as *shim*, with a threshold limit for detecting pathologies set at 3.81%. The formula is given by:

$$shim = \frac{1}{N} \sum_{i=1}^N \left| \frac{A_i - A_{i-1}}{(A_i + A_{i-1})/2} \right| \quad (\text{A.38})$$

Shimmer (Local, dB)

Shimmer (local, dB) represents the average absolute difference of the base 10 logarithm of the amplitude difference between two consecutive periods, denoted as *ShdB*. The threshold limit for detecting pathologies is set at 0.350 dB. The formula is given by:

$$ShdB = \frac{1}{N} \sum_{i=1}^N 20 \cdot \log_{10} \left(\frac{A_i}{A_{i-1}} \right) \quad (\text{A.39})$$

Shimmer (APQ3)

Shimmer (APQ3) represents the quotient of amplitude disturbance within three periods. It is the average absolute difference between the amplitude of a period and the mean amplitudes of its two neighbors, divided by the average amplitude. The formula is given by:

$$apq = \frac{1}{N} \sum_{i=1}^N \left| \frac{A_i - (A_{i-1} + A_{i+1})/2}{(A_i + A_{i-1} + A_{i+1})/3} \right| \quad (\text{A.40})$$

Shimmer (APQ5)

Shimmer (APQ5) represents the ratio of perturbation amplitude within five periods. It is the average absolute difference between the amplitude of a period and the mean amplitudes of it and its four nearest neighbors, divided by the average amplitude. The formula is given by:

$$apq = \frac{1}{N} \sum_{i=1}^N \left| \frac{A_i - (A_{i-2} + A_{i-1} + A_{i+1} + A_{i+2})/4}{(A_i + A_{i-2} + A_{i-1} + A_{i+1} + A_{i+2})/5} \right| \quad (\text{A.41})$$

Signal-to-Noise Ratio

The Signal-to-Noise Ratio stands as a critical metric and is defined as the ratio of signal power to noise power, SNR quantifies the clarity of a signal amidst background noise and is typically expressed in decibels Mathematically expressed as:

$$SNR_{dB} = 10 \cdot \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \quad (\text{A.42})$$

where

$$P_{\text{signal}}$$

is the power of the signal, and

$$P_{\text{noise}}$$

is the power of the noise.

Harmonic-to-Noise Ratio

Harmonic-to-Noise Ratio is a measure used in signal processing and voice analysis to quantify the ratio of harmonics (which represent periodic components such as the fundamental frequency and its overtones) to noise (which represents non-periodic components and disturbances) in a signal, particularly in voice signals. It is computed by analyzing the spectrum of the signal and comparing the energy in harmonic regions to non-harmonic noise regions. The implementation of the Harmonic-to-Noise Ratio was grounded in the mathematical fundamentals presented by Boersma [Boe+93].

The Harmonic-to-Noise Ratio is defined by the following equations:

1. Total Signal Energy (E_{total}):

$$E_{\text{total}} = \sum_{n=1}^N x(n)^2 \quad (\text{A.43})$$

where $x(n)$ represents the signal sample at time n , and N is the total number of samples.

2. Harmonic Energy (E_{harmonic}):

$$E_{\text{harmonic}} = \sum_{k=1}^K X(k)^2 \quad (\text{A.44})$$

Here, $X(k)$ denotes the Fourier coefficients corresponding to the harmonics of the signal.

3. Noise Energy (E_{noise}):

$$E_{\text{noise}} = E_{\text{total}} - E_{\text{harmonic}} \quad (\text{A.45})$$

4. Harmonic-to-Noise Ratio (HNR):

$$\text{HNR} = 10 \log_{10} \left(\frac{E_{\text{harmonic}}}{E_{\text{noise}}} \right) \quad (\text{A.46})$$

This equation computes the ratio of harmonic energy to noise energy in decibels, providing a quantitative measure of the clarity and periodicity of the signal.

The first local peak corresponds to the peak after index 1. The Autocorrelation Value (ACV) (T) of the Equation represents the peak at the index position corresponding to the period of the signal. The expected values for F_0 define a position for that peak between two indices. Considering the set value for the fundamental frequency (e.g., for women: 200 to 300 Hz, for men: 80 to 200 Hz, and for children: 400 to

500 Hz), the first index is calculated as $\frac{f_s}{F_{0\max}}$, and the second index as $\frac{f_s}{F_{0\min}}$. After determining the indices, the local maximum is found within the first and second index, revealing their respective amplitudes.

The value of HNR is then determined using the following formula:

$$HNR = 10 \cdot \log_{10} \left(\frac{V}{AC(T)} \right) \quad (\text{A.47})$$

where V is the amplitude of the first local maximum, and $AC(T)$ is the amplitude of the autocorrelation function at the index position corresponding to the period of the signal.

Despite the usage of the same mathematical formula, the algorithms may differ due to variations in the length of used segments or the incorporation of multiple segments.

A.2.2 Mel Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients are coefficients widely used in speech and audio processing. The motivation behind their development lies in their attempt to emulate the way the human auditory system processes sound. The human ear is more sensitive to certain frequency ranges, and this sensitivity is not linear across the entire spectrum. To capture this non-linear characteristic, MFCC involve a series of steps. First, the linear frequency scale is converted to the mel frequency scale, which approximates human perceptual response. Next, triangular filters are applied in the mel frequency domain, reflecting the ear's sensitivity. The resulting filterbank energies are then logarithmically compressed to align with the ear's response to sound intensity. Finally, a Discrete Cosine Transform is applied to decorrelate the filterbank energies and derive the MFCC coefficients.

MFCC are a representation of the short-term power spectrum of a sound signal, which represents the short-term power spectrum of a sound signal [DM80]. In order to compute MFCC the following steps are necessary:

- **Mel Frequency Scale:** The process begins with the conversion of the linear frequency scale to the mel frequency scale. The mel scale is a perceptual scale of pitches which approximates the human ear's response to different frequencies.
- **Filterbank Calculation:** A set of triangular filters is then applied in the mel frequency domain. These filters are designed to mimic the human ear's sensitivity to different frequencies. The filters are spaced more densely at lower frequencies and become wider as frequency increases. A typical value for the number of filters is usually set to 26, a formula that we followed through this research. The bandwidth distribution in that case is illustrated in Figure A.4.
- **Filterbank Energies:** The output of each filter is computed, and these values represent the energy in different frequency bands. Each filterbank output provides information about the energy distribution across various frequency ranges.

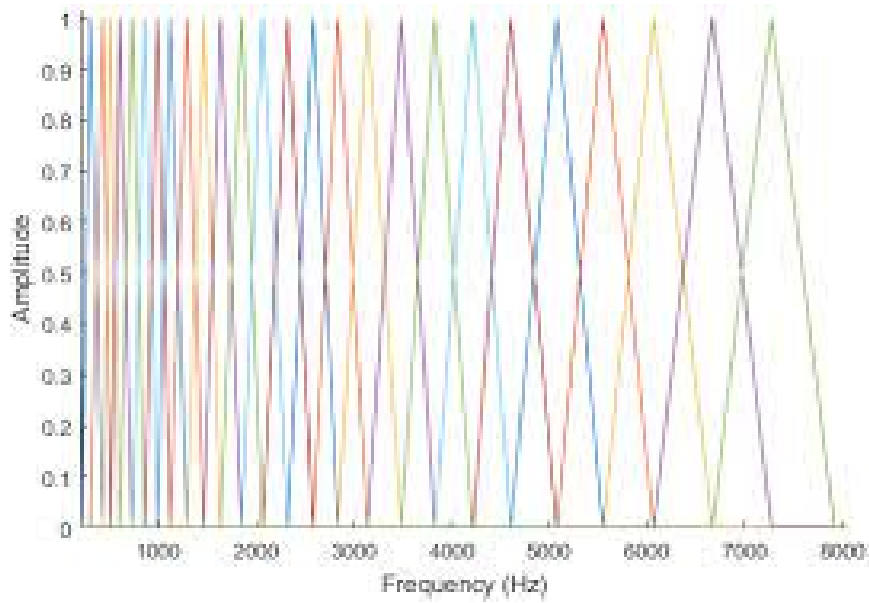


Figure A.4: Caption for your figure goes here.

- **Logarithmic Compression:** The logarithm of the filterbank energies is taken. This step is performed to mimic the logarithmic response of the human ear to sound intensity.
- **Discrete Cosine Transform:** The resulting log filterbank energies are then transformed using a Discrete Cosine Transform. The DCT decorrelates the filterbank energies and produces a set of coefficients. The first few coefficients usually contain the most relevant information.
- **MFCC Coefficients:** The resulting DCT coefficients are the MFCC. Each coefficient corresponds to a specific frequency band in the mel scale and represents a feature that characterizes the spectral content of the audio signal. Lower coefficients often capture the overall spectral shape, while higher coefficients may capture finer details.

