National Technical University of Athens

School of Electrical and Computer Engineering

Division of Information Transmission Systems and Material Technology

Cough sound analysis using Deep Learning methods for COVID-19 diagnosis

DIPLOMA THESIS

CHRISTINA D. NTOURMA

Supervisor: Konstantina S. Nikita
NTUA Professor

Athens, November 2021

National Technical University of Athens

School of Electrical and Computer Engineering

Division of Information Transmission Systems and Material Technology

Cough sound analysis using Deep Learning methods for COVID-19 diagnosis

DIPLOMA THESIS

CHRISTINA D. NTOURMA

Advisory board: Konstantina S. Nikita
Professor, NTUA

Approved by the review board on 04/11/2021

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

Athens, November 2021

.....
Christina D. Ntourma
Graduate Electrical and Computer Engineering N.T.U.A.

.....
Χριστίνα Δ. Ντούρμα

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

Περίληψη

Τον τελευταίο ενάμιση χρόνο η ανθρωπότητα δοκιμάζεται από τον COVID-19 (**CO**rona**VI**rus **D**isease of 20**19**) ο οποίος οφείλεται στον ιό SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) και μπορεί να προκαλέσει βαριά νόσηση και δυσλειτουργία αρκετών ανθρωπίνων οργάνων, με κάποιους από τους ασθενείς τελικά να καταλήγουν. Παρά τη δημιουργία και ευρεία χρήση των εμβολίων ανά την υφήλιο, δεν έχει επιτευχθεί το απαραίτητο ποσοστό ανοσίας του πληθυσμού ώστε να τερματιστεί η διάδοση της νόσου. Από την αρχή της πανδημίας η συχνή διενέργεια διαγνωστικών test σε μεγάλα τμήματα του πληθυσμού διαδραματίζει καθοριστικό ρόλο στον περιορισμό της διασποράς. Ωστόσο, οι δύο πιο διαδεδομένες μέθοδοι ανίχνευσης, η μοριακή μέθοδος ανάλυσης και η ταχεία ανίχνευση του αντιγόνου του ιού, απαιτούν χρόνο και υψηλό κόστος αποτελώντας τροχοπέδη στην εξέταση μεγάλων πληθυσμιακών τμημάτων. Επιπροσθέτως, η μετακίνηση πιθανών κρουσμάτων σε δομές υγείας για τη διενέργεια των test εμπεριέχει τον κίνδυνο διασποράς του ιού. Η παρούσα διπλωματική εργασία εξετάζει μία διαφορετική μέθοδο ανίχνευσης του COVID-19 η οποία δεν καταναλώνει χρόνο και πόρους και δεν απαιτεί τη μετακίνηση του εξεταζόμενου σε κάποια δομή υγείας. Η συγκεκριμένη μέθοδος εκμεταλλεύεται τα πλεονεκτήματα της Μηχανικής Μάθησης και συγκεκριμένα των Συνελικτικών Νευρωνικών Δικτύων για την ανίχνευση του COVID-19 μέσω αρχείων ήχου βήχα που καταγράφονται με το μικρόφωνο του κινητού τηλεφώνου του χρήστη ή μέσω κάποιας διαδικτυακής εφαρμογής. Τα εν λόγω αρχεία μετατρέπονται σε εικόνες και δίνονται ως είσοδος σε κάποια αρχιτεκτονική Συνελικτικών Νευρωνικών Δικτύων η οποία εκπαιδεύεται για την ταξινόμηση τους σε COVID-19 και όχι COVID-19.

Μία από τις βασικότερες προκλήσεις της ανίχνευσης του COVID-19 μέσω ήχων βήχα έγκειται στο γεγονός ότι ο βήχας αποτελεί σύμπτωμα για πληθώρα ιατρικών παθήσεων μη σχετικών με τον COVID-19. Επιπλέον, τα διαθέσιμα σύνολα δεδομένων δεν είναι ισορροπημένα, με τα δείγματα του COVID-19 να είναι σημαντικά λιγότερα από τα υπόλοιπα. Ταυτόχρονα, τα σύνολα δεδομένων είναι crowd-sourced, δηλαδή ο κάθε χρήστης ηχογραφεί ένα δείγμα βήχα σε κάποια εφαρμογή δηλώνοντας εάν νοσεί ή όχι από COVID-19. Ωστόσο, η χρήση τέτοιου είδους δεδομένων σε συνδυασμό με την πιθανότητα μη ορθής δήλωσης σχετικά με τη νόσηση ή όχι του χρήστη από COVID-19, καθιστούν το συγκεκριμένο πρόβλημα απαιτητικό, με τα δείγματα που παρέχονται να είναι πιθανώς ηχητικά αρχεία χαμηλής ποιότητας, ενώ ταυτόχρονα η πληροφορία σχετικά με το εάν ο χρήστης είναι θετικός στον COVID-19 δε μπορεί να επιβεβαιωθεί. Για αυτό τον λόγο δημιουργείται η ανάγκη διερεύνησης διαφόρων μεθόδων, αρχιτεκτονικών και συνόλων δεδομένων. Για την επίλυση του προβλήματος εφαρμόστηκε η μέθοδος διασταυρούμενης επικύρωσης και συγκεκριμένα μια 5-fold cross validation προσέγγιση, δοκιμάζοντας διαφορετικούς συνδυασμούς συνόλων δεδομένων και αρχιτεκτονικών. Για την εξάλειψη των αρνητικών επιπτώσεων της ανισορροπίας των δεδομένων εφαρμόστηκε η μέθοδος συλλογικής μάθησης, ensemble learning, η οποία συνδυάζει τις προβλέψεις μοντέλων εκπαιδευμένων με διαφορετικά υποσύνολα ενός συνόλου δεδομένων. Δεδομένου ότι οι αρχιτεκτονικές Βαθιάς Μάθησης απαιτούν μεγάλο πλήθος δεδομένων εξετάστηκε η εκπαίδευσή τους με πολλαπλά διαφορετικά σύνολα δεδομένων, το οποίο προσέφερε και τα υψηλότερα αποτελέσματα, με την ακρίβεια του μοντέλου να φτάνει το 71.60%. Τα αποτελέσματα αυτά επιβεβαιώνουν τη δυνατότητα ανίχνευσης του COVID-19 μέσω αρχείων ήχου βήχα, επιβεβαιώνοντας ταυτόχρονα τη δυνατότητα χρήσης της Μηχανικής Μάθησης για την ανίχνευση και άλλων ασθενειών του αναπνευστικού συστήματος, γεγονός που θα μπορούσε να διαδραματίσει καταλυτικό ρόλο στην ταχύτερη αντιμετώπιση μελλοντικών πανδημιών.

Λέξεις Κλειδιά

Ανίχνευση του COVID-19, Ταξινόμηση Βήχα, Βαθιά Μηχανική Μάθηση, Ανάλυση Ήχου, Ανάλυση Εικόνας, Συνελικτικά Νευρωνικά Δίκτυα, Συλλογική Μάθηση, Προ-εκπαιδευμένα μοντέλα, Μηχανική Μάθηση, Πανδημία.

Abstract

COVID-19 (**CO**rona**VI**rus **D**isease of 20**19**), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has been challenging humanity for the past one and a half year. It can cause severe illness and dysfunction in multiple human organs, with many patients finally passing away. Although vaccines have been released and are widely used around the globe, the essential amount of immunity in order for the COVID-19 transmission between people to terminate, has not been reached yet. Ever since the beginning of this pandemic, the frequent testing of large portions of the population played a determinant role in the containment of the spread. However, the two widely used testing methods, Nucleic Acid Amplification Tests (NAATs) and antigen tests are time and fund consuming, obstructing the screening process of large groups of people. Moreover, the transmission of possible cases to health structures involves the risk of contaminating both the personnel and the rest of the patients. The current thesis examines a different screening method, which is both time and cost efficient and does not require the transportation of individuals to health facilities. The method used, leverages the success of Machine Learning and especially Convolutional Neural Networks (CNNs) for the detection of COVID-19 through cough samples recorded by the mobile phone of the user or a web application. The cough samples collected are converted to images and fed into a CNN architecture which is trained to classify them between COVID-19 and non-COVID-19.

One of the main challenges of COVID-19 cough classification lies in the fact that cough is a symptom of multiple non-COVID-19 related medical conditions. Moreover, the high imbalance of the available datasets, with the COVID-19 samples being significantly less than the non-COVID-19 samples and the fact that the datasets are crowd-sourced, are two important factors making the current task demanding. That is due to the entailed difficulty of using non clean data, with a ground truth based on the declarations of the users. More specifically, the samples provided by each user may contain sounds of low quality, while the validity of the information relative to the user being positive or negative to COVID-19 cannot be confirmed. To that end, different methods, architectures and datasets were examined. A 5-fold cross validation approach was used examining different combinations of datasets and architectures. In order to deal with the imbalanced nature of the data, an ensemble learning method was implemented. Since Deep Learning architectures are data "hungry", training them with multiple datasets was also examined, providing the highest classification results with an accuracy of 71.60%. The obtained results certify the ability of detecting COVID-19 infection through cough sounds, but more importantly the ability of using Machine Learning for the diagnosis of respiratory diseases. This could play a determining role in the quicker containment of future pandemics.

Keywords

COVID-19 Screening, Cough Classification, Deep Learning, Audio Analysis, Image Analysis, Convolutional Neural Networks (CNNs), Ensemble Learning, Pre-trained Models, Machine Learning, Pandemics.

Acknowledgments

I would first like to thank my thesis advisor Prof. Konstantina Nikita for trusting me and supporting me with the current thesis. I would also like to deeply thank Dr. Eleni Adamidi and Dr. Kalliopi Dalakleidi for their endless support and guidance throughout the whole thesis. Last but not least, I would like to thank my brother George and my parents for their inexhaustible support over the years.

***Εκτεταμένη* Περίληψη**

COVID-19

Τον τελευταίο ενάμιση χρόνο ολόκληρη η υφήλιος έχει έρθει αντιμέτωπη με μία μεταδοτική ασθένεια, την COVID-19 (**CO**rona**VI**rus **D**isease of 20**19**), η οποία οφείλεται στον ιό SARS-CoV-2 και θα αναφέρεται και ως Covid για λόγους ευκολίας. Μέχρι τη στιγμή της συγγραφής έχουν καταγραφεί 225.680.357 επιβεβαιωμένα κρούσματα και 4.644.740 θάνατοι, σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ) [1]. Ο πρώτος θάνατος αναφέρθηκε στις 11 Ιανουαρίου 2020 και ο SARS-CoV-2 ανακηρύχθηκε σε πανδημία από τον ΠΟΥ στις 11 Μαρτίου 2020 [2].

Οι περισσότεροι από τους μολυσμένους από τον ιό ανθρώπους αντιμετωπίζουν ήπια ή μέτρια αναπνευστική νόσηση και αναρρώνουν χωρίς να χρειαστούν ειδική θεραπεία. Ωστόσο, πολλοί φορείς του ιού μπορεί να νοσήσουν σοβαρά και να χρειαστούν ιατρική περίθαλψη. Παρόλο που τα άτομα με υποκείμενα νοσήματα, όπως χρόνιες αναπνευστικές παθήσεις, καρδιαγγειακές παθήσεις, διαβήτη και καρκίνο, καθώς και οι ηλικιωμένοι, είναι επιρρεπείς στην ανάπτυξη σοβαρών συμπτωμάτων, άτομα από κάθε ηλικιακή ομάδα μπορεί να νοσήσουν σοβαρά ή και να πεθάνουν.

Μέθοδοι μετάδοσης

Η μετάδοση του COVID-19 μπορεί να γίνει με ποικίλους τρόπους. Ωστόσο, η κύρια μέθοδος διάδοσης του είναι μέσω σταγονιδίων που εκκρίνονται από τη στοματική και τη ρινική κοιλότητα του ασθενούς και μεταδίδονται κυρίως μέσω της ομιλίας, της αναπνοής, του βήχα και του φτερνίσματος. Η μόλυνση πραγματοποιείται μέσω της εισπνοής μολυσμένων σταγονιδίων ή σωματιδίων ή μέσω της άμεσης επαφής τους με τα μάτια και τη στοματική και ρινική κοιλότητα ενός ατόμου. Ο ιός μεταδίδεται μεταξύ ανθρώπων που βρίσκονται σε κοντινή επαφή, σε απόσταση περίπου ενός μέτρου. Παρόλα αυτά, τα σταγονίδια του ιού μπορούν να μεταφερθούν και σε μεγαλύτερες αποστάσεις σε περιπτώσεις συνωστισμού ή εσωτερικών χώρων χωρίς καλό αερισμό. Αυτό συμβαίνει διότι οι άνθρωποι τείνουν να παραμένουν για μεγαλύτερο χρονικό διάστημα σε τέτοιους χώρους, προκαλώντας την παραμονή των μολυσμένων σωματιδίων στην ατμόσφαιρα για μεγαλύτερη χρονική διάρκεια. Η έμμεση επαφή με τον ιό μπορεί επίσης να προκαλέσει μόλυνση η οποία πραγματοποιείται από μολυσμένα αντικείμενα τα οποία έρχονται σε επαφή με το στόμα, τη μύτη ή τα μάτια. Σημαντικό είναι και το γεγονός ότι ο ιός μπορεί να μεταδοθεί και από άτομα τα οποία δεν εμφανίζουν συμπτώματα. Τα μολυσμένα άτομα τείνουν να είναι πιο μεταδοτικά λίγο πριν εμφανίσουν συμπτώματα, ενώ όσοι νοσούν βαριά μπορούν να μεταδώσουν τον ιό για μεγαλύτερο χρονικό διάστημα [3].

Συμπτώματα και μακροχρόνιες επιπτώσεις

Τα πιο συνηθισμένα συμπτώματα του COVID-19 περιλαμβάνουν πυρετό, ξηρό βήχα, κούραση, έλλειψη γεύσης ή όσφρησης, ενώ κάποια λιγότερο συχνά είναι ο πονοκέφαλος, ο πονόλαιμος η διάρροια και οι πόνοι. Η σοβαρή νόσηση συνοδεύεται από συμπτώματα όπως δυσκολία στην αναπνοή ή δύσπνοια, πόνος ή πίεση στο στήθος, απώλεια ομιλίας ή κίνησης. Τα συμπτώματα εμφανίζονται 2-14 ημέρες μετά την επαφή με τον ιό και κατά μέσο όρο περίπου στις 5-6 ημέρες. Υπάρχουν περιπτώσεις όπου τα συμπτώματα διατηρούνται για περισσότερο από 3 εβδομάδες, τόσο σε άτομα που νόσησαν βαριά όσο και σε άτομα με ήπια συμπτώματα. Παρόλο που τα πιο συνήθη συμπτώματα που παραμένουν για μεγάλο χρονικό διάστημα είναι η κούραση, η δύσπνοια και οι πόνοι, ο κορωνοϊός μπορεί να προκαλέσει δυσλειτουργία σε διάφορα ζωτικά όργανα όπως την καρδιά, τους πνεύμονες και τον εγκέφαλο [4], [5], [6].

Μέτρα πρόληψης και προστασίας

Κάποια από τα σημαντικότερα μέτρα προστασίας κατά του COVID-19 συνοψίζονται ακολούθως:

- Τήρηση αποστάσεων. Η τήρηση απόστασης μεγαλύτερης του ενός μέτρου, τόσο από άτομα που εμφανίζουν πιθανά συμπτώματα κορωνοϊού, όσο και από φαινομενικά υγιή άτομα (πιθανοί ασυμπτωματικοί φορείς)
- Η χρήση προστατευτικής ιατρικής μάσκας σε περιπτώσεις όπου η τήρηση των απαραίτητων αποστάσεων δεν είναι εφικτή, αλλά και σε εσωτερικούς χώρους.
- Η αποφυγή συνωστισμού, παρατεταμένης επαφής με άλλα άτομα και ανεπαρκώς αεριζόμενων εσωτερικών χώρων.
- Ο τακτικός καθαρισμός των χεριών με σαπούνι και νερό ή με αλκοολούχα αντισηπτικά διαλύματα.
- Ο εμβολιασμός. Ο εμβολιασμός αποτελεί το μοναδικό ισχυρό μέσο περιορισμού της διασποράς του ιού και των διαφόρων μεταλλάξεων του.

Πέρα από τα προσωπικά μέτρα προστασίας που μπορεί να λάβει το κάθε άτομο, οι κυβερνήσεις ανά τον κόσμο, έχουν εφαρμόσει πληθώρα μέτρων για τον περιορισμό της διασποράς στην κοινότητα. Στην αρχή της πανδημίας οι περισσότερες κυβερνήσεις εφάρμοσαν καθολικά lockdowns αναστέλλοντας τις περισσότερες δραστηριότητες, τόσο σε εσωτερικούς όσο και σε εξωτερικούς χώρους. Άλλα προστατευτικά μέτρα που εφαρμόστηκαν ανά διαστήματα αποτελούν μεταξύ άλλων η απαγόρευση κυκλοφορίας κατά τις βραδινές ή και απογευματινές ώρες, η απαγόρευση μετακίνησης από νομό σε νομό, η απαραίτητη επίδειξη αρνητικού αποτελέσματος διαγνωστικών εξετάσεων ή πιστοποιητικού εμβολιασμού για τη συμμετοχή σε διάφορες

δραστηριότητες και η υποχρεωτική χρήση μάσκας σε εσωτερικούς χώρους. Επιπλέον, η απομόνωση κρουσμάτων και η ανίχνευση των επαφών τους εφαρμόζεται από τους αρμόδιους φορείς της εκάστοτε κυβέρνησης από την αρχή της πανδημίας, διαδραματίζοντας σημαντικό ρόλο στον περιορισμό της.

Πρόληψη και θεραπεία

Παρόλη την πληθώρα θεραπευτικών μεθόδων που έχουν εφαρμοστεί από τις αρχές της πανδημίας, δεν έχει βρεθεί ακόμα συγκεκριμένη θεραπεία για το κορωνοϊό. Ωστόσο, αποτελεσματικά εμβόλια που δημιουργήθηκαν από την επιστημονική κοινότητα και πλέον βρίσκονται σε ευρεία κυκλοφορία, συνδράμουν καταλυτικά στην έξοδο από την πανδημία.

Το πρώτο μαζικό πρόγραμμα εμβολιασμού ξεκίνησε το Δεκέμβριο του 2020 και μέχρι τώρα έχει εμβολιαστεί το 42.41% του πληθυσμού του πλανήτη με τουλάχιστον μία δόση, με το 30.25% να είναι πλήρως εμβολιασμένο [7].

Μέθοδοι ανίχνευσης

Η ανίχνευση όσο το δυνατόν περισσότερων κρουσμάτων κορωνοϊού είναι καθοριστικής σημασίας για την ταχεία έξοδο από την πανδημία. Οι δύο πιο διαδεδομένοι τρόποι διάγνωσης του ιού είναι η μοριακή μέθοδος ανάλυσης και η ταχεία ανίχνευση του αντιγόνου του ιού. Οι δύο αυτές μέθοδοι χρησιμοποιούνται για τη διάγνωση τρέχουσας μόλυνσης από τον ιό και πραγματοποιούνται συλλέγοντας δείγματα από τη ρινική ή/και στοματική κοιλότητα του ασθενούς. Όσον αφορά τα τεστ αντιγόνου, αυτά μπορούν να πραγματοποιηθούν τόσο σε εργαστηριακές δομές, όσο και από τον ίδιο τον ασθενή, τα επονομαζόμενα "self-tests", με τα αποτελέσματα να είναι διαθέσιμα εντός 15-30 λεπτών. Παρόλο που αυτό τα καθιστά μια αρκετά εύκολη και γρήγορη μέθοδο διάγνωσης του COVID-19, είναι περισσότερο επιρρεπή στη μη ανίχνευση του ιού σε μολυσμένο ασθενή συγκριτικά με τη μοριακή μέθοδο ανάλυσης που σπάνια επιστρέφει λανθασμένα αρνητικό αποτέλεσμα [8], [9].

Κίνητρο της διπλωματικής

Κατά τη διάρκεια της πανδημίας τα συστήματα υγείας πολλών χωρών δέχτηκαν ανυπέρβλητες πιέσεις, με το υγειονομικό προσωπικό να καταβάλλει υπεράνθρωπες προσπάθειες. Αυτό καθιστά την ιχνηλάτηση και την απομόνωση πιθανών κρουσμάτων, θέμα υψίστης σημασίας. Επιπροσθέτως, καινούριες μεταλλάξεις του ιού τον καθιστούν πιο μεταδοτικό, δυσχεραίνοντας την προσπάθεια αποτροπής νέων κυμάτων της πανδημίας. Η δυνατότητα καθημερινού ελέγχου μεγάλου μέρους του πληθυσμού και ιδανικά ολόκληρου του πληθυσμού μιας χώρας, θα μπορούσε να αποδειχθεί παράγοντας καταλυτικής σημασίας για την έξοδο από την πανδημία. Όπως

προαναφέρθηκε, μέχρι στιγμής δύο είδη διαγνωστικών τεστ χρησιμοποιούνται ευρέως. Ωστόσο, η διεξαγωγή μεγάλου αριθμού τέτοιων τεστ είναι χρονοβόρα και ακριβή. Διαφορετικές μέθοδοι Μηχανικής Μάθησης έχουν τεθεί σε εφαρμογή, με στόχο την επίλυση του προαναφερθέντος προβλήματος [10], [11], [12], [13].

Η συνεχής αύξηση της ποσότητας και του είδους των διαθέσιμων δεδομένων με την πάροδο των χρόνων, οδήγησε στην ευρεία χρήση της Μηχανικής Μάθησης σε διάφορες πτυχές της καθημερινότητας, καθώς και σε ζητήματα σχετικά με την υγεία.

Όσον αφορά τον περιορισμό του COVID-19, η εφαρμογή μεθόδων Μηχανικής Μάθησης έχει εξεταστεί σε βάθος. Οι ακτινογραφίες και οι αξονικές τομογραφίες θώρακος χρησιμοποιούνται ευρέως από τους ειδικούς για την ανίχνευση του ιού, καθώς μέσω αυτών γίνονται εμφανή τα χαρακτηριστικά που διαφοροποιούν ένα ασθενή με COVID-19 από έναν ασθενή με κάποιον άλλο τύπο πνευμονίας [10], [14], [15], [16], [17], [18], [19], [20], [21], [11]. Επιπλέον, μπορεί να προβλεφθεί η σοβαρότητα της νόσησης και η ανάγκη εισαγωγής στη Μονάδα Εντατικής Θεραπείας (ΜΕΘ) [22], [23], [24], [25].

Παρόλη την ακρίβεια των προαναφερθέντων μεθόδων στην ανίχνευση του COVID-19, απαιτούν τη φυσική παρουσία του ασθενή ή του πιθανού κρούσματος σε κάποια δομή υγείας ώστε να μπορέσει να διεξαχθεί η εκάστοτε εξέταση. Εκτός από τον απαιτούμενο χρόνο και την προσπάθεια που πρέπει να καταβληθεί για τη διενέργεια τέτοιου είδους εξετάσεων, είναι σημαντική και η αποφυγή διασποράς του ιού μέσω επαφών ενός πιθανού κρούσματος τόσο με το υγειονομικό προσωπικό της δομής, όσο και με τους υπόλοιπους ασθενείς. Επιπλέον, δε μπορεί να θεωρηθεί δεδομένη η δυνατότητα όλων των κρατών να παρέχουν στους πολίτες τους το απαραίτητο πλήθος διαγνωστικών εξετάσεων, ούτε και το γεγονός ότι ο κάθε πολίτης έχει τη δυνατότητα εύκολης και άμεσης πρόσβασης σε παροχές υγείας. Κατά συνέπεια, μία ταχεία και δωρεάν μέθοδος ανίχνευσης του ιού, η οποία θα είναι διαθέσιμη σε όλους μέσω μιας εφαρμογής, διαδικτυακής ή στο κινητό τηλέφωνο, θα μπορούσε να διαδραματίσει καθοριστικό ρόλο στον περιορισμό της πανδημίας.

Βιβλιογραφική ανασκόπηση

Σύμφωνα με έρευνες που έχουν πραγματοποιηθεί, ο COVID-19 μπορεί να ανιχνευθεί μέσω των αναπνευστικών ήχων. Τα τελευταία χρόνια έχει καταγραφεί αξιόλογη πρόοδος στην ανίχνευση αναπνευστικών παθήσεων μέσω τέτοιων ήχων. Κάποιες από τις μεθόδους που εφαρμόζονται επικεντρώνονται στην εξαγωγή χαρακτηριστικών από τα ηχητικά δείγματα, ενώ σε άλλες γίνεται μετατροπή του ήχου σε εικόνα αξιοποιώντας την αποτελεσματικότητα των Συνελικτικών Νευρωνικών Δικτύων στην ταξινόμηση εικόνων. Πληθώρα μελετών έχει πραγματοποιηθεί σχετικά με την απεικόνιση του ήχου ως εικόνα και οι μετασχηματισμοί που έχουν μελετηθεί μεταξύ άλλων περιλαμβάνουν τους εξής: φασματογραφήματα σε κλίμακα mel ή Mel-spectrograms, Continuous Wavelet Transform (CWT), Short Time Fourier Transform (STFT), Mel Frequency Cepstral Coefficient (MFCC), Constant Q-Transform (CQT) και Hybrid Constant

Q-Transform (HCQT) [26], [27], [28], [29], [30], [31], [32], [33], [34], [35]. Όσον αφορά το πιο συγκεκριμένο πρόβλημα της ταξινόμησης ήχων βήχα, διάφορες προσεγγίσεις έχουν παρουσιαστεί με πολλές από αυτές να επιλέγουν τη μετατροπή του ήχου σε εικόνα για την αξιοποίηση των Συνελικτικών Νευρωνικών Δικτύων και των υψηλών επιδόσεων που μπορούν να επιτύχουν. Τέτοια προβλήματα περιλαμβάνουν τόσο την ανίχνευση ήχων βήχα ανάμεσα σε άλλους περιβαλλοντικούς ήχους, όσο και την ταξινόμηση ήχων βήχα σε διάφορες κατηγορίες που σχετίζονται με αναπνευστικές παθήσεις. Ο μετασχηματισμός των ήχων βήχα σε εικόνα στις περιπτώσεις χρήσης Συνελικτικών Νευρωνικών Δικτύων, έχει πραγματοποιηθεί με διάφορους τρόπους οι οποίοι συμπεριλαμβάνουν μεταξύ άλλων τους STFT, Mel-spectrograms, MFCC και RASTA-PLP [36], [37], [38], [39], [40], [41], [42], [43], [44]. Η αποκτηθείσα γνώση σχετικά με τη μετατροπή του ήχου σε εικόνα και με την ανίχνευση και ταξινόμηση ήχων βήχα, έχει εφαρμοστεί για τη διάγνωση του COVID-19. Η πλειοψηφία των μεθόδων εκμεταλλεύεται την επιτυχία των Συνελικτικών Νευρωνικών Δικτύων και της μεταφοράς μάθησης, χρησιμοποιώντας προεκπαιδευμένα Βαθιά Νευρωνικά Δίκτυα. Στις περισσότερες περιπτώσεις χρησιμοποιούνται μόνο αρχεία ήχου βήχα, ενώ υπάρχουν έρευνες στις οποίες δοκιμάστηκε επιπλέον η χρήση αρχείων ήχου αναπνοής και ομιλίας, αποδεικνύοντας ότι και αυτοί οι ήχοι περιέχουν αρκετή πληροφορία σχετική με τον ιό [12], [45], [46], [47], [48], [49], [13], [50], [51].

Στόχος της διπλωματικής εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι η παρουσίαση μίας μεθόδου ανίχνευσης του COVID-19, αναλύοντας αρχεία ήχου βήχα και χρησιμοποιώντας μοντέλα Βαθιάς Μηχανικής Μάθησης.

Ηχητικά Σήματα

Ο ήχος δημιουργείται από τη διατάραξη των σωματιδίων ενός μέσου διάδοσης, όπως είναι ο αέρας. Η διατάραξη των σωματιδίων μεταφέρεται, μέσω του μέσου διάδοσης, από τον πομπό στον δέκτη. Υπάρχουν διάφορα χαρακτηριστικά του ήχου τα οποία προσφέρουν χρήσιμες πληροφορίες για το εκάστοτε σήμα. Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν χαρακτηριστικά που συνδυάζουν το πεδίο του χρόνου και της συχνότητας. Η πλειοψηφία αυτών των χαρακτηριστικών βασίζεται στον STFT και η απεικόνιση τους ως εικόνα ονομάζεται φασματογράφημα. Οι τρόποι απεικόνισης του ήχου σε εικόνα που χρησιμοποιήθηκαν είναι ονομαστικά οι εξής: Short Time Fourier Transform, Mel Spectrograms, Constant-Q Transform και Hybrid Constant-Q Transform .

Μηχανική Μάθηση

Η Μηχανική Μάθηση προσπαθεί να μιμηθεί τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου, ο οποίος αποτελείται από νευρώνες που επικοινωνούν μεταξύ τους με στόχο τη μετάδοση της πληροφορίας. Με όμοιο τρόπο, ένα Νευρωνικό Δίκτυο αποτελείται από νευρώνες οι οποίοι επικοινωνούν μεταξύ τους, με στόχο την απόκτηση γνώσης για την επίλυση κάποιου προβλήματος. Τα πρώτα στάδια των Νευρωνικών Δικτύων προσδιορίζονται τη δεκαετία του 1940, όπου παρουσιάστηκε το πρώτο υπολογιστικό μοντέλο ενός νευρώνα. Κάποια από τα σημαντικότερα σημεία-σταθμοί για την ανάπτυξη του εν λόγω τομέα αναφέρονται στον πίνακα 1.

Αντικείμενο	Χρονολογία	Συγγραφείς
Εισαγωγή στα νευρωνικά δίκτυα	1943	McCulloch and Pitts [52]
Παρουσίαση του **perceptron**	1959	Rosenblatt [53]
Παρουσίαση του "Adaline"	1960	Widrow and Hoff [54]
Μαθηματικές αποδείξεις για τα **perceptrons**	1969	Minsky and Papert [55]
Αυτο-οργανούμενοι χάρτες **(SOM)**	1982	Kohonen [56]
Δίκτυα **Hopfield**	1982	Hopfield [57]
Εισαγωγή στα **CNNs**	1989	LeCun et al. [58]
Τα καλύτερα αποτελέσματα στο διαγωνισμό **ILSVRC2012**	2012	Krizhevsky et al. [59]

Πίνακας 1: Η εξέλιξη της τεχνητής νοημοσύνης

Η συνεχής αύξηση της ποσότητας των διαθέσιμων δεδομένων παρέχει τη δυνατότητα δημιουργίας και εκπαίδευσης μοντέλων Μηχανικής Μάθησης για την υποστήριξη της λήψης αποφάσεων ιατρικού περιεχομένου. Πιο συγκεκριμένα, ένα μοντέλο Μηχανικής Μάθησης το οποίο θα έχει εκπαιδευτεί σε τεράστιο όγκο δεδομένων, σημαντικά περισσότερα από αυτά με τα οποία θα ερχόταν σε επαφή ένας επιστήμονας κατά τη διάρκεια της καριέρας του, μπορεί να κάνει προβλέψεις οι οποίες δεν επηρεάζονται από εξωγενείς παράγοντες που θα μπορούσαν να επηρεάσουν έναν άνθρωπο. Επομένως, τέτοια μοντέλα Μηχανικής Μάθησης μπορούν να χρησιμοποιούνται υποστηρικτικά για τη λήψη αποφάσεων ιατρικής φύσεως και τη μείωση του διαγνωστικού σφάλματος στο ελάχιστο.

