ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

«ΣΧΕΔΙΑΣΗ ΜΕΘΟΔΩΝ ΨΗΦΙΑΚΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΩΝΗΣ ΓΙΑ ΤΗ ΔΙΑΓΝΩΣΗ ΝΟΣΟΥ COVID-19»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παρασκευάς Α. Βογιατζόπουλος

Επιβλέπων : Κωνσταντίνα Νικήτα

Καθηγήτρια Ε.Μ.Π

Αθήνα, Οκτώβριος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ
ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

«ΣΧΕΔΙΑΣΗ ΜΕΘΟΔΩΝ ΨΗΦΙΑΚΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΩΝΗΣ ΓΙΑ ΤΗ ΔΙΑΓΝΩΣΗ ΝΟΣΟΥ COVID-19»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παρασκευάς Α. Βογιατζόπουλος

Επιβλέπων : Κωνσταντίνα Νικήτα

Καθηγήτρια Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 14η Οκτωβρίου 2024:

Νικήτα Κωνσταντίνα	Γεώργιος Στάμμου	Βουλοδήμος Αθανάσιος
Καθηγήτρια Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Επίκουρος καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2024

.....

Παρασκευάς Α. Βογιατζόπουλος

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

Περίληψη

Η Covid-19 είναι μια μεταδοτική αναπνευστική νόσος που εξαπλώθηκε σε παγκόσμιο επίπεδο το 2020 και ακόμα και σήμερα παραμένει μια σημαντική πρόκληση για τη δημόσια υγεία. Η αναγκαιότητα για γρήγορα και αξιόπιστα διαγνωστικά εργαλεία είναι πιο επιτακτική από ποτέ, λαμβάνοντας υπόψιν τις διαρκείς προκλήσεις που επιφέρει η COVID-19. Η συνεισφορά της μηχανικής μάθησης στην ιατρική είναι σημαντική και πολυδιάστατη, και αφορά τόσο τη διάγνωση και πρόβλεψη της νόσου όσο και τη διαχείριση της δημόσιας υγείας. Στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη μιας μεθόδου για την αυτόματη διάγνωση της νόσου COVID-19. Προς αυτήν την κατεύθυνση, μελετάται η δυνατότητα εκπαίδευσης μοντέλων μηχανικής μάθησης με δεδομένα που συλλέγονται από τη διαδικτυακή εφαρμογή Smarty4Covid. Η μέθοδος αποτελείται από τέσσερα κύρια στάδια: επεξεργασία των ηχητικών καταγραφών, απομόνωση συγκεκριμένων φωνημάτων ανά ασθενή, εξαγωγή χαρακτηριστικών μέσω του μοντέλου VGG και διάγνωση βάσει επιλεγμένων χαρακτηριστικών. Η μέθοδος χρησιμοποιεί το μοντέλο Allosaurus για την απομόνωση φωνημάτων ενδιαφέροντος και εκτίμηση του χρόνου έναρξής τους αντίστοιχα. Από τα αποτελέσματα προκύπτει πως το μοντέλο VGG που εκπαιδεύεται σε συγκεκριμένα φωνήματα υπερτερεί σε όλες τις μετρικές αξιολόγησης σε σύγκριση με μοντέλα VGG που εκπαιδεύονται στο σύνολο της ηχητικής καταγραφής φωνής, γεγονός που συνηγορεί προς την προβλεπτική ικανότητα χαρακτηριστικών που εξάγονται από απομονωμένα φωνήματα για τη διάγνωση της COVID-19.

Λέξεις κλειδιά: Smarty4Covid, Μηχανική μάθηση, Covid-19, φωνήματα

Abstract

Covid-19 is a contagious respiratory disease that spread worldwide in 2020 and remains a significant public health challenge even today. The need for a fast and reliable diagnostic process is more urgent than ever, considering the ongoing challenges posed by the coronavirus (COVID-19) pandemic. The contribution of machine learning to healthcare is significant and multidimensional, encompassing both the diagnosis and prediction of the disease, as well as public health management. The aim of the present thesis is to develop a method for the automatic diagnosis of COVID-19. In this direction, the possibility of training machine learning models with data collected from the online application Smarty4Covid is being explored. The method consists of four main stages: processing of audio recordings, isolation of specific phonemes, extraction of audio features through the VGG model, and diagnosis based on the features of specific phonemes. The method uses the Allosaurus model to isolate phonemes of interest and estimate their respective onset times. This phoneme-based system shows a noticeable improvement in every classification metric compared to systems that use the entire recording for feature extraction, advocating towards the predictive capacity of features extracted from isolated phonemes for COVID-19 diagnosis.

Key Words: Smarty4Covid, Machine Learning, Covid-19, phonemes

Πίνακας Περιεχομένων

1. Εισαγωγή 13
1.1 COVID-19 13
1.1.1 Δομή του COVID-19 13
1.1.2 Χρονική Εξέλιξη και Ανάλυση της Πανδημίας COVID-19 15
1.1.3 Η Προέλευση του COVID-19 και η Σχέση του με Προηγούμενες Πανδημίες και Οικολογικές Αλλαγές 17
1.1.4 Κλινικά Χαρακτηριστικά και Συμπτώματα του COVID-19 19
1.1.5 Διαγνωστικές Μέθοδοι για τον COVID-19 20
2. Σχετικές γνώσεις 22
2.1 Στοιχεία ήχου και ψηφιακής επεξεργασίας σήματος 22
2.1.1 Ορισμός του Ήχου: 22
2.1.2 Βασικά Χαρακτηριστικά του Ήχου και των Ηχητικών Κυμάτων 22
2.1.3 Αντίληψη του Ήχου από το Ανθρώπινο Ακουστικό Σύστημα 26
2.1.4 Δομή και Χαρακτηριστικά του Σήματος Ομιλίας: 28
2.1.5 Φωνητική και Φωνολογία (Phonemics and Phonetics) 29
2.1.7 Ψηφιοποίηση του Ήχου: Από την Αναλογική στην Ψηφιακή Μορφή 29
2.1.8 Βασικά Χαρακτηριστικά για την Επεξεργασία Ομιλίας 32
2.2 Η Μηχανική Μάθηση 45
2.2.1 Η Ιστορική Πορεία της Τεχνητής Νοημοσύνης: Από τις Πρώτες Ιδέες στη Σύγχρονη Εποχή 45
2.2.2 Τι είναι η Μηχανική Μάθηση 47
2.2.3 Κατηγορίες και Εφαρμογές Τεχνικών Μηχανικής Μάθησης 48
2.2.4 Ανασκόπηση Αλγορίθμων Μηχανικής Μάθησης και Βαθιάς Μάθησης 52
2.2.5 Υπερπαράμετροι 68
2.2.6 Μεταφορά Μάθησης (Transfer Learning) 76

2.2.7 Μετρικές Αξιολόγησης Απόδοσης 81
3. Βιβλιογραφική ανασκόπηση 86
3.1 Σχετικές εργασίες 86
3.2 Allosaurus 99
3.3 VGG 104
4. Διαδικασία και Μεθοδολογία Έρευνας 114
4.1 Σκοπός του Πειράματος 114
4.2 Σύνολο δεδομένων Smart4Covid 114
4.3 Μεθοδολογία 125
5. Αποτελέσματα 131
6. Συμπεράσματα 145
6.1. Περιορισμοί 145
6.2 Σχολιασμός αποτελεσμάτων 146
6.3 Μελλοντική δουλειά 150
7. Αναφορές 151

Πίνακες

Πίνακας 1.1: «Γεγονότα και εξέλιξη των κρουσμάτων COVID-19 ανά Μήνα» 17
Πίνακας 1.2: «Φυσιολογικές τιμές μετρήσεων παθολογικών στοιχείων»..... 19
Πίνακας 2.1: «Σύγκριση των πιο πολυχρησιμοποιημένων συναρτήσεων κανονικοποίησης»..... 74
Πίνακας 2.2: «Πίνακας σύγχυσης για πρόβλημα ταξινόμησης» 84
Πίνακας 2.3: «Πίνακας σύγχυσης και τιμές TP,FP,TN,FN»..... 84
Πίνακας 3.1: «Πληροφορίες στο main_questionnaire.json του Smart4Covid»..... 120
Πίνακας 3.2: «Πληροφορίες από τους επαγγελματίες υγείας στα αρχεία json»..... 121
Πίνακας 3.3: «Επισημείωση ειδικών επί του βήχα»..... 121
Πίνακας 3.4: «Επισημείωση ειδικών σχετικά με το βήχα»..... 122
Πίνακας 3.5: «Επισημείωση ειδικών για τις ηχογραφήσεις»..... 123
Πίνακας 5.1: «Αποτελέσματα Πειράματος Βάσης»..... 132
Πίνακας 5.2: «Αποτελέσματα Πειράματος με τη χρήση μόνο Vgg χαρακτηριστικών, 0.1 sec»..... 133
Πίνακας 5.3: «Αποτελέσματα Πειράματος με τη χρήση μόνο Vgg χαρακτηριστικών, 0.2 sec»..... 134
Πίνακας 5.4: «Αποτελέσματα Πειράματος με τη χρήση μόνο Vgg χαρακτηριστικών, 0.5 sec»..... 135
Πίνακας 5.5: «Αποτελέσματα Πειράματος με τη χρήση μόνο Vgg χαρακτηριστικών, 1 sec»..... 135
Πίνακας 5.6: «Αποτελέσματα Πειράματος με τη χρήση μόνο Vgg χαρακτηριστικών, 2 sec»..... 136
Πίνακας 5.7: «Αποτελέσματα Πειράματος με τη χρήση μόνο Vgg χαρακτηριστικών, 5 sec»..... 137
Πίνακας 5.8: «Βέλτιστα φωνήματα με τη χρήση μόνο Vgg χαρακτηριστικών»..... 137
Πίνακας 5.9: «Βέλτιστα φωνήματα με τη χρήση συνδυασμού χαρακτηριστικών»..... 138
Πίνακας 5.10: «Αποτελέσματα με χρήση Vgg και Spectral Centroid χαρακτηριστικών για 0.1 sec » 139
Πίνακας 5.11: «Αποτελέσματα με χρήση όλων των χαρακτηριστικών για 0.2 sec»..... 140
Πίνακας 5.12: «Αποτελέσματα με χρήση όλων των χαρακτηριστικών για 0.5 sec»..... 140
Πίνακας 5.13: «Αποτελέσματα με χρήση όλων των χαρακτηριστικών για 1 sec»..... 141
Πίνακας 5.14: «Αποτελέσματα με χρήση όλων των χαρακτηριστικών για 2 sec»..... 142
Πίνακας 5.15: «Αποτελέσματα με χρήση όλων των χαρακτηριστικών για 5 sec»..... 143
Πίνακας 5.16: «Βέλτιστα φωνήματα με πλειοψηφία με χρήση συνδυασμού χαρακτηριστικών»... 144
Πίνακας 5.17: «Βέλτιστα φωνήματα με πλειοψηφία με χρήση μόνο Vgg χαρακτηριστικών»..... 144

Εικόνες

Εικόνα 1.1: «Ο κύκλος ζωής του κορωνοϊού»..... 14
Εικόνα 2.1: «Τρεις φυσικές έννοιες που περιγράφουν τον ήχο»..... 24
Εικόνα 2.2: «Απεικόνιση διαφοράς φάσης δύο κυμάτων» 26
Εικόνα 2.3: «Ακουστικό Φάσμα και Ευαισθησία Ανθρώπινου Αυτιού» 27
Εικόνα 2.4: «Συνεχές και το αντίστοιχο δειγματοληπτούμενο, κβαντισμένο σήμα» 31
Εικόνα 2.5: «Μετατροπή από το πεδίο του χρόνου στο πεδίο συχνοτήτων» 34
Εικόνα 2.6: «Εφαρμογή του STFT» 36
Εικόνα 2.7: «Mel Spectrogram»..... 38
Εικόνα 2.8: «Διαδικασία εξαγωγής των χαρακτηριστικών MFCC»..... 41
Εικόνα 2.9: «Κατηγοριοποίηση προβλημάτων μηχανικής μάθησης» 51
Εικόνα 2.10: «Αναπαράσταση της τοπολογίας ενός τεχνητού νευρωνικού δικτύου»..... 54
Εικόνα 2.11: «Αναπαράσταση απόδοσης Deep Learning, Machine Learning»..... 56
Εικόνα 2.12: «’Ενας νευρώνας Perceptron» 57
Εικόνα 2.13: «Βηματική συνάρτηση (unit step - heavyside)» 58
Εικόνα 2.14: «Γραμμική συνάρτηση » 59
Εικόνα 2.15: «Σιγμοειδής Συνάρτηση» 60
Εικόνα 2.16: «Multi-Layer Perceptron - MLP» 62
Εικόνα 2.17: «Σχηματική απεικόνιση ενός συνελικτικού νευρωνικού δικτύου» 65
Εικόνα 2.18: «Εφαρμογή ενός φίλτρου στη διαδικασία συνέλιξης» 66
Εικόνα 2.19: «: Απεικόνιση της διαδικασίας μεταφοράς μάθησης» 77
Εικόνα 2.20: «Γραφική παράσταση για την κατανόηση της μετρικής ROC-AUC» 85
Εικόνα 3.1: «Απεικόνιση της αντιστοιχίας των φωνημάτων σε διαφορετικές γλώσσες» 100
Εικόνα 4.1: «Σύστηματα καταγραφής δεδομένων – Smart4Covid» 124
Εικόνα 4.2: «Απεικόνιση του τελικού σχήματος που χρησιμοποιήθηκε για τα πειράματα» 126
Εικόνα 4.3: «Ανάλυση των ασθενών για το σύνολο εκπαίδευσης και επαλήθευσης» 128
Εικόνα 5.1: «Συχνότητα εμφάνισης κάθε φωνήεντος στο σύνολο ελέγχου» 131

1. Εισαγωγή

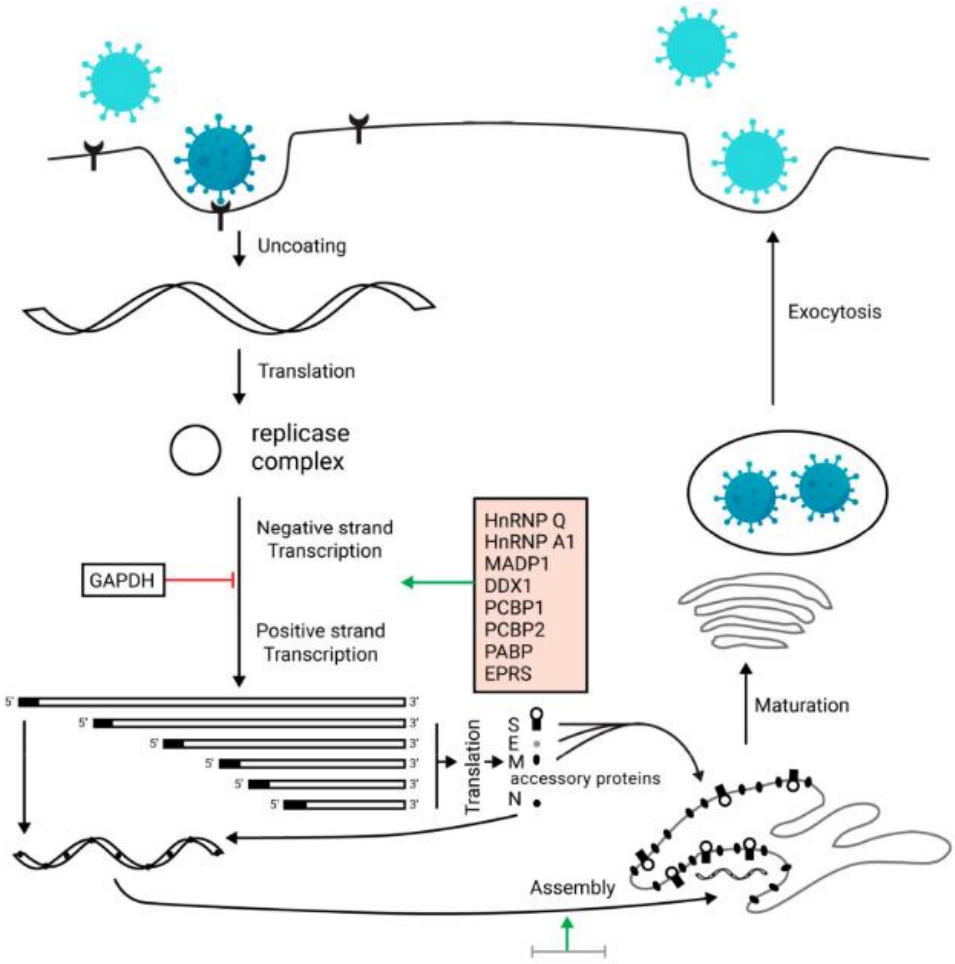
Σε αυτήν την ενότητα πρώτα εξετάζεται η ιστορική και επιδημιολογική προοπτική της COVID-19. Στην συνέχεια γίνεται ανάλυση της συμβολής της μηχανικής μάθησης στην ιατρική και ειδικά στην διάγνωση ασθενειών. Έπειτα, παρουσιάζονται οι στόχοι της εργασίας με έμφαση στην ανάπτυξη μεθόδου για την αυτόματη διάγνωση της COVID-19, γίνεται ανάλυση της μεθοδολογίας που ακολουθήθηκε με εστίαση στην επεξεργασία και ανάλυση ηχητικών δεδομένων. Τέλος, επισημαίνεται η προσδοκώμενη συνεισφορά της εργασίας στην προαγωγή της δημόσιας υγείας και την αντιμετώπιση της πανδημίας.

1.1 COVID-19

1.1.1 Δομή του COVID-19

Πριν από περίπου τρία έτη, η ανθρωπότητα βρέθηκε αντιμέτωπη με την COVID-19, μια νέα νόσο η οποία κηρύχθηκε παγκόσμια υγειονομική έκτακτη ανάγκη τον Ιανουάριο του 2020 και πανδημία τον Μάρτιο του 2020 από τον Παγκόσμιο Οργανισμό Υγείας . Ο νέος κορωνοϊός που ευθύνεται για αυτή την πανδημία ονομάστηκε "2019-nCoV" και αργότερα άλλαξε σε "SARS-CoV-2" λόγω της μεγάλης ομοιότητάς του με τον κορωνοϊό του σοβαρού οξέος αναπνευστικού συνδρόμου (SARS-CoV). Το γονιδίωμα του ιού SARS-CoV-2 αποτελείται από 29.903 νουκλεοτίδια και παρουσιάζει μεγάλη ομοιότητα με τους κορωνοϊούς τύπου SARS που μολύνουν νυχτερίδες, έχοντας 89,1% ομοιότητα στα νουκλεοτίδια. Όπως και άλλοι κορωνοϊοί, ο SARS-CoV-2 διαθέτει μεγάλο γονιδίωμα ριβοζονουκλεϊκού οξέος-Ribonucleic acid (RNA) και παρουσιάζει μια μοναδική στρατηγική αναπαραγωγής. Επιπλέον, το RNA έχει μια ειδική δομή στο ένα άκρο που ονομάζεται "καπέλο 5'," η οποία βοηθά στην προστασία του RNA και υποβοηθά την έναρξη της σύνθεσης πρωτεϊνών. Στο άλλο άκρο, έχει μια "ουρά poly(A) 3'," μια μακριά αλυσίδα από αδενίνες νουκλεοτίδια που επίσης βοηθά στη σταθεροποίηση του RNA και διευκολύνει τη μετάφρασή του σε πρωτεΐνες.

Αυτή η μοναδική δομή επιτρέπει στις πολυπρωτεΐνες της αναπαραγωγής να διαβάσουν το γονιδίωμα για μετάφραση. Τα δύο τρίτα του γονιδιώματος κωδικοποιούν μη δομικές πρωτεΐνες (Non Structural Proteins - NSPs), ενώ το υπόλοιπο ένα τρίτο κωδικοποιεί δομικές και βοηθητικές πρωτεΐνες. Οι σημαντικές δομικές πρωτεΐνες είναι οι πρωτεΐνες spike (S), membrane (M), envelope (E) και nucleocapsid (N). Επιπλέον, δύο πολυπρωτεΐνες ιικής αντιγραφής που ονομάζονται PP1a και PP1ab παράγονται από τον ιό, οι οποίες τροποποιούνται σε 16 ώριμες NSPs [26,27]. Η εικόνα 1.1 απεικονίζει τη ζωή του κύκλου των CoVs.