The calculation involves several steps:

1. **Frame the Signal:** Split the audio signal into short frames.
2. **Apply the Discrete Fourier Transform:** Compute the DFT for each frame to obtain the power spectrum.
3. **Mel Filterbank:** Apply a filterbank of mel filters to the power spectrum. The mel filterbank is defined as follows:

$$H_m(k) = \begin{cases} 0, & \text{if } k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)}, & \text{if } f(m-1) \leq k < f(m) \\ 1, & \text{if } f(m) \leq k \leq f(m+1) \\ \frac{f(m+2)-k}{f(m+2)-f(m+1)}, & \text{if } f(m+1) < k \leq f(m+2) \\ 0, & \text{if } k > f(m+2) \end{cases} \quad (\text{A.48})$$

where $f(m)$ is the mel frequency corresponding to the m -th filter.

4. **Logarithm:** Take the logarithm of the filterbank energies.
5. **Discrete Cosine Transform:** Apply the DCT to the log filterbank energies to obtain the cepstral coefficients.

$$\text{MFCC}_n = \sum_{m=1}^M \log \left(\sum_{k=1}^K X(k) \cos \left[\frac{\pi n(2k-1)}{2K} \right] \right) \cdot H_m(k) \quad (\text{A.49})$$

where $X(k)$ is the log filterbank energy at frequency k , M is the number of mel filters, K is the number of DCT coefficients, and n is the index of the MFCC.

The MFCC coefficients typically capture the overall spectral shape and are often referred to as the “static coefficients.” These coefficients provide information about the energy distribution across different frequency bands.

To represent the rate of change or the gradient of the static coefficients over time, the “dynamic coefficients” or delta (differential) coefficients are introduced. They convey information about the changes in the spectral features between frames and capture dynamic aspects of the audio signal.

The computation of delta coefficients is expressed by the following formula:

$$\delta_t = \frac{\sum_{n=1}^N n(c_{t+n} - c_{t-n})}{\sum_{n=1}^N n^2} \quad (\text{A.50})$$

where δ_t represents a delta coefficient for frame t , calculated based on the static coefficients c_{t-n} to c_{t+n} . Typically, the value of N is set to 2.

In addition to delta coefficients, the incorporation of delta-delta (acceleration) coefficients aims to capture the dynamic aspects of the power spectrum, specifically, the trajectories of delta - MFCC over time.

Similarly, acceleration coefficients are computed using a similar process, but with differentials instead of static coefficients.

A.2.3 Noise Signals

In this subsection we define the basic noise signals that were used for data augmentation for our computational models [Opp+76].

Gaussian Noise

The mathematical equation for Gaussian noise in the time domain is typically represented as:

$$x(t) = A \cdot \exp \left(-\frac{t^2}{2\sigma^2} \right) \quad (\text{A.51})$$

Where:

- $x(t)$ represents the value of the noise at time t .
- A is the amplitude of the noise.

- σ is the standard deviation, controlling the spread or width of the noise distribution.

In the context of digital signal processing or when discrete samples are considered, the Gaussian noise is often characterized by its probability density function (PDF):

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad (\text{A.52})$$

Where:

- $p(x)$ is the probability of the noise having a value x at a certain instant.
- σ is the standard deviation, defining the spread of the Gaussian distribution.

This equation describes the probability of observing a particular value of Gaussian noise. The Gaussian or normal distribution is characterized by its bell-shaped curve centered around a mean value (μ) with a standard deviation (σ) controlling the width of the distribution.

White Noise

White noise is a signal where each sample is an independent, identically distributed random variable. Its power spectral density is purely random and uniformly distributed across all frequencies. White Gaussian noise can be represented as a continuous time signal or a series of discrete samples ($x(t)$ or $x[n]$, respectively) where each sample $x(t)$ or $x[n]$ follows a Gaussian distribution with a mean of zero and a constant variance (σ^2).

Pink Noise

Pink noise, also known as 1/f noise, has a power spectral density inversely proportional to the frequency. Its spectral characteristics result in equal energy per octave. Pink noise is characterized by its power spectral density function. It can be represented as:

$$S(f) \propto \frac{1}{f^\alpha} \quad (\text{A.53})$$

where $S(f)$ is the power spectral density at frequency f and α controls the slope of the spectrum.

Brown Noise

Brown noise, often referred to as Brownian noise, exhibits a power spectral density inversely proportional to the square of the frequency. It has higher energy at lower frequencies compared to pink or white noise. The power spectral density of brown noise can be given by:

$$S(f) \propto \frac{1}{f^2} \quad (\text{A.54})$$

where $S(f)$ is the power spectral density at frequency f .

A.2.4 Time frequency representations: Spectrogram and Mel spectrogram

In the context of audio signal processing, spectrograms and Mel spectrograms are frequently used to analyze and visualize sound signals. They provide a detailed view of how the frequency components of a sound evolve over time. Spectrograms display the intensity of frequencies over time, while Mel spectrograms emphasize perceptually relevant frequency bands using the Mel scale.

These representations have been widely adopted in voice pathology classification, as reported in related work sections. We utilize them in processing of electroglottographic signals 3.3.2 and COVID-19 and respiratory sound classification 5.3.2. They are particularly useful for due to their ability to capture temporal and frequency dynamics effectively.

The spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time [GP14]. The computation of a spectrogram involves short-time Fourier transform (STFT). Initially, the input signal is segmented into overlapping frames to capture its time-varying characteristics. Each segment is then multiplied by a windowing function.

The Fast Fourier Transform is subsequently applied to each windowed segment, converting the signal from the time domain to the frequency domain. This step yields information about the frequency components present in each segment. The magnitude of the resulting complex values is calculated, providing the amplitude of each frequency component.

To create a cohesive representation, the computed spectra from each segment are aligned in time, considering the overlap between adjacent segments. This overlapping ensures a smooth representation of the signal's time-varying frequency content.

Finally, the magnitude values are often represented as intensities on a logarithmic scale and mapped to colors. Darker regions in the resulting spectrogram typically signify lower intensity or less energy at specific frequencies, while lighter regions indicate higher intensity or more energy. The spectrogram visually depicts how the frequency content of the signal changes over time, making it a valuable tool for analyzing time-varying signals like audio waveforms.

The spectrogram computation briefly involves the following steps.

1. **Windowing:** Divide the signal into short-time segments using window functions.
 - Choose a window function (e.g., Hamming, Hanning, etc.).
 - Isolate short-time segments of the signal.
2. **Fast Fourier Transform:** Apply the FFT to each windowed segment.
 - Transform each segment into the frequency domain.
3. **Magnitude Squared:** Calculate the magnitude squared of the FFT output.
 - Obtain the power spectral density.
4. **Overlap and Sum:** Overlap the windowed segments and sum their spectrograms.

- Combine the spectrograms to produce the final spectrogram.

In mathematical terms, the Short-Time Fourier Transform (STFT) segments a signal into overlapping chunks using a sliding window and computes the Fourier transform for each segment. The continuous-time complex-valued signal, denoted as $x(t)$, is expressed in the STFT as:

$$\mathcal{X}(t, f) = \int_{-\infty}^{\infty} x(\tau)w(t - \tau)e^{-2\pi jf\tau} d\tau, \quad (\text{A.55})$$

where $w(\cdot)$ is a complex-valued window function with its complex conjugate denoted as \bar{w} . This expression can be interpreted as the scalar product of x with the window translated by time t and frequency-shifted by f .

For discrete signals $x[n]$ sampled at intervals T_s (inverse of the sampling frequency f_s), the discrete version of STFT is used:

$$\mathcal{X}[n, k] = \sum_{m=-\infty}^{\infty} x[m]w[n - m]e^{-2\pi jk(mT_s)}. \quad (\text{A.56})$$

Here, n represents the time index, k is the frequency index with step size Δf , and w is the sampled window function.

To align with the implementation of Short-Time FFT, the STFT process can be reformulated as a two-step procedure:

1. Extract the n -th slice by windowing with the window composed of M samples centered at nT_s , expressed as:

$$\mathcal{X}_n[k] = \sum_{m=0}^{M-1} x[m + n]w[m]e^{-2\pi jk(mT_s)}. \quad (\text{A.57})$$

2. Perform a discrete Fourier transform of $\mathcal{X}_n[k]$:

$$\text{FFT}\{\mathcal{X}_n[k]\} = \text{DFT}\{\mathcal{X}_n[k]\} = \sum_{k=0}^{N-1} \mathcal{X}_n[k]e^{-\frac{2\pi j}{N}kn}, \quad (\text{A.58})$$

where N is the number of samples in the FFT, and k is the frequency index. This equation represents the computation of the FFT of the sequence $\mathcal{X}_n[k]$. The FFT efficiently calculates the discrete Fourier transform, providing the frequency components of the signal in a computationally efficient manner.

The power spectral density is often computed from the Fourier transform of the autocorrelation function. For a continuous signal $x(t)$, the PSD $S(f)$ is given by:

$$S(f) = \lim_{T \rightarrow \infty} \frac{1}{T} \left| \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)e^{-2\pi jft} dt \right|^2, \quad (\text{A.59})$$

where f is the frequency.