Συνελικτικά Νευρωνικά Δίκτυα

Τα Συνελικτικά Νευρωνικά Δίκτυα αποτελούν ένα πολύ σημαντικό τμήμα της Τεχνητής Νοημοσύνης και της Μηχανικής Μάθησης. Χρησιμοποιούνται ευρέως σε προβλήματα ταξινόμησης εικόνων και εφαρμόζουν την επιβλεπόμενη μάθηση. Η ειδοποιός διαφορά τους από τα υπόλοιπα είδη Νευρωνικών Δικτύων έγκειται στο γεγονός ότι δέχονται εικόνες ως είσοδο και εφαρμόζουν σε αυτές την πράξη της συνέλιξης με κάποιο φίλτρο. Η συνέλιξη ενός πυρήνα w μεγέθους $m \times n$ με μία εικόνα $f(x, y)$ δίνεται από τον τύπο:

$$(w * f)(x, y) = \sum_{s=-a}^{a} \sum_{s=-b}^{b} w(s, t) f(x - s, y - t) \tag{1}$$

Σύνολα Δεδομένων

Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν τρία διαφορετικά σύνολα δεδομένων, ή και υποσύνολα τους, όπως παρουσιάζονται ακολούθως επιγραμματικά:

- Σύνολο δεδομένων του Cambridge
- Σύνολο δεδομένων COUGHVID
- Σύνολο δεδομένων του Coswara

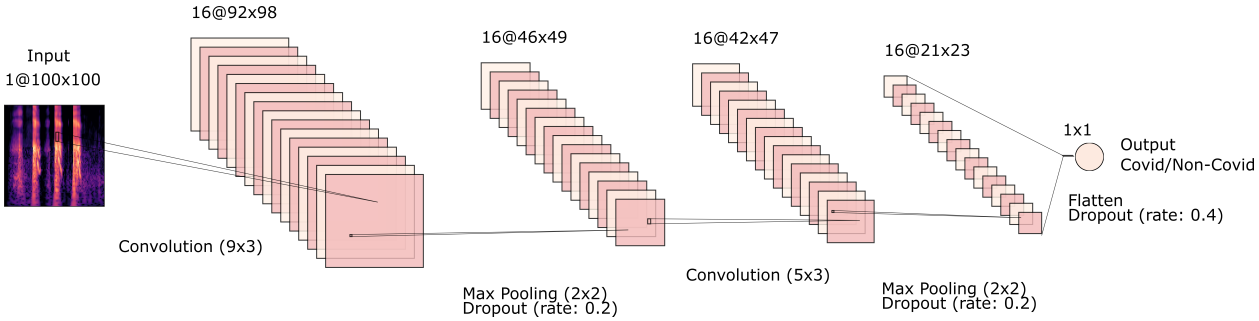
Το σύνολο δεδομένων του Cambridge περιέχει αρχεία βήχα και αναπνοής τόσο από υγιή άτομα, όσο και από άτομα μολυσμένα από τον COVID-19. Χρησιμοποιήθηκαν μόνο τα διαθέσιμα αρχεία βήχα, εκ των οποίων το 31.0% (124 δείγματα) ανήκει σε χρήστες διαγνωσμένους με τον ιό, ενώ το υπόλοιπο 69.0% (276 δείγματα) ανήκει σε υγιείς χρήστες.

Αναφορικά με το σύνολο δεδομένων COUGHVID, αυτό συνολικά περιέχει 27,550 αρχεία ήχου, εκ των οποίων μόνο τα 15,125 δείγματα περιέχουν αρχεία βήχα, λόγω κακής ποιότητας περιεχομένου σε κάποια από τα ηχογραφημένα δείγματα. Πληροφορίες σχετικά με τον αν ο χρήστης έχει ή δεν έχει μολυνθεί από το ιό δίνονται μόνο στα 10,819 αρχεία, τα οποία αποτελούν και τα δείγματα που μπορούν να χρησιμοποιηθούν στο εν λόγω πρόβλημα. Από το σύνολο των 10,819 δειγμάτων, 699 (6.46%) ανήκουν σε χρήστες που δηλώνουν ότι έχουν μολυνθεί από τον ιό, με τα υπόλοιπα 10,120 δείγματα (93.54%) να ανήκουν σε υγιείς χρήστες. Επιπλέον, 2,804 αρχεία εκ των 15,125 που περιέχουν βήχα, έχουν εξεταστεί και ταξινομηθεί από ειδικούς ως προς τη μόλυνση ή όχι του χρήστη από τον ιό. Σε αυτό το υποσύνολο των δεδομένων, 553 δείγματα (19.72%) ανήκουν σε χρήστες που έχουν μολυνθεί από τον ιό, ενώ τα υπόλοιπα 2,251 (80.28%) ανήκουν σε υγιείς χρήστες. Για την επίλυση του προβλήματος δοκιμάστηκε τόσο το συνολικό πλήθος των αρχείων, όσο και το τμήμα των δειγμάτων που έχουν εξεταστεί από κάποιον ειδικό.

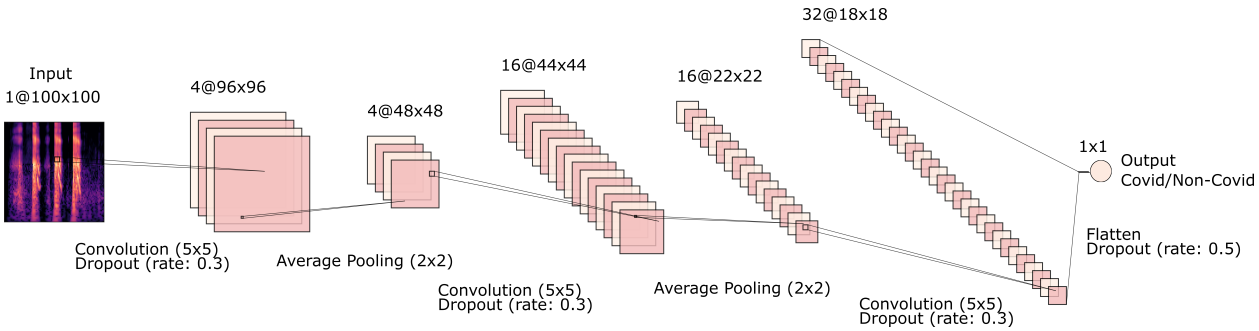
Όσον αφορά το σύνολο δεδομένων του Coswara, αυτό περιέχει εννέα διαφορετικούς αναπνευστικούς ήχους, εκ των οποίων χρησιμοποιούνται μόνο οι δύο που περιέχουν ήχους βήχα (cough heavy, cough shallow). Η κατανομή των δειγμάτων από χρήστες με COVID-19 και υγιείς χρήστες είναι σχεδόν ίδια στα δύο υποσύνολα, με το πρώτο να περιέχει 1,438 δείγματα από υγιείς χρήστες (93.32%) και 103 από άτομα με κορωνοϊό (6,68%), ενώ το δεύτερο περιέχει 1436 δείγματα από υγιή άτομα (93.31%) και 103 από χρήστες με κορωνοϊό (6.69%).

Αρχιτεκτονικές Συνελικτικών Νευρωνικών Δικτύων που χρησιμοποιήθηκαν

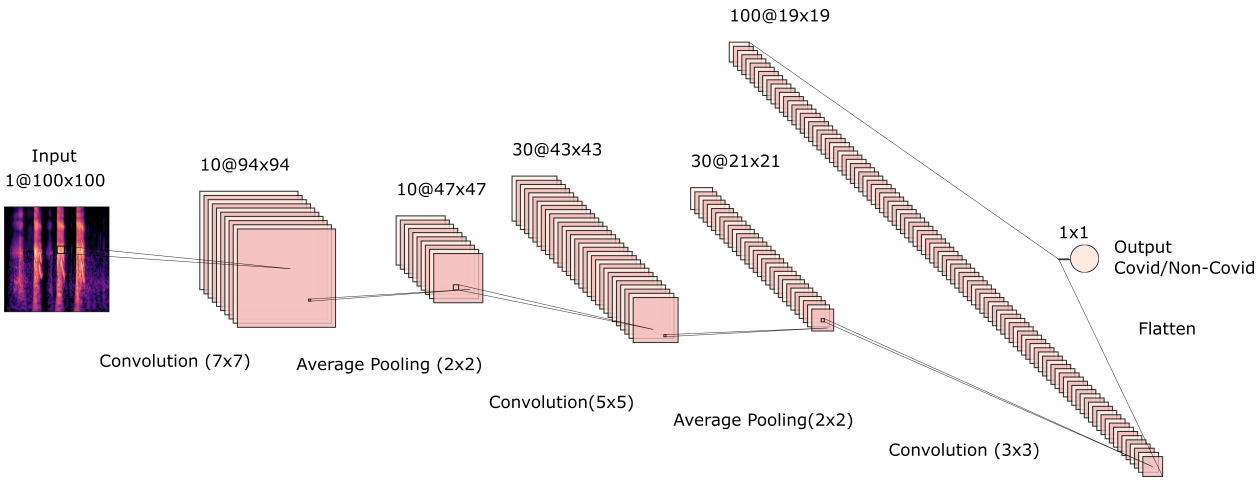
Δημιουργήθηκαν και δοκιμάστηκαν τρία μικρά συνελικτικά δίκτυα, η δομή των οποίων παρουσιάζεται στα σχήματα 1 - 3. Επιπλέον, δοκιμάστηκαν και ήδη υπάρχοντα προ-εκπαιδευμένα μοντέλα στο σύνολο δεδομένων του **ImageNet**, τα οποία ονομαστικά είναι τα: **ResNet-50**, **DenseNet-201** και **Xception**.



Σχήμα 1: Σχηματική απεικόνιση της αρχιτεκτονικής του μοντέλου 1



Σχήμα 2: Σχηματική απεικόνιση της αρχιτεκτονικής του μοντέλου 2

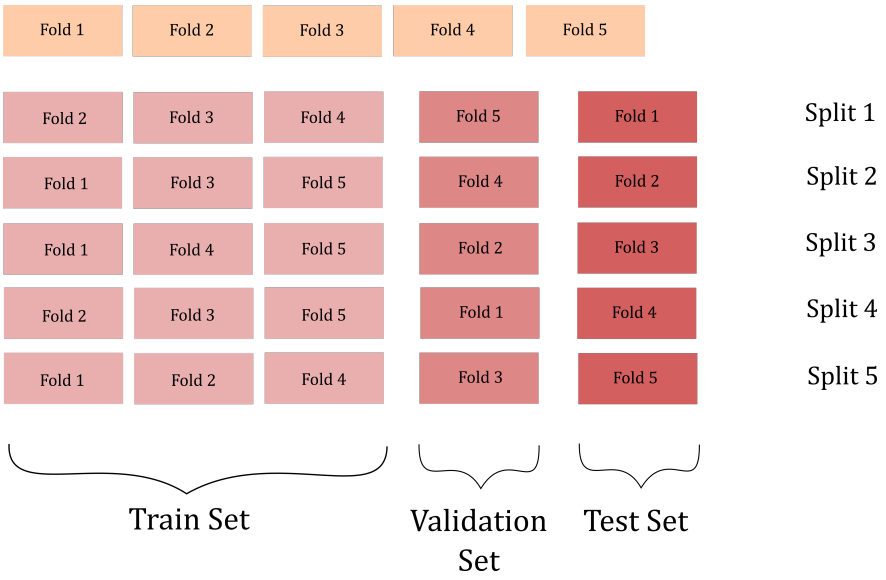


Σχήμα 3: Σχηματική απεικόνιση της αρχιτεκτονικής του μοντέλου 3

Μέθοδοι που εφαρμόστηκαν

Εφαρμόστηκαν τρεις διαφορετικές μέθοδοι για την επίλυση του προβλήματος ανίχνευσης του COVID-19 από αρχεία ήχου βήχα.

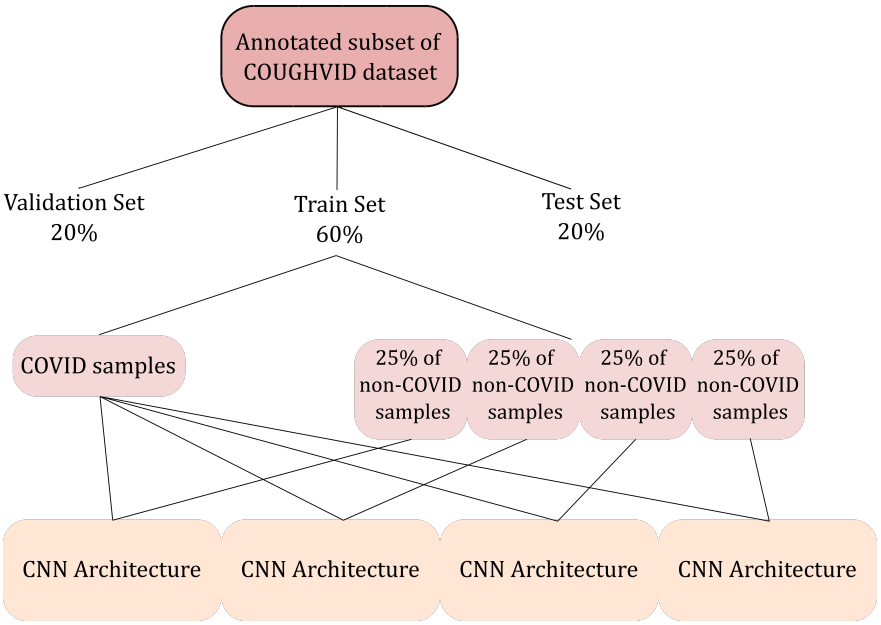
Στην πρώτη, το εκάστοτε μοντέλο εκπαιδεύεται σε ένα από τα διαθέσιμα σύνολα δεδομένων εφαρμόζοντας τη μέθοδο διασταυρούμενης επικύρωσης. Τα δεδομένα χωρίζονται σε 5 τμήματα με το καθένα εκ των οποίων να αποτελεί ακριβώς μία φορά το σύνολο αξιολόγησης (test set), ενώ τα υπόλοιπα 4 τμήματα απαρτίζουν τα σύνολα εκπαίδευσης και επικύρωσης με το πρώτο να αποτελείται από 3 τμήματα του συνόλου των δεδομένων και το δεύτερο από 1. Κατά την εκπαίδευση των μοντέλων εφαρμόστηκε και η μέθοδος Synthetic Minority Oversampling Technique (SMOTE) για την καταπολέμηση του προβλήματος έλλειψης δεδομένων από χρήστες με κορωνοϊό. Ο τρόπος διαχωρισμού των δεδομένων παρουσιάζεται στο σχήμα 4.



Σχήμα 4: Παρουσίαση της μεθόδου διασταυρούμενης επικύρωσης

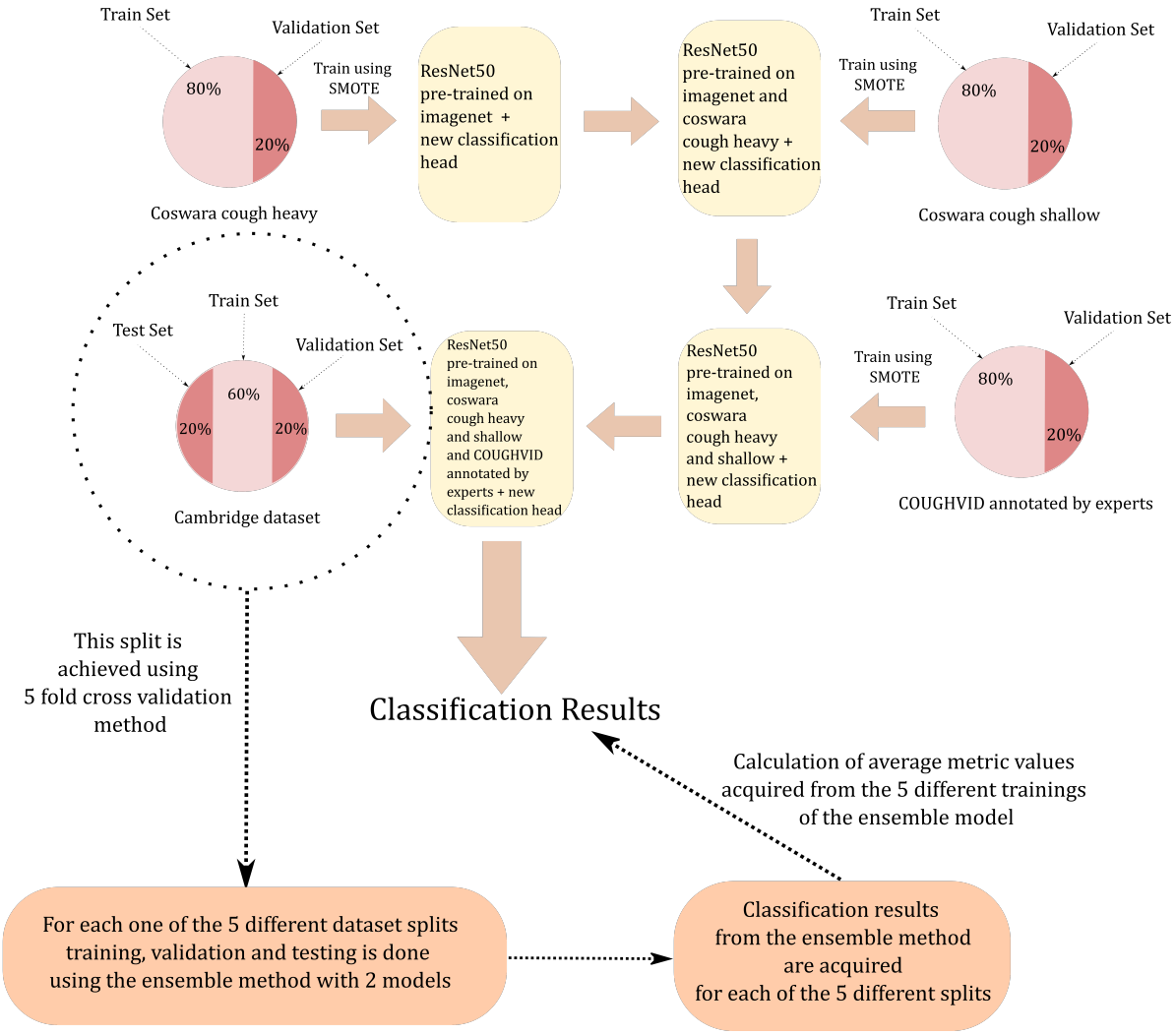
Στη δεύτερη μέθοδο εφαρμόζεται η τεχνική της συλλογικής μάθησης, με στόχο την εξάλειψη των αρνητικών συνεπειών της ανισορροπίας των δεδομένων στα αποτελέσματα της ταξινόμησης. Η ίδια αρχιτεκτονική εκπαιδεύεται δύο ή τέσσερις φορές, ανάλογα με το χρησιμοποιούμενο σύνολο δεδομένων και το ποσοστό δειγμάτων που ανήκουν σε κάθε μία εκ των δύο κλάσεων, covid και όχι-covid, με διαφορετικά υποσύνολα δειγμάτων. Το τελικό αποτέλεσμα αποτελεί συνδυασμό των προβλέψεων των μοντέλων για τα δείγματα που περιέχονται στο σύνολο αξιολόγησης. Ο διαχωρισμός των δεδομένων για την εκπαίδευση των μοντέλων, στην περίπτωση χρήσης τεσσάρων μοντέλων, παρουσιάζεται στο σχήμα 5. Η εν λόγω μέθοδος συνδυάζεται με τη μέθοδο διασταυρούμενης επικύρωσης που περιγράφηκε προηγουμένως, με στόχο την επιβεβαίωση της εγκυρότητας των λαμβανόμενων αποτελεσμάτων.

Τέλος, δοκιμάστηκε η εκπαίδευση των τριών, ήδη υπαρχόντων, προ-εκπαιδευμένων μοντέλων που προαναφέρθηκαν, με τρία διαφορετικά σύνολα δεδομένων, τα δύο υποσύνολα του συνόλου



Σχήμα 5: Διαχωρισμός των δεδομένων για τη μέθοδο συλλογικής μάθησης και για το σύνολο δεδομένων COUGHVID που έχει αξιολογηθεί από τους ειδικούς

του Coswara και το υποσύνολο του COUGHVID που έχει αξιολογηθεί από τους ειδικούς. Στη συνέχεια το μοντέλο αυτό εκπαιδεύεται και αξιολογείται στο σύνολο δεδομένων του Cambridge, χρησιμοποιώντας τη μέθοδο συλλογικής μάθησης σε συνδυασμό με τη μέθοδο διασταυρούμενης επικύρωσης, όπως αυτές περιγράφηκαν προηγουμένως. Η εν λόγω διαδικασία εκπαίδευσης για την αρχιτεκτονική ResNet-50, η οποία μέσω δοκιμών παρατηρήθηκε ότι είναι η καταλληλότερη εκ των τριών για το συγκεκριμένο πρόβλημα, παρουσιάζεται αναλυτικά στο σχήμα 6.



Σχήμα 6: Περιγραφή της μεθόδου εκπαίδευσης του προ-εκπαιδευμένου μοντέλου ResNet-50 με πολλαπλά σύνολα δεδομένων

Διασταυρούμενη επικύρωση

Η μέθοδος της διασταυρούμενης επικύρωσης χρησιμοποιήθηκε για την εκπαίδευση πέντε μοντέλων, των τριών καινούριων μοντέλων που δημιουργήθηκαν, καθώς και δύο προ-εκπαιδευμένων στο ImageNet μοντέλων, των ResNet-50 και DenseNet-201. Τα παραπάνω μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν χρησιμοποιώντας 5 διαφορετικά σύνολα δεδομένων, τα δύο υποσύνολα του συνόλου του Coswara, το σύνολο δεδομένων COUGHVID και το υποσύνολο του που περιέχει μόνο τα δείγματα που αξιολογήθηκαν από κάποιον ειδικό, καθώς και το σύνολο δεδομένων του Cambridge. Οι μετασχηματισμοί των αρχείων ήχου σε εικόνα που εφαρμόστηκαν σε αυτή την περίπτωση είναι ο HCQT και τα φασματογραφήματα Mel. Παρατηρείται ότι η εν λόγω μέθοδος δεν επιτυγχάνει αποδεκτές επιδόσεις, με τις τιμές όλων των μετρικών να είναι σχετικά χαμηλές. Η μετρική precision λαμβάνει εξαιρετικά χαμηλές τιμές οι οποίες, παρόλο που μεταβάλλονται αρκετά ανάλογα με το χρησιμοποιούμενο σύνολο δεδομένων, δεν παύουν να είναι ιδιαίτερα χαμηλές σε όλες τις περιπτώσεις. Επιπροσθέτως, ενδιαφέρουσα είναι η παρατήρηση ότι το Μοντέλο 1 παρουσιάζει τις καλύτερες επιδόσεις σχεδόν σε όλους τους συνδυασμούς συνόλου δεδομένων και μετασχηματισμού του ήχου σε εικόνα. Στον πίνακα 2 παρουσιάζονται τα καλύτερα αποτελέσματα για κάθε σύνολο δεδομένων και για κάθε μετασχηματισμό.

Σύνολο δεδομένων	Μετασχηματισμός	Μοντέλο	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
Coswara cough heavy	HCQT	Μοντέλο 1	67.10	57.09	11.44	65.80	67.80
	Mel spectrograms	Μοντέλο 1	72.48	50.83	12.01	64.34	76.05
Coswara cough shallow	HCQT	Μοντέλο 1	74.15	41.86	11.69	64.27	76.48
	Mel spectrograms	Μοντέλο 1	71.16	46.79	11.16	64.26	72.93
COUGHVID	HCQT	Μοντέλο 1	55.16	42.33	6.33	49.08	56.05
	Mel spectrograms	Μοντέλο 1	52.52	57.94	7.74	55.97	52.14
COUGHVID με αξιολόγηση από τους ειδικούς	HCQT	Μοντέλο 1	50.16	62.01	22.51	55.37	47.24
	Mel spectrograms	Μοντέλο 1	54.48	47.79	21.31	51.63	56.13
Cambridge	HCQT	Μοντέλο 1	62.75	53.82	40.23	63.85	63.93
	Mel spectrograms	Μοντέλο 2	63.75	57.82	42.41	64.28	64.66

Table 2: Οι τιμές των μετρικών για τις καλύτερες επιδόσεις χρησιμοποιώντας τη μέθοδο διασταυρούμενης επικύρωσης

Μέθοδος συλλογικής μάθησης

Η μέθοδος της συλλογικής μάθησης εφαρμόστηκε σε δύο σύνολα δεδομένων, το σύνολο δεδομένων COUGHVID που έχει αξιολογηθεί από τους ειδικούς και το σύνολο δεδομένων του Cambridge. Τα σύνολα αυτά επιλέγονται διότι τα μοντέλα που εκπαιδεύτηκαν και αξιολογήθηκαν σε αυτά, σημείωσαν σχετικά υψηλότερες επιδόσεις στην προηγούμενη μέθοδο. Εφαρμόστηκαν τέσσερις διαφορετικοί τρόποι μετατροπής του ήχου σε εικόνα, ο HCQT, τα φασματογραφήματα mel, ο CQT και ο STFT. Παρατηρείται ελάχιστη αύξηση των βέλτιστων επιδόσεων συγκριτικά με την προηγούμενη μέθοδο, ανεξαρτήτως του χρησιμοποιούμενου συνόλου δεδομένων. Τα αποτελέσματα της ταξινόμησης του μοντέλου με τις καλύτερες επιδόσεις για κάθε σύνολο δεδομένων και για κάθε μετασχηματισμό, παρουσιάζονται στον πίνακα 3.

Σύνολο δεδομένων	Μετασχηματισμός	Μοντέλο	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
COUGHVID με αξιολόγηση από τους ειδικούς	HCQT	DenseNet	55.39	53.51	22.74	57.10	55.84
	Mel spectrograms	Μοντέλο 2	53.18	52.98	21.82	54.86	53.23
	CQT	Μοντέλο 1	57.99	50.83	23.81	57.29	59.75
	STFT	Μοντέλο 2	51.84	55.17	20.05	54.09	51.09
Cambridge	HCQT	Μοντέλο 2	62.55	55.66	41.90	63.68	63.05
	Mel spectrograms	Μοντέλο 1	60.80	54.82	39.96	61.06	60.83
	CQT	Μοντέλο 1	60.05	51.77	38.92	59.16	61.30
	STFT	Μοντέλο 1	59.50	64.35	43.79	63.60	57.64

Table 3: Οι τιμές των μετρικών για τις καλύτερες επιδόσεις χρησιμοποιώντας τη μέθοδο συλλογικής μάθησης

Πολλαπλή εκπαίδευση της αρχιτεκτονικής **ResNet-50** χρησιμοποιώντας 4 διαφορετικά σύνολα δεδομένων

Η πολλαπλή εκπαίδευση προ-εκπαιδευμένων αρχιτεκτονικών με διαφορετικά σύνολα δεδομένων σημείωσε αξιόλογη βελτίωση των τιμών των μετρικών για το μοντέλο ResNet-50, με τις επιδόσεις που επιτυγχάνονται να ξεπερνούν τις βέλτιστες επιδόσεις των προαναφερθέντων μεθόδων. Πραγματοποιήθηκε πληθώρα δοκιμών και συνδυασμών με τις δύο υψηλότερες επιδόσεις να επιτυγχάνονται από την εκπαίδευση του μοντέλου χρησιμοποιώντας τον HCQT και τα φασματογραφήματα mel. Οι τιμές των μετρικών ταξινόμησης για τους δύο αυτούς συνδυασμούς παρουσιάζονται στον πίνακα 4.

Μετασχηματισμός	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
HCQT	71.03	66.58	52.18	73.44	71.51
Mel spectrograms	71.60	62.92	57.21	69.92	74.78

Table 4: Οι τιμές των μετρικών για τις καλύτερες επιδόσεις χρησιμοποιώντας τη μέθοδο πολλαπλής εκπαίδευσης του ResNet-50

Συμπεράσματα

Στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη μίας μεθόδου Βαθιάς Μηχανικής Μάθησης για την ανίχνευση του COVID-19. Δοκιμάστηκε ποικιλία συνόλων δεδομένων, τα οποία περιέχουν αρχεία ήχου βήχα τόσο από άτομα φορείς του κορωνοϊού, όσο και από υγιή άτομα. Εξαιτίας της ανισορροπίας των διαθέσιμων δεδομένων, οι επιδόσεις που σημειώθηκαν χρησιμοποιώντας ένα μόνο συνελικτικό μοντέλο δεν ήταν αρκετά καλές για να θεωρηθεί το μοντέλο αυτό αξιόπιστο. Με στόχο τη βελτίωση των αποτελεσμάτων ταξινόμησης, δοκιμάστηκε η εφαρμογή της συλλογικής μάθησης, για δύο από τα διαθέσιμα σύνολα δεδομένων. Ωστόσο, η βελτίωση της επίδοσης των μοντέλων δεν ήταν αρκετή ώστε να θεωρηθούν αξιόπιστα για την επίλυση του εν λόγω προβλήματος. Τέλος, ο συνδυασμός των παραπάνω μεθόδων με την

πολλαπλή εκπαίδευση της αρχιτεκτονικής ResNet-50 με διαφορετικά σύνολα δεδομένων, επέφερε αξιόλογη αύξηση στα αποτελέσματα της ταξινόμησης με την ακρίβεια να φτάνει το 71.03% στην περίπτωση χρήσης του μετασχηματισμού HCQT και το 71.60% στην περίπτωση χρήσης των φασματογραφημάτων mel. Από όσο γνωρίζουμε, αυτή είναι η πρώτη φορά που χρησιμοποιείται ο μετασχηματισμός HCQT σε πρόβλημα ταξινόμησης ήχων βήχα. Τα αποτελέσματα αυτά επιβεβαιώνουν τη δυνατότητα της Μηχανικής Μάθησης να αποτελέσει ουσιαστικό αρωγό στην επιστήμη της Ιατρικής. Ένα τέτοιο εργαλείο καθιστά το συχνό ή και καθημερινό διαγνωστικό έλεγχο όλου του πληθυσμού για πιθανή μόλυνση από τον κορωνοϊό, άμεσο και μηδενικού κόστους. Επιπλέον, αντίστοιχοι αλγόριθμοι μπορούν να χρησιμοποιηθούν για την ανίχνευση ποικίλων ασθενειών του αναπνευστικού συστήματος, ενώ η επιστημονική κοινότητα θα είναι καλύτερα προετοιμασμένη για την αντιμετώπιση μίας μελλοντικής πανδημίας.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 COVID-19

The last one and a half year, the world has come up against a contagious disease known as COVID-19 (**CO**rona**VI**rus **D**isease of 20**19**), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). COVID-19 is also going to be referred to as Covid for simplicity throughout the rest of this thesis. At the time of writing there have been 225,680,357 confirmed cases of COVID-19 and 4,644,740 deaths, reported to the World Health Organization (WHO) [1]. In 31 December 2019 the government of Wuhan, China, confirmed cases of "viral pneumonia" in Wuhan and nine days later WHO reported the determination that the outbreak is caused by a novel coronavirus. In 11 January 2020 the first death was reported and SARS-CoV-2 was declared as a pandemic by the WHO in 11 March 2020 [2].

