NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF INFORMATION TRANSMISSION SYSTEMS AND MATERIAL TECHNOLOGY

COVID-19 Diagnosis from cough samples using Deep Learning methods

DIPLOMA THESIS

PARASKEVI M. VALERGAKI

Supervisor : Konstantina S. Nikita
NTUA PROFESSOR

Athens, August 2022

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
DIVISION OF INFORMATION TRANSMISSION SYSTEMS AND MATERIAL TECHNOLOGY

COVID-19 Diagnosis from cough samples using Deep Learning methods

DIPLOMA THESIS

PARASKEVI M. VALERGAKI

Advisory Board: Konstantina S. Nikita

Professor, NTUA

Approved by the review board on 30/08/2022.

.....	.....	.....
Konstantina Nikita	Andreas-Georgios Stafylopatis	Giorgos Stamou
Professor, NTUA	Professor, NTUA	Associate Professor, NTUA

Athens, August 2022

.....

Paraskevi M. Valergaki

Graduate Electrical and Computer Engineering N.T.U.A

Copyright © Paraskevi Valergaki, 2022

All rights reserved.

No part of this thesis may be reproduced or transmitted in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) for any commercial purposes without permission in writing from the author. Parts of this thesis may be reproduced, stored or transmitted for any non-commercial purposes provided that the source is referred to and the present copyright notice is retained. Theses and conclusions included in this manuscript are the author's own and do not necessarily reflect the official opinion of the National Technical University of Athens.

.....

Παρασκευή Μ. Βαλεργάκη

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Παρασκευή Μ. Βαλεργάκη, 2022

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

*Η νόσος του κορωνοϊού του 2019 (COVID-19), που προκαλεί το Σοβαρό Οξύ Αναπνευστικό Σύνδρομο τύπου 2 (SARS-CoV-2) έχει επηρεάσει τις ζωές εκατομμυρίων ανθρώπων σε όλο τον κόσμο. Μέχρι τον Ιούλιο του 2022, υπήρχαν 569.771.691 ενεργά κρούσματα COVID-19 παγκοσμίως και είχαν καταγραφεί 6.383.776 θάνατοι. Ο ιός μεταδίδεται κυρίως μέσω σταγονιδίων που δημιουργούνται όταν ένα μολυσμένο άτομο βήχει, φτερνίζεται ή εκπνέει. Τα πιο συχνά εμφανιζόμενα συμπτώματα είναι πυρετός, βήχας και κόπωση. Η τρέχουσα μέθοδος διάγνωσης βασίζεται στη δοκιμή Αλυσιδωτής Αντίδρασης Πολυμεράσης Αντίστροφης Μεταγραφής (RT-PCR). Ωστόσο, η σπανιότητα, το κόστος και ο μεγάλος χρόνος διεκπεραίωσης είναι μερικά μειονεκτήματα της δοκιμής RT-PCR. Επιπλέον, αυτή η διαγνωστική μέθοδος θέτει το ιατρικό προσωπικό σε κίνδυνο λοίμωξης κατά τη διάρκεια της δειγματοληψίας. Ο εμβολιασμός αποτελεί σημαντικό όπλο αντιμετώπισης του κορωνοϊού. Δυστυχώς, οι παραλλαγές της νόσου COVID-19 μπορούν να μειώσουν κάποια στιγμή την αποτελεσματικότητα των εμβολίων, οδηγώντας στη συνέχεια σε επαναλοιμώξεις. Ως εκ τούτου, η ανάγκη για συνεχείς ελέγχους νόσησης παραμένει, καθώς η ανοσία συχνά απειλείται από μεταλλάξεις. Η παρούσα διπλωματική εργασία στοχεύει στη διερεύνηση μεθόδων Βαθιάς Μάθησης για την ανίχνευση της νόσου COVID-19 από ήχους βήχα. Αυτός ο τύπος ελέγχου είναι χωρίς επαφή, είναι εύκολος στην εφαρμογή και μπορεί να μειώσει τον φόρτο εργασίας στα κέντρα ελέγχου καθώς και να περιορίσει τη μετάδοση. Σε αυτή την εργασία έχουν χρησιμοποιηθεί δύο σύνολα δεδομένων, που περιέχουν αρχεία βήχα από συμμετέχοντες από διάφορες χώρες, το σύνολο δεδομένων Coswara και το σύνολο δεδομένων του Cambridge. Η ανισορροπία του συνόλου δεδομένων αντιμετωπίστηκε με την εφαρμογή μιας προσέγγισης εκμάθησης συνόλου, έτσι ώστε η τάξη Covid να μην υποεκπροσωπείται. Το στάδιο προεπεξεργασίας περιλαμβάνει την ανίχνευση βήχα για τον προσδιορισμό του εάν και πότε υπάρχει βήχας στις ακατέργαστες ηχογραφήσεις. Τα σύνολα δεδομένων είναι crowd-sourced, πράγμα που σημαίνει ότι οι ήχοι που συλλέγονται προέρχονται από διαφορετικά περιβάλλοντα και η ποιότητα του μικροφώνου αμφισβητείται. Ως αποτέλεσμα, τα μοντέλα θα μπορούσαν να είναι πολύ επιρρεπή στην υπερβολική προσαρμογή σε ανεπιθύμητα σήματα. Για να αντιμετωπιστεί αυτό το ζήτημα, στα αρχεία ήχου που έχουν ταξινομηθεί ως ηχητικά σήματα βήχα αφαιρείται ο θόρυβος. Η επαύξηση δεδομένων εφαρμόζεται για την αντιμετώπιση του μικρού συνόλου δεδομένων, καθώς οι Αρχιτεκτονικές Βαθιάς Μάθησης (Deep Learning) απαιτούν μεγάλο όγκο δεδομένων. Στη συνέχεια, τα ηχητικά δείγματα μετατρέπονται σε φασματογραφήματα mel (mel spectrograms). Για την ταξινόμηση των δειγμάτων σε COVID-19 ή non COVID-19, δοκιμάζονται και παρουσιάζονται σε αυτή την εργασία εννέα διαφορετικές αρχιτεκτονικές βαθιάς μάθησης. Συγκεκριμένα, υλοποιούνται Συνελικτικά Νευρωνικά Δίκτυα (CNN) σε συνδυασμό με αμφίδρομα δίκτυα Long-Short-Term Memory (BiLSTM) και αμφίδρομα Gated Recurrent Units (BiGRU) σε συνδυασμό με μηχανισμό προσοχής (attention mechanism). Παρουσιάζονται επίσης τρία προεκπαιδευμένα δίκτυα στο ImageNet και ένα μοντέλο συνόλου που αποτελείται από αυτά. Επίσης υλοποιούνται το VGG-13 και το DenseNet Speech, μια αρχιτεκτονική που χρησιμοποιήθηκε σε προηγούμενη μελέτη για την αναγνώριση φωνής και τον εντοπισμό λέξεων-κλειδιών. Η μελέτη ανέδειξε ότι τα CRNN παρέχουν υποσχόμενα αποτελέσματα στην ανίχνευση της νόσου COVID-19. Στη συνέχεια, η διαδικασία εκμάθησης μεταφοράς πολλαπλών σταδίων αποτελείται από τρία στάδια μεταφοράς εκμάθησης και χρησιμοποιεί όλα τα διαθέσιμα σύνολα δεδομένων. Αυτή η προεκπαίδευση σε διαδικασίες που σχετίζονται με τον βήχα οδηγεί σε υψηλότερα αποτελέσματα ταξινόμησης για το σύνολο δεδομένων του Cambridge. Επιπρόσθετα, έγινε μια προσπάθεια ερμηνευσιμότητας του InceptionResnetV2 σε mel spectrograms με χρήση τοπικών ερμηνευτικών μοντέλων LIME (Local Interpretable **Model**-Agnostic Explanations). Τα καλύτερα αποτελέσματα ταξινόμησης, που προέκυψαν μέσω 5-fold cross validation και TCRNN, έχουν φτάσει σε ακρίβεια 76,67% και AUC 76,16%.Τα παραπάνω αποτελέσματα έδειξαν ότι τα αρχεία βήχα μπορούν να χρησιμοποιηθούν ως εργαλείο διαλογής/διάγνωσης για τη νόσο COVID-19.*

Keywords: *Βαθιά μάθηση, Ταξινόμηση Εικόνας, Προεπεξεργασία Ήχου, Πολυεπίπεδη Μεταφορά Μάθησης, Εκμάθηση Συνόλου, Ερμηνευσιμότητα, Συνελικτικά Επαναλαμβανόμενα Νευρωνικά Δίκτυα, Προεκπαιδευμένα Νευρωνικά Δίκτυα, Επαύξηση Δεδομένων, Ταξινόμηση Βήχα*

Abstract

Coronavirus disease of 2019 (COVID-19) has affected the lives of millions of people around the globe. Up until July 2022, there were 569.771.691 active cases of COVID-19 globally, and there had been 6.383.776 deaths. The virus is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. The most common occurring symptoms are fever, cough, and fatigue. The current diagnosis method is performed through Reverse-Transcription Polymer Chain Reaction (RT-PCR) testing. However, scarcity, cost, long turnaround time of clinical testing and the fact that they can lead to another infection if done improperly are some downsides of the RT-PCR testing. Furthermore, the in-person testing methods put the medical staff, particularly those with limited protection, at serious risk of infection. Vaccination remains a key component of the approach needed to reduce the impact of SARS-CoV-2. Unfortunately, variants of Covid-19 reduce at some point the effectiveness of vaccines, subsequently leading to reinfections. Therefore, the need for constant testing remains as immunity is often threatened by mutations. The current thesis aims at demonstrating the feasibility of the automatic detection of COVID-19 from cough sounds. This type of screening is non-contact, easy to apply, and can reduce the workload in testing centres as well as limit transmission. Two datasets have been used in this thesis, containing coughs from people from all continents, namely the “Coswara” and the “Cambridge” dataset. Dataset skew was addressed by applying an ensemble learning approach so that the Covid class is not underrepresented. The preprocessing step involves cough detection to identify if and when the cough is present in the raw audio recordings. The datasets are crowdsourced which means that the collected sounds are from differing environments and the quality of the microphone is disputed. As a result, the models could be highly prone to overfitting to unwanted signals. To address this issue, the sound files classified by the cough detector as cough are denoised. Data augmentation is applied to address data scarcity, since Deep Learning Architectures are data hungry. Then, audio samples are converted to mel spectrograms. For the Covid-19 classification task, nine different deep learning architectures are tested and presented in this thesis. Specifically, CNNs combined with bidirectional Long Short-Term Memory (BiLSTM) and bidirectional Gated Recurrent Units (BiGRU) networks in conjunction with an attention mechanism, are implemented. Three pretrained networks on ImageNet and an ensemble model consisting of them are presented as well. VGG-13 and DenseNet Speech, an architecture used to a prior study for voice recognition and keyword spotting, are also implemented. Temporal CRNNs seem to produce promising and consistent results in Covid-19 detection. Multistage Transfer Learning process consists of three stages of transfer learning and uses all of the available datasets. This pretraining on cough related tasks leads to higher classification results for the Cambridge dataset. Eventually, an interpretability attempt of InceptionResnetV2 has been made on mel spectrograms using Local Interpretable Model-agnostic Explanations. The best classification results, obtained through 5-fold cross validation and TCRNNs, have reached an accuracy of 76,67% and an AUC of 76,16%. These results demonstrate that cough can potentially serve as a helpful triage or diagnostic tool for Covid-19 infection. Since this type of cough audio classification is cost-effective and easy to deploy, it is potentially a useful and viable means of non-contact COVID-19 screening.

Keywords

COVID-19, Multistage Transfer Learning, Deep Learning, Audio Preprocessing, CRNN, Cough Detection, Image Classification, Pretrained Neural Networks, Attention Mechanisms, COVID-19 diagnosis, Ensemble Learning, Data Augmentation, SMOTE

Acknowledgements

I would like to express my deepest gratitude to my thesis advisor Prof. Konstantina Nikita for trusting me, encouraging me and supporting me throughout this thesis. I am extremely thankful to Dr. Eleni Adamidi and Dr. Kalliopi Dalakleidi for their invaluable assistance, their noble guidance and insights leading to the writing of this thesis. It is an honor to learn from them and an inspiration at the beginning of my research journey that I will never forget. I am grateful for my mother, Kalliopi, and my grandmother Georgia, whose endless love and support keep me motivated and confident. I am forever thankful for their unconditional love and encouragement throughout the entire thesis process and everyday.

Εκτεταμένη περίληψη

I. Covid – 19

Η νόσος COVID-19 (Corona Virus Disease of 2019), που προκαλείται από τον ιό του Σοβαρού Οξέος Αναπνευστικού Συνδρόμου Coronavirus 2 (SARS-CoV2), κηρύχθηκε παγκόσμια πανδημία στις 11 Φεβρουαρίου 2020 από τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ). Το Σοβαρό Οξύ Αναπνευστικό Σύνδρομο προηγουμένως ήταν πανδημία το 2003 [1]. Ερευνητικά στοιχεία υποδηλώνουν ότι οι SARS-CoV και MERS-CoV προέρχονται από νυχτερίδες. Μέχρι σήμερα, η προέλευση του SARS-CoV-2 που προκάλεσε την πανδημία COVID-19 δεν έχει εντοπιστεί. Τα μέχρι στιγμής επιστημονικά στοιχεία υποδηλώνουν ότι ο SARS-CoV-2 πιθανότατα προήλθε από την εξέλιξη του ιού στη φύση και μεταπήδησε στους ανθρώπους ή μέσω κάποιου αγνώστου ξενιστή ζώων [2]. Οι αναφορές εντοπίζουν το ξέσπασμα σε μια τεράστια αγορά που πουλούσε ζωντανά ζώα, μεταξύ άλλων αγαθών, στη Γουχάν της Κίνας και υποδηλώνουν ότι ο κορωνοϊός SARS-CoV-2 μεταδόθηκε από ζώα - πιθανώς αυτά που πωλούνται στην αγορά - σε ανθρώπους τουλάχιστον δύο φορές τον Νοέμβριο ή τον Δεκέμβριο του 2019 [3]. Μέχρι τη στιγμή της συγγραφής, υπάρχουν 569.771.691 ενεργά κρούσματα COVID-19 παγκοσμίως και έχουν σημειωθεί 6.383.776 θάνατοι, με τις ΗΠΑ να αναφέρουν τον υψηλότερο αριθμό κρουσμάτων (90.390 .184) και θανάτους (1.026.937)[4]. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, στην Ελλάδα έχουν επιβεβαιωθεί 4,21 εκατομμύρια κρούσματα και έχουν αναφερθεί 30.707 θάνατοι [5].

Οι κορωνοϊοί είναι σημαντικά παθογόνα που μπορούν να επηρεάσουν την κατώτερη αναπνευστική οδό στον άνθρωπο και μπορούν να προκαλέσουν ασθένειες που κυμαίνονται από ένα απλό κρυολόγημα έως σοβαρή μόλυνση με θνησιμότητα έως και 50% [6]. Ο COVID-19 φαίνεται να μην διαφέρει πολύ από τον SARS όσον αφορά τα κλινικά του χαρακτηριστικά. Ωστόσο, έχει ποσοστό θνησιμότητας 1,1%, χαμηλότερο από αυτό του SARS (9,5%) και πολύ χαμηλότερο από αυτό του MERS (34,4%) [7], αλλά μπορεί να διαφέρει σε άτομα που έχουν συννοσηρότητες [8].

Δημογραφικά Στοιχεία

Όπως αναφέρθηκε προηγουμένως, αυτή τη στιγμή υπάρχουν 569.771.691 ενεργά κρούσματα COVID-19 παγκοσμίως και έχουν σημειωθεί 6.383.776 θάνατοι, με τις ΗΠΑ να αναφέρουν τον υψηλότερο αριθμό κρουσμάτων (90.390.184) και θανάτων (1.026.937). [4]. Το Πανεπιστήμιο Johns Hopkins έχει δημιουργήσει και ενημερώνει καθημερινά ένα ανοιχτό αποθετήριο δεδομένων με διεθνή αναλυτικά στοιχεία για την πανδημία SARS-CoV-2 Η μείωση των μέτρων χρήσης μάσκας και οι αυξανόμενες τουριστικές ροές που έχουν ως αποτέλεσμα τον συνωστισμό, έχουν οδηγήσει σε επακόλουθη έξαρση καθημερινών κρουσμάτων τους καλοκαιρινούς μήνες του 2022 στην Ελλάδα.

Θνησιμότητα

Οι περισσότεροι άνθρωποι που έχουν μολυνθεί από τον ιό θα εμφανίσουν ήπια έως μέτρια αναπνευστική νόσο και θα αναρρώσουν χωρίς να χρειάζονται ειδική θεραπεία. Ωστόσο, ορισμένοι θα αρρωστήσουν σοβαρά και θα χρειαστούν ιατρική φροντίδα. Οι ηλικιωμένοι και εκείνοι με υποκείμενες ιατρικές παθήσεις όπως καρδιαγγειακή νόσο, διαβήτης, χρόνια αναπνευστική νόσος ή καρκίνος είναι πιο πιθανό να αναπτύξουν σοβαρή ασθένεια[5]. Το ποσοστό θνησιμότητας κρουσμάτων (CFR) του COVID-19 αναφέρεται ότι είναι 1,1%, αλλά μπορεί να διαφέρει σε ασθενείς που έχουν άλλες προϋπάρχουσες

παθήσεις και διαφέρει επίσης μεταξύ των χωρών. Σε ορισμένους ασθενείς, ειδικά σε αυτούς με άλλα υποκείμενα νοσήματα, μπορεί να υπάρχει αναπνευστική ανεπάρκεια, αρρυθμίες, σοκ, νεφρική ανεπάρκεια, καρδιαγγειακή βλάβη ή ηπατική ανεπάρκεια [12]. Ένας από τους πιο σημαντικούς δείκτες του COVID-19 είναι η θνησιμότητα. Χώρες σε όλο τον κόσμο έχουν αναφέρει πολύ διαφορετικές αναλογίες θνησιμότητας κρουσμάτων (δηλαδή ο αριθμός των θανάτων διαιρεμένος με τον αριθμό των επιβεβαιωμένων κρουσμάτων). Οι διαφορές στους αριθμούς θνησιμότητας μπορεί να προκληθούν από:

• Διαφορές στον αριθμό των ατόμων που εξετάζονται: Με περισσότερες εξετάσεις, εντοπίζονται περισσότερα άτομα με ηπιότερα περιστατικά. Αυτό μειώνει την αναλογία θνησιμότητας.

• Δημογραφικά στοιχεία: Για παράδειγμα, η θνησιμότητα τείνει να είναι υψηλότερη στους ηλικιωμένους πληθυσμούς.

• Χαρακτηριστικά του συστήματος υγειονομικής περίθαλψης: Για παράδειγμα, η θνησιμότητα μπορεί να αυξηθεί καθώς τα νοσοκομεία κατακλύζονται και έχουν λιγότερους πόρους.

Η Ελλάδα κατέχει την 4η θέση στην κατάταξη των χωρών ανάλογα με τους αριθμούς θανάτων ανά 100.000 πληθυσμού.

Μεταλλάξεις

Αν και οι περισσότερες μεταλλάξεις στο γονιδίωμα του κορωνοϊού 2 (SARS-CoV-2) του σοβαρού οξέος αναπνευστικού συνδρόμου αναμένεται να είναι είτε επιβλαβείς και να εξαφανίζονται γρήγορα ή σχετικά ουδέτερες, ένα μικρό ποσοστό θα επηρεάσει τις λειτουργικές ιδιότητες και μπορεί να αλλάξει τη μολυσματικότητα, τη σοβαρότητα της νόσου ή τις αλληλεπιδράσεις με την ανοσία του ξενιστή [14]. Υπάρχουν τρεις κατηγορίες παραλλαγών σύμφωνα με το Ευρωπαϊκό Κέντρο Πρόληψης και Ελέγχου Νόσων:

- Μεταλλάξεις ανησυχίας. Για αυτές τις παραλλαγές είναι διαθέσιμα σαφή στοιχεία που υποδεικνύουν σημαντική επίδραση στη μεταδοτικότητα, τη σοβαρότητα ή/και την ανοσία που είναι πιθανό να έχει αντίκτυπο στην επιδημιολογική κατάσταση.
- Μεταλλάξεις ενδιαφέροντος. Για αυτές τις παραλλαγές, υπάρχουν διαθέσιμα στοιχεία σχετικά με γονιδιωματικές ιδιότητες, επιδημιολογικά στοιχεία ή στοιχεία in vitro που θα μπορούσαν να συνεπάγονται σημαντικό αντίκτυπο στη μεταδοτικότητα, τη σοβαρότητα ή/και την ανοσία, έχοντας ρεαλιστικά αντίκτυπο στην επιδημιολογική κατάσταση. Ωστόσο, τα στοιχεία εξακολουθούν να είναι προκαταρκτικά ή συνδέονται με μεγάλη αβεβαιότητα.
- Μεταλλάξεις υπό παρακολούθηση. Αυτές οι πρόσθετες παραλλαγές του SARS-CoV-2 έχουν εντοπιστεί ως σήματα μέσω επιδημικής νοημοσύνης, ελέγχου γονιδιωματικών παραλλαγών βάσει κανόνων ή προκαταρκτικών επιστημονικών στοιχείων. Υπάρχουν κάποιες ενδείξεις ότι θα μπορούσαν να έχουν ιδιότητες παρόμοιες με αυτές μιας παραλλαγής που προκαλεί ανησυχία, αλλά τα στοιχεία είναι αδύναμα ή δεν έχουν ακόμη αξιολογηθεί από το Ευρωπαϊκό Κέντρο Πρόληψης και Ελέγχου Νόσων (ECDC).

Πρόληψη και Εμβολιασμός

Σε όλο τον κόσμο εφαρμόζονται μέτρα δημόσιας υγείας και κοινωνικής υγείας για την καταστολή της μετάδοσης του SARS-CoV-2 και τη μείωση της θνησιμότητας και της νοσηρότητας από τον COVID-19. Τα μέτρα αυτά περιλαμβάνουν μέτρα ατομικής προστασίας (π.χ. καθαρισμός, απολύμανση, εξαερισμός)· μέτρα επιτήρησης και απόκρισης (π.χ. δοκιμές, γενετική αλληλουχία, ιχνηλάτηση επαφών, απομόνωση

και καραντίνα)· μέτρα φυσικής απόστασης και διεθνή μέτρα που σχετίζονται με τα ταξίδια[17]. Οι οδηγίες του ΠΟΥ για την πρόληψη της μόλυνσης από τον SARS-CoV-2 είναι:

• Διατήρηση φυσικής απόστασης τουλάχιστον 1 μέτρου από τους άλλους, ακόμα κι αν δεν φαίνεται να είναι άρρωστοι. Αποφυγή των πληθών και της στενής επαφής.

• Σωστά τοποθετημένη μάσκα όταν δεν είναι δυνατή η φυσική απόσταση και σε χώρους με ανεπαρκή αερισμό.

• Συχνός καθαρισμός των χεριών με τρίψιμο χεριών με βάση το οινόπνευμα ή σαπούνι και νερό.

• Κάλυψη του στόματος και της μύτης με λυγισμένο αγκώνα ή χαρτομάντιλο σε περίπτωση βήχα ή φτερνίσματος. Απόρριψη των χρησιμοποιημένων χαρτομάντιλων αμέσως και καθαρισμός των χεριών τακτικά.

• Αυτο-απομόνωση εάν αναπτυχθούν συμπτώματα μόλυνσης.

• Εμβολιασμός

Λόγω της πανδημίας κορωνοϊού, έχουν εφαρμοστεί σε πολλές χώρες και περιοχές σε όλο τον κόσμο μια σειρά από μη φαρμακευτικές παρεμβάσεις γνωστές στην καθομιλουμένη ως lockdowns (που περιλαμβάνουν εντολές παραμονής στο σπίτι, απαγόρευση κυκλοφορίας, καραντίνες, υγειονομικά κλεισίματα και παρόμοιους κοινωνικούς περιορισμούς). .

Φυσικά το επικρατέστερο μέτρο κατά της εξάπλωσης του Covid-19 είναι ο εμβολιασμός. Είναι ο ασφαλέστερος και πιο αξιόπιστος τρόπος για να δημιουργηθεί ανοσία σε σύγκριση με τη νόσηση. Ο εμβολιασμός κατά του COVID-19 βοηθά στην προστασία δημιουργώντας μια απόκριση αντισωμάτων χωρίς να χρειάζεται η αντιμετώπιση δυνητικά σοβαρής ασθένειας. Τα πρώτα εμβόλια Covid-19 χορηγήθηκαν με άδεια χρήσης έκτακτης ανάγκης τον Δεκέμβριο του 2020, μόλις ένα χρόνο μετά την πανδημία, ένα «θαύμα» φαρμακευτικής καινοτομίας που έχει σώσει περίπου εκατομμύρια ζωές ή περισσότερες μόνο στις ΗΠΑ. Η ανοσοποίηση με εμβόλια Pfizer-BioNTech και Moderna mRNA προστάτευσε ένα αξιοσημείωτα υψηλό ποσοστό (>90%) των ληπτών από την ανάπτυξη συμπτωματικής λοίμωξης και, σε μικρότερο βαθμό, επίσης από ασυμπτωματική μόλυνση. Κατά το πρώτο εξάμηνο του 2021, όταν η άλφα παραλλαγή του SARS-CoV-2 ήταν κυρίαρχη, το ποσοστό θνησιμότητας από τον Covid-19 μειώθηκε κατά 60%, 75% και 81% σε κομητείες με χαμηλή, μεσαία και υψηλή εμβολιαστική κάλυψη, σε σύγκριση με κομητείες που είχαν πολύ χαμηλή κάλυψη [18]. Τον Μάιο του 2021, η Υπηρεσία Τροφίμων και Φαρμάκων των Ηνωμένων Πολιτειών και ο Ευρωπαϊκός Οργανισμός Φαρμάκων (EMA) ενέκριναν τη χρήση του εμβολίου Pfizer–BioNTech, Comirnaty, για παιδιά ηλικίας 12–15 ετών. Στις 25 Νοεμβρίου 2021, ο EMA επέκτεινε αυτήν την εξουσιοδότηση σε παιδιά ηλικίας 5 - 11 ετών. Τα εμβόλια που είναι εγκεκριμένα για χρήση στην Ευρωπαϊκή Ένωση είναι τα Comirnaty (Pfizer-BioNTech), Jcovden (προηγουμένως COVID-19 Vaccine Janssen), Nuvaxovid (Novanax), Spikevax (Moderna) και Vaxzevria (AstraZeneca). Οι αναμνηστικές δόσεις έχουν χορηγηθεί τουλάχιστον 3 μήνες μετά τη δεύτερη δόση σε άτομα ηλικίας 12 ετών και άνω. Μια τέταρτη αναμνηστική δόση έχει ληφθεί στην Ελλάδα για ευπαθείς ομάδες με συννοσηρότητες. Οι αξιωματούχοι της δημόσιας υγείας ανησυχούν από την αρχή της πανδημίας ότι οι εμβολιασμοί δεν θα κατανέμονται δίκαια σε όλο τον κόσμο. Τα δεδομένα φαίνεται να επιβεβαιώνουν αυτούς τους φόβους καθώς τα ανεπτυγμένα έθνη εμβολιάζουν τους πληθυσμούς τους πολύ πιο γρήγορα από τις λιγότερο ανεπτυγμένες χώρες[4].

Θεραπεία

Ο Οργανισμός Τροφίμων και Φαρμάκων των ΗΠΑ (FDA) ενέκρινε το αντιικό φάρμακο Veklury (remdesivir) για ενήλικες και ορισμένους παιδιατρικούς ασθενείς με COVID-19. Αυτή είναι μια ενδοφλέβια θεραπεία. Ο FDA έχει επίσης εγκρίνει τον ανοσοδιαμορφωτή Olumiant (baricitinib) για ορισμένους νοσηλευόμενους ενήλικες με COVID-19. Σε καταστάσεις έκτακτης ανάγκης για τη δημόσια υγεία, ο FDA μπορεί να εγκρίνει τη χρήση μη εγκεκριμένων φαρμάκων ή μη εγκεκριμένων χρήσεων εγκεκριμένων φαρμάκων υπό ορισμένες προϋποθέσεις. Αυτό ονομάζεται εξουσιοδότηση χρήσης έκτακτης ανάγκης (EUA). Ο FDA έχει εκδώσει EUAs για αρκετές θεραπείες μονοκλωνικών αντισωμάτων, για COVID-19 για τη θεραπεία και σε ορισμένες περιπτώσεις την πρόληψη (προφύλαξη) του COVID-19 σε ενήλικες και παιδιατρικούς ασθενείς. Τα μονοκλωνικά αντισώματα είναι εργαστηριακά κατασκευασμένα μόρια που δρουν ως υποκατάστατα αντισώματα. Υπάρχουν επίσης δύο από του στόματος αντιικά χάπια, το Paxlovid και το Lagevrio (μολνουπιραβίρη), εγκεκριμένα για ασθενείς με ήπιο έως μέτριο COVID-19 [20]. Το Molnupiravir είναι το πρώτο από του στόματος, άμεσης δράσης αντιικό που έχει αποδειχθεί ότι είναι ιδιαίτερα αποτελεσματικό στη μείωση του ρινοφαρυγγικού μολυσματικού ιού SARS-CoV-2 και του ικού RNA και έχει ευνοϊκό προφίλ ασφάλειας και ανεκτικότητας [21]. Ο Ευρωπαϊκός Οργανισμός Φαρμάκων (EMA) έχει εγκρίνει για χρήση στην Ευρωπαϊκή Ένωση τα ακόλουθα φάρμακα: Evusheld(tixagevimab / cilgavimab), Kineret (anakinra), Paxlovid (PF-07321332 / ριτοναβίρη), Regkirona (regdanvimab), RoActemra(tocilizumab), Ronapreve (casirivimab / imdevimab), Veklury (remdesivir), Xevudy (sotrovimab). Ο EMA αξιολογεί επί του παρόντος τις αιτήσεις άδειας κυκλοφορίας για το Olumiant και το Lagevrio [22]. Δυστυχώς, οι μεταβαλλόμενες παραλλαγές του ιού επηρεάζουν την αποτελεσματικότητα των θεραπειών που μπορεί να αποσυρθούν εάν αποδειχθούν αναποτελεσματικές έναντι μιας συγκεκριμένης παραλλαγής.

Μεταδοτικότητα

Ορισμένες δημοσιευμένες μελέτες έχουν υπολογίσει ότι ο R0 (δηλαδή ο αναπαραγωγικός αριθμός) για το SARS φθάνει την τιμή του 4. Είναι ενδιαφέρον ότι μια πρόσφατη ανασκόπηση από τον Liu και τους συνεργάτες του [24] έδειξε ότι ο μέσος αναπαραγωγικός αριθμός του SARS-CoV-2 εκτιμάται ότι είναι 3,28 , με μέση τιμή 2,79, υπερβαίνοντας έτσι τις εκτιμήσεις του ΠΟΥ [7]. Υπάρχουν δύο τρόποι μετάδοσης - άμεση και έμμεση. Η άμεση λειτουργία περιλαμβάνει μετάδοση μέσω αεροζόλ, δακρύων, σάλιου, σπέρματος και από μητέρα σε παιδί. Οι έμμεσοι τρόποι περιλαμβάνουν τη μετάδοση μέσω μολυσμένων αντικειμένων [25]. Οι συγγραφείς καταλήγουν στο συμπέρασμα ότι η μετάδοση από μητέρα σε παιδί μπορεί να είναι σπάνια, αλλά όχι εντελώς απούσα. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ), ο SARS-CoV-2 μεταδίδεται μεταξύ των ανθρώπων με διάφορους τρόπους. Τα τρέχοντα στοιχεία δείχνουν ότι ο ιός εξαπλώνεται κυρίως μεταξύ ατόμων που βρίσκονται σε στενή επαφή μεταξύ τους, για παράδειγμα σε απόσταση συνομιλίας. Ο ιός μπορεί να εξαπλωθεί από το στόμα ή τη μύτη ενός μολυσμένου ατόμου σε μικρά σωματίδια υγρού όταν βήχει, φτερνίζεται, μιλάει, τραγουδάει ή αναπνέει. Ένα άλλο άτομο μπορεί στη συνέχεια να προσβληθεί από τον ιό όταν τα μολυσματικά σωματίδια που διέρχονται από τον αέρα εισπνέονται σε μικρή απόσταση (αυτό ονομάζεται συχνά αεροζόλ μικρής εμβέλειας) ή εάν μολυσματικά σωματίδια έρθουν σε άμεση επαφή με τα μάτια, τη μύτη. ή στο στόμα (μετάδοση σταγονιδίων). Ο ιός μπορεί επίσης να εξαπλωθεί σε ανεπαρκώς αεριζόμενους ή πολυσύχναστους εσωτερικούς χώρους, όπου οι άνθρωποι τείνουν να περνούν μεγαλύτερες χρονικές περιόδους. Αυτό συμβαίνει επειδή τα αερολύματα μπορούν να παραμείνουν αιωρούμενα στον αέρα ή να ταξιδεύουν μακρύτερα από την απόσταση συνομιλίας (αυτό ονομάζεται συχνά αεροζόλ μεγάλης εμβέλειας). Τέλος, οι άνθρωποι μπορεί επίσης να μολυνθούν όταν αγγίζουν τα μάτια, τη μύτη ή το στόμα τους αφού αγγίξουν επιφάνειες ή αντικείμενα που έχουν μολυνθεί από τον ιό.

Μέθοδοι Ανίχνευσης

Περισσότερα από έξι δισεκατομμύρια τεστ για COVID-19 έχουν ήδη πραγματοποιηθεί στον κόσμο. Ο έλεγχος για τον ιό SARS-CoV-2 και τα αντίστοιχα ανθρώπινα αντισώματα είναι απαραίτητος όχι μόνο για τη διάγνωση και τη θεραπεία της λοίμωξης από ιατρικά ιδρύματα, αλλά και ως προϋπόθεση για σημαντικές κοινωνικές δραστηριότητες [26]. Οι τεχνικές ανίχνευσης του ιού περιλαμβάνουν την ανίχνευση ιικών σωματιδίων (virions), ιικού αντιγόνου, αντισωμάτων έναντι του ιού και ιικού νουκλεϊκού οξέος. Υπάρχουν τρεις κατηγορίες μεθόδων εξέτασης: διαγνωστικές δοκιμές PCR, διαγνωστικές εξετάσεις αντιγόνου και δοκιμές αντισωμάτων. Οι κύριες μέθοδοι ανίχνευσης του ιού SARS-CoV-2 βασίζονται στην ανίχνευση ιικού RNA. Η PCR είναι μια από τις κοινές τεχνικές που χρησιμοποιούνται για την ανίχνευση ιικού νουκλεϊκού οξέος. Η εξέταση συνήθως εκτελείται σε δείγμα ρινικού επιχρίσματος ή σάλιου. Το τεστ χρησιμοποιεί μια τεχνολογία γνωστή ως αλυσιδωτή αντίδραση πολυμεράσης (PCR) για να ανιχνεύσει ίχνη γενετικού υλικού του SARS-CoV-2. Η ανίχνευση ιικών σωματιδίων και αντιγόνου είναι μια βιώσιμη εναλλακτική λύση στην RT-PCR [27]. Αυτές οι μέθοδοι είναι δυνητικά φθηνές, φορητές, γρήγορες και μπορούν να χρησιμοποιηθούν για τη διάγνωση ασθενών στο πρώιμο στάδιο της ιογενούς λοίμωξης. Δεν απαιτείται να εκτελούνται από εξειδικευμένο χειριστή και μπορούν να εκτελεστούν από τους ίδιους τους ασθενείς. Η εξέταση πραγματοποιείται συνήθως σε δείγμα ρινικού ή λαιμού. Ανιχνεύει θραύσματα συγκεκριμένων ιικών πρωτεϊνών. Οι εξετάσεις αντισωμάτων ή ορολογικών εξετάσεων μπορούν να βρουν εάν ένα άτομο πιθανότατα είχε προηγούμενη λοίμωξη από SARS-CoV-2. Αυτή η εξέταση αίματος δεν διαγιγνώσκει ενεργή λοίμωξη ούτε παρέχει πληροφορίες σχετικά με τη μακροπρόθεσμη ανοσία. Τα ειδικά αντισώματα IgG, IgM και IgA του SARS-CoV-2 γίνονται συχνότερα αντικείμενα ανίχνευσης χρησιμοποιώντας διαφορετικές μεθόδους. Τα αντισώματα IgM εμφανίζονται στην οξεία φάση της μόλυνσης και αφού φτάσουν στο μέγιστο, μειώνονται σε διαγνωστικά ασήμαντα επίπεδα. Τα αντισώματα IgG συσσωρεύονται πιο αργά από τα αντισώματα IgM, αλλά παραμένουν υψηλά στο αίμα του ασθενούς περισσότερο. Μετά την ανάρρωση, τα αντισώματα IgG μπορούν να παραμείνουν σε χαμηλό επίπεδο επ' αόριστον ως ένδειξη προηγούμενης λοίμωξης.