In practice, for a discrete signal $x[n]$ sampled at intervals T_s (inverse of the sampling frequency f_s), the PSD can be estimated using the periodogram or other methods:

$$S(f_k) = \frac{1}{N \cdot f_s} \left| \sum_{n=0}^{N-1} x[n] e^{-2\pi j f_k n T_s} \right|^2 \quad (\text{A.60})$$

where N is the number of samples in the signal, $f_k = \frac{k}{NT_s}$ is the frequency corresponding to the k -th discrete Fourier transform bin, and k ranges from 0 to $N - 1$. The factor $\frac{1}{N \cdot f_s}$ normalizes the result to represent the power per unit frequency.

In contrast, a Mel spectrogram is a modified representation that incorporates the human auditory system's non-linear frequency perception. It involves mapping the linear frequency scale of the spectrogram to the Mel scale, approximating human auditory perception. The Mel scale is based on human psychoacoustic studies, where the perceived pitch differences are more discriminable at lower frequencies than at higher ones. By transforming the frequency axis to Mel scale and reorganizing the spectral content, Mel spectrograms better align with how humans perceive sound, emphasizing essential lower-frequency information crucial for speech and auditory perception. The algorithm to compute mel spectrograms goes through the following steps:

- **Preprocessing:**
 - **Signal Segmentation:** Divide the audio signal into short overlapping frames. Apply a windowing function (like Hamming or Hanning) to each frame to reduce spectral leakage.
 - **Frame-by-Frame Fourier Transform:** Compute the short-time Fourier transform for each frame to obtain the frequency content.
- **Mel Filterbank:**
 - **Mel Scale Conversion:** Transform the linear frequency scale into the Mel scale
 - **Creating Filterbanks:** Generate triangular filterbanks in the Mel scale, each characterized by center frequency and bandwidth parameters.
- **Filterbank Energies:**
 - **Apply Filters:** Apply Mel filters to the magnitude spectrum obtained from the STFT by multiplying each filterbank with the magnitude spectrum.
 - **Summation:** Sum the energy within each filterbank across all frames.
- **Logarithmic Compression:**
 - **Log Transformation:** Apply a logarithm to the filterbank energies for compression and human perception modeling.
- **Display:**
 - **Normalization:** Normalize the resulting Mel-spectrogram if necessary.

- **Visualization:** Plot the Mel spectrogram with time on the x-axis, frequency on the y-axis, and color intensity indicating the magnitude of log-filterbank energies.

And the necessary equations for generating Mel spectrograms:

Compute the Power Spectrogram: Compute the squared magnitude of the Short-Time Fourier Transform of the signal $x(t)$ using a window function $w(\cdot)$:

$$P(f, t) = |\text{STFT}[x(t)w(t)]|^2 \quad (\text{A.61})$$

Apply the Mel Filterbank: Use a Mel filterbank to transform the power spectrogram to the Mel scale. The Mel filterbank is a set of triangular filters that are applied to the power spectrogram:

$$M(f_m, f) = \begin{cases} 0 & \text{if } f < f_{m-1} \\ \frac{f-f_{m-1}}{f_m-f_{m-1}} & \text{if } f_{m-1} \leq f < f_m \\ 1 & \text{if } f_m \leq f < f_{m+1} \\ \frac{f_{m+1}-f}{f_{m+1}-f_m} & \text{if } f_m \leq f < f_{m+1} \\ 0 & \text{if } f \geq f_{m+1} \end{cases} \quad (\text{A.62})$$

where f_m is the center frequency of the m -th Mel filter.

Compute the Mel Spectrogram: Convolve the power spectrogram with the Mel filterbank to obtain the Mel spectrogram:

$$S_m(t, f_m) = \sum_f P(f, t) \cdot M(f_m, f) \quad (\text{A.63})$$

This operation is typically performed for each time frame t and each Mel frequency bin f_m .

The Mel spectrogram provides a more perceptually relevant representation of the audio signal by emphasizing frequency ranges that are more significant for human hearing.

Bibliography

- [AA14] Ali Akbari and Meisam Khalil Arjmandi. “An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features.” In: *Biomed. Signal Process. Control.* 10 (2014), pp. 209–223.
- [AA22] M. Aly and N. Alotaibi. “A novel deep learning model to detect COVID-19 based on wavelet features extracted from Mel-scale spectrogram of patients’ cough and breathing sounds.” In: *Informatics in Medicine Unlocked* 32 (2022), p. 101049. doi: [10.1016/j.imu.2022.101049](https://doi.org/10.1016/j.imu.2022.101049).
- [Aba+22] O. Abayomi-Alli et al. “Detection of COVID-19 from Deep Breathing Sounds Using Sound Spectrum with Image Augmentation and Deep Learning Techniques.” In: *Electronics (Switzerland)* 11.16 (2022). doi: [10.3390/electronics11162520](https://doi.org/10.3390/electronics11162520).
- [Aga+18] Szkielkowska Agata et al. “Electroglottography in the diagnosis of functional dysphonia.” In: *European Archives of Oto-Rhino-Laryngology* 275 (2018), pp. 2523–2528.
- [Ahm+18] Y. Alnasheri Ahmed et al. “Voice Pathology Detection and Classification Using Auto-Correlation and Entropy Features in Different Frequency Regions.” In: *IEEE Access* 6 (2018), pp. 6961–6974.
- [Ahm+20] Y. Al-nasheri Ahmed et al. “An Investigation of Multi-Dimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification.” In: *Journal of voice : official journal of the Voice Foundation* 31 1 (2020), 113.e9–113.e18. url: <https://api.semanticscholar.org/CorpusID:2730260>.
- [Al+14] Ahmed Y. Al-nasheri et al. “Voice pathology detection using auto-correlation of different filters bank.” In: *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)* (2014), pp. 50–55.
- [Al+17] Ahmed Y. Al-nasheri et al. “An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification.” In: *Journal of voice : official journal of the Voice Foundation* 31 1 (2017), 113.e9–113.e18.
- [Alb+21] Tena Alberto et al. “Automated detection of COVID-19 cough.” In: *Biomedical Signal Processing and Control* 71 (2021), pp. 103175–103175.

- [Ale+22] Ponomarchuk Alexander et al. “Project Achoo: A Practical Model and Application for COVID-19 Detection From Recordings of Breath, Voice, and Cough.” In: *IEEE Journal of Selected Topics in Signal Processing* 16 (2022), pp. 175–187.
- [Alf+21] P. Alfonso et al. “CIoTVID: Towards an Open IoT-Platform for Infective Pandemic Diseases such as COVID-19.” In: *Sensors (Basel, Switzerland)* 21 (2021). url: <https://api.semanticscholar.org/CorpusID:231612018>.
- [All+18] Gary Allwood et al. “Advances in acoustic signal processing techniques for enhanced bowel sound analysis.” In: *IEEE reviews in biomedical engineering* 12 (2018), pp. 240–253.
- [Alm23] S. A. Almutairi. “A multimodal AI-based non-invasive COVID-19 grading framework powered by deep learning, manta ray, and fuzzy inference system from multimedia vital signs.” In: *Heliyon* 9.6 (2023), e16552. doi: [10.1016/j.heliyon.2023.e16552](https://doi.org/10.1016/j.heliyon.2023.e16552).
- [AMA+] A. Al-nasheri, G Muhammad, M Alsulaiman, et al. “An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification.” In: *Journal of Voice* 31 (), 113.e9–113.e18.
- [An96] Guozhong An. “The Effects of Adding Noise During Backpropagation Training on a Generalization Performance.” In: *Neural Computation* 8 (1996), pp. 643–674.
- [Ari+11] Arias-Londono et al. “Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients.” In: *IEEE Transactions on Biomedical Engineering* 58.2 (2011), pp. 370–379.
- [Arj+11] Meisam Khalil Arjmandi et al. “Identification of voice disorders using long-time features and support vector machine with different feature reduction methods.” In: *Journal of voice : official journal of the Voice Foundation* 25 6 (2011), e275–89.
- [Ash+22] A. E. Ashby et al. “Cough-based COVID-19 detection with audio quality clustering and confidence measure based learning.” In: *Proceedings of Machine Learning Research* 179 (2022), pp. 1–20.
- [Awa+23] Muhammad Awais et al. “Optimized DEC: An effective cough detection framework using optimal weighted Features-aided deep Ensemble classifier for COVID-19.” In: *Biomedical Signal Processing and Control* 86 (2023), p. 105026.
- [BA05] Roozbeh Behroozmand and Farshad Almasganj. “Comparison of neural networks and support vector machines applied to optimized features extracted from patients’ speech signal for classification of vocal fold inflammation.” In: *Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005.* (2005), pp. 844–849.