Most of the infected people experience mild or moderate respiratory illness and recover without being in need of any special treatment. However, many contaminated individuals may experience severe illness and require medical assistance. Although people with underlying medical conditions, such as chronic respiratory disease, cardiovascular disease, diabetes and cancer as well as elderly people, are prone to developing serious illness, any age group can sustain severe ailment which can even cause death.

1.1.1 Transmission methods

COVID-19 can be transmitted among people in various ways. The main mean of transmission is air particles which are produced by the mouth and nose of an infected individual and can spread through speaking, breathing, coughing, sneezing and more. The size of these particles can range from larger respiratory droplets to smaller aerosols and can re-

main in the air for up to three hours [65].

A person can be infected when contaminated droplets or aerosols are inhaled or come in direct contact with the person's nasal and oral cavity, or eyes. The virus spreads between people who are in close contact, approximately 1 metre or less, but can also traverse to larger distances in the case of crowded or poorly ventilated indoor spaces. This is due to the fact that people tend to stay for longer periods of time in such places and the particles containing the virus remain in the air for longer or are transmitted for longer distances. Circuitous contact can also cause contamination. This can be due to infected objects coming into direct or indirect (through the hands) contact with the mouth, nose or eyes. The virus can spread from contaminated people regardless of whether they are symptomatic or asymptomatic. Individuals tend to be more contagious shortly before developing symptoms, while people with severe symptoms can infect others for longer periods of time [3]. The virus viability in different surfaces has been examined by Van Doremalen et al. [65]. Viable virus was detected on plastic and stainless steel objects 72 hours after the application, while it did not survive for more than 4 hours on copper and for more than 24 hours on cardboard.

1.1.2 COVID-19 symptoms and long-term consequences

The most common COVID-19 symptoms include fever, cough, tiredness and loss of taste or smell, with some of the less common ones being headache, aches and pains, sore throat and diarrhoea. Serious COVID-19 symptoms include difficulty in breathing or shortness of breath, loss of speech or mobility, confusion and chest pain. Symptoms may appear 2-14 days after contamination, with the average time interval being 5-6 days. The SARS-CoV-2 virus can infect a lot of different cells and systems of the body, with the mostly affected parts being the upper respiratory tract (sinuses, nose, and throat) and the lower respiratory tract (windpipe and lungs) [66]. Symptoms existing for more than 3 weeks are described as post-acute COVID-19 syndrome and except for patients who experienced severe illness, it also affects patients who have had mild or moderate symptoms. Although the most common remaining symptoms are fatigue, dyspnea, joint pain and chest pain, the dysfunction of other organs has also been reported, including the heart, the lungs and the brain. Myocardial injury and thromboembolic disease has been recorded in patients who experienced severe illness. As for the long-term, lung related consequences of COVID-19, studies indicate that even 3 months after discharge, patients could have persistent symptoms, radiological and lung function abnormalities. The most common dysfunctions related to the brain are namely anosmia, ageusia and headache, but other diseases, such as stroke, impairment of consciousness, seizure, and encephalopathy, have also been reported [4], [5], [6].

1.1.3 Contamination prevention

Plenty of personal preventive measures against COVID-19 can be implemented with some of the most valuable being listed below:

- Social distancing. Keeping a distance greater than 1 meter, both from people experiencing possible COVID-19 symptoms but also from healthy individuals (possible asymptomatic carriers)
- Wearing a facial covering when physical distance cannot be kept, or when being in an indoor place
- Avoiding poorly ventilated, indoor locations, crowded places and prolonged contact with others
- Frequently cleaning of hands with soap and water or alcohol-based sanitizer
- Vaccination. Vaccination is the only powerful measure of protection against the spreading of the virus and its mutations.

Except for the protection each individual can take, the governments throughout the globe have implemented various preventive measures in order for the pandemic to be contained. At the outbreak of the pandemic most governments implemented lockdowns, ceasing many indoor and outdoor activities. Other preventive measures applied consist of traffic banning during evening or night hours, prohibition on transportation between different cities, obligatory usage of masks in indoor and outdoor locations, mandatory demonstration of negative COVID-19 testing results or of the vaccination certificate in order to be allowed to travel and many others. Also, quarantining COVID-19 patients and tracking their contacts is of high importance and has been used by the competent institutions of each government ever since the beginning of the pandemic.

1.1.4 Treatments and vaccine

Although many therapeutic strategies have been tried to defeat the pandemic, there is no specific treatment up to now. However, scientists have managed to create effective vaccines giving rise to hopes for a quick exiting from the pandemic. Four vaccines produced by different companies, Pfizer/BioNTech, Moderna, AstraZeneca and Johnson & Johnson/Janssen Pharmaceuticals have been approved by the European Medicines Agency (EMA) and belong to either one of the two available types of vaccine, mRNA and adenovirus. The mRNA types of vaccines contain a part of the "instructions" from SARS-CoV-2 allowing the body cells to create a protein which is unique to the virus. The foreign proteins are detected by the immune system, which produces antibodies and immune cells to

defend it and as a consequence natural defences against COVID-19 infection are created. The Pfizer/BioNTech and the Moderna vaccine implement this technology, while the other two vaccines implement the adenovirus technology. All of the aforementioned vaccines require two doses per person except for the Johnson & Johnson vaccine for which only one dose is required [67]. However, a "commemorative" third dose is scheduled to be given at least to certain social groups that are in greater danger of suffering from COVID-19.

The first mass vaccination program started in December 2020 and by the time of writing and according to [7], 42.41% of the world population has received at least one dose of a COVID-19 vaccine, with 30.25% being fully vaccinated. As for Europe, 55.31% of the population is at least partly vaccinated, with 50.23% being fully vaccinated. Moreover, 62.53% of the United States population has been vaccinated with at least one dose, while 53.31% has been fully vaccinated. Lastly, 60.76% of the Greek population has been at least partly vaccinated, with 56.61% being fully vaccinated. These statistics are also depicted in figure 1.1. By the time of writing a total of 5.79 billion vaccine doses have been administered.

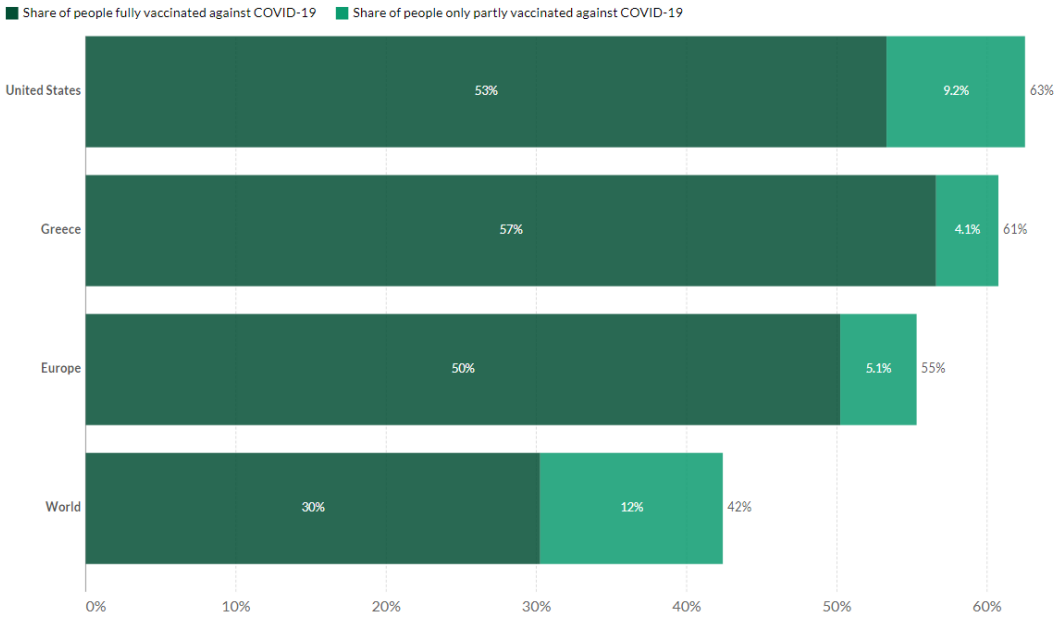


Figure 1.1: Statistics about the progress of vaccinations in certain regions [7]

General statistics about the progress of vaccinations around the world are shown in figures 1.2 and 1.3.

Vaccination is the only method of containing the pandemic and preventing the spread of new and often more contagious virus variants and mutations. By the time of writing, several different variants have been circulating around the globe. Variants Beta, Gamma and Delta constitute the VOC or Variants of Concern, since clear evidence indicating significant increase of transmissibility and severity and a decrease of immunity, is available. The VOI or Variants of Interest are comprised of variants for which evidence for possibly significant

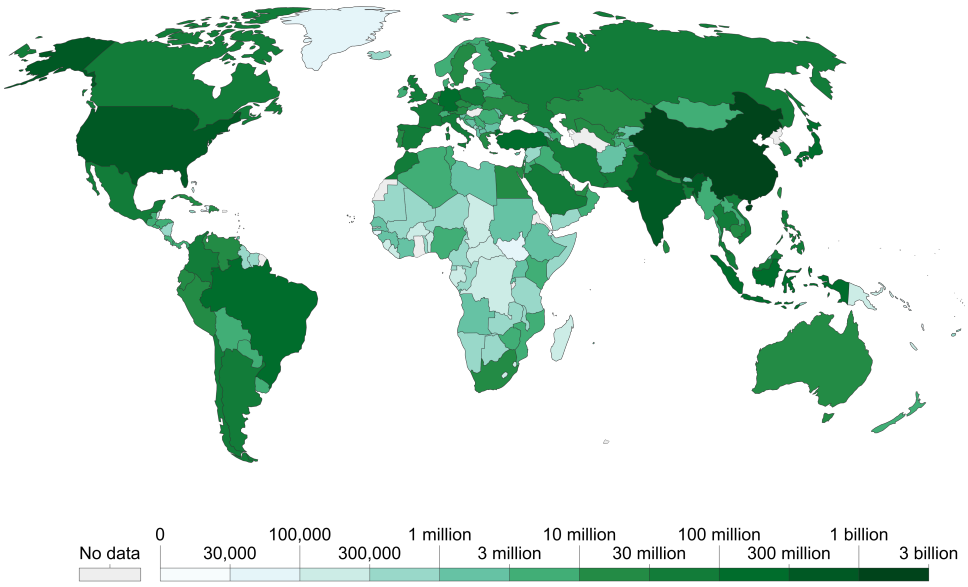


Figure 1.2: Heat map showing the number of vaccine doses given around the globe [60]

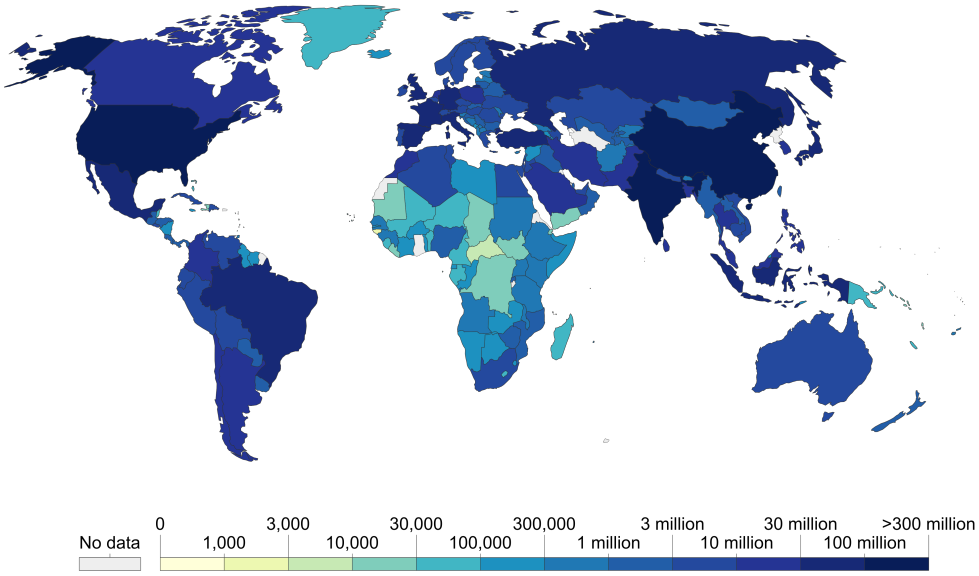


Figure 1.3: Heat map showing the number of fully vaccinated people around the globe [60]

increase of transmissibility and severity and a decrease of immunity is available. Two of them are the Mu and Lambda variants. More detailed information for these variants is provided in table 1.1 [64].

	WHO label	Country first detected (community)	Year and month first detected	Evidence for impact on transmissibility	Evidence for impact on immunity	Evidence for impact on severity
VOC	Beta	South Africa	September 2020	Yes	Yes	Yes
	Gamma	Brazil	December 2020	Yes	Yes	Yes
	Delta	India	December 2020	Yes	Yes	Yes
VOI	Mu	Colombia	January 2021	Yes	Yes	-
	Lambda	Peru	December 2020	-	Yes	-

Table 1.1: Information about some of the mostly spread variants [64]

1.1.5 Testing methods

The ability of detecting COVID-19 infection is of high importance for the containment of spreading. The most prevalent and reliable testing methods are namely: Nucleic Acid Amplification Tests (NAATs) and antigen tests. Both of them are viral tests used for discovering current infection and are performed by collecting nasopharyngeal and/or oropharyngeal specimens from the patient. NAATs are viral diagnostic tests for SARS-CoV-2, detecting genetic material and more specifically the RNA sequences which comprise the genetic material of the virus. One of the most commonly used methods for a NAAT is the Reverse Transcription Polymerase Chain Reaction (RT-PCR). As for the antigen tests, these are immunoassays that detect the presence of a specific viral antigen, implying current viral infection. The currently approved antigen tests include laboratory-based tests and self-tests, with the results being returned in approximately 15-30 minutes, rendering them a quick and easy screening method. However, they are generally less sensitive than NAATs which are unlikely to return a false negative result [8], [9].

1.1.6 General statistics

As mentioned previously, by the time of writing there have been 225,680,357 confirmed cases and 4,644,740 deaths caused by COVID-19 around the globe [2]. A total of 57.58 million confirmed cases and 1.20 million deaths have been reported in Europe, 41.54 million cases and 666,607 deaths in the United States and 622,761 cases and 14,311 deaths in Greece. A heat map presenting the total cases reported since January 22, 2020 is provided in figure 1.4, while statistics related to the age distribution of cases and losses, for those with a known and confirmed age, in Greece are depicted in figure 1.5 [60].

1.2 Motivation

During the COVID-19 pandemic, the health systems of many countries received unprecedented pressure with the hospitals' personnel exerting themselves, making the tracing and isolation of possible COVID-19 cases an issue of paramount importance. Moreover, new

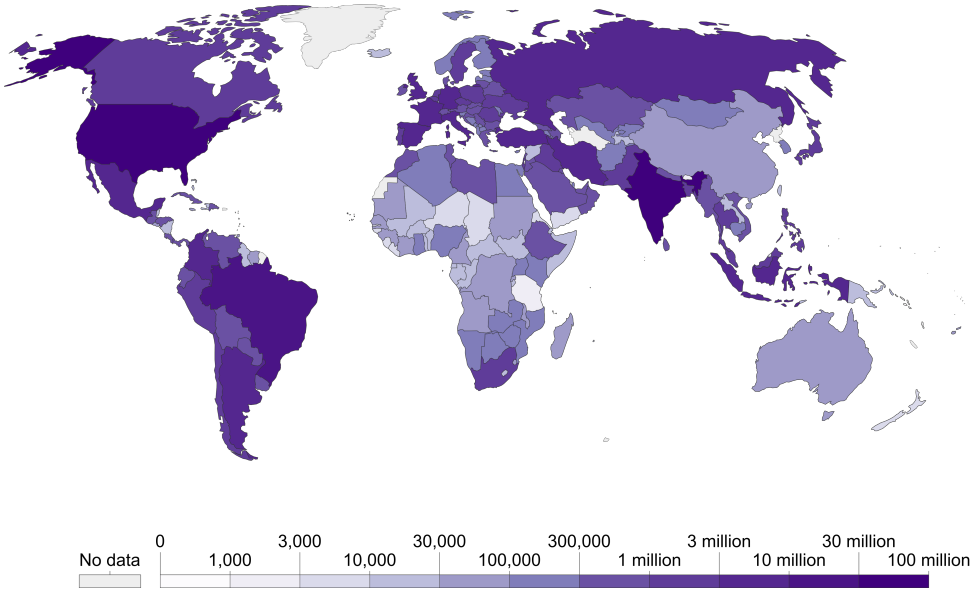


Figure 1.4: Heat map showing the confirmed cases around the globe

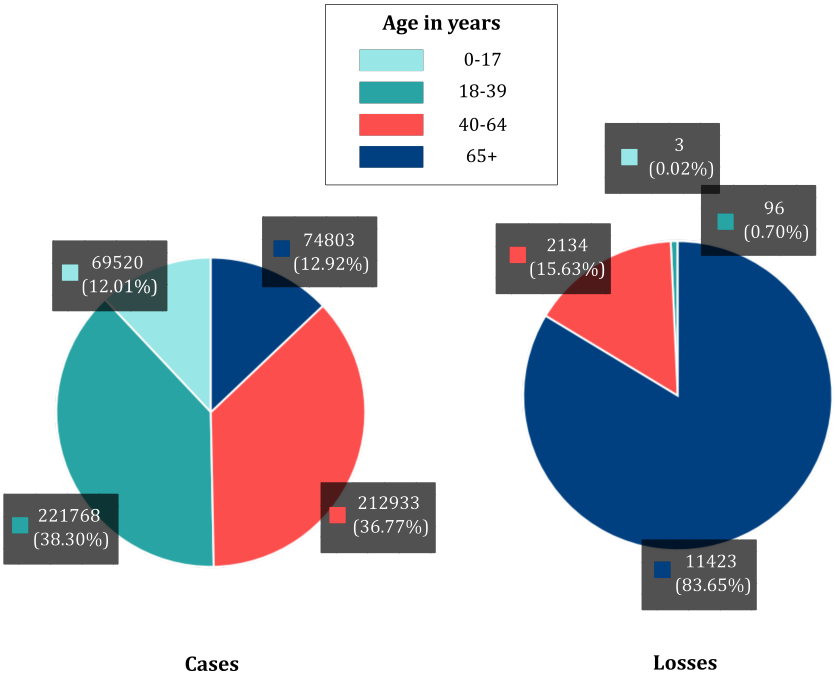


Figure 1.5: COVID-19 age statistics for Greece [61]

virus mutations and variants render COVID-19 more transmissible, making the prevention of new pandemic waves very difficult. The ability of daily testing large amounts of population and ideally the whole population of a country, could be a game changing parameter for the outcome of the pandemic. Currently, two types of viral tests are being used to detect COVID-19: Nucleic Acid Amplification Tests (NAATs) and antigen tests. However,

conducting the large amount of tests needed, is neither time nor cost efficient. In order to deal with this problem, different methods of diagnosing COVID-19 infection using Machine Learning (ML) techniques, have been proposed in recent studies [10], [11], [12], [13].

As the amount and type of available data displays increase during the past years, Machine Learning is widely being used in many aspects of everyday life and is already being implemented in various health related subjects. It is a quick and low cost way of providing high accuracy results, requiring minimum effort both by the user and the doctors.

With regard to COVID-19 controlling, the usage of Machine Learning algorithms has been thoroughly examined. A review of the usage of Artificial Intelligence in Medical Imaging Informatics as well as a systematic review of Artificial Intelligence models utilized for screening, diagnosis and prognosis of COVID-19 has been created [68], [69]. Radiology examination using chest X-ray images is being used by health care experts for the diagnosis of COVID-19. To that end, deep learning approaches have been used for the detection of COVID-19 from X-ray images [10], [14], [15], [16], [17], [18], [19], [20]. Chest computed tomography (CT) images are another mean of diagnosing COVID-19, since they show characteristics that differentiate a patient with COVID-19 from a patient with other types of pneumonia. Machine Learning algorithms for the diagnosis of COVID-19 using CT-images have also been developed [11], [21]. Moreover, Machine Learning approaches have been implemented for predicting the severity of illness and the need for Intensive Care Unit (ICU) admission of patients with COVID-19 [22], [23], [24], [25].

Although the approaches described above can accurately diagnose COVID-19, they require the physical presence of a possible case to a clinical facility, in order for the chest X-ray or the CT scan to be conducted. Apart from the time and effort needed for this to be done, the prevention of the spreading of the virus by needless interactions with the personnel and other patients is of the highest importance. Moreover, it cannot be taken for granted that all countries can provide their citizens with the required amount of testing or that all citizens have easy access to health benefits. Therefore, a quick and free screening method, available to everyone through their smartphones or via web applications, could conduct a very important role in the containment of the pandemic, due to the accessibility and simplicity of such a testing method, as well as the reduction of unnecessary contacts of possibly infected individuals with others.

1.3 Literature review

Research has shown that COVID-19 can be detected from lung sounds and in the recent years, considerable progress has been made regarding the utilization of respiratory sounds for the detection of diseases, using Machine Learning techniques. Some of the methods used, focus on the extraction of characteristics from the sound samples, while others

choose to convert audio to image and leverage the effectiveness of Convolutional Neural Networks (CNNs) in image classification tasks.

1.3.1 Audio to image conversion

Multiple methods of converting audio to image have been employed in research studies. Salamon and Bello [26] implemented the log-scaled Mel spectrogram representation of audio signals in environmental sound classification tasks. Sounds belonging to 10 environmental sound categories which are namely: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music (UrbanSound8K dataset) were converted to mel spectrograms and used to train a deep CNN architecture. The proposed architecture reached a mean accuracy of 73.0%, but the usage of audio data augmentations increased this value to 79%. Kiskin et al. [27] implemented a less popular conversion of sound to image, the Continuous Wavelet Transform (CWT), in two tasks: the detection of mosquitoes and the classification of bird species. The study focuses on the mosquito detection and shows that a robust model can be used to other similar classification tasks with minimal alterations. A performance comparison between Short Time Fourier Transform (STFT), Mel Frequency Cepstral Coefficient (MFCC) and CWT, among other, is conducted with the CWT outperforming both STFT and MFCC with the first one achieving f1-score of 91.3% and the other two achieving scores of 88.3% and 89.5% respectively, with regard to the mosquito detection task. As for the bird species classification task, the combination of the CNN architecture proposed with the CWT outperforms any other experimentation, reaching an f1-score of 92.5%. The usage of CWT has also been examined in other audio related tasks and specifically in the fundamental heart sounds classification task. The scalograms produced by CWT were used to train a CNN architecture, reaching an accuracy of 86.0% when distinguishing between the first and second heart sound [28]. A comparison of different audio to image conversions has also been proposed by Huzaifah [29], where the two environmental sounds datasets ESC-50 and UrbanSound8K and four different approaches to time-frequency representation, i.e. the STFT with both linear and mel-scales, the constant-Q transform (CQT) and the CWT were examined. Three of the four transformations, linear-STFT, Mel-STFT and CQT, performed similarly on both datasets. However, especially for the UrbanSound8K dataset, CWT's performance was lower and closer to MFCC's. Lidy and Schindler [30] examined the usage of CQT for the task of classifying acoustic scenes and urban sound scapes, employing a CNN architecture. A comparison between the Mel-transform and CQT is conducted, with the latter outperforming the first one by achieving an accuracy of 80.25% in contrary to the 76.55% accuracy reached when using the Mel-transform. Environmental sound classification using Deep CNN structures, trained with Mel-spectrograms and transfer learning, has also been examined by Mushtaq et al. [31]. Three datasets are utilised, ESC-10, ESC-50 and Us8k, two CNN architectures and two data augmentation techniques. The first augmentation technique used is the traditional one, where new data is created by applying image transformations such as rotating, flipping, zooming and other. The second one, is

the proposed augmentation technique where new audio samples are created using techniques such as positive and negative pitch shift, slow and fast time stretches and silence trimming. The audio samples are converted to Mel spectrograms to be used with the CNN architectures. In the case of using the CNN models created, the highest accuracy (95.50%) is achieved for the ESC-10 dataset using the proposed augmentation technique, while in the case of using transfer learning techniques the highest accuracy is achieved for the Us8K dataset (99.497%), using the ResNet-152 architecture and the proposed augmentation technique. Sharan and Moir [32] examined time-frequency image representations of sound signals related to an audio surveillance application. A new feature, based on image texture analysis, is proposed and is referred to as the Spectrogram Image Texture Feature (SITF). This feature was observed to be more noise robust than other features applied which are the MFFCs, the Gammatone Cepstral Coefficients (GTCCs), the Spectrogram Image Feature (SIF) and a variation of it with reduced feature dimension referred to as RSIF. Moreover, a gammatone filter-based image, referred to as cochleagram image, was used instead of the spectrogram image for feature extraction, improving the classification accuracy. Cochleagram images have also been utilised for an acoustic event recognition task, using a database comprised of 50 sound classes [33]. Four time-frequency audio representations are examined: the conventional spectrogram, the smoothed spectrogram which is acquired by applying moving average to the spectrogram along the frequency domain, the mel spectrogram and the cochleagram image. The accuracy acquired when using each of the aforementioned representations with a CNN architecture is: 93.46%, 96.34%, 95.35% and 98.03% respectively, with the cochleagram time-frequency representation outperforming the rest. Typical spectrogram audio representations have also been examined in [34]. The DCASE 2016 acoustic scene classification challenge data was utilized for the exploration of the acoustic scene classification task, using a CNN architecture and four different ways of representing sound as image: mel-scaled, logarithmically scaled and linearly scaled filterbank spectrograms as well as Stabilized Auditory Image (SAI) features. Lastly, a novel audio to image conversion has been utilized, in combination with CNN architectures, for the examination of the acoustic scene classification task. Wang et al. [35] examined the usage of CQT, Hybrid Constant-Q Transform (HCQT) as well as MFCCs, log-mel energies and its HPSS. The system proposed is evaluated on the DCASE 2019 challenge and it is observed that HCQT outperforms CQT for different CNN architectures and ensemble models, reaching an accuracy of 77.5% when combined with ensemble methods.

1.3.2 Cough classification

As for the more specific task of cough classification, Amoh and Odame [36] approached the cough detection task using two different Deep Learning methods: image analysis using a CNN and sequence-to-sequence labelling approach using a Recurrent Neural Network (RNN). As for the visual recognition problem, a LeNet-5 inspired CNN architecture, with a smaller number of neurons in each layer, was used combined with the Short Time

Fourier Transform (STFT) of the audio samples to detect cough events. However, since audio signals do not have a fixed size, like image data do, a pre-segmentation step was implemented to ensure the elimination of this problem. A database consisting of various respiratory sounds such as breathing, reading and coughing was created and it was observed that the CNN provided overall higher accuracy (89.7%) than the RNN. In some of their previous work, J. Amoh and K. Odame [37] presented a wearable acoustic sensor that records the person's respiratory sounds combined with a CNN, for the detection of cough. A pre-processing step was also implemented where some preliminary features were extracted and a frame admission process was implemented in order for irrelevant data to be excluded. Spectral segments, created with STFT, were then fed into a CNN offering a classification sensitivity of 95.1% and a specificity of 99.5%. Due to their popularity, CNNs have been used in similar tasks but with different ways of acquiring an image from an audio. Bales et al. [38] used CNN models to initially detect cough sounds between other environmental sounds and then the existence of bronchitis, bronchiolitis and pertussis. The image-frequency representation of the audio signals was obtained using the Mel-spectrograms which were then converted to gray-scale and inputted to the developed CNN architecture. The accuracy reached for the cough detection task equals 89.05%, with the overall accuracy for the cough classification task being 89.60%. Another approach to cough sounds classification is made by Aykanat et al. [39] where both CNN models and classical Support Vector Machines (SVMs) are tested. A total of 17,930 lung sounds from 1,630 individuals were collected using an electronic stethoscope. Mel Frequency Cepstral Coefficients features are inputted in an SVM, while spectrogram images using STFT are fed into a CNN model, to classify respiratory sounds into different categories based on four different tasks: healthy versus pathological classification, rale, rhonchus and normal audio classification, singular respiratory audio type classification and sound type classification. The CNN and SVM performances were extremely close in all four tasks with the accuracy being 86.0% for both classifiers in the first task and 76.0% for the CNN classifier which outperformed the SVM (75.0%) in the second task. As for the third and fourth task, the performances of the two classifiers were the same, with the accuracy reached being 80.0% and 62.0% respectively. In addition to the aforementioned studies, Miranda et al. [40] compared the performance of STFT, Mel Frequency Cepstral Coefficients (MFCC) and Mel-scaled filter banks (MFB) using Deep Neural Networks (DNN), CNNs and long-short term models in the cough detection problem, concluding that considering each cough sample as a single input feature, using longer analysis windows and utilizing the STFT and MFB, in contradiction to the MFCC, improves the classifier's performance. Bardou et al. [41] approached the lung sounds classification problem by training three classifiers (support vector machines, k-nearest neighbour, and Gaussian mixture models) using MFCC coefficients, as well as by experimenting with CNN models and utilizing local binary pattern (LBP) features extracted from the spectrograms. More specifically, these techniques were used to classify respiratory sounds to 7 different classes which are namely the following: normal, coarse crackle, fine crackle, monophonic wheeze, polyphonic wheeze, squawk and stridor, showing CNN's performance was better than this of the classifiers using handcrafted features. Moreover, Barata et al. [42] contributed to the mobile cough detection task by showing that the mean of recording plays a very important role for the model's performance and implemented both a CNN and an ensemble based model, in or-

der for the cross-device deviance to be reduced. Mel spectrograms were inputted into the CNN and the ensemble models, with the classification results indicating that the different quality recordings acquired from various devices plays an important role in the model's performance. The mean accuracies achieved range between [85.9%, 90.9%]. Hui-Hui Wang et al. [43] examined 5 different audio to image representations combined with a CNN architecture for dealing with the cough detection task. Experiments were performed on 70,000 audio samples from 26 patients. The methods used to convert the audio signals to images are namely: the original spectrum, the RASTA-PLP power spectrum, the RASTA-PLP cepstrum, the 12th order PLP power spectrum without RASTA and the 12th order PLP cepstrum without RASTA. The average accuracy achieved in each case is 93.8%, 99.65%, 89.56%, 93.92%, 93.02% respectively, with the RASTA-PLP spectrum outperforming the rest of the methods. Lastly, the advance made in the cough classification task is not only utilized in human cough sounds. Yin et al. [44] leverage the success of CNN architectures on the cough detection task and propose a classification algorithm for a respiratory disease alarm system inside a pig farm. More specifically, a fine-tuned AlexNet model, combined with STFT spectrogram images of pig cough sounds recorded in field situations, is used. The overall accuracy reached equals 95.6%, with the cough accuracy being 96.8% and the f1-score 96.4%.