Εικόνα 1.1: Ο κύκλος ζωής ενός κορωνοϊού. Μετά την είσοδό του στους πνεύμονες του ξενιστή, ο κορωνοϊός προσκολλάται στους υποδοχείς του μετατρεπτικού ενζύμου της αγγειοτενσίνης (ACE-2) και απελευθερώνει το γενετικό του υλικό στο κυτταρόπλασμα. Η μετάφραση του RNA δημιουργεί το σύμπλεγμα της αναπαραγωγής, που έχει ως αποτέλεσμα το σχηματισμό μεταγραφών διαφόρων πρωτεϊνών (ονομάζονται στο ροζ πλαίσιο). Η μετάφραση αυτών των μεταγραφών παράγει τις πρωτεΐνες ακίδας (S), φακέλου (E), μεμβράνης (M), νουκλεοκαψίδας (N) και άλλες βοηθητικές πρωτεΐνες. Η συναρμολόγηση αυτών των πρωτεϊνών παράγει ένα νέο ιόσωμα που ωριμάζει και απελευθερώνεται από το κύτταρο για να μολύνει άλλα κύτταρα. [1]

1.1.2 Χρονική Εξέλιξη και Ανάλυση της Πανδημίας COVID-19

Χρονικά, πέντε ασθενείς εισήχθησαν στο νοσοκομείο στην Κίνα μεταξύ 18 και 29 Δεκεμβρίου 2019, εκ των οποίων ένας απεβίωσε. Το πρώτο δείγμα του ιού ελήφθη από έναν ασθενή στις 7 Ιανουαρίου 2020. Ως απάντηση, η αλληλουχία του γονιδιώματος του ιού δημοσιεύθηκε από Κινέζους επιστήμονες στις 10 Ιανουαρίου 2020 στο Global Initiative on Sharing All Influenza Data (GISAID).

Επιπλέον, 41 ακόμη ασθενείς εισήχθησαν στα νοσοκομεία έως τις 2 Ιανουαρίου 2020. Η διάγνωση αυτών των ασθενών επιβεβαίωσε την παρουσία της COVID-19. Η Ταϊλάνδη ήταν η πρώτη χώρα που μολύνθηκε μετά την Κίνα, και το πρώτο κρούσμα αναφέρθηκε στην Ταϊλάνδη στις 13 Ιανουαρίου 2020, σε έναν ασθενή που ταξίδεψε στην Ταϊλάνδη. Πεντακόσια εβδομήντα ένα νέα κρούσματα COVID-19 αναφέρθηκαν μέχρι τις 22 Ιανουαρίου 2020, σε 25 επαρχίες της Κίνας. Η Εθνική Επιτροπή Υγείας της Κίνας (National Health Commission of People's Republic of China - NHC) επιβεβαίωσε τους πρώτους 17 θανάτους λόγω COVID-19 μέχρι τις 22 Ιανουαρίου 2020. Στις 25 Ιανουαρίου 2020, υπήρχαν συνολικά 1975 μολυσμένοι ασθενείς στην Κίνα, εκ των οποίων 56 απεβίωσαν.

Σύμφωνα με μια άλλη αναφορά από τις 30 Ιανουαρίου 2020, συνολικά 7734 κρούσματα αναφέρθηκαν στην Κίνα, και 90 κρούσματα επιβεβαιώθηκαν από άλλες χώρες, όπως το Βιετνάμ, το Νεπάλ, η Σρι Λάνκα, η Ιαπωνία, η Ταϊλάνδη, η Νότια Κορέα, οι Ηνωμένες Πολιτείες, τα Ηνωμένα Αραβικά Εμιράτα, η Μαλαισία, οι Φιλιππίνες, η Ινδία, η Αυστραλία, ο Καναδάς, η Καμπότζη, η Φινλανδία, η Σιγκαπούρη, η Γαλλία, η Γερμανία και η Ταϊβάν.

Το Κινεζικό Κέντρο Ελέγχου και Πρόληψης Νοσημάτων ανέφερε 44.672 επιβεβαιωμένα κρούσματα μέχρι τις 11 Φεβρουαρίου 2020, εκ των οποίων το 81% των θανάτων σημειώθηκε σε ασθενείς ηλικίας άνω των 60 ετών. Τα ποσοστά θνησιμότητας σε ασθενείς ηλικίας 70-79 ετών και άνω των 80 ετών ήταν 8,0% και 14,8% αντίστοιχα. Στις 28 Φεβρουαρίου 2020, ο ΠΟΥ ανέφερε 82.000 επιβεβαιωμένα κρούσματα παγκοσμίως, και η έξαρση είχε φτάσει σε 45 χώρες εκτός της Κίνας. Ακολουθώντας την ταχεία πορεία μετάδοσης, 90.000 μολυσμένα κρούσματα και 60 χώρες αντιμετώπιζαν την έξαρση του COVID-19 μέχρι τις 2 Μαρτίου 2020. Προηγουμένως, θεωρούνταν ότι τα παιδιά ήταν λιγότερο ευαίσθητα στη μόλυνση λόγω έλλειψης στοιχείων, αλλά στις 5 Μαρτίου 2020, οι κινεζικές μελέτες κατέληξαν ότι τα παιδιά είναι εξίσου ευαίσθητα με τους ενήλικες.

Μέχρι τις 13 Μαρτίου 2020, η Ευρώπη είχε αναγνωριστεί ως το επίκεντρο της πανδημίας από τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ), με τα περισσότερα αναφερόμενα κρούσματα να εμφανίζονται στην ήπειρο, ιδίως στην Ιταλία, η οποία αντιμετώπιζε τη μεγαλύτερη έξαρση. Ωστόσο, στις 13 Μαρτίου 2020, οι Ηνωμένες Πολιτείες κήρυξαν εθνική έκτακτη ανάγκη, καθώς η εξάπλωση της πανδημίας επιταχύνθηκε και άρχισε να επηρεάζει σοβαρά τη χώρα. Μέχρι τις 27 Μαρτίου 2020, ο συνολικός αριθμός των μολυσμένων ασθενών παγκοσμίως έφτασε το μισό εκατομμύριο, με την πανδημία να έχει εξαπλωθεί σε

περίπου 175 χώρες. Στις 2 Απριλίου 2020, ο αριθμός των κρουσμάτων ξεπέρασε το ένα εκατομμύριο. Η εξάπλωση συνεχίστηκε με ραγδαίο ρυθμό, και στις 15 Απριλίου 2020, τα κρούσματα παγκοσμίως ανήλθαν στα δύο εκατομμύρια. Πλέον, το επίκεντρο της πανδημίας είχε μετατοπιστεί στις Ηνωμένες Πολιτείες, οι οποίες κατέλαβαν την πρώτη θέση σε αριθμό κρουσμάτων, ακολουθούμενες από την Ιταλία και την Ισπανία.

Από τα δεδομένα που παρουσιάστηκαν, παρατηρείται μια εκθετική αύξηση της εξάπλωσης του COVID-19. Ξεκινώντας από μεμονωμένες περιπτώσεις κρουσμάτων στην Κίνα στα τέλη Δεκεμβρίου του 2019, ο ιός εξαπλώθηκε γρήγορα και, μέσα σε περίπου δύο μήνες, έφτασε σε περίπου 90.000 κρούσματα σε 60 χώρες. Στη συνέχεια, μόλις μέσα σε ένα ακόμη διάστημα ενός μήνα, τα κρούσματα αυξήθηκαν ραγδαία, καλύπτοντας περισσότερες από 175 χώρες και ξεπερνώντας το φράγμα των 2 εκατομμυρίων μολυσμένων ατόμων μέχρι τα μέσα Απριλίου 2020. Αυτή η εκθετική αύξηση των κρουσμάτων υπογραμμίζει την ταχεία μετάδοση του ιού και τη σοβαρότητα της πανδημίας, η οποία αντικατοπτρίζει την υψηλή μεταδοτικότητα και τον αναπαραγωγικό αριθμό (R0). Οι κινεζικές αναφορές δείχνουν μια τιμή R0 από 2,2 έως 2,7, που σημαίνει ότι τα μολυσμένα κρούσματα διπλασιάζονταν κάθε 6-7 ημέρες. Ο ρυθμός μετάδοσης ήταν εξαιρετικά υψηλός, με αποτέλεσμα να ιδρυθούν αρκετές ιατρικές εγκαταστάσεις και να αναφερθούν κατευθυντήριες γραμμές από το αμερικάνικο Centers for Disease Control and Prevention (CDC) και τον ΠΟΥ για τη διαχείριση της αυξανόμενης νόσου.

Ένας άλλος παράγοντας για την κατανόηση της σοβαρότητας και της πρόγνωσης της λοίμωξης είναι ο λόγος θνησιμότητας (CFR), που υπολογίζεται διαιρώντας τον αριθμό των αποβιωσάντων ασθενών με τον συνολικό αριθμό των διαγνωσμένων ασθενών και πολλαπλασιάζοντας με το 100. Σύμφωνα με την έκθεση κατάστασης-185 του ΠΟΥ για τον COVID-19, ο συνολικός CFR ήταν 5,12 μέχρι τις 23 Ιουλίου 2020.

(18-19)/12/2019: 5 ασθενείς εισήχθησαν στο νοσοκομείο στην Κίνα. 1 θάνατος.
2/01/2020: 41 ασθενείς εισήχθηκαν στα νοσοκομεία, επιβεβαιώθηκε η παρουσία του Covid.
13/01/2020: Το πρώτο κρούσμα εκτός Κίνας αναφέρθηκε στην Ταϊλάνδη
22/01/2020: 571 νέα κρούσματα COVID-19 αναφέρθηκαν σε 25 επαρχίες της Κίνας. 17 θάνατοι.
25/01/2020: Συνολικά 1975 μολυσμένοι ασθενείς στην Κίνα, εκ των οποίων 56 απεβίωσαν.
31/01/2020: Συνολικά 7734 κρούσματα στην Κίνα. 90 κρούσματα επιβεβαιώθηκαν σε άλλες χώρες.
11/02/2020: 44.672 επιβεβαιωμένα κρούσματα στην Κίνα.Το 81% των θανάτων σημειώθηκε σε ασθενείς ηλικίας άνω των 60 ετών.
28/02/2020: 82.000 επιβεβαιωμένα κρούσματα παγκοσμίως. Έξαρση σε 45 χώρες εκτός Κίνας.
2/03/2020: 90.000 μολυσμένα κρούσματα σε 60 χώρες.

5/03/2020: Μελέτες δείχνουν ότι τα παιδιά είναι εξίσου ευαίσθητα με τους ενήλικες στη μόλυνση.
13/03/2020: Η Ευρώπη κηρύχθηκε επίκεντρο της έξαρσης από τον ΠΟΥ. Ο Πρόεδρος των ΗΠΑ κήρυξε εθνική έκτακτη ανάγκη στις ΗΠΑ.
27/03/2020: Συνολικός αριθμός μολυσμένων ασθενών έφτασε το μισό εκατομμύριο. Η πανδημία επεκτάθηκε σε περίπου 175 χώρες.
2/04/2020: Παγκόσμιο όριο του ενός εκατομμυρίου κρουσμάτων επιτεύχθηκε.
15/04/2020: 2 εκατομμύρια κρούσματα παγκοσμίως. Οι ΗΠΑ στην κορυφή της λίστας, ακολουθούμενες από την Ιταλία και την Ισπανία.
23/07/2020: Ο ΠΟΥ αναφέρει συνολικό CFR 5,12%.

Πίνακας 1.1: Σημαντικά γεγονότα και εξέλιξη των κρουσμάτων COVID-19 ανά Μήνα.

1.1.3 Η Προέλευση του COVID-19 και η Σχέση του με Προηγούμενες Πανδημίες και Οικολογικές Αλλαγές

Μια σημαντική πρότερη ασθένεια που σχετίζεται με τον COVID-19 είναι το Σύνδρομο Επίκτητης Ανοσοανεπάρκειας (Acquired Immune Deficien - AIDS), το οποίο προκαλείται από τον ιό της ανθρώπινης ανοσοανεπάρκειας (Human Immunodeficiency Virus - HIV). Αυτή η ασθένεια ανακαλύφθηκε το 1981 και μέχρι το 2018 είχε επηρεάσει περίπου 40 εκατομμύρια ανθρώπους και είχε προκαλέσει πάνω από 750.000 θανάτους. Οι ιοί του HIV είναι το αποτέλεσμα πολλαπλών μεταδόσεων μεταξύ ειδών, που αρχικά μολύνουν πίθηκους και άλλα θηλαστικά της Αφρικής. Αυτές οι μεταδόσεις συχνά οδηγούσαν σε ιούς που περιορισμένα εξαπλώνονταν στους ανθρώπους, μέχρι μια συγκεκριμένη μετάδοση από χιμπατζήδες στο νοτιοανατολικό Καμερούν, που οδήγησε στην πανδημία στους ανθρώπους.

Η μετάδοση ιών από άγρια ζώα στους ανθρώπους δεν είναι σπάνια. Πολλοί ανθρώπινοι παθογόνοι ιοί προέρχονται από ιούς που αρχικά μολύνουν ζώα πριν αρχίσουν να μεταδίδονται αποκλειστικά στους ανθρώπους. Από την εμφάνιση του AIDS, πολλές άλλες επιδημικές λοιμώδεις ασθένειες, όπως ο Έμπολα, ο SARS και ο MERS, έχουν προκληθεί από τη μετάδοση ιών από άγρια ζώα στους ανθρώπους.

Οι μεταδόσεις αυτές δεν είναι απαραίτητα τυχαίες. Υπάρχουν ισχυρές ενδείξεις ότι οι οικολογικές αλλαγές έχουν οδηγήσει σε αυξημένα ποσοστά αναδυόμενων και επανερχόμενων ασθενειών, όπως η μαλάρια, το σύνδρομο πνευμονικού ιού Hanta, ο ιός Nipah και η νόσος του ιού Έμπολα. Οι ανθρώπινες δραστηριότητες διαταράσσουν ολοένα και περισσότερο τους φυσικούς οικότοπους και τα οικοσυστήματα της γης, μεταβάλλοντας έντονα τα πρότυπα και τους μηχανισμούς αλληλεπίδρασης μεταξύ των ειδών και ανθρώπων.

Υπάρχουν ενδείξεις ότι μέχρι το 2050 θα κατασκευαστούν 25 εκατομμύρια χιλιόμετρα νέων δρόμων, και οι 9 στους 10 θα κατασκευαστούν σε αναπτυσσόμενες χώρες, συμπεριλαμβανομένων περιοχών με εξαιρετική βιοποικιλότητα και ζωτικής σημασίας οικοσυστήματα. Αυτοί οι δρόμοι συχνά προκαλούν

απώλεια και κατακερματισμό οικοτόπων, πυρκαγιές δασών και άλλες περιβαλλοντικές υποβαθμίσεις, συχνά με μη αναστρέψιμες επιπτώσεις στα οικοσυστήματα.

Στην περίπτωση του COVID-19, το Κινεζικό Κέντρο Ελέγχου και Πρόληψης Νοσημάτων επιβεβαίωσε ότι ο ιός που προκάλεσε την έξαρση του COVID-19 στην Ουχάν προήλθε από άγρια ζώα, το κρέας των οποίων πωλούνταν στην αγορά Χανκού. Οι πρώτοι ασθενείς με SARS-CoV-2 στην Ουχάν ήταν κυρίως έμποροι αυτής της αγοράς. Οι συνθήκες της προέλευσης του COVID-19 είναι παρόμοιες με αυτές του SARS, μια έξαρση κατά την οποία οι πρώτοι ασθενείς ήταν επίσης έμποροι άγριων ζώων στην πόλη Γκουανγκντόνγκ.

Ο John Vidal ανέδειξε τη σύνδεση μεταξύ του COVID-19 και της υγείας του πλανήτη. Ο COVID-19 αποτελεί χαρακτηριστικό παράδειγμα ασθένειας της Ανθρωποκαίνου Εποχής, μιας περιόδου στην οποία οι ανθρώπινες δραστηριότητες έχουν σημαντικό και πολλές φορές αρνητικό αντίκτυπο στο περιβάλλον και το κλίμα του πλανήτη. Η καταστροφή των φυσικών οικοτόπων και η εξαφάνιση των ειδών, η ανεπαρκώς ρυθμισμένη σύλληψη, εμπορία και κατανάλωση μη ανθρώπινων ζώων, η επιρροή των λόμπι που ακυρώνουν ή καθυστερούν μέτρα προστασίας των φυσικών και κοινωνικών συστημάτων, η περιορισμένη τρέχουσα επιστημονική γνώση και η περιφρόνηση από κυβερνήσεις και εταιρείες των διαθέσιμων αποδεικτικών στοιχείων, έχουν λειτουργήσει συντονισμένα για να διευκολύνουν και να οδηγήσουν στην πανδημία COVID-19.[2]

1.1.4 Κλινικά Χαρακτηριστικά και Συμπτώματα του COVID-19

Ο SARS-CoV-2 έχει μια περίοδο επώασης (το χρονικό διάστημα που μεσολαβεί από τη στιγμή που κάποιος μολύνεται από έναν παθογόνο οργανισμό μέχρι την εμφάνιση των πρώτων συμπτωμάτων) που διαρκεί περίπου 4 ημέρες. Σε αυτήν την περίοδο υπάρχει πιθανότητα μετάδοσης από ασυμπτωματικούς φορείς. Το διάστημα από την έναρξη της νόσου μέχρι το τέλος της, είτε πρόκειται για ανάρρωση είτε για θάνατο, κυμαίνεται από 6 έως 41 ημέρες, με μέσο όρο τις 14 ημέρες. Η ηλικία του ασθενούς και η κατάσταση του ανοσοποιητικού συστήματος είναι οι κύριοι παράγοντες που καθορίζουν αυτή την περίοδο. Άτομα με αδύναμο ανοσοποιητικό σύστημα, όπως είναι ηλικιωμένοι άνω των 60 ετών, είναι πιο ευάλωτα.

Τα πιο κοινά συμπτώματα του COVID-19 είναι ξηρός βήχας, μυαλγία, διάρροια, πονοκέφαλος και υψηλός πυρετός. Ο πυρετός αναφέρεται σε περισσότερα από το 90% των περιπτώσεων, ο βήχας στο 75%, και η δύσπνοια στο 50% των ασθενών. Μικρότερος αριθμός ασθενών μπορεί επίσης να εμφανίσει νεφρική βλάβη, σύνδρομο οξείας αναπνευστικής δυσχέρειας (ARDS), σηπτικό σοκ και οξεία καρδιακή βλάβη που απαιτεί νοσηλεία.

Οι κλινικές εκδηλώσεις, δηλαδή τα συμπτώματα και οι ενδείξεις που παρουσιάζει ένας ασθενής κατά τη διάρκεια της νόσου, ενσωματώνονται με τη σοβαρότητα της νόσου. Όταν η νόσος είναι ήπια, κάτι που συμβαίνει στο 81% των ασθενών, δεν υπάρχουν σημάδια πνευμονίας. Σε περιπτώσεις με μέτρια σοβαρότητα, οι οποίες παρατηρούνται στο 14% των περιπτώσεων, ο κορεσμός οξυγόνου στο αίμα είναι ≤93%, υπάρχει διήθηση των πνευμόνων ≥50% εντός μίας ή δύο ημερών, και η αναλογία μερικής πίεσης οξυγόνου προς κλάσμα εισπνεόμενου οξυγόνου (PaO2/FiO2) είναι <300. Οι "αδιαφανείς περιοχές στο έδαφος υάλου" (ground-glass opacities - GGO) είναι επίσης χαρακτηριστικό των σοβαρών συμπτωμάτων. Σε περιπτώσεις όπου η νόσος είναι κρίσιμη, κάτι που συμβαίνει στο 5% των ασθενών, παρατηρείται αποτυχία του αναπνευστικού συστήματος, σηπτικό σοκ και πολυοργανική ανεπάρκεια.

Μέτρηση	**Φυσιολογική Τιμή**	**Πηγή**
Κορεσμός Οξυγόνου στο αίμα	95-100%	Harrison's Principles of Internal Medicine
Αναλογία PaO2/FiO2	300-500	Oxford Handbook of Critical Care
Διήθηση των πνευμόνων	<10%	ARDS diagnosis protocols
Αδιαφανείς περιοχές στο έδαφος υάλου(GGO)	Καμία ή ελάχιστες	Felson's Principles of Chest Roentgenology

Πινακας 1.2: Φυσιολογικές τιμές μετρήσεων παθολογικών στοιχείων

Τα συνήθη εργαστηριακά ευρήματα περιλαμβάνουν λεμφοπενία και λευκοπενία, που παρατηρήθηκαν σε μεγάλο αριθμό περιπτώσεων. Η αξονική τομογραφία θώρακα δείχνει κλινικά χαρακτηριστικά όπως πνευμονία και άλλες ανωμαλίες, όπως RNAemia, οξεία καρδιακή βλάβη, και εμφάνιση GGO, που μπορούν να οδηγήσουν σε θάνατο. Σε πολλές περιπτώσεις, παρατηρήθηκαν πολλαπλές περιφερειακές GGO και στους δύο πνεύμονες, που πιθανότατα προκαλούν τη συστημική και τοπική ανοσολογική αντίδραση, προκαλώντας σοβαρή φλεγμονή.