- [Bab+22] Naseem Babu et al. “Multiclass Categorisation of Respiratory Sound Signals Using Neural Network.” In: *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (2022), pp. 228–232. url: <https://api.semanticscholar.org/CorpusID:253555602>.
- [Bak92] Ronald J Baken. “Electroglottography.” In: *Journal of Voice* 6.2 (1992), pp. 98–110.
- [BAM06] Roozbeh Behroozmand, Farshad Almasganj, and Mohammad Hassan Moradi. “Pathological Assesment of Vocal Fold Nodules and Polyp Using Accoustic Perturbation and Phase Space Features.” In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings 2* (2006), pp. II–II.
- [BdD01] Nathalie Henrich Bernardoni, Christophe d’Alessandro, and Boris Doval. “GLOTTAL FLOW MODELS : WAVEFORMS, SPECTRA AND PHYSICAL MEASUREMENTS.” In: 2001. url: <https://api.semanticscholar.org/CorpusID:1179508>.
- [Bha+23] Debarpan Bhattacharya et al. “Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection.” In: *Scientific Data* 10.1 (2023), p. 397.
- [BK18] Chitralekha Bhat and Sunil Kumar Kopparapu. “FEMH Voice Data Challenge: Voice disorder Detection and Classification using Acoustic Descriptors.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5233–5237.
- [BKH16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization.” In: *arXiv preprint arXiv:1607.06450* (2016).
- [BMM18] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. “A systematic study of the class imbalance problem in convolutional neural networks.” In: *Neural networks : the official journal of the International Neural Network Society* 106 (2018), pp. 249–259.
- [Boe+93] Paul Boersma et al. “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.” In: *Proceedings of the institute of phonetic sciences*. Vol. 17. 1193. Amsterdam. 1993, pp. 97–110.
- [Boe93] Paul Boersma. “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.” In: *IFA Proceedings 17*. 1993, pp. 97–110.
- [BP07] WJ Barry and M Putzer. “Saarbrucken voice database. Institute of Phonetics, University of Saarland.” In: *Institute of Phonetics University of Saarland* (2007).
- [CAO23] NR Calvo-Ariza, T Arias-Vergara, and JR Orozco-Arroyave. “Automatic Assessment of Voice Disorders Using Phase Plots.” In: *Workshop on Engineering Applications*. Springer. 2023, pp. 127–138.

- [Cel23] Gaffari Celik. “CovidCoughNet: A new method based on convolutional neural networks and deep feature extraction using pitch-shifting data augmentation for covid-19 detection from cough, breath, and voice signals.” In: *Computers in Biology and Medicine* (2023), p. 107153.
- [Ces+18] Ugo Cesari et al. “A new database of healthy and pathological voices.” In: *Comput. Electr. Eng.* 68 (2018), pp. 310–321. url: <https://api.semanticscholar.org/CorpusID:49407702>.
- [Cha+22] Y. Chang et al. “CovNet: A Transfer Learning Framework for Automatic COVID-19 Detection From Crowd-Sourced Cough Sounds.” In: *Frontiers in Digital Health* 3.August 2021 (2022), pp. 1–11. doi: [10.3389/fdgth.2021.799067](https://doi.org/10.3389/fdgth.2021.799067).
- [Che+23] Srikanth Raj Chetupalli et al. “Multi-modal point-of-care diagnostics for COVID-19 based on acoustics and symptoms.” In: *IEEE Journal of Translational Engineering in Health and Medicine* 11 (2023), pp. 199–210.
- [Cho+12] Se-Jin Choi et al. “Comparison of maximum phonation time associated with the changes in vocal intensity in patients with unilateral vocal fold palsy and sulcus vocalis.” In: *Phonetics and Speech Sciences* 4.1 (2012), pp. 125–131.
- [Cho+22a] N K Chowdhury et al. “Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method.” In: *Computers in Biology and Medicine* 145 (2022), p. 105405. doi: [10.1016/j.combiomed.2022.105405](https://doi.org/10.1016/j.combiomed.2022.105405).
- [Cho+22b] Nihad Karim Chowdhury et al. “Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method.” In: *Computers in Biology and Medicine* 145 (2022), pp. 105405–105405.
- [Cox58] D. R. Cox. “The Regression Analysis of Binary Sequences.” In: *Journal of the Royal Statistical Society. Series B (Methodological)* 20.2 (1958), pp. 215–232. doi: [10.1111/j.2517-6161.1958.tb00304.x](https://doi.org/10.1111/j.2517-6161.1958.tb00304.x).
- [CS07] César David Paredes Crovato and Adalberto Schuck. “The use of wavelet packet transform and artificial neural networks in analysis and classification of dysphonic voices.” In: *IEEE Transactions on Biomedical Engineering* 54.10 (2007), pp. 1898–1900.
- [CSN22] Michele Cozzatti, Federico Simonetta, and Stavros Ntalampiras. “Variational Autoencoders for Anomaly Detection in Respiratory Sounds.” In: *International Conference on Artificial Neural Networks*. 2022. url: <https://api.semanticscholar.org/CorpusID:251402903>.
- [Dar+15] Panek Daria et al. “Acoustic analysis assessment in speech pathology detection.” In: *International Journal of Applied Mathematics and Computer Science* 25 (2015), pp. 631–643.

- [DB14] Khalid Daoudi and Blaise Bertrac. “On classification between normal and pathological voices using the MEEI-KayPENTAX database: Issues and consequences.” In: *Interspeech* (2014). url: <https://api.semanticscholar.org/CorpusID:29957096>.
- [DBC+01] P. Dejonckere, P. Bradeley, P. Clemente, et al. “A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques.” In: *European Archives of Otorhinolaryngology* 258 (2001), pp. 77–82.
- [DEH18] Kevin Degila, Rahhal Errattahi, and Asmaa El Hannani. “The UCD System for the 2018 FEMH Voice Data Challenge.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5242–5246.
- [DM80] Steven Davis and Paul Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.” In: *IEEE transactions on acoustics, speech, and signal processing* 28.4 (1980), pp. 357–366.
- [DNB02] Alireza A Dibazar, S Narayanan, and Theodore W Berger. “Feature analysis for automatic detection of pathological speech.” In: *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*. Vol. 1. IEEE. 2002, pp. 182–183.
- [DQM22] Ting Dang, Thomas Quinnell, and Cecilia Mascolo. “Exploring Semi-supervised Learning for Audio-based COVID-19 Detection using Fix-Match.” In: *Interspeech*. 2022. url: <https://api.semanticscholar.org/CorpusID:252338722>.
- [Dut+22] Debottam Dutta et al. “Acoustic Representation Learning on Breathing and Speech Signals for COVID-19 Detection.” In: *Interspeech*. 2022. url: <https://api.semanticscholar.org/CorpusID:250073069>.
- [EI94] Massachusetts Eye and Ear Infirmary. “Elemetrics Disordered Voice Database (Version 1.03).” In: *Voice and Speech Lab, Boston, MA* (1994).
- [EN23] Meysam Effati and Goldie Nejat. “A Performance Study of CNN Architectures for the Autonomous Detection of COVID-19 Symptoms Using Cough and Breathing.” In: *Computers* 12.2 (2023), p. 44.
- [Esi+22] Darici Esin et al. “Using Deep Learning with Large Aggregated Datasets for COVID-19 Classification from Cough.” In: *ArXiv abs/2201.01669* (2022).
- [Fan+19] Shih-Hau Fang et al. “Detection of pathological voice using cepstrum vectors: A deep learning approach.” In: *Journal of voice* 33.5 (2019), pp. 634–641.
- [Fan+21a] Fenglei Fan et al. “On Interpretability of Artificial Neural Networks: A Survey.” In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 5 (2021), pp. 741–760.

- [Fan+21b] Ziqi Fan et al. “Class-Imbalanced Voice Pathology Detection and Classification Using Fuzzy Cluster Oversampling Method.” In: *Applied Sciences* 11 (2021), p. 3450. doi: [10.3390/APP11083450](https://doi.org/10.3390/APP11083450).
- [FM21] Amara Fethi and Fezari Mohamed. “Voice Pathologies Classification Using GMM And SVM Classifiers.” In: *International Journal of Mathematics and Computers in Simulation* (2021). url: <https://api.semanticscholar.org/CorpusID:17853464>.
- [Fou71] Adrian J Fourcin. “First applications of a new laryngograph.” In: *Med Biol Illus* 21 (1971), pp. 172–182.
- [Fra+15] D.H. François et al. “The Physiologic Impact of Unilateral Recurrent Laryngeal Nerve (RLN) Lesion on Infant Oropharyngeal and Esophageal Performance.” In: *Dysphagia* 30 (2015), pp. 714–722.
- [Fra+21] Mohammad Fraiwan et al. “A dataset of lung sounds recorded from the chest wall using an electronic stethoscope.” In: *Data in Brief* 35 (2021), p. 106913.
- [Gal+22] Carlos A. Galindo-Meza et al. “Detection of COVID-19 in Respiratory Sounds using End-to-End Deep Audio Embeddings.” In: *International Journal of Health Science* (2022). url: <https://api.semanticscholar.org/CorpusID:250233031>.
- [GD23] Praveen Gupta and Sheshang Degadwala. “A Comprehensive Review on COVID-19 Cough Audio Classification through Deep Learning.” In: *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 4 (2023), p. 6.
- [Gen+21] Lei Geng et al. “Voice pathology detection and classification from speech signals and EGG signals based on a multimodal fusion method.” In: *Biomedical Engineering / Biomedizinische Technik* 66 (2021), pp. 613–625. url: <https://api.semanticscholar.org/CorpusID:244730024>.
- [GGH19] Ning Gui, Danni Ge, and Ziyin Hu. “AFS: An attention-based mechanism for supervised feature selection.” In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 3705–3713.
- [Ghu+17] Muhammad Ghulam et al. “Enhanced Living by Assessing Voice Pathology Using a Co-Occurrence Matrix.” In: *Sensors (Basel, Switzerland)* 17 (2017). url: <https://api.semanticscholar.org/CorpusID:17989125>.
- [Gok+21] A. Gokcen et al. “Artificial Intelligence–Based COVID-19 Detection Using Cough Records.” In: *Electrica* (2021). url: <https://api.semanticscholar.org/CorpusID:236404699>.
- [Gon+12] David Martínez González et al. “Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit.” In: *IberSPEECH*. 2012. url: <https://api.semanticscholar.org/CorpusID:1115177>.
- [GP14] Theodoros Giannakopoulos and Aggelos Pikrakis. *Introduction to audio analysis: a MATLAB® approach*. Academic Press, 2014.