1.3.3 COVID-19 classification using cough samples

The knowledge acquired by studies on respiratory sound classification tasks, gave rise to Machine Learning approaches using respiratory sound samples for the containment of the COVID-19 pandemic. Imran et al. [12] examined the differential pathomorphological alternations caused by COVID-19, relative to other cough causing medical conditions. A simple mobile app called AI4COVID-19 that collects a sound, detects the existence of cough and classifies it as COVID or non-COVID case, is created. The data used was collected from COVID-19, pertussis and bronchitis patients and healthy people. For the cough detection task, the mel-spectrograms of the sound signal are passed into the CNN based classifier. For the final decision, regarding the existence of COVID-19 infection, to be taken, a combination of three parallel classifiers was implemented. A Deep Transfer Learning-based multi-class classifier, a classical Machine Learning-based multi-class classifier and a Deep Transfer Learning-based binary class classifier were used, with their outcomes being passed into a mediator. More specifically, the input of the first and last, out of the three aforementioned, classifier was acquired by computing the Mel-spectrograms of the cough samples, while the input of the second classifier was acquired using MFCC and Principle Components Analysis (PCA) based feature extraction. The overall accuracy for each of the three classifiers respectively is: 92.64%, 88.76%, 92.85%. Another approach to the COVID-19 classification task is made by Brown et al. [45], where three different tasks regarding COVID-19 classification were examined, creating and using a crowdsourced dataset containing breath and cough sounds. The first task focuses on samples from users who have declared they tested positive for COVID-19 and users who have not declared a positive

test and have a clean medical history. The second task uses samples from users who have tested positive and have cough as a symptom and healthy users, while the third task focuses on distinguishing users who have tested positive in COVID-19 having cough as a symptom, from users who have not tested positive and have reported asthma as a pre-existing medical condition. Logistic Regression is used with handcrafted features and with features automatically extracted by VGGish, giving AUC values of about 80%. Classification of forced cough sounds can also be achieved by inputting their transformation with Mel Frequency Cepstral Coefficients to a combination of 3 pre-trained ResNet50's as shown in [46]. More specifically, a Poisson biomarker layer was combined with three pre-trained ResNet50 models in parallel. The first ResNet50 model was trained to distinguish the word "Them" from other words using an audiobook dataset containing approximately 1,000 hours of speech [47]. The second model was trained to learn sentiment features on a dataset including actors that intonate in 8 emotional states which are namely: neutral, calm, happy, sad, angry, fearful, disgust and surprised [48]. The last ResNet50 was trained to distinguish the spoken language of the person coughing (English or Spanish) on the cough dataset used, after taking into consideration only the metadata referring to the spoken language of the person. Classification results provided by the pre-trained models are higher than the ones acquired using not pre-trained ResNet50 models, reaching a sensitivity of 98.5%, a specificity of 94.2% and an AUC value of 97.0%. Moreover, the approach of Chaudhari et al. [49] has proven that the usage of an ensemble model combining Mel Frequency Cepstral Coefficients, Mel spectrograms and a label denoting the presence of respiratory diseases can provide a robust model, independent of the dataset used. The dataset employed consists of cough audio samples recorded with smartphones and more specifically, the proposed deep neural network architecture was trained on the Coswara and COUGHVID dataset containing 1,543 and 20,072 cough samples respectively. In order to obtain a more robust evaluation of the model's performance, two additional datasets (Virufy Latin American Crowdsourced Test Dataset and Virufy South Asian Clinical Test Datasets) consisting of data labelled using COVID-19 PCR results, were utilized. The results prove that the model constructed is robust enough for the change of the evaluation dataset to not significantly impact the performance. The highest accuracy is reached when using the Coswara and COUGHVID datasets and equals 77.1%, with the Virufy crowdsourced dataset achieving an accuracy of 72.1%. A combination of models has also been tested by Schuller et al. [13]. Spectrograms derived from raw breath and cough audio, contained in the dataset provided by [45], are inputted into a CNN architecture consisting of two branches, one for each of the two respiratory sound types available. The learned features of the two models are combined using fully connected layers in order for the final classification to be made. The best AUC score achieved is 80.7%, while an observation that breathing sounds could contain more COVID-19 information and thus provide slightly better results compared to cough sounds, was made. Other research studies utilizing cough, breath and speech sounds for the diagnosis of COVID-19 include the work of Pahar and Niesler [50] where two datasets, Coswara and ComParE, containing audio samples of the aforementioned categories were used, considering seven different classifiers: Logistic Regression (LR), Support Vector Machines (SVM), Multilayer Perceptrons (MLP), K-Nearest Neighbour (KNN), CNNs, Long Short-Term Memory (LSTM), RNNs and a residual based network (Resnet-50). Pre-processing was also implemented to remove periods of silence

in the signal. It is concluded that all three audio types can be used to successfully detect COVID-19, with the cough sounds carrying more COVID-19 information and reaching an AUC value of 93.0%, followed by the breath samples with an AUC of 92.0% and the speech segments with an AUC of 91.0%. Another approach to the COVID-19 diagnosis using respiratory sounds task is made by Bagad et al. [51], where the ResNet-18 was used as the base of their CNN architecture and was pre-trained on three open source cough datasets. The first one is the FreeSound Database 2018, containing 11,073 audio files belonging to 41 possible categories with 273 of them being cough samples [70]. The second one is the Flusense dataset where 11,687 samples of various categories were used, with 2,486 of them being cough samples [71]. Finally, Coswara dataset was also used containing 2,034 cough sounds and 7,115 non-cough sounds [72]. The data were split in train and validation sets and the model was trained to predict the presence or the absence of cough in an audio sample. This model was then used with a Covid dataset created, which was labelled using RT-PCR test results, containing 3,117 cough samples from 1,039 individuals, showing that pre-training improves the mean value of AUC by 17%. The log-scaled mel-spectrograms constitute the input of the model. Moreover, label smoothing is implemented on the final task, which is the COVID-19 classification task, since although the labels are obtained by RT-PCR testing results, this test is not completely accurate and hence the ground truth of the problem should not be based solely on that. The highest AUC value achieved using pre-trained models reaches 68.0%.

1.4 Scope of Thesis

One of the most challenging aspects of facing this pandemic, is the rapid and horizontal screening of citizens for possible COVID-19 positive cases. The current thesis presents a Deep Learning approach for the detection of COVID-19 positive cases using cough sound samples. More specifically, cough sound samples are converted to images, which are fed into Convolutional Neural Network (CNN) architectures that classify the sample to a Covid or non-Covid case. CNNs are widely used in image classification problems showing very promising results. A typical example is the state-of-the-art accuracy that has been achieved on the ImageNet task using the ResNet-152 model where the errors made during predictions were less than these made by a human [73]. On that end, we leverage the success of CNNs in image classification tasks and deal with the COVID-19 screening problem as such.

The structure followed in the current thesis is described below:
Chapter 2: Theoretical background regarding the audio signal characteristics, time-frequency representations of audio, Machine Learning (evolution and utilization in medical applications), classification metrics used for the assessment of the produced results.

Chapter 3: Analysis of the datasets, the architectures and the methods used.

Chapter 4: Presentation of the results obtained by implementing the methods described

in Chapter 3.

Chapter 5: Conclusions obtained and possible future research.

Chapter 2

Theoretical Background

2.1 Audio signals

Sound is created by the perturbation of a transmission medium's particles, such as air. The vibration of the particles propagates through the medium, from the transmitter to the receivers, as a mechanical wave. The waveform is one of the most common representations of sound and it provides information about the particles' displacement over the time. As shown in Figure 2.1, the y-axis represents the displacements, with the amplitude being the maximum displacement, while the x-axis provides time information. The wave's amplitude is an important characteristic, since it can be related to the sound's loudness, intensity and the energy transmitted. On the other hand, the time axis provides information about the period and frequency of a signal, which can either be periodic or aperiodic. However, frequency is an objective measure of sound's change over time and is perceived by humans logarithmically. This peculiarity of frequency perception is described by the "pitch". Apart from the information acquired from the amplitude and the frequency of a sound signal, the energy transferred via a sound wave can provide important characteristics of a particular audio signal. More specifically, sound power is used to provide information about the rate of energy transferring, while sound intensity about the sound power per unit area, with the intensity level of a sound being in logarithmic scale and measured in decibels (dB). Although a sound signal could be a simple sinusoidal, most of everyday sounds are complex signals composed of a superposition of sinusoidals, the harmonic partials.

2.1.1 Audio features

Audio features describe different characteristics and aspects of a sound and can be classified based on the Signal Domain in the following categories among others: Time domain

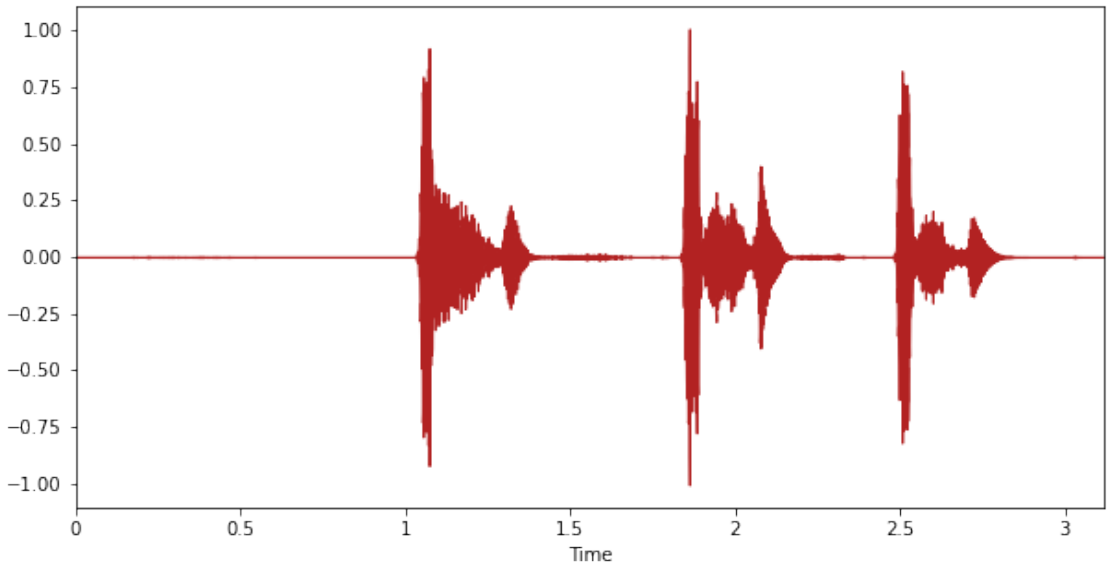


Figure 2.1: An example of a cough waveform

features, Frequency domain features, Cepstral domain features, Time-Frequency domain features [74]. Time domain features, such as Zero Crossing Rate (ZCR), provide information about the changes of the signal's amplitude over time, whereas frequency domain features are acquired by applying the Fourier Transform on the audio signal and supply information about its frequency components. However, time-frequency domain features are the ones of the highest interest since they provide combined knowledge of both time and frequency characteristics of an audio signal. Most time-frequency representations of a signal are based on the Short Time Fourier Transform and enable the visualization of the signal as a heat map known as spectrogram.

2.1.2 Audio to image transformations

Short Time Fourier Transform (STFT)

Short Time Fourier Transform is one of the most common methods used to depict audio signals. The signal is divided into smaller segments, where Discrete Fourier Transform (DFT) is applied and the Fourier spectrum of each specific segment is acquired. STFT provides information about the frequency variance over time, whereas DFT about the frequency over the whole time interval of the signal. The STFT of a segment is given by [75]:

$$X_{STFT}[m,n] = \sum_{k=0}^{L-1} x[k]g[k-m]e^{-j2\pi nk/L} \tag{2.1}$$

where $x[k]$ denotes the signal, $g[k]$ a windowing function and L the number of samples in each segment.

Mel Spectrograms

Humans do not perceive frequencies linearly but in a logarithmic scale. Although the difference between two pairs of sounds, with the first one containing sounds of 500 and 1000 Hz and the second one of 7500 and 8000 Hz, equals 500 Hz in both cases, the difference between the second pair of sounds is almost not noticeable. The Mel Scale (named after the word *melody*) is the result of transforming the frequency scale and constitutes a perceptual scale of pitches, which are judged by listeners to be equal in distance from one another [76]. One of the most commonly used formulas to convert f Hz into m mel is given by [77]:

$$m = 2595 log_{10}(1 + \frac{f}{700}) \tag{2.2}$$

As a result, Mel Spectrograms differ from regular spectrograms in the representation of the frequency, which in this case is achieved using the Mel scale.

Constant-Q Transform (CQT)

CQT is a transform widely used with music audio signals, since it resembles the human auditory system and was introduced by [78]. The purpose of CQT is to overcome the resolution problems interrelated with the DFT. To that end, the ratio of the frequency to the filter bandwidth, known as quality factor, is constant:

$$Q = f/\delta_f \tag{2.3}$$

showing the need for the resolution, or bandwidth, to vary as the frequencies also vary. Assuming that K different filters are used, their window lengths are given by

$$N[k] = \frac{f_s}{\delta f_k} = \frac{f_s}{f_k} Q \tag{2.4}$$

where f_k denotes the center frequency of the k th filter, f_s the sampling frequency and δf_k the width of the k th filter. Since $\frac{f_s}{f_k}$ equals the number of samples processed per cycle at frequency f_k, Q equals the number of cycles processed at the central frequency f_k. The

window function used can be denoted as $W[n,k]$, since its length is determined by $N[k]$, although its shape is the same for all frequency components. As a result, the CQT of the k th spectral component is given by:

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k,n]x[n]e^{\frac{-j2\pi Qn}{N[k]}} \tag{2.5}$$

A high time-resolution is observed at high frequencies and a high frequency-resolution at low frequency bins.

Hybrid Constant-Q Transform (HCQT)

Hybrid Constant-Q Transform is a variation of CQT and has been mainly used in Acoustic Scene Classification tasks [35], [79]. HCQT results from two CQTs with different resolutions, aiming at solving the high-frequency bins issue of the CQT. Assuming that the frame shift contains L samples in the time domain and selecting the k_c th filter where:

$$N[k_c] = 2L \tag{2.6}$$

the frequencies higher than f_{k_c} are treated as high frequencies and the rest as low frequencies. For the high frequency part, the STFT spectrogram is filtered by the filter bank of the high frequency part of CQT, while for the low frequency part the standard CQT is used.

2.2 Machine Learning

2.2.1 The evolution of Artificial Intelligence and Machine Learning

Machine Learning is the ability of a computer program to improve automatically using new data. As stated by Mitchell [80], a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Machine learning can be divided into the following general categories: Supervised Learning, Unsupervised Learning, Competitive Learning and Reinforcement Learning. Artificial Neural Networks (ANNs), ever since their early history, have drawn inspiration from the human brain and have been based on the fact that it executes calculations very differently compared to that of a common computer. What makes the human brain fast and efficient is its complexity and non linearity. The

building blocks of the brain are the neurons, which are organised in such a manner so that certain calculations, for example face recognition, can be executed impressively fast. The human nervous system is a network of 10^{11} neurons, where each of them receives and transmits information. Some of the neuron's main elements are the soma, the dendrites and the axon. The dendrites collect information from other neurons and send it to the soma, while the axon transmits this information to other neurons, as seen in figure 2.2 [62]. The key to the success of the human brain in operations where even the fastest computers cannot cope with, springs from experience. Brains, even from very young ages, have the ability to create their own behavioral rules through the experience acquired, an action tightly connected to the goal of ML models which is not other than learning. The brain's capabilities have motivated scientists to imitate it through ANNs [81].

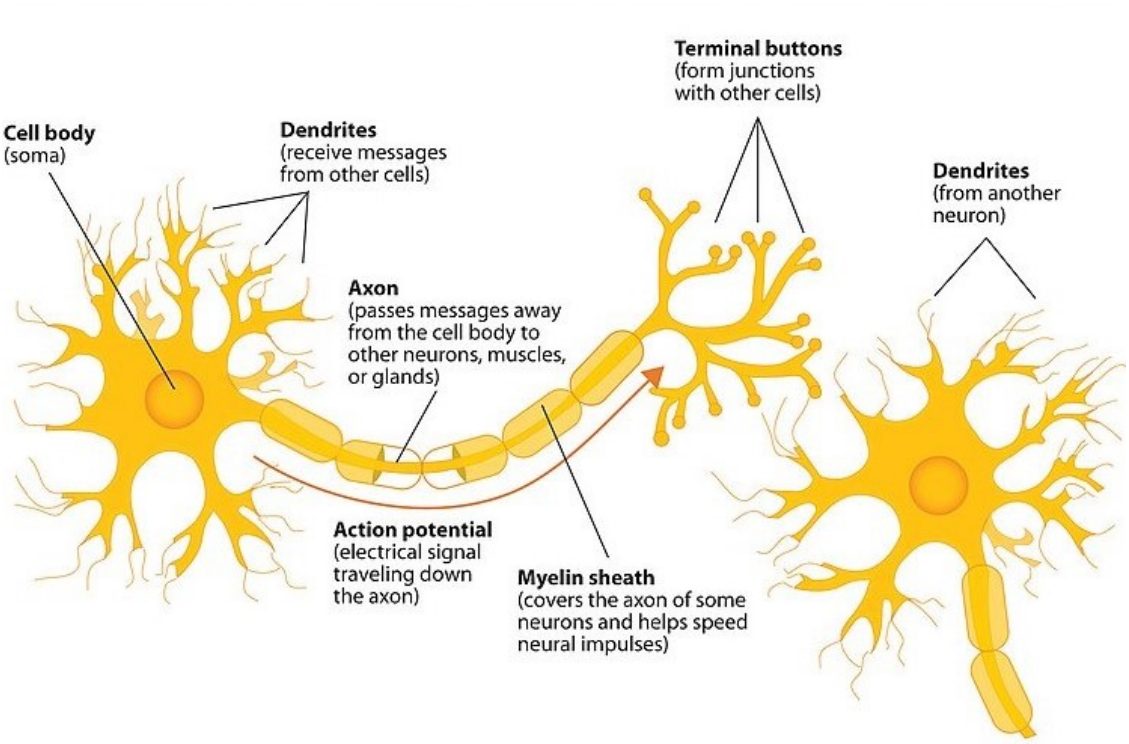


Figure 2.2: The components of a neuron [62]

According to Khan [82] an ANN, or a Neural Network, consists of an input and an output layer of neurons between which, one or several hidden layers of neurons exist. Neurons are the fundamental processing units of an ANN and their in-between connections are associated with a value called weight. Figure 2.4 shows an ANN with the input layer, a hidden layer and the output layer. The structure of a single neuron can be seen in figure 2.3.

The initial stages of Neural Networks can be specified around 1940s with McCulloch and Pitts [52] presenting the first computational model of a neuron. In 1960 Widrow and Hoff [54] created "Adaline", **Ada**ptive **line**ar neuron, a single-layer neural network, based on

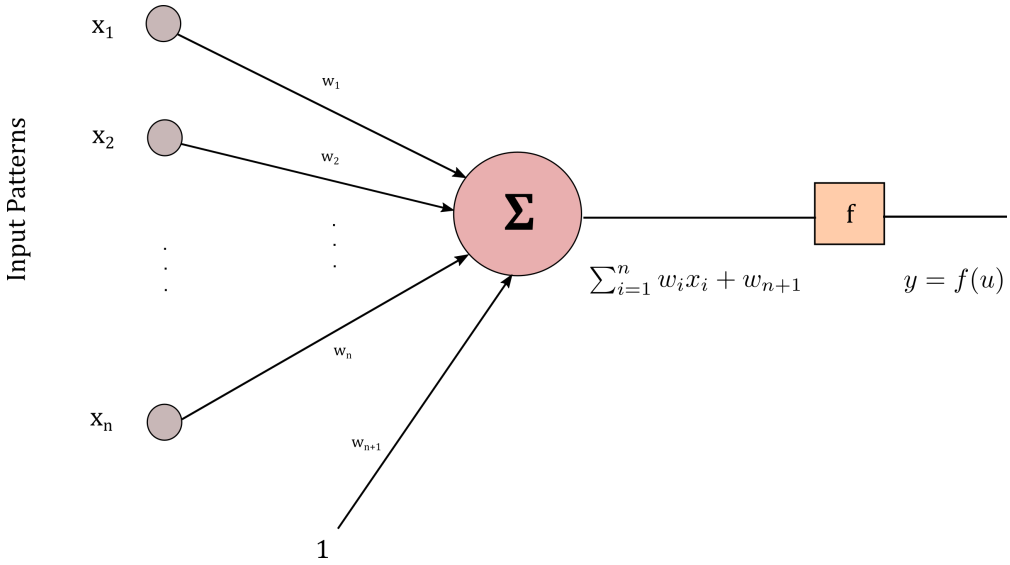


Figure 2.3: The structure of a neuron

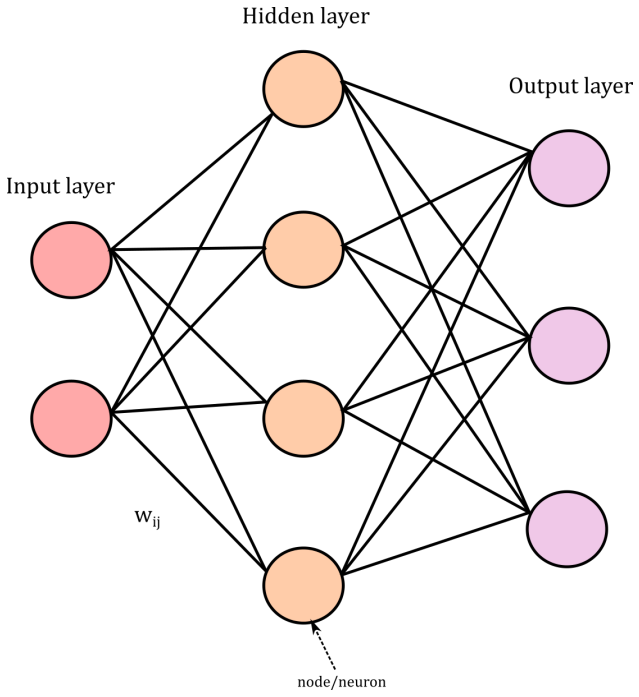


Figure 2.4: An example of an ANN architecture

[52] but differing from it in the learning phase during which the weights are adapted. Around the same era, two research works [53], [83] that presented perceptrons, constituted another important milestone. Perceptrons' innovation lies in the mathematical proof that they always converge to a solution, if the problem is linearly separable. However, Minsky and Papert [55] questioned Rosenblatt's aforementioned work, something of great impact on the AI history, since it discouraged scientists from further research. It was in the 1980's when progress started happening again, with important examples being

the work of Kohonen [56] and Hopfield [57] with the first one proposing Self-organizing maps (SOMs), or Kohonen maps and the second one proposing the Hopfield network, a form of recurrent neural network. One of the first research studies implementing a CNN belongs to LeCun et al. [58], proposing a CNN to identify handwritten postal codes. However, wide usage of CNNs emerged in 2012, after the remarkable classification results on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC2012 [84]) achieved by Krizhevsky et al. [59] and the CNN proposed, named AlexNet.

Some of the cornerstones of artificial intelligence and its evolution over the years can be seen in table 2.1

Subject	Year	Authors
Introduction to Neural Networks	1943	McCulloch and Pitts [52]
The perceptron	1959	Rosenblatt [53]
Adaptive pattern classification machine "Adaline"	1960	Widrow and Hoff [54]
Mathematical proofs about perceptrons	1969	Minsky and Papert [55]
Self-Organizing Map (SOM)	1982	Kohonen [56]
Hopfield Networks	1982	Hopfield [57]
Introduction to CNNs	1989	LeCun et al. [58]
Best results in ILSVRC2012	2012	Krizhevsky et al. [59]

Table 2.1: The evolution of Artificial Intelligence

2.2.2 Convolutional Neural Networks

Convolutional Neural Networks constitute one of the most important advances of Artificial Intelligence and Machine Learning. They are widely used in image classification tasks and implement supervised learning, where each input value is associated with an output value or target. The aim of this type of Neural Networks is to reduce the overall classification error, in order for the model to be reliable. As stated by Gonzalez and Woods [85], a fundamental difference between CNNs and other Neural Networks is the type of expected input, which for a CNN must be 2D arrays, making them highly suitable for tasks related to images. The main operation differentiating CNNs from other NNs is the convolution, a sum of multiplications. The convolution of a kernel w of size $m \times n$ with an image $f(x, y)$ is given by:

$$(w * f)(x, y) = \sum_{s=-a}^{a} \sum_{s=-b}^{b} w(s, t) f(x - s, y - t) \tag{2.7}$$

The kernel slides over all spatial locations of the image in order for all the elements of the 2D array representing the image, to participate in the operation. A bias is added to every

value coming from the convolution of the kernel with each part of the image and the final result is passed through an activation function to acquire one simple value. These values resulting from all the convolutions of the kernel with the image, create a new 2D array, the feature map, which constitutes the input of the following layers. The operations described above are conducted by the cornerstone of a CNN, the convolutional layer.

Other layers typically used in CNNs are the pooling layers. Their goal is reducing the spatial dimensions of the feature maps. Consequently, the number of parameters to learn and the computational effort needed by the network decrease. These layers summarise the features of a region of the feature map at which they are applied. The new, summarised features become the input of the following layers. Pooling is conducted in small regions of the feature map, usually 2×2 areas, while the pooling method can vary, with the two most common methods being max-pooling and average pooling. Max-pooling is conducted by sliding a max filter of size $m \times n$ over all values of the feature map and keeping only the maximum of the values contained in the $m \times n$ patch of the feature map, on which the filter is applied. Average pooling works in the same way as max-pooling, but differing from it on the value being kept. In this case, the average of the values belonging to each patch is kept.

A common problem related to machine learning is overfitting. A model is overfitting when it learns how to efficiently classify the training data but cannot reach such high performance on the evaluation data. The dropout layers are introduced to deal with this problem, by randomly setting input units to 0 following a specific rate.

The final goal of a CNN is the classification of the images according to the task. This can be achieved by inputting the final extracted features into a classifier which consists of fully connected layers, with the conversion of the 2D feature maps acquired to a 1D vector being essential, in order for the features extracted to be inputted in the fully connected layers. The input of a fully connected layer is passed through the activation function used, in order for the output to be acquired. These basic operations executed during a CNN's training are schematically shown in figure 2.5.

As already mentioned, an activation function is needed for the output of a convolutional or a fully connected layer to be acquired. The activation function of a node defines its output, given a specific input. The Rectified Linear Unit (ReLU) activation function is one of the mostly used activation functions in a convolutional layer and is the one used in the models implemented in the current thesis. ReLU is defined by:

$$f(x) = max(0, x) \tag{2.8}$$

Hence the output is the same as the input, if it is positive, otherwise it equals 0. The activation function used for the output layer of the proposed architectures is the sigmoid function, since the task examined is a binary classification task, which is given by:

$$sigmoid(x) = \frac{1}{1 + exp(-x)} \tag{2.9}$$

The implementation of the CNN architectures utilized in the current thesis is achieved using TensorFlow [86], an open source library widely used in ML and Keras [87], a deep learning API written in Python.

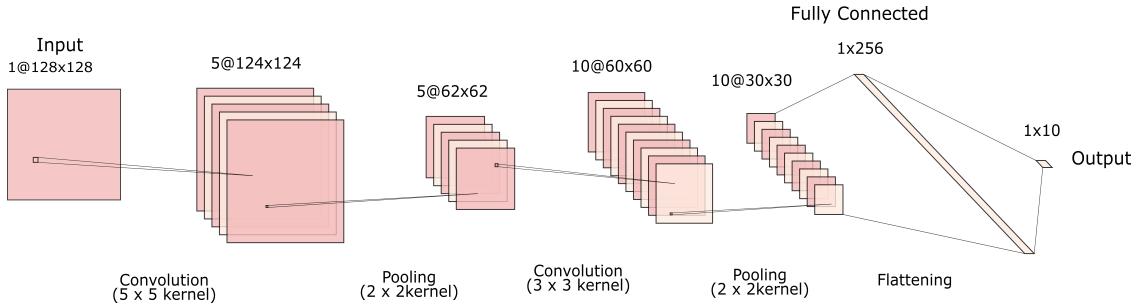


Figure 2.5: An example of a fundamental CNN architecture

2.2.3 Machine Learning and Medicine

Many computer-based algorithms used in medicine are sets of rules encoding knowledge on a specific topic. These rules are implemented in order for decisions, such as the most suitable medicine for a case, to be made. Nonetheless, the continuous rise in the available health care data has provided the opportunity to create and train machine learning models for the assistance of medical decisions in multiple different fields. Machine Learning approaches problems in a human like manner, by processing the available information and learning rules stemming from the data. The main difference between human and machine learning is the data needed in order for the ability of diagnosing a medical condition to be acquired. The human brain is able to learn and recognise patterns using very little data, in contradiction to a machine that requires tremendous amounts of data to reach acceptable performance levels with the quantity and quality of it playing a decisive role. However, a machine that can be trained on thousands or millions of data will eventually be able to recognise diseases or classify X-rays and CT-scans depending on the patient's condition, in opposition to a physician who will not even be able to see such an amount of data throughout his or her career. Nevertheless, great amounts of data are useless without a suitable pre-processing and the right algorithms. As this field of science evolves and the available data improves in quality and quantity, machine learning algorithms could reach better results and lower error rates than trained experts, since they will have been exposed to extremely larger amounts of data with the predictions made being independent of other factors affecting a human's decision, such as fatigue or else exogenous conditions.

According to the Institute of Medicine, diagnosis error is likely to occur in the care of every patient during his or her life, with possibly disastrous consequences [88]. To that end, the

best utilization of the available data and technologies, would prevent such mistakes on the maximum possible degree. Although machine learning models can be trained to diagnose diseases or suggest treatments with extremely high accuracy, errors are always likely to occur and the model's suggestion should be examined by an expert. Thus, machines should be used as a supportive mean, making suggestions about tests that could be conducted, questions that could be asked to a patient, as well as the relevant possible health conditions [89], [90], [91], [92], [93].