Ατυπική αξονική τομογραφία δείχνει μονούς, πολλαπλούς συμπαγείς και συμπαγείς όζους στο μέσο λοβό του δεξιού ή αριστερού ή και των δύο πνευμόνων, που περιβάλλονται από GGO σε έναν αριθμό ασθενών με COVID-19. [1]

1.1.5 Διαγνωστικές Μέθοδοι για τον COVID-19

Η διάγνωση του COVID-19 είναι κρίσιμη για την έγκαιρη έναρξη της κατάλληλης θεραπείας και τον περιορισμό της εξάπλωσής του. Δεδομένης της μεγάλης ποικιλίας των κλινικών συμπτωμάτων που παρουσιάζουν οι ασθενείς με COVID-19 (π.χ. βήχας, πυρετός και δύσπνοια) τα οποία είναι παρόμοια με αυτά της γρίπης ή άλλων αναπνευστικών λοιμώξεων, οδηγεί σε αυξημένη δυσκολία διάγνωσης. Για την επιβεβαίωση της λοίμωξης χρησιμοποιούνται διάφορες διαγνωστικές μέθοδοι, όπως οι μοριακές μέθοδοι ανίχνευσης νουκλεϊκών οξέων μέσω Real-Time Quantitative Polymerase Chain Reaction (RT-qPCR) και τεχνολογία CRISPR, οι μέθοδοι ανίχνευσης αντισωμάτων όπως η Enzyme-Linked Immunosorbent Assay (ELISA) και η ταχεία ανίχνευση αντισωμάτων IgM/IgG, καθώς και οι ταχείες διαγνωστικές δοκιμές (Rapid Diagnostic Tests - RDT).

Οι μοριακές μέθοδοι, όπως η ανίχνευση νουκλεϊκών οξέων μέσω της μεθόδου RT-qPCR αποτελούν τις πρώτες γραμμές για την επιβεβαίωση των ύποπτων περιπτώσεων. Η RT-qPCR είναι η πιο κοινή και άμεση μέθοδος για την ανίχνευση του SARS-CoV-2 λόγω της υψηλής ειδικότητας (specificity) και ευαισθησίας (sensitivity)της. Η υψηλή ειδικότητα σημαίνει ότι η μέθοδος μπορεί να αναγνωρίζει σωστά τα άτομα που δεν έχουν την ασθένεια, ενώ η υψηλή ευαισθησία σημαίνει ότι μπορεί να αναγνωρίζει σωστά τα άτομα που έχουν την ασθένεια. Λεπτομερής επεξήγηση των μετρικών αυτών αλλά και άλλων που χρησιμοποιούνται στην παρούσα εργασία θα γίνει στο κεφάλαιο 2.2.7. Ωστόσο, αυτή η μέθοδος είναι δαπανηρή και χρονοβόρα και μπορεί να εμφανίσει ψευδώς αρνητικά αποτελέσματα, ειδικά στα αρχικά στάδια της λοίμωξης ή λόγω αστοχίας στη συλλογή των δειγμάτων.

Η τεχνολογία CRISPR έχει αναπτυχθεί πρόσφατα ως εργαλείο για τη διάγνωση του COVID-19. Προσφέρει υψηλή ευαισθησία και ειδικότητα χωρίς την ανάγκη για εκτεταμένο εξοπλισμό, καθιστώντας την μια πολλά υποσχόμενη μέθοδο. Παρά τα πλεονεκτήματα αυτά, η χρήση δειγμάτων από ασθενείς μπορεί να φέρει κάποιους βιολογικούς κινδύνους ασφαλείας. Αυτοί οι κίνδυνοι περιλαμβάνουν την πιθανότητα μετάδοσης του ιού στους εργαζόμενους στον τομέα της υγείας κατά τη συλλογή και επεξεργασία των δειγμάτων. Επιπλέον, υπάρχει ο κίνδυνος μόλυνσης των εργαστηριακών χώρων και του

εξοπλισμού, κάτι που θα μπορούσε να οδηγήσει σε λανθασμένα αποτελέσματα και περαιτέρω διασπορά του ιού. Ωστόσο, αυτοί οι βιολογικοί κίνδυνοι δεν περιορίζονται αποκλειστικά στην τεχνολογία CRISPR αλλά είναι παρόντες σε όλες τις μεθόδους που περιλαμβάνουν τη χρήση μολυσματικών δειγμάτων, όπως η RT-qPCR και οι μέθοδοι ανίχνευσης αντισωμάτων.

Οι μέθοδοι ανίχνευσης αντισωμάτων, όπως η ELISA και η ταχεία ανίχνευση αντισωμάτων IgM/IgG, είναι γρήγορες, εύκολες στη χρήση και απαιτούν μικρό δείγμα. Αυτές οι μέθοδοι είναι ιδιαίτερα χρήσιμες για την αναγνώριση ασυμπτωματικών λοιμώξεων και την επιβεβαίωση των ύποπτων περιπτώσεων με αρνητικά αποτελέσματα RT-qPCR. Ωστόσο, δεν μπορούν να ανιχνεύσουν την παρουσία του ιού στα αρχικά στάδια της νόσου και μπορεί να εμφανίσουν διασταυρούμενες αντιδράσεις με άλλες λοιμώξεις.

Οι RDTs που ανιχνεύουν την παρουσία ιικών αντιγόνων σε δείγματα του αναπνευστικού συστήματος είναι απλές και παρέχουν αποτελέσματα μέσα σε περίπου 30 λεπτά. Αυτές οι δοκιμές είναι καλύτερες για την αναγνώριση οξείας ή πρώιμης λοίμωξης, αλλά είναι επιρρεπείς σε ψευδώς αρνητικά αποτελέσματα.

Συμπερασματικά, η διάγνωση του COVID-19 απαιτεί τη χρήση πολλαπλών διαγνωστικών μεθόδων για την επιβεβαίωση των κρουσμάτων και την παροχή της κατάλληλης θεραπείας. Παρά τα οφέλη των διαφόρων μεθόδων, κάθε μία έχει τις δικές της προκλήσεις και περιορισμούς που πρέπει να ληφθούν υπόψη για την αποτελεσματική διαχείριση της πανδημίας.[3]

2. Σχετικές γνώσεις

2.1 Στοιχεία ήχου και ψηφιακής επεξεργασίας σήματος

2.1.1 Ορισμός του Ήχου:

Το φαινόμενο του ήχου περιγράφεται από διαφορές πίεσης που μεταδίδονται μέσω ενός μέσου, όπως ο αέρας, το νερό ή τα στερεά. Οι παλμικές δονήσεις που παράγονται από ένα σώμα σε κίνηση ή κραδασμό προκαλούν αυτές τις διαφορές πίεσης. Τα ηχητικά κύματα, που ξεκινούν από τους παλμούς του ηχογόνου σώματος και μεταδίδονται σφαιρικά προς όλες τις κατευθύνσεις, φτάνουν και προσκρούουν στη μεμβράνη του ακουστικού τυμπάνου και θέτουν σε ενέργεια τον μηχανισμό της ακοής. Με άλλα λόγια, ο ήχος θα μπορούσε να περιγραφεί από ταχύτατες μεταβολές της πίεσης του ατμοσφαιρικού αέρα. Αυτές οι μεταβολές πίεσης, διαδίδονται με τη μορφή ηχητικών κυμάτων τα οποία ενεργοποιούν ή καλύτερα διεγείρουν το αισθητήριο όργανο της ακοής, δηλαδή το ανθρώπινο αυτί.

Η ταχύτητα μετάδοσης του ήχου διαφέρει ανάλογα με το είδος και τις ιδιότητες του μέσου. Η ταχύτητα του ήχου επηρεάζεται από τη θερμοκρασία, την πίεση και την υγρασία της ατμόσφαιρας για ηχητικά κύματα που . Για παράδειγμα, σε μια ατμόσφαιρα με μέση πίεση και θερμοκρασία περίπου 15°C, η ταχύτητα του ήχου είναι περίπου 340.45 m/sec.

Για να γίνουν κατανοητά τα ηχητικά κύματα, είναι σημαντικό να εξεταστούν και άλλες έννοιες όπως η συχνότητα, το μήκος κύματος, το πλάτος και η φάση.

2.1.2 Βασικά Χαρακτηριστικά του Ήχου και των Ηχητικών Κυμάτων

Συχνότητα (Frequency), Ύψος (Pitch), Τόνος(Tone):

Η συχνότητα αναφέρεται στον αριθμό των επαναλαμβανόμενων μοτίβων μιας περιοδικής κυματομορφής που παρατηρούνται μέσα σε μια συγκεκριμένη χρονική περίοδο. Συνήθως συμβολίζεται με το μικρό λατινικό γράμμα f και η μονάδα μέτρησής της είναι τα Hertz (Hz) ή κύκλοι ανά δευτερόλεπτο (cycles per second - cps). Έτσι, όταν περιγράφεται μια περιοδική κίνηση συχνότητας 1 Hz, περιγράφετα κίνηση που επαναλαμβάνεται μία φορά ανά δευτερόλεπτο.

Το ύψος (pitch) είναι η υποκειμενική αντίληψη της συχνότητας από το ανθρώπινο αυτί. Με άλλα λόγια, το ύψος είναι ο τρόπος που αντιλαμβανόμαστε έναν ήχο ως "ψηλό" ή "χαμηλό". Όταν η συχνότητα αυξάνεται, το ηχητικό κύμα γίνεται αντιληπτό ως υψηλότερο, ενώ όταν η συχνότητα μειώνεται ως χαμηλότερο. Για παράδειγμα, ήχοι με υψηλές συχνότητες, όπως 1.000 Hz, ακούγονται "πιο ψηλοί" σε σύγκριση με ήχους με χαμηλότερες συχνότητες, όπως 500 Hz.

Ο τόνος (tone) είναι η συνολική ποιότητα του ήχου που σχετίζεται τόσο με τη συχνότητα όσο και με το ύψος (pitch). Ένας τόνος μπορεί να είναι ένας απλός ήχος, γνωστός ως απλός τόνος, ο οποίος έχει μία μόνο συχνότητα, ή μπορεί να είναι σύνθετος, γνωστός ως σύνθετος τόνος, που περιλαμβάνει πολλές συχνότητες. Στην περίπτωση των σύνθετων τόνων, οι πολλαπλές συχνότητες περιλαμβάνουν τη θεμελιώδη συχνότητα και τις αρμονικές. Η θεμελιώδης συχνότητα είναι η χαμηλότερη συχνότητα στον τόνο και καθορίζει την αντιληπτή "βάση" ή "κεντρικό" ύψος. Οι αρμονικές είναι ακέραια πολλαπλάσια της θεμελιώδους συχνότητας και προσδίδουν στον τόνο την ιδιαίτερη χροιά και ποιότητά του, συμβάλλοντας στην αναγνώριση του ηχητικού χαρακτήρα και του βάθους του ήχου. Όσο υψηλότερη είναι η θεμελιώδης συχνότητα, τόσο υψηλότερο είναι το ύψος του ήχου που αντιλαμβανόμαστε. Οι αρμονικές, από την άλλη πλευρά, προσδίδουν στον τόνο την ιδιαίτερη χροιά και ποιότητά του, αλλά δεν επηρεάζουν άμεσα το ύψος όπως το καθορίζει η θεμελιώδης συχνότητα. Έτσι, το ύψος (pitch) ενός σύνθετου τόνου εξαρτάται κυρίως από τη θεμελιώδη συχνότητα, ενώ ο τόνος ως σύνολο περιλαμβάνει και τις αρμονικές που προσδίδουν τον μοναδικό χαρακτήρα στον ήχο.

Συνολικά, η συχνότητα καθορίζει το ύψος του ήχου, το οποίο είναι η αντίληψη της συχνότητας από το ανθρώπινο αυτί. Ο τόνος, ως ευρύτερη έννοια, περιλαμβάνει την ποιότητα του ήχου που συνδέεται τόσο με τη συχνότητα όσο και με το ύψος, καθιστώντας την ακουστική εμπειρία πολύπλοκη και πολυδιάστατη.[4]

Μήκος (Wavelength), Περίοδος (Period), Ταχύτητα Κύματος:

Το μήκος κύματος και η περίοδος είναι έννοιες που μπορούν να θεωρηθούν παρόμοιες λόγω της εξίσωσης 2.1. Σε μια απλή κυματομορφή, το μήκος κύματος είναι η απόσταση ή το μήκος από ένα σημείο της κυματομορφής έως το αντίστοιχο σημείο στην επόμενη επανάληψη του μοτίβου. Συμβολίζεται με το ελληνικό γράμμα λ (λάμδα). Αντίθετα, η περίοδος είναι η χρονική διάρκεια που απαιτείται για να ολοκληρωθεί ένας πλήρης κύκλος της κυματομορφής και συμβολίζεται με το κεφαλαίο γράμμα Τ.

Η ταχύτητα του κύματος μπορεί να προσδιοριστεί από αυτές τις δύο έννοιες. Το κύμα κινείται κατά ένα μήκος κύματος λ σε μία περίοδο Τ, και έτσι η ταχύτητά v ορίζεται ως:

$v = \lambda/T$	(2.1)

Αυτή η σχέση δείχνει πώς το μήκος κύματος και η περίοδος συνδέονται με την ταχύτητα διάδοσης του κύματος, κάνοντας εμφανές ότι η ταχύτητα εξαρτάται άμεσα από το μήκος κύματος και την περίοδο του κύματος.

Η περίοδος είναι αντιστρόφως ανάλογη με τη συχνότητα, δηλαδή καθώς η συχνότητα αυξάνεται, η περίοδος μειώνεται και το αντίστροφο όπως παρουσιάζεται στην εξίσωση 2.2.

$T=1/f$ ή $f=1/T$	(2.2)

Πλάτος (Amplitude), Decibel, Ηχητική Πίεση (Sound Pressure), Ηχητική Ισχύς (Sound Power), Ηχητική Ένταση (Sound Intensity):

Το πλάτος είναι η μετατόπιση των σωματιδίων του μέσου από την κατάσταση ηρεμίας τους. Σε γραφική αναπαράσταση της κυματομορφής, το πλάτος αντιπροσωπεύεται από το ύψος του κύματος. Η μονάδα μέτρησης του πλάτους είναι το μέτρο (m).

Τα decibels (dB) είναι μια λογαριθμική μονάδα μέτρησης που χρησιμοποιείται για να εκφράσει αναλογίες μεταξύ δύο τιμών. Σε αντίθεση με τις απόλυτες μονάδες, όπως το μέτρο ή το κιλό, τα dB δεν μετρούν απόλυτες τιμές, αλλά εκφράζουν τη σχέση μεταξύ μιας μετρούμενης ποσότητας και μιας τιμής αναφοράς. Αυτό βοηθά στην εύκολη κατανόηση πολύ μεγάλων ή πολύ μικρών αριθμών, καθιστώντας τα dB χρήσιμα στην ακουστική για για να περιγράφουν πόσο δυνατός ή απαλός είναι ένας ήχος ή θόρυβος.

Οι τρεις όροι ηχητική πίεση, ηχητική ισχύς και ηχητική ένταση μετρούν διαφορετικές πτυχές του ήχου, αλλά όλοι μπορούν να εκφραστούν σε dB, όπως φαίνεται στην εικόνα 2.1:

Measurement	Measurement Unit	Commonly Reported In
Sound Pressure	Pa (Pascal)	dB (ref = 20×10^{-6} Pa)
Sound Power	W (Watts)	dB (ref = 1×10^{-12} W)
Sound Intensity	W/m^2 (Watts/Area)	dB (ref = 1×10^{-12} W/m^2)

Εικόνα 2.1: Τρεις φυσικές έννοιες που περιγράφουν τον ήχο, οι μονάδες μέτρησής τους και το κατώφλι ακουστότητας για τον ορισμό της κλίμακας

Μπορεί να χρησιμοποιηθεί η αναλογία ενός θερμαντικού σώματος σε ένα κρύο δωμάτιο με ένα αντικείμενο που εκπέμπει ήχο σε ένα ήσυχο δωμάτιο για να εξηγηθούν οι διαφορές μεταξύ ηχητικής πίεσης, ισχύος και έντασης. Όπως ο θερμαντήρας διαχέει θερμότητα σε όλο το δωμάτιο, έτσι και ένα αντικείμενο που εκπέμπει ήχο διαχέει ηχητικά κύματα. Σε κάθε σημείο ενός δωματίου, υπάρχει ένα συγκεκριμένο επίπεδο θερμοκρασίας, το οποίο μετριέται σε βαθμούς Κελσίου. Ομοίως, σε κάθε σημείο ενός δωματίου με πηγή ήχου, υπάρχει ένα συγκεκριμένο επίπεδο ηχητικής πίεσης, το οποίο μετριέται

σε Pascals. Όσο πιο κοντά βρισκόμαστε στη θερμαντική πηγή, τόσο υψηλότερη είναι η θερμοκρασία. Παρομοίως, όσο πιο κοντά βρισκόμαστε στην πηγή του ήχου, τόσο υψηλότερο είναι το επίπεδο της ηχητικής πίεσης. Και τα δύο, η θερμοκρασία και η ηχητική πίεση, εξαρτώνται από τη θέση και την απόσταση από την πηγή.

Ο θερμαντήρας παράγει μια συγκεκριμένη ποσότητα θερμότητας ανά ώρα. Η απαιτούμενη ισχύς για την παραγωγή αυτής της θερμότητας παραμένει σταθερή, ανεξάρτητα από τη θερμοκρασία στο δωμάτιο. Η ισχύς του θερμαντήρα μετριέται σε ενέργεια ανά μονάδα χρόνου, δηλαδή σε Watt. Ομοίως, η ηχητική ισχύς ενός αντικειμένου είναι ένα χαρακτηριστικό του αντικειμένου και είναι ανεξάρτητη από τα επίπεδα ηχητικής πίεσης στο δωμάτιο. Η ηχητική ισχύς είναι ο ρυθμός με τον οποίο εκπέμπεται η ηχητική ενέργεια ανά μονάδα χρόνου και μετριέται επίσης σε Watt.

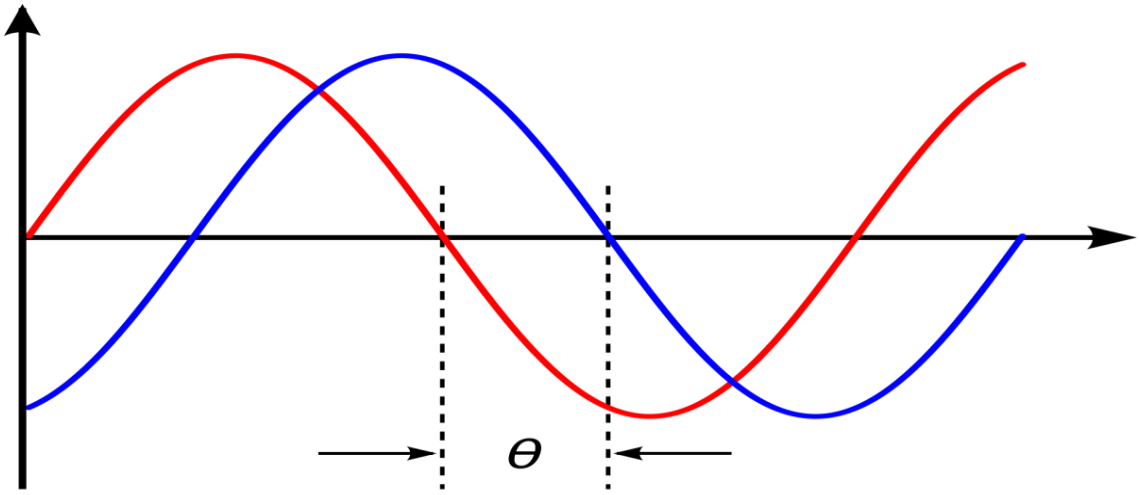
Η θερμότητα διαχέεται και ρέει σε όλο το δωμάτιο. Η ροή της θερμότητας έχει ένα επίπεδο θερμοκρασίας και μια κατεύθυνση. Η ηχητική ένταση μετρά τη ροή του ήχου και επίσης έχει επίπεδο και κατεύθυνση. Αυτή η ροή παρατηρείται σε μια συγκεκριμένη επιφάνεια, γι' αυτό οι μονάδες μέτρησης της ηχητικής έντασης είναι Watt ανά τετραγωνικό μέτρο (W/m^2).

Συνοψίζοντας, η ηχητική πίεση μετρά το επίπεδο έντασης του ήχου σε μια συγκεκριμένη θέση και εξαρτάται από την απόσταση και την τοποθεσία σε σχέση με την πηγή του ήχου. Η ηχητική ισχύς αναφέρεται στον ρυθμό εκπομπής ήχου από ένα αντικείμενο, ανεξάρτητα από την απόσταση ή τη θέση παρατήρησης, και μετριέται σε Watt. Τέλος, η ηχητική ένταση είναι η ηχητική ισχύς ανά μονάδα επιφάνειας, δείχνοντας τη ροή του ήχου μέσα από μια συγκεκριμένη περιοχή και μετριέται σε Watt ανά τετραγωνικό μέτρο (W/m^2).

Φάση (Phase):

Η φάση ενός κύματος αναφέρεται στην ακριβή θέση ή κατάσταση ενός σημείου του κύματος σε σχέση με ένα σταθερό σημείο αναφοράς. Για παράδειγμα, σε ένα περιοδικό κύμα, όπως ένα αρμονικό κύμα που ταξιδεύει κατά μήκος μιας χορδής, η φάση καθορίζει τη θέση ενός συγκεκριμένου σημείου του κύματος.

Αυτή η θέση μπορεί να εκφραστεί είτε σε μοίρες είτε σε ακτίνια. Ένας πλήρης κύκλος, που περιλαμβάνει μια άνοδο και μια κάθοδο, αντιστοιχεί σε 360° ή 2π ακτίνια. Έτσι, μια φάση 90° (ή π/2 ακτίνια) υποδεικνύει ότι το σημείο βρίσκεται στο ένα τέταρτο της διαδρομής του κύκλου, συχνά στην ανιούσα φάση προς την κορυφή του κύματος.