- [Gu+18] Jiuxiang Gu et al. “Recent advances in convolutional neural networks.” In: *Pattern recognition* 77 (2018), pp. 354–377.
- [Gue+19] Victor Guedes et al. “Transfer learning with audioset to voice pathologies identification in continuous speech.” In: *Procedia Computer Science* 164 (2019), pp. 662–669.
- [GVT13] Pavel Grill, Josef Vavřina, and Jana Tučková. “Databases and their applications for diagnosis of developmental dysphasia.” In: *2013 IEEE 11th International Workshop of Electronics, Control, Measurement, Signals and their application to Mechatronics*. IEEE. 2013, pp. 1–4.
- [Heg+19] Sarika Hegde et al. “A survey on machine learning approaches for automatic detection of voice disorders.” In: *Journal of Voice* 33.6 (2019), 947–e11.
- [Hem17] D. Hemmerling. “Voice pathology distinction using autoassociative neural networks.” In: *EUSIPCO 2017* (2017), pp. 1844–1847.
- [Her20] Christian T Herbst. “Electroglottography—an update.” In: *Journal of Voice* 34.4 (2020), pp. 503–526.
- [HFS10] C. Herbst, W. Fitch, and J. Svec. “Electroglottographic wavegrams: a technique for visualizing vocal fold dynamics noninvasively.” In: *The Journal of the Acoustical Society of America* 128 5 (2010), pp. 3070–8.
- [HFŠ10] Christian T Herbst, W Fitch, and Jan G Švec. “Electroglottographic wavegrams: a technique for visualizing vocal fold dynamics noninvasively.” In: *The Journal of the Acoustical Society of America* 128.5 (2010), pp. 3070–3078.
- [HHC20] Rabeh Hamdi, Salah Hajji, and Adnen Cherif. “Recognition of Pathological Voices by Human Factor Cepstral Coefficients (HFCC).” In: *Journal of Computer Science* (2020). url: <https://api.semanticscholar.org/CorpusID:221147935>.
- [HK92] L. Holmström and P. Koistinen. “Using additive noise in back-propagation training.” In: *IEEE transactions on neural networks* 3 1 (1992), pp. 24–38.
- [Hoa+22] T. Hoang et al. “A Cough-based deep learning framework for detecting COVID-19.” In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Vol. 2022-July. 2022, pp. 3422–3425. doi: [10.1109/EMBC48229.2022.9871179](https://doi.org/10.1109/EMBC48229.2022.9871179).
- [How95] David M Howard. “Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers.” In: *Journal of Voice* 9.2 (1995), pp. 163–172.
- [Hsu+21] Fu-Shun Hsu et al. “Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database—HF_Lung_V1.” In: *PLoS One* 16.7 (2021), e0254134.

- [HT06] Christian T. Herbst and Sten Ternström. “A comparison of different methods to measure the EGG contact quotient.” In: *Logopedics Phoniatrics Vocology* 31 (2006), pp. 126–138.
- [Hu+23] Jinhai Hu et al. “Supervised Contrastive Pretrained ResNet with MixUp to Enhance Respiratory Sound Classification on Imbalanced and Limited Dataset.” In: *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2023, pp. 1–5.
- [IAT22] Rumana Islam, Esam Abdel-Raheem, and Mohammed Tarique. “A study of using cough sounds and deep neural networks for the early detection of COVID-19.” In: *Biomedical Engineering Advances* 3 (2022), p. 100025.
- [IPL18] Kazi Aminul Islam, Daniel Pérez, and Jiang Li. “A Transfer Learning Approach for the 2018 FEMH Voice Data Challenge.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5252–5257.
- [IRT22] Rumana Islam, Esam Abdel Raheem, and Mohammed Tarique. “Deep Learning Based Pathological Voice Detection Algorithm Using Speech and Electroglottographic (EGG) Signals.” In: *2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA)* (2022), pp. 127–131.
- [IS15] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In: *International conference on machine learning*. PMLR, 2015, pp. 448–456.
- [Isl+22] Rumana Islam et al. “Voice pathology detection using convolutional neural networks with electroglottographic (EGG) and speech signals.” In: *Computer Methods and Programs in Biomedicine Update* 2 (2022), p. 100074.
- [ITA20] Rumana Islam, Mohammed Tarique, and Esam Abdel-Raheem. “A Survey on Signal Processing Based Pathological Voice Detection Techniques.” In: *IEEE Access* 8 (2020), pp. 66749–66776. url: <https://api.semanticscholar.org/CorpusID:216044130>.
- [Jas+15] Yosinski Jason et al. “Understanding Neural Networks Through Deep Visualization.” In: *ArXiv abs/1506.06579* (2015).
- [Jes+18] B. Alonso-Hernández Jesús et al. “New Feature Extraction from Electroglottographic Signals Applied to Automatic Detection of Laryngeal Pathologies.” In: *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)* (2018), pp. 365–371.
- [Jin+22] Han Jing et al. “Sounds of COVID-19: exploring realistic performance of audio-based digital testing.” In: *NPJ Digital Medicine* 5 (2022). url: <https://api.semanticscholar.org/CorpusID:235670069>.
- [Jor+20] Laguarda Jordi et al. “COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings.” In: *IEEE Open Journal of Engineering in Medicine and Biology* 1 (2020), pp. 275–281.

- [JR22] Amlu Anna Joshy and Rajeev Rajan. “Automated dysarthria severity classification: A study on acoustic features and deep learning techniques.” In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), pp. 1147–1157.
- [Ju+18] Mingxuan Ju et al. “A Multi-Representation Ensemble Approach to Classifying Vocal Diseases.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5258–5262.
- [JU23] Rallapalli Jhansi and G. Uganya. “Detection of COVID-19 Patients using Speech Recognition with Support Vector Machine” and Comparing with “K Nearest Neighbour Algorithm.” In: *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)* (2023), pp. 1–5. url: <https://api.semanticscholar.org/CorpusID:257934120>.
- [Jul+18] D. Arias-Londoño Julián et al. “ByoVoz Automatic Voice Condition Analysis System for the 2018 FEMH Challenge.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5228–5232.
- [Jun+22] Li Jun et al. “Improving The ResNet-based Respiratory Sound Classification Systems With Focal Loss.” In: *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (2022), pp. 223–227. url: <https://api.semanticscholar.org/CorpusID:253555603>.
- [KA20] Antony Koroulakis and Manuj Agarwal. *Laryngeal Cancer*. 2020. url: <https://journals.lww.com/co-otolaryngology/toc/2014/04000>.
- [Kas88] Hideki Kasuya. “Voice evaluation by acoustic analysis.” In: *The Japan Journal of Logopedics and Phoniatics* 29.2 (1988), pp. 194–199.
- [KBL23] Sera Kim, Ji-Young Baek, and Seok-Pil Lee. “COVID-19 detection model with acoustic features from cough sound and its application.” In: *Applied Sciences* 13.4 (2023), p. 2378.
- [Kim+08] Heejin Kim et al. “Dysarthric speech database for universal access research.” In: *Interspeech*. Vol. 2008. 2008, pp. 1741–1744.
- [Kim+21] Yoonjoo Kim et al. “Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning.” In: *Scientific reports* 11.1 (2021), p. 17186.
- [Ksi+23] Amel Ksibi et al. “Voice Pathology Detection Using a Two-Level Classifier Based on Combined CNN–RNN Architecture.” In: *Sustainability* 15.4 (2023), p. 3204.
- [Lal23] Kumari Nidhi Lal. “A lung sound recognition model to diagnoses the respiratory diseases by using transfer learning.” In: *Multimedia Tools and Applications* (2023), pp. 1–17.
- [Lar+21] Orlandic Lara et al. “The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms.” In: *Scientific Data* 8 (2021). url: <https://api.semanticscholar.org/CorpusID:221878789>.