2.3 Metrics used for classification assessment

There are four important values produced during predicting the class in which the evaluation samples belong and these are the number of true positive (TP) and true negative (TN) predictions as well as the number of false positive (FP) and false negative (FN) predictions. These values are used to calculate different classification metrics. Some of these metrics used in the current thesis to assess the performance of each model in combination with the dataset examined, are namely: accuracy, sensitivity, precision and specificity. Accuracy is the ratio of the number of correct predictions to the number of total predictions made and can be calculated using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

However, accuracy, solely used, is not a good indicator of a model's performance when the dataset used is imbalanced, as it is in the task examined. This is due to the fact that predicting all the samples as non-Covid could produce very high accuracy results, but by using a model that has not acquired any knowledge on the problem it is required to solve. In health related tasks, a metric that is highly indicative of a model's performance is the Sensitivity, or Recall, which is calculated by formula:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.11}$$

Sensitivity provides information about the number of positive (covid) samples correctly predicted as positive, out of the total number of samples belonging to the positive class. Another metric used is the Precision metric which calculates the number of correct predictions of samples belonging to the positive class out of the total number of samples predicted to belong to this class and is calculated using the following formula:

$$Precision = \frac{TP}{TP + FP} \tag{2.12}$$

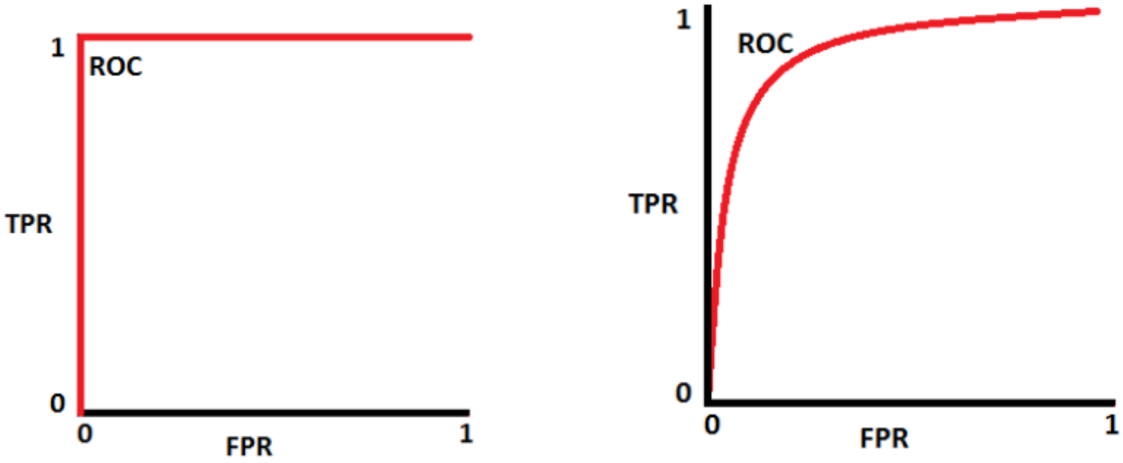
In contrast with Precision, Sensitivity is a metric of higher importance, since predicting a covid sample as non-covid can cause more undesirable consequences than predicting a non-covid sample as covid. Another metric used to assess the performance of a model is Specificity, which is calculated using formula 2.13 and is indicative of the number of negative (non-covid) samples predicted correctly by the classifier.

$$Specificity = \frac{TN}{TN + FP} \tag{2.13}$$

Lastly, the Area Under the Curve (AUC) metric is also calculated. The AUC-ROC curve (Area Under the Curve of Receiver Characteristic Operator) is a probability curve which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold values, with the Area Under the Curve (AUC) value measuring the ability of a classifier to distinguish between the positive and negative class. The TPR equals the sensitivity value, while the FPR can be calculated using formula 2.14 and depicts the percentage of the negative class that was incorrectly classified.

$$False\ Positive\ Rate = \frac{FP}{TN + FP} = 1 - Specificity \tag{2.14}$$

In the ideal situation where the model can completely distinguish between the two classes, the value of the AUC equals 1.0 and the AUC-ROC curve is given by figure 2.6(a). However, when the model cannot fully distinguish the two classes, the value of the AUC metric would range between [0.0,1.0). The higher the AUC value, the higher the possibility of the model distinguishing between the two classes. For instance, an AUC value of 0.7 means that the chance of the model distinguishing the two classes equals 70%. The AUC-ROC curve in such a situation can be seen in figure 2.6(b). In the worst possible situation where the model cannot distinguish between the two classes, the AUC value would equal 0.5 and the relative AUC-ROC curve is presented in figure 2.6(c). Lastly, when the AUC value equals 0.0 the model predicts all the positive samples as negatives and vice versa, with the relative AUC-ROC curve being depicted in figure 2.6(d).



(a) AUC-ROC curve for AUC = 1.0

(b) AUC-ROC curve for AUC = 0.7

(c) AUC-ROC curve for AUC = 0.5

(d) AUC-ROC curve for AUC = 0.0

Figure 2.6: An explanation of the AUC-ROC curve [63]

Chapter 3

Deep Learning Methods for the detection of COVID-19

3.1 Datasets

Three different datasets or subsets of them containing respiratory sounds from Covid and non-Covid users have been used.

3.1.1 Cambridge dataset

The Cambridge dataset is a crowd-sourced dataset containing breath and cough sounds recorded via an android and a web application [45]. The dataset is shared with us under a data-sharing agreement and contains breath and cough audio samples from users declaring to belong in one of the following categories: healthy with no symptoms, healthy with cough as a symptom, tested positive in Covid-19 with cough as a symptom, tested positive in Covid-19 but do not have cough as a symptom and users with asthma reporting to have cough. The specific distribution of the samples in each of the different users' categories, as well as the recording method used (android or web), is shown in figure 3.1.

Only three of the above categories were used. More specifically, cough samples from healthy users without any symptoms, recorded either through the android or the web application comprised the non-Covid samples used, while cough samples from users declaring to have tested positive in COVID-19 that may or may not have had cough as a symptom and have been recorded either via the android or the web application, comprise the Covid samples used. The part of the dataset used, contains 141 Covid samples acquired from 66 different users and 298 non-Covid samples acquired from 220 different users. No other

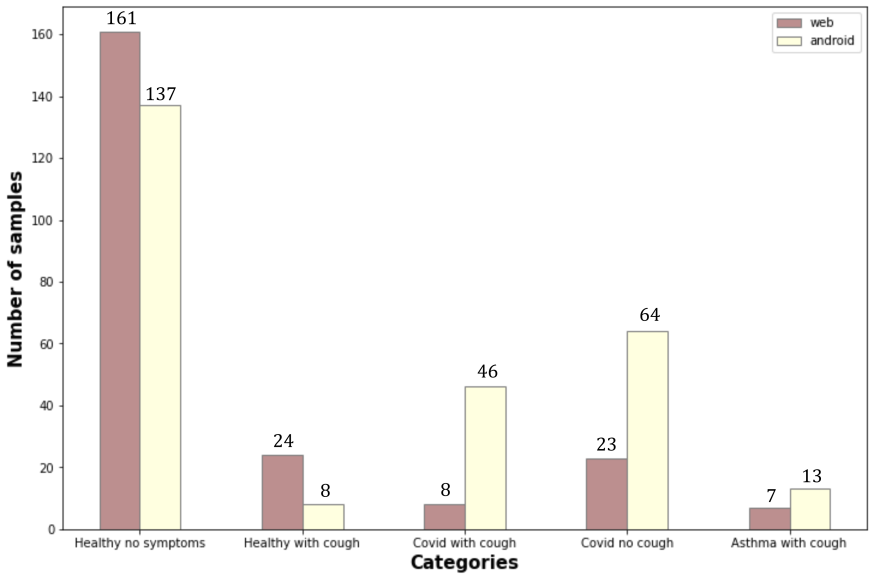
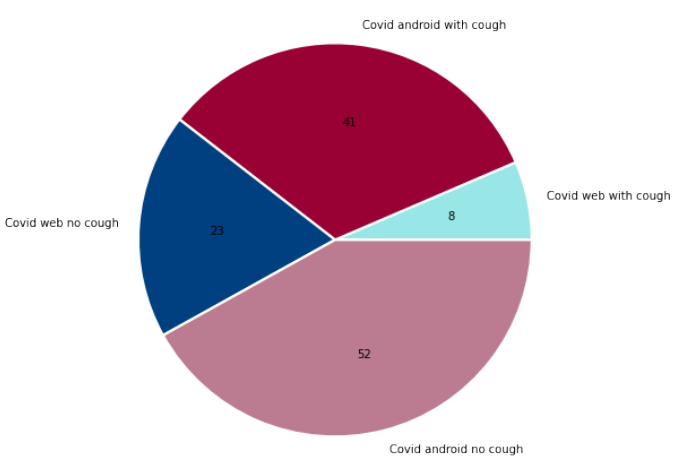


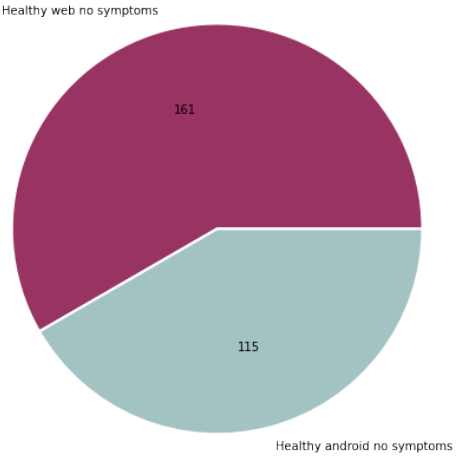
Figure 3.1: Distribution of cough and breath samples in the different categories

metadata such as the age or the gender of the user have been shared.

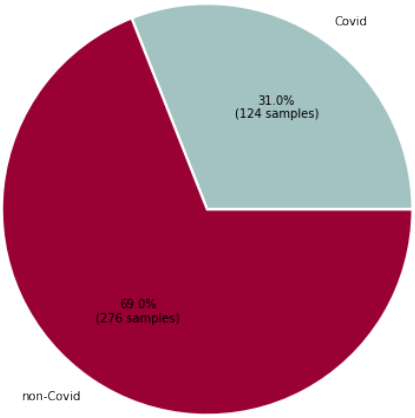
Each of the samples given is associated with a unique user ID, whether it has been recorded through the android or the web application. No user has recorded more than one samples using the web application. However, some users gave more than one sample using the android application, providing the ability to track possible changes in their physical health. These users will be referred to as "returning users". As far as the three android categories used are concerned, none of the returning users' condition changed in the new audio recordings. From the total 247 android users, 30 recorded samples more than once. Each sample is accompanied by a unix timestamp in milliseconds, providing information about the exact date and time at which it was recorded. These timestamps are used for tracking the time interval between samples provided by the same user. Nevertheless, some of the returning users recorded their samples in less than 24 hours after the previous recording. These samples were considered to not offer any new information about the user's condition and were not included. In total, 40 samples were not used, creating a new distribution of the covid and non-covid samples in each category, shown in figures 3.2(a) and 3.2(b). More specifically, concerning the audio files recorded via android, 12 audio samples from users tested positive but without cough as a symptom, 5 samples from users who tested positive and had cough as a symptom and 22 audio samples from healthy users without symptoms were not included in the classification task, because of the time interval between them and the previous or the following recording being smaller than 24 hours. The total number of Covid samples used, regardless the mean of recording and the existence of cough as a symptom, are 124 with the total number of non-Covid samples being 276 as shown in figure 3.2(c).



(a) Distribution of covid cough samples



(b) Distribution of non covid cough samples



(c) Distribution of covid and non-covid samples

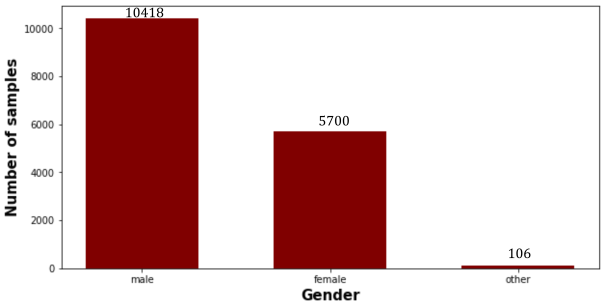
Figure 3.2: Number of samples in each category for the Cambridge dataset

3.1.2 COUGHVID dataset

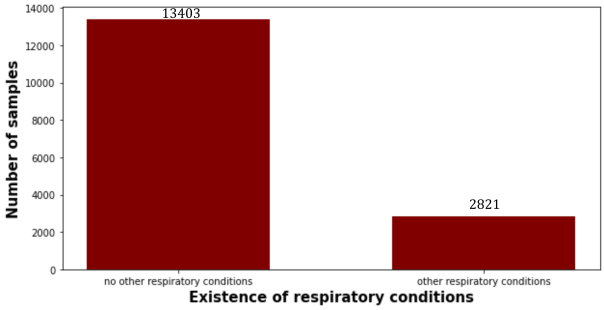
The COUGHVID dataset [94] is another crowd-sourced dataset containing cough audio samples recorded through a web application. The version of the dataset used in the current thesis contains 27,550 cough samples, each from a different user. Metadata information about each user recording a cough sample is also provided. More specifically, information about the geographical coordinates, the age, the gender and the respiratory condition of the user, is collected. Moreover, the users can self-report information about their health status as COVID-19, symptomatic, i.e. declaring they have symptoms but no diagnosis, and healthy. Not all metadata information was given by all users. Only 55.24% (15,218 samples) of the users have provided age information with the average age being 36.8 years. The frequency of audio samples based on the user's age is shown in figure 3.3(e). From the total available samples, 58.89% (16,224 samples) contain information about the gender, the user's respiratory condition and the existence of fever or muscle pain. The same amount of samples are accompanied by health status information. More details about the distribution of the samples in each of the aforementioned categories are shown in figures 3.3(a)- 3.3(d).

Since crowd-sourced data can contain samples of poor quality, a classifier that provides the probability with which a given audio sample contains cough is also shared. Moreover, the results of the classifier for each of the shared audio samples is contained in the metadata information. A probability threshold of 0.8 is used, as suggested by Orlandic et al. [94], in order for audio samples that do not contain cough to be excluded in the maximum possible extent. Out of the total 27,550 samples 54.9% (15,125 samples) are considered to contain cough according to the aforementioned cough classifier. Some metadata statistics have been calculated only for the samples classified as cough samples. The user's age was given in 10,291 samples (68.04% of the total cough samples), with the average age being 36.44 years. Information about the gender, the user's respiratory condition, the existence of fever or muscle pain and the health status was included in 10,819 samples (71.53% of the total cough samples). The distribution of the cough samples according to the metadata information is shown in figures 3.4(a)- 3.4(d). Due to the fact that 10,819 samples, out of the 15,125 cough samples, contain status information as it has been declared by the user, these are the samples that can be used and they are classified as Covid or non-Covid based on the user defined health status. As shown in figure 3.5 this dataset contains 699 Covid samples and 10,120 non-Covid samples, considering as non-Covid samples those deriving from users declaring to be either healthy or symptomatic, but without a diagnosis.

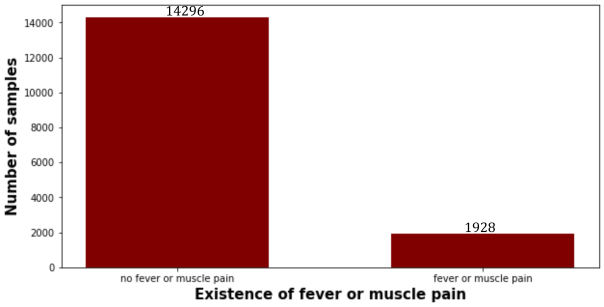
Apart from the publication of the crowd-sourced data and metadata information about them, some of these samples were annotated by four expert physicians to improve the quality of the dataset. Each expert annotated 1,000 samples. The total number of samples annotated by at least one expert amounts to 2,804, with expert 1 and expert 2 having annotated 802 samples, expert 3, 796 samples and expert 4, 803 samples with 129 of these samples being annotated by all four experts. Figure 3.6 shows the distribution of samples in the two classes as labelled by each expert.



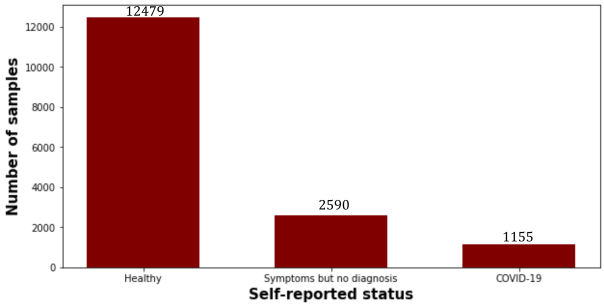
(a) Gender distribution of samples



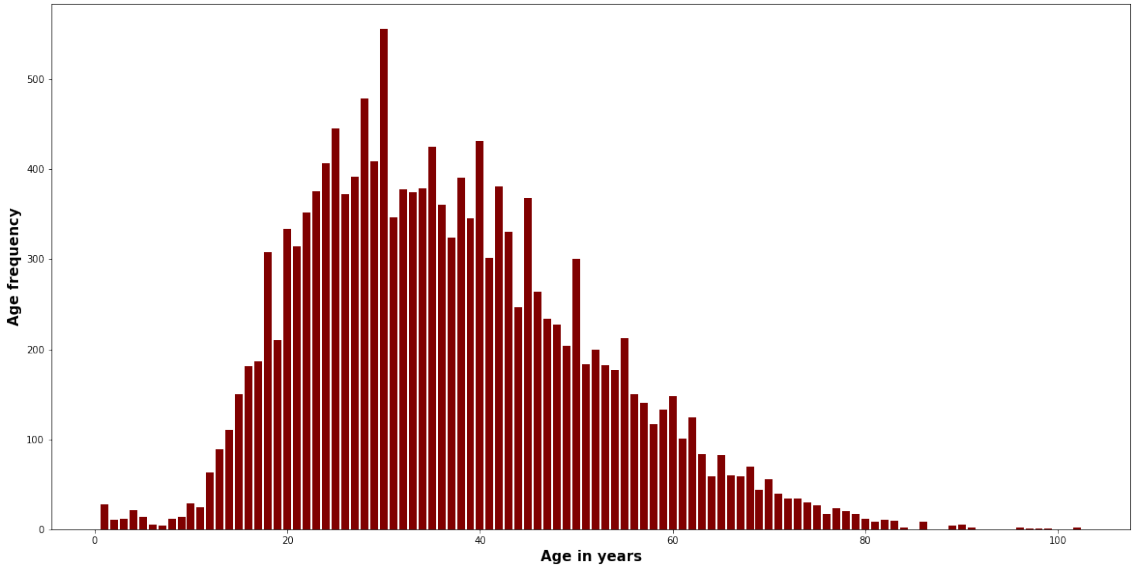
(b) Existence of respiratory conditions



(c) Existence of fever or muscle pain



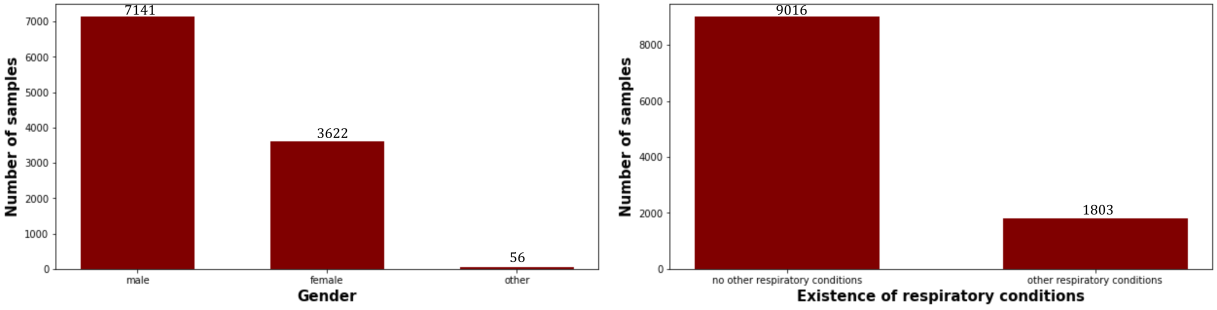
(d) Self-reported health status distribution



(e) Samples per age distribution

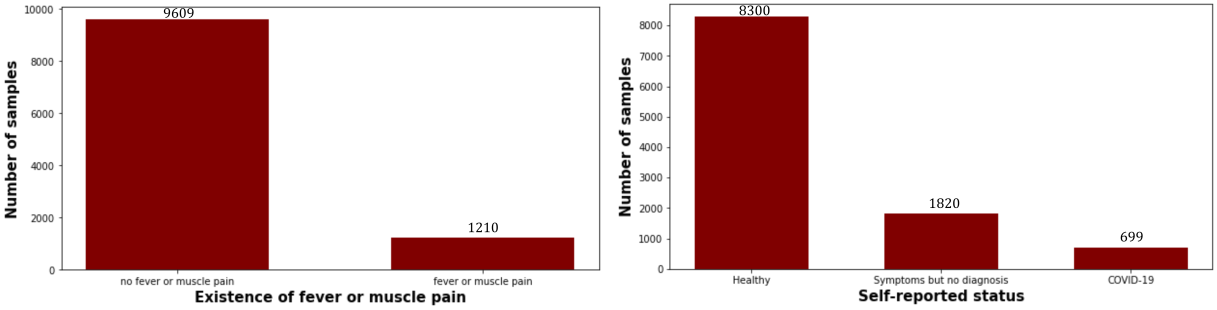
Figure 3.3: Metadata statistics for the COUGHVID dataset

An interesting observation on the experts' annotation is the fact that 621 samples have been labelled as samples from users infected by COVID-19 by at least one expert, 26 samples by at least two experts, 2 samples by at least three experts, while no sample has been labelled as a COVID-19 sample by all experts. This shows a high discordance between the four experts, confirming the difficulty existing in the classification of a cough sample as



(a) Gender distribution of cough samples

(b) Existence of respiratory conditions



(c) Existence of fever or muscle pain

(d) Self-reported health status distribution

Figure 3.4: Metadata statistics for the cough samples of the COUGHVID dataset

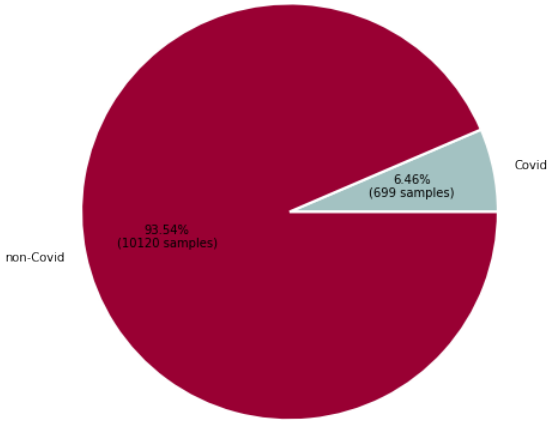


Figure 3.5: Distribution of Covid and non-Covid samples for the COUGHVID dataset

Covid or non-Covid. Figure 3.7 presents the number of samples annotated as Covid that have also been labelled with some other audible respiratory condition by at least one expert. It is observed that most of the Covid samples do not present any audible respiratory condition.

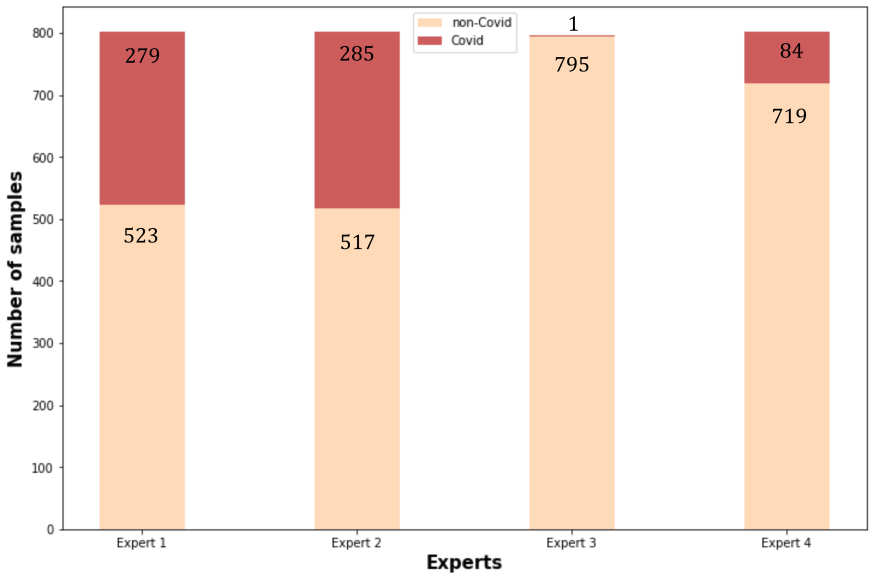


Figure 3.6: Annotations of samples as covid and non-covid by each expert

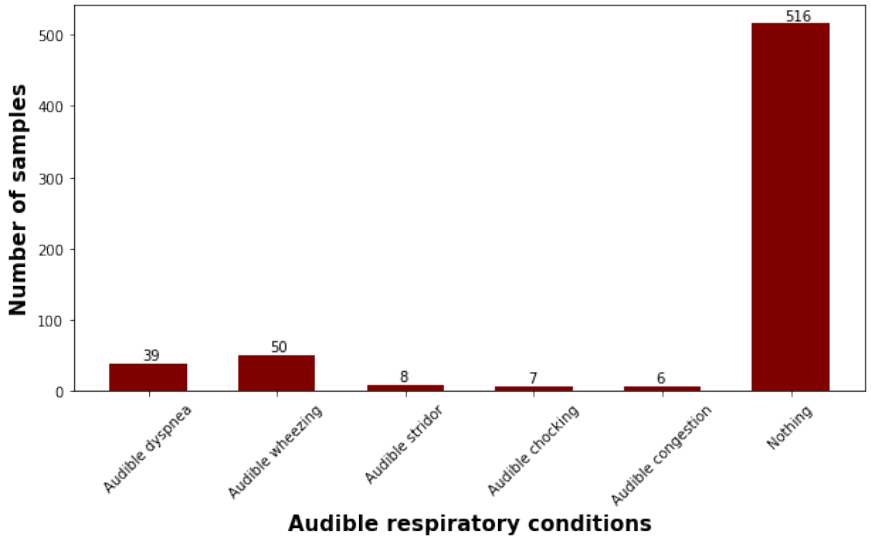


Figure 3.7: Number of samples with other audible respiratory diseases as annotated by the experts

There occur 124 contradictions between the experts' annotations as they can be seen by the different labels given to the same sample when annotated by more than one expert. To that end, in order to classify a sample in the covid or the non-covid class, the agreement of each expert's annotations with the status given by the user was calculated. When a sample is annotated by more than one experts and the labels given are not the same, the sample is classified in the class indicated by the expert with the higher rate of agreement with the status given by the users. As shown in figure 3.8 this subset of the COUGHVID dataset contains 553 covid samples and 2,251 non-covid samples.

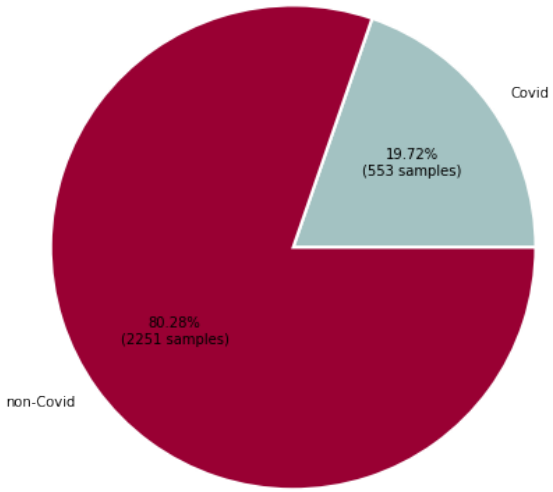


Figure 3.8: Distribution of covid and non-covid samples in the annotated subset of the COUGHVID dataset

Although the annotated dataset is noticeably smaller than the original dataset, it is less imbalanced.

3.1.3 Coswara dataset

The Coswara dataset [72] is an open access dataset containing cough, breath and speech sounds both from healthy and from COVID-19 infected users. It contains two types of cough sounds, heavy and shallow, two types of breath sounds, shallow and deep, sustained vowel phonation and two types of one to twenty digit counting, normal and fast paced. Each user provided 9 different audio samples, but for the purposes of the current thesis only the two types of cough sounds, heavy and shallow, have been used as separate datasets. Each audio sample is accompanied by metadata information including the age, the gender, the location, the current health status of the user, as well as information about the presence of co-morbidity. All audio files have been manually assessed, with regard to the quality of the audio sample and the category it belongs to, by 13 annotators with each file being annotated once. Each of the 9 categories contains 1,569 samples and their distribution in the possible health status (healthy, no respiratory illness exposed, not identified respiratory illness, positive mild, positive moderate, positive asymptomatic and fully recovered) is shown in figure 3.9. The samples with a status in one of the three categories i.e. positive mild, positive moderate and positive asymptomatic are classified as Covid, with the rest of the samples being classified as non-Covid.

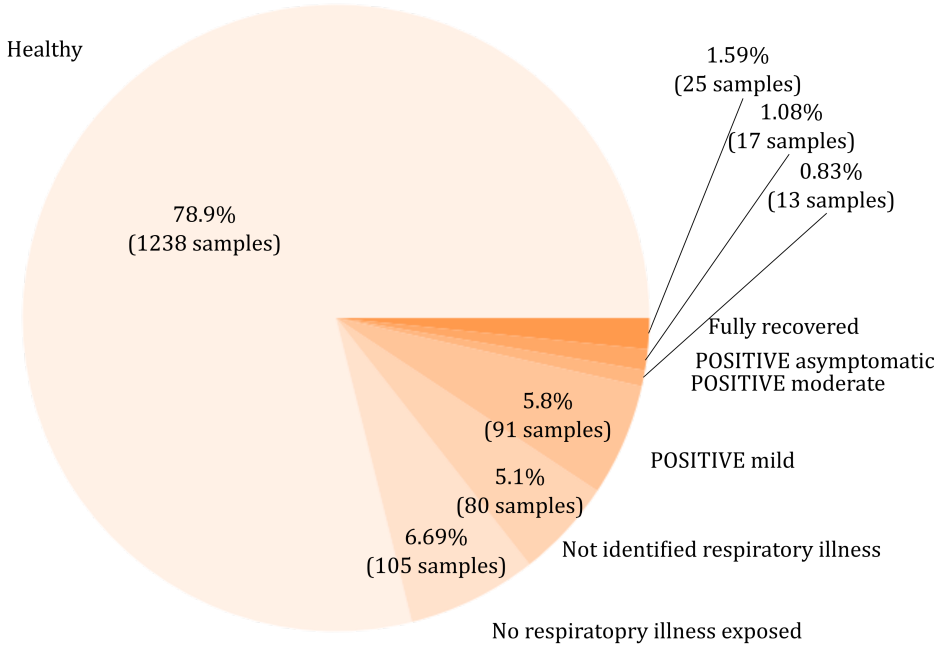


Figure 3.9: The health status distribution of the samples in the Coswara dataset

As for the metadata provided, the user's age was given in all available samples, with the average age being 33.22 years. The frequency of audio recordings based on the age of the

user is shown in figure 3.10.