Εικόνα 2.2: Απεικόνιση διαφοράς φάσης δύο κυμάτων ίδιας συχνότητας συναρτήσει του χρόνου. Η διαφορά φάσης ορίζεται ως η διαφορά χρόνου μεταξύ δύο ακρότατων σημείων του κύματος

Η εικόνα 2.2 παρουσιάζει δύο ημιτονοειδή κύματα με διαφορά φάσης. Τα κύματα, κόκκινο και μπλε, έχουν το ίδιο πλάτος και συχνότητα, αλλά διαφέρουν ως προς τη φάση τους, δηλαδή εμφανίζουν χρονική μετατόπιση το ένα σε σχέση με το άλλο. Ο οριζόντιος άξονας αντιπροσωπεύει τον χρόνο, ενώ ο κατακόρυφος το πλάτος των κυμάτων.

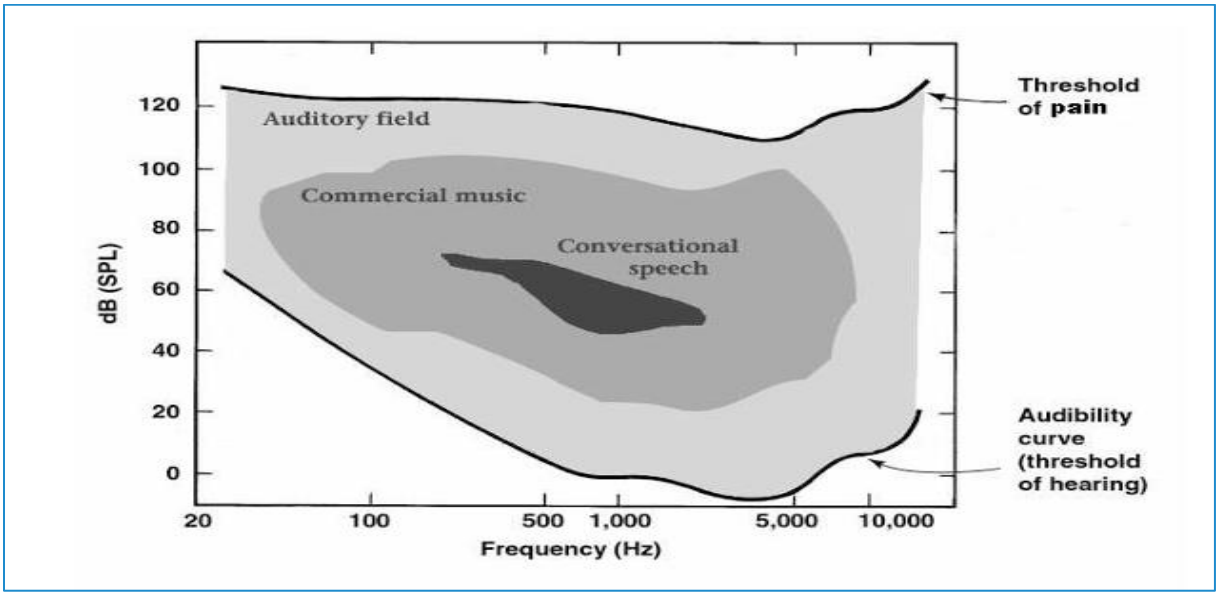
Η διαφορά φάσης μεταξύ των δύο κυμάτων φαίνεται από την απόσταση κατά μήκος του χρονικού άξονα, όπου τα κύματα φτάνουν τις ίδιες θέσεις (πλάτος) σε διαφορετικές χρονικές στιγμές. Συγκεκριμένα, η διαφορά φάσης ορίζεται ως η διαφορά των χρονικών στιγμών κατά τις οποίες τα δύο κύματα παρουσιάζουν το ίδιο πλάτος, όπως τα μέγιστα, τα ελάχιστα ή οποιαδήποτε άλλη χρονική στιγμή. Αυτή η χρονική μετατόπιση καθορίζει πώς τα κύματα του ίδιου πλάτους και συχνότητας διαφέρουν στην χρονική στιγμή που επιτυγχάνουν την ίδια τιμή πλάτους. Αυτό το παράδειγμα της διαφοράς φάσης βοηθά στην κατανόηση του πώς η φάση περιγράφει την χρονική θέση δύο κυμάτων με ίδια χαρακτηριστικά, αλλά με διαφορετική χρονική στιγμή κατά την οποία επιτυγχάνουν το ίδιο πλάτος.[5]

2.1.3 Αντίληψη του Ήχου από το Ανθρώπινο Ακουστικό Σύστημα

Αφού έγινε περιγραφή του ήχου από φυσική άποψη, είναι σημαντικό να γίνει κατανοητό το πώς ο άνθρωπος τον αντιλαμβάνεται. Το εξωτερικό αυτί και ο ακουστικός πόρος συλλέγουν τα φθίνοντα κύματα πίεσης του αέρα, τα οποία μεταδίδονται ως μηχανικές δονήσεις μέσω του τύμπανου του αυτιού και του εσωτερικού αυτιού στον κοχλία που είναι γεμάτος με υγρό. Ο κοχλίας μοιάζει με μια κωνική

σπείρα, σπειροειδές σαλιγκάρι (η ονομασία προέρχεται από την αρχαία ελληνική λέξη για το "κέλυφος σαλιγκαριού"). Η εσωτερική επιφάνεια του κοχλία επενδύεται από αισθητήρια κύτταρα που αποτελούνται από μικρές τρίχες που συνδέονται με νευρικές απολήξεις, τα λεγόμενα τριχωτά κύτταρα. Ανάλογα με τη συχνότητα του εισερχόμενου ήχου, διαφορετικές περιοχές του κοχλία συντονίζονται, προκαλώντας τη δόνηση των τριχωτών κυττάρων στην αντίστοιχη περιοχή (τονοτοπική οργάνωση). Με τη σειρά τους, οι διεγερμένες νευρικές απολήξεις στέλνουν ερεθίσματα στον εγκέφαλο, επιτρέποντάς του να αντιλαμβάνεται και να διακρίνει ήχους διαφορετικής συχνότητας, στάθμης και χροιάς.[6]

Το ανθρώπινο ακουστικό σύστημα ανταποκρίνεται σε ήχους στο εύρος συχνοτήτων από 20 Hz έως 20 kHz, εφόσον η ένταση βρίσκεται πάνω από την εξαρτώμενη συχνότητα, το λεγόμενο “κατώφλι ακοής”. Το εύρος της ακουστής έντασης είναι περίπου 120 dB, το οποίο αντιπροσωπεύει το εύρος μεταξύ του ψιθύρου των φύλλων και του βόμβου της απογείωσης ενός αεροσκάφους. Το Εικόνα 2.3 εμφανίζει το ανθρώπινο ακουστικό πεδίο στο επίπεδο συχνότητας-έντασης.



Εικόνα 2.3: Σχήμα που παρουσιάζει ακουστικά φαινόμενα (ομιλία, μουσική) και το ανθρώπινο φάσμα ακοής. Σημειώνονται και τα αντίστοιχα όρια ακουστότητας (όριο πάνω από το οποίο γίνεται αντιληπτός ήχος) και το όριο του πόνου (όριο πάνω από το οποίο προκαλείται η αίσθηση του πόνου).

Το Σχήμα 2.3 παρουσιάζει τη δυνατότητα του ανθρώπινου ακουστικού συστήματος να ανιχνεύει ήχους, διευκρινίζοντας ένα φάσμα συχνοτήτων από 20 Hz έως 20 kHz και ένα φάσμα έντασης έως 120

dB. Οπτικοποιεί το κατώφλι της ακοής, το οποίο περιγράφει την ελάχιστη ένταση του ήχου που απαιτείται για αντίληψη σε διαφορετικές συχνότητες. Αυτή η αναπαράσταση τονίζει τη μη γραμμική ανταπόκριση του ακουστικού συστήματος στην ένταση και τη συχνότητα του ήχου, υπογραμμίζοντας περιοχές αυξημένης ευαισθησίας, ιδιαίτερα στο εύρος συχνοτήτων κρίσιμο για την κατανόηση της ανθρώπινης ομιλίας. Λειτουργεί ως θεμελιώδες εργαλείο για την κατανόηση της ανθρώπινης ακοής, τον σχεδιασμό ηχητικού εξοπλισμού και τη βελτιστοποίηση του ήχου για διάφορες εφαρμογές, απεικονίζοντας τα όρια του ακουστικού πεδίου και την ευαισθησία του αυτιού σε όλο το ακουστικό φάσμα. [7]

2.1.4 Δομή και Χαρακτηριστικά του Σήματος Ομιλίας:

Αφού έγινε περιγραφή του ήχου από φυσικής άποψης και της αντίληψης του από τον άνθρωπο, θα γίνει εστίαση σε ένα συγκεκριμένο είδος ήχου: την ομιλία. Ένα συνεχές σήμα ομιλίας αποτελείται από δύο κύρια μέρη: το ένα μεταφέρει τις πληροφορίες της ομιλίας και το άλλο περιλαμβάνει τα τμήματα σιγής ή θορύβου που υπάρχουν μεταξύ των εκφωνήσεων, χωρίς καμία λεκτική πληροφορία.

Το λεκτικό μέρος της ομιλίας μπορεί να χωριστεί περαιτέρω σε δύο κύριους τύπους: έμφωνη και άφωνη ομιλία. Όταν οι άνθρωποι μιλούν, ο αέρας περνάει μέσω του λάρυγγα. Οι έμφωνοι ήχοι παράγονται όταν οι φωνητικές χορδές που πάλλονται, ενώ όταν είναι ανοιχτές και χωρίς να πάλλονται, παράγεται άφωνη ομιλία.

Η έμφωνη ομιλία είναι ένα σχετικά αργά μεταβαλλόμενο περιοδικό σήμα με θεμελιώδη συχνότητα την βασική συχνότητα δόνησης των φωνητικών χορδών, που ονομάζεται τόνος. Για τους άνδρες, η συχνότητα του τόνου συνήθως κυμαίνεται μεταξύ 50 Hz και 250 Hz, ενώ για τις γυναίκες κυμαίνεται μεταξύ 120 Hz και 500 Hz.

Οι άφωνοι ήχοι ομιλίας παράγονται όταν ο αέρας περνά απευθείας μέσω των σχηματισμών του φωνητικού σωλήνα. Σε αντίθεση με τη φωνητική ομιλία, η άφωνη ομιλία δεν παρουσιάζει περιοδικότητα και χαρακτηρίζεται από ένα θορυβώδες σήμα.

Η διαδικασία παραγωγής ομιλίας περιλαμβάνει την παραγωγή εκφωνήσεων, που είναι ομάδες διαδοχικών φωνητικών και/ή άφωνων ήχων. Αυτοί οι ήχοι συνήθως συνδέονται πολύ ομαλά μεταξύ τους.

Οι εκφωνήσεις διαχωρίζονται από περιοχές σιγής. Κατά τη διάρκεια αυτών των περιοχών σιγής, δεν παρέχεται διέγερση στον φωνητικό σωλήνα και, επομένως, δεν υπάρχει ομιλία. Ωστόσο, η σιγή θεωρείται αναπόσπαστο μέρος του σήματος ομιλίας. Τα ακουστικά σήματα κατά τη διάρκεια των περιοχών σιγής είναι κυρίως θόρυβος παρασκηνίου και άσχετοι με την ομιλία ήχοι του στόματος.[8]

2.1.5 Φωνητική και Φωνολογία (Phonemics and Phonetics)

Η παραγωγή της ομιλίας αρχίζει στο μυαλό του ανθρώπου όταν σχηματίζει μια σκέψη που θέλει να μεταδώσει στον ακροατή. Αφού σχηματιστεί η επιθυμητή σκέψη, ο ομιλητής δημιουργεί μια φράση ή πρόταση επιλέγοντας μια συλλογή αμοιβαίων αποκλειστικών ήχων. Η βασική θεωρητική μονάδα για την περιγραφή της μετάδοσης της γλωσσικής σημασίας στην ομιλία ονομάζεται φώνημα.

Τα φωνήματα μπορούν να θεωρηθούν ως ένας τρόπος αναπαράστασης των διαφόρων μερών ενός ηχητικού κύματος ομιλίας, που παράγονται μέσω του ανθρώπινου φωνητικού μηχανισμού και χωρίζονται σε συνεχείς ή μη συνεχείς ήχους. Ένα φώνημα είναι συνεχές αν ο ήχος παράγεται όταν ο φωνητικός αγωγός βρίσκεται σε σταθερή κατάσταση. Αντίθετα, το φώνημα είναι μη συνεχές όταν ο φωνητικός αγωγός αλλάζει τα χαρακτηριστικά του κατά την παραγωγή ομιλίας. Για παράδειγμα, αν η περιοχή στον φωνητικό αγωγό αλλάζει με το άνοιγμα και κλείσιμο του στόματος ή την κίνηση της γλώσσας σε διαφορετικές καταστάσεις, το φώνημα που περιγράφει την παραγόμενη ομιλία είναι μη συνεχές.

Τα φωνήματα μπορούν να ομαδοποιηθούν με βάση τις ιδιότητες είτε του χρονικού κύματος είτε των χαρακτηριστικών συχνοτήτων και να ταξινομηθούν σε διαφορετικούς ήχους που παράγονται από τον άνθρωπο. Η παρακάτω ομαδοποίηση περιλαμβάνει τις κύριες κατηγορίες φωνημάτων:

1. Συνεχή (Continuant), τα οποία περιλαμβάνουν Vowels, Fricatives, Affricates, και Nasals, και
2. Μη Συνεχή (Non-Continuant), τα οποία περιλαμβάνουν Diphthongs, Semivowels, και Stops.

Για περισσότερες πληροφορίες για τις παραπάνω κατηγορίες φωνημάτων, ο αναγνώστης μπορεί να ανατρέξει στο ακόλουθο άρθρο.[9]

2.1.7 Ψηφιοποίηση του Ήχου: Από την Αναλογική στην Ψηφιακή Μορφή

Αφού έγινε περιγραφή της φύσης του ήχου και της ομιλίας, θα παρουσιαστεί η διαδικασία ψηφιοποίησης της φωνής, από τη στιγμή που ο ήχος εισέρχεται στο μικρόφωνο, μέχρι τη μετατροπή του σε ψηφιακό σήμα. Η ψηφιοποίηση του ήχου είναι κρίσιμη για την αποθήκευση, επεξεργασία και μετάδοση του ήχου μέσω ψηφιακών συστημάτων.

Η διαδικασία ξεκινά όταν παράγεται οποιοσδήποτε ήχος κοντά σε ένα μικρόφωνο. Το μικρόφωνο είναι ένας μετατροπέας που μετατρέπει τις ηχητικές δονήσεις στον αέρα σε ηλεκτρικά σήματα. Οι ηχητικές δονήσεις, που είναι κύματα πίεσης, προκαλούν τη μετατόπιση της μεμβράνης του μικροφώνου. Η μεμβράνη κινείται μέσα σε ένα μαγνητικό πεδίο όπου έχει προσαρτημένο έναν αγωγό και μέσω του φαινομένου Faraday, παράγεται ρεύμα ανάλογο με τη μετατόπιση της μεμβράνης και, συνεπώς, ανάλογο με το ηχητικό κύμα. Αυτές οι μεταβολές δημιουργούν ένα αναλογικό ηλεκτρικό σήμα.

Το αναλογικό σήμα που παράγεται από το μικρόφωνο είναι συνεχές, δηλαδή έχει άπειρες δυνατές τιμές μέσα σε ένα συγκεκριμένο εύρος και αλλάζει συνεχώς με τον χρόνο. Αυτό το σήμα, παρόλο που είναι μια πιστή αναπαράσταση του πρωτότυπου ήχου, δεν μπορεί να επεξεργαστεί από ψηφιακά συστήματα όπως υπολογιστές. Για να μπορέσει να αποθηκευτεί και να επεξεργαστεί ψηφιακά, το αναλογικό σήμα πρέπει να μετατραπεί σε ψηφιακή μορφή. Η ψηφιακή μορφή είναι μια αναπαράσταση του σήματος σε δυαδική μορφή, δηλαδή ως ακολουθία από bits (0 και 1). Η μετατροπή του αναλογικού σήματος σε ψηφιακή μορφή γίνεται μέσω των διαδικασιών της δειγματοληψίας, του κβαντισμού και της κωδικοποίησης.

Η πρώτη κρίσιμη φάση στη διαδικασία ψηφιοποίησης είναι η δειγματοληψία. Κατά τη δειγματοληψία, το συνεχές αναλογικό σήμα μετράται σε τακτά χρονικά διαστήματα. Αυτές οι μετρήσεις ονομάζονται δείγματα και η συχνότητα με την οποία λαμβάνονται τα δείγματα ονομάζεται συχνότητα δειγματοληψίας. Η συχνότητα δειγματοληψίας μετράται σε Hertz (Hz) και καθορίζει τον αριθμό των δειγμάτων που λαμβάνονται ανά δευτερόλεπτο.

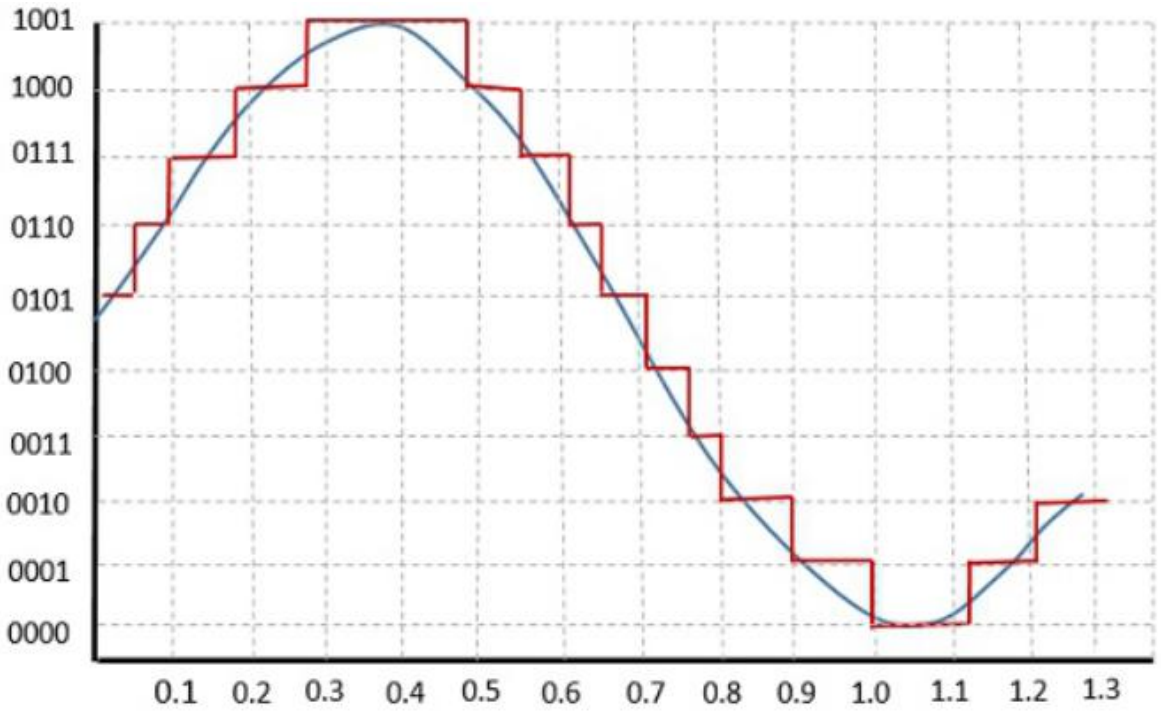
Σύμφωνα με το θεώρημα Nyquist-Shannon, για την πλήρη αναπαράσταση ενός αναλογικού σήματος χωρίς απώλειες πληροφοριών, η συχνότητα δειγματοληψίας πρέπει να είναι υπερδιπλάσια από την ανώτατη συχνότητα του αναλογικού σήματος.

Η αναδίπλωση (aliasing) είναι ένα φαινόμενο που συμβαίνει όταν ένα αναλογικό σήμα δειγματοληπτείται σε συχνότητα χαμηλότερη από την απαιτούμενη σύμφωνα με το θεώρημα Nyquist-Shannon. Όταν το αναλογικό σήμα δειγματοληπτείται σε χαμηλότερη συχνότητα από την απαιτούμενη, οι υψηλές συχνότητες του σήματος "αναδιπλώνονται" ή "επικαλύπτονται". Αυτό σημαίνει ότι οι υψηλές συχνότητες, που δεν μπορούν να αναπαραχθούν σωστά λόγω της ανεπαρκούς δειγματοληψίας, "μεταφράζονται" σε λανθασμένες συχνότητες στο ψηφιακό σήμα, χαμηλότερης συχνότητας. Με άλλα λόγια, το ψηφιακό σήμα περιέχει επιπρόσθετες, ψευδείς συχνότητες που δεν υπήρχαν στο αρχικό αναλογικό σήμα, προκαλώντας παραμόρφωση και απώλεια της ακρίβειας του σήματος.