- [Lar02] Jean Laroche. “Time and pitch scale modification of audio signals.” In: *Applications of digital signal processing to audio and acoustics*. Springer, 2002, pp. 279–309.
- [LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning.” In: *nature* 521.7553 (2015), pp. 436–444.
- [Lec77] FLE Lecluse. “Elektroglottografie: een experimenteel onderzoek betreffende de elektrische impedantie van het mannelij // jk strottehoofd.” In: (1977). url: <https://api.semanticscholar.org/CorpusID:86798606>.
- [Lin+22] Zhang Lin et al. “A Feature Polymerized Based Two-Level Ensemble Model for Respiratory Sound Classification.” In: *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (2022), pp. 238–242. url: <https://api.semanticscholar.org/CorpusID:253557206>.
- [LM21] M. Loey and S. Mirjalili. “COVID-19 cough sound symptoms classification from scalogram image representation using deep learning models.” In: *Computers in Biology and Medicine* 139 (2021), pp. 105020–105020. url: <https://api.semanticscholar.org/CorpusID:244005841>.
- [LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.
- [Lui+17] Jesus Luis M. T. et al. “The Advanced Voice Function Assessment Databases (AVFAD): Tools for Voice Clinicians and Speech Research.” In: *Advances in Speech-Language Pathology* (2017), pp. 221–236. url: <https://api.semanticscholar.org/CorpusID:59467790>.
- [MA21] Ghulam Muhammad and MUSAED Alhussein. “Convergence of artificial intelligence and internet of things in smart healthcare: a case study of voice pathology detection.” In: *Ieee Access* 9 (2021), pp. 89198–89209.
- [Mad+21] Pahar Madhurananda et al. “COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features.” In: *Computers in Biology and Medicine* 141 (2021), pp. 105153–105153.
- [Mad+22] R. Kamble Madhu et al. “Exploring Auditory Acoustic Features for The Diagnosis of Covid-19.” In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022), pp. 566–570.
- [MAD24] Ahlam Fadhil Mahmood, Ahmed M. Alkababji, and Amar Daood. “Resilient embedded system for classification respiratory diseases in a real time.” In: *Biomedical Signal Processing and Control* (2024). url: <https://api.semanticscholar.org/CorpusID:266711173>.
- [Man+23] F. Manzella et al. “The voice of COVID-19: Breath and cough recording classification with temporal decision trees and random forests.” In: *Artificial Intelligence in Medicine* 137 (2023), p. 102486.

- [Man21] Negin Melek Manshouri. “Identifying COVID-19 by using spectral analysis of cough recordings: a distinctive classification study.” In: *Cognitive Neurodynamics* 16 (2021), pp. 239–253.
- [Mar+10a] Maria Markaki et al. “Dysphonia detection based on modulation spectral features and cepstral coefficients.” In: *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5162–5165.
- [Mar+10b] Maria E. Markaki et al. “Dysphonia detection based on modulation spectral features and cepstral coefficients.” In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (2010), pp. 5162–5165.
- [Mar+18] Pishgar Maryam et al. “Pathological Voice Classification Using Mel-Cepstrum Vectors and Support Vector Machine.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5267–5271.
- [MD14] Matthias Mauch and S. Dixon. “PYIN: A fundamental frequency estimator using probabilistic threshold distributions.” In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014), pp. 659–663.
- [MDP23] Soumya Ranjan Mishra, Tusar Kanti Dash, and Ganapati Panda. “Speech phoneme and spectral smearing based non-invasive COVID-19 detection.” In: *Frontiers in Artificial Intelligence* 5 (2023). url: <https://api.semanticscholar.org/CorpusID:255497531>.
- [MGA14] Malak Al Mojaly, Muhammad Ghulam, and Mansour Alsulaiman. “Detection and classification of voice pathology using feature selection.” In: *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)* (2014), pp. 571–577.
- [Mid07] Janet H Middendorf. “Phonotrauma in children: Management and treatment.” In: *The ASHA leader* 12.15 (2007), pp. 14–17.
- [MK18] Jihye Moon and Sanghun Kim. “An approach on a combination of higher-order statistics and higher-order differential energy operator for detecting pathological voice with machine learning.” In: *ICTC 2018* (2018), pp. 46–51.
- [Moh+21] E A Mohammed et al. “An ensemble learning approach to digital coronavirus preliminary screening from cough sounds.” In: *Scientific Reports* 11.1 (2021), pp. 1–11. doi: [10.1038/s41598-021-95042-2](https://doi.org/10.1038/s41598-021-95042-2).
- [Mon+20] Alfonso Monaco et al. “Multi-time-scale features for accurate respiratory sound classification.” In: *Applied Sciences* 10.23 (2020), p. 8606.
- [Mos+22] Zohreh Mostaani et al. “Modeling of pre-trained neural network embeddings learned from raw waveform for covid-19 infection detection.” In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8482–8486.

- [MQS05] Nicolas Malyska, Thomas F Quatieri, and Douglas Sturim. “Automatic dysphonia recognition using biologically-inspired amplitude-modulation features.” In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*. Vol. 1. IEEE. 2005, pp. I–873.
- [MS11a] Maria Markaki and Yannis Stylianou. “Voice pathology detection and discrimination based on modulation spectral features.” In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 1938–1948.
- [MS11b] Maria E. Markaki and Yannis Stylianou. “Voice Pathology Detection and Discrimination Based on Modulation Spectral Features.” In: *IEEE Transactions on Audio, Speech, and Language Processing* 19 (2011), pp. 1938–1948.
- [Mug+21] Ananya Muguli et al. “DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics.” In: *Interspeech*. 2021. url: <https://api.semanticscholar.org/CorpusID:232240233>.
- [Muh+12] Ghulam Muhammad et al. “Multidirectional regression (MDR)-based features for automatic voice disorder detection.” In: *Journal of voice : official journal of the Voice Foundation* 26 6 (2012), 817.e19–27.
- [Muh+16] Ghulam Muhammad et al. “Automatic voice pathology detection and classification using vocal tract area irregularity.” In: *Biocybernetics and Biomedical Engineering* 36 (2016), pp. 309–317.
- [Muh+22] Enamul Hoque Chowdhury Muhammad et al. “QUCoughScope: An Intelligent Application to Detect COVID-19 Patients Using Cough and Breath Sounds.” In: *Diagnostics* 12 (2022).
- [NA16a] Souissi Nawel and Ch. Adnane. “Speech recognition system based on short-term cepstral parameters, feature reduction method and Artificial Neural Networks.” In: *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* (2016), pp. 667–671.
- [NA16b] Souissi Nawel and Cherif Adnane. “Artificial Neural Networks and Support Vector Machine for Voice Disorders Identification.” In: *International Journal of Advanced Computer Science and Applications* 7 (2016).
- [Naf+23] Muhammad Fauzan Nafiz et al. “Automated Detection of COVID-19 Cough Sound using Mel-Spectrogram Images and Convolutional Neural Network.” In: *J. Ilm. Tek. Elektro Komput. dan Inform* 9.3 (2023), pp. 535–548.
- [NDS23] Sudhansu Sekhar Nayak, Anand D Darji, and Prashant K Shah. “Machine learning approach for detecting Covid-19 from speech signal using Mel frequency magnitude coefficient.” In: *Signal, Image and Video Processing* (2023), pp. 1–8.

- [Nee+20] Kumar Neeraj et al. “Coswara - A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis.” In: *ArXiv* abs/2005.10548 (2020).
- [Nee+21] Kumar Neeraj et al. “Towards sound based testing of COVID-19 — Summary of the first Diagnostics of COVID-19 using Acoustics (DiCOVA) Challenge.” In: *Computer Speech & Language* 73 (2021), pp. 101320–101320.
- [Ngo+23] Dat Ngo et al. “A Deep Learning Architecture with Spatio-Temporal Focusing for Detecting Respiratory Anomalies.” In: *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE. 2023, pp. 1–5.
- [NH08] Brett Ninness and Soren John Henriksen. “Time-scale modification of speech signals.” In: *IEEE Transactions on Signal Processing* 56.4 (2008), pp. 1479–1488.
- [NP20] Truc The Nguyen and Franz Pernkopf. “Lung Sound Classification Using Snapshot Ensemble of Convolutional Neural Networks.” In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2020), pp. 760–763. url: <https://api.semanticscholar.org/CorpusID:221387767>.
- [NP21] Truc Kim Thi Nguyen and Franz Pernkopf. “Lung Sound Classification Using Co-tuning and Stochastic Normalization.” In: *IEEE Transactions on Biomedical Engineering* PP (2021), pp. 1–1. url: <https://api.semanticscholar.org/CorpusID:236912776>.
- [Nta23] Stavros Ntalampiras. “Explainable Siamese Neural Network for Classifying Pediatric Respiratory Sounds.” In: *IEEE Journal of Biomedical and Health Informatics* 27 (2023), pp. 4728–4735. url: <https://api.semanticscholar.org/CorpusID:260246527>.
- [OMO22] Asli Nur Omeroglu, Hussein MA Mohammed, and Emin Argun Oral. “Multi-modal voice pathology detection architecture based on deep and handcrafted feature fusion.” In: *Engineering Science and Technology, an International Journal* 36 (2022), p. 101148.
- [Opp+76] AV Oppenheim et al. *Digital signal processing and theory and application of digital signal processing*. 1976.
- [Orl91] Robert F Orlikoff. “Assessment of the dynamics of vocal fold contact from the electroglottogram: data from normal male subjects.” In: *Journal of Speech, Language, and Hearing Research* 34.5 (1991), pp. 1066–1072.
- [Pah+21] M Pahar et al. “COVID-19 cough classification using machine learning and global smartphone recordings.” In: *Computers in Biology and Medicine*. Vol. 135. 2021, p. 104572. doi: [10.1016/j.compbimed.2021.104572](https://doi.org/10.1016/j.compbimed.2021.104572).