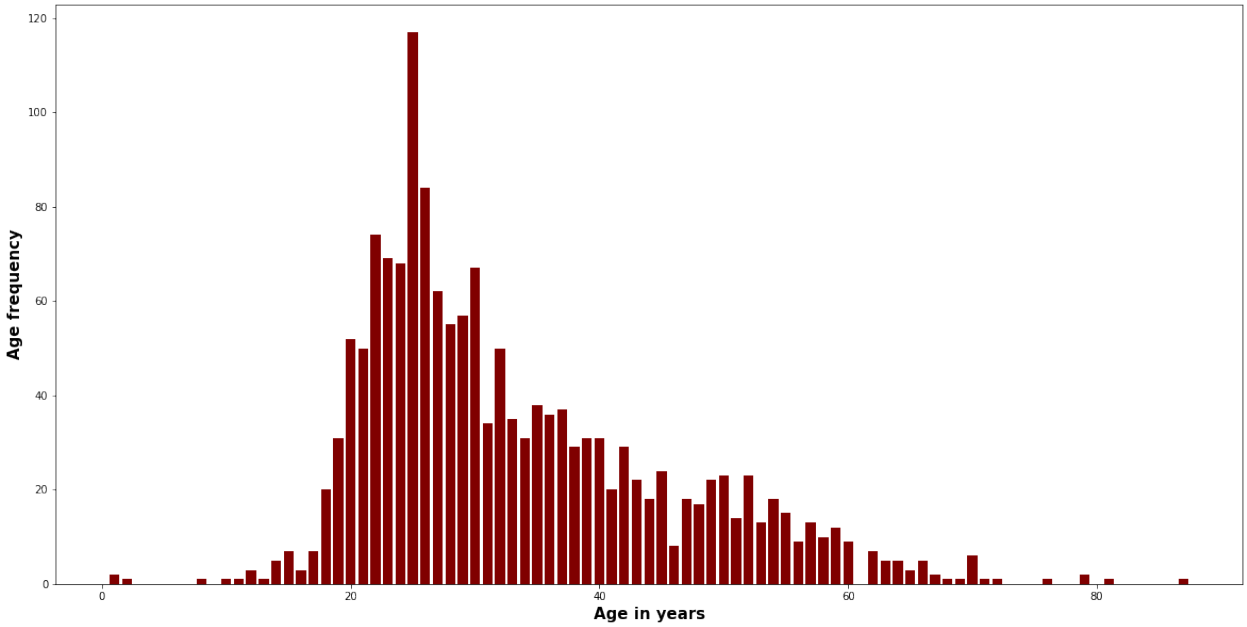
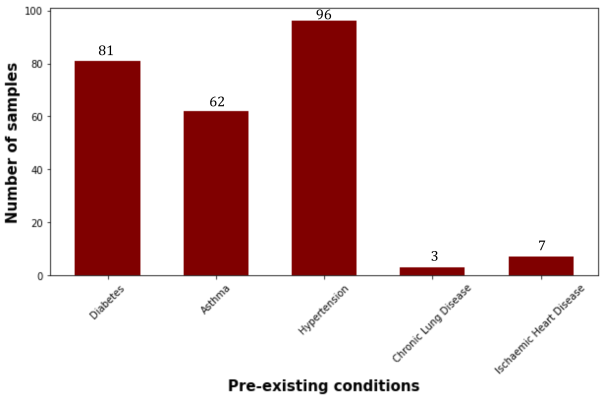


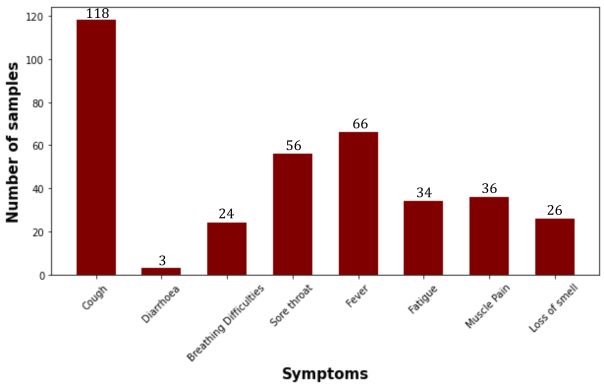
Figure 3.10: The samples per age distribution for the Coswara dataset

As for the gender distribution, 1,175 samples belong to male users and 395 to female. Moreover, 81 users declare to have diabetes as a pre-existing condition, 62 users declare to have asthma, 96 users declare hypertension, 3 users reported having a chronic lung disease and 7 users declared having ischaemic heart disease as a pre-existing condition. As for the symptoms reported, 118 users declared cough as a symptom, 3 users declared diarrhoea, 24 users reported breathing difficulties, 56 users declared sore throat, 66 users reported fever, 34 users fatigue, 36 users muscle pain and 26 users reported loss of smell as a symptom. Last but not least, 113 users declared to be smokers, 5 users reported having pneumonia and 88 users a cold. Out of the total 1,569 different samples available, information about the Covid test status is provided only in 156 of them, with 26 users declaring to have been tested positive, 44 users declaring having been tested negative and 86 users declaring to not have taken a test. Although some users have declared to be returning users, i.e. having previously recorded samples, a unique ID is assigned to each sample making the recognition of samples deriving from the same user impossible. However, the users that have denoted to be returning users are only 40 (2.55% of the total number of users). The above information is schematically presented in figures 3.11(a)-3.11(d).

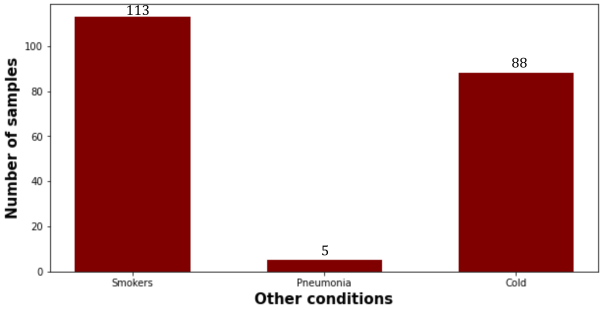
Due to some files containing bad audio quality, 1,541 cough-heavy samples and 1,539 cough-shallow samples are used. From the total of 1,541 cough-heavy samples, 103 of them are Covid samples (6.68%) with the other 1,438 being non-Covid samples (93.32%). As for the distribution of cough-shallow samples in the Covid and non-Covid classes, 103 of them belong to the Covid class (6.69%) and the other 1,436 samples (93.31%) belong to the non-Covid class. These can also be seen in figures 3.12(a) and 3.12(b).



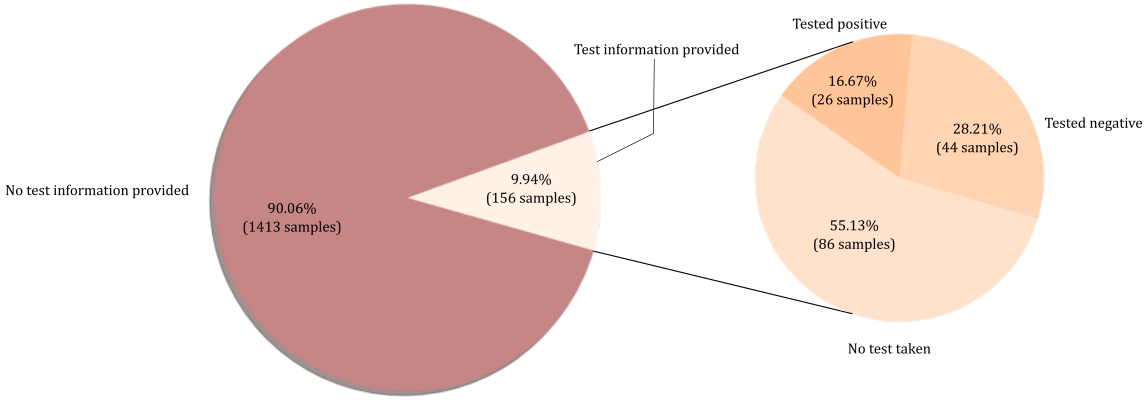
(a) Pre-existing conditions



(b) Symptoms



(c) Other conditions

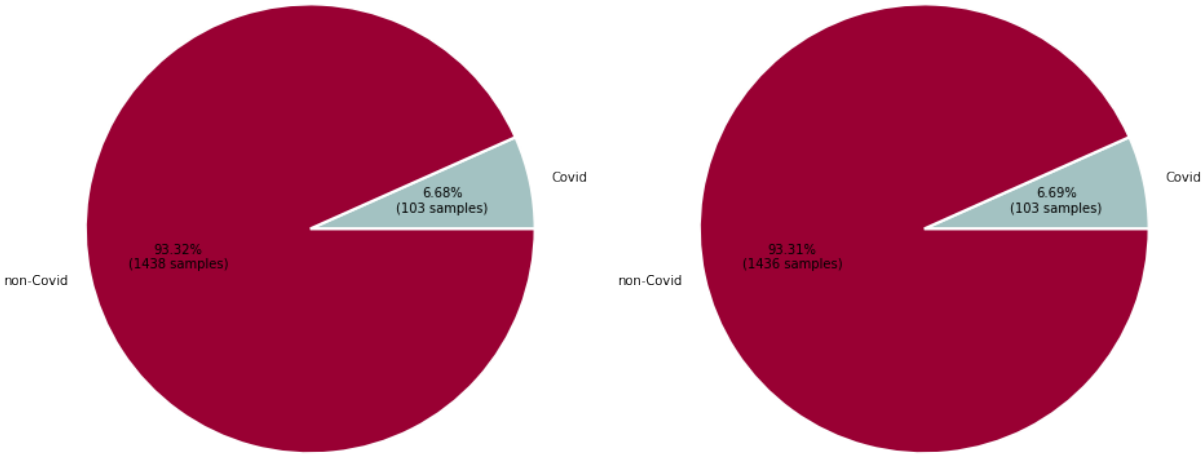


(d) Test information

Figure 3.11: Metadata statistics for the Coswara dataset

3.2 CNN architectures used

Six CNN architectures were used in total, with three of them being small CNN models created for the current task and the rest being already existing pre-trained models. The



(a) Distribution of Covid and non-Covid samples in the Coswara cough-heavy dataset

(b) Distribution of Covid and non-Covid samples in the Coswara cough-shallow dataset

Figure 3.12: Statistics about the Coswara cough heavy and shallow datasets

three CNN architectures created, will be referred to as Model 1, Model 2 and Model 3, as detailed in subsections 3.2.1, 3.2.2 and 3.2.3 respectively. As for the pre-trained models, ResNet-50, DenseNet-201 and Xception were examined, since they account for three of the highest accuracy scoring models on the ImageNet classification task.

3.2.1 Model 1

Model 1 consists of two convolutional layers, each one containing 16 nodes, with the kernel used in the first one being 9×3 and in the second one being 5×3. The activation function used in both convolutional layers is the Rectified Linear Unit (ReLU). Each convolutional layer is followed by a max-pooling layer, with a pooling window of size 2×2 and an equal stride and a dropout layer with a dropout rate of 0.2. These layers are followed by a flatten layer, a dropout layer with a dropout rate equal to 0.4 and a dense layer with one node and sigmoid activation function. The Adam optimiser is used with a learning rate of 0.0001 and the loss function used is the binary cross entropy. This model constitutes a variation of the architecture utilized by Amoh and Odame [36] for a cough detection related task. This architecture is schematically described in figure 3.13.

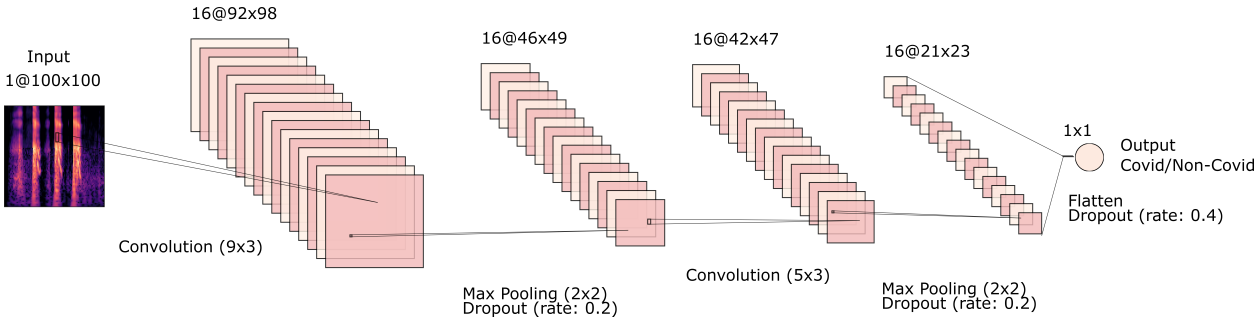


Figure 3.13: The architecture of Model 1

3.2.2 Model 2

Model 2 consists of three convolutional layers. The first one contains 4 nodes, using a kernel of size 5×5 and ReLU activation function. It is followed by a dropout layer with a dropout rate equal to 0.3 and an average pooling layer with the size of the pooling window and the stride being 2×2. These three layers are repeated and the difference between the second and the first convolutional layer lies in the number of nodes which equals 16. The third convolutional layer contains 32 nodes, with the kernel size being 5×5 and the activation function used being ReLU. It is followed by a dropout layer with a dropout rate of 0.3, a flattening layer, a dropout layer with a dropout rate equal to 0.5 and the dense layer containing one node and using sigmoid activation function. The Adam optimiser is used with a learning rate of 0.0001 and the binary cross entropy loss. A schematic illustration of Model 2 can be found in figure 3.14.

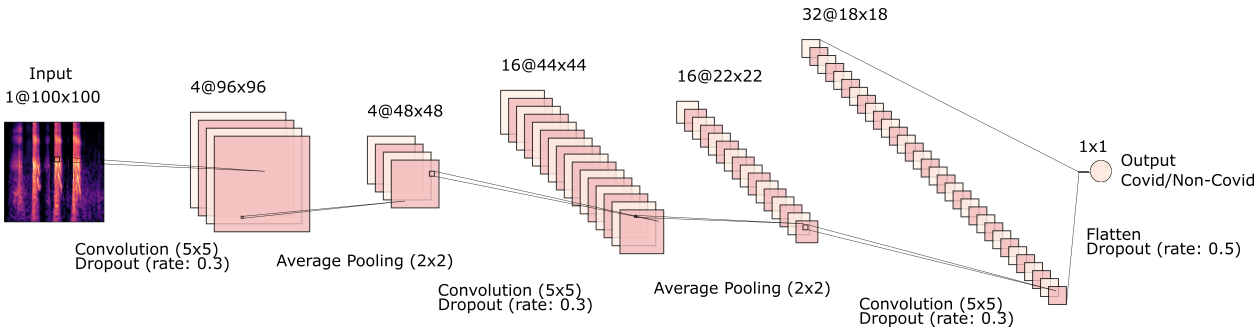


Figure 3.14: The architecture of Model 2

3.2.3 Model 3

Model 3 differs from Model 2 in the configuration of the parameters. More specifically, the differences lie in the following parameters: No dropout layers are used and the number of nodes used in the first convolutional layer equals 10, with a 7×7 kernel size. The nodes used in the second convolutional layer are 30 with the size of the kernel being set to 5×5.

In the third convolutional layer, the number of nodes used equal 100, with a kernel of size 3×3. The RMSprop optimiser is used with a learning rate of 0.001 and the discounting factor for the gradient being 0.4. Moreover, the loss function used is mean squared error. The modifications in the above parameters are a result of implementing hyper-parameter optimisation in Model 2 using the annotated COUGHVID dataset transformed with HCQT. The model is also depicted in figure 3.15.

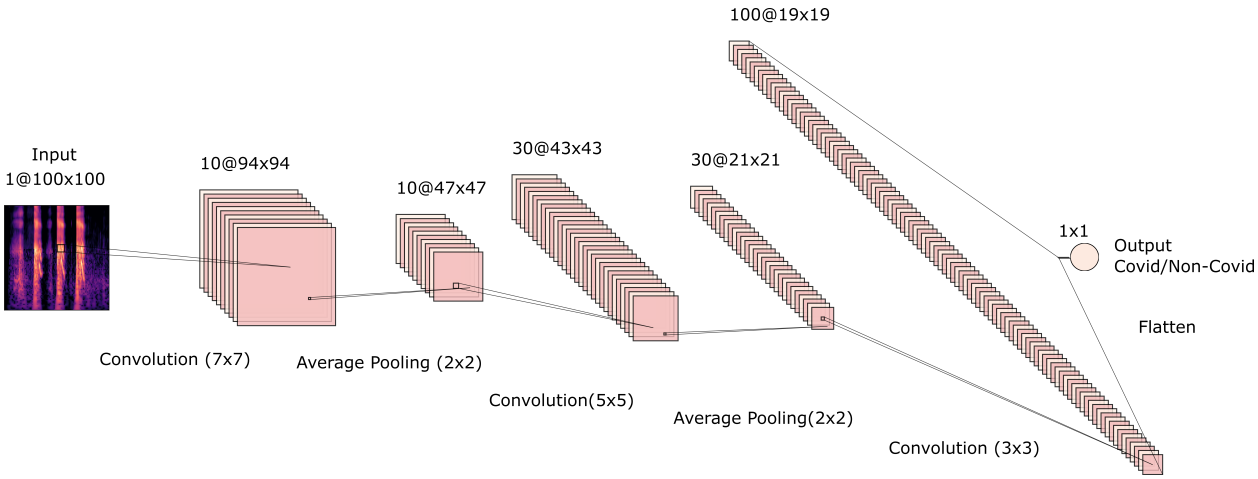


Figure 3.15: The architecture of Model 3

3.2.4 ResNet model

The ResNet architecture implements residual learning and won the 1st place on the ILSVRC 2015 classification task [95]. The current thesis utilizes the ResNet-50 architecture which has been widely used in cough classification tasks [31] and especially in Covid related tasks achieving remarkable results [46], [50]. The initial weights used are the ones acquired when training the model on the ImageNet dataset. A classification head consisting of a dropout layer with a dropout rate equal to 0.5, a global average pooling layer and a dense layer with sigmoid activation function, is added to the convolutional base of the model. The Adam optimiser is used with a learning rate of 0.0001 and the binary cross entropy loss.

3.2.5 DenseNet model

The Dense Convolutional Network (DenseNet) was introduced by Huang et al. [96] and its innovation lies in the fact that each layer is connected to every other layer in a feed-forward fashion. A traditional convolutional network with L layers has L connections while DenseNet has $\frac{L(L+1)}{2}$ connections. The DenseNet-201 architecture is implemented in the current thesis. The initial weights used are the ones acquired when training the model

on the ImageNet dataset. The classification head of the model is replaced by a dropout layer with dropout rate equal to 0.5, a global average pooling layer and a dense layer with sigmoid activation function. The Adam optimiser is used with a learning rate of 0.0001 and the binary cross entropy loss.

3.2.6 Xception model

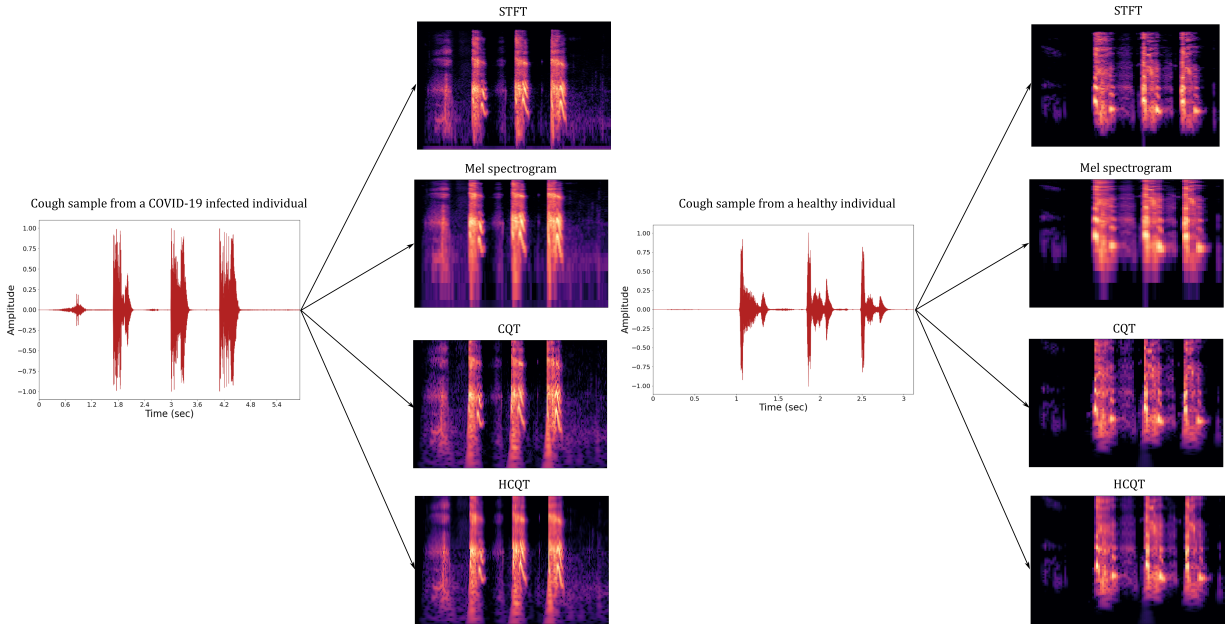
Xception is a deep CNN architecture inspired by Inception [97]. However, the Inception modules have been replaced with depthwise seperable convolutions, achieving slightly better performances than InceptionV3 architecture on the ImageNet dataset [98]. The initial weights used are the ones acquired when training the model on the ImageNet dataset. A classification head consisting of a dropout layer with dropout rate equal to 0.5, a global average pooling layer and a dense layer with sigmoid activation function is added to the convolutional base of the model. The Adam optimiser is used with a learning rate of 0.0001 and the binary cross entropy loss.

3.3 Implemented Methods

The aforementioned datasets, or subsets of them, have been used to train and test multiple CNN architectures. The general method followed is a 5-fold cross validation method, experimenting with different CNN architectures and combinations of them, amalgamating them with different datasets and audio to image transformations. The audio samples are initially converted to images using one of the four transformations described in section 2.1.2. The obtained images are then used as the dataset and the following methods described are implemented on them. An example of a cough audio sample from a Covid infected individual and from a healthy individual, converted using all four transformations described, is presented in figure 3.16(a) and 3.16(b) respectively.

3.3.1 5-fold cross validation

In this method, the dataset is divided in 5 different folds, each one containing approximately $1/5$ of the dataset's samples. Each one of the five folds is used as a test set exactly one time, while the four remaining folds comprise the train and validation set. The train set contains three folds while the validation set is comprised of one fold, creating a 60%-20%-20% split for the train, validation and test set respectively, as shown in figure 3.17. Each fold is also used as a validation set exactly once, reassuring that all the data will be used precisely one time in the validation and the test set.



(a) Conversion of a cough sample from a COVID-19 infected individual with cough symptoms using multiple transformations

(b) Conversion of a cough sample from a healthy individual without any symptoms using multiple transformations

Figure 3.16: Examples of converting audio to image

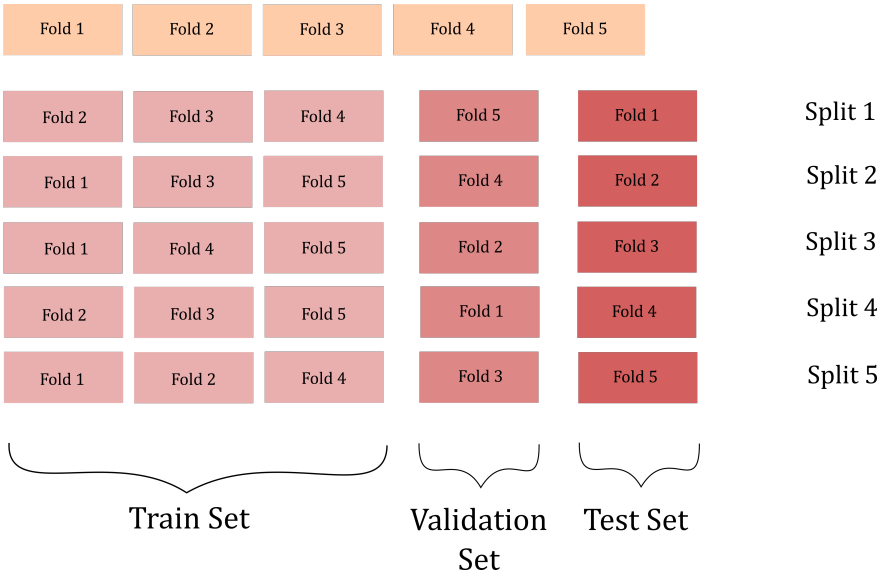


Figure 3.17: Description of the 5-fold cross validation data split

During training, Synthetic Minority Oversampling Technique (SMOTE) is applied for dealing with the under representation of the minority class. SMOTE is an over-sampling technique that has been used in cough classification tasks for dealing with imbalanced datasets [99]. It has also been implemented in Covid related, image classification tasks as well as cough classification tasks using CNN architectures [100], [50]. SMOTE over-samples the

minority class by creating synthetic examples. These are samples created along the line segments that join some or all of the k nearest neighbours. The neighbours to be used are randomly chosen from the k nearest neighbours of a sample, depending on the amount of oversampling required in the minority class [101]. The creation of synthetic samples using SMOTE is schematically depicted in figure 3.18. The SMOTE class from the imbalanced-learn library [102] is employed, with the number of nearest neighbours used to construct synthetic samples being set to 5 and the random state used to control the randomization of the algorithm being set to 42.

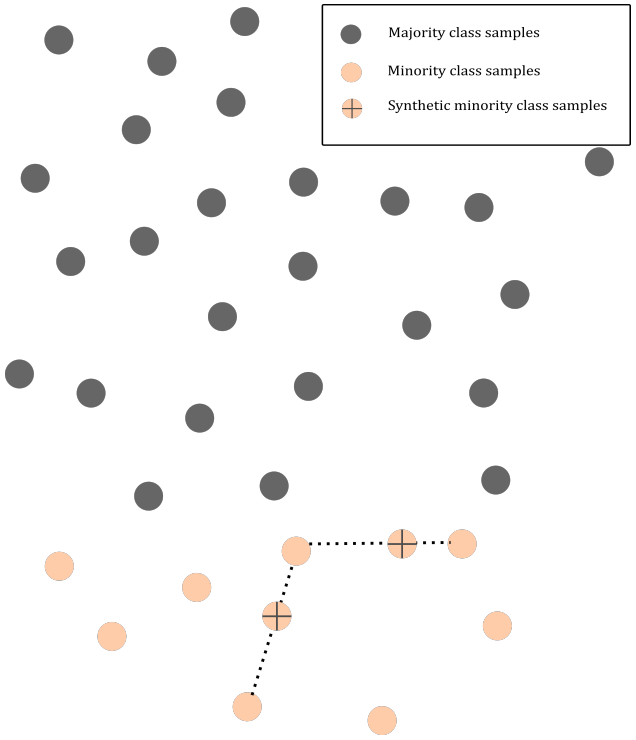


Figure 3.18: Synthetic Minority Oversampling Technique (SMOTE)

The Hybrid Constant-Q Transform was implemented in all datasets, with the sampling rate being equal to the initial sampling rate of the signal, the number of samples between successive CQT columns being set to 512, while the window function used is Hann window. Moreover, the Mel Spectrograms were also examined and the conversion of the audio samples was conducted, with the sampling rate being equal to the signal's initial sampling rate. In order for these audio transformations to be obtained, Librosa, a Python package for music and audio analysis was used [103]. The three CNNs created, as well as the ResNet-50 and the DenseNet-201 architectures, were tested. Since the available datasets are imbalanced, an ensemble method is also examined and is explained in detail in section 3.3.2.

3.3.2 Ensemble method

Due to the nature of the problem, the available datasets are highly imbalanced with the vast majority of the samples belonging to the non-Covid class. As a consequence, it can be difficult for a model to become able to distinguish between the two classes, rendering itself inappropriate for the specific task. However, ensemble methods have been widely used with imbalanced data, showing promising results in enhancing a model's performance [104], [105]. The ensemble method was implemented for the subset of the COUGHVID dataset that contains only annotated cough samples and for the Cambridge dataset. Four different audio to image conversions were applied in this method and they namely are: HCQT, Mel Spectrograms, CQT and STFT. The configuration of the first two is the same as the one described in section 3.3.1. As for the CQT transform, the sampling rate is equal to the initial sampling rate of the signal, the number of samples between successive CQT columns is set to 512, while the window function used is Hann window. The STFT transformation is obtained by setting the n_fft parameter, which denotes the length of the windowed signal after padding with zeros, to 2048 with the window function used being the Hann window. All of the aforementioned transformations were acquired using Librosa [103]. In order for the validity of the acquired classification results to be ascertained, the ensemble method is combined with the 5-fold cross validation method described in section 3.3.1. To that end, the separation of the train set used in the ensemble method refers to the train set regarding a particular split out of the 5 different dataset splits used.

Ensemble method for the annotated COUGHVID dataset

Since the percentage of Covid samples in the annotated COUGHVID dataset equals 19.72% of the total samples with the non-Covid samples constituting the other 80.28%, the ensemble model consists of four models, in order for the train samples to be almost equally distributed between the two classes. The same CNN architecture is used four separate times, trained using a different part of the dataset each time. The validation and test set are preserved the same. Testing each of the four models using the same test set provides four probability outputs for each one of the test samples. The average of these four probabilities acquired is used as the classification probability for the corresponding test sample. This probability is converted to the final prediction referring to this sample using a probability threshold 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. The splitting of the dataset for the current case is schematically depicted in figure 3.19.

The CNN architecture, depicted as a "black box" in figure 3.19, is the same for the four models. The difference between these four models lies in the dataset used for their training. Each model is trained on 1/4 of the non-Covid samples and on all of the Covid samples belonging to the train set, as it results from the initial split of the dataset. The assignment of data samples to each model can also be seen in figure 3.20.

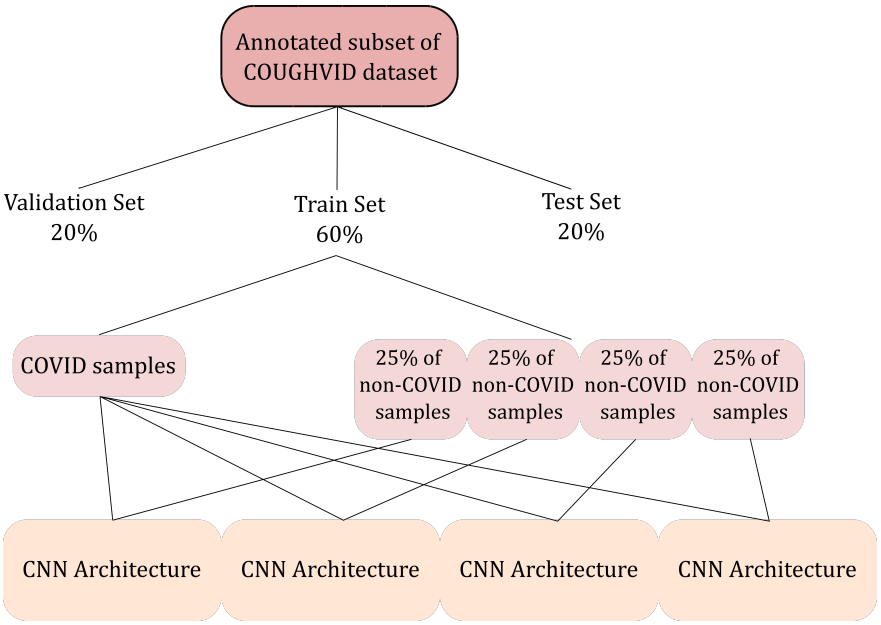


Figure 3.19: Dataset split for the ensemble method using the annotated COUGHVID dataset

Ensemble method for the Cambridge dataset

The ensemble method when using the Cambridge dataset follows the exact same principles described regarding the annotated COUGHVID dataset, but uses two CNN architectures instead of four. That is because the Covid class contains 31.0% of the total samples and the non-Covid the other 69.0%, so splitting the negative samples into two equally sized subsets and combining them with the positive samples would provide a balanced train set. More specifically, two probability outputs are provided for each one of the test samples. The average of these two probabilities acquired is used as the classification probability for the corresponding test sample. This probability is converted to the final prediction referring to this sample using a probability threshold 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. The splitting of the Cambridge dataset is schematically depicted in figure 3.21.