Συμπτώματα

Σύμφωνα με τους Russel M Viner et al. πυρετός και βήχας ήταν τα πιο κοινά συμπτώματα. τα ποσοστά με πυρετό κυμαίνονταν από 46% έως 64,2% και με βήχα από 32% έως 55,9%. Όλα τα άλλα συμπτώματα ή σημεία, συμπεριλαμβανομένης της ρινόρροιας, του πονόλαιμου, του πονοκεφάλου, της κόπωσης/μυαλγίας και των γαστρεντερικών συμπτωμάτων, συμπεριλαμβανομένων της διάρροιας και του εμέτου, ήταν σπάνια και εμφανίζονταν σε λιγότερο από 10%-20% [28]. Τα συμπτώματα του Covid-19 κυμαίνονται καθώς επικρατούν νέες παραλλαγές. Επιπλέον, ο ΠΟΥ έχει συμπεριλάβει ως πολύ κοινά συμπτώματα, εκτός από πυρετό και βήχα, κόπωση, δύσπνοια και απώλεια γεύσης ή όσφρησης. . Το Long Covid αναφέρεται όταν οι άνθρωποι συνεχίζουν να εμφανίζουν συμπτώματα του COVID-19 και δεν αναρρώνουν πλήρως για αρκετές εβδομάδες ή μήνες μετά την έναρξη των συμπτωμάτων τους. Τα πέντε πιο κοινά συμπτώματα ήταν κόπωση (58%), πονοκέφαλος (44%), διαταραχή προσοχής (27%), απώλεια μαλλιών (25%) και δύσπνοια (24%). Άλλα συμπτώματα σχετίζονταν με πνευμονική νόσο (βήχας, δυσφορία στο στήθος, μειωμένη πνευμονική ικανότητα διάχυσης, άπνοια ύπνου και πνευμονική ίνωση), καρδιαγγειακά (αρρυθμίες, μυοκαρδίτιδα), νευρολογικά (άνοια, κατάθλιψη, άγχος, διαταραχή προσοχής, ιδεοψυχαναγκαστικές διαταραχές). και άλλα ήταν μη ειδικά όπως η απώλεια μαλλιών, οι εμβοές και ο νυχτερινός ιδρώτας. Εντόπισαν συνολικά 55 μακροπρόθεσμες επιδράσεις που σχετίζονται με τον COVID-19 στη βιβλιογραφία που ανασκοπήθηκε. Οι περισσότερες από τις επιδράσεις αντιστοιχούν σε

κλινικά συμπτώματα όπως κόπωση, πονοκέφαλος, πόνος στις αρθρώσεις, ανοσμία κ.λπ. Επιπλέον, υπήρχαν επίσης ασθένειες όπως το εγκεφαλικό και ο σακχαρώδης διαβήτης [29].

II. Βιβλιογραφική Ανασκόπηση

Προηγούμενες μελέτες έχουν δείξει ότι ο βήχας από διακριτά αναπνευστικά σύνδρομα έχει διακριτά λανθάνοντα χαρακτηριστικά. Η ανίχνευση βήχα είναι ένα βήμα προεπεξεργασίας που καθορίζει εάν υπάρχει ή όχι βήχας σε ένα αρχείο ήχου. Η δυσκολία αναγνώρισης του βήχα έγκειται κυρίως στο θόρυβο του περιβάλλοντος. Μερικά από τα βαθιά νευρωνικά δίκτυα που δοκιμάστηκαν για την ανίχνευση βήχα ήταν το YAMNNet [50] και το Ubicoustics [51]. Αυτές οι αρχιτεκτονικές είναι σε θέση να ταξινομήσουν ένα ευρύ φάσμα ήχων που εμφανίζονται συχνά στο περιβάλλον. Επιπλέον, εκτός από τα βαθιά νευρωνικά δίκτυα, οι ερευνητές εφάρμοσαν έναν ταξινομητή XGBoost στο σύνολο δεδομένων Coughvid για να αφαιρέσουν εγγραφές χωρίς βήχα χρησιμοποιώντας το 78% των διαθέσιμων δεδομένων [54] . Στην τρέχουσα εργασία, για το συγκεκριμένο έργο της ανίχνευσης βήχα έχει χρησιμοποιηθεί το σύστημα ανίχνευσης βήχα που περιγράφεται εδώ [55]. Σε αυτό το σύστημα οι Simou et al. πέτυχαν ακρίβεια της τάξης του 90% και specificity της τάξης του 99% σε οικιακό περιβάλλον με την αξιοποίηση της αρχιτεκτονικής βαθιάς νευρωνικών δικτύων Long-Short-Term-Memory. Οι Bales et al. [56] πρότειναν ένα σύστημα που μπόρεσε με επιτυχία να ανιχνεύσει και να διαχωρίσει τα συμβάντα βήχα από τον θόρυβο του περιβάλλοντος. Χρησιμοποίησαν CNN για να εντοπίσουν πρώτα και να διαχωρίσουν τους ήχους βήχα από διαφορετικούς τύπους ήχων. Στη συνέχεια χρησιμοποίησαν τους ήχους βήχα που ανιχνεύτηκαν για να διαγνώσουν τρεις πιθανές ασθένειες (δηλαδή, βρογχίτιδα, βρογχιολίτιδα και κοκκύτη) με βάση τα μοναδικά χαρακτηριστικά ήχου του βήχα σε ένα ενοποιημένο πλαίσιο. Οι Quan et al. [39] πρότειναν μια μέθοδο αναγνώρισης βήχα που βασίζεται σε φασματογράμματα Mel και ένα Συνελικτικό Νευρωνικό Δίκτυο. Οι Infante et al. χρησιμοποίησαν μια μέθοδο μηχανικής μάθησης για να αναγνωρίσουν τον ξηρό/υγρό βήχα [57]. Η ημι-εποπτευόμενη διανυσματική μηχανή υποστήριξης δέντρων προτείνεται για αναγνώριση και ανίχνευση βήχα. Το K-NN είναι επίσης ένα αποτελεσματικό εργαλείο που χρησιμοποιείται συχνά για την αναγνώριση του βήχα [58]. Επιπλέον, το Τεχνητό Νευρωνικό Δίκτυο (ANN), το Gaussian Mixture Model (GMM), το Support Vector Machine (SVM) και άλλες μέθοδοι χρησιμοποιούνται επίσης για την αναγνώριση του βήχα [59]. Πολλές μελέτες έχουν προσπαθήσει να χρησιμοποιήσουν την τεχνητή νοημοσύνη για την ταξινόμηση του Covid-19 χρησιμοποιώντας ήχους βήχα. Οι ερευνητές του MIT ανέπτυξαν ένα πλαίσιο ομιλίας AI για την ανίχνευση του Covid-19 από καταγραφές βήχα[60]. Οι Imran et al. [61] παρουσίασαν ένα αναπτυσσόμενο εργαλείο προκαταρκτικής διάγνωσης βασισμένο σε AI για τον COVID-19 χρησιμοποιώντας ήχους βήχα μέσω μιας εφαρμογής που ονομάζεται AI4COVID-19. Ο Brown et.al [62] χρησιμοποίησαν έναν αλγόριθμο που βασίζεται στη μηχανική μάθηση για να διακρίνει τους ήχους βήχα υγιούς και COVID-19 (πληθυσμιακά δεδομένα) από ασθενείς ακόμη και με προϋπάρχουσες καταστάσεις άσθματος. Οι συγγραφείς αναφέρουν μια μέση μέτρηση AUC 70% για τις εργασίες που αναφέρονται στη μελέτη. Μετά από αυτήν την προσπάθεια, το έργο Coswara [63] συνέταξε ένα σύνολο δεδομένων που περιέχει μια ποικιλία ήχων, συμπεριλαμβανομένων των παρατεταμένων φωνημάτων, του βήχα και των μοτίβων αναπνοής. Οι Pahar et al.[64] ανέπτυξαν ταξινομητές βήχα COVID-19 χρησιμοποιώντας ηχογραφήσεις smartphone και επτά αρχιτεκτονικές μηχανικής εκμάθησης. Οι Chaudhari et al. [67] διαπίστωσαν ότι ένα μοντέλο συνόλου τριών χαρακτηριστικών έδειξε την καλύτερη απόδοση. Το πρώτο χαρακτηριστικό ήταν τα MFCCs, το δεύτερο τα mel spectrograms και το τελευταίο ήταν μια δυαδική ετικέτα σχετικά με την παρουσία ή την απουσία τρεχουσών αναπνευστικών ασθενειών.

III. Στόχος της διπλωματικής

Σε αυτή την εργασία, ο Covid-19 ανιχνεύεται με μεθόδους βαθιάς μάθησης από δείγματα βήχα. Τα βήματα προεπεξεργασίας αποτελούνται από αυτόματη αναγνώριση βήχα και μείωση θορύβου για μια

ευαίσθητη ανάλυση. Η επαύξηση δεδομένων υλοποιείται επίσης στα αρχεία ήχου χωρίς θόρυβο πριν αυτά τροφοδοτηθούν στις περισσότερες αρχιτεκτονικές βαθιάς εκμάθησης. Στη συνέχεια, τα αρχεία ήχου μετατρέπονται σε mel spectrograms και το πρόβλημα αντιμετωπίζεται ως εργασία ταξινόμησης δυαδικής εικόνας. Παρουσιάζονται εννέα αρχιτεκτονικές και μία από αυτές είναι μια συνολική προσέγγιση τριών προεκπαιδευμένων μοντέλων. Η ανισορροπία δεδομένων αντιμετωπίζεται με διάφορες τεχνικές, όπως η εκμάθηση συνόλου, η SMOTE και η τυχαία υπερδειγματοληψία, με τις δύο τελευταίες να εφαρμόζονται στη μάθηση μεταφοράς πολλαπλών σταδίων. Τέλος, το LIME χρησιμοποιείται για την ερμηνεία των αποτελεσμάτων των παραπάνω αρχιτεκτονικών βαθιάς μάθησης.

IV. Βαθιά Μάθηση

Η βαθιά μάθηση (DL), ένας κλάδος της μηχανικής μάθησης (ML) και της τεχνητής νοημοσύνης (AI) θεωρείται σήμερα ως η βασική τεχνολογία της σημερινής Τέταρτης Βιομηχανικής Επανάστασης. Λόγω των δυνατοτήτων εκμάθησής της από δεδομένα, η τεχνολογία DL εφαρμόζεται ευρέως σε διάφορους τομείς εφαρμογών όπως η υγειονομική περίθαλψη, η οπτική αναγνώριση, η ασφάλεια στον κυβερνοχώρο και πολλά άλλα [75]. Στα τέλη της δεκαετίας του 1980, τα νευρωνικά δίκτυα έγιναν ένα διαδεδομένο θέμα στον τομέα της Μηχανικής Μάθησης (ML) καθώς και της Τεχνητής Νοημοσύνης (AI), λόγω της εφεύρεσης διαφόρων αποτελεσματικών μεθόδων εκμάθησης και δομών δικτύου, όπως τα δίκτυα perceptron πολλαπλών στρωμάτων που εκπαιδεύτηκαν από αλγόριθμους τύπου 'Backpropagation', χάρτες αυτοοργάνωσης και δίκτυα συναρτήσεων ακτινικής βάσης [76]. Το 2006, το «Deep Learning» (DL) εισήχθη από τους Hinton et al.[77] , το οποίο βασίστηκε στην έννοια του τεχνητού νευρωνικού δικτύου (ANN). Η τεχνολογία DL χρησιμοποιεί πολλαπλά επίπεδα για να αναπαραστήσει τις αφαιρέσεις δεδομένων για τη δημιουργία υπολογιστικών μοντέλων. Ένα τυπικό νευρωνικό δίκτυο αποτελείται κυρίως από πολλά απλά, συνδεδεμένα στοιχεία επεξεργασίας ή επεξεργαστές που ονομάζονται νευρώνες, καθένας από τους οποίους δημιουργεί μια σειρά από ενεργοποιήσεις πραγματικής αξίας για το αποτέλεσμα-στόχο. Ο Sarker [78] στην εργασία του περιέγραψε τις διαφορετικές εργασίες βαθιάς μάθησης σύμφωνα με τις οποίες προκύπτει η ταξινόμηση των δικτύων βαθιάς μάθησης.Οι τεχνικές DL χωρίζονται σε τρεις μεγάλες κατηγορίες: (i) βαθιά δίκτυα για εποπτευόμενη ή διακριτική μάθηση. (ii) βαθιά δίκτυα για μη εποπτευόμενη ή παραγωγική μάθηση. και (iii) βαθιά δίκτυα για υβριδική μάθηση που συνδυάζει και τα δύο και τα σχετικά άλλα.

Συνελικτικά Νευρωνικά Δίκτυα

Ως ένα είδος μεθόδου βαθιάς μάθησης, τα Συνελικτικά Νευρωνικά Δίκτυα (CNN) χρησιμοποιούνται ευρέως στον τομέα της όρασης υπολογιστών. Τα CNN προορίζονται ειδικά για την αντιμετώπιση μιας ποικιλίας σχημάτων 2D και επομένως χρησιμοποιούνται ευρέως στην οπτική αναγνώριση, την ανάλυση ιατρικών εικόνων, την κατάτμηση εικόνας, την επεξεργασία φυσικής γλώσσας και πολλά άλλα. Τα CNN αποτελούνται από τρεις τύπους επιπέδων. Αυτά είναι τα συνελικτικά στρώματα, τα pooling στρώματα και τα πλήρως συνδεδεμένα στρώματα. Όταν αυτά τα επίπεδα στοιβάζονται, έχει διαμορφωθεί μια αρχιτεκτονική CNN. Το συνελικτικό στρώμα θα καθορίσει την έξοδο των νευρώνων που συνδέονται με τις τοπικές περιοχές της εισόδου μέσω του υπολογισμού του βαθμωτού γινομένου μεταξύ των βαρών τους και της περιοχής που συνδέεται με τον όγκο εισόδου. Στη συνέχεια, το επίπεδο συγκέντρωσης θα πραγματοποιήσει απλώς μείωση δειγματοληψίας κατά μήκος της χωρικής διάστασης της δεδομένης εισόδου, μειώνοντας περαιτέρω τον αριθμό των παραμέτρων εντός της ενεργοποίησης.

Επαναλαμβανόμενα Νευρωνικά Δίκτυα

Ένα επαναλαμβανόμενο νευρωνικό δίκτυο (RNN) είναι ένα άλλο δημοφιλές νευρωνικό δίκτυο, το οποίο χρησιμοποιεί διαδοχικά ή δεδομένα χρονοσειρών και τροφοδοτεί την έξοδο από το προηγούμενο βήμα ως είσοδο στο τρέχον στάδιο [75]. Όπως το feedforward και το CNN, τα επαναλαμβανόμενα δίκτυα

μαθαίνουν από την εισαγωγή εκπαίδευσης, ωστόσο, διακρίνονται από τη «μνήμη» τους, η οποία τους επιτρέπει να επηρεάσουν την τρέχουσα είσοδο και έξοδο χρησιμοποιώντας πληροφορίες από προηγούμενες εισόδους. Σε αντίθεση με το τυπικό DNN, το οποίο υποθέτει ότι οι είσοδοι και οι έξοδοι είναι ανεξάρτητες η μία από την άλλη, η έξοδος του RNN εξαρτάται από προηγούμενα στοιχεία εντός της ακολουθίας. Οι πιο διαδεδομένες παραλλαγές των RNN είναι η Long Short Term Memory (LSTM), η αμφίδρομη RNN/LSTM και οι Gated recurrent units (GRU).

V. Σύνολα Δεδομένων

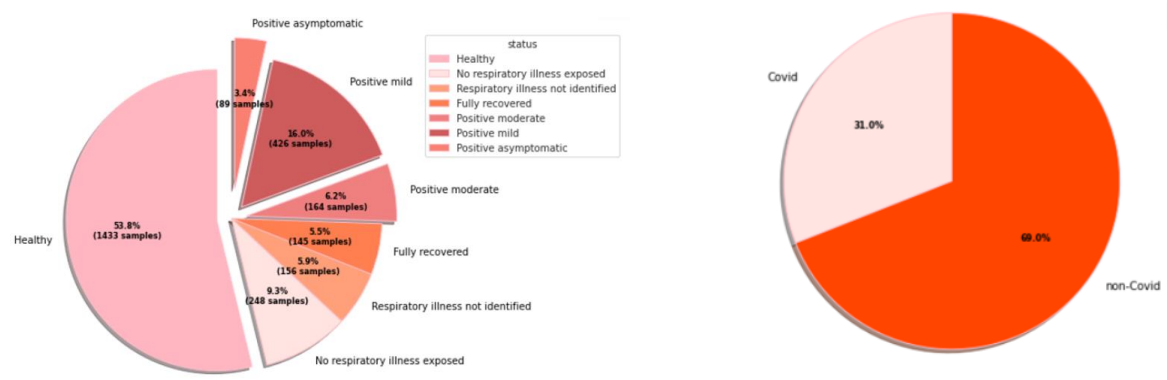
Δύο σύνολα δεδομένων έχουν χρησιμοποιηθεί για την εργασία ταξινόμησης της διάγνωσης του Covid-19 από δείγματα βήχα. Πιο συγκεκριμένα, πρόκειται για το σύνολο δεδομένων Coswara και το σύνολο δεδομένων Cambridge. Το Coswara στοχεύει στην ανάπτυξη ενός διαγνωστικού εργαλείου για τον Covid-19 με βάση τους αναπνευστικούς ήχους, τον βήχα και τους ήχους ομιλίας[63],[113]. Η συλλογή δεδομένων έγινε μέσω web εφαρμογή όπου ζητήθηκε από τους χρήστες να παράσχουν μεταδεδομένα και να προχωρήσουν στην εγγραφή των δειγμάτων ήχου χρησιμοποιώντας το μικρόφωνο της συσκευής. Οι δημόσιοι συμμετέχοντες παρείχαν 9 αρχεία ήχου ένα για κάθε κατηγορία ήχου: αναπνοή (δύο τύποι, ρηχά και βαθιά), βήχας (δύο τύποι, ρηχός και βαρύς), παρατεταμένη προφορά φωνηέντων (τρεις τύποι, / ey /, / i/,/ u: /), και μετρώντας από ένα έως είκοσι ψηφία (δύο ειδών, κανονικό και γρήγορο). Για κάθε χρήστη, τα μεταδεδομένα μπορούν να ομαδοποιηθούν σε πέντε διακριτές κατηγορίες: ηλικία, φύλο, τοποθεσία (χώρα, πολιτεία/επαρχία), τρέχουσα κατάσταση υγείας και παρουσία συννοσηροτήτων (προϋπάρχουσες ιατρικές παθήσεις). Η κατάσταση της υγείας περιλαμβάνει «υγιή», «εκτεθειμένη», «θεραπευμένη» ή «μολυσμένη». Σε αυτή τη μελέτη, χρησιμοποιήσαμε τις ακατέργαστες ηχογραφήσεις του ρηχού βήχα και του έντονου βήχα ως δύο ξεχωριστά σύνολα δεδομένων και εφαρμόσαμε προεπεξεργασία. Υπάρχουν 2744 δείγματα συνολικά, αλλά μόνο 2661 από αυτά έχουν περιγραφή κατάστασης υγείας, αφού υπάρχουν 83 δείγματα υπό επικύρωση. Το Σχήμα 1α παρουσιάζει την κατανομή των δειγμάτων με βάση την κατάσταση υγείας.

Το σύνολο δεδομένων του Cambridge είναι ένα σύνολο δεδομένων που συλλέγεται μέσω μιας εφαρμογής (Android και Web) που ζητούσε από εθελοντές δείγματα της φωνής, του βήχα και της αναπνοής τους, καθώς και το ιατρικό ιστορικό και τα συμπτώματά τους[62]. Ο χρήστης καλείται να εισαγάγει την ηλικία και το φύλο του καθώς και ένα σύντομο ιατρικό ιστορικό και εάν νοσηλεύεται στο νοσοκομείο. Στη συνέχεια, οι χρήστες εισάγουν τα συμπτώματά τους (αν υπάρχουν) και καταγράφουν αναπνευστικούς ήχους: καλούνται να βήξουν τρεις φορές, να αναπνεύσουν βαθιά από το στόμα τους τρεις έως πέντε φορές και να διαβάσουν μια σύντομη πρόταση που εμφανίζεται στην οθόνη τρεις φορές. Τέλος, οι χρήστες ερωτώνται εάν έχουν ελεγχθεί για COVID-19 και συλλέγεται δείγμα τοποθεσίας με άδεια. Υπάρχουν 141 δείγματα Covid από 66 μοναδικούς χρήστες και 298 δείγματα μη Covid που αποκτήθηκαν από 220 μοναδικούς χρήστες. Τελικά, χρησιμοποιούνται 124 δείγματα covid και 276 μη Covid (Σχήμα 1β) και αυτό γιατί καταργούνται οι εγγραφές των ίδιων χρηστών που ανεβαίνουν σε λιγότερο από 24 ώρες.

VI. Προεπεξεργασία Δεδομένων

Αναγνώριση Βήχα

Το Universal System for Cough Detection in Domestic Acoustic Environments [55],[114] χρησιμοποιήθηκε για την αυτόματη αναγνώριση των δειγμάτων βήχα που καταχωρήθηκαν στα ακατέργαστα αρχεία ήχου. Το Universal System for Cough Detection σε οικιακά ακουστικά περιβάλλοντα προσφέρει το πλεονέκτημα της ισχυρής σήμανσης των ηχητικών γεγονότων. Χρησιμοποιεί έναν ανιχνευτή ακουστικής έναρξης ως βήμα προεπεξεργασίας, με στόχο να ανιχνεύσει παρορμητικά μοτίβα στη ροή ήχου. Σε ένα επόμενο βήμα, η διάκριση των συμβάντων βήχα από άλλους παρορμητικούς ήχους αντιμετωπίζεται ως εργασία δυαδικής ταξινόμησης.



(α) (β)

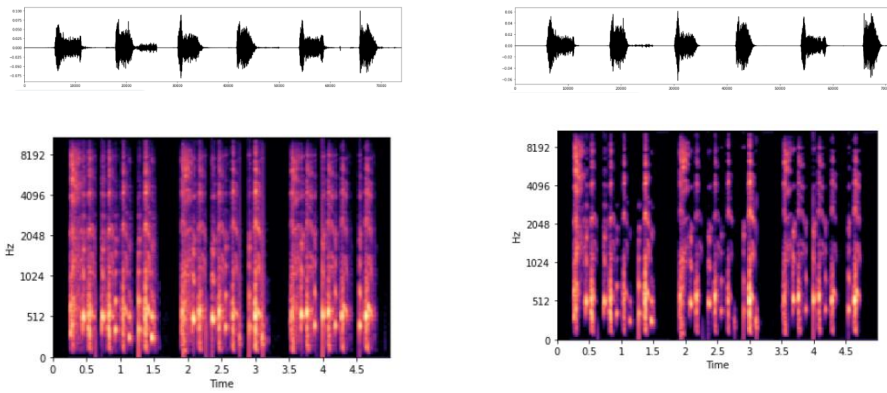
Σχημα 1: Κατανομή των δειγμάτων ανάλογα με την κατάσταση υγείας για το (α) Coswara (β) Cambridge σύνολο δεδομένων

Αφαίρεση θορύβου

Ένα σημαντικό πρόβλημα που προέρχεται από δεδομένα που προέρχονται από πλήθος είναι η έλλειψη ικανότητας ελέγχου των ήχων στο περιβάλλον και της ποιότητας του μικροφώνου. Για το σκοπό αυτό, χρησιμοποιείται μια βιβλιοθήκη που ονομάζεται Noisereduce[116]. Το Noisereduce είναι ένας αλγόριθμος μείωσης θορύβου σε python που μειώνει το θόρυβο σε σήματα στον τομέα του χρόνου όπως ομιλία, βιοακουστική και φυσιολογικά σήματα. Βασίζεται σε μια μέθοδο που ονομάζεται "φασματική πύλη" που είναι μια μορφή Θορύβου. Το Σχήμα 2 απεικονίζει τις κυματομορφές και τα φασματογράμματα του σήματος βήχα ασθενούς με Covid α) πριν και β) μετά τη μείωση του θορύβου.

Επαύξηση Δεδομένων

Η επαύξηση δεδομένων είναι υποχρεωτική για μικρά σύνολα δεδομένων όταν χρησιμοποιούνται συνελικτικά νευρωνικά δίκτυα επειδή αντιμετωπίζει τη σπανιότητα δεδομένων, αυξάνει την ευρωστία των μοντέλων, βελτιώνει την ακρίβεια των μοντέλων, μειώνει την υπερπροσαρμογή και εξοικονομεί πόρους για τη συλλογή και την επισήμανση δεδομένων.). Για τους σκοπούς αυτής της εργασίας χρησιμοποιώ δύο βιβλιοθήκες python, τη librosa[119] και τις audiomentations[120]. Μια βαρύτερη αύξηση δεδομένων έχει εφαρμοστεί στο σύνολο δεδομένων του Cambridge λόγω του γεγονότος ότι είναι μικρότερο και πιο επιρρεπές σε υπερπροσαρμογή από το Coswara heavy και το Coswara shallow.

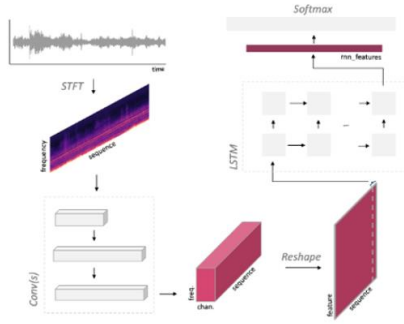


(α) (β)

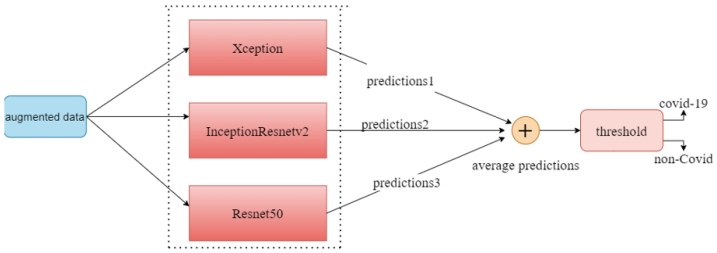
Σχήμα 2: Οι κυματομορφές και τα φασματογράμματα του σήματος βήχα ασθενούς με Covid α) πριν και β) μετά τη μείωση του θορύβου.

VII. *Αρχιτεκτονικές Βαθιάς Μάθησης που χρησιμοποιήθηκαν*

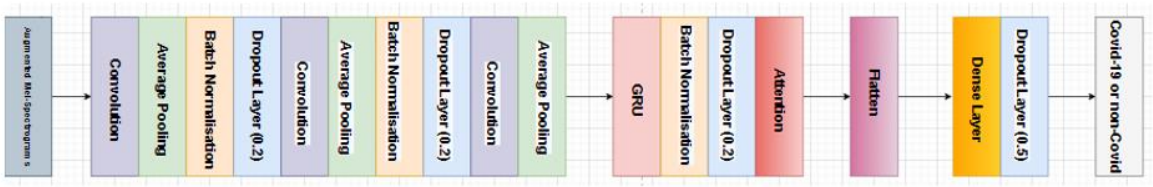
Εννέα μοντέλα με τα καλύτερα αποτελέσματα για την ταξινόμηση του Covid-19 χρησιμοποιήθηκαν. Τρία από αυτά είναι προεκπαιδευμένα νευρωνικά δίκτυα στο σύνολο δεδομένων ImageNet[122] και τα υπόλοιπα είναι τροποποιήσεις υπαρχόντων μοντέλων στη βιβλιογραφία καθώς και ένα στοιβαγμένο νευρωνικό δίκτυο, το οποίο συνδυάζει τις προβλέψεις των προαναφερθέντων προεκπαιδευμένων δικτύων. Τα τρία προεκπαιδευμένα δίκτυα είναι τα Xception, InceptionResnetv2 και Resnet50. Εξετάστηκαν επίσης και άλλα προεκπαιδευμένα δίκτυα, όπως τα Inceptionv3, VGG-16, EfficientNetB0, MobileNetv2, Densenet121, Resnet18 αλλά δεν πέτυχαν αξιόλογα αποτελέσματα. Τα υπόλοιπα μοντέλα είναι το VGG-13, ένα CNN σε συνδυασμό με Bi-LSTM, ένα CNN σε συνδυασμό με BiGRU, ένα συνελικτικό επαναλαμβανόμενο νευρωνικό δίκτυο και η παραλλαγή ενός DenseNet δικτύου που χρησιμοποιείται σε εφαρμογές αναγνώρισης ομιλίας, το DenseNet Speech. Οι αρχιτεκτονικές παρουσιάζονται στα Σχήματα 3 – 7.



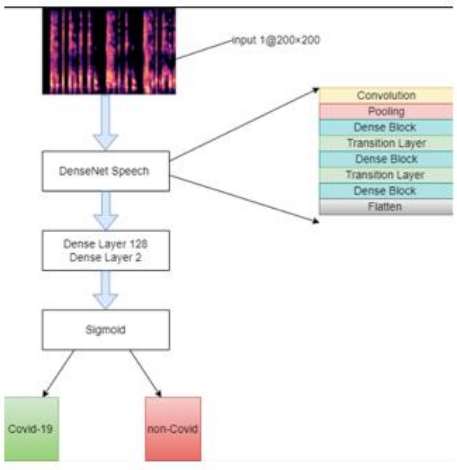
Σχήμα 3: TCRNN αρχιτεκτονική



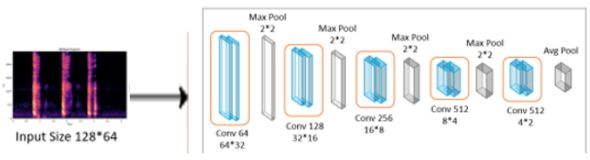
Σχήμα 4: Stacked CNN αρχιτεκτονική



Σχήμα 5: CNN-BiGRU αρχιτεκτονική



Σχήμα 6: DenseNet Speech αρχιτεκτονική



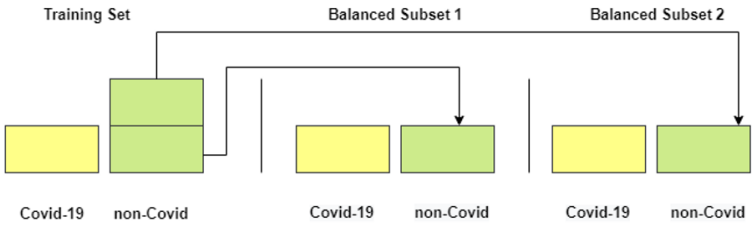
Σχήμα 7: VGG-13 αρχιτεκτονική

Διασταυρούμενη Επικύρωση

Το σύνολο δεδομένων χωρίζεται σε 5 πτυχές, με το καθένα να περιέχει το 20% του συνολικού συνόλου και, στη συνέχεια, το μοντέλο εκπαιδεύεται στις πτυχές k −1, ενώ μένει μία πτυχή για τη δοκιμή ενός μοντέλου. Αυτή η διαδικασία επαναλαμβάνεται 5 φορές. Η παραπάνω διαδικασία ακολουθείται για το σύνολο δεδομένων του Cambridge, αλλά για το σύνολο δεδομένων Coswara (ρηχός βήχας και βαρύς βήχας), ένα σύνολο δοκιμών (20% του συνόλου του συνόλου δεδομένων) διατηρείται από την αρχή και στη συνέχεια τα δεδομένα εκπαίδευσης χωρίζονται σύμφωνα με τις αρχές της διασταυρούμενης επικύρωσης. Και τα δύο σύνολα δεδομένων έχουν την ιδιαιτερότητα των ίδιων χρηστών (δηλαδή των χρηστών που συνδέονται με πολλαπλές εγγραφές). Η εκπαίδευση και οι δοκιμές στο ίδιο σύνολο χρηστών μπορούν να δώσουν τρομερά παραπλανητικά αποτελέσματα που δεν θα προβλέψουν την απόδοση του δείγματος σε νέους χρήστες. Η εκπαίδευση σε πολλαπλές εγγραφές/παρατηρήσεις από τον ίδιο χρήστη γίνεται αποδεκτή, αλλά τα δεδομένα δοκιμής πρέπει να είναι ανεξάρτητα από τα δεδομένα εκπαίδευσης. Για το λόγο αυτό επιστρατεύεται το GroupShuffleSplit κατά το διαχωρισμό σε σύνολα εκπαίδευσης, επικύρωσης, δοκιμής.

Ανισορροπία Κλάσεων

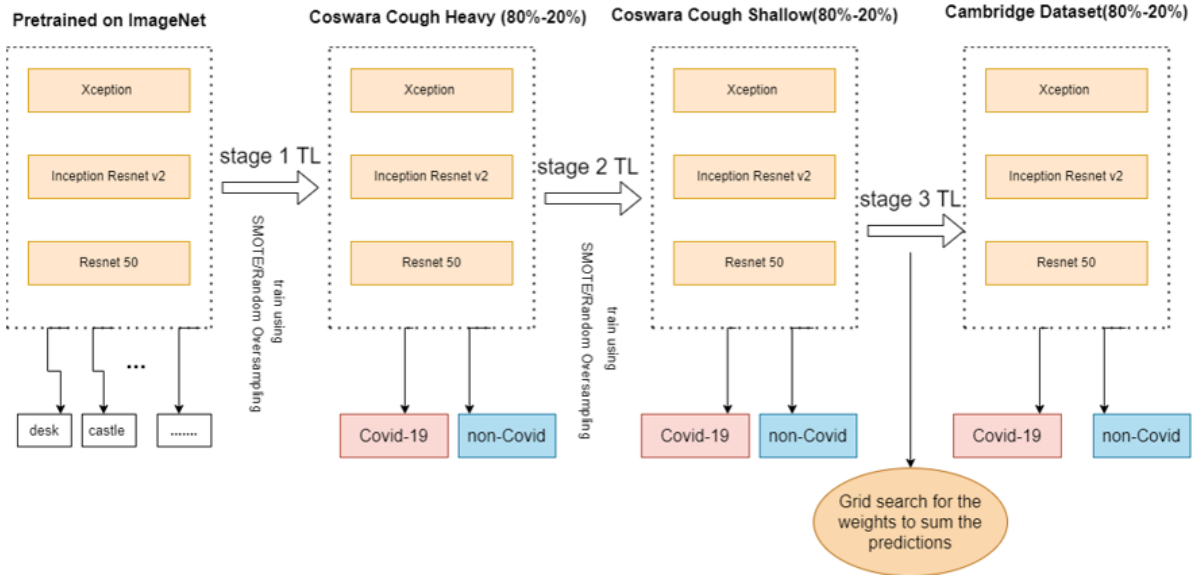
Στην ταξινόμηση κλινικών δεδομένων, ο μη ισορροπημένος αριθμός δειγμάτων δεδομένων, όπου τουλάχιστον μία από τις κατηγορίες αποτελεί μόνο μια πολύ μικρή μειοψηφία των δεδομένων, εμφανίζεται πολύ συχνά[138]. Στο σύνολο δεδομένων Cambridge και στα σύνολα δεδομένων Coswara για βαρύ βήχα και Coswara βήχα ρηχά, η κατηγορία Covid-19 υποεκπροσωπείται και η αναλογία μεταξύ ατόμων χωρίς Covid και Covid-19 είναι 2:1. Για την αντιμετώπιση αυτού του ζητήματος, εφαρμόστηκε μια μέθοδος εκμάθησης συνόλου. Πρώτον, το σύνολο δεδομένων χωρίζεται σε σύνολα εκπαίδευσης, επικύρωσης και δοκιμών. Το σετ εκπαίδευσης περιέχει το 60% των δεδομένων, ενώ η δοκιμή και η επικύρωση περιέχουν 20% το καθένα. Στη συνέχεια, η αύξηση δεδομένων εφαρμόζεται στο σετ εκπαίδευσης μόνο και για τις δύο κατηγορίες. Τα σετ επικύρωσης και δοκιμών παρέμειναν αμετάβλητα διατηρώντας την αρχική κατανομή των δειγμάτων στις δύο κατηγορίες (δηλαδή 2:1). Στη συνέχεια, υιοθετήθηκε μια ισορροπημένη προσέγγιση υποδειγματοληψίας, όπου δημιουργήθηκαν υποσύνολα εκπαίδευσης, διατηρώντας μια αναλογία 1:1 μεταξύ της τάξης της πλειοψηφίας (μη Covid) και της μειοψηφίας (Covid-19) [139]. Η τάξη μη Covid του σετ εκπαίδευσης χωρίστηκε σε δύο σετ ίσου μεγέθους και συγχωνεύτηκε με τα δείγματα ολόκληρης της μειονοτικής τάξης. Ως εκ τούτου, δημιουργήθηκαν δύο ισορροπημένα υποσύνολα που περιέχουν τα μισά από τα αρχικά δείγματα μη Covid και όλα τα δείγματα Covid-19, όπως φαίνεται στο Σχήμα 8.