Αφού ληφθούν τα δείγματα του αναλογικού σήματος, η επόμενη φάση είναι ο κβαντισμός. Ο κβαντισμός είναι η διαδικασία κατά την οποία τα πλάτη των δειγμάτων στρογγυλοποιούνται σε διακριτές τιμές που μπορούν να αναπαρασταθούν ψηφιακά. Κάθε τιμή πλάτους αντιστοιχείται στην πλησιέστερη τιμή από ένα προκαθορισμένο σύνολο διακριτών τιμών, γνωστές ως επίπεδα κβαντισμού.

Ο αριθμός των επιπέδων κβαντισμού καθορίζει την ακρίβεια της ψηφιοποίησης. Περισσότερα επίπεδα κβαντισμού σημαίνουν μεγαλύτερη ακρίβεια και καλύτερη αναπαράσταση του αρχικού σήματος, αλλά απαιτούν περισσότερα bits για την αποθήκευση κάθε δείγματος. Αντίθετα, λιγότερα επίπεδα κβαντισμού μειώνουν την ακρίβεια και μπορούν να εισαγάγουν θόρυβο κβαντισμού, που επηρεάζει την ποιότητα του ψηφιακού σήματος.

Η εικόνα 2.4 δείχνει το πώς γίνεται ο κβαντισμός ενός αναλογικού σήματος. Η μπλε γραμμή αντιπροσωπεύει το αναλογικό σήμα, ενώ η κόκκινη γραμμή αντιπροσωπεύει το κβαντισμένο σήμα.



Εικόνα 2.4: Συνεχές σήμα (μπλέ χρώματος) και το αντίστοιχο δειγματοληπτούμενο και κβαντισμένο σήμα (κόκκινου χρώματος).

Μετά τον κβαντισμό, τα κβαντισμένα δείγματα πρέπει να κωδικοποιηθούν σε δυαδική μορφή (bits) για να μπορούν να αποθηκευτούν και να επεξεργαστούν από ψηφιακά συστήματα. Αυτή η διαδικασία πραγματοποιείται μέσω ενός αναλογικού προς ψηφιακό μετατροπέα (ADC - Analog to Digital Converter). Ο ADC λαμβάνει τις διακριτές τιμές των δειγμάτων και τις μετατρέπει σε ακολουθίες από bits. Για παράδειγμα, ένα δείγμα μπορεί να κωδικοποιηθεί ως μια ακολουθία 8, 16 ή περισσότερων bits, ανάλογα με την απαιτούμενη ακρίβεια και το διαθέσιμο εύρος ζώνης.

Η ψηφιοποίηση του ήχου μέσω μικροφώνων, δειγματοληψίας, κβαντισμού και κωδικοποίησης είναι μια περίπλοκη διαδικασία που μετατρέπει το αναλογικό ηχητικό σήμα σε ψηφιακή μορφή, επιτρέποντας την αποθήκευση, την επεξεργασία και τη μετάδοση του ήχου μέσω ψηφιακών συστημάτων. Αυτή η διαδικασία εξασφαλίζει ότι ο αρχικός ήχος αναπαρίσταται με ακρίβεια και μπορεί να χρησιμοποιηθεί σε μια ποικιλία εφαρμογών, από την αναπαραγωγή μουσικής μέχρι την αναγνώριση ομιλίας.[10]

2.1.8 Βασικά Χαρακτηριστικά για την Επεξεργασία Ομιλίας

Έχοντας αναλύσει τη φύση του ήχου και της ομιλίας, καθώς και τη διαδικασία ψηφιοποίησής τους, είναι πλέον κρίσιμο να εξεταστούν τα βασικά χαρακτηριστικά που απαιτούνται για την αποτελεσματική επεξεργασία της ομιλίας. Τα χαρακτηριστικά ομιλίας είναι αριθμητικές αναπαραστάσεις των σημάτων ομιλίας που χρησιμοποιούνται για ανάλυση, αναγνώριση και σύνθεση. Γενικά, τα χαρακτηριστικά των σημάτων ομιλίας μπορούν να ταξινομηθούν σε δύο κατηγορίες: χαρακτηριστικά χρονικού πεδίου και χαρακτηριστικά συχνοτικού πεδίου.

Χαρακτηριστικά Χρονικού Πεδίου

Τα χαρακτηριστικά χρονικού πεδίου προκύπτουν άμεσα από το πλάτος του σήματος ομιλίας στον χρόνο. Είναι απλά στον υπολογισμό και συχνά χρησιμοποιούνται σε εφαρμογές επεξεργασίας ομιλίας σε πραγματικό χρόνο. Ορισμένα κοινά χαρακτηριστικά χρονικού πεδίου περιλαμβάνουν:

1. Ενέργεια: Η ενέργεια είναι ένα ποσοτικό μέτρο των χαρακτηριστικών πλάτους ενός σήματος ομιλίας στον χρόνο. Υπολογίζεται με τον τετραγωνισμό κάθε δείγματος στο σήμα και το άθροισμά τους μέσα σε ένα συγκεκριμένο χρονικό παράθυρο. Αυτή η διαδικασία αποκαλύπτει τη συνολική δυναμική του σήματος, αποτυπώνοντας τις χρονικές μεταβολές στην ένταση.

Η μέτρηση της ενέργειας παρέχει πληροφορίες για τμήματα του σήματος με μεγαλύτερα ή μικρότερα πλάτη, διευκολύνοντας την αναγνώριση ομιλίας, την τμηματοποίηση ήχου και τη διαχείριση ομιλητών. Επιπλέον, βοηθά στον εντοπισμό γεγονότων και μεταβάσεων που υποδηλώνουν αλλαγές στη φωνητική δραστηριότητα. Μέσω της ποσοτικοποίησης των μεταβολών του πλάτους, η ανάλυση της ενέργειας συμβάλλει σε μια ολοκληρωμένη κατανόηση των σημάτων ομιλίας και των ακουστικών τους ιδιοτήτων.

2. Zero Crossing Rate (ZCR): Ο ρυθμός διασχίσεων μηδενός (Zero Crossing Rate - ZCR) αποτελεί ένα σημαντικό χαρακτηριστικό του σήματος ομιλίας στον τομέα της επεξεργασίας ομιλίας. Ο ZCR δείχνει πόσο συχνά το σήμα ομιλίας διασχίζει τον άξονα του μηδενός μέσα σε ένα καθορισμένο χρονικό πλαίσιο. Η τιμή του ZCR υπολογίζεται μετρώντας τον αριθμό των αλλαγών πολικότητας του σήματος κατά τη διάρκεια ενός συγκεκριμένου παραθύρου.

Ο ZCR είναι μια σημαντική παράμετρος καθώς μπορεί να δώσει πληροφορίες για την ένταση και τη συχνότητα του σήματος. Υψηλός ρυθμός διασχίσεων μηδενός συχνά υποδεικνύει ένα σήμα με υψηλότερη συχνότητα ή θόρυβο, ενώ χαμηλός ρυθμός διασχίσεων μηδενός συνδέεται συνήθως με πιο χαμηλές συχνότητες ή πιο σταθερά σήματα. Αυτή η παράμετρος είναι χρήσιμη για διάφορες εφαρμογές όπως η αναγνώριση ομιλίας, η ανάλυση μουσικής και ο εντοπισμός φωνής.

3. Pitch: Αναφέρεται στην αντιληπτή τονική ποιότητα της φωνής ενός ομιλητή, η οποία καθορίζεται από την ανάλυση της θεμελιώδους συχνότητας του σήματος ομιλίας.

Είναι ένα από τα χαρακτηριστικά του σήματος στο χρονικό πεδίο που εξάγονται απευθείας από το πλάτος του σήματος ομιλίας με την πάροδο του χρόνου. Αυτό το χαρακτηριστικό είναι ιδιαίτερα

σημαντικό καθώς προσφέρει πληροφορίες σχετικά με τον τόνο και τον ρυθμό της φωνής, που είναι κρίσιμα για διάφορες εφαρμογές επεξεργασίας ομιλίας, όπως η αναγνώριση ομιλίας, η ανάλυση φωνής και η συνθετική ομιλία.

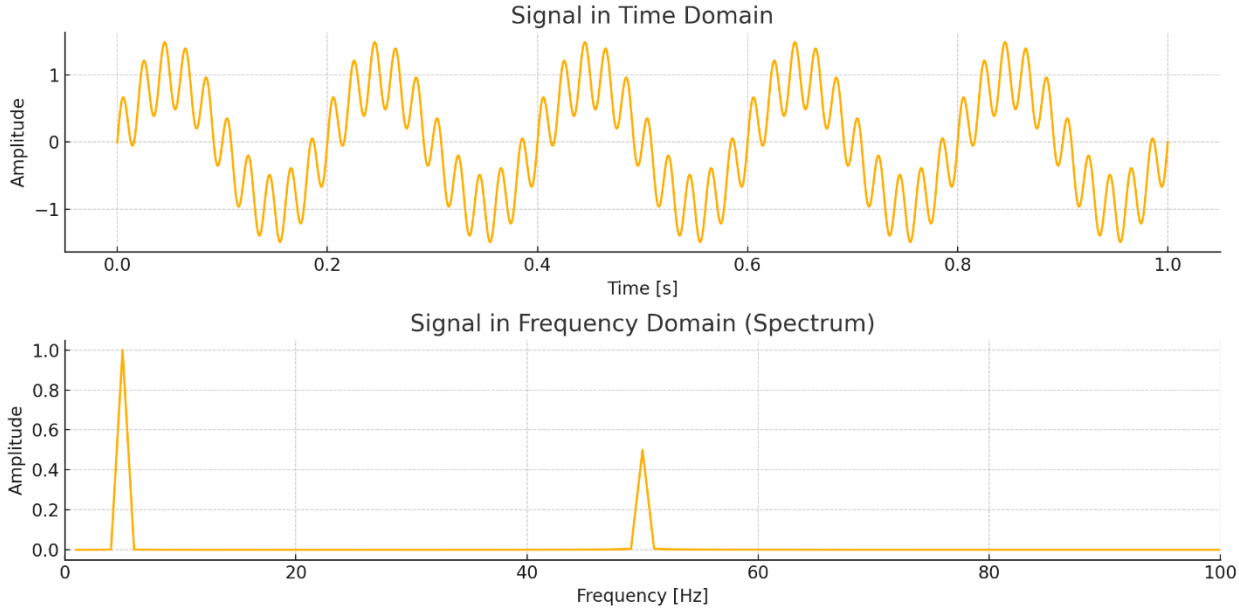
Χαρακτηριστικά Πεδίου Συχνότητας

1. Φάσμα (Spectrum): Ο μετασχηματισμός Fourier, το όνομά του οποίου προήλθε από τον Ζοζέφ Φουριέ, είναι ένας μαθηματικός μετασχηματισμός με πολλές εφαρμογές στη φυσική και την μηχανική. Σύμφωνα με το θεώρημα του Fourier, κάθε περιοδικό σήμα μπορεί να προσεγγιστεί ως άθροισμα ημιτονοειδών κυμάτων, το καθένα με μια συγκεκριμένη συχνότητα και πλάτος. Έτσι, ένα ακουστικό σήμα μπορεί να προσομοιωθεί ως άθροισμα ηχητικών κυμάτων με διαφορετικές συχνότητες και πλάτη. Όταν παίρνουμε δείγματα του σήματος σε σχέση με το χρόνο, καταγράφουμε μόνο τα προκύπτοντα πλάτη. Ο μετασχηματισμός Fourier μετατρέπει μια συνάρτηση που εξαρτάται από τον χρόνο σε μια νέα συνάρτηση που εξαρτάται από τη συχνότητα. Η νέα συνάρτηση είναι τότε γνωστή ως μετασχηματισμός Fourier ή και ως φάσμα συχνοτήτων της συνάρτησης που εξαρτάται από τον χρόνο. Με άλλα λόγια, μετατρέπει το σήμα από το πεδίο του χρόνου στο πεδίο των συχνοτήτων. Το αποτέλεσμα ονομάζεται φάσμα (Spectrum).

Για να κατανοηθεί καλύτερα ο μετασχηματισμός Fourier, ακολουθεί ένα συγκεκριμένο παράδειγμα. Έστω ότι υπάρχει ένα σήμα που αποτελείται από δύο ημιτονοειδή κύματα. Το πρώτο κύμα έχει συχνότητα 5 Hz και πλάτος 1, ενώ το δεύτερο κύμα έχει συχνότητα 50 Hz και πλάτος 0.5. Η συνάρτηση στο χρονικό πεδίο μπορεί να γραφτεί ως:

$f(t) = 1.0 \sin(2\pi 5t) + 0.5\sin(2\pi 50t)$	2

Από αυτήν τη συνάρτηση προκύπτει η γραφική παράσταση στο πεδίο του χρόνου και εφαρμόζοντας τον μετασχηματισμό Fourier προκύπτει η συνάρτηση στο πεδίο της συχνότητας.



Εικόνα 2.5: Ημιτονοειδές σήμα ως άθροισμα δύο επιμέρους ημιτονοειδών συνιστωσών (άνω διάγραμμα) και η ανάλυση του στο πεδίο της συχνότητας (κάτω διάγραμμα)

Από το πεδίο του χρόνου καταγράφονται μόνο τα πλάτη του σήματος σε σχέση με το χρόνο, ενώ από το φάσμα στο πεδίο της συχνότητας διακρίνονται οι συχνότητες που περιέχει το σήμα και οι αντίστοιχες εντάσεις τους. Με αυτόν τον τρόπο, ο μετασχηματισμός Fourier παρέχει μια πολύτιμη μέθοδο για την ανάλυση και κατανόηση των σύνθετων σημάτων, καθώς αποκαλύπτει τη συχνότητα και την ένταση των επιμέρους συχνοτήτων που συνθέτουν το αρχικό σήμα.

Ωστόσο, ο μετασχηματισμός Fourier μπορεί να εφαρμοστεί μόνο σε μια συνεχή κυματομορφή. Η επεξεργασία των ηχητικών κυμάτων γίνεται μετατρέποντας το αναλογικό σήμα σε ψηφιακό, και στην ουσία καταγράφονται δείγματα πλάτους της κυματομορφής, δηλαδή διακριτές τιμές και όχι μια συνεχή κυματομορφή. Έτσι, σε διακριτές κυματομορφές εφαρμόζεται ο διακριτός μετασχηματισμός Fourier (Discrete Fourier Transform - DFT), ο οποίος χρησιμοποιείται για την ανάλυση πεπερασμένων διακριτών σημάτων, τόσο κατά πλάτος όσο και κατά χρόνο.

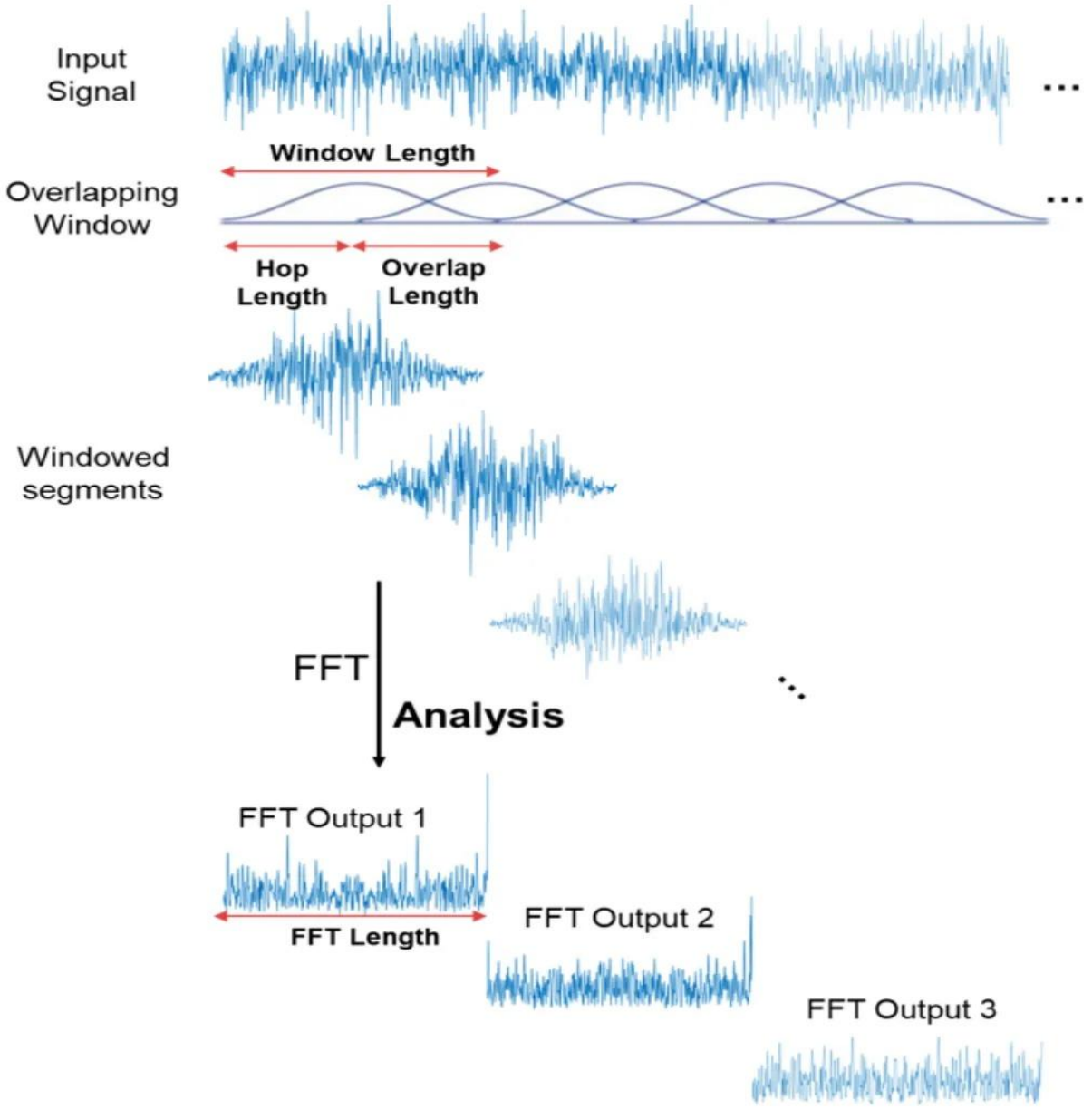
Ο DFT λοιπόν είναι ένα ισχυρό εργαλείο για την ανάλυση πεπερασμένων συνόλων δεδομένων, μετατρέποντας μια ακολουθία διακριτών σημείων από το πεδίο του χρόνου στο πεδίο της συχνότητας. Ωστόσο, ο υπολογισμός του DFT για μεγάλα σύνολα δεδομένων μπορεί να είναι πολύ απαιτητικός σε πόρους και χρόνο, καθώς απαιτεί O(N^2) πράξεις, όπου N είναι ο αριθμός των δειγμάτων. Αυτό σημαίνει ότι αν υπάρχουν 1000 δείγματα, ο υπολογισμός θα απαιτήσει περίπου 1.000.000 πράξεις. Όσο

μεγαλύτερος είναι ο αριθμός των δειγμάτων, τόσο περισσότερες πράξεις απαιτούνται, γεγονός που καθιστά τη διαδικασία υπολογιστικά πολύ απαιτητική για μεγάλα σύνολα δεδομένων.

Για να αντιμετωπιστεί αυτό το πρόβλημα, αναπτύχθηκε ο ταχύς μετασχηματισμός Fourier (Fast Fourier Transform - FFT), ένας αλγόριθμος που βελτιστοποιεί τον υπολογισμό του DFT, μειώνοντας σημαντικά τον χρόνο επεξεργασίας. Ο αλγόριθμος FFT χρησιμοποιεί την τεχνική "διαίρει και βασίλευε" (divide-and-conquer) για να διασπάσει το πρόβλημα του υπολογισμού του DFT σε μικρότερα υποπροβλήματα, τα οποία μπορούν να επιλυθούν πιο αποτελεσματικά.

Ο FFT μειώνει την πολυπλοκότητα του υπολογισμού από O(N^2) σε O(NlogN), καθιστώντας τον υπολογισμό του DFT πολύ ταχύτερο για μεγάλα σύνολα δεδομένων. Αυτό σημαίνει ότι ο υπολογισμός του DFT γίνεται πολύ ταχύτερος για μεγάλα σύνολα δεδομένων. Για παράδειγμα, αν υπάρχουν 1000 δείγματα, ο FFT θα απαιτήσει περίπου 10.000 πράξεις αντί για 1.000.000, κάνοντάς τον πολύ πιο αποδοτικό. Επομένως, ο FFT αποτελεί την αποδοτική εκδοχή του DFT, επιτρέποντας την ταχύτερη και πιο αποδοτική ανάλυση των συχνοτήτων σε μεγάλα σύνολα δεδομένων, καθιστώντας τον ένα ουσιώδες εργαλείο στη σύγχρονη ψηφιακή επεξεργασία σήματος.[11]

2. Mel-spectrogram: Αν το περιεχόμενο συχνοτήτων του σήματος μεταβάλλεται με την πάροδο του χρόνου, όπως συμβαίνει με τα περισσότερα ακουστικά σήματα όπως η μουσική και η ομιλία, ο FFT δεν επαρκεί. Αυτά τα σήματα είναι γνωστά ως μη στατικά (non-stationary). Υπάρχει ανάγκη για μια μέθοδο που να αναπαριστά το φάσμα αυτών των σημάτων καθώς μεταβάλλονται με τον χρόνο. Μια προσεγγιστική μέθοδος είναι ο υπολογισμός πολλαπλών φασμάτων με την εφαρμογή



Εικόνα 2.6: Σχήμα που παρουσιάζει σήμα, εφαρμογή παραθύρωσης που οδηγεί σε επιμέρους φάσματα. Τελικά, το φασματογράφημα είναι η παράθεση του φάσματος για κάθε παράθυρο.