3.3.3 Multiple trainings of ResNet architecture with different cough datasets

Four different datasets are used in this method and are namely the following: Coswara cough heavy, Coswara cough shallow, annotated by experts subset of COUGHVID dataset and Cambridge dataset. As previously stated, the Coswara cough heavy dataset contains 1,541 samples, with 103 of them being Covid samples (6.68%) and the other 1,438 being non-Covid samples (93.32%). The Coswara cough shallow dataset contains 1,539

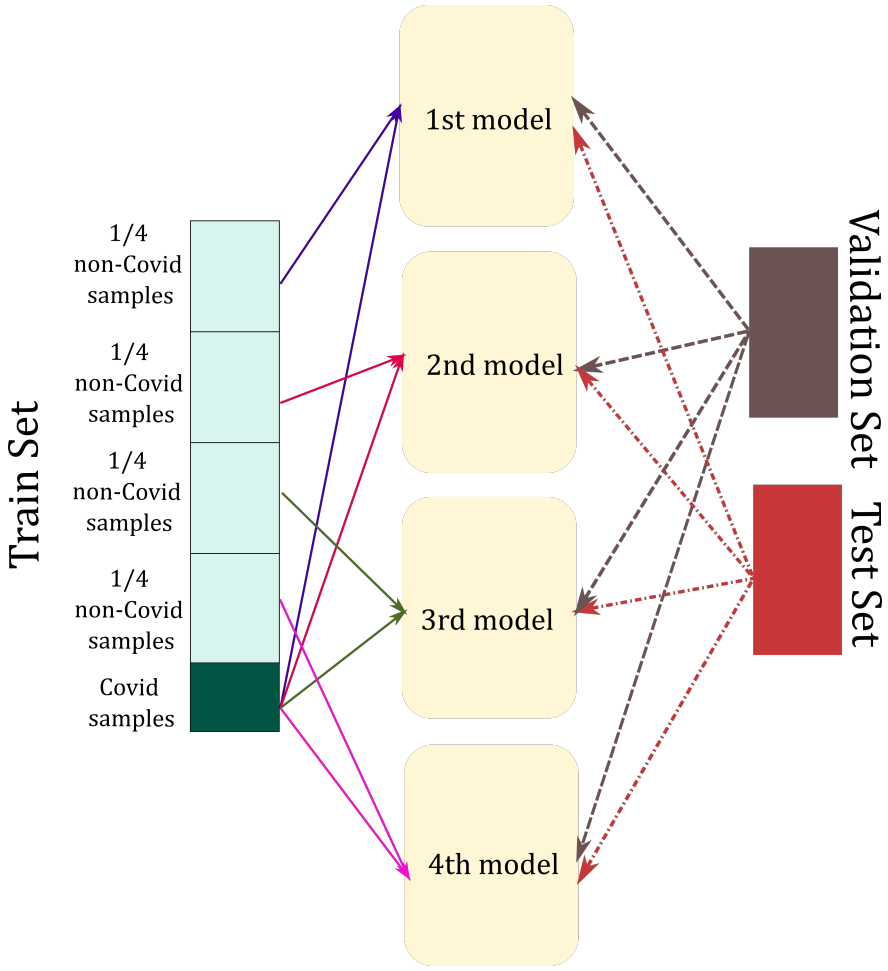


Figure 3.20: Assignment of data samples in each one of the ensemble models when using the annotated COUGHVID dataset

samples, 103 of which are Covid samples (6.69%) and the other 1,436 are non-Covid samples (93.31%). The annotated by the experts subset of COUGHVID dataset contains 2,804 cough samples in total, 553 of which are Covid samples (19.72%) and 2,251 are non-Covid samples (80.28%). The Cambridge dataset contains 400 samples in total, with 124 of them being Covid samples (31.0%) and the other 276 being non-Covid samples (69.0%).

Three pre-existing, Deep Learning architectures, ResNet-50, DenseNet-201 and Xception, have been examined and the reason for choosing to examine this method using deep, already existing CNN architectures, lies in the fact that they are provably able to achieve significant classification results. Throughout the majority of experimentations the ResNet-50 model is used and the following pipeline is implemented during training: The model is trained on the four aforementioned datasets. It is initially trained using the Coswara cough heavy dataset, initializing the weights with the ones acquired by pre-training the model on ImageNet. The dataset is split into train and validation set with the separation rate being 80%-20%. During training Synthetic Minority Oversampling Technique (SMOTE) is used,

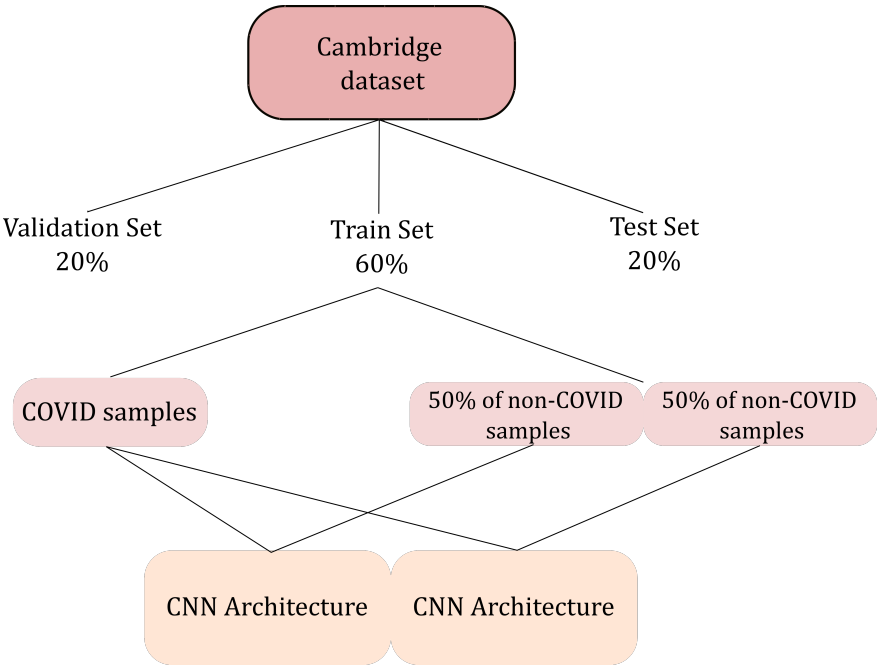


Figure 3.21: Dataset split for the ensemble method using the Cambridge dataset

in order for the train set to become balanced and contain the same number of Covid and non-Covid samples. The ResNet-50 model trained on the Coswara cough heavy dataset is then trained on the Coswara cough shallow dataset using the exact same method of training, that being using 80% of the dataset as the train set and the other 20% as the validation set and implementing SMOTE during training to balance the train set. The new model, resulting from training the pre-trained on ImageNet ResNet-50 model with the Coswara cough heavy and shallow dataset, is then trained using the subset of the COUGHVID dataset containing only samples that have been annotated by experts. The same, previously described, training method is also followed with this dataset. The model acquired by training ResNet-50 with the three aforementioned datasets is then used with the Cambridge dataset. The 5-fold cross validation method is combined with the ensemble method. The dataset is divided into 5 folds, with 3 folds being used as the train set, 1 fold as the validation set and 1 fold as the test set. Each fold is used as the test and the validation set exactly once. As for the ensemble method, two of the aforementioned ResNet-50 models are used, in order for the train set to become balanced and reduce the negative effects of an imbalanced dataset in the classification results, as described in section 3.3.2. The steps used in order for the final classification results to be obtained, are schematically shown in 3.22. This combination of datasets and architectures is mostly examined, because it provided the higher classification results.

In the first three trainings of ResNet-50, the convolutional base of the model is trained for 20 epochs. A classification head consisting of a dropout layer with a dropout rate equal to 0.5, a global average pooling layer and a dense layer with sigmoid activation function, is added to the convolutional base. Different trials have been made as for the

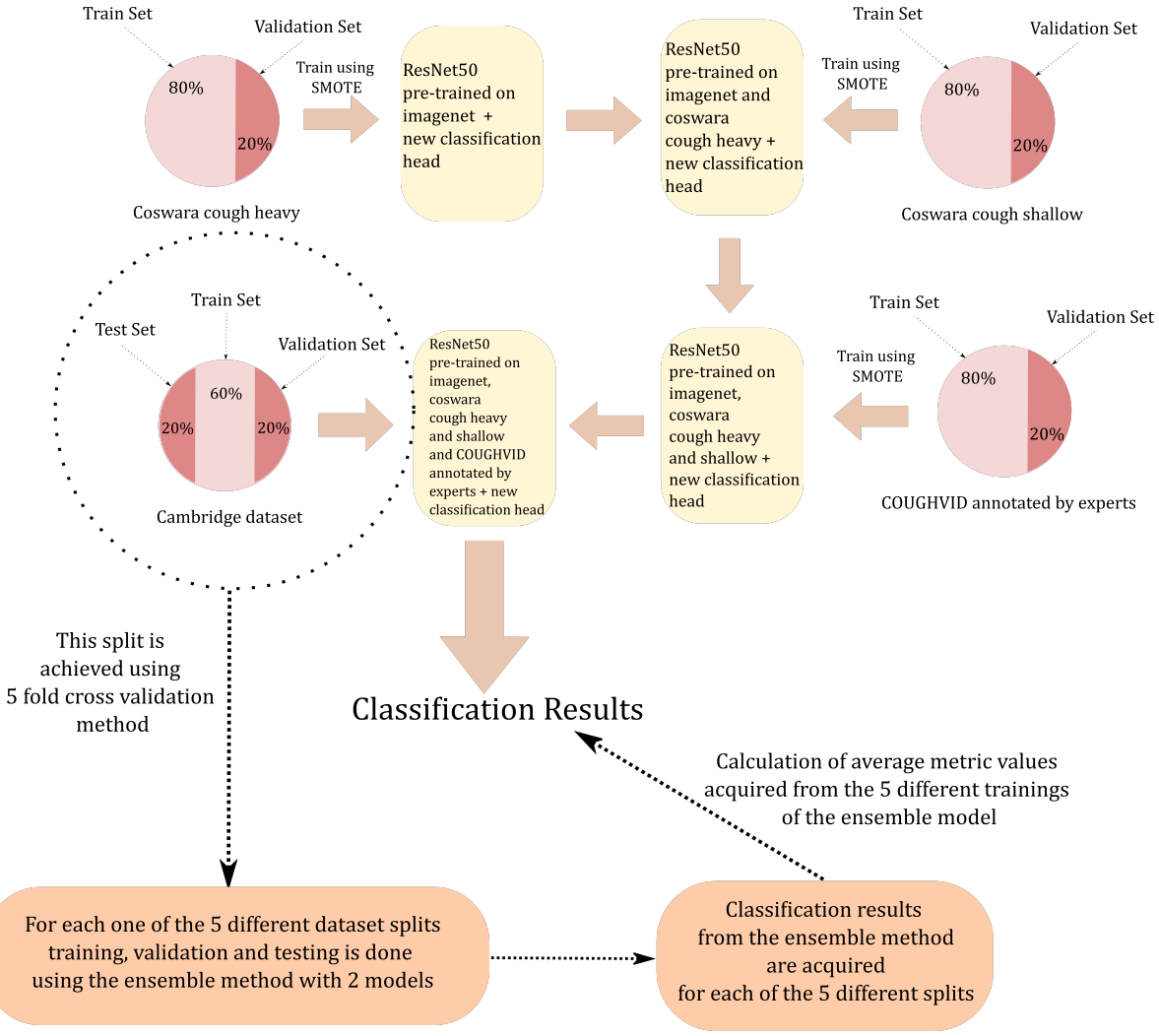


Figure 3.22: Description of the steps followed in the method described in section 3.3.3

number of epochs used for the last training of the model using the Cambridge dataset. Moreover, three different audio to image transformations, HCQT, Mel Spectrograms and STFT, were used. In all cases, all datasets used for training the ResNet-50 have been converted using the same transformation, either the HCQT, the Mel Spectrograms or the STFT. The optimiser used is Adam optimiser and the learning rate is set to 0.0001. The loss function used is Binary Cross Entropy loss and the $label\_smoothing$ parameter is set to 0.4. Label smoothing is a regularisation technique which introduces noise for the labels. When this parameter is larger than 0, the loss is computed between the predicted labels and a smoothed version of the true labels. As a result, it prevents the model from making very confident predictions during training, which could lead to bad generalisation.

The initial choice of this specific method and the idea of training the same model with many different task-related datasets stems from the fact that the ResNet-50 architecture is a deep, data "hungry" architecture and although it is pre-trained on ImageNet, a dataset containing more than 14 million images, these do not display audio samples and hence the

model does not have any previous knowledge on an audio related image classification task. However, pre-training it on three relevant to the task datasets offers a better initialisation of the model's weights, enabling it to effectively learn features of the fourth dataset and provide better testing results. Similar approaches to Covid classification tasks, analysing respiratory audio samples using CNN architectures, have shown very promising results and confirm the observation made that a ResNet architecture pre-trained on audio samples can provide very promising results in such a classification task. Researchers from MIT combined a Poisson biomarker layer with three pre-trained ResNet-50 models in parallel. The first ResNet-50 model was trained to distinguish the word "Them" from other words using LibriSpeech, an audiobook dataset containing approximately 1,000 hours of speech [47]. The second model was trained to learn sentiment features on the RAVDESS speech dataset, a dataset including actors that intonate in 8 emotional states which are namely: neutral, calm, happy, sad, angry, fearful, disgust and surprised [48]. The last ResNet-50 was trained to distinguish the spoken language of the person coughing (English or Spanish) on the cough dataset used, after taking into consideration only the metadata referring to the spoken language of the person. The classification results provided by the pre-trained models are higher than the ones acquired using not pre-trained ResNet-50 models [46]. Bagad et al. [51] used ResNet-18 as the base of their CNN architecture, pre-training it on three open source cough datasets. The first one is the FreeSound Database 2018, containing 11,073 audio files belonging to 41 possible categories with 273 of them being cough samples [70]. The second one is the Flusense dataset where 11,687 samples of various categories were used, with 2,486 of them being cough samples [71]. Finally, Coswara dataset was also used containing 2,034 cough sounds and 7,115 non-cough sounds [72]. The data were split in train and validation sets and the model was trained to predict the presence or the absence of cough in an audio sample. This model was then used with a Covid dataset created, containing 3,117 cough samples from 1,039 individuals, showing that pre-training improves the mean value of AUC by 17%.

Chapter 4

Results

The results presented have been acquired using a batch size of 32, with the shape of the input images being equal to (100, 100, 3).

4.1 5-fold cross validation method using one single model

The initial results obtained concern the 5-fold cross validation method. Five different datasets were examined using this method, as they are enumerated below.

1. Coswara cough heavy dataset
2. Coswara cough shallow dataset
3. COUGHVID dataset
4. The annotated by an expert subset of COUGHVID dataset
5. The Cambridge dataset

The classification results acquired by training the 5 models described in sections 3.2.1-3.2.5 with each of the 5 different datasets, after transforming audio to image using both the HCQT and the Mel Spectrograms, are presented in tables 4.1-4.5. The results using these two audio to image conversions are chosen, since it was observed that they generally provide higher values for the classification metrics than other transformations examined. The values of the metrics presented have been acquired by training each model for 30 epochs. The model providing the highest values for the classification metrics, per transformation and dataset has been highlighted.

Transformation	Model	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
HCQT	**Model 1**	**67.10**	**57.09**	**11.44**	**65.80**	**67.80**
	Model 2	73.46	48.42	12.50	67.17	75.25
	Model 3	90.01	7.82	12.13	54.79	95.90
	DenseNet	92.86	3.91	25.71	56.83	99.24
	ResNet	29.55	78.10	7.28	53.01	26.12
Mel	**Model 1**	**72.48**	**50.83**	**12.01**	**64.34**	**74.05**
	Model 2	75.59	46.82	13.04	63.54	77.67
	Model 3	86.70	19.65	16.29	58.57	91.52
	DenseNet	92.28	2.86	11.67	55.10	98.68
	ResNet	53.10	57.49	7.90	59.23	52.79

Table 4.1: Performance metrics for the Coswara Cough Heavy dataset

It is observed that Model 1 achieves the highest values for the classification metrics, although the precision is extremely low. This indicates that most of the samples predicted as positive, did not belong to the positive class, i.e. the number of TPs is low and the number of FPs is high. Moreover, the sensitivity value is also low, ranging from 2.86% to 57.49% depending on the model and the transformation used, indicating that the models cannot predict the positive (covid) class correctly.

Transformation	Model	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
HCQT	**Model 1**	**74.15**	**41.86**	**11.69**	**64.27**	**76.48**
	Model 2	68.88	45.71	10.12	64.05	70.54
	Model 3	89.99	7.71	12.11	54.32	95.89
	DenseNet	92.79	1.90	4.00	58.04	99.30
	ResNet	34.20	73.81	7.19	54.53	31.35
Mel	**Model 1**	**71.16**	**46.79**	**11.16**	**64.26**	**72.93**
	Model 2	71.61	42.70	10.35	64.17	73.68
	Model 3	89.28	7.90	9.30	53.94	95.13
	DenseNet	91.82	8.81	40.21	60.01	97.78
	ResNet	49.30	57.19	9.06	61.46	48.74

Table 4.2: Performance metrics for the Coswara Cough Shallow dataset

The results acquired using the Coswara cough-shallow dataset are very similar to the ones obtained when using the Coswara cough-heavy dataset, an anticipated observation taking into consideration the fact that the two datasets contain cough samples from the same users with the only difference being the type of cough recorded. Model 1 once again provides the highest classification results for this dataset. Comparing the best results obtained using the HCQT in both datasets, the accuracy and specificity show an increase of 7-9%, while the sensitivity presents a decrease of 16% when using the cough shallow samples with the precision remaining invariable. Therefore, in the case of using the HCQT, Model 1 seems to be better at predicting the positive class when trained on cough-heavy samples, while it achieves a higher specificity and overall accuracy when trained with cough-shallow samples, indicating the ability of the model to predict the negative class more efficiently.

As for the case of using the Mel spectrograms, the differences in the results acquired from Model 1 and for the two datasets are minor.

Transformation	Model	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
HCQT	**Model 1**	**55.16**	**42.33**	**6.33**	**49.08**	**56.05**
	Model 2	**50.52**	**47.90**	**6.15**	**50.68**	**50.70**
	Model 3	80.40	16.57	6.05	52.00	84.80
	DenseNet	91.79	4.30	13.18	54.65	97.84
	ResNet	6.84	99.57	6.46	50.41	0.43
Mel	**Model 1**	**52.52**	**57.94**	**7.74**	**55.97**	**52.14**
	Model 2	49.04	55.04	6.85	53.58	48.62
	Model 3	85.46	11.00	9.75	56.39	90.60
	DenseNet	91.95	3.43	10.60	55.71	98.06
	ResNet	56.50	20.00	3.75	51.30	80.00

Table 4.3: Performance metrics for the COUGHVID dataset

Compared to the previously presented results regarding the two subsets of the Coswara dataset, the values of the classification metrics acquired when using the COUGHVID dataset are even lower. This confirms the fact that the quantity of the data and the distribution of samples in the two classes are not the only factors affecting the classification ability of a model, since the COUGHVID dataset contains a lot more samples with their distribution between Covid and non-Covid being almost the same as the one in the other two datasets, i.e. approximately 6% of the data belongs to the Covid class with the other 94% belonging to the non-Covid class. Although the value of most of the measured metrics is low, the best performances are obtained by Models 1 and 2 when using the HCQT and by Model 1 when using the Mel Spectrograms. Comparing Models 1 and 2 in the occasion of HCQT, their difference lies in the accuracy, sensitivity and specificity values, with Model 1 achieving a slightly higher accuracy and specificity and thus predicting the negative class better than Model 2 which reaches moderately higher sensitivity values and therefore predicts the positive class marginally better. As for the ResNet and DenseNet architectures used, either the sensitivity or the precision, or both, in most cases are significantly lower than the ones provided by the rest of the architectures, indicating that the model has not been trained well and the predictions made are random, classifying almost all samples either as Covid or as non-Covid. It is observed that Model 1, trained with the Mel spectrograms of the audio samples, outperforms the rest of the models independently of the transformation used with regard to the COUGHVID dataset.

In comparison to the datasets previously examined, the annotated COUGHVID dataset achieves a noticeable increase in the precision's values, which are twice the ones reached when using the Coswara cough-heavy and shallow datasets and approximately three times bigger than the values achieved when training with all of the samples of the COUGHVID dataset. The values of the rest of the metrics are kept at relatively the same levels in most cases. The best performance is achieved by Model 1, independently of the transformation used. A noticeably higher sensitivity value is reached when implementing the HCQT transform, with the model being able to predict the positive class more efficiently. However,

Transformation	Model	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
HCQT	**Model 1**	**50.16**	**62.01**	**22.51**	**55.37**	**47.24**
	Model 2	60.04	41.02	22.30	53.91	64.69
	Model 3	74.25	12.65	22.35	52.78	89.38
	DenseNet	70.68	19.50	22.26	53.86	83.23
	ResNet	20.68	99.64	19.87	50.21	1.29
Mel	**Model 1**	**54.48**	**47.79**	**21.31**	**51.63**	**56.13**
	Model 2	52.21	46.84	19.35	49.88	53.53
	Model 3	71.30	15.53	20.95	52.03	85.00
	DenseNet	76.07	10.68	22.05	53.38	92.14
	ResNet	22.00	96.92	19.81	49.57	3.60

Table 4.4: Performance metrics for the annotated COUGHVID dataset

the model achieves slightly higher accuracy and specificity values when employing Mel Spectrograms, indicating an ability to predict the negative class slightly better.

Transformation	Model	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
HCQT	**Model 1**	**62.75**	**53.82**	**40.23**	**63.85**	**63.93**
	Model 2	61.75	55.76	39.98	61.97	61.97
	Model 3	66.00	35.67	43.70	64.47	77.96
	DenseNet	47.50	87.74	36.18	65.22	29.93
	ResNet	69.00	0.0	0.0	43.82	100.00
Mel	Model 1	62.00	55.51	40.66	63.57	63.47
	Model 2	**63.75**	**57.82**	**42.41**	**64.28**	**64.66**
	Model 3	67.75	41.93	42.77	65.81	76.91
	DenseNet	57.22	83.57	40.86	67.17	45.33
	ResNet	52.80	40.00	11.87	49.43	60.00

Table 4.5: Performance metrics for the Cambridge dataset

As for the Cambridge dataset, when using the HCQT transform the higher classification results are acquired by Model 1, while Model 2 outperforms the rest of the models when trained using Mel Spectrograms. Although the values of the metrics do not generally indicate a large increase, the precision value is almost twice the respective value when training the models using the annotated COUGHVID dataset and is the highest acquired so far.

It is observed that Model 1 and 2 highly outperform the rest of the models independently of the dataset used. The very low values in the classification metrics can be justified by the high rate of imbalance between the two classes. This can be explained by the slightly better results provided by the annotated COUGHVID and the Cambridge datasets, which are the less unbalanced of the five, with the last one achieving precision values two times larger than the ones achieved using the annotated COUGHVID dataset. Another factor that complicates the examined task is the existence of bad audio quality samples, as well as the fact that the data is labelled based on the user's declarations, so the ground truth

of the problem is not solid. An interesting observation are the classification results of the ResNet-50 and DenseNet-201 models which are pre-trained on the ImageNet dataset. In most cases, the sensitivity and precision of these models are significantly lower than the ones provided by the rest of the architectures, whereas in some of the experiments, the sensitivity value is extremely high with the rest of the metrics taking extremely low values. In both situations, this behaviour indicates that the model has not been trained well and the predictions made are random, classifying almost all samples either as Covid or as non-Covid. Since these two pre-trained models have achieved very promising results in other image classification tasks, the most possible reason for them not being trained well, is the fact that the datasets they are trained on are not large enough for such deep CNN architectures to become able to generalise well on unseen data.

4.2 5-fold cross validation method using ensemble models

Since the imbalance of the data plays a very important role on the ability of the model to learn how to distinguish between the two classes, the ensemble method is implemented in order for the negative results of this imbalance to be abated. The ensemble method was implemented on the two datasets providing better results, which as previously mentioned are the annotated COUGHVID dataset and the Cambridge dataset. The results obtained from this method are presented in tables 4.6 and 4.7. More image to audio transformations were examined in this subsection, in search for the most suitable one for the task addressed. More specifically, apart from the Mel Spectrograms and the HCQT, the CQT and STFT transforms are also examined. The model achieving the highest classification results for each dataset and transformation is highlighted.

It is observed that the values acquired by all models and transforms used are more consistent in comparison to the ones obtained using one single model, with the general performance of the best models being slightly better in the ensemble method. When implementing the HCQT transform, the DenseNet model outperforms the rest. Model 2 achieves the highest results when trained with Mel Spectrograms and STFT spectrograms, while Model 1 in the case of CQT. The overall best results for the ensemble method using the annotated COUGHVID dataset are acquired using the CQT transform and Model 1. Although, a small increase in the classification metrics is generally observed using the ensemble method with the annotated COUGHVID dataset and Models 1 and 2, remarkable is the change observed in the classification metrics of the rest of the models with the values obtained showing a more consistent form, compared to these acquired by the previous method.

As for the implementation of the ensemble method using the Cambridge dataset, Model 1 provided the best results both when using the Mel Spectrograms and the STFT, Model 2 outperforms the rest of the models when using the HCQT, while Models 1 and 2 provide the

Transformation	Model	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
HCQT	Model 1	56.52	47.73	22.15	55.19	58.68
	Model 2	53.70	55.18	22.49	55.25	53.34
	Model 3	52.25	45.92	19.67	50.96	53.80
	DenseNet	**55.39**	**53.51**	**22.74**	**57.10**	**55.84**
	ResNet	52.36	54.60	21.89	56.40	51.81
Mel	Model 1	57.10	47.19	22.18	55.20	59.53
	Model 2	**53.18**	**52.98**	**21.82**	**54.86**	**53.23**
	Model 3	51.78	52.64	21.09	53.70	51.58
	DenseNet	57.73	48.11	22.99	56.62	60.10
	ResNet	61.74	32.59	20.30	53.30	68.88
CQT	**Model 1**	**57.99**	**50.83**	**23.81**	**57.29**	**59.75**
	Model 2	56.88	52.99	23.63	56.65	57.84
	Model 3	53.79	49.53	21.21	52.74	54.83
	DenseNet	54.10	56.39	22.91	57.20	53.53
	ResNet	36.94	72.28	19.84	51.30	28.22
STFT	Model 1	49.70	51.42	18.62	53.16	49.32
	Model 2	**51.84**	**55.17**	**20.05**	**54.09**	**51.09**
	Model 3	51.50	43.98	17.40	48.51	53.18
	DenseNet	52.88	50.41	19.44	54.14	53.43
	ResNet	52.29	50.52	19.29	51.55	52.70

Table 4.6: Performance metrics for the annotated COUGHVID dataset in the ensemble method

best values for the classification metrics when employing the CQT transform. The results acquired by the ensemble method are almost the same with these obtained when using one single model architecture and the 5-fold cross validation training method. This behaviour can be explained by taking into consideration the fact that the Cambridge dataset is not as imbalanced as the rest. Nevertheless, the ResNet model is still not able to generalise well in the test set. Although a generally remarkable increase in the value of the classification metrics has been achieved by all models when implementing the ensemble method, the classification results are still neither noteworthy nor reliable.

4.3 Multiple training of ResNet-50 architecture using 4 different datasets

The method described in section 3.3.3 leverages the availability of multiple different datasets, taking into consideration the large amount of data needed for a Deep Convolutional Neural Network to be trained. The pre-trained on ImageNet ResNet-50 architecture is trained using the Coswara Cough Heavy, Coswara Cough Shallow, annotated COUGHVID and Cambridge datasets as explained in section 3.3.3. In the first three trainings, the model is trained for 20 epochs. The number of training epochs used for the final training of the

Transformation	Model	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
HCQT	Model 1	60.30	54.68	40.07	61.49	60.57
	Model 2	**62.55**	**55.66**	**41.90**	**63.68**	**63.05**
	Model 3	62.30	48.71	42.02	63.29	67.01
	DenseNet	49.17	79.73	34.44	61.08	35.89
	ResNet	56.52	25.71	8.77	47.31	70.16
Mel	**Model 1**	**60.80**	**54.82**	**39.96**	**61.06**	**60.83**
	Model 2	60.04	53.15	38.72	58.16	60.52
	Model 3	57.30	44.93	34.42	56.06	61.19
	DenseNet	55.79	68.34	40.95	66.07	51.20
	ResNet	54.31	60.00	22.31	51.63	40.00
CQT	**Model 1**	**60.05**	**51.77**	**38.92**	**59.16**	**61.30**
	Model 2	**59.28**	**53.99**	**38.73**	**62.75**	**59.27**
	Model 3	61.30	43.27	40.88	59.86	66.87
	DenseNet	56.19	67.92	37.59	61.22	50.33
	ResNet	45.47	39.17	7.90	45.66	60.00
STFT	**Model 1**	**59.50**	**64.35**	**43.79**	**63.60**	**57.64**
	Model 2	56.56	52.55	36.70	60.63	58.29
	Model 3	55.78	52.41	34.71	57.44	55.57
	DenseNet	67.42	58.51	47.92	69.78	70.94
	ResNet	58.27	17.14	2.31	50.28	85.40

Table 4.7: Performance metrics for the Cambridge dataset in the ensemble method

model, using the Cambridge dataset, varies. The outcome of these trials is presented in tables 4.10 and 4.11. However, the final choice of datasets to be used for pre-training the ResNet-50 model was decided after trials conducted using different combinations of datasets as they are presented in table 4.8. More specifically, table 4.8 contains the results of initially training the ResNet-50 architecture for 20 epochs using either the Coswara cough heavy dataset or the Coswara cough heavy and shallow datasets and then training and testing the pre-trained model using the Cambridge dataset, utilizing the HCQT transform for all of the different trainings of the model. Table 4.9 contains the results acquired by training the pre-trained on ImageNet ResNet-50 architecture on the Coswara cough heavy, Coswara cough shallow and Cambridge datasets for 20 epochs and finally training and testing it, using four ensemble models as described in section 3.3.2, on the annotated COUGHVID dataset, using the Mel-Spectrograms of the audio samples contained in these datasets.

As it can be observed, the classification results in the case of training the model with less datasets are noticeably lower than these obtained when training it using all available datasets. This behaviour confirms the knowledge that such Deep CNN architectures require large amounts of data in order to get efficiently trained. Moreover, changing the order with which the datasets are used to train the model, i.e. using the annotated COUGHVID dataset for the final training and the testing of the model, remarkably decreases the model's performance. To that end, most of the experimentation was conducted using all four of the datasets and in the order previously described, as it is presented in tables 4.10, 4.11

Datasets used	Epochs	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
Coswara cough heavy	25	33.38	95.00	30.61	50.84	6.44
and	50	35.65	93.27	31.11	61.42	10.95
Cambridge	100	64.56	57.17	44.97	64.78	67.36
Coswara cough heavy-shallow	25	68.63	46.36	47.18	67.34	77.32
and	50	65.09	52.52	43.70	67.33	69.48
Cambridge	100	67.52	53.35	46.00	65.56	73.64

Table 4.8: Performance metrics using the HCQT and different combinations of datasets

Epochs	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
25	54.96	52.25	22.43	54.33	55.63
50	53.32	54.04	22.05	55.09	53.14

Table 4.9: Performance metrics using Mel Spectrograms and testing the model on the annotated COUGHVID dataset

Epochs	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
10	71.25	45.35	49.73	72.58	81.53
15	71.53	52.62	51.46	72.09	78.03
20	69.55	54.76	51.80	73.65	74.69
25	**71.03**	**66.58**	**52.18**	**73.44**	**71.51**
28	68.75	59.63	48.34	70.45	71.40
30	65.75	61.23	44.96	70.14	66.51
35	70.84	63.71	52.38	72.95	72.50
50	69.06	63.03	50.69	73.42	70.86
150	67.58	63.98	49.59	73.50	69.24
200	68.83	60.00	51.27	72.72	72.98

Table 4.10: Performance metrics using HCQT transformation in all datasets

and in the rest of the section.