Σχήμα 8: Μέθοδος εκμάθησης συνόλου

VIII. *Μάθηση Μεταφοράς Πολλαπλών Σταδίων*

Δύο από τις stacked CNN αρχιτεκτονικές που περιγράφηκαν προηγουμένως χρησιμοποιούνται με τις προβλέψεις συνόλου των ταξινομητών να υπολογίζονται κατά μέσο όρο. Εφαρμόζεται πενταπλάσια διασταυρούμενη επικύρωση προκειμένου να καθοριστεί η ακρίβεια, το precision, η ευαισθησία, η ειδικότητα, η AUC και το F1-score του μοντέλου. Τα αποτελέσματα στο μη επαυξημένο σύνολο δεδομένων του Cambridge επιβεβαίωσαν ότι η προεκπαίδευση σε δύο σύνολα δεδομένων που σχετίζονται με τον βήχα επιτυγχάνει υψηλότερα αποτελέσματα δοκιμών. Το Coswara cough heavy και το Coswara Cough Shallow εκπαιδεύονται χρησιμοποιώντας τόσο τυχαία υπερδειγματοληψία όσο και SMOTE. Ο στόχος του MSTL είναι να επωφεληθούμε από τη γνώση που αποκτάται μέσω της μάθησης σε διαφορετικά στάδια του TL. Η διαδικασία MSTL περιλαμβάνει Μεταφορά Μάθησης 3 σταδίων και τα τρία προεκπαιδευμένα μοντέλα είναι τα Xception, InceptionResnet-v2, ResNet50. Όλα τα μοντέλα εκπαιδεύτηκαν για 20 εποχές. Στο πρώτο στάδιο, τα βάρη που είναι προεκπαιδευμένα στο ImageNet φορτώνονται χρησιμοποιώντας το Keras. Στο δεύτερο στάδιο, τα βάρη αρχικοποιούνται σε αυτά που αποκτώνται με την εκπαίδευση τύπου δεδομένων Coswara για το βαρύ βήχα. Επιπλέον, όταν εκπαιδεύεται το Coswara Cough Shallow, εφαρμόζεται grid search προκειμένου να καθοριστούν τα βέλτιστα βάρη για την άθροιση των προβλέψεων κάθε ταξινομητή. Ένα σταθμισμένο σύνολο είναι μια επέκταση ενός συνόλου μέσου όρου μοντέλων όπου η συμβολή κάθε μέλους στην τελική πρόβλεψη σταθμίζεται από την απόδοση του μοντέλου. Αυτό οδήγησε στον ορισμό των ακόλουθων ως βαρών [0.2, 0.2, 0.6] για τα Xception, InceptionResnetv2, ResNet50 αντίστοιχα, πράγμα που σημαίνει ότι το ResNet50 συμβάλλει τα μέγιστα στο άθροισμα των προβλέψεων με συντελεστή 0.6. Στο τρίτο στάδιο, το σύνολο δεδομένων του Cambridge εκπαιδεύεται χρησιμοποιώντας τα βάρη που αποκτήθηκαν από την εκπαίδευση του Coswara για το ρηχό βήχα. Η διαδικασία φαίνεται στο Σχήμα 9.



Σχήμα 9: MSTL διαδικασία

IX. *Αποτελέσματα*

Το μοντέλο Stacked CNN είναι το μοντέλο συνόλου των τριών προεκπαιδευμένων δικτύων και έχει καλύτερη απόδοση από τα άλλα μοντέλα όταν εκπαιδεύεται στο σύνολο δεδομένων Coswara Cough Heavy. Το TCRNN, το οποίο εισήχθη αρχικά για την ταξινόμηση περιβαλλοντικού ήχου και τροποποιήθηκε για τις ανάγκες της τρέχουσας εργασίας, έχει συνεπή συμπεριφορά, καθώς είναι το καλύτερο μοντέλο για το σύνολο δεδομένων Coswara Cough Shallow και το δεύτερο καλύτερο για το Coswara Cough Shallow. Αυτό θα μπορούσε να αποδοθεί στο γεγονός ότι τα CRNN εκμεταλλεύονται τα

συνελικτικά επίπεδα έτσι ώστε να εξάγουν τοπικές πληροφορίες και τα επαναλαμβανόμενα επίπεδα για να τις συνδυάσουν σε ένα μεγαλύτερο χρονικό πλαίσιο. Το ResNet50 είναι το προεκπαιδευμένο δίκτυο στο ImageNet που επιτυγχάνει καλύτερα αποτελέσματα σε σύγκριση με τα υπόλοιπα προεκπαιδευμένα NN για τα σύνολα δεδομένων Coswara Cough Heavy και Coswara Cough Shallow, ενώ το InceptionResNetV2 ξεπερνά τα υπόλοιπα προεκπαιδευμένα δίκτυα για το σύνολο δεδομένων του Cambridge. Το υβριδικό CRNN με μηχανισμό που βασίζεται στην προσοχή επιτυγχάνει υψηλότερα αποτελέσματα όταν συνδυάζεται με ένα BiLSTM για το σύνολο δεδομένων Coswara Cough Heavy και με μια μονάδα BiGRU για τα άλλα δύο σύνολα δεδομένων. Τα αποτελέσματα που λαμβάνονται από το VGG13 είναι προφανώς καλύτερα για το σύνολο δεδομένων Coswara Cough Heavy από τα άλλα σύνολα δεδομένων. Γενικά, το Coswara Cough Heavy προσφέρει καλύτερα αποτελέσματα ταξινόμησης. Τα αποτελέσματα παρουσιάζονται στους πίνακες 1-4.

Πίνακας 1: Μετρικές αξιολόγησης για το σύνολο δεδομένων Coswara Cough heavy

Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Xception	65,84	61,82	55,31	48,26	75,13	51,38
InceptionResnetV2	69,24	65	52,1	53,14	76,86	52,6
Resnet50	69,14	65,17	54,62	52,85	77,5	53,72
Stacked CNN	**74,1**	**70,86**	**64,7**	**59,68**	**82,05**	**62,1**
TCRNN	71,1	66,4	66,1	54,8	78,5	59,9
VGG13	70,52	66,9	58,82	54,68	79,14	56,7
CRNN+Att+BiLSTM	69,15	65,08	53,78	52,89	77,28	53,33
CRNN+Att+BiGRU	65	63,81	67,22	47,62	80	55,75
DenseNet Speech	70,25	65,39	35,29	57,53	73,45	43,69

Πίνακας 2: Μετρικές αξιολόγησης για το σύνολο δεδομένων Coswara Cough Shallow

Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
XCeption	60,61	58,11	52,25	44,27	71,98	47,93
InceptionResnetV2	59,69	58,1	55,86	43,66	72,47	49,01
ResNet50	62,18	60,77	60,36	46,53	75	52,55
Stacked CNN	63,12	61,16	58,55	47,45	74,86	52,42
VGG13	56,87	59,74	72,73	42,78	76,69	53,87
TCRNN	**76,67**	**76,16**	**74,02**	**71,32**	**81**	**72,65**
CRNN+Att+BiLSTM	60,16	59,4	61,63	43,44	75,37	50,96
CRNN+Att+BiGRU	64,1	61,2	55,81	47,06	75,3	51,1
DenseNet Speech	63,67	64,34	73,25	47,37	81,3	57,53

X. Ερμηνευσιμότητα προβλέψεων

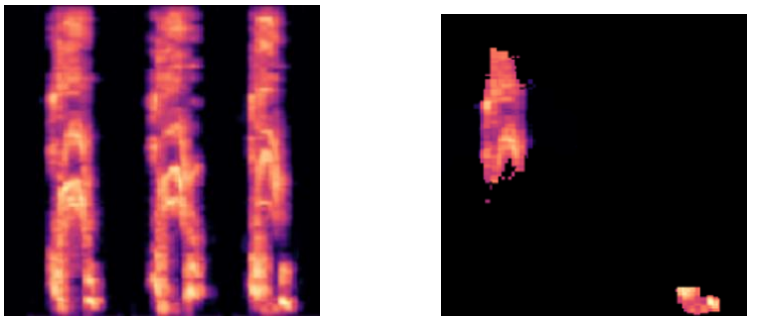
Ορισμένες από τις προβλέψεις στο σετ δοκιμών εξετάστηκαν για την ποιοτική αξιολόγηση της απόδοσης του InceptionResnetV2 για το σύνολο δεδομένων Coswara Cough Heavy. Για το σκοπό αυτό χρησιμοποιείται τοπική ερμηνεύσιμη μέθοδος αγνωστικών επεξηγήσεων μοντέλου (LIME). Στο Σχήμα 10 παρουσιάζονται παραδείγματα ερμηνείας για μια αληθή αρνητική (α) και μια ψευδώς αρνητική περίπτωση (β). Εδώ το αρνητικό σημαίνει χρήστες μη Covid. Το παραπάνω σχήμα δείχνει τι επιστρέφει το LIME ως εξήγηση στην πρόβλεψη ταξινόμησης εικόνων.

Πίνακας 3: Μετρικές αξιολόγησης για το σύνολο δεδομένων Cambridge

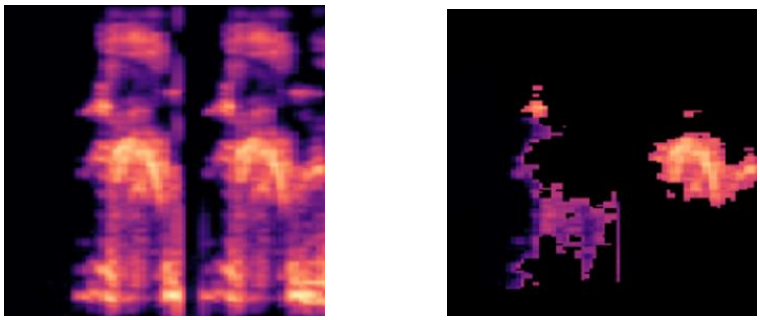
Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Xception	59,72	58,15	59,09	39,39	76,92	47,27
InceptionResNetV2	62,5	60,95	63,63	42,42	79,49	50,9
ResNet50	57	59,83	72,73	39,02	80,64	50,79
Stacked CNN	62,32	62,12	66,7	48,48	75,75	56,15
VGG13	62,5	61,97	68,18	42,86	81,08	52,63
TCRNN	70,2	52,2	57,93	54,4	72,3	56,1
CRNN+Att+BiLSTM	64	63,86	72,72	44,44	83,33	55,14
CRNN+Att+BiGRU	**62,5**	**64,22**	**77,27**	**43,59**	**84,85**	**55,74**
DenseNet Speech	59,82	59,18	63,63	40	78,37	49,12

Πίνακας 4: Μετρικές Αξιολόγησης για το Cambridge dataset κατά το multistage transfer learning

Feature	Imblance Handling	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Mel Spectrogram	SMOTE	67,87	65,9	63	54,49	77,3	58,44
	Random Oversampling	**69,34**	**70,2**	**78,03**	**56,3**	**81,56**	**65,4**



(a)



(b)

Σχήμα 10: (Αριστερά) το αρχικό mel spectrogram και (δεξιά) η εξήγηση του LIME για (α) αληθώς αρνητικό, (β) ψευδώς αρνητικό δείγμα.

Το Σχήμα 10 (α) δείχνει την περιοχή της εικόνας (super-pixel) που έχουν ισχυρότερη σχέση με την πρόβλεψη "non Covid", ενώ το 10(b) δείχνει τα super-pixels που έχουν ισχυρότερη σχέση με την κατηγορία "non Covid" αλλά η εικόνα ήταν εσφαλμένη ταξινόμηση. Η έξοδος του LIME είναι μια λίστα

επεξηγήσεων, που αντικατοπτρίζει τη συμβολή κάθε χαρακτηριστικού στην πρόβλεψη ενός δείγματος δεδομένων. Αυτό παρέχει τοπική ερμηνευσιμότητα και επιτρέπει επίσης να προσδιοριστούν ποιες αλλαγές χαρακτηριστικών θα έχουν τον μεγαλύτερο αντίκτυπο στην πρόβλεψη. Από το παραπάνω παράδειγμα προσδιορίζουμε ότι τα υψηλότερα ντεσιμπέλ παίζουν πιο σημαντικό ρόλο στην ταξινόμηση ενός mel spectrogram ως μη Covid, καθώς το mel spectrogram ταξινομήθηκε σωστά. Το ψευδώς αρνητικό παράδειγμα λαμβάνει υπόψη χαμηλότερη ένταση ήχου (dB), πράγμα που σημαίνει ότι αυτά τα χαρακτηριστικά έχουν μικρότερο αντίκτυπο.

XI. Συμπεράσματα – Μελλοντική Έρευνα

Ο τρόπος με τον οποίο η COVID-19 επηρεάζει το αναπνευστικό σύστημα είναι ουσιαστικά μοναδικός και ως εκ τούτου, ο βήχας που σχετίζεται με αυτήν είναι πιθανό να έχει επίσης μοναδικά λανθάνοντα χαρακτηριστικά. Συνοπτικά, στην παρούσα διπλωματική εργασία, παρουσιάζεται μία μέθοδος για την επεξεργασία των ηχογραφήσεων, την ανίχνευση του βήχα, την εξαγωγή φασματογραμμάτων mel και την ταξινόμηση των δειγμάτων σε COVID-19 ή non COVID-19. Έχουν χρησιμοποιηθεί τρία σύνολα δεδομένων, το συνόλο δεδομένων Cambridge και δύο ακόμη που προέρχονται από το σύνολο δεδομένων Coswara. Μία από τις κύριες προκλήσεις ήταν η υπερπροσαρμογή (overfitting) η οποία αντιμετωπίστηκε με διάφορες τεχνικές, όπως η L2 κανονικοποίηση και η διερεύνηση ρηχότερων αρχιτεκτονικών, για παράδειγμα VGG13 αντί για VGG16 ή VGG19, DenseNet Speech αντί για DenseNet201. Μια άλλη πρόκληση σχετίζεται με την ανισορροπία κλάσεων. Διερευνήθηκαν διαφορετικές προσεγγίσεις για την ταξινόμηση των δειγμάτων σε COVID-19 ή non COVID-19. Για το σκοπό αυτό, δοκιμάστηκαν εννέα διαφορετικές αρχιτεκτονικές βαθιάς μάθησης. Ορισμένες από αυτές περιλαμβάνουν υβριδικές τεχνικές εκμάθησης που συνδυάζουν CNN και BiLSTM ή BiGRU. Ένα μοντέλο συλλογικής μάθησης που αποτελείται από τρία προεκπαιδευμένα μοντέλα στο ImageNet πέτυχε ικανοποιητικά αποτελέσματα για το σύνολο δεδομένων Coswara Cough Heavy, και συγκεκριμένα ακρίβεια 74,1%, AUC 70,86%, Precision 64,7%, Recall 59,68%, Specificity 82,05% και F1-score 62,1%. Το TCRNN ξεπέρασε τις επιδόσεις των υπόλοιπων αρχιτεκτονικών σε δύο σύνολα δεδομένων, το σύνολο δεδομένων Cambridge και το Coswara Cough Shallow, στο οποίο πέτυχε ακρίβεια 76,67%, AUC 76,16%, Precision 74,02%, Recall 71,32%, Specificity 81% και F1-score 72,65%. Το MSTL αξιοποιεί όλα τα διαθέσιμα σύνολα δεδομένων προκειμένου να επωφεληθεί από τη γνώση που αποκτήθηκε μέσω της μάθησης σε διαφορετικά στάδια της διαδικασίας Transfer Learning. Συνδυάζεται με την εκμάθηση συνόλου για ανισορροπία τάξης για το σύνολο δεδομένων του Cambridge. Μετά από αυτή τη διαδικασία, οι μετρήσεις αξιολόγησης και ειδικά η AUC, η ακρίβεια και το F1-score έδειξαν αξιοσημείωτη βελτίωση. Συμπεραίνεται ότι η προεκπαίδευση σε δύο συναφή με την εργασία ταξινόμησης σύνολα δεδομένων προσφέρει καλύτερη αρχικοποίηση των βαρών του μοντέλου και άρα μαθαίνει πιο αποτελεσματικά τα χαρακτηριστικά στο τρίτο σύνολο δεδομένων.

Η μελλοντική έρευνα θα μπορούσε να περιλαμβάνει άλλες διαθέσιμες φωνητικές λειτουργίες, όπως η αναπνοή και η ομιλία, πέρα από τον βήχα που αναλύεται σε αυτή την εργασία. Οι προκλήσεις που σχετίζονται με την αποσαφήνιση με άλλες παθολογίες του αναπνευστικού με παρόμοια συμπτώματα παραμένουν προς αντιμετώπιση. Δεδομένου ότι οι αρχιτεκτονικές βαθιάς μάθησης μπορούν να εξάγουν πολλαπλά χαρακτηριστικά, η συνένωση χαρακτηριστικών θα μπορούσε να είναι ένας αποτελεσματικός τρόπος για να βελτιωθεί η διαδικασία ταξινόμησης. Επιπλέον, θα μπορούσαν να χρησιμοποιηθούν βιοδείκτες ως είσοδοι, μαζί με φασματογράμματα, MFCC ή εικόνες, σε παράλληλες αρχιτεκτονικές. Οι βιοδείκτες θα μπορούσαν να περιέχουν επιπλέον χαρακτηριστικά σχετικά με τα συμπτώματα της COVID-19.

Contents

List Of Figures

Chapter 1

Introduction

1.1 Covid-19

COVID-19 (COrona VIrus Disease of 2019), caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2) virus, was declared a global pandemic on February 11, 2020 by the World Health Organization (WHO). The Severe Acute Respiratory Syndrome previously was pandemic in 2003[11]. Research evidence suggests that SARS-CoV and MERS-CoV originated in bats. SARS-CoV then spread from infected civets to people. To date, the origin of SARS-CoV-2 which caused the COVID-19 pandemic has not been identified. The scientific evidence thus far suggests that SARS-CoV-2 likely resulted from viral evolution in nature and jumped to people or through some unidentified animal host [12]. Reports trace the outbreak back to a massive market that sold live animals, among other goods, in Wuhan, China , and a third suggests that the coronavirus SARS-CoV-2 spilled over from animals — possibly those sold at the market — to humans at least twice in November or December 2019 [13].At the moment of writing, there were 569.771.691 active cases of COVID-19 globally, and there had been 6.383.776 deaths, with the USA reporting the highest number of cases (90.390.184) and deaths (1.026.937)[14]. According to the World Health Organization, in Greece 4,21 million cases have been confirmed and 30,707 deaths have been reported[15].

Coronaviruses are important pathogens that can affect the lower respiratory tract in humans and can cause diseases ranging from a simple cold to severe infection with up to 50% lethality[16]. COVID-19 seems not to be very different from SARS regarding its clinical features. However, it has a fatality rate of 1.1%, lower than that of SARS (9.5%) and much lower than that of MERS (34.4%) [17], but it can differ in people who have comorbidities [18].

The most common symptoms in COVID-19 patients include fever, cough, fatigue, dyspnea and the sputum [19]. According to WHO, loss of taste or smell is also a very common symptom while less common symptoms are sore throat, headache, aches and pains, diarrhea, a rash on skin, or discoloration of fingers or toes, red or irritated eyes. Diarrhea is more common in SARS [17]. Among serious symptoms are difficulty in breathing or shortness of breath, loss of speech or mobility, or confusion and chest pain.

The reproductive number (R_0) of the novel infection is estimated by the World Health Organization (WHO) to range between 2 and 2.5, which is higher than that for SARS (1.7-1.9) and MERS (<1), suggesting that Covid-19 has a higher pandemic potential [17]. Due to incorporation of more individual case information and travel data, the estimate for R_0 in Wuhan was revised upward from 2.2–2.7 to 5.7 [20]. The virus can spread from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak, sing or breathe.

From the beginning of its appearance, Covid-19 has mutated several times, which is a common behavior of viruses. Most changes have little to no impact on the virus' properties. However, some changes may affect the virus's properties, such as how easily it spreads, the associated disease severity, or the

performance of vaccines, therapeutic medicines, diagnostic tools, or other public health and social measures[21].

1.1.1 Demographics

As stated earlier, at the moment there are 569.771.691 active cases of COVID-19 globally, and there had been 6.383.776 deaths, with the USA reporting the highest number of cases (90.390.184) and deaths (1.026.937) [14]. The overall reported number of Covid-19 cases from the beginning of the pandemics until July 24 2022 is depicted in Figure 1.1. Johns Hopkins University has created and daily updates an open data repository with international analytics on the SARS-CoV-2 pandemic. Figure 1.2 shows the daily number of cases in Greece for the last two years. The abatement of mask use measures and the increasing tourist flows which results in overcrowding, have led to a subsequent outbreak of daily cases on the summer months of 2022.

Figure 1.1: Heat map showing the number of cases across the world since the beginning of Covid-19 pandemics

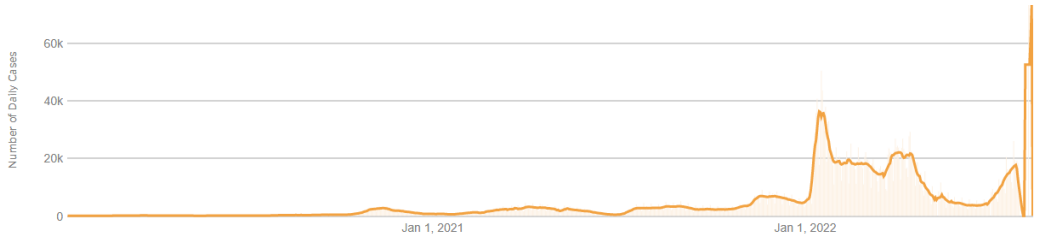


Figure 1.2:Number of daily cases in Greece since the beginning of Covid-19 pandemics

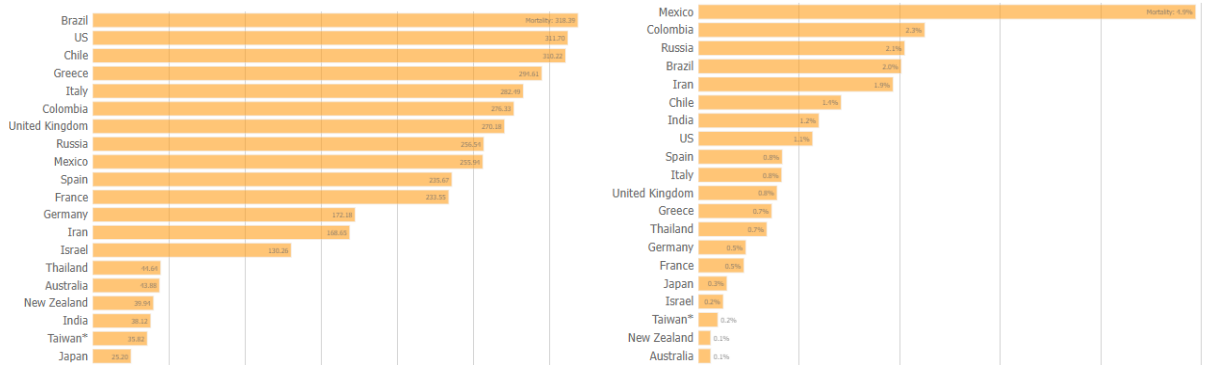
1.1.2 Fatality Rates

Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are more likely to develop serious illness[15]. The case fatality rate (CFR) of COVID-19 is reported to be 1.1% but it can differ in patients who have other pre-

existing conditions and it differs across countries, as well. In some patients, especially those with other underlying diseases, there may be a respiratory failure, arrhythmias, shock, kidney failure, cardiovascular damage, or liver failure [22]. One of the most important ways to measure the burden of COVID-19 is mortality. Countries throughout the world have reported very different case fatality ratios (i.e. the number of deaths divided by the number of confirmed cases). Differences in mortality numbers can be caused by:

- Differences in the number of people tested: With more testing, more people with milder cases are identified. This lowers the case fatality ratio.
- Demographics: For example, mortality tends to be higher in older populations.
- Characteristics of the healthcare system: For example, mortality may rise as hospitals become overwhelmed and have fewer resources.

For the twenty countries currently most affected by COVID-19 worldwide, the bars in the chart below show the number of deaths. Figure 1.3 (b) depicts the number of deaths per 100 confirmed cases (i.e. observed case-fatality ratio) and Figure 1.3 (a) the number of deaths per 100,000 population (this represents a country's general population, with both confirmed cases and healthy people). Countries at the top of this figure have the most deaths proportionally to their COVID-19 cases or population, not necessarily the most deaths overall. Greece holds the 4th position in Figure 1.3 (a) which means that the overall number of deaths is big in proportion to the general population since the CFR is equal to 0.7, probably because of the high number of daily tests. Figure 1.4 shows the number of daily deaths in Greece, which is currently in recession despite the outbreak of cases. Figure 1.5 details the number of deaths by age in United States as of July 13, 2022 [23].



(a) (b)

Figure 1.3:Number of deaths (a) per 100 confirmed cases (CFR) (b) per 100,000 population [4]

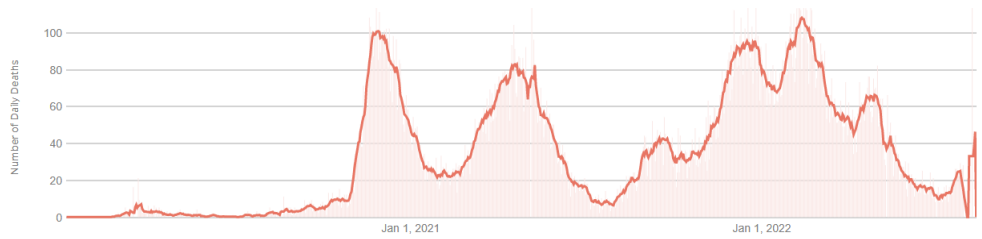


Figure 1.4: Number of daily deaths in Greece since the beginning of the pandemics[4]

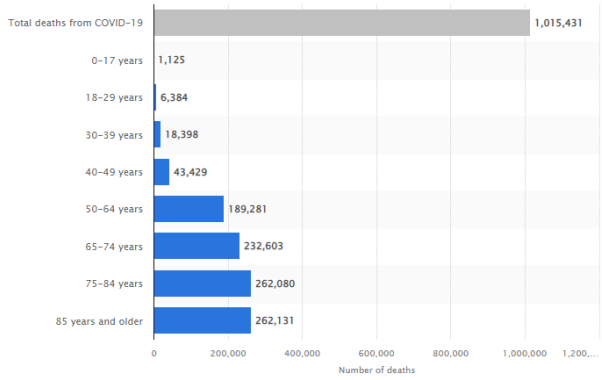


Figure 1.5:Number of Covid-19 deaths in the U.S. by age as of July 13, 2022[13].

1.1.3 SARS-CoV-2 variants

Although most mutations in the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome are expected to be either deleterious and swiftly purged or relatively neutral, a small proportion will affect functional properties and may alter infectivity, disease severity or interactions with host immunity [24]. There are three categories of variants according to the European Centre for Disease prevention and Control:

- Variants of concern. For these variants clear evidence is available indicating a significant impact on transmissibility, severity and/or immunity that is likely to have an impact on the epidemiological situation.
- Variants of interest. For these variants, evidence is available on genomic properties, epidemiological evidence or in-vitro evidence that could imply a significant impact on transmissibility, severity and/or immunity, realistically having an impact on the epidemiological situation. However, the evidence is still preliminary or is associated with major uncertainty.
- Variants under monitoring. These additional variants of SARS-CoV-2 have been detected as signals through epidemic intelligence, rules-based genomic variant screening, or preliminary scientific evidence. There is some indication that they could have properties similar to those of a variant of concern, but the evidence is weak or has not yet been assessed by European Centre for Disease prevention and Control (ECDC).

Table 1 shows the variants of concern and table 2 shows the variants under monitoring. With regards to the titles of each column, these are conventions that have been made for describing mutations. As of 31st May 2021, WHO proposed labels for global SARS-CoV-2 variants of concern and variants of interest to be used alongside the scientific nomenclature in communications about variants to the public [25]. Lineage and additional mutations αρε the variant designation specified by one or more PANGO lineages and any additional characteristic spike protein changes. The Phylogenetic Assignment of Named Global Outbreak Lineages (PANGOLIN) is a software tool developed by Dr. Áine O'Toole in order to implement a dynamic nomenclature (known as the PANGO nomenclature) to classify genetic lineages for SARS-CoV-2 [26]. Country first detected is only present if there is enough evidence linking the mutation with the first country if detection. Year and month first detected as reported in the GISAID EpiCoV database. Transmission in the EU/EEA is categorised as dominant, community, outbreak(s), and sporadic/travel. Evidence is given on three different areas, transmissibility, immunity and infection. Each category is described as:

- increased or reduced, if there is enough evidence that the variant is different enough from previous variants and hence will have an impact on the epidemiological situation.
- Similar if there is enough evidence that the variant is similar to previous circulating variants

- Unclear if the evidence is incomplete or contradictory
- No evidence

WHO LABEL	LINEAGE	COUNTRY FIRST DETECTED	YEAR AND MONTH FIRST DETECTED	TRANSMISSIBILITY IMPACT	IMMUNITY IMPACT	SEVERITY IMPACT	TRANSMISSIO N IN EU
Omicron	BA.1	South Africa and Botswana	November 2021	increased	increased	Reduced	Community
Omicron	BA.2	South Africa	November 2021	increased	increased	Reduced	Dominant
Omicron	BA.3	South Africa	January 2022	No evidence	increased	No evidence	Community
Omicron	BA.4	South Africa	February 2022	No evidence	increased	No evidence	Community

Table 1: Variants of Concern at the time of writing

WHO LABEL	LINEAGE	COUNTRY FIRST DETECTED	YEAR AND MONTH FIRST DETECTED	TRANSMISSIBILITY IMPACT	IMMUNITY IMPACT	SEVERITY IMPACT	TRANSMISSION IN EU
Omicron	BA.3	South Africa	November 2021	No evidence	No evidence	No evidence	detected

Table 2: Variants under Monitoring at the time of writing

It is worth noting that at the time of writing of this thesis, the omicron BA.2 variant is dominant in Europe.

1.1.4 Infection Prevention and Vaccination

Public health and social measures (PHSMs) are being implemented across the world to suppress SARS-CoV-2 transmission and reduce mortality and morbidity from COVID-19. PHSMs include personal protective measures (e.g. physical distancing, avoiding crowded settings, hand hygiene, respiratory etiquette, mask-wearing); environmental measures (e.g. cleaning, disinfection, ventilation); surveillance and response measures (e.g. testing, genetic sequencing, contact tracing, isolation, and quarantine); physical distancing measures (e.g. regulating the number and flow of people attending gatherings, maintaining distance in public or workplaces, domestic movement restrictions); and international travel-related measures[27]. The guidelines of WHO to prevent the contamination from SARS-CoV-2 are:

- Keep physical distance of at least 1 metre from others, even if they don't appear to be sick. Avoid crowds and close contact.
- Wear a properly fitted mask when physical distancing is not possible and in poorly ventilated settings.
- Hands cleaning frequently with alcohol-based hand rub or soap and water.
- Cover your mouth and nose with a bent elbow or tissue when you cough or sneeze. Dispose of used tissues immediately and clean hands regularly.
- Self-isolation if symptoms of the infection are developed.

- Vaccination

Due to the COVID-19 pandemic, a number of non-pharmaceutical interventions colloquially known as lockdowns (encompassing stay-at-home orders, curfews, quarantines, cordons sanitaires and similar societal restrictions) have been implemented in numerous countries and territories around the world.

So far, the most prevailing measure against the spread of Covid-19 is vaccination. It is a safer and more reliable way to build protection than getting sick with COVID-19. COVID-19 vaccination helps protect by creating an antibody response without having to experience potentially severe illness or post-COVID conditions. The first covid-19 vaccines were administered under emergency use authorisation in December 2020, just one year into the pandemic, a "miracle" of pharmaceutical innovation that has saved an estimated million lives or more in the US alone. Immunization with Pfizer-BioNTech and Moderna mRNA vaccines protected a remarkably high percentage (>90%) of recipients from developing symptomatic infection and, to a lesser extent, from asymptomatic infection too. During the first half of 2021, when the alpha variant of SARS-CoV-2 was dominant, the covid-19 mortality rate was reduced by 60%, 75%, and 81% in counties with low, medium, and high vaccination coverage, compared with counties that had very low coverage [28]. In May 2021, the United States Food and Drug Administration and the European Medicines Agency (EMA) authorised the use of the Pfizer–BioNTech vaccine, Comirnaty, for children aged 12–15 years. On 25 November 2021, the EMA extended that authorisation to children aged 5 - 11 years. The vaccines that are authorized for use in the European Union are Comirnaty (Pfizer-BioNTech), Jcovden (previously COVID-19 Vaccine Janssen), Nuvaxovid (Novanax), Spikevax (Moderna) and Vaxzevria (AstraZeneca). Comirnaty contains tozinameran, a messenger RNA (mRNA) molecule with instructions for producing a protein from SARS-CoV-2 and it is given as two injections 3 weeks apart. Jcovden is made up of another virus (of the adenovirus family) that has been modified to contain the gene for making a protein found on SARS-CoV-2. A booster dose has been given at least 2 months after the first dose of Jcovden in people aged 18 years and older. Nuvaxovid contains a version of a protein found on the surface of SARS-CoV-2 which has been produced in the laboratory and it is given as two injections 3 weeks apart. Spikevax contains elasomeran, a messenger RNA (mRNA) molecule with instructions for producing a protein from SARS-CoV-2 and it is given as two injections 28 days apart. Finally, Vaxzevria is made up of another virus (of the adenovirus family) that has been modified to contain the gene for making a protein from SARS-CoV-2. Vaxzevria is given as two injections, usually into the muscle of the upper arm. The second dose should be given between 4 and 12 weeks after the first dose. Booster doses have been delivered at least 3 months after the second dose to people aged 12 years and older. A fourth booster dose has been taken in Greece for vulnerable groups with comorbidities. Public health officials were concerned since the start of the pandemic that vaccinations would not be equitably distributed around the world. The data appears to be confirming those fears as developed nations are vaccinating their populations far faster than less developed countries[14]. Figure 1.6 shows a heatmap of the percentage of people fully vaccinated around the world. In Greece, 73.2 % of people are fully vaccinated with 21.23 million doses. The alarming low rate of COVID-19 vaccination in Africa has made the continent trail behind in the vaccination campaign, thereby putting the global vaccination progress under threat [29]. This might be the reason why current variants of concern and variants under monitoring were first detected in South Africa (Table 1, 2).

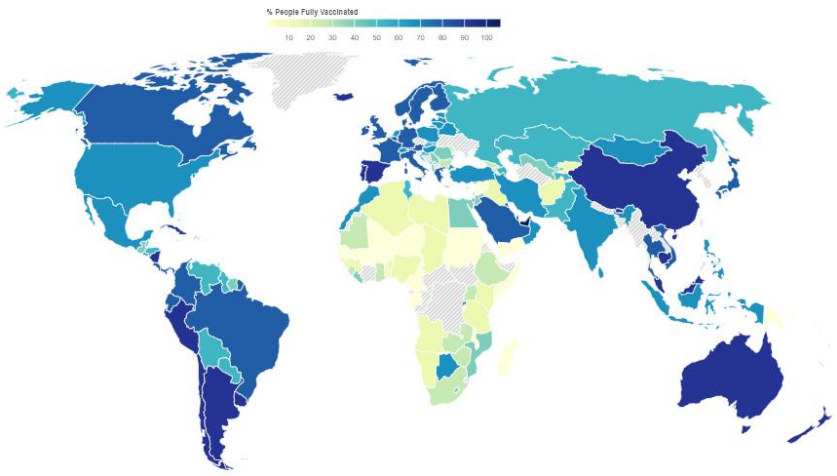


Figure 1.6:Heat map showing the percentage of vaccinations around the world [4].

1.1.5 Treatments

The U.S. Food and Drugs Administration (FDA) has approved the antiviral drug Veklury (remdesivir) for adults and certain pediatric patients with COVID-19. This is an intravenous therapy. The FDA has also approved the immune modulator Olumiant (baricitinib) for certain hospitalized adults with COVID-19. During public health emergencies, the FDA may authorize the use of unapproved drugs or unapproved uses of approved drugs under certain conditions. This is called an Emergency Use Authorization (EUA). The FDA has issued EUAs for several monoclonal antibody treatments, for COVID-19 for the treatment, and in some cases prevention (prophylaxis), of COVID-19 in adults and pediatric patients. Monoclonal antibodies are laboratory-made molecules that act as substitute antibodies. There are also two oral antiviral pills, Paxlovid and Lagevrio (molnupiravir), authorized for patients with mild-to-moderate COVID-19 [30]. Molnupiravir is the first oral, direct-acting antiviral shown to be highly effective at reducing nasopharyngeal SARS-CoV-2 infectious virus and viral RNA and has a favorable safety and tolerability profile [31]. The European Medicines Agency (EMA) has approved for use in the European Union the following medications: Evusheld(tixagevimab / cilgavimab), Kineret (anakinra), Paxlovid (PF-07321332 / ritonavir), Regkirona (regdanvimab), RoActemra(tocilizumab), Ronapreve (casirivimab / imdevimab), Veklury (remdesivir), Xevudy (sotrovimab). EMA is currently evaluating marketing authorization applications for Olumiant and Lagevrio [32]. Unfortunately, changing variants of the virus affect the efficacy of treatments which may be withdrawn if proven ineffective against a certain variant. Table 3 analyses some of the most common antiviral treatments as described from Centers for Disease Control and Prevention [33].