- [Pap+23] Charalampos Papadakis et al. “AuscultNET: A Deep Learning framework for Adventitious Lung Sounds Classification.” In: *2023 30th IEEE International Conference on Electronics, Circuits and Systems (ICECS)* (2023), pp. 1–4. url: <https://api.semanticscholar.org/CorpusID:266905400>.
- [Par+19] Daniel S Park et al. “SpecAugment: A simple data augmentation method for automatic speech recognition.” In: *arXiv preprint arXiv:1904.08779* (2019).
- [Pav+17] Harár Pavol et al. “Voice Pathology Detection Using Deep Learning: a Preliminary Study.” In: *IWOBI 2017* (2017), pp. 1–4.
- [Pay+22] Christopher L Payten et al. “Frameworks, Terminology and Definitions Used for the Classification of Voice Disorders: A Scoping Review.” In: *Journal of voice : official journal of the Voice Foundation* (2022). url: <https://api.semanticscholar.org/CorpusID:247590640>.
- [Pes+23] Diogo Pessoa et al. “Pediatric Respiratory Sound Classification Using a Dual Input Deep Learning Architecture.” In: *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE. 2023, pp. 1–5.
- [Pha+22] Thi Viet Pham et al. “Classification of lung sounds using scalogram representation of sound segments and convolutional neural network.” In: *Journal of Medical Engineering & Technology* 46.4 (2022), pp. 270–279.
- [PLZ18] Minh Pham, Jing Lin, and Yanjia Zhang. “Diagnosing Voice Disorder with Machine Learning.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5263–5266.
- [PN21a] Madhurananda Pahar and Thomas Niesler. “Machine learning based COVID-19 detection from smartphone recordings: cough, breath and speech.” In: abs/2104.02477 (2021). url: <https://api.semanticscholar.org/CorpusID:233033565>.
- [PN21b] Madhurananda Pahar and Thomas R. Niesler. “Deep Transfer Learning based COVID-19 Detection in Cough, Breath and Speech using Bottleneck Features.” In: 2021. url: <https://api.semanticscholar.org/CorpusID:236456536>.
- [Poł+19] Dawid Połap et al. “Bio-inspired voice evaluation mechanism.” In: *Appl. Soft Comput.* 80 (2019), pp. 342–357.
- [Pre+23] VK Preetha et al. “Covid-19 Detection Through Cough Sounds: A Comparative Study Using XGboost and RESNET.” In: *2023 International Conference on Innovations in Engineering and Technology (ICIET)*. IEEE. 2023, pp. 1–5.
- [Qin+23] Zhang Qing et al. “Grand Challenge on Respiratory Sound Classification for SPRSound Dataset.” In: *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (2023), pp. 1–5. url: <https://api.semanticscholar.org/CorpusID:253556967>.

- [Rah+22] Tawsifur Rahman et al. “QUCoughScope: an intelligent application to detect COVID-19 patients using cough and breath sounds.” In: *Diagnostics* 12.4 (2022), p. 920.
- [Ras+23] Nicholas Rasmussen et al. “Cough Sound Analysis for the Evidence of Covid-19.” In: *Computer Vision and Machine Intelligence: Proceedings of CVMI 2022*. Springer, 2023, pp. 501–512.
- [Raz+22] Haroldas Razvadauskas et al. “Exploring traditional machine learning for identification of pathological auscultations.” In: *arXiv:2209.00672* (2022).
- [Ren+22] Z. Ren et al. “Learning complementary representations via attention-based ensemble learning for cough-based COVID-19 recognition.” In: *Acta Acustica* 6 (2022), pp. 0–4. doi: [10.1051/aacus/2022029](https://doi.org/10.1051/aacus/2022029).
- [RG72] M Reinsch and H Gobsch. “Zur quantitativen Auswertung elektroglogtographischer Kurven bei Normalpersonen.” In: *Folia Phoniatica et Logopaedica* 24.1 (1972), pp. 1–6.
- [Rib+23a] D. Ribas et al. “Automatic Voice Disorder Detection Using Self - Supervised Representations.” In: *IEEE Access* 11 (2023), pp. 14915–14927. doi: [10.1109/ACCESS.2023.3243986](https://doi.org/10.1109/ACCESS.2023.3243986).
- [Rib+23b] Dayana Ribas et al. “Automatic voice disorder detection using self-supervised representations.” In: *IEEE Access* 11 (2023), pp. 14915–14927.
- [RNW11] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. “The TORGO database of acoustic and articulatory speech from speakers with dysarthria.” In: *Language Resources and Evaluation* 46 (2011), pp. 523–541. url: <https://api.semanticscholar.org/CorpusID:15481029>.
- [Roc+18] BM Rocha et al. “A respiratory sound database for the development of automated classification.” In: *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*. Springer. 2018, pp. 33–37.
- [Ros58] Frank Rosenblatt. “The Perceptron: A probabilistic model for information storage and organization in the brain.” In: *Psychological Review* 65.6 (1958), pp. 386–408. doi: [10.1037/h0042519](https://doi.org/10.1037/h0042519).
- [Rot92] Martin Rothenberg. “A multichannel electroglottograph.” In: *Journal of Voice* 6.1 (1992), pp. 36–43.
- [RSA20] Hamdi Rabeh, Hajji Salah, and Cherif Adnen. *Recognition of pathological voices by Human Factor Cepstral Coefficients (HFCC)*. 2020. url: <https://www.researchsquare.com/article/rs-23108/v1>.
- [RU23] Jhansi Rallapalli and G. Uganya. “Improved Accuracy in Speech Recognition System for Detection of Covid-19 using K Nearest Neighbour and Comparing with Artificial Neural Network.” In: *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)* (2023), pp. 1–5. url: <https://api.semanticscholar.org/CorpusID:256904137>.

- [Rum+22] Islam Rumana et al. “A study of using cough sounds and deep neural networks for the early detection of Covid-19.” In: *Biomedical Engineering Advances* 3 (2022), pp. 100025–100025.
- [S+21] Syed S et al. “Inter classifier comparison to detect voice pathologies.” In: *Mathematical biosciences and engineering : MBE* 18 3 (2021), pp. 2258–2273.
- [Sae+06] Nicolas Saenz-Lechon et al. “Methodological issues in the development of automatic systems for voice pathology detection.” In: *Biomedical Signal Processing and Control* 1.2 (2006), pp. 120–128.
- [Saf+23] Saima Safdar et al. “Prediction of Specific Language Impairment in Children using Cepstral Domain Coefficients.” In: *2023 International Conference on Business Analytics for Technology and Security (IC-BATS)*. IEEE. 2023, pp. 1–11.
- [San+23] Bae Sangmin et al. “Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification.” In: *ArXiv abs/2305.14032* (2023). url: <https://api.semanticscholar.org/CorpusID:258841333>.
- [SB16] Mausumi N Syamal and M. Benninger. “Vocal fold paresis: a review of clinical presentation, differential diagnosis, and prognostic indicators.” In: *Current Opinion in Otolaryngology & Head and Neck Surgery* 24 (2016), pp. 197–202.
- [SC15] Nawel Souissi and Adnane Cherif. “Dimensionality reduction for voice disorders identification system based on Mel Frequency Cepstral Coefficients and Support Vector Machine.” In: *ICMIC 2015* (2015), pp. 1–6.
- [SG23] Sandeep B. Sangle and Chandrakant J. Gaikwad. “COVID-19 Respiratory Sound Signal Detection Using HOS-Based Linear Frequency Cepstral Coefficients and Deep Learning.” In: *Circuits, Systems, and Signal Processing* 43 (2023), pp. 331–347. url: <https://api.semanticscholar.org/CorpusID:261056806>.
- [SH20] Based Shuvo Samiul and othersl Hassan Bhuiyan. “A Lightweight CNN Model for Detecting Respiratory Diseases From Lung Auscultation Sounds Using EMD-CWT-Based Hybrid Scalogram.” In: *IEEE Journal of Biomedical and Health Informatics* 25 (2020), pp. 2595–2603. url: <https://api.semanticscholar.org/CorpusID:221556950>.
- [She+18] Tsui Sheng-Yang et al. “Demographic and Symptomatic Features of Voice Disorders and Their Potential Application in Classification Using Machine Learning Algorithms.” In: *Folia phoniatica et logopaedica : official organ of the International Association of Logopedics and Phoniatics* 70 3-4 (2018), pp. 174–182.