The best classification results provided for each one of the transformations are boldly marked. A significant increase in all metric values is accomplished, with an accuracy of 71.03%, a sensitivity of 66.58%, a precision of 52.18%, an AUC value of 73.44% and a specificity of 71.51% in the case of transforming the audio samples using the HCQT. When using the Mel Spectrograms as inputs to the model, the highest results achieved provide

Epochs	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
15	73.03	48.01	62.91	69.13	83.92
20	72.29	54.26	59.13	70.12	79.94
25	70.35	60.49	55.14	68.98	74.36
30	71.37	61.14	57.98	68.60	75.36
35	70.05	59.43	54.19	69.96	74.26
40	69.28	58.45	52.38	69.53	73.52
50	**71.60**	**62.92**	**57.21**	**69.92**	**74.78**
60	70.02	59.60	52.63	69.84	73.94
70	71.30	60.68	56.37	69.44	75.50
80	70.05	59.15	53.17	68.99	74.28
100	69.35	60.23	52.97	68.84	72.89
150	71.28	56.28	55.84	69.67	76.32

Table 4.11: Performance metrics using Mel Spectrograms in all datasets

an accuracy of 71.60%, a sensitivity of 62.92%, a precision of 57.21%, an AUC of 69.92% and a specificity of 74.78%. Although a small decrease in the AUC and sensitivity values is observed in the case of using Mel Spectrograms, a small increase in the precision and specificity values is noticed, with the precision reaching the highest value achieved compared to all the other trials made. The number of epochs for which the model is trained plays an important role to its performance on unseen data. In the case of using HCQT, when training the model on the final stage a smaller number of epochs provides higher classification results. As for the case of using the Mel spectrograms, the best performance is again acquired when training the model for a relatively small number of epochs.

Since the STFT reached the highest performance when utilized with the Cambridge dataset and the ensemble method, it was also examined in the current method. The ResNet-50 model was trained for 20 epochs using the Coswara cough heavy dataset, then for 20 epochs using the Coswara cough shallow dataset, then trained for 20 epochs utilizing the annotated COUGHVID dataset and lastly it was trained and tested using the Cambridge dataset. All samples in these datasets were represented as images using the STFT. However, the values of the metrics acquired from the trials made were not as high as the ones obtained when using the HCQT or the Mel spectrograms, rendering the STFT image-frequency representation unsuitable for this combination of datasets and architecture. The results obtained from the aforementioned experiments can be seen in table 4.12.

Another interesting observation in respect to the CNN architecture used, is that two other pre-trained Deep CNN architectures, DenseNet-201 and Xception, which were also examined, provided appreciably lower classification results. These results are presented in table

Epochs	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
25	67.50	43.82	49.83	64.92	77.75
50	58.75	56.48	39.57	65.71	59.21
100	58.75	54.11	38.33	64.90	60.36

Table 4.12: Performance metrics using the STFT in all datasets

4.13 and confirm the wide usage of ResNet architectures in cough related classification tasks.

Model	Epochs	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
DenseNet-201	15	64.66	48.80	45.11	63.33	71.88
	25	66.99	55.74	48.05	66.96	71.92
	50	59.88	37.88	34.09	59.22	69.46
Xception	15	66.92	48.38	46.31	66.68	75.10
	25	66.39	59.48	47.00	69.44	69.19
	50	62.19	56.40	41.36	68.19	65.39

Table 4.13: Performance metrics using HCQT transform with the DenseNet and the Xception model

The label smoothing parameter is set to 0.4, regarding all of the aforementioned experiments. However, more values for the label smoothing parameter were examined, in order for the most appropriate one to be found. The HCQT was used for all of the datasets, in combination with the ResNet-50 architecture. Different rates of label smoothing were examined and the metrics acquired from these trials are presented in table 4.14. It is observed that other rates of label smoothing provide lower values for the classification metrics compared, to the ones presented in table 4.10, rendering themselves inappropriate for the current task.

The number of epochs for which the model is trained, throughout all of the experiments made when implementing the current method, was chosen by observing the model's behaviour and the performance achieved after each trial.

Label Smoothing Rate	Epochs	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
0.0	25	66.10	49.91	45.52	65.80	72.20
0.2	25	67.38	55.31	49.97	67.87	72.50
	50	66.83	53.50	47.48	67.76	72.48
0.5	25	68.02	52.54	48.52	69.69	74.64
0.6	25	69.32	49.52	52.01	68.11	77.91
	50	68.11	50.48	49.88	69.82	75.73
0.9	25	63.86	54.70	43.84	67.31	67.79
	40	65.92	56.58	46.87	66.19	69.95

Table 4.14: Performance metrics using the HCQT, the ResNet-50 architecture and multiple values for label smoothing

4.4 Summary of the acquired classification results

A summary of the best classification results acquired by each of the three methods and for each dataset is presented in table 4.15. It is observed that Model 1 generally performs better independently of the transform used. Moreover, the final method implemented, reaches adequate performance levels when ResNet-50 model is trained with Coswara cough heavy, Coswara cough shallow, annotated COUGHVID and Cambridge datasets following the order with which they are mentioned.

Method	Dataset	Transform	Model	Accuracy (%)	Sensitivity (%)	Precision (%)	AUC (%)	Specificity (%)
5-fold cross validation using one CNN model	Coswara cough heavy	HCQT	Model 1	67.10	57.09	11.44	65.80	67.80
	Coswara cough shallow	HCQT	Model 1	74.15	41.86	11.69	64.27	76.48
	COUGHVID	Mel	Model 1	52.52	57.94	7.74	55.97	52.14
	Annotated COUGHVID	HCQT	Model 1	50.16	62.01	22.51	55.37	47.24
	Cambridge	Mel	Model 2	63.75	57.82	42.41	64.28	64.66
5-fold cross validation using ensemble models	Annotated COUGHVID	CQT	Model 1	57.99	50.83	23.81	57.29	59.75
	Cambridge	STFT	Model 1	59.50	64.35	43.79	63.60	57.64
Multiple trainings of ResNet-50 model	Four datasets*	HCQT	ResNet-50	**71.03**	**66.58**	**52.18**	**73.44**	**71.51**
	Four datasets*	Mel	ResNet-50	**71.60**	**62.92**	**57.21**	**69.92**	**74.78**

Table 4.15: Summarised results (*Four datasets refer to Coswara cough heavy, Coswara cough shallow, annotated COUGHVID and Cambridge datasets)

The fluctuation of the values of the five examined metrics when using the three different methods, for the best result acquired from each, is depicted in figure 4.1. The results presented for the first method are obtained by training Model 2 using the Cambridge dataset and Mel spectrograms, while the results presented for the second method are acquired by training Model 1, using the Cambridge dataset and STFT. The results presented for the final method are the ones obtained using the HCQT.

To the best of our knowledge, this is the first time that the HCQT is utilized in a cough classification task and especially in a COVID-19 detection task.

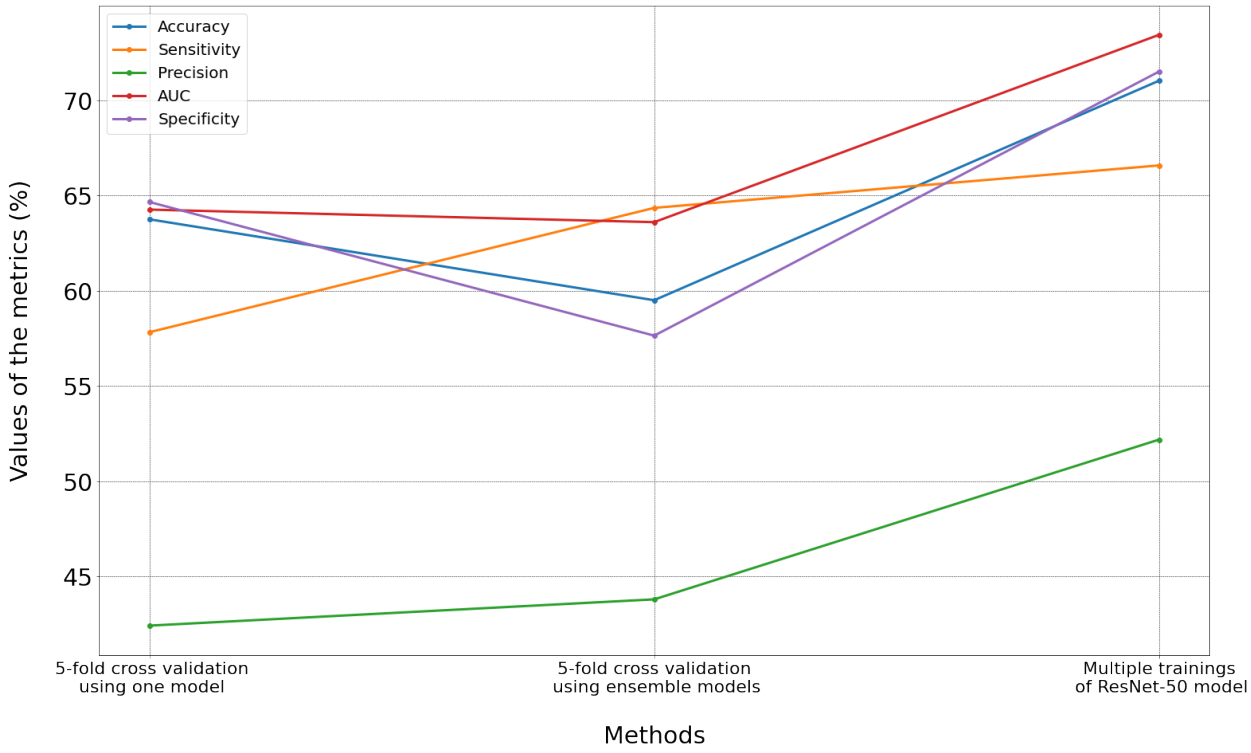


Figure 4.1: Comparison of the best results acquired by each method

Chapter 5

Conclusion and future research

5.1 Conclusion

The goal of the current thesis is the development of a Deep Learning method for the screening of COVID-19. Different datasets containing cough samples from healthy and COVID-19 infected individuals, multiple audio to image conversions, various CNN architectures and different methods of implementing them have been examined, trying multiple possible values for each of the model's parameters and the training parameters.

Due to the imbalanced nature of the problem and the available datasets, a single CNN model cannot provide reliable results for the task in question. The best classification results using one single CNN architecture and the 5-fold cross validation method were provided by training Model 2 using the Cambridge dataset, with the audio samples being converted into Mel Spectrograms. This combination reached an accuracy value of 63.75%, sensitivity of 57.82%, precision of 42.41%, an AUC value of 64.28% and a specificity of 64.66%. Due to the high imbalance of most of the available datasets, an ensemble method was also implemented as a means of reducing the negative impact of this data feature. This method was tested with the two datasets providing slightly better results when used for training one single CNN model. These are the annotated COUGHVID dataset and the Cambridge dataset, with the best performance achieved reaching an accuracy of 57.99%, a sensitivity of 50.83%, a precision of 23.81%, an AUC of 57.29% and a specificity of 59.75% when training Model 1 using the CQT transform for the annotated COUGHVID dataset. As for the Cambridge dataset, the best performance was attained when training Model 2 using the HCQT transform, where an accuracy of 62.55%, a sensitivity of 55.66%, a precision of 41.90%, an AUC of 63.68% and a specificity of 63.05% were achieved during testing. Nonetheless, these results cannot be reliable and thus these models cannot be used to effectively detect COVID-19. To that end, multiple combinations of architectures and datasets were examined.

The combination of ensemble learning and multiple trainings of a pre-trained Deep CNN architecture, ResNet-50, introduced impressive improvement in the values of the examined classification metrics reaching an accuracy of 71.60% when using the Mel-spectrograms and an accuracy of 71.03% when using the HCQT. More specifically, in the case of using the HCQT transformation, the best results were acquired by training the model for 25 epochs, when using the Cambridge dataset, providing an accuracy score of 71.03%, a sensitivity of 66.58%, a precision of 52.18%, an AUC of 73.44% and a specificity of 71.51%. When using the Mel Spectrograms, the best results were obtained by training the model for 50 epochs, when using the Cambridge dataset. This provided an accuracy score of 71.60%, a sensitivity of 62.92% a precision of 57.21%, an AUC of 69.92% and a specificity of 74.78%. To the best of our knowledge, it is the first time that the HCQT transform is utilized for a cough classification task. The results acquired using this method, which combines four different datasets, are significantly better than the ones obtained when using the same model but solely one of these datasets. Moreover, two different pre-trained models were also examined using the exact same method and replacing the ResNet-50 model with either the DenseNet-201 or the Xception model. However, the obtained results are worse, with the DenseNet-201 model achieving an accuracy score of 66.99%, a sensitivity of 55.74%, a precision of 48.05%, an AUC of 66.96% and a specificity of 71.92% and the Xception model achieving an accuracy of 66.39%, a sensitivity of 59.48%, a precision of 47.00%, an AUC of 69.44% and a specificity of 69.19%.

These results prove the ability of Machine Learning to decisively assist Medicine in multiple domains and especially in the diagnosis of diseases. Except for the speed of testing possible COVID-19 cases when using such methods, the biggest asset is the ease of access to free testing by the vast majority of the public. Although such methods were not available at the beginning of the pandemic, partly due to the absence of data, their novelty could now be availed aiming for a sooner exit from the pandemic. Moreover, they could be widely used in the diagnosis and classification of other respiratory illnesses, while the research community is now better prepared for future pandemic outbreaks.

5.2 Future Research

Future research could include the analysis and examination of other respiratory sounds, such as breathing and speaking, since they could also provide valuable information about the health status of the user. Various features extracted from different types of respiratory sounds and combinations of them, when used to train a CNN architecture, could possibly increase its performance. This is due to the fact that a Deep Learning architecture can extract multiple features from each respiratory sound analysed and use them to arrive at better classification results, leveraging the ability of CNN architectures to learn discriminative spectro-temporal patterns.

Moreover, for users infected with COVID-19, re-sampling every one or two days, for a

specific time interval or for as long as they are infected, could provide valuable information about the deterioration or the improvement of the user's health status. Except for that, supplying a model with samples of the same user, but from different phases of the illness would contribute to the better understanding of COVID-19, independently of the phase of the user's illness or its severity. Thus, this could result to a more robust model, achieving highly reliable classification outcomes.

The obtained results could also be noticeably improved by involving more and higher quality data. Datasets with more samples and especially more Covid samples, labelled by experts with respect to the quality and with the label related to the health status of the user being assigned by using the results of PCR testing, could possibly lead to much better classification results since the ground truth on which the model will be trained would be more reliable.

Bibliography

[1] W. H. Organization. Who coronavirus (covid-19) dashboard. [Online]. Available: https://covid19.who.int/ pages 1, 23

[2] W. H. Organization. Listings of who's response to covid-19. [Online]. Available: https://www.who.int/news/item/29-06-2020-covidtimeline pages 1, 23, 28

[3] W. H. Organization. Coronavirus disease (covid-19): How is it transmitted? [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted pages 1, 24

[4] Y.-m. Zhao, Y.-m. Shang, W.-b. Song, Q.-q. Li, H. Xie, Q.-f. Xu, J.-l. Jia, L.-m. Li, H.-l. Mao, X.-m. Zhou *et al.*, "Follow-up study of the pulmonary function and related physiological characteristics of covid-19 survivors three months after recovery," *EClinicalMedicine*, vol. 25, p. 100463, 2020. pages 2, 24

[5] A. S. Zubair, L. S. McAlpine, T. Gardin, S. Farhadian, D. E. Kuruvilla, and S. Spudich, "Neuropathogenesis and neurologic manifestations of the coronaviruses in the age of coronavirus disease 2019: a review," *JAMA neurology*, vol. 77, no. 8, pp. 1018–1027, 2020. pages 2, 24

[6] C. del Rio, L. F. Collins, and P. Malani, "Long-term Health Consequences of COVID-19," *JAMA*, vol. 324, no. 17, pp. 1723–1724, 11 2020. [Online]. Available: https://doi.org/10.1001/jama.2020.19719 pages 2, 24

[7] O. W. in Data. Coronavirus (covid-19) vaccinations. [Online]. Available: https://ourworldindata.org/covid-vaccinations pages 3, 19, 26

[8] C. for Disease Control and P. (CDC). Nucleic acid amplification tests (naats). [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/lab/naats.html pages 3, 28

[9] C. for Disease Control and P. (CDC). Interim guidance for antigen testing for sars-cov-2. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antigen-tests-guidelines.html pages 3, 28

[10] V. Bansal, G. Pahwa, and N. Kannan, "Cough classification for COVID-19 based on audio mfcc features using convolutional neural networks," *2020 IEEE International Conference on Computing, Power and Communication Technologies, GUCON 2020*, pp. 604–608, 2020. pages 4, 30

[11] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Q. Ni, Y. Chen, J. Su *et al.*, "A deep learning system to screen novel coronavirus disease 2019 pneumonia," *Engineering*, vol. 6, no. 10, pp. 1122–1129, 2020. pages 4, 30

[12] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. D. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020. pages 4, 5, 30, 34

[13] B. W. Schuller, H. Coppock, and A. Gaskell, "Detecting covid-19 from breathing and coughing sounds using deep neural networks," *arXiv preprint arXiv:2012.14553*, 2020. pages 4, 5, 30, 35

[14] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020. pages 4, 30

[15] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *Pattern Analysis and Applications*, pp. 1–14, 2021. pages 4, 30

[16] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," *arXiv preprint arXiv:2003.11055*, 2020. pages 4, 30

[17] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 635–640, 2020. pages 4, 30

[18] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images," *Pattern Recognition Letters*, vol. 138, pp. 638–643, 2020. pages 4, 30

[19] N. Tsiknakis, E. Trivizakis, E. E. Vassalou, G. Z. Papadakis, D. A. Spandidos, A. Tsatsakis, J. Sánchez-García, R. López-González, N. Papanikolaou, A. H. Karantanas *et al.*, "Interpretable artificial intelligence framework for covid-19 screening on chest x-rays," *Experimental and Therapeutic Medicine*, vol. 20, no. 2, pp. 727–735, 2020. pages 4, 30

[20] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, and Y. Xia, "COVID-19 Screening on Chest X-ray Images Using Deep Learning based Anomaly Detection," pp. 1–1, 2020. pages 4, 30

[21] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct," *Radiology*, 2020. pages 4, 30

[22] J. L. Izquierdo, J. Ancochea, J. B. Soriano, S. C.-. R. Group *et al.*, "Clinical characteristics and prognostic factors for intensive care unit admission of patients with covid-19: retrospective study using machine learning and natural language processing," *Journal of medical Internet research*, vol. 22, no. 10, p. e21801, 2020. pages 4, 30

[23] M. Marcos, M. Belhassen-García, A. Sánchez-Puente, J. Sampedro-Gomez, R. Azibeiro, P.-I. Dorado-Díaz, E. Marcano-Millán, C. García-Vidal, M.-T. Moreiro-Barroso, N. Cubino-Bóveda *et al.*, "Development of a severity of disease score and classification model by machine learning for hospitalized covid-19 patients," *PloS one*, vol. 16, no. 4, p. e0240200, 2021. pages 4, 30

[24] A. Alotaibi, M. Shiblee, and A. Alshahrani, "Prediction of severity of covid-19-infected patients using machine learning techniques," *Computers*, vol. 10, no. 3, p. 31, 2021. pages 4, 30

[25] K. Ikemura, E. Bellin, Y. Yagi, H. Billett, M. Saada, K. Simone, L. Stahl, J. Szymanski, D. Goldstein, and M. R. Gil, "Using automated machine learning to predict the mortality of patients with covid-19: Prediction model development study," *Journal of medical Internet research*, vol. 23, no. 2, p. e23458, 2021. pages 4, 30

[26] J. Salamon and J. P. Bello, "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, mar 2017. pages 5, 31

[27] I. Kiskin, D. Zilli, Y. Li, M. Sinka, K. Willis, and S. Roberts, "Bioacoustic detection with wavelet-conditioned convolutional neural networks," *Neural Computing and Applications*, vol. 32, no. 4, pp. 915–927, 2020. pages 5, 31

[28] A. Meintjes, A. Lowe, and M. Legget, "Fundamental Heart Sound Classification using the Continuous Wavelet Transform and Convolutional Neural Networks," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2018-July, pp. 409–412, 2018. pages 5, 31

[29] M. Huzaifah, "Comparison of Time-Frequency Representations for Environmental Sound Classification using Convolutional Neural Networks," *arXiv*, pp. 1–5, 2017. pages 5, 31

[30] T. Lidy and A. Schindler, "Cqt-based convolutional neural networks for audio scene classification," in *Proceedings of the detection and classification of acoustic scenes and events 2016 workshop (DCASE2016)*, vol. 90. IEEE Budapest, 2016, pp. 1032–1048. pages 5, 31

[31] Z. Mushtaq, S. F. Su, and Q. V. Tran, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," *Applied Acoustics*, vol. 172, p. 107581, 2021. pages 5, 31, 63

[32] R. V. Sharan and T. J. Moir, "Subband Time-Frequency Image Texture Features for Robust Audio Surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 12, pp. 2605–2615, 2015. pages 5, 32

[33] R. V. Sharan and T. J. Moir, "Acoustic event recognition using cochleagram image and convolutional neural networks," vol. 148. Elsevier, 2019, pp. 62–66. pages 5, 32

[34] R. Hyder, S. Ghaffarzadegan, Z. Feng, J. H. Hansen, and T. Hasan, "Acoustic scene classification using a CNN-Supervector system trained with auditory and spectrogram image features," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-Augus, no. August, pp. 3073–3077, 2017. pages 5, 32

[35] M. Wang, R. Wang, X. L. Zhang, and S. Rahardja, "Hybrid constant-Q transform based CNN ensemble for acoustic scene classification," *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, no. November, pp. 1511–1516, 2019. pages 5, 32, 41

[36] J. Amoh and K. Odame, "Deep Neural Networks for Identifying Cough Sounds," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 5, pp. 1003–1011, oct 2016. pages 5, 32, 61

[37] J. Amoh and K. Odame, "Deepcough: A deep convolutional neural network in a wearable cough detection system," pp. 1–4, 2015. pages 5, 33

[38] C. Bales, M. Nabeel, C. N. John, U. Masood, H. N. Qureshi, H. Farooq, I. Posokhova, and A. Imran, "Can Machine Learning Be Used to Recognize and Diagnose Coughs?" in *2020 International Conference on e-Health and Bioengineering (EHB)*, oct 2020, pp. 1–4. pages 5, 33

[39] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, p. 65, 2017. pages 5, 33

[40] I. D. Miranda, A. H. DIacon, and T. R. Niesler, "A Comparative Study of Features for Acoustic Cough Detection Using Deep Architectures*," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 2601–2605, 2019. pages 5, 33

[41] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung sounds classification using convolutional neural networks," *Artificial Intelligence in Medicine*, vol. 88, pp. 58–69, 2018. pages 5, 33

[42] F. Barata, K. Kipfer, M. Weber, P. Tinschert, E. Fleisch, and T. Kowatsch, "Towards Device-Agnostic Mobile Cough Detection with Convolutional Neural Networks," in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, jun 2019, pp. 1–11. pages 5, 33

[43] Hui-Hui Wang, Jia-Ming Liu, Mingyu You, and Guo-Zheng Li, "Audio signals encoding for cough classification using convolutional neural networks: A comparative study," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, nov 2015, pp. 442–445. pages 5, 34

[44] Y. Yin, D. Tu, W. Shen, and J. Bao, "Recognition of sick pig cough sounds based on convolutional neural network in field situations," *Information Processing in Agriculture*, 2020. pages 5, 34

[45] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3474–3484, 2020. pages 5, 34, 35, 50

[46] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020. pages 5, 35, 63, 72

[47] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210. pages 5, 35, 72

[48] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018. pages 5, 35, 72

[49] G. Chaudhari, X. Jiang, A. Fakhry, A. Han, J. Xiao, S. Shen, and A. Khanzada, "Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough audio samples," *arXiv*, 2020. pages 5, 35

[50] M. Pahar and T. Niesler, "Machine Learning based COVID-19 Detection from Smartphone Recordings: Cough, Breath and Speech," 2021. pages 5, 35, 63, 65

[51] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, and R. Panicker, "Cough against COVID: evidence of COVID-19 signature in cough sounds," *CoRR*, vol. abs/2009.08790, 2020. pages 5, 36, 72

[52] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec 1943. pages 6, 42, 43, 44

[53] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958. pages 6, 43, 44

[54] B. Widrow and M. E. Hoff, "Adaptive switching circuits," Stanford Univ Ca Stanford Electronics Labs, Tech. Rep., 1960. pages 6, 42, 44

[55] M. Minsky and S. Papert, "An introduction to computational geometry," *Cambridge tiass., HIT*, 1969. pages 6, 43, 44

[56] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, vol. 43, no. 1, pp. 59–69, 1982. pages 6, 44

[57] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982. pages 6, 44

[58] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. pages 6, 44

[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., 2012, p. 1097–1105. pages 6, 44

[60] O. W. in Data. Coronavirus (covid-19) cases. [Online]. Available: https://ourworldindata.org/covid-cases pages 19, 27, 28

[61] govgr. Daily overview. [Online]. Available: https://covid19.gov.gr/covid19-live-analytics/ pages 19, 29

[62] C. Stangor, J. Walinga *et al.*, "Introduction to psychology-1st canadian edition," 2014. pages 19, 42

[63] S. Narkhede, "Understanding auc-roc curve," *Towards Data Science*, vol. 26, pp. 220–227, 2018. pages 19, 49

[64] E. C. for Disease Prevention and Control. Sars-cov-2 variants of concern as of 9 september 2021. [Online]. Available: https://www.ecdc.europa.eu/en/covid-19/variants-concern pages 21, 27, 28

[65] N. Van Doremalen, T. Bushmaker, D. H. Morris, M. G. Holbrook, A. Gamble, B. N. Williamson, A. Tamin, J. L. Harcourt, N. J. Thornburg, S. I. Gerber *et al.*, "Aerosol and surface stability of sars-cov-2 as compared with sars-cov-1," *New England journal of medicine*, vol. 382, no. 16, pp. 1564–1567, 2020. pages 24

[66] A. G. Harrison, T. Lin, and P. Wang, "Mechanisms of sars-cov-2 transmission and pathogenesis," *Trends in immunology*, 2020. pages 24

[67] E. Commission. How do vaccines work? [Online]. Available: https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans_en pages 26

[68] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc, K. Huang, K. S. Nikita, B. P. Veasey, M. Zervakis, J. H. Saltz, and C. S. Pattichis, "Ai in medical imaging informatics: Current challenges and future directions," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 7, pp. 1837–1857, 2020. pages 30

[69] E. S. Adamidi, K. Mitsis, and K. S. Nikita, "Artificial intelligence in clinical care amidst covid-19 pandemic: A systematic review," *Computational and Structural Biotechnology Journal*, 2021. pages 30

[70] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," *arXiv preprint arXiv:1807.09902*, 2018. pages 36, 72

[71] F. Al Hossain, A. A. Lover, G. A. Corey, N. G. Reich, and T. Rahman, "Flusense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–28, 2020. pages 36, 72

[72] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, R. Nirmala, P. Kumar Ghosh, and S. Ganapathy, "Coswara - A database of breathing, cough, and voice sounds for COVID-19 diagnosis," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-Octob, pp. 4811–4815, 2020. pages 36, 58, 72

[73] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," 2018. pages 36

[74] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," 2020. pages 39

[75] N. Kehtarnavaz, "Chapter 7 - frequency domain processing," in *Digital Signal Processing System Design (Second Edition)*, second edition ed., N. Kehtarnavaz, Ed. Burlington: Academic Press, 2008, pp. 175–196. pages 39

[76] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The journal of the acoustical society of america*, vol. 8, no. 3, pp. 185–190, 1937. pages 40

[77] D. O'Shaughnessy, "Speech communication, human and machine addison wesley," *Reading MA*, 1987. pages 40

[78] "1990 - Brown - Calculation of a constant Q spectral transform.pdf." pages 40

[79] M. Wan, R. Wang, B. Wang, J. Bai, C. Chen, Z. Fu, J. Chen, X. Zhang, and S. Rahardja, "Ciaic-asc system for dcase 2019 challenge task1," *Tech. Rep., DCASE2019 Challenge*, 2019. pages 41

[80] T. M. Mitchell, "The Discipline of Machine Learning," pp. 1–7, 2006. pages 41

[81] S. S. Haykin, *Neural networks and learning machines*, 3rd ed. Upper Saddle River, NJ: Pearson Education, 2009. pages 42

[82] G. M. Khan, "Artificial neural network (ANNs)," pp. 39–55, 2018. pages 42

[83] J. Sandler and B. Rosenblatt, "The concept of the representational world," *The Psychoanalytic Study of the Child*, vol. 17, no. 1, pp. 128–145, 1962. pages 43

[84] ImageNet. Imagenet large scale visual recognition challenge 2012 (ilsvrc2012). pages 44

[85] R. Gonzalez and R. Woods, *Digital Image Processing*. Pearson, 2018. pages 44

[86] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. pages 46

[87] F. Chollet *et al.*, "Keras," 2015. pages 46

[88] E. A. McGlynn, K. M. McDonald, and C. K. Cassel, "Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the institute of medicine," *Jama*, vol. 314, no. 23, pp. 2501–2502, 2015. pages 46

[89] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019. pages 47

[90] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *The New England journal of medicine*, vol. 375, no. 13, p. 1216, 2016. pages 47

[91] 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," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2020, pp. 859–864. pages 47

[92] M. Skevofilakas, K. Zarkogianni, B. G. Karamanos, and K. S. Nikita, "A hybrid decision support system for the risk assessment of retinopathy development as a long

term complication of type 1 diabetes mellitus," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 6713–6716. pages 47

[93] I. K. Valavanis, S. G. Mougiakakou, K. A. Grimaldi, and K. S. Nikita, "A multifactorial analysis of obesity as cvd risk factor: use of neural network based methods in a nutrigenetics context," *BMC bioinformatics*, vol. 11, no. 1, pp. 1–10, 2010. pages 47

[94] L. Orlandic, T. Teijeiro, and D. Atienza, "The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021. pages 53

[95] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. pages 63

[96] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708. pages 63

[97] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258. pages 64

[98] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826. pages 64

[99] A. Windmon, M. Minakshi, P. Bharti, S. Chellappan, M. Johansson, B. A. Jenkins, and P. R. Athilingam, "Tussiswatch: A smart-phone system to identify cough episodes as early symptoms of chronic obstructive pulmonary disease and congestive heart failure," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1566–1573, 2018. pages 65

[100] R. Kumar, R. Arora, V. Bansal, V. Sahayasheela, H. Buckchash, J. Imran, N. Narayanan, G. Pandian, and B. Raman, "Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers," 2020. pages 65

[101] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. pages 66

[102] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. pages 66

[103] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25. pages 66, 67

[104] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognition*, vol. 48, no. 5, pp. 1623–1637, 2015. pages 67

[105] M. Bahrami and H. Sajedi, "Image concept detection in imbalanced datasets with ensemble of convolutional neural networks," *Intelligent Data Analysis*, vol. 23, no. 5, pp. 1131–1144, 2019. pages 67