TREATMENT	WHO CAN TAKE IT	WHEN IT SHOULD BE TAKEN	HOW
Paxlovid	Adults, children ages 12 years or older	Within 5 days of when symptoms start	Orally
Veklury	Adults and children	Within 7 days of when symptoms start	Intravenous
Bebtelovimab (monoclonal antibody)	Adults, children ages 12 years or older	Within 7 days of when symptoms start	Single intravenous injection
Lagevrio	Adults	Within 5 days of when symptoms start	Orally

Table 3: Treatments of Covid19

1.1.6 Transmissibility

The basic reproduction number (R_0) is a well-known epidemiological concept to measure the spread of an infectious disease. Some published studies have estimated an R_0 for SARS reaching the value of 4. Interestingly, a recent review by Liu and colleagues [34] has shown that the average reproductive number of SARS-CoV-2 is estimated to be 3.28, with a median value of 2.79, thus exceeding the WHO estimates [17]. Two modes of transmission exist - direct and indirect. The direct mode includes transmission via aerosols, tears, saliva, semen, and mother-to-child. Indirect modes include transmission via fomites [35]. Authors conclude that transmission from mother-to-child may be rare, but not completely absent. Further data is needed to find the details on this mode of transmission. Transmission from mother-to-child can be prevented by delivering the neonates in negative pressure isolation rooms. Figure 1.1 demonstrates the dominant transmission modes. According to the World Health Organization (WHO), the SARS-CoV-2 is spread between people in several ways. Current evidence suggests that the virus spreads mainly between people who are in close contact with each other, for example at a conversational distance. The virus can spread from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak, sing or breathe. Another person can then contract the virus when infectious particles that pass through the air are inhaled at short range (this is often called short-range aerosol or short-range airborne transmission) or if infectious particles come into direct contact with the eyes, nose, or mouth (droplet transmission). The virus can also spread in poorly ventilated aor crowded indoor settings, where people tend to spend longer periods of time. This is because aerosols can remain suspended in the air or travel farther than conversational distance (this is often called long-range aerosol or long-range airborne transmission). Finally, people may also become infected when touching their eyes, nose or mouth after touching surfaces or objects that have been contaminated by the virus.

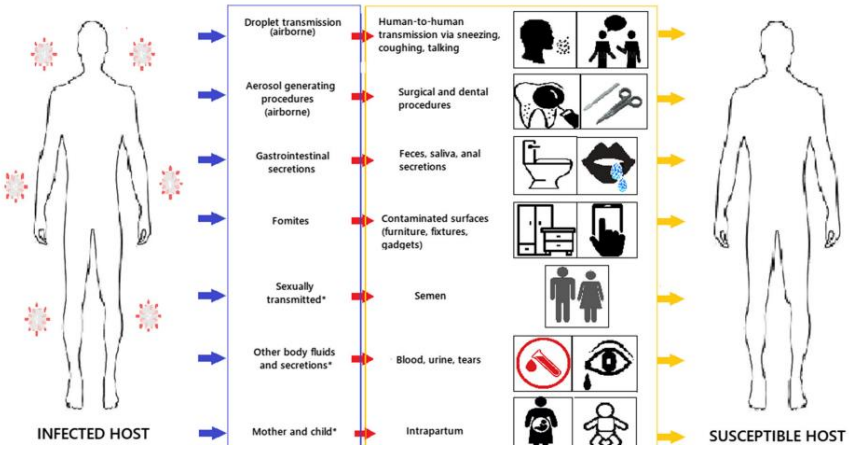


Figure 1.7:Transmission modes of Covid-19 infection

1.1.7 Testing Methods

More than six billion tests for COVID-19 has been already performed in the world. The testing for SARS-CoV-2 virus and corresponding human antibodies is essential not only for diagnostics and treatment of the infection by medical institutions, but also as a pre-requisite for major semi-normal economic and social activities [36]. Techniques of viral detection include detection of viral particles (virions), viral antigen, antibodies to the virus, and viral nucleic acid. There are three categories of testing methods: diagnostic PCR tests, diagnostic antigen tests and antibody tests. Major methods of detection of SARS-CoV-2 virus is based on detection of viral RNA. PCR is one of the common techniques used to detect viral nucleic acid. The test is typically performed on a nasal swab or saliva sample. The test uses a technology known as polymerase chain reaction (PCR) to detect trace amounts of genetic material of SARS-CoV-2. Once a swab is taken, its viral RNA is isolated from the sample and then converted into a complimentary strand of DNA. Then using the PCR technique, the DNA is multiplied to create thousands of copies, allowing a large enough sample to test for SARS-CoV-2 genes. Detection of viral particles and antigen is a viable alternative to RT-PCR [37]. These methods are potentially inexpensive, portable, rapid, and can be used to diagnose patients at the early stage of viral infection. They are not required to be performed by a skilled operator and can be run by patients themselves. The test is typically performed on a nasal or throat swab sample. It detects fragments of specific viral proteins. After a swab is collected, the sample is mixed with a liquid and then placed on a testing strip. As the sample flows down the test strip, SARS-CoV-2 antibodies in the test can recognize and bind to viral protein fragments, if present. This protein fragment-antibody complex appears as a visible, colored line. Antibody or serology tests can find whether a person likely had a previous SARS-CoV-2 infection. This blood test does not diagnose an active infection or provide information about long-term immunity. SARS-CoV-2 specific IgG, IgM, and IgA antibodies most often become objects of detection using different methods. IgM antibodies appear in the acute phase of infection, and after reaching the maximum, they decrease to diagnostically insignificant levels. IgG antibodies build up more slowly than IgM antibodies, but they remain high in the patient's blood longer. After recovery, IgG antibodies can remain at a low level indefinitely as evidence of a previous illness.

1.1.8 Symptoms and Long-Covid

According to Russel M Viner et al. fever and cough were the most common symptoms; proportions with fever ranged from 46% to 64.2% and with cough from 32% to 55.9%. All other symptoms or signs including rhinorrhoea, sore throat, headache, fatigue/myalgia and gastrointestinal symptoms including diarrhoea and vomiting were infrequent, occurring in less than 10%–20% [38]. The symptoms of Covid-19 range as new variants prevail. Moreover, WHO has included as very common symptoms, other than fever and cough, fatigue, dyspnea and loss of taste or smell. CDC urges to seek emergency medical attention if trouble breathing, persistent pain or pressure in the chest, confusion, inability to stay awake or wake, pale, gray, or blue-colored skin, lips, or nail beds, depending on skin tone are some of the symptoms.

COVID-19 can involve persistence, sequelae, and other medical complications that last weeks to months after initial recovery. According to Sandra Lopez Lion et al. it was estimated that 80% of the infected patients with SARS-CoV-2 developed one or more long-term symptoms. Long Covid refers to when people continue to experience symptoms of COVID-19 and do not fully recover for several weeks or

months after the start of their symptoms. The five most common symptoms were fatigue (58%), headache (44%), attention disorder (27%), hair loss (25%), and dyspnea (24%). Other symptoms were related to lung disease (cough, chest discomfort, reduced pulmonary diffusing capacity, sleep apnea, and pulmonary fibrosis), cardiovascular (arrhythmias, myocarditis), neurological (dementia, depression, anxiety, attention disorder, obsessive– compulsive disorders), and others were unspecific such as hair loss, tinnitus, and night sweat. A couple of studies reported that fatigue was more common in females, and one study reported that post-activity polypnea and alopecia were more common in females. They identified a total of 55 long-term effects associated with COVID-19 in the literature reviewed. Most of the effects correspond to clinical symptoms such as fatigue, headache, joint pain, anosmia, ageusia, etc. In addition, diseases such as stroke and diabetes mellitus were also present [39]. Figure 1.8 presents all the symptoms of long covid according to the above research.

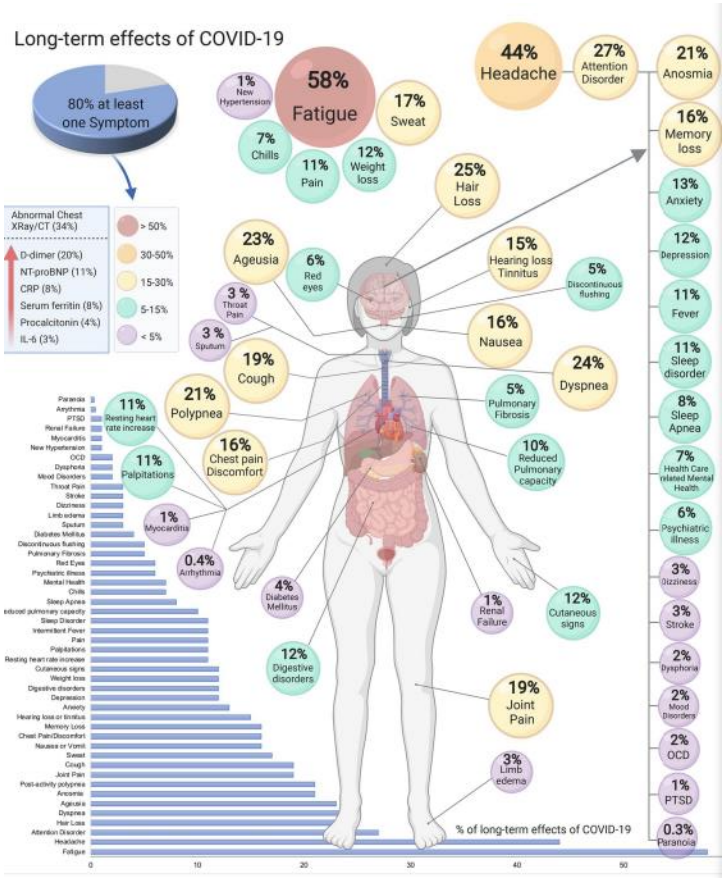


Figure 1.8:Long-term effects of coronavirus disease 2019 (COVID-19)[29]

1.1.9 Consequences

COVID-19 has rapidly affected the day to day life, businesses, disrupted the world trade and movements. The impacts of COVID-19 in daily life are extensive and have far reaching consequences. These can be divided into downsides on economy, on healthcare system, on psychology/society and on environmental issues. Covid-19 has burdened the existing medical system, overloading doctors who are at very high risk.

Meanwhile, patients with other health problems were neglected, pandemic led to cancellation of many health care visits resulting in non prevention or diagnosis of serious illnesses and cancer. Poor cash flow in the market, due to the continuous lockdowns for example, led to huge losses in national and international businesses. On a society level, quarantines and lockdowns had detrimental effects on mental health by increasing stress and social distancing from the peers and family members [40]. Furthermore, the environmental toll of disposable masks is dramatically increased in a planet suffering from climate change due to overproduction and overconsumption.

1.2 Related Works

The inability to test at scale has become humanity's Achilles' heel in the ongoing war against the COVID-19 pandemic [41]. The limited availability of testing due to geographical and temporal factors, the scarcity and expense of clinical tests needed to cover the massive time-sensitive demand combined with the fact that vaccinations are not highly effective in the current variants of concern make imperative the need of investigating the feasibility of a preliminary diagnosis tool. Therefore, various intelligent diagnostic approaches have been proposed in the literature to fight against this pandemic situation. Several of them involve the use of artificial intelligence applied to images. For example, it has been demonstrated that COVID-19 can be detected from computed tomography (CT) images[42] and from X-ray images [43] with deep learning methods. Another school of thought, proposes the extraction of features from cough, breath and speech samples (either handcrafted or automatically extracted from VGGish networks [44]) which are fed into logistic regression, support vector machines and neural networks .

1.2.1 Feature Extraction in Audio Recognition

Several researchers have studied how to extract features of sound and recognize the sound. Feature extraction is the key of speech and audio processing. Spectral features computed from windowed Discrete Fourier Transform (i.e. DFT) or Linear Predictive (i.e. LP) models are used in most of speech processing. The DFT and LP models perform good under clean conditions but verification accuracy degrades under different surrounding. Shintri et al. [45] used Mel Frequency Spectrum Coefficient (MFCC) as a method of extracting audio features. They proposed a multitaper MFCC feature extraction method which extracts low variance MFCC features from the speaker's voice samples. For speaker verification the extracted feature is used to design a model using classifier Gaussian Mixture Models (i.e. GMM), in order to decide whether to accept or deny the registered speaker. They achieved an accuracy of 87.5% with multitaper MFCC extraction as a method which is higher compared to the typical MFCC extraction. Xie et al. [46] used MFCC to recognize abnormal voice. Many attempts have been made to analyze the cough type, its intensity and its sound from its acoustic properties. In the study of Singh et al.[47] a Linear Predictive analysis of cough signal is done. For a short duration of milliseconds the cough signal was considered as a stationary system for carrying out linear prediction analysis. The main aim of this study was to classify the ailment cough and healthy cough and compare it with normal voice of speakers. Swankar et al. [48] applied logistic regression on a comprehensive set of features including MFCC, Formant Frequencies, non Gaussianity on 178 cough instances from 46 subjects collected using a bed-side microphone. They were able to achieve 80% sensitivity and 73% specificity. Quan et al.[49] use mel spectrograms and a convolutional neural network in order to distinguish cough sounds. Cai et al. [50] compared the impact of

different input features on the performance of cough recognition. The input features were namely Short-Time Fourier Transform, Mel Spectrograms, Log Mel Spectrograms and MFCC. Among the four input features they selected Mel spectrograms because they achieved the best accuracy (92.67%). And it has the best sensitivity, which may be largely due to the enhancement of certain features of cough during Mel conversion, which increases the accuracy of recognition. Other studies have introduced the concept of visual multi feature fusion according to which multiple features are combined for image classification tasks. Peng et al. [51] analyzed the effect of single features such as Mel Scale Spectrogram, Log-Mel Scale Spectrogram, and Mel frequency cepstral coefficient as well as multi-feature such as Mel-MFCC, LogMel-MFCC, and Mel-LogMel-MFCC. The experiment results showed that in the environmental sound classification tasks, multi-features are better than the single features in the same dimensions, and LogMel-MFCC has the strongest robustness. Xie et al. [52] has used both visual and acoustic features to train a CNN and combine the results of both domains to classify a bird sound. They have transformed the audio data through Constant Q-Transform (CQT), which is the input feature to CNN. For acoustic features, they have chosen spectral centroid, spectral bandwidth, spectral contrast, spectral flatness, spectral rolloff, zero-crossing rate, the energy of the signal, and Mel Frequency Cepstral Coefficients (MFCC). For acoustic and visual feature classification, they have compared the results of K-NN and Random -Forest classifier. Non Matrix Factorization spectrograms and MFCCs have been used for diagnosing covid-19 from cough samples [53]. To calculate the NMF-spectrogram feature, they first did a fourier transform on the audio files to get the spectrum features. After that, they performed a Non-negative Matrix Factorization on the spectrum. Then, they took the resulting non-negative matrix without the temporal values.

1.2.2 Cough Classification and Cough Detection

Prior studies have shown that coughs from distinct respiratory syndromes have distinct latent features. Cough is a powerful reflex mechanism for the clearance of the central airways of inhaled and secreted material. Typically, it follows a well-defined pattern, with an initial inspiration, glottal closure and development of high thoracic pressure, followed by an explosive expiratory flow as the glottis opens with continued expiratory effort [54]. Air from the lungs passes through the trachea and larynx and into the vocal tract pharyngeal, oral and nasal cavities. The way in which we breathe while speaking, including the rate and length of an exhalation and its intensity and variability, highly affects the quality of our voice. The respiratory system is highly coordinated with these primarily laryngeal-based subsystems [55]. Likewise, in turn, laryngeal activity is finely coupled to articulation in the oral and nasal cavities [56]. Coughing is one of the predominant symptoms of COVID-19 as described earlier and also a symptom of more than 100 other diseases caused by bacterial or viral respiratory infections apart from Covid-19. Trained physicians have been using cough sounds to perform a differential diagnosis among several respiratory conditions such as pneumonia, asthma, COPD, laryngitis and Tracheitis. It has also been postulated that the glottis behaves differently under different pathological conditions and this makes it possible to distinguish between coughs due to Tuberculosis, asthma , bronchitis and pertussis (whooping cough) [57]. This is possible because in all these diseases the nature and location of the underlying irritant in the respiratory system is quite different leading to audibly distinct cough sounds [41]. For example, lung diseases can cause the airway to be either restricted or obstructed and this can influence the acoustics of the cough [58]. COVID-19 disease is commonly distinguished by disorders in respiratory physiology counting with the diaphragm and other sections of the lower respiratory tract, thereby influencing breathing patterns in the course of inhalation and exhalation of air from the lungs. The cough associated with COVID-19 has been reported to be dry (non-productive) in the early phases of the disease

and more productive (wet) as the disease begins to affect the lungs. This possible transition from a dry to wet cough is notable among diseases with cough as a symptom [59].

Cough detection is a preprocessing step which determines whether a cough sound exists or not in a sound file. The difficulty of cough recognition mainly lies in the distinction of background noise. There are many kinds of sound mixed together in daily scenes. How to effectively distinguish between coughing and other sounds has become a difficult problem to be solved. Some of the deep neuronal networks which were tested for the task of cough detection were YAMNNet[60] and Ubicoustics[61]. These architectures are capable of classifying a wide range of frequently occurring sounds in the environment. Ubicoustics employs a model based on the pre-trained YouTube-8M VGG-16 architecture with a modified last layer, and trained on sound set with substantial audio augmentation including amplification augmentation, persistence augmentation and mixing augmentation using sound effect libraries such as AudioSet and Freesound[62]. YAMNet classifies audio segments into sound classes described by the AudioSet ontology employing MobileNet[63]. Both models need a post-processing step since they perform a weak labelling of sound events as shown in figure 2.12. In post-processing the confidence scores for each of the sound classes are converted into binary masks, then a selection of the best threshold of detecting cough samples must be done, while the boundaries of the sound regions are found through a detect speech algorithm. Moreover, apart from deep neuronal networks, researchers applied an XGBoost classifier in the Coughvid dataset to remove non-cough recordings using 78% of the available data [64]. In the current thesis, for the specific task of cough detection the universal system for cough detection has been used [65]. In this system Simou et al. achieved a sensitivity in the order of 90% and a specificity in the order of 99% in a domestic environment with the utilization of Long-Short-Term-Memory deep neural network architecture. Their methodology employs onset detection which is used for spotting impulsive events in the audio stream. For each detection, a short signal segment is extracted around the onset which is subsequently passed as input to the feature extraction step. The feature representation of the audio segment is then passed to the LSTM that decides whether a cough or a non-cough event occurred. Furthermore, Bales et al. [66] proposed a system which was successfully able to detect and separate cough events from background noise. They used CNNs to first detect and separate cough sounds from different types of sounds. They then used the detected cough sounds to diagnose three potential illnesses (i.e., bronchitis, bronchiolitis and pertussis) based on their unique cough audio features in a unified framework. For the cough detection task, the input raw audio clips were transformed into Mel-spectrogram, resulting in a 2-dimensional image where one dimension represents time, other dimension represents frequency and the value of pixels in the image represent the amplitude. The resulting images were then converted to grayscale. The CNN structure had three max pooling layers and two convolutional layers with 32 filters and a size of 5×5. The features learned from this convolutional block were flattened before passing them to two fully connected layers, each having 128 neurons and ReLU activation function. An accuracy of 89.05% was achieved. For the task of diagnosis of potential illnesses the sound files are cut to a single cough event which lasts two seconds and then they are converted to mel spectrograms and turned into gray. Mel spectrograms are fed into a similar CNN structure to the one used for the cough detection task but after the 2×2 max-pooling layer at the end of convolutional block, another similar convolutional block comprising of two convolutional layers is added. This method achieved an F1-score of 94.43% for diagnosing pertussis, 85.74% for the detection of bronchitis and 88.89% for the detection of bronchiolitis. Quan et al. [49] proposed a cough recognition method based on Mel-spectrograms and a Convolutional Neural Network. First, they enhanced the audio data and mix the voice in various complex scenes. Then, they preprocessed the data to ensure the consistency of data length and convert it into a Mel-spectrogram. At last, they built a CNN-based model to classify the cough using the Mel-spectrogram. After the experiment result comparison, it can be seen that this method can effectively identify and detect coughing in complex scenes. An accuracy of 98.18% is achieved. The architecture has four convolutional blocks

each one comprising of two convolutional layers except for the first one, which has one convolutional layer, a kernel size of 5 and a stride of 2. The rest of them have a kernel size of 3 and a stride of 1. The last fully connected layers have 256 and 2 neurons respectively. Infante et al. used a machine learning method to recognize dry/wet cough [67]. Semi-supervised Tree Support Vector Machine is proposed for cough recognition and detection. K-NN is also an efficient tool that is often used for cough recognition [68]. In addition, the Artificial Neural Network (ANN), Gaussian Mixture Model (GMM), Support Vector Machine (SVM), and other methods are also used for cough recognition [69].

1.2.3 Covid-19 Diagnosis from cough samples

Many studies have tried to utilize AI to classify Covid-19 using cough sounds. The detection of COVID-19 by cough sound is very economical, does not require contact, thereby reducing the risk of COVID-19 transmission, can be carried out in bulk and the results are fast. The first related work was developed by MIT researchers. They developed an AI speech framework for the detection of Covid-19 from cough recordings[70]. This model was trained using a dataset that they collected themselves, Opensigma dataset. Cough recordings were transformed with Mel Frequency Cepstral Coefficient and inputted into a Convolutional Neural Network (CNN) based architecture made up of one Poisson biomarker layer and 3 pre-trained ResNet50's in parallel, outputting a binary prescreening diagnostic. Firstly, they created a vocal cord biomarker model capable of detecting changes in basic features of vocal cord sounds in continuous speech and they trained a ResNet50 with input shape (300, 200) from MFCC to discriminate the word 'Them' from others using LibriSpeech, which is an audiobook dataset. Secondly, they trained a Sentiment Speech classifier model to learn sentiment features on the RAVDESS speech dataset. Finally, a ResNet50 was trained on binary classification of English vs Spanish spoken language with input shape (600, 200) from MFCC. The results revealed that when using pretrained ResNet50 on audio datasets, higher results were accomplished, compared to the results given by non-pretrained ResNet50. There are several authors that have concluded pretraining on audio datasets in covid-19 classification tasks from cough sounds offers better results. Imran et al. [71] presented a deployable AI-based preliminary diagnosis tool for COVID-19 using cough sounds through an application called AI4COVID-19. The smartphone app records cough when prompted by the press and release button. The recorded sounds are forwarded to the server when the diagnosis button is pressed. At the server, the sounds are first fed into the cough detector. In case, the sound is detected as a cough, the sound is forwarded to three parallel, different classifier systems, i.e., Deep Transfer Learning-based Multi Class classifier (DTL-MC), Classical Machine Learning-based Multi Class classifier (CML-MC) and Deep Transfer Learning-based Binary Class classifier (DTL-BC). The results of all these three classifiers are then passed on to a mediator. The app reports a diagnosis only if all three classifiers return identical classification results. If the classifiers do not agree, the app returns 'test inconclusive'. The AI4COVID-19 engine displays three results as output to the user which are the following: Covid-19 likely, Covid-19 not likely and test inconclusive. The architecture minimizes the misdiagnosis error since it demands that the three classifiers converge to the same output. The cough detector acts as a filter before the diagnosis engine and is capable to distinguish cough sound from 50 types of common environmental noises. To train and test this detector, they used the ESC-50 dataset and the cough and non-cough sounds recorded from their smartphone app and in particular they used 1838 cough sounds and 3597 non-cough environmental sounds for training and testing. The recorded cough sample is forwarded to their cloud-based server where the cough detector engine first computes its Mel-spectrogram with 128 Mel components (bands). This image is then resized and converted into grayscale. The resultant image is then fed into a Convolutional Neural Network (CNN) based classifier to decide whether the recorded input sound is of cough or not. Results demonstrate that

the cough detection algorithm can classify between cough evet and no cough event with an overall accuracy of 95.60%. When the input sound is detected to be cough by the cough detection engine, it is forwarded to the tri-pronged mediator-centered AI engine to diagnose between COVID-19 and non-COVID-19 coughs. To train their cough diagnosis system, they collected cough samples from COVID-19 patients as well as pertussis and bronchitis patients. They also collected normal coughs, i.e., cough sounds from healthy people. At the time of writing, they had access to 96 bronchitis, 130 pertussis, 70 COVID-19, and 247 normal cough samples from different people, to train and test the diagnosis system. The first classifier leverages a CNN-based four class classifier, using Mel spectrograms as input. The four classes here are cough caused by 1) COVID-19, 2) pertussis, 3) bronchitis or 4) normal person with no known respiratory infection. The architecture is similar to the one implemented in the cough detection step, but the number of neurons in the last layer was modified since there are 4 classes and not two. the second classifier begins with a different pre-processing of cough sounds. Instead of using a spectrogram like the first classifier, it uses MFCC and Principal Component Analysis based feature extraction. These smart features are then fed into a multi-class support vector machine (SVM) for classification. Class imbalance is handled with random undersampling. The third parallel diagnosis test also uses deep transfer learning based CNN on the Mel spectrogram image of the input cough samples, similar to the first branch of the AI engine, but performs only binary classification, i.e. Covid-19 or not. The performance of the two deep learning-based classifiers (DTL-MC and DTL-BC) is superior than the manual feature extraction based classic machine learning classifier (CML-MC). In particular, DTL-MC achieved an accuracy of 92.64%, CML-MC achieved an accuracy of 88.76% and DTL-BC 92.85%. Brown et.al [72] used a machine learning-based algorithm to distinguish between healthy and COVID-19 cough sounds (crowdsourced data) from patients even with pre-existing asthma conditions. The authors report an average AUC metric of 70% for the tasks reported in the study. Subsequent to this effort, the Coswara project [73] compiled a crowdsourced dataset containing a variety of sounds including sustained phonations, coughs and breathing patterns. Utilizing classical features such as Mel-frequency cepstral coefficients (MFCCs), spectral centroid and mean square energy features to train a random forest classifier for the sound classification task, the authors report a test accuracy of 66%. Pahar et al.[74] developed COVID-19 cough classifiers using smartphone audio recordings and seven machine learning architectures. To train and evaluate these classifiers, they used two datasets, the Coswara Dataset and the Sarcos Dataset. Data imbalance was handled with Synthetic Minority Oversamling Technique. With regards to the features extraction process, MFCCs along with the velocity (first-order difference, Δ) and acceleration (second-order difference, $\Delta\Delta$) were extracted as well as kurtosis, log energies and zero crossing rates. The best-performing classifier is the Resnet50 architecture and is able to discriminate between COVID-19 coughs and healthy coughs with an AUC of 0.98 on the Coswara dataset. When testing on the Sarcos dataset, the LSTM model trained on the Coswara dataset exhibit the best performance, discriminating COVID-19 positive coughs from COVID-19 negative coughs with an AUC of 0.94 while using the best 13 features determined by sequential forward selection (SFS). The LSTM model has 128 LSTM units, each with rectified linear activation functions and a dropout rate of 0.3. This is followed by two dense layers with 32 and 8 units respectively and rectified linear activation functions. Despotovic et al. [75] created the CDCVA dataset repository [76] (i.e. COVID-19 Detection by Cough and Voice Analysis). Five vocal tasks were asked from the participants: sustained phonation of a vowel, coughing (3 times), breathing deeply in and out through mouth (3 times), number counting from 1 to 20, and reading a specified text. Health status of the participant (positive or negative to COVID-19) is determined based on the self-declaration confirmed by the standard RT-qPCR or RAT test, with the date of testing. They opted to experiment with standard acoustic feature sets, such as the Geneva Minimalistic Acoustic Parameter Set (GeMaps), extended Geneva Minimalistic Acoustic Parameter Set (eGeMaps) and ComParE feature set, which are used as baseline feature sets for various acoustic tasks. They furthermore experimented with the wavelet scaterring features which are used to extract low-

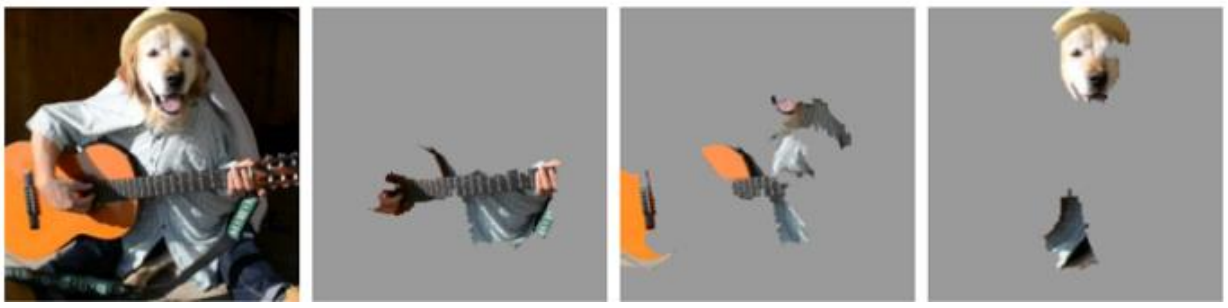
variance representations from audio signal by applying a wavelet scattering transform. The features that are extracted are fed into three ensemble models: random forests, boosted and bagged decision trees. Obtained results provided in reveal that although minimalistic sets of acoustic features, such as GeMaps and eGeMaps, are capable of learning intrinsic features from coughs, substantially better results are obtained using a brute force audio feature extraction approach with ComParE features, leading to accuracy and sensitivity approximately equal to 87% for random forests, whereas specificity goes up to 90.87% in case of bagging. Artificial Intelligence methods are critical tools for utilizing the rapidly growing body of COVID-19 positive patient datasets, with a vast contribution in the fight against this pandemic [77]. Chaudhari et al. [78] found that an ensemble model of three features showed the best performance. The first feature was mel-frequency cepstral coefficients (MFCCs), the second mel spectrograms and the last one was a binary label about the presence or absence of current respiratory diseases. The best performing network was an ensemble of 3 separate networks, each one for the three inputs described earlier, whose structure and hyperparameters were fine-tuned using grid search to minimize overfitting. Outputs from each network were aggregated to predict the probability of having COVID-19. An AUC of 77% was achieved for the Coswara/Coughvid dataset.

1.2.4 Interpretability Methods

The improved predictive accuracy of deep learning methods has often been achieved through increased model complexity. Deep Learning has achieved state of the art performance, similar to that of human experts in solving classification tasks in computer vision from lung disease classification, metastasis detection for breast cancer, skin lesion classification, identifying diabetic retinopathy, attention deficit hyperactivity disorder (ADHD), Alzheimer's disease and improving reconstruction for MRI, PET/CT imaging. One of the most widely known definitions of interpretability is the one of Doshi-Velez and Kim, who, in their work [79] , define it as "the ability to explain or to present in understandable terms to a human". The more interpretable a machine learning system is, the easier it is to identify cause-and-effect relationships within the system's inputs and outputs. For example, in image recognition tasks, part of the reason that led a system to decide that a specific object is part of an image (output) could be certain dominant patterns in the image (input) [80]. According to the type of algorithms that could be applied, the scale of interpretation and the type of data, interpretable methods can be divided into model-agnostic/model-specific, local/global and for tabular/text data/images respectively. Local methods explain a single prediction whereas global explain the overall model. Model agnostic methods can be applied to any model while model specific can be applied to a single model or a group of models. A substantial portion of attention regarding python tools is focused on deep learning for images more specifically on the concept of saliency. Saliency refers to unique features, such as pixels or resolution of the image in the context of visual processing. The local interpretable model-agnostic explanations (LIME) method is one of the most popular interpretability methods for black-box models. LIME samples input data used to train a classification model, slightly perturbs the training data, and evaluates the perturbed data with the classification model to evaluate how changes to input impact output. Figure 1.9 illustrates the explanation returned by LIME in an image classification prediction made by Google's Inception neural network**.** 1.9 (b) shows the area of the image (super-pixels) that have a stronger association with the prediction of "Electric guitar", 1.9(c) shows the super-pixels that have a stronger association with the class 'Acoustic guitar' and 1.9(d) explains the class 'Labrador'. Another interpretability technique that can be applied to any black-box model is Shapley Additive explanations (SHAP). SHAP provides explanations on individual models' decisions in the form of particular feature contributions. In[81] an innovative

approach, based on the XGBoost algorithm and a variant of the SHAP method named Tree SHAP, towards the development of an explainable Cardiovascular Disease risk prediction model, was proposed.

Despotovic et al.[75] went deeper to further investigate the explainability of the ensemble models by trying to discover the exact mechanisms that alter the acoustic parameters of coughs in people with COVID-19. They analyzed the ten most informative features in "ComParE" acoustic feature set. The best indicator of COVID-19 coughs according to the mutual information criterion is the root mean square signal frame energy. They considered that cough has bursts of energy increase in a short interval of time which are more evident in signals produced by people with COVID-19. Spectral harmonicity, the second most informative feature, describes the harmonic structure of an audio signal in which the sound frequencies are integer multiples of the fundamental frequency. Ghoshal et al. [82] trained Bayesian Deep Learning classifier using transfer learning method on COVID-19 X-Ray images to estimate model uncertainty. Their experiment has shown a strong correlation between model uncertainty and accuracy of prediction. The estimated uncertainty in deep learning yields more reliable prediction, which can alert radiologists on false predictions, which will increase the acceptance of deep learning into clinical practice in disease detection. Furthermore, they qualitatively compared, the saliency maps produced by various state-of-the-art methods e.g. Class Activation Map (CAM), Guided Backpropagation and Guided Gradient CAM and Gradients. Chatterjee et al. [83] used ResNet18, ResNet34, InceptionV3, DenseNet161, InceptionResNetV2 to detect Covid-19 from chest X-ray images and used Occlusion, Saliency, Input X Gradient, Integrated Gradients, Guided Backpropagation, DeepLIFT to interpret their results. ResNet18 is the most outstanding model, as it yielded high evaluation scores, despite having the least number of network parameters. The interpretability analysis of this model showed where the lesion was located and also the network can be utilized for the follow-up or severity estimations.



(a) (b) (c) (d)

Figure 1.9:(a) Shows the original image, (b) explains the electric guitar, (c) explains the acoustic guitar and (d) explains Labrador[72].

1.3 Scope of Thesis

In this thesis, Covid-19 is detected with deep learning methods from cough samples. The preprocessing steps consist of automatic cough recognition and noise reduction for a sensitive analysis. Data augmentation is also implemented to the denoised audio files before they are fed into most of the deep learning architectures. Then, the audio files are converted into mel spectrograms and the problem is handled as a binary image classification task. Nine architectures are presented and one of them is an ensemble approach of three pretrained models. Data imbalance is handled with various techniques

including ensemble learning, SMOTE and Random Oversampling with the last two being applied to multistage transfer learning. Finally, LIME is used to interpret the outputs of the aforementioned deep learning architectures.

The structure of the current thesis is presented as follows. Chapter 2 describes the evolution of Deep Learning, audio features and mel spectrograms, basic concepts of Convolutional Neural Networks and Recurrent Neural Networks, as well as performance measurements. Chapter 3 contains the dataset description, the data preprocessing analysis, the methodology and the DNN architectures used and LIME description. In Chapter 4 the results about the performance of classifiers and interpretability models are presented. Chapter 5 conclusions are reached and future research is suggested.

Chapter 2

Theoretical Framework

2.1 Audio Features and Mel Spectrograms

Sound is produced by the vibration of an object. Vibrations cause air molecules to oscillate. The subsequent change in the air pressure creates a wave. The waveform carries multifactorial information about the frequency, the intensity and the timbre of the sound. The audio features describe the sound and its each one captures a different aspect of the sound. Based on the level of abstraction, there are three categories of audio features: high level features (i.e. instrumentation, key, chords, lyrics, melody, rhythm, mood, tempo, genre etc.), mid-level features (i.e. pitch related descriptors such as MFCCs, onsets, fluctuation patterns) and low level (i.e. amplitude envelope, energy, spectral centroid, spectral flux etc.)[84]. Based on the signal domain, audio features are divided into time domain, frequency domain, time frequency domain and cepstral domain features. Examples of time domain features are amplitude envelope, root mean square energy and zero crossing rate. Some frequency domain features are namely band energy ratio, spectral centroid and spectral flux. The time frequency representation of a signal is done through spectrograms, mel spectrograms and constant Q transform for example. Cepstral domain based features are MFCCs, Liner Prediction Cepstral Coefficients, Perceptual Linear Prediction and Gammatone Cepstral Coefficients [85].

In the current thesis, mel spectrograms will be used as input images to deep learning architectures. The Mel spectrum contains a short-time Fourier transform (STFT) for each frame of the spectrum (energy/amplitude spectrum), from the linear frequency scale to the logarithmic Mel-scale, and then goes through the filter bank to get the eigenvector, these eigenvalues can be roughly expressed as the distribution of signal energy on the Mel-scale frequency. After the audio data are processed, only the ones containing cough are denoised, augmented and transformed into Mel-spectrograms so that we can train the convolutional neural networks for detection of Covid-19. Audio data usually have complex features, so it is necessary to extract useful features to recognize the audio. The Mel- spectrogram is one of the efficient methods for audio processing. In the experiment, we employ the Python package called librosa for data processing and all parameters are as follows: number of mels = 128 and f_{max} = 8kHz. Then we call the power_to_db function to convert the power spectrum (amplitude square) to decibel (DB) units. Figure 2.1 presents the formula to convert f hertz into m mels and figure 2.2 presents the spectrogram of a healthy user.