- [She+23] Jiakun Shen et al. “Piecewise Position Encoding in Convolutional Neural Network for Cough-Based Covid-19 Detection.” In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.
- [Shi+19] Fang Shih-Hau et al. “Combining acoustic signals and medical records to improve pathological voice classification.” In: *APSIPA Transactions on Signal and Information Processing* 8 (2019), e14.
- [Sid+21] Abid Syed Sidra et al. “Comparative Analysis of CNN and RNN for Voice Pathology Detection.” In: *BioMed Research International* 2021 (2021). url: <https://api.semanticscholar.org/CorpusID:233455810>.
- [SK16] Tim Salimans and Diederik P. Kingma. “Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks.” In: *Advances in neural information processing systems* 29 (2016), pp. 279–309.
- [SL22] Myoung-Jin Son and Seok-Pil Lee. “COVID-19 Diagnosis from Crowdsourced Cough Sound Data.” In: *Applied Sciences* (2022). url: <https://api.semanticscholar.org/CorpusID:246756027>.
- [Sob+22] Nebras Sobahi et al. “Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds.” In: *Biocybernetics and Biomedical Engineering* 42.3 (2022), pp. 1066–1080.
- [SRH20a] S. Syed, Munaf Rashid, and S. Hussain. “Meta-analysis of voice disorders databases and applied machine learning techniques.” In: *Mathematical biosciences and engineering : MBE* 17 6 (2020), pp. 7958–7979.
- [SRH20b] Sidra Abid Syed, Munaf Rashid, and Samreen Hussain. “Meta-analysis of voice disorders databases and applied machine learning techniques.” In: *Mathematical Biosciences and Engineering: MBE* 17.6 (2020), pp. 7958–7979.
- [SS24] Irum Sindhu and Mohd Shamrie Sainin. “Automatic Speech and Voice Disorder Detection Using Deep Learning—A Systematic Literature Review.” In: *IEEE Access* 12 (2024), pp. 49667–49681. url: <https://api.semanticscholar.org/CorpusID:268214932>.
- [SSD98] Christine M Sapienza, Elaine T Stathopoulos, and Christopher Dromey. “Approximations of open quotient and speed quotient from glottal air-flow and EGG waveforms: Effects of measurement criteria and sound pressure level.” In: *Journal of Voice* 12.1 (1998), pp. 31–43.
- [Sye+21a] S Syed et al. “Inter classifier comparison to detect voice pathologies.” In: *Mathematical Biosciences and Engineering* 18.3 (2021), pp. 2258–2273.

- [Sye+21b] Sidra Abid Syed et al. “Comparative analysis of CNN and RNN for voice pathology detection.” In: *BioMed Research International* 2021 (2021), pp. 1–8. url: <https://api.semanticscholar.org/CorpusID:233455810>.
- [Syr+22] Ghrabli Syrine et al. “Challenges and Opportunities of Deep Learning for Cough-Based COVID-19 Diagnosis: A Scoping Review.” In: *Diagnostics* 12 (2022).
- [Tei+15] Paulo Teixeira et al. “Acoustic analysis of vocal dysphonia.” In: *Procedia Computer Science* 64 (2015), pp. 466–473.
- [Ter19] S. Ternström. “Normalized time-domain parameters for electroglottographic waveforms.” In: *The Journal of the Acoustical Society of America* 146 1 (2019), EL65.
- [TF15] João Paulo Teixeira and Paula Odete Fernandes. “Acoustic Analysis of Vocal Dysphonia.” In: *Procedia Computer Science* 64 (2015), pp. 466–473.
- [TK95] J. Timmer and M. König. “On Generating Power Law Noise.” In: *A&A* 300 (1995), pp. 707–710.
- [TOL13] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. “Vocal acoustic analysis—jitter, shimmer and hnr parameters.” In: *Procedia Technology* 9 (2013), pp. 1112–1122.
- [Tom+18] Grzywalski Tomasz et al. “Parameterization of Sequence of MFCCs for DNN-based voice disorder detection.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5247–5251.
- [Ton+21] Xia Tong et al. “COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening.” In: *NeurIPS Datasets and Benchmarks*. 2021. url: <https://api.semanticscholar.org/CorpusID:244306367>.
- [Ton20] Han Tong. “Automatic assessment of dysarthric severity level using audio-video cross-modal approach in deep learning.” MA thesis. 2020.
- [TSP03] Ingo R Titze, Jan G Svec, and Peter S Popolo. “Vocal dose measures: quantifying accumulated vibration exposure in vocal fold tissues.” In: *Journal of Speech, Language, and Hearing Research* 46 (2003), pp. 919–932.
- [TV12] Ingo R Titze and Katherine Verdolini. *Vocology: The Science and Practice of Voice Habilitation*. Salt Lake City, UT: National Center for Voice and Speech, 2012.
- [Ulu+23] S Ulukaya et al. “MSCCov19Net: multi-branch deep learning model for COVID-19 detection from cough sounds.” In: *Medical and Biological Engineering and Computing* 61.7 (2023), pp. 1619–1629. doi: [10.1007/s11517-023-02803-4](https://doi.org/10.1007/s11517-023-02803-4).
- [Uma+02] Karthikeyan Umapathy et al. “Discrimination of pathological voices using an adaptive time-frequency approach.” In: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* 4 (2002), pp. IV-3852-IV–3855.

- [UVL16] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. “Instance normalization: The missing ingredient for fast stylization.” In: *arXiv preprint arXiv:1607.08022* (2016).
- [Ver+06] K Verdolini et al. *Classification Manual for Voice Disorders— I*. Lawrence Erlbaum Associates, Inc, 2006. doi: [10 . 4324 / 9781410617293](https://doi.org/10.4324/9781410617293).
- [Ver+21] Laura Verde et al. “Exploring the use of artificial intelligence techniques to detect the presence of coronavirus covid-19 through speech and voice analysis.” In: *Ieee Access* 9 (2021), pp. 65750–65757.
- [Vic+19] de Oliveira Guedes Victor et al. “Transfer Learning with AudioSet to Voice Pathologies Identification in Continuous Speech.” In: *Procedia Computer Science* 164 (2019), pp. 662–669.
- [Wal+21] Conor Wall et al. “Deep recurrent neural networks with attention mechanisms for respiratory anomaly classification.” In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8.
- [Wal+22] Conor Wall et al. “A Deep Ensemble Neural Network with Attention Mechanisms for Lung Abnormality Classification Using Audio Inputs.” In: *Sensors (Basel, Switzerland)* 22 (2022).
- [Wei+22] Ma Weijie et al. “An Effective Lung Sound Classification System for Respiratory Disease Diagnosis Using DenseNet CNN Model with Sound Pre-processing Engine.” In: *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (2022), pp. 218–222. url: <https://api.semanticscholar.org/CorpusID:253557838>.
- [XHM22] Tong Xia, Jing Han, and Cecilia Mascolo. “Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues.” In: *Experimental Biology and Medicine* 247.22 (2022), pp. 2053–2061.
- [Xin+22] Chen Xing-Yu et al. “Supervised and Self-Supervised Pretraining Based Covid-19 Detection Using Acoustic Breathing/Cough/Speech Signals.” In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022), pp. 561–565.
- [Yan+23] Runze Yang et al. “Respiratory Sound Classification by Applying Deep Neural Network with a Blocking Variable.” In: *Applied Sciences* 13.12 (2023), p. 6956.
- [YDM+23] Niharika Yerramsetty, Kusupati Deekshitha, Mahathi Mantrala, et al. “Preliminary Diagnosis of COVID-19 using Speech Processing Techniques.” In: *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*. IEEE. 2023, pp. 157–161.
- [Yi+23] Zhu Yi et al. “COVID-19 Detection via Fusion of Modulation Spectrum and Linear Prediction Speech Features.” In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 1536–1549. url: <https://api.semanticscholar.org/CorpusID:258033151>.

- [Zha+22a] Qing Zhang et al. “SPRSound: Open-Source SJTU Paediatric Respiratory Sound Database.” In: *IEEE Transactions on Biomedical Circuits and Systems* 16.5 (2022), pp. 867–881.
- [Zha+22b] Ren Zhao et al. “The Acoustic Dissection of Cough: Diving Into Machine Listening-based COVID-19 Analysis and Detection.” In: *Journal of Voice* (2022). url: <https://api.semanticscholar.org/CorpusID:247359177>.
- [Zhu+22] Yilun Zhu et al. “How Generalizable and Interpretable are Speech-Based COVID-19 Detection Systems?: A Comparative Analysis and New System Proposal.” In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (2022), pp. 1–5.
- [Ziq+21] Fan Ziqi et al. “Class-Imbalanced Voice Pathology Detection and Classification Using Fuzzy Cluster Oversampling Method.” In: *Applied Sciences* 11 (2021), p. 3450.
- [Ziz+22] Chen Zizhao et al. “Classify Respiratory Abnormality in Lung Sounds Using STFT and a Fine-Tuned ResNet18 Network.” In: *2022 IEEE Biomedical Circuits and Systems Conference (BioCAS)* (2022), pp. 233–237. url: <https://api.semanticscholar.org/CorpusID:251929163>.
- [ZJ08a] Yu Zhang and Jack Jiang. “Acoustic analyses of sustained and running voices from patients with laryngeal pathologies.” In: *Journal of Voice* 22.1 (2008), pp. 1–9.
- [ZJ08b] Yu Zhang and Jack J. Jiang. “Acoustic analyses of sustained and running voices from patients with laryngeal pathologies.” In: *Journal of voice : official journal of the Voice Foundation* 22 1 (2008), pp. 1–9.
- [Zon+18] Chuanget Zong-Ying et al. “DNN-based Approach to Detect and Classify Pathological Voice.” In: *2018 IEEE International Conference on Big Data (Big Data)* (2018), pp. 5238–5241.