του FFT σε διαδοχικά τμηματικά παράθυρα του σήματος. Αυτή η διαδικασία ονομάζεται μετασχηματισμός Fourier μικρού χρόνου (short time Fourier transform - STFT). Στον STFT, ο FFT υπολογίζεται σε επικαλυπτόμενα τμηματικά παράθυρα του σήματος, με αποτέλεσμα τον σχηματισμό του φασματογράφου.

Η εικόνα 2.6 απεικονίζει τη διαδικασία υπολογισμού του φασματογράφου χρησιμοποιώντας τον STFT. Αρχικά, το σήμα εισόδου αποτελείται από μια σειρά από ηχητικά κύματα, τα οποία είναι μη στατικά (non-stationary). Αυτό σημαίνει ότι οι διάφορες συχνότητες και οι εντάσεις τους δεν παραμένουν σταθερές, αλλά μεταβάλλονται συνεχώς καθώς το σήμα εξελίσσεται.

Στη συνέχεια, το σήμα χωρίζεται σε μικρά τμήματα, γνωστά ως παράθυρα. Κάθε παράθυρο έχει ένα συγκεκριμένο μήκος. Αυτά τα παράθυρα επικαλύπτονται με τα διπλανά τους, δηλαδή κάθε παράθυρο καλύπτει μέρος τόσο του προηγούμενου όσο και του επόμενου παραθύρου. Το μήκος του βήματος (hop length) είναι η απόσταση μεταξύ των διαδοχικών παραθύρων και καθορίζει πόσο μακριά μετακινείται κάθε παράθυρο σε σχέση με το προηγούμενο. Το μέγεθος της επικάλυψης των παραθύρων επηρεάζεται τόσο από το μήκος του παραθύρου όσο και από το μήκος του βήματος.

Σε κάθε παράθυρο του σήματος εφαρμόζεται ο FFT, ο οποίος μετατρέπει το σήμα από το χρονικό πεδίο στο πεδίο της συχνότητας. Τα αποτελέσματα του FFT για κάθε παράθυρο δείχνουν το φάσμα συχνοτήτων για το αντίστοιχο τμήμα του σήματος. Μετά τον υπολογισμό του FFT, εφαρμόζονται φίλτρα διέλευσης ζώνης σε κάθε παράθυρο του σήματος στο πεδίο της συχνότητας. Τα φίλτρα διέλευσης ζώνης επιτρέπουν μόνο συγκεκριμένα εύρη συχνοτήτων να περάσουν, ενώ καταστέλλουν τις υπόλοιπες. Αυτά τα φίλτρα τοποθετούνται με βάση την κλίμακα Mel.

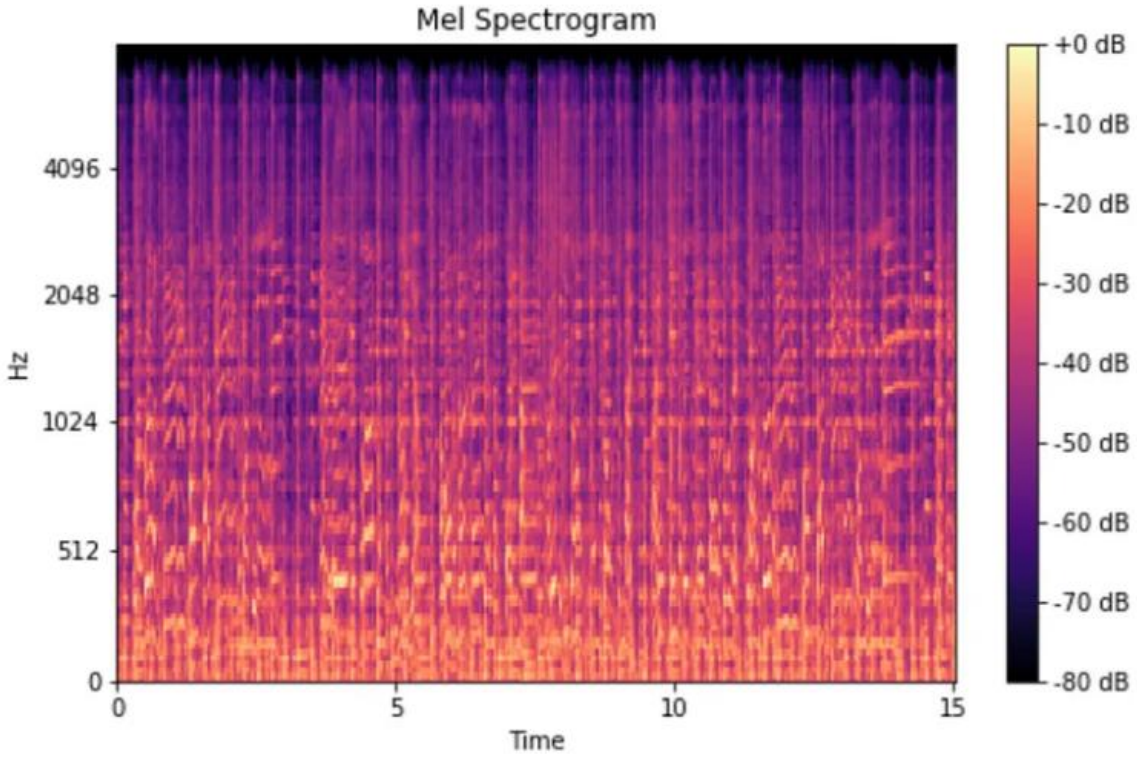
Οι άνθρωποι δεν αντιλαμβάνονται τις συχνότητες με τον ίδιο τρόπο σε όλο το φάσμα των ακουστικών συχνοτήτων. Η ακοή είναι πιο ευαίσθητη στις αλλαγές των χαμηλότερων συχνοτήτων σε σχέση με τις υψηλότερες συχνότητες. Για παράδειγμα, οι άνθρωποι μπορούν να ξεχωρίσουν εύκολα τη διαφορά μεταξύ δύο συχνοτήτων όπως 500 Hz και 1000 Hz, αλλά είναι πολύ πιο δύσκολο να ξεχωρίσουμε τη διαφορά μεταξύ 10.000 Hz και 10.500 Hz, ακόμη και αν η διαφορά μεταξύ των δύο ζευγών συχνοτήτων είναι η ίδια (500 Hz).

Για να αντιμετωπιστεί αυτή η ασυμμετρία στην ανθρώπινη αντίληψη των συχνοτήτων, οι επιστήμονες Stevens, Volkmann και Newmann εισήγαγαν το 1937 την κλίμακα Mel. Η κλίμακα Mel προσαρμόζει τις συχνότητες έτσι ώστε οι διαφορές στις συχνότητες να γίνονται αντιληπτές με τον ίδιο τρόπο από τον ακροατή, ανεξάρτητα από το αν είναι χαμηλές ή υψηλές. [12]

Μετά την εφαρμογή των φίλτρων, εφαρμόζεται μια λογαριθμική συνάρτηση στην έξοδο κάθε φίλτρου. Αυτό γίνεται για να μειωθεί το δυναμικό εύρος των σημάτων, δηλαδή η διαφορά μεταξύ των πιο δυνατών και των πιο αδύναμων σημάτων. Η λογαριθμική συνάρτηση μειώνει αυτή τη διαφορά, καθιστώντας τις διαφορές στο πλάτος πιο εύκολα αντιληπτές.

Τέλος, τα επεξεργασμένα φάσματα συχνοτήτων που προέκυψαν από τα παραπάνω βήματα συνδυάζονται ευθυγραμμίζοντας τα αποτελέσματα από κάθε παράθυρο του σήματος σε ένα κοινό χρονικό πλαίσιο. Για κάθε παράθυρο, οι τιμές συχνοτήτων τοποθετούνται σε διαδοχικά χρονικά διαστήματα, δημιουργώντας μια συνεχή ροή πληροφορίας. Το Mel-Spectrogram σχηματίζεται από τη σύνθεση αυτών των συνεχόμενων φασμάτων, με κάθε χρονική στιγμή να αντιπροσωπεύει τις συχνότητες και τις εντάσεις που υπάρχουν σε εκείνο το συγκεκριμένο χρονικό διάστημα.

Επομένως, το Mel-Spectrogram είναι μια δισδιάστατη οπτική αναπαράσταση όπου ο οριζόντιος άξονας αντιπροσωπεύει τον χρόνο, ο κατακόρυφος άξονας αντιπροσωπεύει τη συχνότητα, και το χρώμα κάθε σημείου δείχνει την ένταση της ενέργειας σε συγκεκριμένες συχνότητες και χρονικές στιγμές (Εικόνα 2.7). Σε αντίθεση με ένα τυπικό φασματογράφημα (Spectrogram), το οποίο συνήθως αναπαριστά μόνο την κατανομή της έντασης στις συχνότητες (άξονας y), το Mel-Spectrogram παρέχει μια δυναμική αναπαράσταση που δείχνει πώς οι συχνότητες και οι εντάσεις τους μεταβάλλονται με την πάροδο του χρόνου. Αυτή η μέθοδος προσφέρει μια λεπτομερή εικόνα των χρονικών μεταβολών στις συχνότητες και τις εντάσεις τους, καθιστώντας τη χρήσιμη για την ανάλυση μη στατικών ακουστικών σημάτων όπως η μουσική και η ομιλία.[13]



Εικόνα 2.7: Τρόπος απεικόνισης του φασματογραφήματος. Ο άξονας x απεικονίζει τον χρόνο, ο y τις

συχνότητες και ως πλάτος του αντίστοιχου ημιτόνου χρησιμοποιείται ένας χρωματικός κώδικας (μπάρα στα δεξιά του φασματογραφήματος). Σύμφωνα με την συγκεκριμένη, όσο πιο υψηλής έντασης είναι το αντίστοιχο ημίτονο, τόσο πιο θερμό είναι και το χρώμα που χρησιμοποιείται.

3. Power Spectrum: Όπως αναφέρθηκε, εφαρμόζοντας τον μετασχηματισμό Fourier στο ηχητικό σήμα προκύπτει το φάσμα, το οποίο παρέχει πληροφορίες για τις συχνότητες και τα πλάτη τους. Στη συνέχεια, το Power Spectrum προκύπτει με τον υπολογισμό του τετραγώνου των απόλυτων τιμών του φάσματος. Αυτή η διαδικασία προσφέρει μια λεπτομερή αναπαράσταση της κατανομής της ισχύος στις διάφορες συχνότητες, μετρώντας την ενέργεια που συνεισφέρει κάθε συχνότητα στο συνολικό σήμα. Με άλλα λόγια, το Power Spectrum αποκαλύπτει πόση ισχύς (ή ενέργεια) υπάρχει σε κάθε συχνότητα, δίνοντας μια εικόνα της ενεργειακής δομής του ηχητικού σήματος.

Στην περίπτωση των φωνητικών σημάτων, το Power Spectrum είναι ιδιαίτερα χρήσιμο, καθώς μπορεί να αποκαλύψει σημαντικά χαρακτηριστικά της φωνής, όπως η ταυτότητα των φωνηέντων. Τα φωνήεντα διακρίνονται από άλλους ήχους λόγω των ιδιαίτερων συχνοτήτων που ενισχύονται στο φωνητικό σύστημα, οι οποίες είναι γνωστές ως formants. Οι formants είναι συγκεκριμένες συχνότητες που καθορίζονται από τη μορφή και τη θέση των φωνητικών οργάνων, όπως η γλώσσα και τα χείλη, κατά τη διάρκεια της παραγωγής των φωνηέντων. Κάθε φωνήεν έχει ένα μοναδικό σύνολο formants, που λειτουργούν σαν "δακτυλικά αποτυπώματα" για την αναγνώρισή του. Το Power Spectrum μπορεί να χρησιμοποιηθεί για να εντοπίσει αυτές τις formants, αναλύοντας τις κορυφές στο φάσμα ισχύος που αντιστοιχούν στις ενισχυμένες συχνότητες.

Ωστόσο, η ανάλυση του Power Spectrum ενός φωνητικού σήματος δεν είναι πάντα εύκολη, καθώς το εύρος των τιμών μπορεί να είναι πολύ άνισο. Αυτό σημαίνει ότι μπορεί να υπάρχουν μεγάλες διαφορές στις τιμές της ισχύος για διαφορετικές συχνότητες, καθιστώντας δύσκολη την αναγνώριση των σημαντικών πληροφοριών με μια απλή οπτική εξέταση. Αυτό το ζήτημα δημιουργεί προβλήματα στην εξαγωγή πληροφοριών από το Power Spectrum, καθώς οι κρίσιμες λεπτομέρειες μπορεί να είναι δύσκολο να εντοπιστούν χωρίς κατάλληλη επεξεργασία ή μετασχηματισμό των δεδομένων.

4. Logarithmic Spectrum: Από την άλλη, το logarithm Spectrum, το οποίο υπολογίζεται λογαριθμίζοντας το φάσμα, αποτελεί μια πολύ πιο προσιτή αναπαράσταση. Είναι όχι μόνο πιο οπτικά κατανοητό, αλλά και το γεγονός ότι ο λογαριθμικός υπολογισμός προσεγγίζει την ευαισθησία του ανθρώπινου αυτιού, επιτρέπει τη χρήση των λογαριθμικών φασμάτων για την εκτίμηση της ακουστικής σημασίας των φασματικών χαρακτηριστικών. Το logarithmic Spectrum απεικονίζει το φασματικό περιεχόμενο με τρόπο ώστε το μέγεθος των τιμών να είναι περίπου ομοιόμορφο σε όλο το φάσμα.

Η μόνη εξαίρεση σε αυτό είναι τα μηδενικά και άλλες πολύ μικρές τιμές στο Spectrum, οι οποίες δίνουν αρνητικά άπειρα ή αυθαίρετα μεγάλες αρνητικές τιμές. Αν και αυτές οι τιμές είναι «δύσκολες»

για την οπτικοποίησή τους, είναι ασήμαντες για την ακουστική αντίληψη και μπορούν συχνά να αγνοηθούν. Ωστόσο, οι αυθαίρετα μεγάλες αρνητικές τιμές αποτελούν πρόβλημα στο να υπολογιστεί το Logarithmic Spectrum. Για να μειώσουμε την εμφάνιση προβληματικών τιμών στο Logarithmic Spectrum, μπορούμε να προσθέσουμε έναν μικρό θετικό αριθμό, ε, πριν πάρουμε τον λογάριθμο του Spectrum, δηλαδή $y=\log(|x|^2+\varepsilon)$. Αυτό αποτρέπει την τιμή y από το να γίνει αρνητικά άπειρη ή πολύ αρνητική.

Εφαρμόζοντας και άλλες παρόμοιες τεχνικές, επιτυγχάνουμε ένα πιο σταθερό και ακριβές logarithmic Spectrum, το οποίο μπορεί να χρησιμοποιηθεί με αξιοπιστία σε διάφορες εφαρμογές ανάλυσης ήχου, εξασφαλίζοντας την καλύτερη δυνατή αναπαράσταση των ακουστικών χαρακτηριστικών.

5. Cepstrum: Το cepstrum προκύπτει από το logarithmic Spectrum, εφαρμόζοντας ένα δεύτερο μετασχηματισμό πάνω στο ήδη logarithmic Spectrum. Συγκεκριμένα, μπορεί να χρησιμοποιηθεί είτε ο διακριτός μετασχηματισμός Fourier (DFT) είτε ο διακριτός συνημιτονικός μετασχηματισμός (Discrete Cosine Transform - DCT).

Το cepstrum λοιπόν προκύπτει μέσω δύο διαδοχικών μετασχηματισμών χρόνου-συχνότητας. Αρχικά, εφαρμόζεται ο μετασχηματισμός Fourier στο αρχικό χρονικό σήμα για να προκύψει το φάσμα, έπειτα λογαριθμίζοντας προκύπτει το logarithmic Spectrum και τέλος, για την εξαγωγή του ceptsrum, εφαρμόζεται ένας δεύτερος μετασχηματισμός Fourier στο logarithmic spectrum. Αυτός ο δεύτερος μετασχηματισμός μετατρέπει το σήμα σε ένα νέο πεδίο, το οποίο αναπαριστά την κατανομή των περιόδων και όχι των συχνοτήτων, καθιστώντας το cepstrum παρόμοιο με το πεδίο του χρόνου.

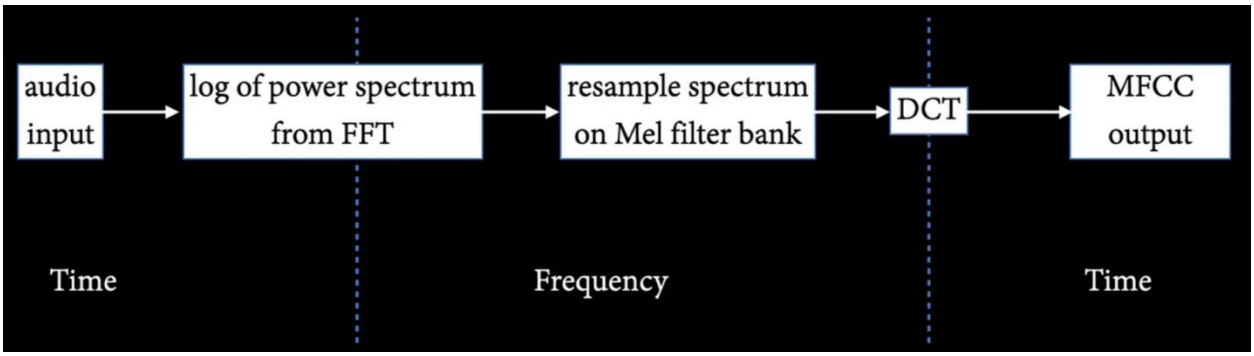
Αυτό συμβαίνει επειδή ο άξονας του cepstrum, γνωστός ως άξονας quefrency, μετρά τον "χρόνο" που απαιτείται για να επαναληφθεί μια συγκεκριμένη δομή μέσα στο σήμα. Έτσι, οι χαμηλές τιμές quefrency αντιστοιχούν σε αργά μεταβαλλόμενα χαρακτηριστικά, όπως οι formants, που αντιστοιχούν σε δομές που επαναλαμβάνονται αργά στον χρόνο. Με αυτή την έννοια, το cepstrum προσφέρει μια οπτική αναπαράσταση που μοιάζει με αυτή του πεδίου του χρόνου, αλλά αντί για πραγματικό χρόνο, αναπαριστά την "περιοδικότητα" ή την "επαναληψιμότητα" των στοιχείων του σήματος.

Το logarithmic Spectrum διαθέτει αρμονική δομή, στην οποία η θεμελιώδης συχνότητα εμφανίζεται ως μια χαρακτηριστική "δομή κτενίσματος" (comb-structure), δηλαδή μια σειρά από κορυφές που επαναλαμβάνονται περιοδικά στο φάσμα. Ο μετασχηματισμός Fourier που εφαρμόζεται για τον υπολογισμό του cepstrum είναι ιδιαίτερα κατάλληλος για την εξαγωγή χαρακτηριστικών από τέτοιες δομές. Στο cepstrum, λοιπόν, αναμένεται να υπάρχει μια κορυφή στο quefrency που αντιστοιχεί στο μήκος της περιόδου του τόνου (pitch-period length) σε δευτερόλεπτα. Αν υποτεθεί ότι οι θεμελιώδεις συχνότητες F0 βρίσκονται στο εύρος 80 έως 450 Hz, τότε η αντίστοιχη κορυφή στο cepstrum θα πρέπει να βρίσκεται στο quefrency 1/F0, δηλαδή μεταξύ 2,2 και 12,5 χιλιοστών του δευτερολέπτου.

Η εκτίμηση της θεμελιώδους συχνότητας στο cepstrum είναι στην πραγματικότητα πολύ απλή και σχετικά αξιόπιστη. Το μόνο που χρειάζεται είναι να βρεθεί η υψηλότερη κορυφή στο κατάλληλο εύρος quefrency.

6. Mel-frequency cepstral coefficients (MFCCs): Οι MFCCs (Mel-Frequency Cepstral Coefficients) είναι μια ειδική περίπτωση του cepstrum, όπου πριν από τον υπολογισμό του cepstrum

εφαρμόζεται η κλίμακα Mel. Συγκεκριμένα, μετά τον υπολογισμό του log-spectrum του ηχητικού σήματος, χρησιμοποιείται ένα φίλτρο Mel για να μετατρέψει τις συχνότητες σε αντιληπτικές μονάδες, προσαρμόζοντας τις συχνότητες σύμφωνα με την κλίμακα Mel. Στη συνέχεια, εφαρμόζεται ο διακριτός συνημιτονικός μετασχηματισμός (DCT) στα αποτελέσματα αυτά για να παραχθούν οι MFCCs.