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

Figure 2.1: Formula of conversion from hertz to mels

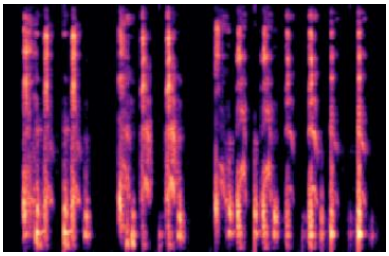


Figure 2.2: A healthy mel spectrogram

2.2 Deep Learning

2.2.1 History and Expansion of Deep Learning

Deep learning (DL), a branch of machine learning (ML) and artificial intelligence (AI) is nowadays considered as a core technology of today's Fourth Industrial Revolution . Due to its learning capabilities from data, DL technology originated from artificial neural network (ANN), has become a hot topic in the context of computing [2]. However, the primary attribute behind deep learning success has been the unprecedented accuracy in classification, segmentation, and image synthesis performance, consistently, across imaging modalities [86]. In the late 1980s, neural networks became a prevalent topic in the area of Machine Learning (ML) as well as Artificial Intelligence (AI), due to the invention of various efficient learning methods and network structures such as multilayer perceptron networks trained by "Backpropagation" type algorithms, self-organizing maps, and radial basis function networks [87]. In 2006, "Deep Learning" (DL) was introduced by Hinton et al.[88] , which was based on the concept of artificial neural network (ANN). DL technology uses multiple layers to represent the abstractions of data to build computational models. A typical neural network is mainly composed of many simple, connected processing elements or processors called neurons, each of which generates a series of real-valued activations for the target outcome. Figure 2.3 shows a schematic representation of the mathematical model of an artificial neuron, i.e., processing element, highlighting input (Xi), weight (w), bias (b), summation function (P), activation function (f) and corresponding output signal (y) [2]. Sarker [1] in his work described the different deep learning tasks according to which the taxonomy of deep learning networks occurs. Supervised is a task-driven approach that uses labeled training data while unsupervised is a data-driven process that analyzes unlabeled datasets. Semi-supervised is a hybridization of both the supervised and unsupervised methods and reinforcement is an environment driven approach. DL techniques are divided into three major categories: (i) deep networks for supervised or discriminative learning; (ii) deep networks for unsupervised or generative learning; and (iii) deep networks for hybrid learning combing both and relevant others. Discriminative architectures mainly include Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN or ConvNet), Recurrent Neural Networks (RNN), along with their variants. Commonly used deep neural network techniques for unsupervised or generative learning are Generative Adversarial Network (GAN), Autoencoder (AE), Restricted Boltzmann Machine (RBM), Self-Organizing Map (SOM), and Deep Belief Network (DBN) along with their variants. Hybrid deep networks and several other approaches such as deep transfer learning (DTL) and deep reinforcement learning (DRL) are popular.

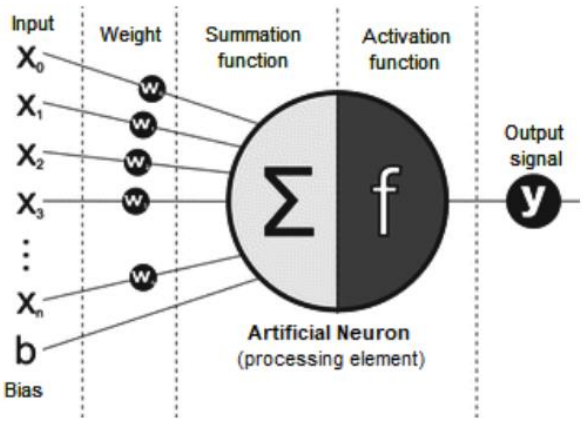


Figure 2.3: The mathematical model of an artificial neuron [1]

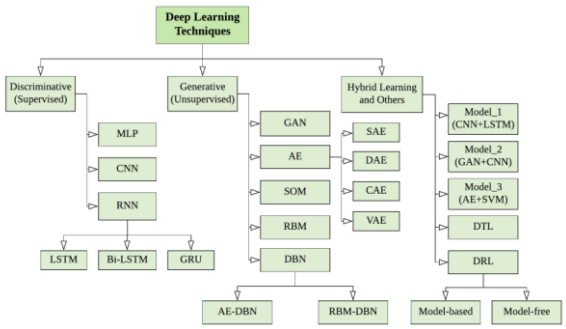


Figure 2.4: Taxonomy of DL techniques [2]

Figure 2.4 depicts the taxonomy of DL techniques described earlier. Supervised learning is used in classification and regression tasks, unsupervised learning is used for dimensionality reduction, clustering and associations. Semi supervised learning is used for classification and clustering while reinforcement learning is used for classification and control. In 1989 a novel method combining expertise of neural networks with speech recognition was used to realize a speech recognition system [89]. In 1992 Frasconi et al. analyzed the limitations and characteristics of a local feedback multi-layered network with feedback connections allowed only from neurons to itself [90]. In 1998 a multilayer neural network was trained through backpropogation algorithm applied to create a complex decision surface to classify highdimensional patterns like handwritten characters [91]. In 2010 an innovative framework of comprehending a deep neural network through layers of denoising encoders trained to denoise corrupted versions of their inputs [92]. 2012 was a thriving year for deep learning breakthroughs. ImageNet Classification with Deep Convolutional Neural Networks was a testimony that applied convolutional nets to halve the error rate for object recognition, resulting in brisk implementation of deep learning by the computer vision commune [93]. The same year DistBelief software framework was developed to train large, distributed models and significant results were obtained about large-scale nonconvex optimizations [94], Hinton et al. [95] utilized a feed-forward neural network for speech recognition, Le et al. [96] prepared a 9- layered locally connected thin autoencoder with pooling and local contrast normalization on a large dataset of images and Hinton et al. [97] reduced overfitting on large feed-forward neural networks through randomly omitting half feature detectors on each training set by overcoming complex co-adaptations for many routine tasks in speech and object recognition. In 2013 a network by Zeiler et. al. [98], winner of ILSVRC 2013 achieved top 5 error rate of 11.2%. AlexNet was fine tuned to improve performance. They examined different feature activations and their relations to the input space. The same year Graves et al.[99] trained Deep Long Short-term Memory RNN on the TIMIT phoneme recognition benchmark for speech recognition, Sutskever et al. [100] used Stochastic Gradient Descent with Momentum applied to train DNNs as well as RNNs on datasets with long term dependencies to achieve considerable performance, Vanhoucke et al. [101]achieved reduction of the neural network computational cost using speech signal stationarity for tying neural network parameters across frames. In 2014, Srivastana et al. [102] introduced dropout technique which uses random unit dropping during training to avoid coadaption. These resultant thinned networks optimize neural net performance on supervised learning tasks in vision, speech recognition, document categorization as well as computational biology. The same year two different ImageNet applications were published. Simonyan et al. [103]evaluated very deep convolutional networks (up to 19 weight layers) for largescale image classification. It was demonstrated that the representation depth is beneficial for the classification accuracy, and that state-of-the-art performance on the ImageNet challenge dataset can be achieved using a conventional ConvNet architecture with substantially increased depth. Szegedy et al. introduced GoogLeNet a 22-layer CNN, winner of ILSVRC 2014 with 6.7% error rate. Nine inception modules were utilized with over 100 layers. In 2015, an approach for conversational modelling based on sequence-to-sequence framework was developed. It was to predict the next sentence in a conversation [104]. The same year a Deep Recurrent Attentive Writer (DRAW) utilizing spatial attention mechanism to mimic the foveation of human eye with a sequential variational auto-encoding framework to construct complex images was designed [105]. Moreover, He et al. [8] introduced Microsoft ResNet, a 152-layer network architecture winner of ILSVRC 2015 with error rate 3.6%. In 2016, Gulshan et al. [106]applied deep learning to propose an algorithm for automated detection of diabetic retinopathy and diabetic retinal fundus photographs. In 2017, an algorithm applying 121-layer CNN to detect pneumonia from chest x-rays was proposed [107]. The algorithm was tested and found to exceed average radiologist performance on pneumonia detection. Table 2.1 shows a summary of deep learning tasks and methods in healthcare and medical applications.

APPLICATION AREAS	TASKS	METHODS	REFERENCES
Healthcare and Medical applications	Regular health factors analysis	CNN-based	Ismail et al. [108]
	Identifying malicious behaviors	RNN-based	Xue et al. [109]
	Coronary heart disease risk prediction	Autoencoder based	Amarbayasgalan et al. [110]
	Cancer classification	Transfer learning based	Sevakula et al. [111]
	Diagnosis of COVID-19	CNN and BiLSTM based	Aslan et al. [112]
	Detection of COVID-19	CNN-LSTM based	Islam et al. [113]

Table 2 : Summary of popular deep learning tasks and methods in healthcare and medical applications [2]

2.2.2 Convolutional Neural Networks (CNNs or ConvNets)

With the development of deep learning, more and more deep learning methods are applied to various scenarios, such as image recognition, image classification, speech recognition, machine translation, etc. As a kind of deep learning method, Convolutional Neural Networks (CNN) are widely used in the field of computer vision. CNNs are specifically intended to deal with a variety of 2D shapes and are thus widely employed in visual recognition, medical image analysis, image segmentation, natural language processing, and many more. It is worth mentioning that ultrasound imaging[114] and feature extraction from image regions[115] combined with kNN and statistical descriptors have been a breakthrough in medical image analysis before the wide application of CNNs. In this section, the components of the proposed CNN-based network are introduced. CNNs are comprised of three types of layers. These are convolutional layers, pooling layers and fully-connected layers. When these layers are stacked, a CNN architecture has been formed. A simplified CNN architecture is illustrated in Figure 2.5. According to Shea et al. [116] the input layer of a CNN will hold the pixel values of an image. The convolutional layer will determine the output of neurons which are connected to local regions of the input through the calculation of the scalar product between their weights and the region connected to the input volume. The pooling layer will then simply perform downsampling along the spatial dimensionality of the given input, further reducing the number of parameters within the activation. The fully-connected layers will perform the same duties found in standard ANNs and attempt to produce class scores from the activations, to be used for classification. The layers' parameters focus around the use of learnable kernels. These kernels are usually small in spatial dimensionality, but spreads along the entirety of the depth of the input. When the data hits a convolutional layer, the layer convolves each filter across the spatial dimensionality of the input to produce a 2D activation map. As we glide through the input, the scalar product is calculated for each value in that kernel. From this the network will learn kernels that 'fire' when they see a specific feature at a given spatial position of the input. These are commonly known as activations. Every kernel will have a corresponding activation map, of which will be stacked along the depth dimension to form the full output volume from the convolutional layer. Zero-padding is the simple process of padding the border of the input, and is an effective method to give further control as to the dimensionality of the output volumes. The calculation formula for the convolutional layer is as follows:

$$x_j^n = f\left(\sum_{i \in M_j} x_i^{n-1} * \mathrm{k}_{\mathrm{ij}}^{\mathrm{n}} + \mathrm{b}_{\mathrm{j}}^{\mathrm{n}} \right),$$

where x_j^n is the output feature map, x_i^{n-1} is the input feature map, Mj is the selected area in the n − 1 layer, k_{ij}^n is weight parameter, b_j^n is bias, and f is the activation function.

Sometimes after each convolutional layer, we conduct batch normalization to make the outputs of the convolutional layer stay identically distributed, which can improve the performance of the model. The batch normalization formula is as follows:

$$y_i = \gamma \frac{x_i - u}{\sqrt{\sigma^2 + \varepsilon}} + \beta,$$

,

where x_i is the output of convolutional layer without activation, u is the mean of x, σ^2 is the variance of x, and γ and β are parameters to learn.

Pooling layers aim to gradually reduce the dimensionality of the representation, and thus further reduce the number of parameters and the computational complexity of the model. Max pooling is a mathematical operation that works by taking the largest value from a portion of the image with a certain size, while average pooling is a mathematical operation that works by taking the average value of a portion of the image with a certain size [4]. Average Pooling is a pooling operation that calculates the average value for patches of a feature map, and uses it to create a downsampled (pooled) feature map. It is usually used after a convolutional layer. It adds a small amount of translation invariance - meaning translating the image by a small amount does not significantly affect the values of most pooled outputs. It extracts features more smoothly than Max Pooling, whereas max pooling extracts more pronounced features like edges. Figure 2.6 illustrates an example of the max pooling and average pooling operations.

Classification layer is a layer consisting of flattening, hidden layer and activation functions. Hidden layers in artificial neural networks are layers between input layer and output layer, where artificial neurons take a set of weight inputs and produce output through activation functions such as sigmoid, ReLU, or Softmax. The calculation for the fully connected layer is:

$$y_j = f\left(\sum_{i=1}^{N} x_i * w_{ij} + b_j \right),$$

where x is the input layer, N is the number of input layer nodes, w_{ij} is the weight between the links x_i and y_j, b_j is the bias, and f is the activation function.

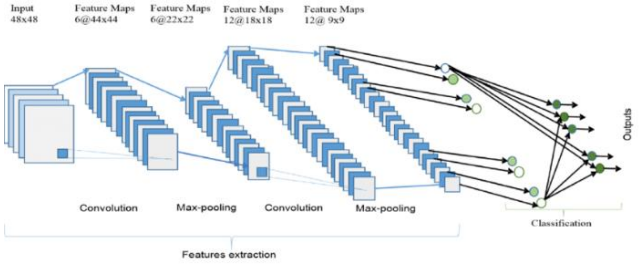


Figure 2.5 : A typical CNN architecture [3]

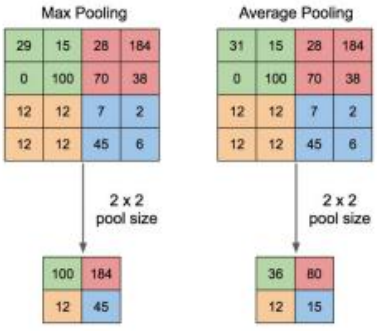


Figure 2.6 : Max Pooling and Average Pooling[4]

Since convolution is a linear operation and images are far from linear, non-linearity layers are often placed directly after the convolutional layer to introduce non-linearity to the activation map. A proper activation function significantly improves the performance of a CNN for a certain task. Rectified linear unit (ReLU) is one of the most notable non-saturated activation functions. The ReLU activation function is defined as:

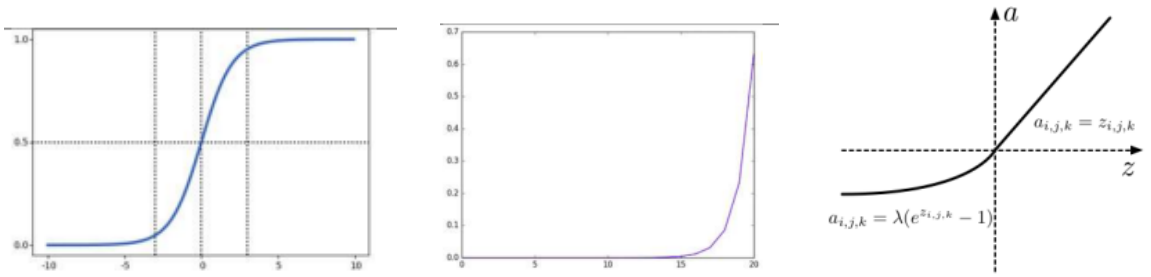
$$a_{i,j,k} = \max(z_{i,j,k}, 0)$$

where $z_{i,j,k}$ is the input of the activation function at location (i, j) on the k-th channel. ReLU is a piecewise linear function which prunes the negative part to zero and retains the positive part. The simple max(·) operation of ReLU allows it to compute much faster than sigmoid or tanh activation functions. Even though the discontinuity of ReLU at 0 may hurt the performance of backpropagation, many works have shown that ReLU works better than sigmoid and tanh activation functions empirically [117] [118]. Exponential Linear Unit (ELU) enables faster learning of deep neural networks and leads to higher classification accuracies. ELU avoids the vanishing gradient problem by setting the positive part to identity. In contrast to ReLU, ELU has a negative part which is beneficial for fast learning. As the saturation function will decrease the variation of the units if deactivated, it makes ELU more robust to noise. The function of ELU is defined as:

$$a_{i,j,k} = \max(z_{i,j,k}, 0) + \min(\lambda(e^{z_{i,j,k}} - 1), 0)$$

where λ is a predefined parameter for controlling the value to which an ELU saturate for negative inputs [117].

The sigmoid non-linearity has the mathematical form $\sigma(\kappa) = 1/(1+e^{-\kappa})$. It takes a real-valued number and compresses it into a range between 0 and 1. However, a very undesirable property of sigmoid is that when the activation is at either tail, the gradient becomes almost zero. If the local gradient becomes very small, then in backpropagation it will effectively terminate the gradient. Also, if the data coming into the neuron is always positive, then the output of sigmoid will be either all positives or all negatives, resulting in a zig-zag dynamic of gradient updates for weight. Figure 2.7 shows the sigmoid plot. Since in the current thesis the classification process is a binary task, sigmoid function is used in the last dense layer for the output. Tanh activation function squashes a real-valued number to the range [-1, 1]. Like sigmoid, the activation saturates, but — unlike the sigmoid neurons — its output is zero centered. Finally, Softmax assigns decimal probabilities to each class in a multi-class problem. Those decimal probabilities must add up to 1.0. This additional constraint helps training converge more quickly than it otherwise would. Softmax is implemented through a neural network layer just before the output layer.



(a) (b) (c)

Figure 2.7: Activation functions (a)sigmoid activation function , (b) softmax activation function, (c) ELU activation function.

It is important to choose an appropriate loss function for a specific task. Binary cross-entropy loss function is chosen in this binary classification task.

$$\text{Loss} = -\frac{1}{\text{output size}} \sum_{i=1}^{\text{output size}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)$$

,

where y_i is the true output label of the model and $\hat{y}_i$ the predicted label.

Overfitting is an unneglectable problem in deep CNNs, which can be effectively reduced by regularization. Some regularization techniques are L2 regularization and Dropout. We can quantify complexity using the L_2 regularization formula, which defines the regularization term as the sum of the squares of all the feature weights:

$$\text{L2 regularization term} = \|w\|_2^2 = w_1^2 + w_2^2 + \ldots + w_n^2$$

In this formula, weights close to zero have little effect on model complexity, while outlier weights can have a huge impact. Regularization modifies the objective function by adding additional terms that penalize the model complexity. Formally, if the loss function is L(θ, x, y), then the regularized loss will be:

$$E(\theta, x, y) = L(\theta, x, y) + \lambda R(\theta)$$

where R(θ) is the regularization term, and λ is the regularization strength. For p = 2, the l2-norm regularization is commonly referred to as weight decay. The output of Dropout is $y = r * a(W^T x)$, where

$x = [x_1, x_2, \ldots, x_n]^T$ is the input to fully-connected layer, $W \in \mathbb{R}^{n \times d}$ is a weight matrix, and r is a binary vector of size d whose elements are independently drawn from a Bernoulli distribution with parameter p, i.e. $r_i \sim$ Bernoulli(p). Dropout can prevent the network from becoming too dependent on any one (or any small combination) of neurons, and can force the network to be accurate even in the absence of certain information [117].

2.2.3 Recurrent Neural Networks and its variants

A Recurrent Neural Network (RNN) is another popular neural network, which employs sequential or timeseries data and feeds the output from the previous step as input to the current stage [2]. Like feedforward and CNN, recurrent networks learn from training input, however, distinguish by their "memory", which allows them to impact current input and output through using information from previous inputs. Unlike typical DNN, which assumes that inputs and outputs are independent of one another, the output of RNN is reliant on prior elements within the sequence. The most prevalent variants of RNNs are Long short-term memory (LSTMs), Bidirectional RNN/LSTM and Gated recurrent units (GRUs). LSTM is a popular form of RNN architecture that uses special units to deal with the vanishing gradient problem, which was introduced by Hochreiter et al. [119]. A memory cell in an LSTM unit can store data for long periods and the flow of information into and out of the cell is managed by three gates. The 'Forget Gate' determines what information from the previous state cell will be memorized and what information will be removed that is no longer useful, while the 'Input Gate' determines which information should enter the cell state and the 'Output Gate' determines and controls the outputs. LSTM is a recurrent

model, in which the current time prediction depends on all past time inputs[5]. For each layer, the LSTM processes at time t by computing:

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t c_{t-1} + i_t \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \sigma_h(c_t)
\end{aligned}
$$

where, σ_g is sigmoid function, σ_c and σ_h is hyperbolic tangent function, f , i, o are gates, c is the internal cell states, h is the hidden states. LSTM only uses past information. Bidirectional RNNs connect two hidden layers that run in opposite directions to a single output, allowing them to accept data from both the past and future. Bidirectional RNNs, unlike traditional recurrent networks, are trained to predict both positive and negative time directions at the same time. It is a sequence processing model comprising of two LSTMs: one takes the input forward and the other takes it backward. BiLSTM can also take advantage of future information. In each BiLSTM layer, there are a forward pass and a backward pass. The forward pass gets feature-maps , whereas the backward pass does the opposite by changing all t − 1 to t + 1 in the above equations. Figure 2.8 depicts the structure of a biLSTM. A Gated Recurrent Unit (GRU) is another popular variant of the recurrent network that uses gating methods to control and manage information flow between cells in the neural network. The GRU is like an LSTM, however, has fewer parameters, as it has a reset gate and an update gate but lacks the output gate. The GRU's structure enables it to capture dependencies from large sequences of data in an adaptive manner, without discarding information from earlier parts of the sequence.

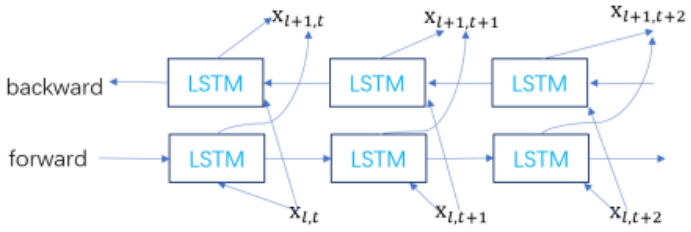


Figure 2.8: Structure of a BiLSTM [5]

A GRU unit is composed of a reset gate r_t and an update gate z_t [6]. The output h_t is determined by both current input x_t and previous state h_{t-1} under the control of these two gates. The outputs of the gates and the GRU unit are calculated as follows:

$$r_t=\sigma(W_r x_t+U_r h_{t-1}+b_r)$$

$$z_t=\sigma(W_z x_t+U_z h_{t-1}+b_z)\tilde{h}$$

$$t=\tanh[W_h x_t+U_h(r_t \odot h_{t-1})+b_h]$$

$$ht=(1-z_t)\odot h_t-1+z_t\odot \tilde{h}_t$$

where Wr , Ur , Wz , Uz , Wh and Uh are the weight matrices. br , bz , bh are the synthesis of bias vectors for input xt and previous state $ht-1$, σ is the logistic sigmoid function, tanh is the hyperbolic tangent activation function, ⊙denotes the Hadamard product.

Models with bi-directional structure have the ability to learn information from previous and subsequent data when dealing with the current data. The structure of the bi-GRU model diagram is shown in Fig. 2.9 . The bi-GRU model is determined based on the state of two GRUs, which are unidirectional in opposite directions . One GRU that moves forward, beginning from the start of the data sequence, the other GRU that moves backward, beginning from the end of the data sequence. This allows the information from both future and past to impact the current states. The bi-GRU is defined as follows:

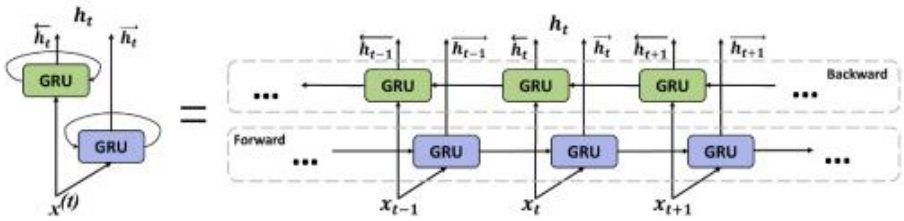


Figure 2.9: Structure of a BiGRU unit[6]

2.3 Performance Measurements

In order to better evaluate the performance of the model, several indicators were used to evaluate the model. The values used for classification assessment are True Positive (TP), True Negative(TN), False Positive (FP), False Negative (FN). Accuracy is the indicator that the samples with a correct reaction classification account for the total samples. In other words, accuracy simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Recall is the ratio of the number of samples recognized correctly to the total number of samples recognized. It is a useful metric in cases where False Negative is of higher concern than False Positive. It is important in medical cases where it doesn't matter whether we raise a false alarm but the actual positive cases should not go undetected. Recall for a label is defined as the number of true positives divided by the total number of actual positives.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Precision is the ratio of the number of samples recognized correctly to the number of samples that should be recognized. Precision explains how many of the correctly predicted cases actually turned out to be positive. Precision is useful in the cases where False Positive is a higher concern than False Negatives. Precision for a label is defined as the number of true positives divided by the number of predicted positives.

$$\text{Precision} = \frac{TP}{TP+FP}$$

F1-score gives a combined idea about Precision and Recall metrics. It is maximum when Precision is equal to Recall. F1 Score is the harmonic mean of precision and recall. F1 Score could be an effective evaluation metric in the following cases:

- When FP and FN are equally costly.
- Adding more data doesn't effectively change the outcome
- True Negative is high

$$\text{F1-score} = \frac{2*Precision*Recall}{Precision+Recall}$$

The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR(True Positive Rate) against the FPR(False Positive Rate) at various threshold values and separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes. From the graph shown in Figure 2.10 below, the greater the AUC, the better is the performance of the model at different threshold points between positive and negative classes. This simply means that When AUC is equal to 1, the classifier is able to perfectly distinguish between all Positive and Negative class points. When AUC is equal to 0, the classifier would be predicting all Negatives as Positives and vice versa. When AUC is 0.5, the classifier is not able to distinguish between the Positive and Negative classes. The TPR equals to the recall evaluation metric while FPR is computed as follows:

$$\text{False Positive Rate} = \frac{FP}{TN+FP}$$

Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). It is computed as follows:

$$\text{Specificity} = \frac{TN}{TN+FP}$$

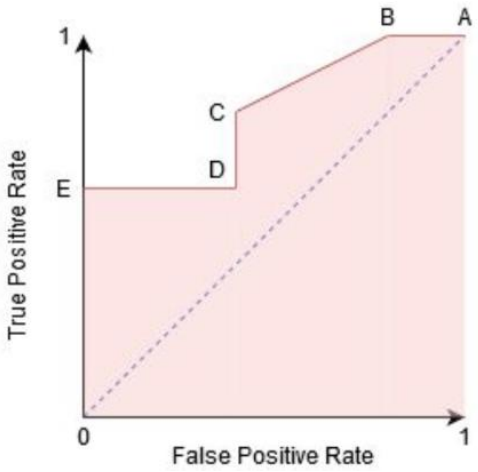


Figure 2.10: AUC curve

Chapter 3

Datasets and Methods

3.1 Dataset Description

Two datasets have been used for the classification task of diagnosing Covid-19 from cough samples. More specifically, these are the Coswara dataset and the Cambridge dataset.

3.1.1 The Coswara Dataset

The Coswara project is aimed at developing a diagnostic tool for Covid-19 based on respiratory, cough and speech sounds[73],[120].The data collection was done via a web application where users were asked to provide metadata, and proceed to recording the sound samples using the device microphone. Public participants provided 9 sound files one for each sound category: breathing (two types; shallow and deep), cough (two types; shallow and heavy), prolonged vowel pronunciation (three types; / ey /, / i/,/ u:/), and counting from one to twenty digits (of two kinds, normal and fast). For each user, metadata can be grouped into five distinctive categories: age, sex, location (country, state/province), current health status and the presence of comorbidities (pre-existing medical conditions). Health status includes 'healthy', 'exposed', 'cured' or 'infected'. The samples with a status in one of the three categories
i.e. positive mild, positive moderate and positive asymptomatic are classified as Covid, and the samples with status healthy are classified as non-Covid. Audio recordings were sampled at 44.1 KHz. In this study, we have made use of the raw audio recordings of shallow cough and heavy cough as two separate datasets and applied pre-processing as described in Section 3.2. There are 2744 samples in total but only 2661 of them have a health status description, since there are 83 samples under validation. Figure 3.1 depicts the health status of samples meaning their distribution in each of the seven categories (healthy, no respiratory illness exposed, respiratory illness not identified, fully recovered, positive moderate, positive mild, positive asymptomatic).

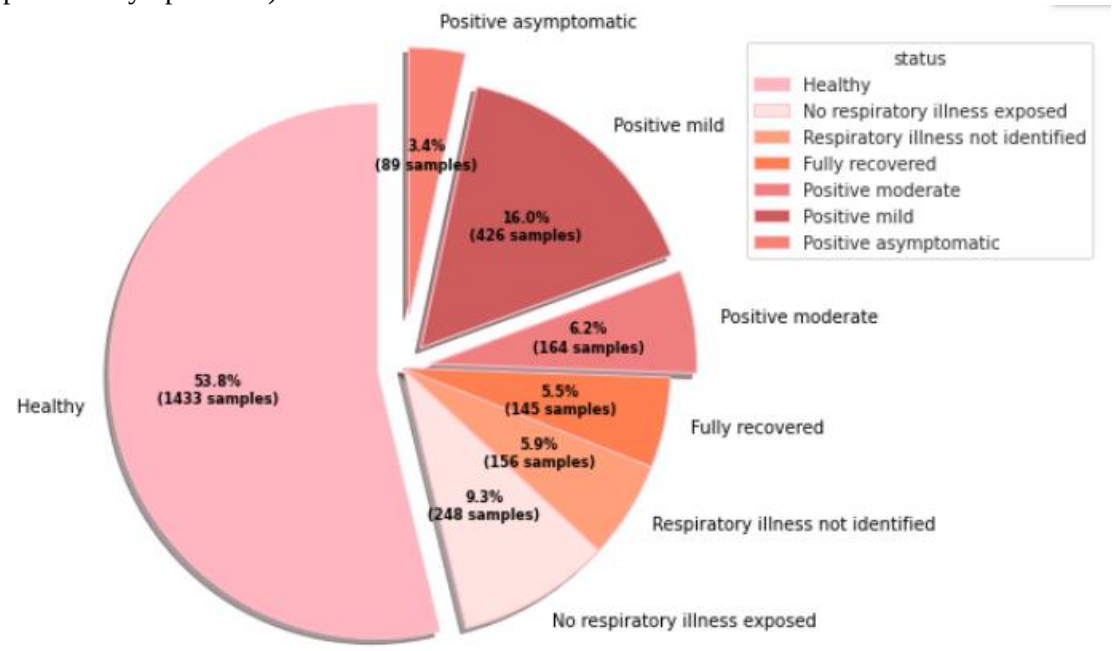


Figure 3.1: Health status distribution of samples

Information captured from the metadata are analyzed in figures 3.2 – 3.10. The age distribution of participants is shown in figure 3.2, most of them are aged between 20 and 50 with the average age being 35.15 years.

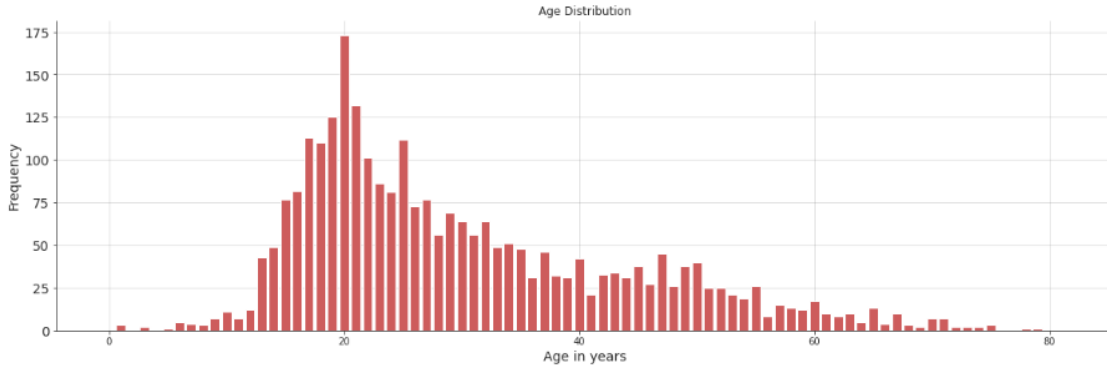


Figure 3.2: The age distribution of samples for the Coswara Dataset

With regards to the country of origin, subjects are from six continents: Asia (Bahrain, Bangladesh, China, India, Indonesia, Iran, Japan, Malaysia, Oman, Philippines, Qatar, Saudi Arabia, Singapore, Sri Lanka, United Arab Emirates, Israel, Vietnam, Thailand, Russia, Turkey, South Korea), Australia, Europe (Belgium, Finland, France, Germany, Ireland, Netherlands, Norway, Romania, Spain, Sweden, Switzerland, Ukraine, United Kingdom, Greece, Italy, Ukraine), North America (Canada, United States, Mexico), South America (Argentina, Brazil, Ecuador, Peru), Africa(Mozambique, Egypt, South Africa). The greatest proportion of samples (2515) come from India and figure 3.3 shows the top 3 continents of origin. Furthermore, there are 1900 male and 844 female subjects at the time of experimentation (figure 3.4).

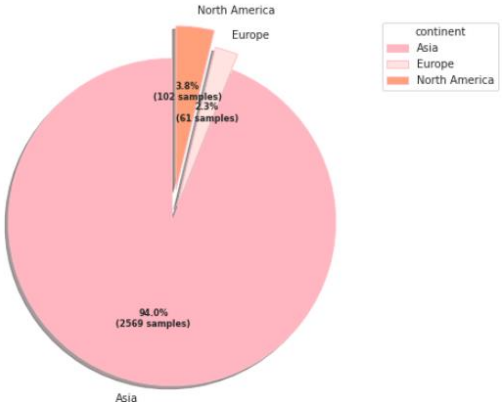


Figure 3.3: The origin of samples for the Coswara Dataset.

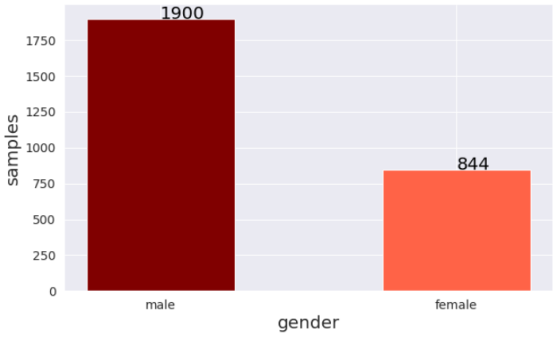


Figure 3.4: Gender distribution among subjects for the Coswara Dataset

Figure 3.5 is a violin plot of the age of participants by health status, separated by sex. The grouped violin plot shows male subjects tend to have the same age or slightly bigger than female subjects in each health status category. Further, conclusions are drawn about how the sex delta varies across categories: the median weight difference is more pronounced for positive asymptomatic and fully recovered than the rest categories.

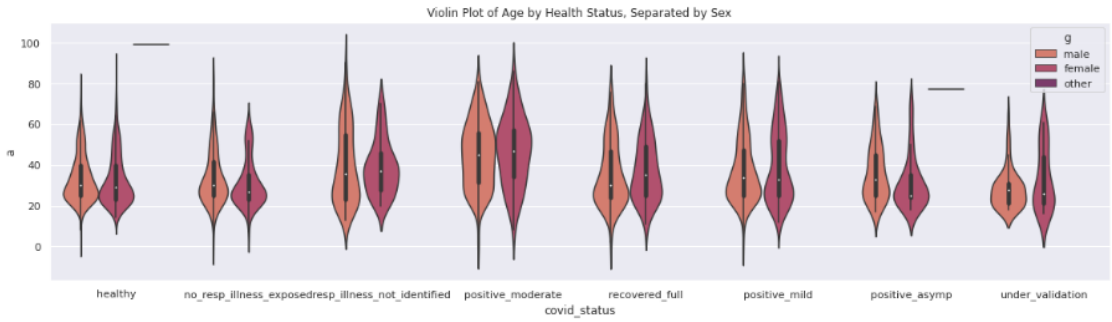


Figure 3.5: Violin plot of age of participants by health status, separated by sex.

The health status of vaccinated and non-vaccinated subjects is examined in figure 3.6. There are 752 subjects that claim to be partially (one dose) or fully vaccinated and 211 subjects that are not vaccinated at all. Health status is provided to only 700 of these subjects. In each category, except for respiratory illness not identified, vaccinated subjects prevail as expected since they are the majority.

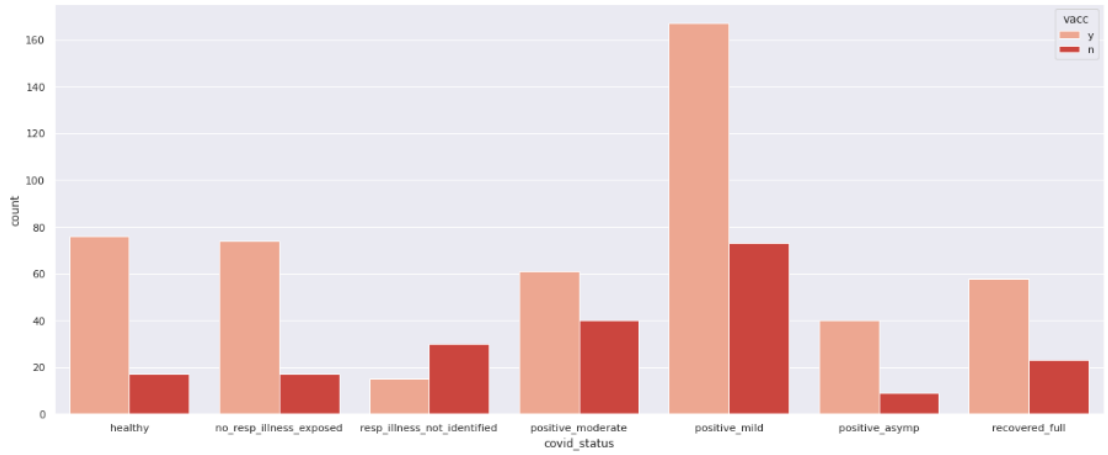


Figure 3.6: Health status of vaccinated and non vaccinated subjects

Questioned whether they are using a mask or not, 57.3% of subjects replied negatively as shown in figure 3.7. As for the referred symptoms of Covid-19, 652 users declared cough as a symptom, 52 users declared diarrhea, 211 users reported breathing difficulties, 296 users declared sore throat, 406 users reported fever, 382 users fatigue, 325 users muscle pain and 169 users reported loss of smell as a symptom (figure 3.8). As for the pre-existing conditions, 232 users claimed to have diabetes as a pre-existing condition, 134 users claimed to have asthma, 233 users hypertension, 44 users reported having a chronic lung disease and 37 users ischemic heart disease as a pre-existing condition(figure 3.9). Other conditions that have been reported include cold, at a frequency of 488 samples, pneumonia, at a frequency of 45 samples, whilst there are 225 smokers (figure 3.10).

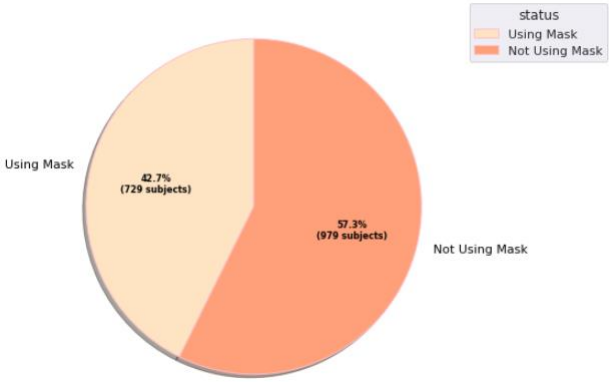


Figure 3.7: Mask and non mask users

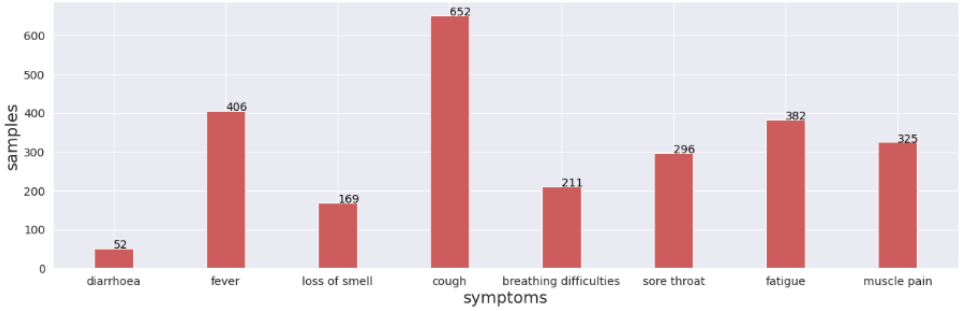


Figure 3.8: Referred symptoms of Covid-19 infection

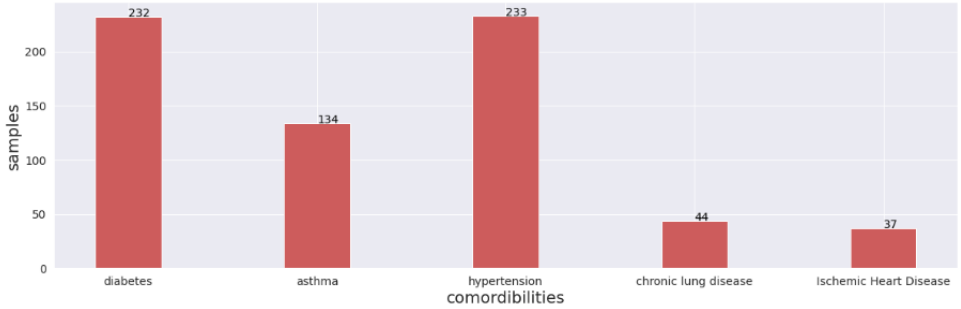


Figure 3.9:Pre-existing conditions

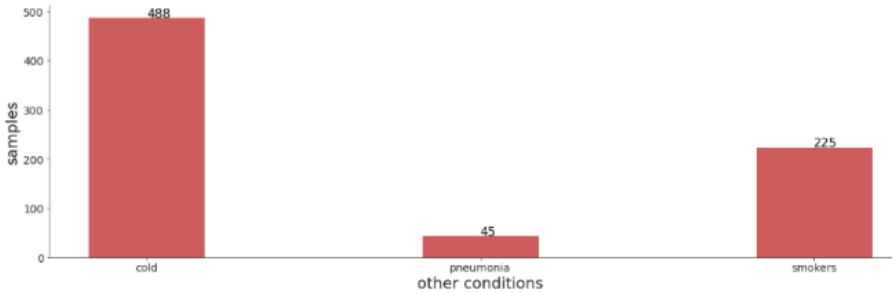


Figure 3.10:Other conditions

3.1.2 The Cambridge Dataset

The Cambridge dataset is a crowdsourced dataset collected through an app (Android and Web) that asked volunteers for samples of their voice, coughs and breathing as well as their medical history and symptoms[72]. The user is asked to input their age and gender as well as a brief medical history and whether they are in hospital. Users then input their symptoms (if any) and record respiratory sounds: they are asked to cough three times, to breathe deeply through their mouth three to five times and to read a short sentence appearing on the screen three times. Finally, users are asked if they have been tested for COVID-19, and a location sample is gathered with permission. Given the data is sensitive (i.e., containing voice) sharing agreements were set up for the data. The shared data originated both from web and android app, contained breath and cough samples and was divided to five categories: healthy with no symptoms, healthy with cough as a symptom, tested positive for Covid-19 with cough as a symptom, tested positive for Covid-19 but do not have cough as a symptom and users with asthma with cough. For the task of distinguishing users who have declared they tested positive for COVID-19 (COVID-positive), from users who have not declared a positive test for COVID-19, have a clean medical history, have never smoked and have no symptoms, the following folders were used : covid android no cough, covid android with cough, covid web no cough, covid web with cough, healthy android no symptoms and healthy web no symptoms. There are 141 Covid samples from 66 unique users and 298 non-Covid samples acquired from 220 unique users. Eventually, 124 covid samples are used and 276 non-Covid and that is because recordings of same users that are uploaded within less than 24 hours are eliminated. Each audio file contains a timestamp in milliseconds. By comparing the timestamps of same users' recordings, forty samples are excluded. Figure 3.11 shows the number of samples in each class.

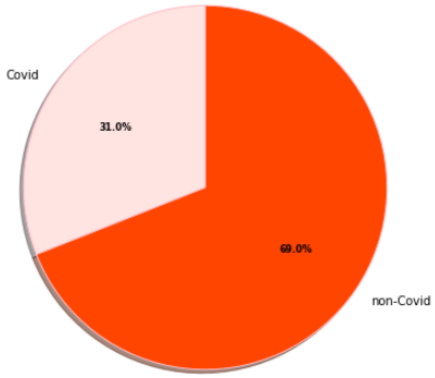


Figure 3.11: Covid and non-Covid samples for the Cambridge dataset

3.2 Data pre-processing

Preprocessing is performed to account for the potential missing, incomplete or noisy data in the dataset. Various problems can be observed in a dataset, such as missing data instances for particular vocal tasks or substitution of vocal tasks (e.g. coughing recorded instead of breathing)[75]. To tackle with the risk of the model overfitting to unwanted signals, such as the method of recording or other environmental sounds, prior to the cough-specific audio analysis and classification, it is necessary to identify if and when the cough is present in the recording, crop the audio file accordingly and denoise the raw audio files.

3.2.1 Cough Detection

The Universal System for Cough Detection in Domestic Acoustic Environments [65],[121] was used for the automatic identification of the cough samples registered in the raw audio files. The Universal System for Cough Detection in Domestic Acoustic Environments offers the benefit of strong labelling of sound events. It utilizes an acoustic onset detector as a pre-processing step, aiming to detect impulsive patterns in the audio stream. In a subsequent step, discrimination of coughing events from other impulsive sounds is handled as a binary classification task. The method for onset detection relies on measures of spectral energy across successive time-frames. Frames are formed by windowing the signal with a short-length Hanning window of length 672, moving on a continuous time-grid with hop-size 512. At each frame, the short-time Fourier transform (STFT) is calculated and the frequency bins k $\in$ [120Hz, 6kHz] corresponding to a specified spectral range are used for further processing. Various DNN architectures were tested, among which the most promising appeared to be the one based on Long Short Term Memory (LSTM) units. Specifically, two LSTM layers of 256 units each, were used, followed by one fully connected layer of 64 units and finally a dropout layer with 0.3 probability and an output softmax layer. For each .wav file of the raw audio recordings, if one or more coughs are detected firstly a.txt file will be exported which contains each cough instance's timestamp as well as the corresponding level of confidence the classifier has. Additionally, a .wav file containing all the cough detections concatenated, is exported. All of the final wav files are down sampled to 16kHz. Figure 3.13 illustrates the waveforms of the cough signal before and after the algorithm is implemented.

In Coswara cough heavy dataset there are originally 659 covid and 1376 non-covid samples due to the fact that 77 files were empty. Cough was detected in 572 covid samples (86,8%) and 1241 healthy samples (90,19%). In Coswara cough shallow, 532 out of 659 Covid samples were classified as cough(80,72%) and 1065 out of 1375 non-Covid samples (77,45%). It was expected that Coswara cough heavy would achieve higher classification results since the heavier cough has a more explosive phase compared to the shallow, so the onset detection algorithm performs better. Last but not least, in Cambridge dataset cough was detected in 119 out of 124 Covid samples (95,97%) and in 250 out of 276 non-Covid samples (90,57%).

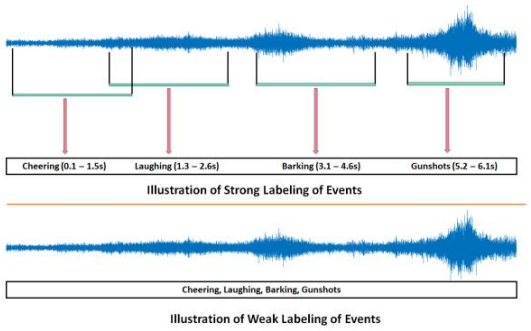
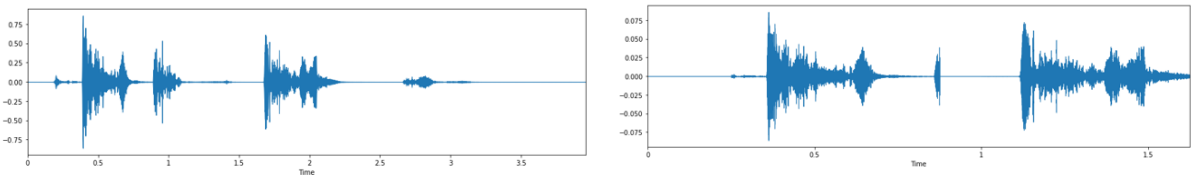


Figure 3.12: Strongly labeled vs weakly labeled data. The first ones contain time stamps of the occurrences of the events. Weakly labeled, on the other hand, only requires one to mark whether the event is present or not[122]

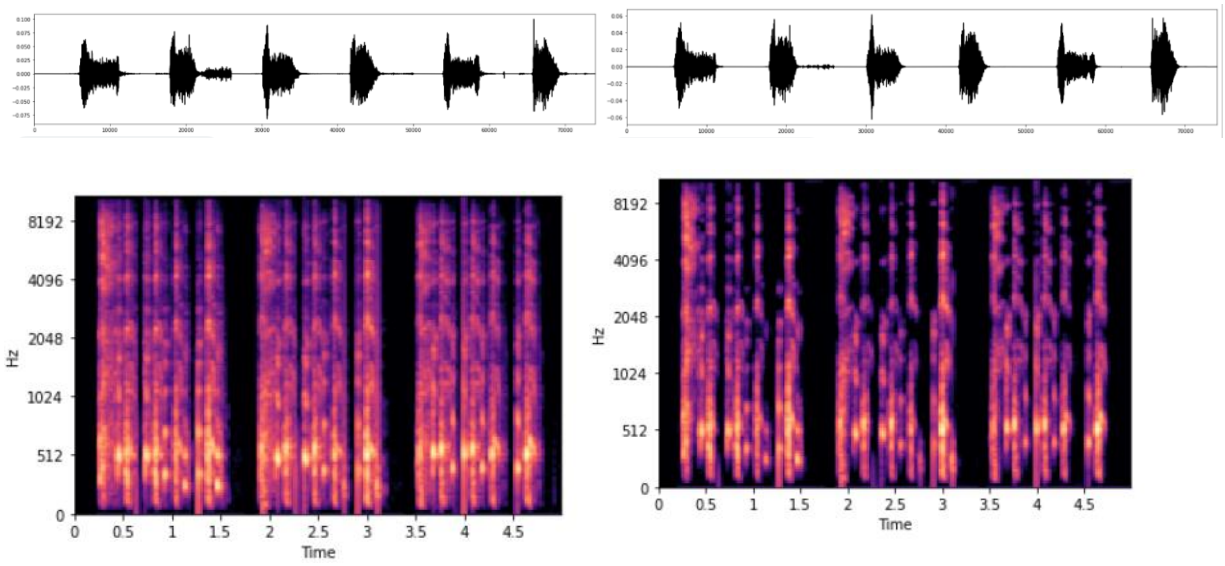


(a) (b)

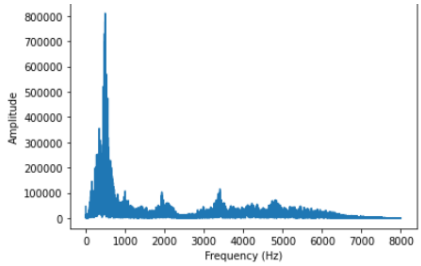
Figure 3.13: The waveforms of the cough signal a) before and b) after cough detection and cropping.

3.2.2 Noise Reduction

A major problem that comes from crowdsourced data is the lack of ability to control for the sounds in the environment and the quality of the microphone. For this purpose, a library called Noisereduce[123] is used. Noisereduce is a noise reduction algorithm in python that reduces noise in time-domain signals like speech, bioacoustics, and physiological signals. It relies on a method called "spectral gating" which is a form of Noise Gate. It works by computing a spectrogram of a signal and estimating a noise threshold (or gate) for each frequency band of that signal. That threshold is used to compute a mask, which gates noise below the frequency-varying threshold. The library works by removing a certain frequency from the target sound clip by isolating the signal using Fast Fourier Transforms. The basic algorithm which leads to data cleaning with the FFT in python is described in [124]. Firstly, an FFT is calculated over the noise audio clip. Secondly, statistics are calculated over FFT of the noise (in frequency). Then, a threshold is calculated based upon the statistics of the noise (and the desired sensitivity of the algorithm). Moreover, an FFT is calculated over the signal and a mask is determined by comparing the signal FFT to the threshold. The mask is smoothed with a filter over frequency and time, applied to the FFT of the signal, and is inverted.



(a) (b)



(c)

Figure 3.14: The waveforms and spectrograms of the cough signal of a covid patient a) before and b) after noise reduction. All spectrum resolution is kept after denoising as depicted in (c).

Overall, the noise reduction plays an important role in standardizing the datasets. Figure 3.14 compares the spectrograms and the waveforms between a noisy and the corresponding denoised audio file of a covid patient and depicts its spectrum.

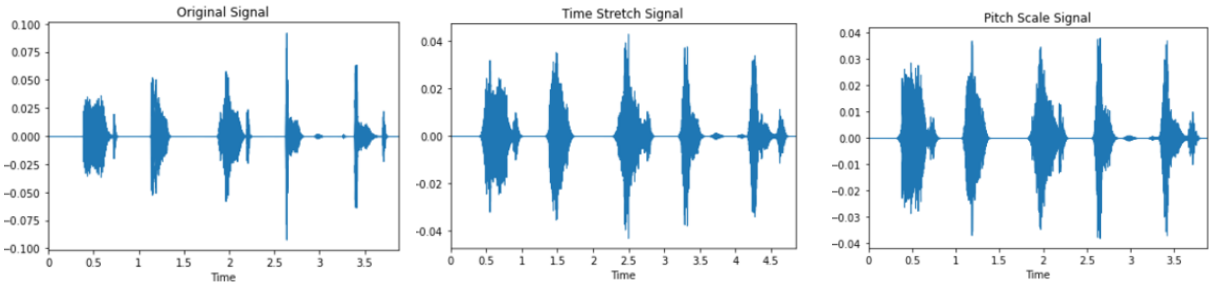
3.2.3 Data Augmentation

Data augmentation is applied to train set only , because otherwise a data leakage would occur. There are two ways of implementing augmentation to audio data: raw audio and spectrogram augmentation such as SpecAugment[125]**.** In this study, raw audio data augmentation is implemented. Data augmentation is compulsory for small datasets when using convolutional neural networks because it addresses data scarcity, it increases models' robustness, it improves models' accuracy, it reduces overfitting and it saves resources to collect and label data. There are many techniques with the most dominant ones being time shifting, time stretching (i.e. change speed without affecting pitch or frequency), pitch scaling, noise addition, impulse response addition, low/high/band pass filtering, polarity inversion, random gain (i.e. change the amplitude). For the purposes of this task I am using two python libraries, librosa[126] and audiomentations[127]. A heavier data augmentation has been applied to the Cambridge dataset due to the fact that it is smaller and more prone to overfitting than the Coswara heavy and the Coswara shallow. Furthermore, audio data augmentation can be used to improve the recall performance according to [72]. With regards to the techniques, random gain, pitch shifting[128] and time stretching have been used. The waveforms of the original and the augmented data is shown in Figure 3.15(a-c). The following has been applied:

i. Stretching Time: reduce the sample sound signal (to unchanged running pitch). Based on the factors {0.80,1.20} the duration is stretched.

ii. Shift Pitch: Sound/audio samples can be increased or decreased (to unchanged running pitch), and every sample can be shifted differently by 1 or 2 semitones up.

iii. Random gain : Multiply the audio by a random amplitude factor between 10 and 15 to increase the volume. This technique can help a model become invariant to the overall gain of the input audio.



(a) (b) (c)

Figure 3.15: Data augmentation. The waveforms of (a) the original signal b) augmented signal by a time stretching factor of 0.8 and b) augmented signal by one semitone up.

Data augmentation is done for both classes Covid and non-Covid on the training set. Eventually, this method has led to 1456 Covid samples and 3184 non-Covid samples in Coswara's heavy training set, 1352 Covid samples and 2732 non-Covid samples in Coswara's shallow training set, 616 Covid and 1113 non-Covid samples in Cambridge dataset.

3.3 CNN models

Nine models with the best results for the task of Covid-19 classification will be described in this section. Three of them are pretrained neural networks on the ImageNet dataset[129] and the rest of them are modifications of existing models in bibliography as well as a stacked neural network, which combines the predictions of the aforementioned pretrained networks. The three pretrained networks are namely Xception, InceptionResnetv2 and Resnet50. Other pretrained networks were examined as well, such as Inceptionv3, VGG-16, EfficientNetB0, MobileNetv2, Densenet121, Resnet18 but they did not achieve remarkable results. In all pretrained models, for each dataset, grid search has been done for hyperparameter tuning in order to define the optimal learning rate, activation function, momentum, optimizer, dropout rate, number of neurons in the last dense layer, batch size and number of epochs. The initial weights used are the ones of training the model on the ImageNet dataset for the pretrained networks.

3.3.1 Xception

Xception is an extension of the Inception architecture which replaces the standard Inception modules with depthwise separable convolutions[130]. It is based on the hypothesis that the mapping of cross-channels correlations and spatial correlations in the feature maps of convolutional neural networks can be entirely decoupled. Because this hypothesis is a stronger version of the hypothesis underlying the Inception architecture, the proposed architecture(Figure 3.16) was named Xception, which stands for "Extreme Inception".

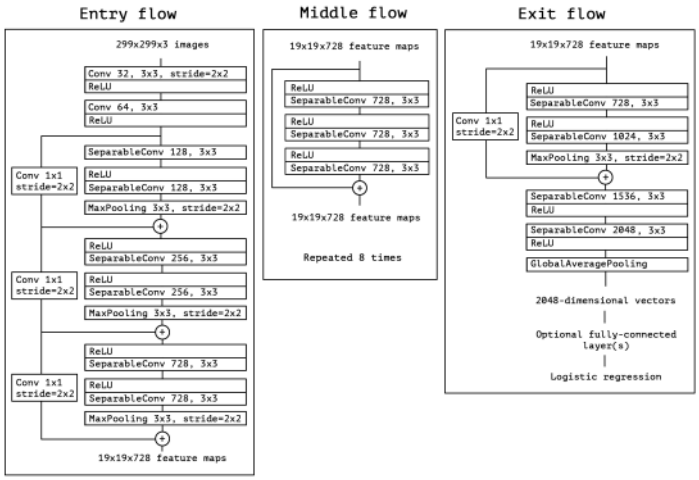


Figure 3.16: Xception architecture [130]

After grid search, the optimal hyperparameters were chosen. For the Coswara cough heavy dataset, XCeption's last fully connected layer was replaced by a dense layer consisting of 256 neurons with softmax activation function and a 0.3 dropout rate before being finally connected to a single neuron, responsible for the binary classification. Batch normalization is introduced after the activation function. The learning rate is set to 0.0001, the optimizer is RMSprop, the momentum is 0.7, the batch size is equal to 16 and trained for 10 epochs. It is important to underline the usage of the batch normalization layer. Batch normalization not only speeds up the training process and improves model generalization, but also helps reduce sensitivity to bad parameters initialization which could undermine models' training process[131]. For the Coswara cough shallow dataset, the base model is used with average pooling and batch normalization, Adam optimizer and learning rate equal to 0.0001 trained for 20 epochs. For the Cambridge dataset, the last dense layer consists of 1024 neurons and the dropout rate is set to 0.5, the optimizer chosen is Stochastic Gradient Descent and the learning rate is 0.0001, while the training process lasts for 10 epochs. The loss function is binary cross entropy in all of the above datasets.

3.3.2 InceptionResnetv2

Inception-ResNet-v2 is a convolutional neural architecture that builds on the Inception family of architectures but incorporates residual connections (replacing the filter concatenation stage of the Inception architecture)[7]. In the Inception-Resnet block multiple sized convolutional filters are combined by residual connections. The usage of residual connections not only avoids the degradation problem caused by deep structures but also reduces the training time[132]. This model is trained on more than a million images from the ImageNet database and it is 164 layers deep(Figure 3.17). With regards to the Coswara cough heavy dataset, the classification head is substituted with a dense layer of 512 neurons, a 'relu' activation and a dropout layer with dropout rate equal to 0.5. The optimizer is Adam and the learning rate is set to 0.001. Furthermore, the batch size is equal to 16 and the model achieves better results when trained for 10 epochs. In the Coswara cough shallow dataset, the base model is used with the same optimizer and learning rate as before, but it is trained for 20 epochs. Finally, regarding to the Cambridge dataset, the last dense layer has 1024 neurons with the other hyperparameters remaining the same as in the Coswara cough heavy.

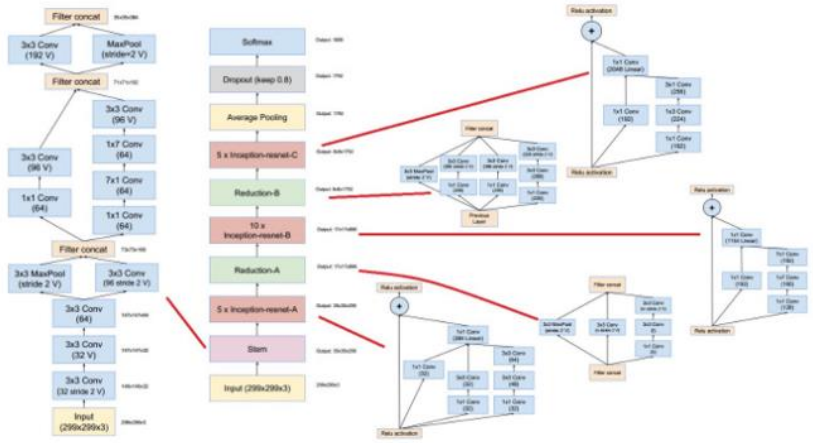


Figure 10.17: InceptionResnetv2 architecture [7]

3.3.3 ResNet50

Resnet50 has been reported as one of the models with the highest accuracy in diagnosing Covid-19 from cough samples [74],[133], while Resnet18 has extraordinary results in the classification of Covid-19 from chest X-Ray and CT images[134]. The fundamental breakthrough with ResNet was it allowed to train extremely deep neural networks with more than 150 layers successfully. Prior to ResNet training very deep neural networks was difficult due to the problem of vanishing gradients. ResNet uses skip connection to add the output from an earlier layer to a later layer. This helps it mitigate the vanishing gradient problem. The ResNet-50 model consists of 5 stages each with a convolution and Identity block. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers. The ResNet-50 has over 23 million trainable parameters[8]. With regards to the Coswara cough heavy dataset, the neurons of the last dense layer are 256, and the dropout rate is 0.5. The model is trained for 20 epochs. As for the Coswara cough shallow and the Cambridge dataset, the base model as described in Figure 3.18 is used, with the classification head being replaced with an average pooling layer and a sigmoid activation. Nadam is used as an optimizer in all three datasets, as well as the same learning rate which is equal to 0.001 and the batch size is 16.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 3.18: ResNet architectures and base models [8]

3.3.4 Ensemble Model of Pretrained Networks

Deep ensemble learning models combine the advantages of both the deep learning models as well as the ensemble learning such that the final model has better generalization performance[135]. Combination of several different predictions from different models to make the final prediction is known as ensemble learning or ensemble model. The ensemble learning involves multiple models combined in some fashion like averaging, voting such that the ensemble model is better than any of the individual models. Deep Neural Networks offer increased flexibility and can scale in proportion to the amount of training data available. A downside of this flexibility is that they learn via a stochastic training algorithm which means that they are sensitive to the specifics of the training data and may find a different set of weights each time they are trained, which in turn produce different predictions. Generally, this is referred to as neural networks having a high variance and it can be frustrating when trying to develop a final model to use for making predictions. A successful approach to reducing the variance of neural network models is to train multiple models instead of a single model and to combine the predictions from these models. This can lead, subsequently, to an increased robustness of the model. Ensemble models have been used for Covid-19 classification tasks[136]. Each model uses the optimal parameters for training as defined in 3.1.2 – 3.1.3. Then, the predictions of each model are summed with the same weights (i.e. contributing equally) and an average prediction is calculated. The probability threshold is set to 0.5. Samples with probability greater than or equal to the threshold are classified as Covid and the ones with a probability smaller than the threshold as non-Covid. Figure 3.19 depicts the model's architecture.

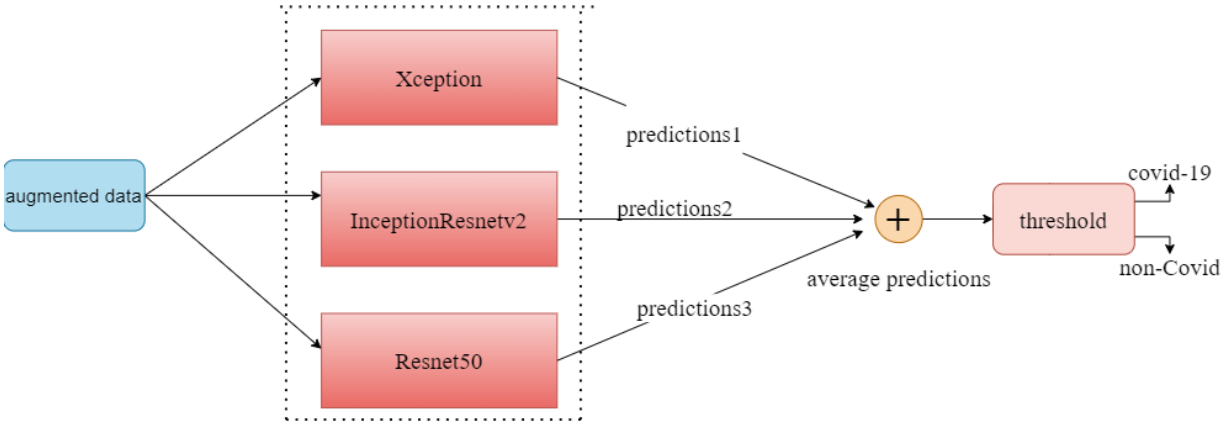


Figure 3.19: The ensemble model which combines the predictions of the three pretrained models on ImageNet.

3.3.5 Temporal Convolutional Recurrent Neural Networks

This model differs from the previous models in that the input is the full cough sounds themselves, thereby preserving much of the information in the data. The model was implemented in the Early Detection of COVID-19 from Cough Sounds, Symptoms, and Context[62]. The input was the raw .wav file of the denoised cough audio files. This input was fed into a Temporal Convolutional Recurrent Neural Network, termed TCRNN with a combined CNN and LSTM architecture, outputting a binary classification of COVID-19 positive or negative. In more detail, the data is first transformed into a spectrogram, a temporal sequence of spectra. As in images, neighboring spectrogram bins of natural sounds in time and frequency are correlated; but in sound production, so are harmonics, or frequencies that are multiples of the same base frequency. Therefore, the short-time Fourier transform (STFT) of the signal is obtained in order to

calculate the spectrogram, which serves as the features for our model. CNNs and RNNs both have their respective advantages and disadvantages in audio classification. CNNs have a fixed receptive field, which can be limiting but also modified, while RNNs can in theory utilize an unlimited temporal context, but in practice may require modifications to achieve this. Ideally, CRNNs offer the best of both words by using the convolutional layers to extract local information, and the recurrent layers to combine it over a longer temporal context. This classification model employs an LSTM to better capture long term temporal dependencies. In summary, the CNN takes the spectrogram as input and consists of a sequence of 2D convolutions, followed by a bi-directional LSTM. A maxpool layer was removed to accommodate the data size**.** Moreover, the CNN consists of three convolutional blocks followed by a batch normalization and a max pooling. The bidirectional LSTM contains 2 layers with hidden size equal to 64. The model structure is illustrated in Figure 3.20.

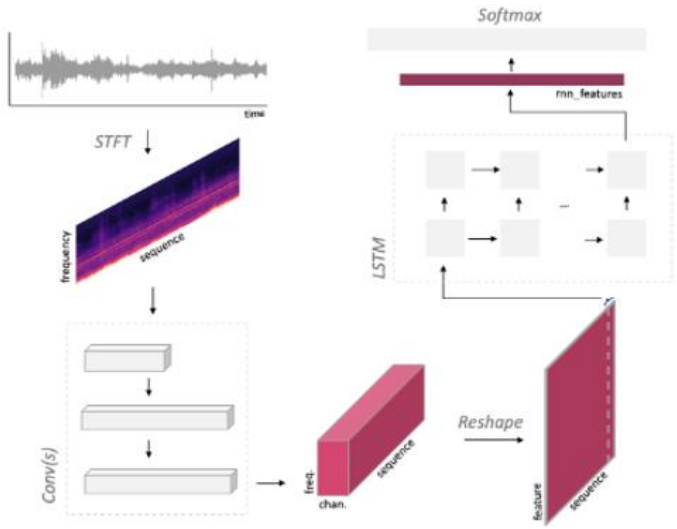


Figure 3.20: TCRNN step-by-step. Firstly, the STFTs are calculated from the raw denoised audio files. Then, a 3 block convolutional network is applied. Images are reshaped and fed into a bidirectional LSTM that captures the longer temporal context.

3.3.6 VGG-13

VGG-13 network is structured as a series of layers, including convolutional layers and pooling layers. Compared with traditional feature extraction methods convolutional layers can automatically extract features from data [9]. Following convolutional layers, max pooling layers, which compute the maximum of a local patch of units in one feature map, are added to reduce the dimension of representation and create invariance to small translations or rotations. Particularly, all 13 convolutional layers contained in VGG-Base adapt 3 × 3 kernel size. VGG16 contains 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers; however in VGG-13 the 3 fully-connected layers and the last max-pooling layer are removed. Each convolutional block consists of two convolutional layers followed by a max pooling layer that halves each spatial dimension. After each convolution, which uses the ReLU activation function, batch normalization is applied as a form of regularization. After the convolutional blocks, each channel is averaged to a scalar value. Finally, a softmax layer is used to generate the predictions [137]. The input images of mel spectrograms have size (128,64,3). The modifications of the VGG-13 base model that have been done include the introduction of batch normalization after each convolution, one dense layer instead

of two because the datasets are prone to overfitting, the replacement of the last max pooling layer with an average pooling and the reduction of the number of units of the last dense layer from 4096 to 1024 for the two datasets of Coswara repository and to 2048 for the Cambridge Dataset. The batch size is set to 20 for the Coswara Cough Heavy and Shallow Datasets and 32 for the Cambridge Dataset and the epochs are 13, 13 and 150 respectively. The architecture used is shown in Figure 3.21 while the original architecture proposed in the paper[9] is shown in Figure 3.22.

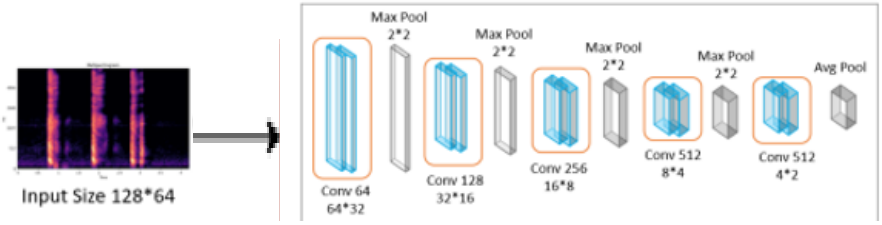


Figure 3.21: The VGG-13 architecture used

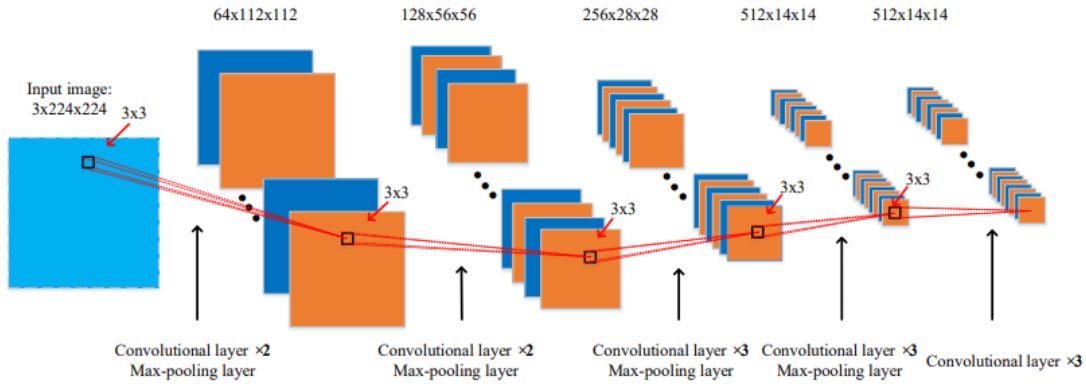


Figure 3.22: The VGG-13 architecture proposed in [9]. It contains 13 convolutional layers and 4 max-pooling layers.

3.3.7 CRNN with an attention mechanism and Bi-directional LSTM

The attention based hybrid CNN-LSTM architecture was introduced in [10]. The architecture can be divided into three blocks. The first block uses a CNN architecture, which receives augmented mel-spectrograms as input of shape (39×88×3). Then, the most relevant and informative features are extracted by the convolution layers. In the second block, Attention-LSTM feature maps are passed to LSTM block, where the deep features that have high temporal correlation are selected to be passed to the attention block in order to capture more useful patterns. In the third block, a simple fully connected layer is used for feature learning and classification.

The attention mechanism enables neural networks (NNs) to select the portions in speech that are more likely to contain keywords, while ignoring the irrelevant parts. Soft attention is introduced to automatically learn how to describe the speech content. Firstly, it learns a scalar score as

$$e_t = v^T tanh(Wh_t + b)$$

where, h_t is the hidden states. Then softmax is applied to compute the normalized weight as

$$\alpha_t = \frac{\exp(e_t)}{\Sigma_{j=1}^{T} \exp(e_j)}$$

where, α_t stands for the attention score, and is applied to further extract content of the feature-maps from BiLSTM layers [138], [5].

The modification that has been made in this architecture is the reduction of convolutional blocks from four to three. Additionally, the Exponential Linear Unit has been used as an activation function instead of ReLU. In contrast to ReLUs, ELUs have negative values which allows them to push mean unit activations closer to zero like batch normalization but with lower computational complexity [139]. The batch size is 256 in all three datasets, but the LSTM units are 128, 64, 32 for the Coswara cough heavy, the Coswara cough shallow and the Cambridge dataset respectively. The dropout rate is set to 0.5 and the neurons of the last dense layer are 50. The optimizer is Adam and the learning rate is 0.001. Moreover, the number of epochs are 100, 30, 100 for the aforementioned datasets respectively. Figure 3.23 shows the above structure in detail.

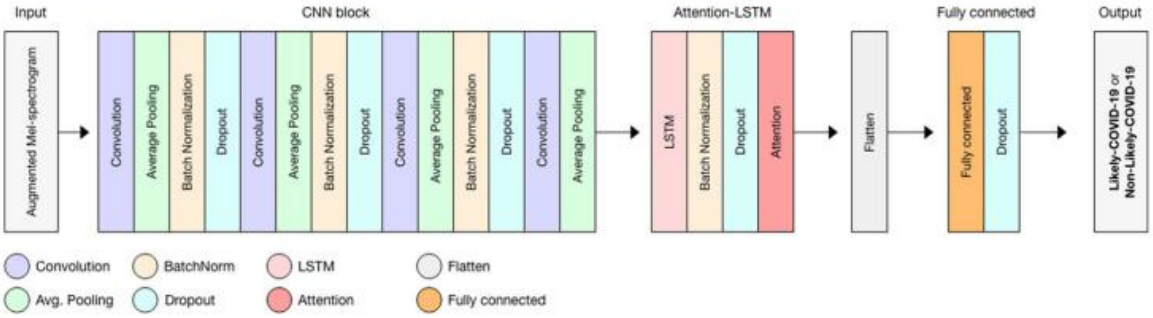


Figure 3.23: Structure of the proposed Attention Hybrid CNN-LSTM architecture [10].