Εικόνα 2.8: Διαδικασία εξαγωγής των χαρακτηριστικών Mel-Frequency Cepstral Coefficient (MFCC). Ο ήχος εισόδου αναλύεται με τον Fast Fourier Transform (FFT) και εξάγεται το λογαριθμοποιημένο πλάτος. Για την καλύτερη προσομοίωση της ανθρώπινης ακοής, γίνεται επιμέρους ομαδοποίηση των συχνοτήτων σύμφωνα με την κλίμακα Mel. Τελος εφαρμόζεται ο μετασχηματισμός συνημιτόνου (Discrete Cosine Transform - DCT) και παράγονται τα υπό έρευνα χαρακτηριστικά.

Οι MFCCs έχουν καταστεί το πιο ευρέως χρησιμοποιούμενο χαρακτηριστικό στις εφαρμογές αναγνώρισης ομιλίας και ήχου. Χρησιμοποιούνται επειδή είναι αποτελεσματικοί και έχουν σχετικά χαμηλή πολυπλοκότητα, ενώ είναι εύκολοι στην υλοποίηση.[14]

7. Spectral Centroid: Το spectral centroid είναι μια μέτρηση που περιγράφει το κέντρο βάρους του φάσματος ενός ηχητικού σήματος. Στην ουσία, αντιπροσωπεύει τη "μέση" συχνότητα του φάσματος, δηλαδή το σημείο στο οποίο η ενέργεια του σήματος είναι ισομερώς κατανεμημένη. Αυτό σημαίνει ότι οι συχνότητες που βρίσκονται πιο κοντά στο spectral centroid έχουν σημαντική συμμετοχή στην ενέργεια του σήματος.

Για να υπολογιστεί το spectral centroid, λαμβάνεται υπόψη η ένταση (ή το πλάτος) κάθε συχνότητας στο σήμα και σταθμίζεται η συμβολή της στο συνολικό φάσμα. Ουσιαστικά, κάθε συχνότητα πολλαπλασιάζεται με την αντίστοιχη έντασή της και το αποτέλεσμα αυτού του πολλαπλασιασμού προστίθεται για όλες τις συχνότητες. Αυτό το άθροισμα στη συνέχεια διαιρείται με το συνολικό άθροισμα των εντάσεων όλων των συχνοτήτων:

$$SC_{Hz} = \frac{\sum_{k=1}^{N-1} k \cdot X^d[k]}{\sum_{k=1}^{N-1} X^d[k]}$$	2.4

Όπου SC_{Hz} είναι το spectral centroid σε Hertz, X_d[k] είναι το μέγεθος (πλάτος τάσης) που αντιστοιχεί στο frequency bin k, k είναι το frequency bin σε Hertz (δηλαδή, το εύρος συγκεκριμένων συχνοτήτων που προκύπτουν από τη χρήση παραθύρων), fs είναι η συχνότητα δειγματοληψίας και N είναι ο αριθμός των δειγμάτων.

Αυτός ο υπολογισμός δίνει μια τιμή που αντιπροσωπεύει τη συχνότητα στην οποία συγκεντρώνεται το μεγαλύτερο μέρος της ενέργειας του σήματος. Το spectral centroid είναι μια σημαντική ένδειξη για τον χαρακτήρα του ήχου: υψηλότερες τιμές του spectral centroid συνδέονται με πιο "λαμπερούς" ήχους, που περιέχουν περισσότερες υψηλές συχνότητες, ενώ χαμηλότερες τιμές δείχνουν πιο "σκοτεινούς" ήχους με περισσότερες χαμηλές συχνότητες.[15]

8. Spectral Bandwidth: Το spectral bandwidth είναι ένα μέτρο που χρησιμοποιείται για να περιγράψει το εύρος των συχνοτήτων σε ένα σήμα. Αυτό το μέτρο δείχνει πόσο "πλατιά" ή "στενή" είναι η κατανομή των συχνοτήτων στο σήμα. Συγκεκριμένα, το spectral bandwidth μπορεί να βοηθήσει να κατανοήσουμε πόσο διασκορπισμένες είναι οι συχνότητες που συνθέτουν το σήμα.

Για να υπολογιστεί το spectral bandwidth, χρησιμοποιείται συχνά η μέση τετραγωνική απόκλιση των συχνοτήτων από μια κεντρική συχνότητα, η οποία ονομάζεται επίσης και spectral centroid. Η κεντρική συχνότητα είναι η μέση συχνότητα γύρω από την οποία συγκεντρώνεται η ενέργεια του σήματος. Το spectral bandwidth μετράει το εύρος των συχνοτήτων γύρω από αυτήν την κεντρική συχνότητα, δείχνοντας πόσο "διασκορπισμένες" είναι αυτές οι συχνότητες.

Μια κοινή μέθοδος για τον υπολογισμό του spectral bandwidth είναι να υπολογιστεί η ρίζα της μέσης τετραγωνικής απόκλισης των συχνοτήτων από το spectral centroid, λαμβάνοντας υπόψη τις εντάσεις (ή πλάτη) των αντίστοιχων συχνοτήτων. Η εξίσωση που χρησιμοποιείται είναι:

$$\Delta\omega_2 = \sqrt{\sum_{k=1}^{N} (f_k - SC_{Hz})^2 \cdot X_d[k]}$$ **1.**	2.5

Όπου Δω_2 είναι το spectral bandwidth, f_k είναι η συχνότητα στο bin k, k είναι το frequency bin σε Hertz (δηλαδή, το εύρος συγκεκριμένων συχνοτήτων που προκύπτουν από τη χρήση παραθύρων), SC_Hz είναι το spectral centroid σε Hertz, X_d[k] είναι η ένταση (ή πλάτος) στη συχνότητα f_k, N είναι ο συνολικός αριθμός των συχνοτήτων.[16]

9. Spectral_contrast: Το Spectral Contrast είναι ένα χαρακτηριστικό που βοηθά στην ανάλυση της σχετικής έντασης των συχνοτήτων ενός ηχητικού σήματος, εξετάζοντας συγκεκριμένες περιοχές του φάσματος (αυτές οι περιοχές προκύπτουν από τη διαίρεση του συνολικού Spectrum σε μικρότερα τμήματα, χρησιμοποιώντας τεχνικές που αναφέρθηκαν όπως το windowing). Περιλαμβάνει πληροφορίες για την κατανομή των φασματικών κορυφών και κοιλάδων σε αυτές τις διάφορες περιοχές του Spectrum ενός σήματος.

Οι φασματικές κορυφές και κοιλάδες αναφέρονται στις μέγιστες και ελάχιστες τιμές της έντασης του Spectrum αντίστοιχα. Οι φασματικές κορυφές αντιπροσωπεύουν τις συχνότητες με την υψηλότερη ένταση, οι οποίες συχνά συνδέονται με αρμονικές του σήματος, ενώ οι φασματικές κοιλάδες είναι περιοχές χαμηλής έντασης, που μπορεί να αντιστοιχούν σε θόρυβο ή άλλες μη αρμονικές συνιστώσες.

Για τον υπολογισμό του spectral contrast, αρχικά υπολογίζεται το φάσμα Στη συνέχεια, το φάσμα διαιρείται σε περιοχές με χρήση φίλτρων κλίμακας οκτάβας (octave-scale filters). Σε κάθε περιοχή, υπολογίζεται η ισχύς των φασματικών κορυφών και κοιλάδων, καθώς και η διαφορά τους. Η εκτίμηση αυτή γίνεται λαμβάνοντας τη μέση τιμή σε μια μικρή γειτονιά γύρω από τις μέγιστες και ελάχιστες τιμές. Τέλος, εφαρμόζεται ο μετασχηματισμός Karhunen-Loeve.

Ο μετασχηματισμός Karhunen-Loeve (K-L), γνωστός και ως ανάλυση κύριων συνιστωσών (Principal Component Analysis - PCA), είναι μια τεχνική στατιστικής ανάλυσης που χρησιμοποιείται για τη μείωση των διαστάσεων ενός συνόλου δεδομένων. Η βασική του λειτουργία είναι να μετασχηματίσει τα δεδομένα σε ένα νέο σύνολο μεταβλητών, τις λεγόμενες κύριες συνιστώσες, οι οποίες είναι διατεταγμένες κατά φθίνουσα σειρά με βάση την ποσότητα της μεταβλητότητας που εξηγούν στα δεδομένα.

Συγκεκριμένα, ο μετασχηματισμός K-L επιδιώκει να βρει μια νέα βάση για το δεδομένο διανυσματικό χώρο, όπου οι διαστάσεις (άξονες) είναι ορθογώνιες μεταξύ τους και δεν είναι συσχετισμένες. Η "απομάκρυνση συσχέτισης" σημαίνει ότι οι νέες διαστάσεις ή οι κύριες συνιστώσες δεν επηρεάζονται η μία από την άλλη — δηλαδή, η τιμή σε μια διάσταση δεν εξαρτάται από τις τιμές στις άλλες διαστάσεις.

Στην περίπτωση του Spectral Contrast, ο μετασχηματισμός Karhunen-Loeve χρησιμοποιείται για να διασφαλιστεί ότι τα χαρακτηριστικά που εξάγονται από τις φασματικές κορυφές και κοιλάδες δεν έχουν αλληλεξαρτήσεις μεταξύ τους. Αυτό επιτρέπει στα δεδομένα να αναλυθούν πιο αποτελεσματικά και να βελτιωθεί η απόδοση των αλγορίθμων κατηγοριοποίησης, καθώς οι αλληλεξαρτήσεις μεταξύ χαρακτηριστικών μπορεί να οδηγήσουν σε μειωμένη ακρίβεια στην αναγνώριση των κατηγοριών.

Συνοψίζοντας, το Spectral Contrast παρέχει πληροφορίες για την ένταση και τη διαφορά μεταξύ φασματικών κορυφών και κοιλάδων σε υπο-ζώνες του σήματος, επιτρέποντας την αναγνώριση

αρμονικών και μη αρμονικών στοιχείων. Οι αρμονικές συνιστώσες είναι συχνότητες που σχετίζονται με την βασική συχνότητα του σήματος και δημιουργούν έναν "καθαρό" ήχο, ενώ οι μη αρμονικές συνιστώσες περιλαμβάνουν θόρυβο και άλλες παραμορφώσεις. Αυτές οι πληροφορίες είναι πολύτιμες σε εφαρμογές όπως η κατηγοριοποίηση μουσικής, η ανάλυση της ποιότητας του ήχου και η ανίχνευση θορύβου, καθώς επιτρέπουν την κατανόηση της φασματικής δομής και της υφής του σήματος.[17]

10. Spectral Flatness: Το Spectral Flatness είναι ένας δείκτης που χρησιμοποιείται για να περιγράψει την κατανομή της ενέργειας σε ένα Spectrum. Αξιολογεί πόσο ομοιόμορφα κατανεμημένη είναι η ενέργεια ενός ηχητικού σήματος στις διάφορες συχνότητές του. Με άλλα λόγια, εξετάζει αν η ενέργεια είναι συγκεντρωμένη σε λίγες συχνότητες (όπως συμβαίνει με τους καθαρούς τόνους) ή αν είναι διασκορπισμένη ομοιόμορφα σε ένα ευρύ φάσμα συχνοτήτων (όπως συμβαίνει με τον θόρυβο). Ένα υψηλό επίπεδο ομοιομορφίας υποδεικνύει ότι το σήμα περιέχει πολλές συχνότητες με παρόμοια επίπεδα ενέργειας, χωρίς ιδιαίτερα έντονες κορυφές, ενώ ένα χαμηλό επίπεδο ομοιομορφίας υποδηλώνει την παρουσία συγκεκριμένων συχνοτήτων με υψηλή ενέργεια σε σχέση με τις υπόλοιπες.

Για τον υπολογισμό του Spectral Flatness, χρησιμοποιείται ο λόγος της γεωμετρικής μέσης τιμής προς την αριθμητική μέση τιμή του Power Spectrum. Συγκεκριμένα, η γεωμετρική μέση τιμή υπολογίζεται παίρνοντας το γινόμενο όλων των τιμών του Power Spectrum και στη συνέχεια την νιοστή ρίζα αυτού του γινομένου, όπου n είναι ο συνολικός αριθμός των τιμών. Η γεωμετρική μέση τιμή είναι ευαίσθητη σε μικρές τιμές και επηρεάζεται σημαντικά από αυτές.

Η αριθμητική μέση τιμή υπολογίζεται παίρνοντας το άθροισμα όλων των τιμών του Power Spectrum και στη συνέχεια διαιρώντας το άθροισμα με τον αριθμό των τιμών. Η αριθμητική μέση τιμή δείχνει το μέσο επίπεδο της ενέργειας στο φάσμα.

$$\text{spectral flatness} = \frac{\text{geometric mean}}{\text{arithmetic mean}} = \frac{\sqrt[N]{\prod_k x(k)}}{\frac{1}{N}\sum_k x(k)}$$	2.6

Όπου, x είναι το φάσμα, k είναι ένας δείκτης στο φάσμα και N είναι ο αριθμός των (μη μηδενικών) στοιχείων στο φάσμα.

Ο λόγος αυτών των δύο μέτρων δίνει το Spectrum Flatness, το οποίο κυμαίνεται μεταξύ 0 και 1. Ένα Spectral Flatness κοντά στο 0 υποδηλώνει ότι το φάσμα έχει έντονες κορυφές, όπως σε καθαρούς τόνους, ενώ ένα Spectral Flatness κοντά στο 1 υποδηλώνει ότι το φάσμα είναι περισσότερο επίπεδο, όπως στον λευκό θόρυβο.[18]

11. Spectral-Roll off: Το Spectral Roll-off είναι ένα μέτρο που χρησιμοποιείται για να καθορίσει τη συχνότητα κάτω από την οποία βρίσκεται ένα συγκεκριμένο ποσοστό της συνολικής ενέργειας του Power Spectrum ενός σήματος. Συνήθως, αυτό το ποσοστό είναι 85% ή 95%. Αυτό σημαίνει ότι υπολογίζεται η συχνότητα κάτω από την οποία συγκεντρώνεται το προκαθορισμένο ποσοστό της συνολικής ενέργειας του σήματος, με το υπόλοιπο ποσοστό της ενέργειας να βρίσκεται σε υψηλότερες συχνότητες.

Για τον υπολογισμό του Spectral Roll-off, αρχικά έχουμε το φάσμα του σήματος, το οποίο προκύπτει μέσω του μετασχηματισμού Fourier. Από αυτό το φάσμα, υπολογίζεται το Power Spectrum. Στη συνέχεια, υπολογίζεται το σωρευτικό άθροισμα των τιμών ισχύος για όλες τις συχνότητες, ξεκινώντας από τη χαμηλότερη μέχρι την υψηλότερη. Το ποσοστό του Spectral Roll-off (π.χ., 85%) καθορίζει το σημείο στο σωρευτικό άθροισμα όπου βρίσκεται το συγκεκριμένο ποσοστό της συνολικής ενέργειας. Η αντίστοιχη συχνότητα είναι αυτή που ονομάζεται συχνότητα του Spectral Roll-off.

Είναι χρήσιμο στην ανάλυση και την κατηγοριοποίηση ηχητικών σημάτων, καθώς μπορεί να δώσει πληροφορίες για τη φύση του σήματος. Σήματα με υψηλότερο Spectral Roll-off τείνουν να έχουν περισσότερες υψηλές συχνότητες και μπορούν να υποδεικνύουν την παρουσία θορύβου ή αρμονικών, ενώ σήματα με χαμηλότερο Spectral Roll-off υποδεικνύουν κυρίως χαμηλές συχνότητες, όπως στη φωνή ή σε συγκεκριμένα μουσικά όργανα. Αυτό το χαρακτηριστικό χρησιμοποιείται ευρέως σε εφαρμογές όπως η αναγνώριση μουσικών ειδών, η ανίχνευση φωνής, και η ανάλυση της ποιότητας του ήχου.[19]

2.2 Η Μηχανική Μάθηση

2.2.1 Η Ιστορική Πορεία της Τεχνητής Νοημοσύνης: Από τις Πρώτες Ιδέες στη Σύγχρονη Εποχή

Η ιστορία της τεχνητής νοημοσύνης (Artificial Intelligence - AI) είναι πλούσια και πολυδιάστατη, καλύπτοντας πάνω από επτά δεκαετίες από τις πρώτες θεωρητικές ιδέες έως τις σύγχρονες εφαρμογές. Ξεκινώντας από το 1943, οι Warren McCulloch και Walter Pitts παρουσίασαν το πρώτο αναγνωρισμένο έργο στον τομέα της AI, δημιουργώντας ένα μοντέλο τεχνητών νευρωνικών δικτύων (Neural Network – NN), το οποίο αποτέλεσε τη βάση για μελλοντικές έρευνες στον εγκέφαλο και τις νευρωνικές λειτουργίες.

Η έννοια της AI καθιερώθηκε επίσημα το 1956 κατά τη διάρκεια του καλοκαιρινού εργαστηρίου στο Dartmouth College, όπου ο John McCarthy, μαζί με άλλους ερευνητές, συναντήθηκαν για να συζητήσουν

την προοπτική των μηχανών να μιμούνται την ανθρώπινη νοημοσύνη. Αυτό το συνέδριο σημάδεψε τη γέννηση ενός νέου επιστημονικού πεδίου, οδηγώντας στην πρώτη χρυσή εποχή της AI, όπου τα πρώτα συστήματα και θεωρίες αναπτύχθηκαν με ενθουσιασμό και μεγάλες προσδοκίες.

Οι δεκαετίες του 1960 και 1970 έφεραν τόσο την πρόοδο όσο και την απογοήτευση, καθώς οι πρώτες υποσχέσεις της AI δεν επιτεύχθηκαν πλήρως. Προγράμματα όπως το MYCIN, το οποίο αναπτύχθηκε για τη διάγνωση λοιμωδών νοσημάτων του αίματος, και το DENDRAL, που χρησιμοποιήθηκε για χημική ανάλυση, εισήγαγαν την ιδέα των συστημάτων εμπειρογνωμόνων και της μηχανικής γνώσεων, αναδεικνύοντας τη σημασία της ειδικής γνώσης στενά περιορισμένων τομέων.

Στα τέλη της δεκαετίας του 1980 και στη δεκαετία του 1990, η επανάσταση των NN και της ασαφούς λογικής (fuzzy logic) έφεραν νέες προοπτικές και εφαρμογές, με την τεχνητή νοημοσύνη να ενσωματώνει ικανότητες μάθησης και προσαρμογής.

Η AI έχει γνωρίσει τεράστια ανάπτυξη και έχει γίνει ολοένα και πιο θεσμοθετημένη κατά τον 21ο αιώνα. Στοχεύει στην επέκταση και βελτίωση των δυνατοτήτων του ανθρώπου σε καθήκοντα που αφορούν τη φύση και τη διακυβέρνηση της κοινωνίας μέσω εξυπνότερων μηχανών, με τελικό στόχο τη συνύπαρξη ανθρώπων και μηχανών σε αρμονία. Η AI χρησιμοποιεί υπολογιστές για να προσομοιώσει ανθρώπινες νοητικές συμπεριφορές όπως η μάθηση, η κρίση και η λήψη αποφάσεων και έχει σημειώσει αξιοσημείωτα αποτελέσματα σε εφαρμογές όπως η αναγνώριση ομιλίας, η επεξεργασία εικόνας, η φυσική επεξεργασία γλώσσας, η απόδειξη αυτόματων θεωρημάτων και τα έξυπνα ρομπότ. Η ταχεία εξέλιξη της AI έχει αλλάξει τον τρόπο ζωής των ανθρώπων και έχει γίνει μια σημαντική στρατηγική ανάπτυξης για πολλές χώρες, βελτιώνοντας την εθνική ανταγωνιστικότητα και διατηρώντας την ασφάλεια. Κορυφαίες εταιρείες όπως η Google, η Microsoft και η IBM έχουν δεσμευτεί στην εφαρμογή της AI σε πολλούς τομείς, ενισχύοντας περαιτέρω την τεχνολογική πρόοδο και τη διεθνή ανταγωνιστικότητα.

Η AI έχει βιώσει μια εκρηκτική ανάπτυξη τα τελευταία χρόνια, χάρη σε μια σειρά καταλυτικών παραγόντων που έχουν συνδυάσει τεχνολογικές καινοτομίες, ακαδημαϊκή έρευνα και οικονομικές επενδύσεις σε ένα ισχυρό ρεύμα προόδου. Καθοριστικός παράγοντας στην πρόοδο της AI ήταν η ταχεία εξέλιξη και διαθεσιμότητα μεγάλων δεδομένων (Big Data), τα οποία παρέχουν την απαραίτητη ύλη για την εκπαίδευση και τη βελτίωση των αλγορίθμων μηχανικής μάθησης. Η ευρεία εφαρμογή του διαδικτύου των Πραγμάτων (Internet of Things - IoT) έχει επίσης πολλαπλασιάσει την ποσότητα δεδομένων που παράγονται, επεκτείνοντας τις διαστάσεις και την ποικιλομορφία τους.