3.3.8 CRNN with an Attention Mechanism and BiGRU

This model was inspired from the architecture described in 3.1.7 but here instead of LSTM, GRU units have been used. The idea for the GRU units came from a paper for a music auto tagging task proposed in [140] which combined with a CNN achieved remarkable results. The architecture is shown in Figure 3.25. The GRU units are equal to 256 for all the datasets and the batch size is set to 256, 32, 32 for the Coswara cough heavy, the Coswara cough shallow and the Cambridge dataset respectively. The model is trained for 100, 20 and 100 epochs for the aforementioned datasets respectively. The best optimizer for this task is Adam. The learning rate scheduler has been used, with the initial learning rate being 0.002 and it drops to half of its value every 10 epochs.

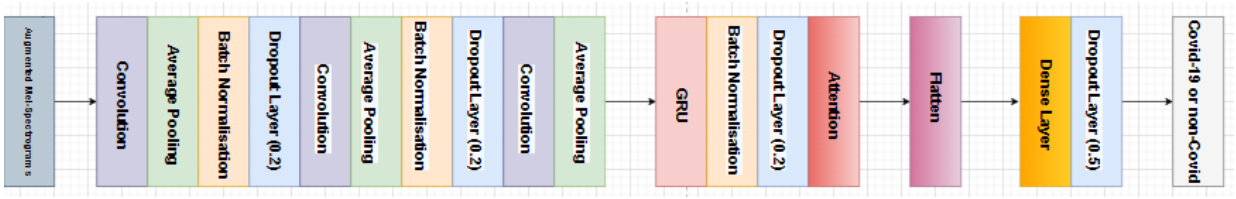


Figure 3.24: Structure of the proposed Attention Hybrid CNN-GRU architecture.

3.3.9 DenseNet Speech

The effective combination of DenseNet Speech and BiLSTM was introduced by Mengjun Zeng and Nanfeng Xiao for keyword spotting[5]. The DenseNet is primarily applied to obtain local features, while the BiLSTM is used to grab time series features. The DenseNet Speech architecture that they proposed is different from the original DenseNet, as they removed the pool on the time dimension in transition layers to preserve speech time series information. In addition, DenseNet-Speech uses less dense blocks and filters to keep the model small, thereby reducing time consumption. In DenseNet-Speech, they only performed a convolution operation along the time dimension to get the basic feature-maps of the time series. The convolution has a kernel of 5×1, without any pooling operation. The DenseNet Speech that is used for the purposes of Covid-19 classification task is detailed in Table 3.1. There is no use of the BiLSTM and attention layer that they used and the model was built from scratch. The growth rate is set to 10, this determines the number of feature maps output into individual layers inside dense blocks. The transition layers aggregate the feature maps from a dense block and reduce its dimensions. Here, we have two transition layers with 1×2 average pooling enabled. The Reshape Layer is substituted with a flatten layer, followed by a dense layer with 128 neurons. As for the hyperparameters, batch size is 16. Adam optimizer is utilized for training with a learning rate of 0.0001. All of the datasets achieve higher results when trained for 20 epochs. Figure 3.26 illustrates the model architecture.

Table 3.1: DenseNet-Speech Architecture. There are 3 dense blocks. The growth rate k = 10. Each 'conv' shown on the table corresponds the sequence BN-ReLU-Conv.

LAYERS	OUTPUT SIZE	DENSENET SPEECH
Convolution	196 × 200 × 10	5 × 1 × conv(10)
Pooling	98 × 100 × 10	2×2 average pooling, stride 2×2
Dense Block (1)	98 × 100 × 10	$\begin{bmatrix} 1x\ 1conv(40) \\ 3x\ 3conv(10) \end{bmatrix} \times 6$
Transition Layer (1)	98 × 100 × 10	1 × 1 × conv(10)
	98 ×50 × 10	1×2 average pooling, stride 1×2
Dense Block (2)	98 ×50 × 10	$\begin{bmatrix} 1x\ 1conv(40) \\ 3x\ 3conv(10) \end{bmatrix} \times 6$
Transition Layer (2)	98 ×50 × 10	1 × 1 × conv(10)
	98 ×25 × 10	1×2 average pooling, stride 1×2
Dense Block (3)	98 ×25 × 10	$\begin{bmatrix} 1x\ 1conv(40) \\ 3x\ 3conv(10) \end{bmatrix} \times 6$
Flatten	(None, 24500)	
Dense	(None, 128)	
Dense	(None,2)	

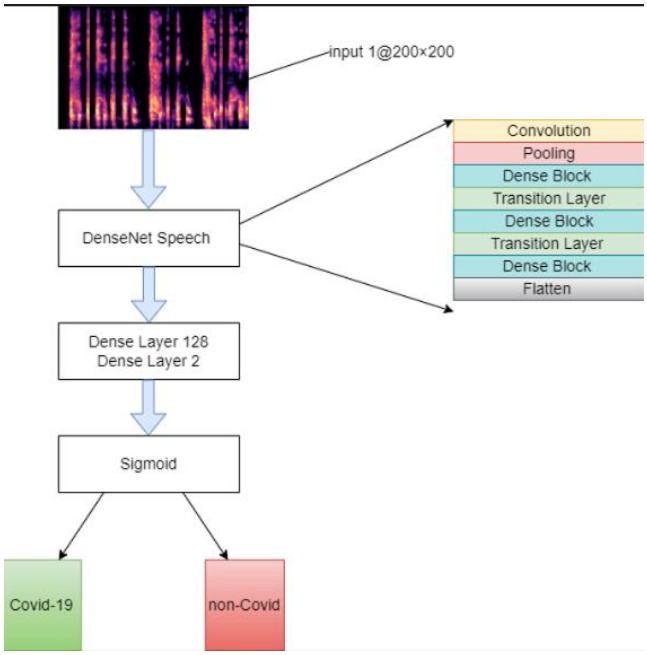


Figure 3.25: The Dense net speech model

3.4 Implemented Methods

3.4.1 5-fold Cross validation

Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting [48] . To avoid it, a test set should be held out. Moreover, a validation set is needed so as to tackle the risk that knowledge about the test set can "leak" into the model and evaluation metrics no longer report on generalization performance. Finally, a solution to the bias that would occur from a random selection of train and validation set is k fold cross validation. We have chosen k =5. The dataset is split into 5 folds, with each one containing 20% of the total dataset and then the model is trained on the k −1 folds, while one fold is left to test a model. This procedure is repeated 5 times. The above procedure is followed for the Cambridge dataset but for the Coswara dataset (cough shallow and cough heavy), a test set (20% of the whole dataset) is held out from the beginning and then the training data are split according to the 5 fold cross validation principles (as shown in Figure 3.27). Both datasets demonstrate the singularity of same users (i.e. users linked to multiple recordings). Training and testing on the same set of users can give horribly misleading results that will not predict out of sample performance on new users. Training on multiple records/observations from the same user/subject is accepted, but test data must be independent of the training data. There is the need for a subject-wise cross validation strategy [141] (Figure 3.28). CSV files were created where each recording was linked to a unique user id. Hence, GroupShuffleSplit iterator is used. It behaves as a combination of ShuffleSplit and LeavePGroupsOut, and generates a sequence of randomized partitions in which a subset of groups are held out for each split. GroupShuffleSplit tends to repeat the same splits as the splits are chosen after shuffling the data. This repetitive behavior may generalize the model (Figure 3.29). Furthermore, grid search for the random state

seed has effectively accomplished that the proportion of classes at each fold is same for most of the folds as the initial dataset (non-Covid/Covid 2:1).

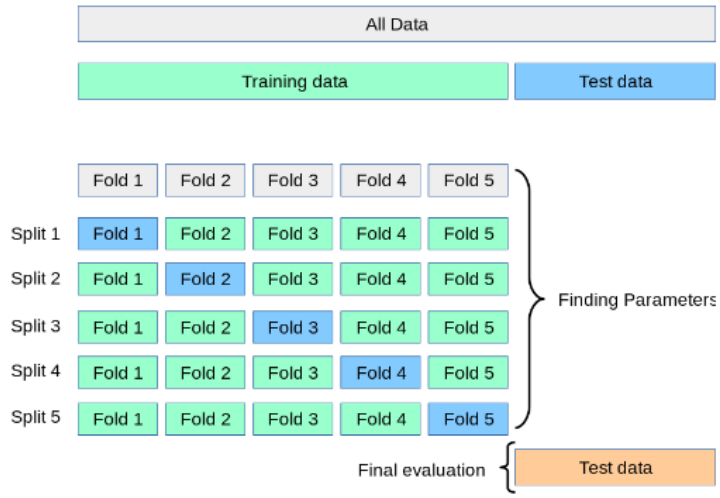


Figure 3.26: 5-fold cv for the Coswara cough heavy and Coswara cough shallow datasets

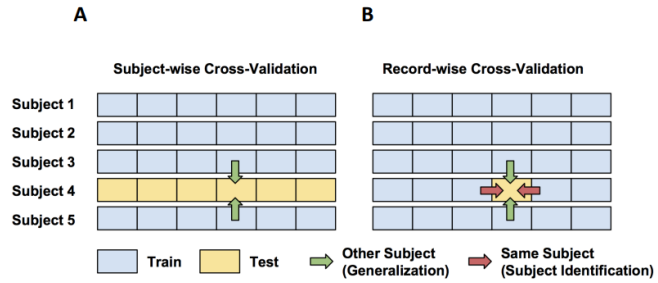


Figure 3.27: Subject-wise/record-wise cross validation

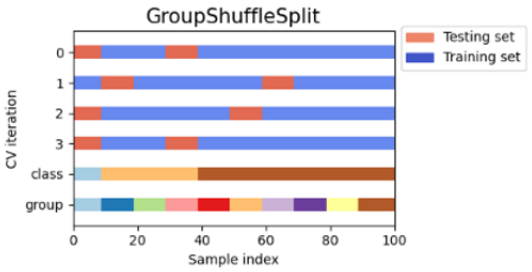


Figure 3.28: GroupShuffleSplit

3.4.2 Handling Class Imbalance

In general, the imbalanced dataset is a problem often found in health applications. In clinical data classification, the imbalanced number of data samples, where at least one of the classes constitutes only a very small minority of the data, occurs very often[142]. To reduce the overfitting, appropriate testing and training datasets should be created. A sub-sampling approach similar to the one developed by Zarkogianni et al. was implemented [143]. In Cambridge, Coswara cough heavy and Coswara cough shallow datasets, the Covid-19 class is underrepresented and the ratio between non-Covid and Covid-19 subjects is 2:1. To address this issue, an ensemble learning method was applied. Firstly, the dataset is split into training, validation and testing sets, as described in 3.2.1. Training set contains 60% of the data, while testing and validation contain 20% each. Then, data augmentation is applied to training set only for both classes. The validation and testing sets remained unchanged retaining the original distribution of samples in the two classes (i.e. 2:1). Subsequently, a balanced sub-sampling approach was adopted, where training sub-sets were generated, preserving a 1:1 ratio between the majority (non-Covid) and the minority (Covid-19) class [144]. The non-Covid class of the training set was split into two equally sized sets and was merged with the whole minority class samples. Hence, two balanced subsets were created containing half

of the original non-Covid samples and the whole Covid-19 samples. These subsets are trained individually on the models described in section 3.3, with the architectures being identical in terms of hyperparameters. The classification results are acquired from averaging the probabilities of the two subsets. Of course, the final evaluation metrics are a result of this procedure combined with 5-fold cross validation. The ensemble learning procedure is detailed in Figure 3.30. Figure 3.31 shows the split of the dataset and how the two subsets are fed into deep learning architectures.

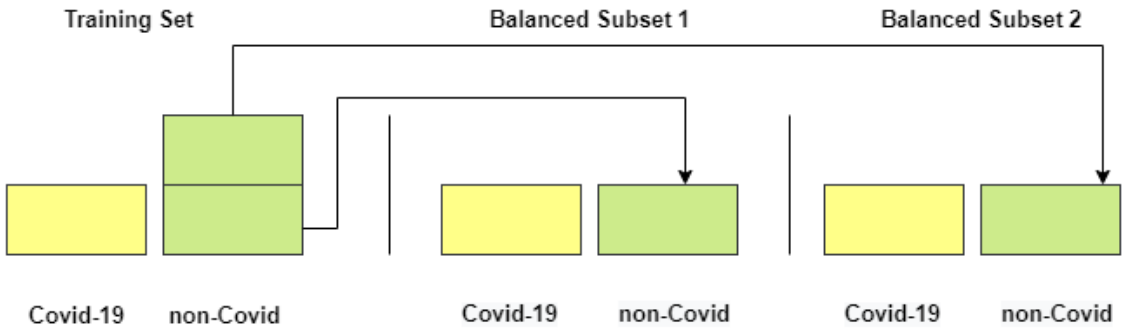


Figure 3.29: Training set divided into two balanced subsets.

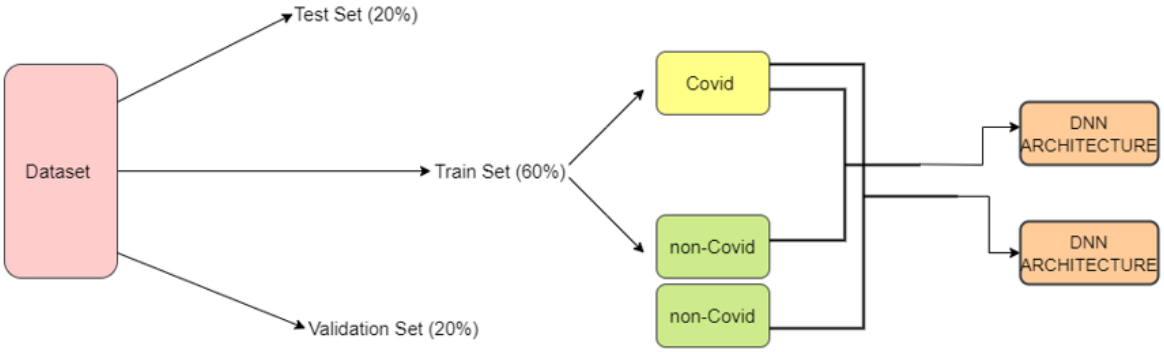


Figure 3.30: The dataset split into training/testing/validation sets and the ensemble method through which the classification results are acquired.

The class imbalance of the datasets in multistage transfer learning, which will be described thoroughly in section 3.4.3, is handled with two oversampling techniques. Namely these are: Synthetic Minority Oversampling TEchnique (SMOTE) and Random Oversampling. Both techniques have been used in Covid-19 related tasks as well [53].SMOTE's main idea is to create new minority class examples by interpolating several minority class instances that lie together[145]. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier. SMOTE performs better than simple oversampling with structured data, but not always in image classification tasks and a practical explanation of this is that SMOTE is applied on flattened images, therefore the localized information obtained from convolutions is lost. For each sample from the minority class (x) samples from the minority class with the smallest Euclidean distance from the original sample were identified (nearest neighbors), and one of them was randomly chosen (x^R). The new synthetic SMOTE sample was defined as:

$$S = x + u \cdot (x^R - x),$$

where u was randomly chosen from U(0, 1). u was the same for all variables, but differed for each SMOTE sample; this choice guarantees that the SMOTE sample lies on the line joining the two original samples used to generate it[146]. Synthetic Minority Oversampling Technique (SMOTE) algorithm applies KNN approach where it selects K nearest neighbors, joins them and creates the synthetic samples in the space.

The algorithm takes the feature vectors and its nearest neighbors, computes the distance between these vectors. The difference is multiplied by random number between (0, 1) and it is added back to feature. Since the Covid class represents the 50% of the non-Covid, the k_neighbors parameter is set to 5. Random oversampling tries to balance class distribution by randomly replicating minority class instances. However, several authors agree that this method can increase the likelihood of overfitting occuring, since it makes exact copies of existing instances[145]. The python library imblearn is used for the implementation of both techniques.

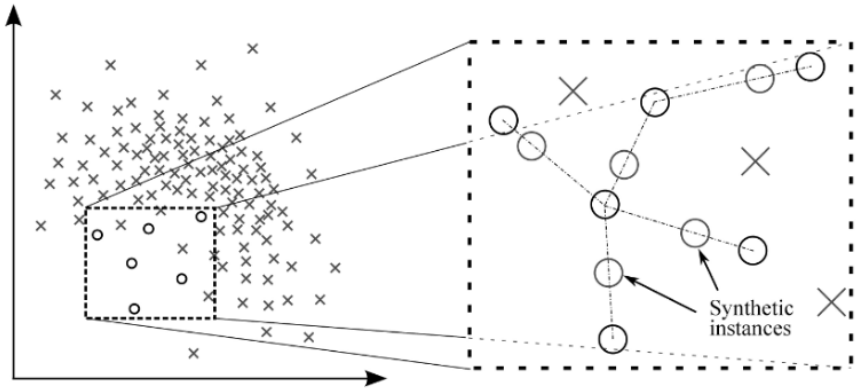


Figure 3.31: SMOTE

3.4.3 Multistage Transfer Learning

The multistage transfer learning architecture was inspired from an application on an ultrasound breast cancer image classification[147]. Recently, multistage transfer learning (MSTL), where a model pre-trained on a large dataset is further pre-trained on a given domain with a relatively small dataset size compared to ImageNet, before fine-tuning it on a given target task with a much smaller dataset size, has become popular. Their proposed MSTL method involves TL from an ImageNet (dataset containing 1000 categories and 1.2 million images) pre-trained model to cancer cell line microscopic images (dataset containing three categories and 20,400 images), which is in turn used as a pre-trained model for TL on US breast cancer images (200 Mendeley and 400 MT-SmallDataset images) to classify them as malignant or benign. MSTL was implemented using three pre-trained models: EfficientNetB2, InceptionV3, and ResNet50. For the purposes of the current analysis, the MSTL procedure involves 3-stage Transfer Learning and the three pretrained models are Xception, InceptionResnet-v2, ResNet50 as shown in Figure 3.18 . No data augmentation is used, just the mel spectrograms of the denoised audio recordings. In the first stage, we applied TL from ImageNet to Coswara cough heavy dataset which contains the most samples among Coswara cough shallow and Cambridge. In the second stage, we utilized the first-stage TL as a starting point and assigned weights to the model that classifies mel spectrograms images as Covid-19 or non-Covid by applying TL from Coswara cough heavy to Coswara cough shallow dataset. In the third stage, (i.e. TL from Coswara cough shallow to Cambridge dataset), we used the previous stages as starting points and acquired the classification results. The objective of our MSTL task is to benefit from knowledge acquired through learning at different stages of TL. More specifically, with regards to the model, the same protocol was utilized for all three CNN models, Xception, InceptionResnet-v2, and ResNet50, at each stage of transfer learning. XCeption's last fully connected layer was replaced by a dense layer consisting of 256 neurons with softmax activation function and a 0.3 dropout rate before being finally connected to a single neuron, responsible for the binary classification. The optimizer is Adam and the learning rate is set to 0.0001. InceptionResnet-v2 had its classification head substituted with a dense layer of 512 neurons, a 'relu' activation and a dropout layer with dropout rate equal to 0.5. The optimizer is Adam and the learning rate is set to 0.001. As for the ResNet50 architecture, the neurons of the last dense layer are 256, and the dropout rate is 0.5. The optimizer is Nadam with a learning rate equal to

0.001. All models were trained for 20 epochs. To overcome the class imbalance issue, SMOTE and Random Oversampling are used during training for the Coswara cough heavy and Coswara cough shallow datasets, as described in section 3.4.2. In the first stage, the weights pretrained on ImageNet are loaded using Keras. In the second stage, the weights are initialized to the ones acquired by training the Coswara cough heavy dataset. Moreover, when Coswara cough shallow is trained, grid search is implemented in order to define the optimal weights to sum each classifier's predictions. A weighted ensemble is an extension of a model averaging ensemble where the contribution of each member to the final prediction is weighted by the performance of the model. This led to defining the following as weights [0.2, 0.2, 0.6] for Xception,InceptionResnetv2,ResNet50 respectively, which means that ResNet50 contributes the most to summing the predictions by a factor 0.6. In the third stage, Cambridge dataset is trained using the weights acquired from the training of Coswara cough shallow. Each classifier contributes to the predictions according to the weights defined in the previous stage. Coswara cough shallow and Coswara cough heavy are split into 80% training set and 20% validation set but the Cambridge dataset is split as detailed in Figure 3.33. To overcome the issue of class imbalance in the Cambridge dataset an ensemble learning approach has been used. Two of the architectures described are used with the ensemble predictions of the classifiers being averaged. 5-fold cross validation is implemented in order to define accuracy, precision, sensitivity, specificity, AUC and F1-score of the model. The results on the non-augmented Cambridge dataset have revealed that pretraining on two cough related datasets achieves higher testing results.

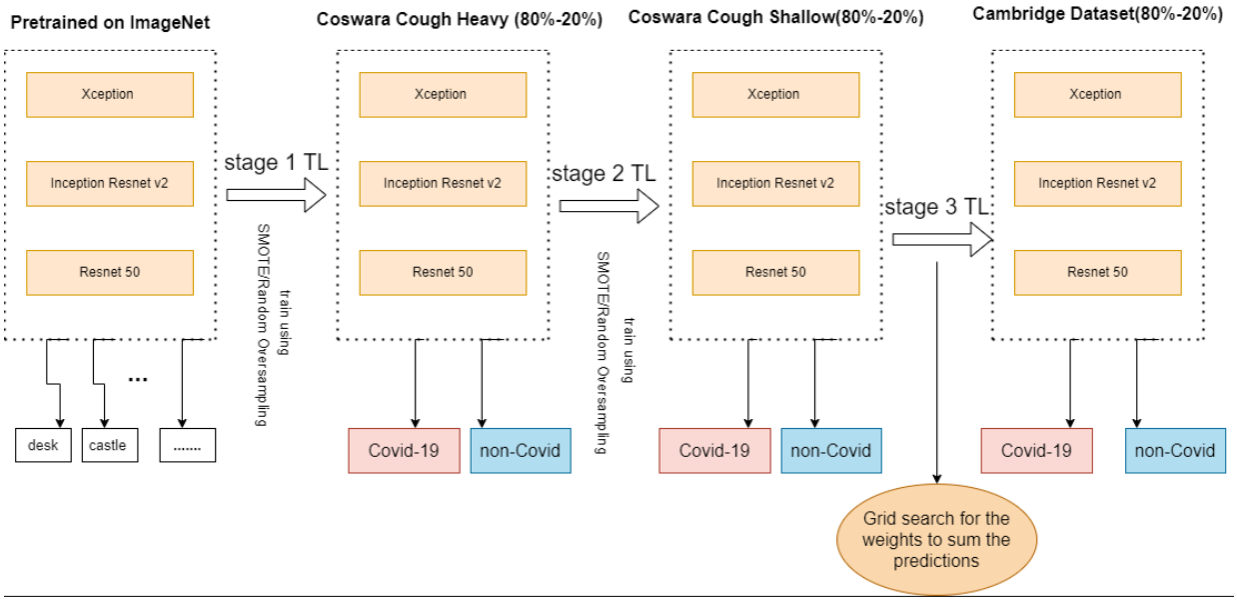


Figure 3.32: The Multistage Transfer Learning architecture.

3.5 Interpretability

Miller in his work [148] defines interpretability as "the degree to which a human can understand the cause of a decision". The widespread adoption of deep learning methods, combined with the fact that it is in their very nature to produce black-box machine learning systems, has led to a considerable amount of experiments and scientific work around them and, therefore, tools regarding their interpretability [80]. In this thesis, the local interpretable model-agnostic explanations (LIME) method has been used. The term

local is used to describe an interpretability method, which explains a single prediction rather than the overall model. A model agnostic tool can be applied to any model. It is used so as to highlight the super-pixels with positive weight towards a specific class, as they give intuition as to why the model would think that class may be present [149] . LIME creates explanations by generating a new dataset of random perturbations (with their respective predictions) around the neighbourhood of input instance, and then fitting a weighted local surrogate model. This local model is usually a simpler model with intrinsic interpretability such as a linear regression model. LIME generates perturbations by turning on and off some of the super-pixels in the image. Additionally, we use the InceptionResnet-v2 model to predict the class of each of the perturbed images. We use a distance metric to evaluate how far is each perturbation from the original image. The original image is just a perturbation with all the super-pixels active (all elements in one). Given that the perturbations are multidimensional vectors, the cosine distance is a metric that can be used for this purpose. After the cosine distance has been computed, a kernel function is used to translate such distance to a value between zero and one (a weight). At the end of this process we have a weight (importance) for each perturbation in the dataset. Finally, we fit a weighted linear model using the information obtained in the previous steps. We get a coefficient for each super-pixel in the image that represents how strong is the effect of the super-pixel in the prediction of Covid-19 or non-Covid class. Then, we sort these coefficients to determine what are the most important super-pixels (number of features) for the prediction of each class. In our task, the number of features is set to 10.

Although an explanation of a single prediction provides some understanding into the reliability of the classifier to the user, it is not sufficient to evaluate and assess trust in the model as a whole. As proposed in [149], a global understanding of the model is given by explaining a set of individual instances. We examined some of the predictions on the test set, and results are presented for a true negative and a false negative case. The results will be demonstrated in Section 4.3.

Chapter 4

Results

4.1 Evaluation of models' performance

The classification results for the models presented in Section 3.3 are acquired, as stated earlier, through 5-fold cross validation. The data imbalance (i.e. the prevalence of the negative – non Covid class at a ration 2 : 1) is handled with ensemble learning. The individual models predict on the five validation folds and on the test dataset. The prediction probabilities are finally obtained for all test samples by averaging the predictions from the two subsets that contain equal number of Covid and non-Covid cases as described thoroughly in 3.4.2. The classification results for the Coswara Cough Heavy, the Coswara Cough Shallow and the Cambridge datasets are shown in Tables 4.1 – 4.3 respectively. The Stacked CNN model is the ensemble model of the three pretrained networks defined in Section 3.3.4 and it outperforms the other models when trained on the Coswara Cough Heavy Dataset. TCRNN, which was originally introduced for environmental sound classification and was modified for the needs of the current task, has a consistent behavior since it is the best model for the Coswara Cough Shallow dataset and the second best for the Coswara Cough Shallow. This could be attributed to the fact that CRNNs take advantage of the convolutional layers so that they extract local information, and the recurrent layers to combine it over a longer temporal context. ResNet50 is the pretrained network on ImageNet that achieves better results compared to the rest pretrained NNs for the Coswara Cough Heavy and Coswara Cough Shallow Datasets, while InceptionResNetV2 outperforms the rest pretrained networks for the Cambridge Dataset. The hybrid CRNN with an attention-based mechanism achieves higher results when combined with a BiLSTM for the Coswara Cough Heavy Dataset and with a BiGRU unit for the other two datasets. The results obtained from VGG13 are obviously better for the Coswara Cough Heavy Dataset than the other datasets. Generally, Coswara Cough Heavy offers better classification results. One apparent reason is the numerical superiority of samples. Secondly, the lack of ability to control for the sounds in the environment and the quality of the microphone affect the classification outcome despite the fact that sounds have been

Table 4.1: Evaluation metrics for the Coswara Cough heavy dataset

Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Xception	65,84	61,82	55,31	48,26	75,13	51,38
InceptionResnetV2	69,24	65	52,1	53,14	76,86	52,6
Resnet50	69,14	65,17	54,62	52,85	77,5	53,72
Stacked CNN	**74,1**	**70,86**	**64,7**	**59,68**	**82,05**	**62,1**
TCRNN	71,1	66,4	66,1	54,8	78,5	59,9
VGG13	70,52	66,9	58,82	54,68	79,14	56,7
CRNN+Att+BiLSTM	69,15	65,08	53,78	52,89	77,28	53,33
CRNN+Att+BiGRU	65	63,81	67,22	47,62	80	55,75
DenseNet Speech	70,25	65,39	35,29	57,53	73,45	43,69

denoised. A heavy cough file probably encodes more information about the cough than a shallow one. This is the reason why in the Coswara's paper [73] confusion matrix heavy cough is predicted more accurately than shallow cough.

Table 4.2: Evaluation metrics for the Coswara Cough Shallow dataset

Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
XCeption	60,61	58,11	52,25	44,27	71,98	47,93
InceptionResnetV2	59,69	58,1	55,86	43,66	72,47	49,01
ResNet50	62,18	60,77	60,36	46,53	75	52,55
Stacked CNN	63,12	61,16	58,55	47,45	74,86	52,42
VGG13	56,87	59,74	72,73	42,78	76,69	53,87
TCRNN	**76,67**	**76,16**	**74,02**	**71,32**	**81**	**72,65**
CRNN+Att+BiLSTM	60,16	59,4	61,63	43,44	75,37	50,96
CRNN+Att+BiGRU	64,1	61,2	55,81	47,06	75,3	51,1
DenseNet Speech	63,67	64,34	73,25	47,37	81,3	57,53

Table 4.3: Evaluation metrics for the Cambridge dataset

Model	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Xception	59,72	58,15	59,09	39,39	76,92	47,27
InceptionResNetV2	62,5	60,95	63,63	42,42	79,49	50,9
ResNet50	57	59,83	72,73	39,02	80,64	50,79
Stacked CNN	62,32	62,12	66,7	48,48	75,75	56,15
VGG13	62,5	61,97	68,18	42,86	81,08	52,63
TCRNN	70,2	52,2	57,93	54,4	72,3	56,1
CRNN+Att+BiLSTM	64	63,86	72,72	44,44	83,33	55,14
CRNN+Att+BiGRU	**62,5**	**64,22**	**77,27**	**43,59**	**84,85**	**55,74**
DenseNet Speech	59,82	59,18	63,63	40	78,37	49,12

On the other hand, the Cambridge Dataset is a much smaller dataset while Deep Learning methods, at the same time, are data hungry. However, the data augmentation combined with the grid search for hyperparameter tuning has shown some promising results. DenseNet Speech architecture has the second best AUC score for the Coswara Cough Shallow dataset, though it did not perform as expected for the other datasets.

4.2 Evaluation Metrics for Multistage Transfer Learning

To handle class imbalance in the Cambridge dataset an ensemble learning approach has been used. Two of the stacked CNN architectures described earlier are used with the ensemble predictions of the classifiers being averaged. 5-fold cross validation is implemented in order to define accuracy, precision, sensitivity, specificity, AUC and F1-score of the model. The results on the non- augmented Cambridge dataset have confirmed that pretraining on two cough related datasets achieves higher testing results. Classification results are presented in Table 4.4. The imbalance handling refers to the previous from Cambridge datasets

techniques. Coswara cough heavy and Coswara Cough Shallow are trained using both Random Oversampling and SMOTE. These techniques are described in detail in Section 3.4.2. The aim of our MSTL task is to benefit from knowledge acquired through learning at different stages of TL. The MSTL procedure involves 3-stage Transfer Learning and the three pretrained models are Xception, InceptionResnet-v2, ResNet50. All models were trained for 20 epochs. In the first stage, the weights pretrained on ImageNet are loaded using Keras. In the second stage, the weights are initialized to the ones acquired by training the Coswara cough heavy dataset. Moreover, when Coswara cough shallow is trained, grid search is implemented in order to define the optimal weights to sum each classifier's predictions. A weighted ensemble is an extension of a model averaging ensemble where the contribution of each member to the final prediction is weighted by the performance of the model. This led to defining the following as weights [0.2, 0.2, 0.6] for Xception,InceptionResnetv2,ResNet50 respectively, which means that ResNet50 contributes the most to summing the predictions by a factor 0.6. In the third stage, Cambridge dataset is trained using the weights acquired from the training of Coswara cough shallow.

Among the two imbalance handling strategies, Random Oversampling offers higher and more consistent results. SMOTE performs better than simple oversampling with structured data, but not always in image classification tasks and a practical explanation of this is that SMOTE is applied on flattened images, therefore the localized information obtained from convolutions is lost. Finally, an AUC of 70,2% and an F1-score of 65,4% is achieved for the Cambridge dataset. A noticeable increase has been accomplished since multistage transfer learning clearly outperformed the ensemble models for the Cambridge dataset. As it can be observed in Figure 4.1 most of the evaluation metrics are increased with MSTL even when compared to the best DNN model which was CRNN with the attention mechanism and BiGRU for the Cambridge dataset.

Table 4.4: Evaluation metrics for the Cambridge dataset after the multistage transfer learning process

Feature	Imblance Handling	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Mel Spectrogram	SMOTE	67,87	65,9	63	54,49	77,3	58,44
	Random Oversampling	**69,34**	**70,2**	**78,03**	**56,3**	**81,56**	**65,4**

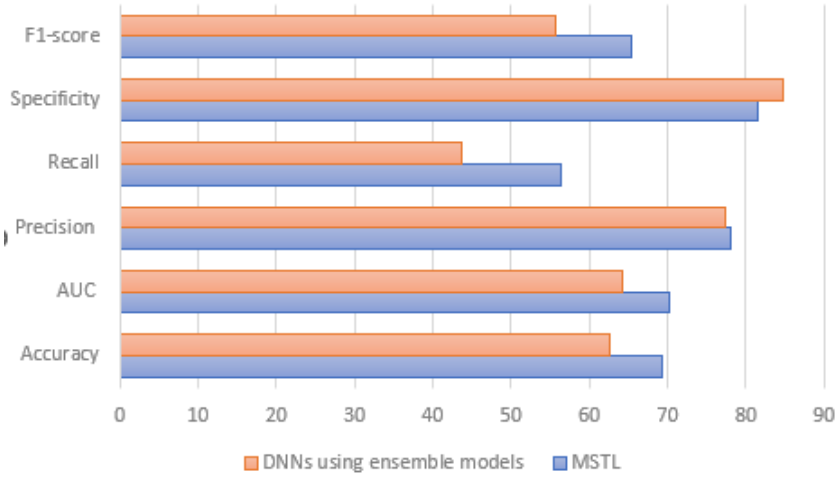
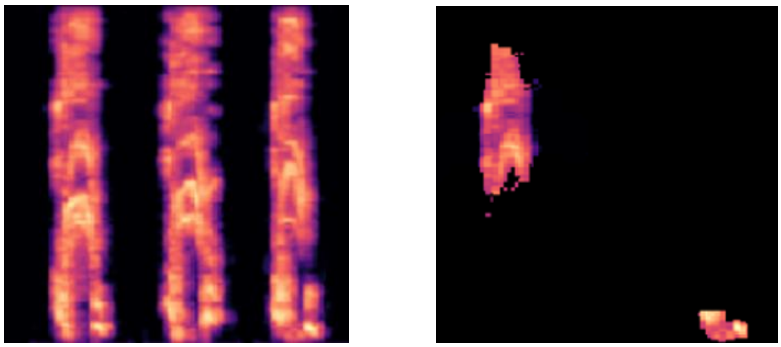


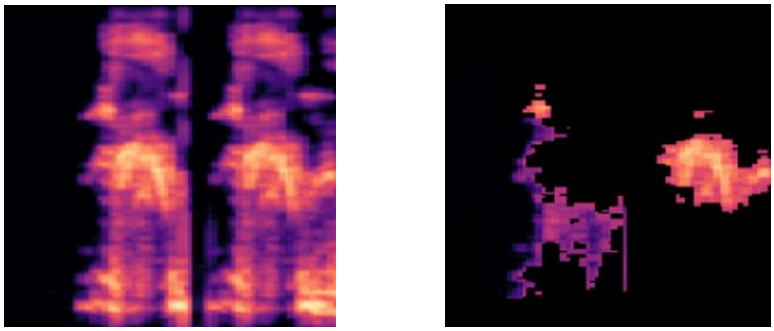
Figure 4.1: Comparison of evaluation metrics of the Cambridge dataset between the best DNN model and Multistage Transfer Learning).

4.3 Prediction Interpretation

Some of the predictions on the test set were examined to qualitatively evaluate InceptionResnetV2 performance for the Coswara Cough Heavy dataset. To this end local interpretable model-agnostic explanations (LIME) method is employed. Fig. 4.1 examples of interpretation are shown for a true negative (a) and a false negative case (b). Here negative stands for non-Covid users. The figure above illustrates what LIME returns as explanation in the image classification prediction**.** 4.2 (a) shows the area of the image (super-pixels) that have a stronger association with the prediction of "non Covid", 4.1(b) shows the super-pixels that have a stronger association with the class 'non Covid' but the image was misclassified. The output of LIME is a list of explanations, reflecting the contribution of each feature to the prediction of a data sample. This provides local interpretability, and it also allows to determine which feature changes will have most impact on the prediction. From the above example we determine that higher decibels play a more important role to classify a mel spectrogram as non-Covid since the mel spectrogram was correctly classified. The false negative example takes in account lower decibels which means that these features have less impact. Figure 4.3 illustrates a true positive case and what LIME returns as an explanation. In Figure 4.3(b) only the super-pixels that are responsible for COVID-19 classification are shown. This means that our model classifies our image as COVID-19 because of these. In Figure 4.3(c) the area of super-pixels colored in green are the ones that increase the probability of our image belongs to COVID-19 class, while the super-pixels colored in red are the ones that decrease the probability. In Figure 4.4 a false positive case is demonstrated and the explanation of LIME.

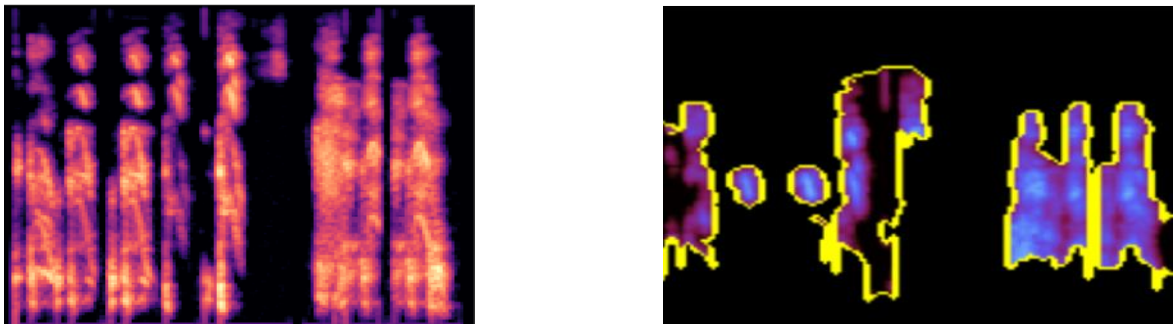


(a)

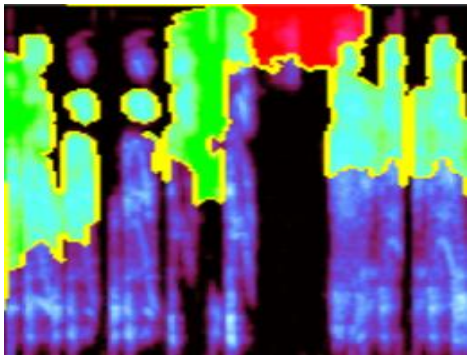


(b)

Figure 4.2 : (Left) the original mel spectrogram and (right) LIME's explanation for (a) a non Covid user which was predicted non Covid, (b) a Covid user which was misclassified as non Covid



(a) (b)



(c)

Figure 4.3: (a) Original mel spectrogram of COVID-19 patient, (b) super pixels that accounted for the COVID-19 class, (c) super pixels that accounted for COVID-19 (in green color) and the ones that decreased the probability in red.

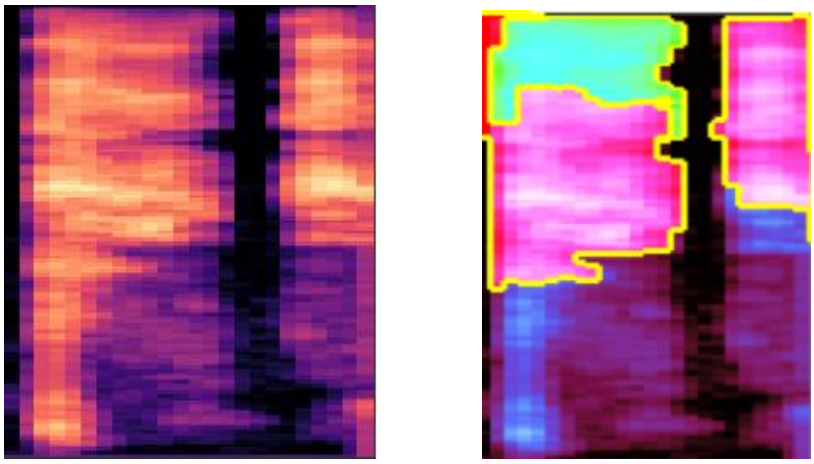


Figure 4.4: (Left) original mel spectrogram of a false positive case, (right) LIME'S explanation

Chapter 5

Conclusion and Future Research

The current thesis aims at demonstrating the feasibility of the automatic detection of COVID-19 from coughs. The core idea of the tool is inspired by prior studies that show cough can be used as a test medium for diagnosis of a variety of respiratory diseases using AI. In summary, a pipeline has been provided for processing the audio recordings, segmenting the cough sound, extracting mel spectrograms and classifying the presence of COVID-19. Three datasets have been used while two of them are derived from the Coswara dataset and the last one is Cambridge dataset. One of the main challenges was overfitting which was tackled with various techniques, such as dropout, L2 regularization and the search for shallower neural networks, for instance VGG13 instead of VGG16 or VGG19, DenseNet speech instead of DenseNet201. Another challenge occurred from class imbalance.

We focused on breadth and explored several possibilities especially for the classification task. To this end nine different deep learning architectures have been tested. Some of them involve hybrid learning techniques which combine CNNs and BiLSTMs or BiGRUs. A stacked model consisting of three pretrained models on ImageNet has achieved great results for the Coswara Cough Heavy dataset, and in particular an accuracy of 74,1%, an AUC of 70,86%, a Precision score of 64,7%, a Recall reaching 59,68%, a specificity at about 82,05% and an F1-score of 62,1%. The results from the Temporal CRNN model demonstrate potential promise on the feasibility of using cough sound to detect COVID-19. Its effectiveness lies on the fact that it employs an LSTM to better capture long term temporal dependencies along with a CNN for local extraction of features. TCRNN has outperformed in the Coswara Cough Shallow dataset. It has accomplished an accuracy of 76,67%, an AUC of 76,16%, a precision of 74,02%, a recall of 71,32%, a specificity of 81% and an F1-score of 72,65% for the Coswara Cough shallow dataset. CRNN with the BiGRU has outperformed in Cambridge dataset achieving an AUC of 64,22%.

Cambridge dataset had limited cough samples and a multistage transfer method was used to improve the classification results. Multistage Transfer Learning (MSTL) was conducted from the dataset containing the most samples (i.e. Coswara Cough Heavy) to the dataset with the fewest (i.e. Cambridge Dataset). MSTL reclaims all available datasets in order to benefit from knowledge acquired through learning at different stages of TL. It is combined with ensemble learning for class imbalance for the Cambridge dataset. After this process, the evaluation metrics and especially AUC, precision and F1-score have remarkable increase confirming the knowledge that an architecture pre-trained on audio samples can provide very promising results in such a classification task. It is concluded that pre-training on two relevant to the task datasets offers a better initialization of the model's weights and so it effectively learns features of the third dataset and provides better testing results. When using random oversampling to handle class imbalance on the two first datasets, the best classification results are obtained. Specifically, an accuracy of 69,34%, an AUC of 70,2%, a precision of 78,03%, a recall of 56,3%, a specificity of 81,56% and an F1-score of 65,4% have been acquired. Although the accuracy is not as high as in certain prior literature works, an essential conclusion is extracted. Firstly, cough can potentially serve as a helpful triage or diagnostic tool for Covid-19 infection. Secondly, it is observed that higher results occur from pretraining on similar cough related tasks or using the weights of audio classifications than using pretrained networks on ImageNet.

Eventually, an interpretability attempt has been made on mel spectrograms using LIME. LIME generates perturbations by turning on and off some of the super-pixels in the image. The InceptionResnet-v2 model

is used to predict the class of each of the perturbed images. The results of a true negative and a false negative case are presented.
The way COVID- 19 affects the respiratory system is substantially unique and hence, cough associated with it is likely to have unique latent features as well. This idea is confirmed from the above results. Cough sound screening tools may not replace testing but they are functional for timely, cost-effective and most importantly safe monitoring, tracing, tracking and thus, controlling the spread of the pandemic by enabling testing for everyone.

Future research could include other vocal modalities available, such as breathing, speech and sustained vowel phonations, in addition to cough analysed in this thesis. Challenges related to disambiguation with other respiratory pathologies with similar symptoms remain to be addressed. Since, deep learning architectures can extract multiple features, feature concatenation could be an effective way to add different features together in order to enhance the classification process. Furthermore, biomarkers could be used as inputs, along with spectrograms, MFCCs or images, into parallel architectures. Biomarkers could contain extra features concerning symptoms of Covid-19. Additionally, the credibility of health status declaration could be enhanced based on the confirmation by the standard RT-qPCR or RAT test, with the date of testing. Many of the available datasets are not annotated by experts and they don't demand a test to submit the health status, which challenges the classification results. The aspects to be considered for improvement range from time complexity, space complexity, data quality, recording, record keeping, transfer learning, and the aspect of transition networks.

Bibliography

[1] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[2] I. H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 6, 2021, doi: 10.1007/s42979-021-00815-1.

[3] M. Z. Alom *et al.*, “A state-of-the-art survey on deep learning theory and architectures,” *Electron.*, vol. 8, no. 3, pp. 1–67, 2019, doi: 10.3390/electronics8030292.

[4] M. Yani, B. Irawan, and C. Setiningsih, “Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry’s Nail,” *J. Phys. Conf. Ser.*, vol. 1201, no. 1, 2019, doi: 10.1088/1742-6596/1201/1/012052.

[5] M. Zeng and N. Xiao, “Effective combination of DenseNet and BiLSTM for keyword spotting,” *IEEE Access*, vol. 7, pp. 10767–10775, 2019, doi: 10.1109/ACCESS.2019.2891838.

[6] X. Liu, Y. Wang, X. Wang, H. Xu, C. Li, and X. Xin, “Bi-directional gated recurrent unit neural network based nonlinear equalizer for coherent optical communication system,” *Opt. Express*, vol. 29, no. 4, p. 5923, 2021, doi: 10.1364/oe.416672.

[7] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-ResNet and the impact of residual connections on learning,” *31st AAAI Conf. Artif. Intell. AAAI 2017*, pp. 4278–4284, 2017, doi: 10.1609/aaai.v31i1.11231.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[9] D. Zeng, S. Chen, B. Chen, and S. Li, “Improving remote sensing scene classification by integrating global-context and local-object features,” *Remote Sens.*, vol. 10, no. 5, pp. 1–19, 2018, doi: 10.3390/rs10050734.

[10] S. Hamdi, M. Oussalah, A. Moussaoui, and M. Saidi, “Attention-based hybrid CNN-LSTM and spectral data augmentation for COVID-19 diagnosis from cough sound,” *J. Intell. Inf. Syst.*, 2022, doi: 10.1007/s10844-022-00707-7.

[11] D. K. Bonilla-Aldana, K. Dhama, and A. J. Rodriguez-Morales, “Revisiting the one health approach in the context of COVID-19: A look into the ecology of this emerging disease,” *Adv. Anim. Vet. Sci.*, vol. 8, no. 3, pp. 234–237, 2020, doi: 10.17582/journal.aavs/2020/8.3.234.237.

[12] NIH, “Origins of Coronaviruses | NIH: National Institute of Allergy and Infectious Diseases.” 2022, [Online]. Available: https://www.niaid.nih.gov/diseases-conditions/origins-coronaviruses.

[13] A. Maxmen, “Wuhan market was epicentre of pandemib’s start, study suggests,” *Nature*, vol. 603, no. January, 2022.

[14] John Hopkins University, “COVID-19 dashboard by the center for systems science and engineering (CSSE).” [Online]. Available: https://coronavirus.jhu.edu.

[15] World Health Organization, “WHO Coronavirus Disease (COVID-19) Dashboard With Vaccination Data | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data,” *World*

Health Organization. pp. 1–5, 2021, [Online]. Available: https://covid19.who.int/%0Ahttps://covid19.who.int/%0Ahttps://covid19.who.int/region/searo/country/bd.

[16] C. Drosten *et al.*, "Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome," *N. Engl. J. Med.*, vol. 348, no. 20, pp. 1967–1976, 2003, doi: 10.1056/nejmoa030747.

[17] N. Petrosillo, G. Viceconte, O. Ergonul, G. Ippolito, and E. Petersen, "COVID-19, SARS and MERS: are they closely related?," no. January, 2020.

[18] WHO, "Naming the coronavirus disease (COVID-19) and the virus that causes it". https://www.who.int/emergencies/diseases/novel-coronavirus-," *Brazilian J. …*, no. February, p. 2019, 2019, [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it.

[19] Y. Alimohamadi, M. Sepandi, M. Taghdir, and H. Hosamirudsari, "Determine the most common clinical symptoms in COVID-19 patients: A systematic review and meta-analysis," *J. Prev. Med. Hyg.*, vol. 61, no. 3, pp. E304–E312, 2020, doi: 10.15167/2421-4248/jpmh2020.61.3.1530.

[20] A. Hasan *et al.*, "A new estimation method for COVID-19 time-varying reproduction number using active cases," *Sci. Rep.*, vol. 12, no. 1, pp. 1–9, 2022, doi: 10.1038/s41598-022-10723-w.

[21] FDA-U.S. FOOD AND DRUGS, "SARS-CoV-2 Viral Mutations_ Impact on COVID-19 Tests _ FDA." .

[22] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth, "How will country-based mitigation measures influence the course of the COVID-19 epidemic?," *Lancet*, vol. 395, no. 10228, pp. 931–934, 2020, doi: 10.1016/S0140-6736(20)30567-5.

[23] D. Visualization, "COVID-19 Deaths by Age," *Heritage.org*. 2022, [Online]. Available: https://www.heritage.org/data-visualizations/public-health/covid-19-deaths-by-age/.

[24] Ecdc, "SARS-CoV-2 variants, spike mutations and immune escape," *Nat. Rev. Microbiol.*, vol. 19, no. 7, pp. 409–424, 2021, doi: 10.1038/s41579-021-00573-0.

[25] ECDC, "SARS-CoV-2 variants of concern as of 15 July 2022." 2022, [Online]. Available: https://www.ecdc.europa.eu/en/covid-19/variants-concern.

[26] Á. O'Toole *et al.*, "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool," *Virus Evol.*, vol. 7, no. 2, pp. 1–9, 2021, doi: 10.1093/ve/veab064.

[27] WHO, "Infection prevention and control in the context of coronavirus disease (COVID-19): A living guideline," *World Heal. Organ.*, no. April, pp. 1–74, 2022.

[28] C. Dye, "The benefits of large scale covid-19 vaccination," *BMJ*, pp. 1–2, 2022, doi: 10.1136/bmj.o867.

[29] M. Al-Kassim Hassan, A. Adam Bala, and A. I. Jatau, "Low rate of COVID-19 vaccination in Africa: a cause for concern," *Ther. Adv. Vaccines Immunother.*, vol. 10, pp. 1–3, 2022, doi: 10.1177/25151355221088159.

[30] FDA-U.S. FOOD AND DRUGS, "Know Your Treatment Options for COVID-19," *Fda*. 2021.

[31] W. Fischer *et al.*, "Molnupiravir, an Oral Antiviral Treatment for COVID-19.," *medRxiv Prepr. Serv. Heal. Sci.*, 2021, doi: 10.1101/2021.06.17.21258639.

[32] EMA, "COVID-19 treatments: under evaluation | European Medicines Agency." [Online].

Available: https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/treatments-vaccines/treatments-covid-19/covid-19-treatments-under-evaluation#covid-19-treatments-under-rolling-review-section.

[33] D. of V. D. National Center for Immunization and Respiratory Diseases (NCIRD), "COVID-19 Treatments and Medications | CDC," *April 29*. 2022, [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/your-health/treatments-for-severe-illness.html.

[34] Y. Liu, A. A. Gayle, A. Wilder-Smith, and J. Rocklöv, "The reproductive number of COVID-19 is higher compared to SARS coronavirus," *J. Travel Med.*, vol. 27, no. 2, pp. 1–4, 2020, doi: 10.1093/jtm/taaa021.

[35] R. Karia, I. Gupta, H. Khandait, A. Yadav, and A. Yadav, "COVID-19 and its Modes of Transmission," *SN Compr. Clin. Med.*, vol. 2, no. 10, pp. 1798–1801, 2020, doi: 10.1007/s42399-020-00498-4.

[36] O. Filchakova, D. Dossym, A. Ilyas, T. Kuanysheva, A. Abdizhamil, and R. Bukasov, "Review of COVID-19 testing and diagnostic methods," *Talanta*, vol. 244, no. January, p. 123409, 2022, doi: 10.1016/j.talanta.2022.123409.

[37] Pfizer, "Understanding COVID-19 Testing Methods." [Online]. Available: https://www.pfizer.com/news/articles/understanding_covid_19_testing_methods.

[38] R. M. Viner *et al.*, "Systematic review of reviews of symptoms and signs of COVID-19 in children and adolescents," *Arch. Dis. Child.*, vol. 106, no. 8, pp. 802–807, 2021, doi: 10.1136/archdischild-2020-320972.

[39] S. Lopez-Leon *et al.*, "More than 50 long-term effects of COVID-19: a systematic review and meta-analysis," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021, doi: 10.1038/s41598-021-95565-8.

[40] A. Haleem, M. Javaid, and R. Vaisha, "Effects of COVID-19 pandemic in daily life," *Curr. Med. Res. Pract.*, vol. 10, no. January, pp. 78–79, 2020.

[41] A. Imran *et al.*, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics Med. Unlocked*, vol. 20, p. 100378, 2020, doi: 10.1016/j.imu.2020.100378.

[42] S. Walvekar and D. S. Shinde, "Detection of COVID-19 from CT Images Using resnet50," 2020.

[43] M. Yildirim and A. Cinar, "A deep learning based hybrid approach for covid-19 disease detections," *Trait. du Signal*, vol. 37, no. 3, pp. 461–468, 2020, doi: 10.18280/ts.370313.

[44] T. Xia *et al.*, "COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening," *Thirty-fifth Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 2)*, no. NeurIPS, pp. 1–13, 2021, [Online]. Available: https://covid19.who.int/.

[45] R. G. and S. K. Bhatia, "Analysis of MFCC and Multitaper MFCC Feature Extraction Methods," *Int. J. Comput. Appl.*, vol. 131, no. 4, pp. 7–10, 2015, doi: 10.5120/ijca2015906883.

[46] C. Xie, X. Cao, and L. He, "Algorithm of abnormal audio recognition based on improved MFCC," *Procedia Eng.*, vol. 29, pp. 731–737, 2012, doi: 10.1016/j.proeng.2012.01.032.

[47] V. P. Singh, J. M. S. Rohith, and V. K. Mittal, "Preliminary analysis of cough sounds," *12th IEEE Int. Conf. Electron. Energy, Environ. Commun. Comput. Control (E3-C3), INDICON 2015*, no. December 2015, 2016, doi: 10.1109/INDICON.2015.7443512.

[48] Y. A. A. and A. C. V. Swarnkar, U. R. Abeyratne, "Automated algorithm for Wet/Dry cough sounds classification," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 3147–3150, doi: 10.1109/EMBC.2012.6346632.

[49] Q. Zhou *et al.*, “Cough Recognition Based on Mel-Spectrogram and Convolutional Neural Network,” *Front. Robot. AI*, vol. 8, no. May, pp. 1–7, 2021, doi: 10.3389/frobt.2021.580080.

[50] Y. Cai and W. Xu, “The best input feature when using convolutional neural network for cough recognition,” *J. Phys. Conf. Ser.*, vol. 1865, no. 4, 2021, doi: 10.1088/1742-6596/1865/4/042111.

[51] N. Peng *et al.*, “Environment sound classification based on visual multi-feature fusion and GRU-AWS,” *IEEE Access*, vol. 8, pp. 191100–191114, 2020, doi: 10.1109/ACCESS.2020.3032226.

[52] J. Xie and M. Zhu, “Handcrafted features and late fusion with deep learning for bird sound classification,” *Ecol. Inform.*, vol. 52, no. November 2018, pp. 74–81, 2019, doi: 10.1016/j.ecoinf.2019.05.007.

[53] D. A. Rahman and D. P. Lestari, “COVID-19 Classification Using Cough Sounds,” *Proc. - 2021 8th Int. Conf. Adv. Informatics Concepts, Theory, Appl. ICAICTA 2021*, 2021, doi: 10.1109/ICAICTA53211.2021.9640278.

[54] W. Thorpe, M. Kurver, G. King, and C. Salome, “Acoustic analysis of cough,” *ANZIIS 2001 - Proc. 7th Aust. New Zeal. Intell. Inf. Syst. Conf.*, no. February 2017, pp. 391–394, 2001, doi: 10.1109/ANZIIS.2001.974110.

[55] Zhaoyan Zhang, “Respiratory Laryngeal Coordination in Airflow Conservation and Reduction of Respiratory Effort of Phonation,” *J. Voice*, vol. 30, no. 6, pp. 760.e7-760.e13, 2016, doi: https://doi.org/10.1016/j.jvoice.2015.09.015.

[56] Vincent L. Gracco and Anders Lofqvist, “Speech Motor Coordination and Control: Evidence from Lip, Jaw, and Laryngeal Movements,” *J. Neurosci.*, vol. 14, no. 11, pp. 6585–6597, 1994, [Online]. Available: https://about.jstor.org/terms%0Ahttps://www.jstor.org/stable/44486767%0Awww.pnas.orgcgidoi10.1073pnas.0506072102%0Awww.pnas.org/cgi/content/full/%0Ahttps://doi.org/10.1101/851147%0Awww.preprints.org.

[57] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, “A cough-based algorithm for automatic diagnosis of pertussis,” *PLoS One*, vol. 11, no. 9, pp. 1–20, 2016, doi: 10.1371/journal.pone.0162128.

[58] P. I. Chung KF, “Prevalence, pathogenesis, and causes of chronic cough,” *Lancet*, vol. 371, no. 9621, pp. 1364–1374, 2008, doi: https://doi.org/10.1016/S0140-6736(08)60595-4.

[59] M. Cohen-Mcfarlane, R. Goubran, and F. Knoefel, “Novel Coronavirus Cough Database: NoCoCoDa,” *IEEE Access*, vol. 8, pp. 154087–154094, 2020, doi: 10.1109/ACCESS.2020.3018028.

[60] “Sound classification with YAMNet,” *TensorFlow*. 2022, [Online]. Available: https://www.tensorflow.org/hub/tutorials/yamnet.

[61] G. Laput, K. Ahuja, M. Goel, and C. Harrison, “Ubicoustics: Plug-and-play acoustic activity recognition,” *UIST 2018 - Proc. 31st Annu. ACM Symp. User Interface Softw. Technol.*, pp. 213–224, 2018, doi: 10.1145/3242587.3242609.

[62] B. Lange, D. Li, E. Nehoran, E. Tuzhilina, and M. Lu2, “Early Detection of COVID-19 from Cough Sounds, Symptoms, and Context Machine Learning / Signal Processing Sub-Team CS 472: Data science and AI for COVID-19,” p. 12, 2020, [Online]. Available: https://web.stanford.edu/~elenatuz/CS472_report.pdf.

[63] A. Tena, F. Clarià, and F. Solsona, “Automated detection of COVID-19 cough,” *Biomed. Signal Process. Control*, vol. 71, 2022, doi: 10.1016/j.bspc.2021.103175.

[64] L. Orlandic, T. Teijeiro, and D. Atienza, “The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms,” *Sci. Data*, vol. 8, no. 1, 2021, doi: 10.1038/s41597-021-00937-4.

[65] N. Simou, N. Stefanakis, and P. Zervas, “A universal system for cough detection in domestic acoustic environments,” *Eur. Signal Process. Conf.*, vol. 2021-Janua, pp. 111–115, 2021, doi: 10.23919/Eusipco47968.2020.9287659.

[66] C. Bales *et al.*, “Can machine learning be used to recognize and diagnose coughs?,” *2020 8th E-Health Bioeng. Conf. EHB 2020*, 2020, doi: 10.1109/EHB50910.2020.9280115.

[67] R. K. and R. R. F. C. Infante, D. B. Chamberlain, “Classification of voluntary coughs applied to the screening of respiratory disease,2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),” pp. 1413–1416, 2017, doi: 10.1109/EMBC.2017.8037098.

[68] S. Vhaduri, T. Van Kessel, B. Ko, D. Wood, S. Wang, and T. Brunschwiler, “Nocturnal cough and snore detection in noisy environments using smartphone-microphones,” *2019 IEEE Int. Conf. Healthc. Informatics, ICHI 2019*, 2019, doi: 10.1109/ICHI.2019.8904563.

[69] T. Drugman *et al.*, “Objective study of sensor relevance for automatic cough detection,” *IEEE J. Biomed. Heal. Informatics*, vol. 17, no. 3, pp. 699–707, 2013, doi: 10.1109/JBHI.2013.2239303.

[70] J. Laguarta, F. Hueto, and B. Subirana, “COVID-19 Artificial Intelligence Diagnosis Using only Cough Recordings,” *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 275–281, 2020, doi: 10.1109/OJEMB.2020.3026928.

[71] A. Imran *et al.*, “AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app,” *Informatics Med. Unlocked*, vol. 20, p. 100378, 2020, doi: 10.1016/j.imu.2020.100378.

[72] C. Brown *et al.*, “Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 3474–3484, 2020, doi: 10.1145/3394486.3412865.

[73] N. Sharma *et al.*, “Coswara - A database of breathing, cough, and voice sounds for COVID-19 diagnosis,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 4811–4815, 2020, doi: 10.21437/Interspeech.2020-2768.

[74] M. Pahar, M. Klopper, R. Warren, and T. Niesler, “COVID-19 cough classification using machine learning and global smartphone recordings,” *Comput. Biol. Med.*, vol. 135, no. June, p. 104572, 2021, doi: 10.1016/j.compbiomed.2021.104572.

[75] V. Despotovic, M. Ismael, M. Cornil, R. M. Call, and G. Fagherazzi, “Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results,” *Comput. Biol. Med.*, vol. 138, 2021, doi: 10.1016/j.compbiomed.2021.104944.

[76] CDCVA, “Covid-19 Survey.” 2021, [Online]. Available: https://cdcva.list.lu/.

[77] E. S. Adamidi, K. Mitsis, and K. S. Nikita, “Artificial intelligence in clinical care amidst COVID-19 pandemic: A systematic review,” *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 2833–2850, 2021, doi: 10.1016/j.csbj.2021.05.010.

[78] G. Chaudhari *et al.*, “Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough Audio Samples,” 2020, [Online]. Available: http://arxiv.org/abs/2011.13320.

[79] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,”

no. Ml, pp. 1–13, 2017, [Online]. Available: http://arxiv.org/abs/1702.08608.

[80] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” *Entropy*, vol. 23, no. 1, pp. 1–45, 2021, doi: 10.3390/e23010018.

[81] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A. C. Thanopoulou, and K. S. Nikita, “An explainable XGBoost-based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus,” *Proc. - IEEE 20th Int. Conf. Bioinforma. Bioeng. BIBE 2020*, pp. 859–864, 2020, doi: 10.1109/BIBE50027.2020.00146.

[82] B. Ghoshal and A. Tucker, “Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection,” pp. 1–14, 2020, [Online]. Available: http://arxiv.org/abs/2003.10769.

[83] O. Speck, C. Sarasaen, G. Rose, N. Andreas, S. Ghosh, and K. E. Group, “EXPLORATION OF INTERPRETABILITY,” 2020.

[84] M. Schedl, “Music Similarity and Retrieval Peter Knees Goals of this tutorial,” 2013.

[85] A. Badr and A. Abdul-Hassan, “A Review on Voice-based Interface for Human-Robot Interaction,” *Iraqi J. Electr. Electron. Eng.*, vol. 16, no. 2, pp. 1–12, 2020, doi: 10.37917/ijeee.16.2.10.

[86] P. A.S. *et al.*, “AI in Medical Imaging Informatics: Current Challenges and Future Directions,” *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 7, pp. 1837–1857, 2020, doi: 10.1109/JBHI.2020.2991043.AI.

[87] J. Karhunen, T. Raiko, and K. H. Cho, “Unsupervised deep learning: A short review,” *Adv. Indep. Compon. Anal. Learn. Mach.*, pp. 125–142, 2015, doi: 10.1016/B978-0-12-802806-3.00007-5.

[88] S. Palit, “Studies on Ozone-oxidation of Dye in a Bubble Column Reactor at Different pH and Different Oxidation-reduction Potential.,” *Int. J. Environ. Sci. Dev.*, vol. 1554, pp. 341–346, 2010, doi: 10.7763/ijesd.2010.v1.67.

[89] Y. Bengio, R. Cardin, R. De Mori, P. Cosi, and V. Oberdan, “USE OF MULTI-LAYERED NETWORKS FOR CODING SPEECH WITH PHONETIC FEATURES,” pp. 224–231.

[90] G. S. P. Frasconi, M. Gori, “Local Feedback Multilayered Networks,” *Neural Comput.*, vol. 4, pp. 120–130, 1992.

[91] Y. Lecun, L. Bottou, Y. Bengio, and P. Ha, “LeNet,” *Proc. IEEE*, no. November, pp. 1–46, 1998.

[92] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, “Stacked denoising autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion,” *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.

[93] T. F. Gonzalez, “Handbook of approximation algorithms and metaheuristics,” *Handb. Approx. Algorithms Metaheuristics*, pp. 1–1432, 2007, doi: 10.1201/9781420010749.

[94] J. Dean *et al.*, “Large scale distributed deep networks,” *Adv. Neural Inf. Process. Syst.*, vol. 2, pp. 1223–1231, 2012.

[95] and B. K. Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, “Deep Neural Networks for Acoustic Modeling in Speech Recognition,” 2012, doi: 10.1016/0022-0728(82)80038-7.

[96] Q. V. Le, “Building high-level features using large scale unsupervised learning,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 8595–8598, 2013, doi: 10.1109/ICASSP.2013.6639343.

[97] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” pp. 1–18, 2012, [Online]. Available: http://arxiv.org/abs/1207.0580.

[98] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8689 LNCS, no. PART 1, pp. 818–833, 2014, doi: 10.1007/978-3-319-10590-1_53.

[99] A. Graves, A. R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 3, pp. 6645–6649, 2013, doi: 10.1109/ICASSP.2013.6638947.

[100] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” *30th Int. Conf. Mach. Learn. ICML 2013*, no. PART 3, pp. 2176–2184, 2013.

[101] V. Vanhoucke, M. Devin, and G. Heigold, “Multiframe deep neural networks for acoustic modeling,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 7582–7585, 2013, doi: 10.1109/ICASSP.2013.6639137.

[102] N. S. G. Hinton, A. K. I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” 2014, doi: 10.1016/0010-4361(73)90803-3.

[103] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.

[104] O. Vinyals and Q. Le, “A Neural Conversational Model,” vol. 37, 2015, [Online]. Available: http://arxiv.org/abs/1506.05869.

[105] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, “DRAW: A recurrent neural network for image generation,” *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 2, pp. 1462–1471, 2015.

[106] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA - J. Am. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016, doi: 10.1001/jama.2016.17216.

[107] P. Rajpurkar *et al.*, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” pp. 3–9, 2017, [Online]. Available: http://arxiv.org/abs/1711.05225.

[108] W. Ismail, M. Hassan, H. Alsalamah, and G. Fortino, “CNN-Based Health Model for Regular Health Factors Analysis in Internet-of-Medical Things Environment,” *IEEE Access*, pp. 52541–52549, 2020.

[109] Q. Xue and M. C. Chuah, “New attacks on RNN based healthcare learning system and their detections,” *Smart Heal.*, vol. 9–10, pp. 144–157, 2018, doi: 10.1016/j.smhl.2018.07.015.

[110] T. Amarbayasgalan, J. Y. Lee, K. R. Kim, and K. H. Ryu, “Deep Autoencoder Based Neural Networks for Coronary Heart Disease Risk Prediction,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11721 LNCS, pp. 237–248, 2019, doi: 10.1007/978-3-030-33752-0_17.

[111] R. K. Sevakula, V. Singh, N. K. Verma, C. Kumar, and Y. Cui, “Transfer Learning for Molecular Cancer Classification Using Deep Neural Networks,” *IEEE/ACM Trans. Comput.*

Biol. Bioinforma., vol. 16, no. 6, pp. 2089–2100, 2019, doi: 10.1109/TCBB.2018.2822803.

[112] M. Aslan, M. Unlersen, K. Sabanci, and A. Durdu, "CNN-based transfer learning–BiLSTM network: A novel approach for COVID-19 infection detection," *Elsevier*, no. January, 2020.

[113] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics Med. Unlocked*, vol. 20, p. 100412, 2020, doi: 10.1016/j.imu.2020.100412.

[114] P. Asvestas, S. Golemati, G. K. Matsopoulos, K. S. Nikita, and A. N. Nicolaides, "Fractal dimension estimation of carotid atherosclerotic plaques from B-mode ultrasound: A pilot study," *Ultrasound Med. Biol.*, vol. 28, no. 9, pp. 1129–1136, 2002, doi: 10.1016/S0301-5629(02)00550-1.

[115] S. Golemati, T. J. Tegos, A. Sassano, K. S. Nikita, and A. N. Nicolaides, "Sonographic Images of the," pp. 659–669, 2004.

[116] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," pp. 1–11, 2015, [Online]. Available: http://arxiv.org/abs/1511.08458.

[117] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, 2019, doi: 10.1016/j.neucom.2018.09.038.

[118] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," *ICML Work. Deep Learn. Audio, Speech Lang. Process.*, vol. 28, 2013.

[119] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[120] I. GitHub, "GitHub - iiscleap/Coswara-Data: Data repository of Project Coswara." 2022, [Online]. Available: https://github.com/iiscleap/Coswara-Data.

[121] "GitHub - spl-icsforth_CoughDetection." .

[122] A. Kumar, "Acoustic Intelligence in Machines," no. September 2018, p. 36, 2018, [Online]. Available: www.lti.cs.cmu.edu.

[123] "Tim Sainburg – Noise reduction using spectral gating in python." [Online]. Available: https://timsainburg.com/noise-reduction-python.html.

[124] S. L. Brunton and J. N. Kutz, "Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control," *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. pp. 54–96, 2019, doi: 10.1017/9781108380690.

[125] D. S. Park *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Septe, pp. 2613–2617, 2019, doi: 10.21437/Interspeech.2019-2680.

[126] "librosa — librosa 0." .

[127] "GitHub - iver56_audiomentations_ A Python library for audio data augmentation." .

[128] K. K. Lella and A. Pja, "Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: Cough, voice, and breath," *Alexandria Eng. J.*, vol. 61, no. 2, pp. 1319–1334, 2022, doi: 10.1016/j.aej.2021.06.024.

[129] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," no. June, pp. 248–255, 2010, doi: 10.1109/cvpr.2009.5206848.

[130] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1800–1807, 2017, doi: 10.1109/CVPR.2017.195.

[131] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, 2015.

[132] L. D. Nguyen, D. Lin, Z. Lin, and J. Cao, “Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation,” *Proc. - IEEE Int. Symp. Circuits Syst.*, vol. 2018-May, no. May, 2018, doi: 10.1109/ISCAS.2018.8351550.

[133] M. Pahar *et al.*, “Automatic Tuberculosis and COVID-19 cough classification using deep learning,” 2022, [Online]. Available: http://arxiv.org/abs/2205.05480.

[134] G. Jia, H. Lam, and Y. Xu, “Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19 . The COVID-19 resource centre is hosted on Elsevier Connect , the company ’ s public news and information ,” no. January, 2020.

[135] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, “Ensemble deep learning: A review,” 2021, [Online]. Available: http://arxiv.org/abs/2104.02395.

[136] J. F. Hernández Santa Cruz, “An ensemble approach for multi-stage transfer learning models for COVID-19 detection from chest CT scans,” *Intell. Med.*, vol. 5, no. November 2020, p. 100027, 2021, doi: 10.1016/j.ibmed.2021.100027.

[137] S. Rao, V. Narayanaswamy, M. Esposito, J. Thiagarajan, and A. Spanias, “Deep Learning with hyper-parameter tuning for COVID-19 Cough Detection,” *IISA 2021 - 12th Int. Conf. Information, Intell. Syst. Appl.*, 2021, doi: 10.1109/IISA52424.2021.9555564.

[138] D. Bahdanau, K. H. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[139] D. A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (ELUs),” *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, pp. 1–14, 2016.

[140] K. Choi, G. Fazekas, M. Sandler, and K. Cho, “Convolutional recurrent neural networks for music classification,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 2392–2396, 2017, doi: 10.1109/ICASSP.2017.7952585.

[141] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, “Voodoo Machine Learning for Clinical Predictions,” *bioRxiv*, p. 059774, 2016, doi: 10.1101/059774.

[142] S. Belarouci and M. A. Chikh, “Medical imbalanced data classification,” *Adv. Sci. Technol. Eng. Syst.*, vol. 2, no. 3, pp. 116–124, 2017, doi: 10.25046/aj020316.

[143] K. Zarkogianni, M. Athanasiou, and A. C. Thanopoulou, “Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication,” *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 5, pp. 1637–1647, 2018, doi: 10.1109/JBHI.2017.2765639.

[144] T. Ganitidis, M. Athanasiou, K. Dalakleidi, N. Melanitis, S. Golemati, and K. S. Nikita, “Stratification of carotid atheromatous plaque using interpretable deep learning methods on B-mode ultrasound images,” *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 3902–3905, 2021, doi: 10.1109/EMBC46164.2021.9630402.

[145] W. Feng, W. Huang, and J. Ren, “Class imbalance ensemble learning based on the margin

theory," *Appl. Sci.*, vol. 8, no. 5, 2018, doi: 10.3390/app8050815.

[146] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, 2013, doi: 10.1186/1471-2105-14-106.

[147] G. Ayana, J. Park, J. W. Jeong, and S. W. Choe, "A Novel Multistage Transfer Learning for Ultrasound Breast Cancer Image Classification," *Diagnostics*, vol. 12, no. 1, pp. 1–14, 2022, doi: 10.3390/diagnostics12010135.

[148] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.

[149] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," *NAACL-HLT 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Demonstr. Sess.*, pp. 97–101, 2016, doi: 10.18653/v1/n16-3020.