Παράλληλα, οι αλγόριθμοι που αναπτύχθηκαν για την αναγνώριση μοτίβων και τη μάθηση από δεδομένα έχουν γίνει πιο αποτελεσματικοί και εξελιγμένοι. Μηχανική μάθηση (Machine Learning – ML), μια υποκατηγορία της AI, χρησιμοποιεί αλγορίθμους που βελτιώνουν την απόδοσή τους με την εμπειρία, διευρύνοντας τις εφαρμογές της AI σε πεδία όπως η αναγνώριση ομιλίας, η επεξεργασία εικόνας, η φυσική επεξεργασία γλώσσας και άλλα.

Ταυτόχρονα, η συνεχής αύξηση της υπολογιστικής ισχύος, ειδικά μέσω της εξέλιξης των καρτών γραφικών (Graphics Processing Unit – GPU) και της προσαρμοσμένης ανάπτυξης επεξεργαστών, έχει

σπάσει τους περιορισμούς που είχαν καθυστερήσει παλαιότερα την ανάπτυξη της AI. Αυτό έχει οδηγήσει σε μια εκρηκτική ανάπτυξη και εξάπλωση της τεχνητής νοημοσύνης σε παγκόσμιο επίπεδο, με κράτη και οργανισμούς να εντάσσουν την AI στρατηγικά στις εθνικές τους αναπτυξιακές πολιτικές, ενισχύοντας την εθνική ανταγωνιστικότητα και την ασφάλεια. [20]

2.2.2 Τι είναι η Μηχανική Μάθηση

Η Μάθηση (Learning) είναι μία από τις θεμελιώδεις ιδιότητες της νοήμονος συμπεριφοράς του ανθρώπου. Παρά τις μελέτες και τις έρευνες επί χρόνια από τους επιστήμονες του πεδίου της Γνωστικής Ψυχολογίας και τους φιλοσόφους, η έννοια της μάθησης δεν έχει γίνει πλήρως κατανοητή. Πώς, λοιπόν, θα μπορούσαν οι επιστήμονες του χώρου της AI να δημιουργήσουν υπολογιστικά συστήματα ικανά να μάθουν. Η ML μπορεί να οριστεί ως το φαινόμενο κατά το οποίο ένα σύστημα βελτιώνει την απόδοσή του κατά την εκτέλεση μιας συγκεκριμένης εργασίας, χωρίς να υπάρχει ανάγκη να προγραμματιστεί εκ νέου. Βάσει του ορισμού αυτού, η ML έχει ως σκοπό τη δημιουργία μηχανών ικανών να μαθαίνουν, να βελτιώνουν, δηλαδή, την απόδοσή τους σε κάποιους τομείς μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας. Ένας σχετικός γενικός ορισμός ML δίνεται από τον Mitchell το 1997: «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E.» Η εκπαίδευση δεν περιορίζεται σε μια αρχική προσαρμογή κατά τη διάρκεια ενός ορισμένου χρονικού διαστήματος. Όπως συμβαίνει και με τους ανθρώπους, ένας καλός αλγόριθμος μπορεί να εξασκεί "δια βίου" μάθηση καθώς επεξεργάζεται νέα δεδομένα και μαθαίνει από τα λάθη του. [21]

2.2.3 Κατηγορίες και Εφαρμογές Τεχνικών Μηχανικής Μάθησης

Οι τεχνικές ML χωρίζονται κυρίως σε τέσσερις κατηγορίες: Επιβλεπόμενη μάθηση, Μη επιβλεπόμενη μάθηση, ημι-επιβλεπόμενη μάθηση και Ενισχυτική μάθηση.

Επιβλεπόμενη Μάθηση:

Η επιβλεπόμενη μάθηση (superv αφορά την εκμάθηση μιας συνάρτησης που αντιστοιχίζει μια είσοδο σε μια έξοδο, χρησιμοποιώντας παραδείγματα εισόδου-εξόδου που έχουν ήδη χαρακτηριστεί με σωστές απαντήσεις. Αυτό σημαίνει ότι τα δεδομένα εκπαίδευσης περιλαμβάνουν τόσο τις εισόδους όσο και τις σωστές εξόδους, επιτρέποντας στο σύστημα να μάθει πώς να μετατρέπει μια είσοδο στην επιθυμητή έξοδο. Η επιβλεπόμενη μάθηση εφαρμόζεται όταν έχουν καθοριστεί σαφείς στόχοι που πρέπει να επιτευχθούν μέσω των εισόδων, ακολουθώντας έτσι μια προσέγγιση που βασίζεται σε συγκεκριμένα καθήκοντα.

Ένα κατανοητό παράδειγμα επιβλεπόμενης μάθησης είναι η πρόβλεψη της ετικέτας κλάσης ή του συναισθήματος ενός κειμένου. Για παράδειγμα, μπορεί να χρησιμοποιηθεί η επιβλεπόμενη μάθηση για να προσδιοριστεί αν ένα tweet είναι θετικό, αρνητικό ή ουδέτερο, ή να αναλύσουμε τις κριτικές προϊόντων για να αποφανθούμε αν είναι θετικές ή αρνητικές. Αυτή η διαδικασία περιλαμβάνει την εκπαίδευση του μοντέλου με δεδομένα που έχουν ήδη κατηγοριοποιηθεί σωστά, ώστε να μπορεί στη συνέχεια να εφαρμόσει τις γνώσεις αυτές σε νέα, άγνωστα δεδομένα.

Εν κατακλείδι, η επιβλεπόμενη μάθηση βασίζεται στην ύπαρξη δεδομένων εκπαίδευσης που έχουν ήδη χαρακτηριστεί με τις σωστές απαντήσεις και αποσκοπεί στην εκμάθηση συναρτήσεων που επιτρέπουν την αξιόπιστη πρόβλεψη εξόδων για νέες εισόδους, καθιστώντας την μια ισχυρή προσέγγιση για πολλές πρακτικές εφαρμογές.

Μη-Επιβλεπόμενη Μάθηση:

Η μη επιβλεπόμενη μάθηση (unsupervised learning) είναι μια τεχνική εκμάθησης μηχανών που αναλύει σύνολα δεδομένων χωρίς ετικέτες, δηλαδή δεδομένα που δεν έχουν προηγουμένως κατηγοριοποιηθεί ή επισημανθεί από ανθρώπους. Σε αντίθεση με την επιβλεπόμενη μάθηση, η μη επιβλεπόμενη μάθηση δεν απαιτεί ανθρώπινη παρέμβαση για την εκπαίδευση του μοντέλου. Αντίθετα, το σύστημα μαθαίνει μόνο του από τα δεδομένα, προσπαθώντας να ανακαλύψει κρυφές δομές και σχέσεις μέσα σε αυτά. Αυτή η προσέγγιση βασίζεται αποκλειστικά στα δεδομένα που παρέχονται.

Η μη επιβλεπόμενη μάθηση χρησιμοποιείται ευρέως για την εξαγωγή γενικών χαρακτηριστικών από δεδομένα, τον εντοπισμό σημαντικών τάσεων και δομών, και την ομαδοποίηση δεδομένων σε κατηγορίες για διερευνητικούς σκοπούς. Για παράδειγμα, στην ομαδοποίηση, τα δεδομένα διαχωρίζονται σε ομάδες ή κατηγορίες που μοιράζονται κοινά χαρακτηριστικά. Ένα παράδειγμα

ομαδοποίησης είναι η ανάλυση των αγοραστικών συνηθειών πελατών σε ένα κατάστημα λιανικής. Με τη μη επιβλεπόμενη μάθηση, μπορούμε να ομαδοποιήσουμε τους πελάτες σε κατηγορίες, όπως "τακτικοί αγοραστές", "περιστασιακοί αγοραστές" και "νέοι πελάτες", με βάση τις αγορές που έχουν κάνει.

Άλλοι αλγόριθμοι μη επιβλεπόμενης μάθησης περιλαμβάνουν την εκτίμηση πυκνότητας, τη μάθηση χαρακτηριστικών και τη μείωση διαστάσεων. Επιγραμματικά, η εκτίμηση πυκνότητας αφορά την κατανόηση της κατανομής των δεδομένων στον χώρο χαρακτηριστικών. Η μάθηση χαρακτηριστικών περιλαμβάνει την ανακάλυψη των πιο σημαντικών χαρακτηριστικών που περιγράφουν τα δεδομένα, ενώ η μείωση διαστάσεων απλοποιεί τα δεδομένα διατηρώντας τις πιο κρίσιμες πληροφορίες.

Συνοψίζοντας, η μη επιβλεπόμενη μάθηση βασίζεται στην ανάλυση μη επισημειωμένων δεδομένων (non annotated data) για την ανακάλυψη κρυφών δομών και τάσεων, παρέχοντας πολύτιμες πληροφορίες χωρίς την ανάγκη ανθρώπινης παρέμβασης.

Ημι-Επιβλεπόμενη Μάθηση:

Η ημι-επιβλεπόμενη μάθηση (semi-supervised learning) μπορεί να οριστεί ως ένας υβριδισμός των μεθόδων επιβλεπόμενης και μη επιβλεπόμενης μάθησης, καθώς χρησιμοποιεί τόσο ετικετοποιημένα όσο και μη ετικετοποιημένα δεδομένα. Με άλλα λόγια, βρίσκεται μεταξύ της μάθησης "χωρίς επίβλεψη" και της μάθησης "με επίβλεψη". Στον πραγματικό κόσμο, τα ετικετοποιημένα δεδομένα μπορεί να είναι σπάνια, ενώ τα μη ετικετοποιημένα δεδομένα είναι άφθονα. Σε τέτοιες περιπτώσεις, η ημι-επιβλεπόμενη μάθηση είναι ιδιαίτερα χρήσιμη.

Η κύρια ιδέα της ημι-επιβλεπόμενης μάθησης είναι να εκμεταλλευτεί τα μικρά σύνολα ετικετοποιημένων δεδομένων για να καθοδηγήσει την εκμάθηση, ενώ ταυτόχρονα χρησιμοποιεί τα μεγάλα σύνολα μη ετικετοποιημένων δεδομένων για να βελτιώσει την απόδοση του μοντέλου. Ο τελικός στόχος ενός μοντέλου ημι-επιβλεπόμενης μάθησης είναι να παρέχει καλύτερο αποτέλεσμα πρόβλεψης από αυτό που θα επιτυγχανόταν αν χρησιμοποιούνταν μόνο τα ετικετοποιημένα δεδομένα. Η μέθοδος αυτή επιτρέπει στο σύστημα να μάθει περισσότερα χαρακτηριστικά και σχέσεις μέσα στα δεδομένα, βελτιώνοντας την ακρίβεια και την αξιοπιστία των προβλέψεων του.

Μερικές εφαρμογές όπου χρησιμοποιείται η ημι-επιβλεπόμενη μάθηση περιλαμβάνουν τη μηχανική μετάφραση, την ανίχνευση απάτης, την ετικετοποίηση δεδομένων και την ταξινόμηση κειμένου.

Στη μηχανική μετάφραση ένα σύστημα μπορεί να μάθει να μεταφράζει κείμενα από τη μία γλώσσα στην άλλη χρησιμοποιώντας ένα μικρό αριθμό μεταφρασμένων παραδειγμάτων και πολλά μη μεταφρασμένα κείμενα. Επίσης, στην ανίχνευση απάτης, η ημι-επιβλεπόμενη μάθηση μπορεί να χρησιμοποιηθεί για να εντοπίσει ασυνήθιστες συναλλαγές συνδυάζοντας λίγες ετικετοποιημένες περιπτώσεις απάτης με πολλές κανονικές συναλλαγές. Επιπλέον, μπορεί να χρησιμοποιηθεί για την ετικετοποίηση δεδομένων, όπου ένα μικρό σύνολο ετικετοποιημένων δεδομένων βοηθά στη διαδικασία κατηγοριοποίησης μεγάλων μη ετικετοποιημένων συνόλων, και στην ταξινόμηση κειμένου, όπου τα

ετικετοποιημένα κείμενα βοηθούν στην κατηγοριοποίηση μεγάλων ποσοτήτων μη ετικετοποιημένου κειμένου.

Επομένως, η ημι-επιβλεπόμενη μάθηση προσφέρει έναν αποδοτικό τρόπο να αξιοποιηθούν τα δεδομένα στον πραγματικό κόσμο, συνδυάζοντας τα καλύτερα στοιχεία της επιβλεπόμενης και μη επιβλεπόμενης μάθησης, για να παραχθούν πιο ακριβείς και αξιόπιστες προβλέψεις.

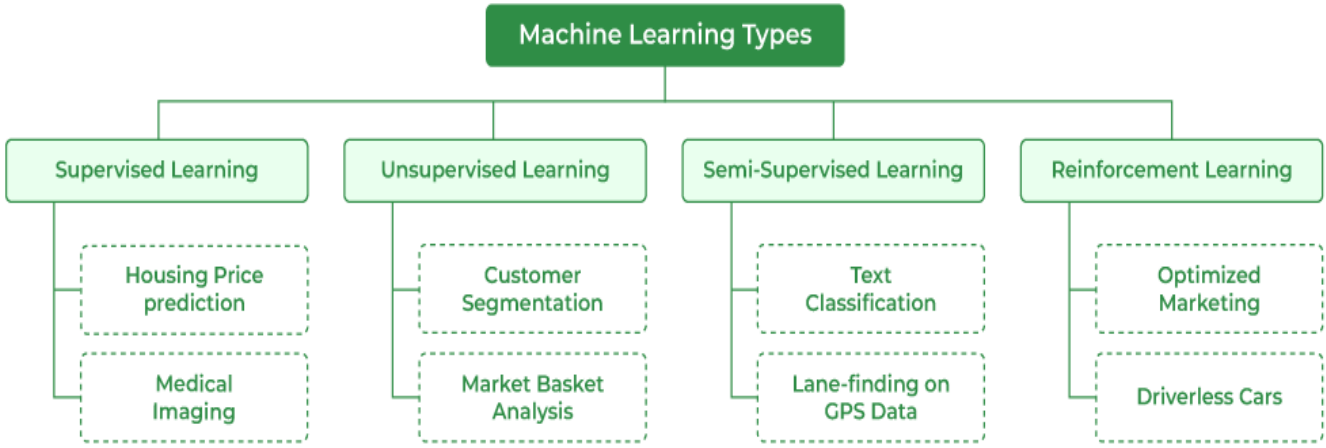
Ενισχυτική Μάθηση:

Η ενισχυτική μάθηση (reinforcement learning) βοηθάει τις μηχανές να μάθουν ποια είναι η βέλτιστη συμπεριφορά σε ένα συγκεκριμένο πλαίσιο ή περιβάλλον, προκειμένου να βελτιώσουν την απόδοσή τους. Αυτή η προσέγγιση βασίζεται στην αλληλεπίδραση του συστήματος με το περιβάλλον του. Το σύστημα λαμβάνει ενέργειες και, ανάλογα με τις συνέπειες αυτών των ενεργειών, λαμβάνει θετική ή αρνητική ανατροφοδότηση. Ο τελικός στόχος είναι να χρησιμοποιήσει τις πληροφορίες που αποκτά από την ανατροφοδότηση αυτή για να βελτιστοποιήσει τις ενέργειές του και να μεγιστοποιήσει τα θετικά αποτελέσματα ή να ελαχιστοποιήσει τα αρνητικά αποτελέσματα.

Η ενισχυτική μάθηση είναι ένα ισχυρό εργαλείο για την εκπαίδευση μοντέλων τεχνητής νοημοσύνης που μπορούν να βοηθήσουν στην αύξηση του αυτοματισμού ή στη βελτιστοποίηση της αποδοτικότητας των σύνθετων συστημάτων. Για παράδειγμα, χρησιμοποιείται ευρέως στη ρομποτική για να διδάξει τα ρομπότ πώς να εκτελούν εργασίες, όπως το να πλοηγούνται σε ένα χώρο ή να συναρμολογούν αντικείμενα. Επίσης, είναι κρίσιμη για τις αυτόνομες εργασίες οδήγησης, όπου τα οχήματα μαθαίνουν να λαμβάνουν αποφάσεις σε πραγματικό χρόνο για να κινούνται με ασφάλεια και αποδοτικότητα. Επιπλέον, η ενισχυτική μάθηση εφαρμόζεται στην παραγωγή και την εφοδιαστική αλυσίδα, βοηθώντας στην βελτιστοποίηση των διαδικασιών.

Ωστόσο, η ενισχυτική μάθηση δεν είναι πάντα η κατάλληλη προσέγγιση για την επίλυση βασικών ή απλών προβλημάτων. Η φύση της, που απαιτεί συνεχή αλληλεπίδραση με το περιβάλλον και μάθηση μέσω δοκιμής και σφάλματος, την καθιστά πιο κατάλληλη για περίπλοκα συστήματα όπου οι βέλτιστες ενέργειες δεν είναι προφανείς από την αρχή.

Επομένως, η ενισχυτική μάθηση προσφέρει έναν αποτελεσματικό τρόπο για τις μηχανές να μάθουν και να βελτιώσουν την απόδοσή τους μέσω αλληλεπίδρασης με το περιβάλλον, καθιστώντας την μια κρίσιμη τεχνολογία για την ανάπτυξη προηγμένων και αυτόνομων συστημάτων.



Εικόνα 2.9: Κατηγοριοποίηση προβλημάτων μηχανικής μάθησης σύμφωνα με τον τρόπο εκμάθησης και παραδείγματα εφαρμογών ανά κατηγορία.

2.2.4 Ανασκόπηση Αλγορίθμων Μηχανικής Μάθησης και Βαθιάς Μάθησης

Ανάλυση ταξινόμησης

Η ταξινόμηση (classification) θεωρείται μια μέθοδος επιβλεπόμενης μάθησης και σχετίζεται με την πρόβλεψη της ετικέτας μιας κλάσης για ένα συγκεκριμένο παράδειγμα δεδομένων. Η διαδικασία αυτή αναφέρεται στον καθορισμό μιας συνάρτησης f που συνδέει τις μεταβλητές εισόδου X με τις μεταβλητές εξόδου Y, οι οποίες αντιπροσωπεύουν τους στόχους, τις ετικέτες ή τις κατηγορίες. Η ταξινόμηση μπορεί να εφαρμοστεί σε δομημένα ή αδόμητα δεδομένα για να προβλέψει την κλάση στην οποία ανήκουν συγκεκριμένα δεδομένα.

Τα δομημένα δεδομένα (structured data) είναι οργανωμένα σε πίνακες με σαφώς καθορισμένες στήλες και γραμμές, όπως τα δεδομένα σε βάσεις δεδομένων ή τα υπολογιστικά φύλλα. Κάθε στήλη αντιπροσωπεύει μια συγκεκριμένη κατηγορία πληροφοριών (όπως ηλικία, ύψος, βάρος), και κάθε γραμμή αντιπροσωπεύει μια καταγραφή ή μια είσοδο.

Από την άλλη πλευρά, τα μη δομημένα δεδομένα (unstructured data) δεν έχουν συγκεκριμένη δομή ή μορφή. Παραδείγματα αδόμητων δεδομένων περιλαμβάνουν κείμενα, εικόνες, βίντεο και ηχητικά αρχεία. Αυτά τα δεδομένα δεν είναι οργανωμένα σε πίνακες και μπορεί να είναι πιο δύσκολο να αναλυθούν και να ταξινομηθούν χωρίς την κατάλληλη επεξεργασία.

Για παράδειγμα, ένα σύστημα ταξινόμησης θα μπορούσε να χρησιμοποιηθεί για να προβλέψει την κατηγορία ενός προϊόντος (δομημένα δεδομένα) ή να αναγνωρίσει το περιεχόμενο μιας φωτογραφίας (αδόμητα δεδομένα).

Στην συνέχεια, συνοψίζουμε τα κοινά προβλήματα ταξινόμησης.

Δυαδική Ταξινόμηση:

Η δυαδική ταξινόμηση (binary αναφέρεται σε εργασίες ταξινόμησης που έχουν δύο ετικέτες κλάσης, όπως "αληθές και ψευδές" ή "ναι και όχι". Σε τέτοιες δυαδικές εργασίες ταξινόμησης, μία κλάση μπορεί να είναι η κανονική κατάσταση, ενώ η άλλη κλάση μπορεί να αντιπροσωπεύει μια ανώμαλη κατάσταση. Για παράδειγμα, σε ένα σύστημα αναγνώρισης εικόνων, το πρόβλημα μπορεί να οριστεί ως "ανίχνευση αυτοκινήτου". Η κανονική κατάσταση μπορεί να είναι "αυτοκίνητο υπάρχει στην εικόνα", ενώ η αντίθετη περίπτωση μπορεί να είναι "οτιδήποτε άλλο εκτός από αυτοκίνητο". Σε αυτό το παράδειγμα το σύστημα πρέπει να αναγνωρίζει αν υπάρχει ένα αυτοκίνητο στην εικόνα, παρόλο που η εικόνα μπορεί να περιέχει πολλά άλλα αντικείμενα που δεν είναι ενδιαφέροντος. Αυτή η προσέγγιση δυαδικής ταξινόμησης εστιάζει στη διάκριση μεταξύ των εικόνων που περιέχουν αυτοκίνητα και όλων των άλλων περιπτώσεων.

Πολυκλασική Ταξινόμηση:

Η πολυκλασική ταξινόμηση (multiclass classification) είναι μια προσέγγιση στη μηχανική μάθηση όπου ένα παράδειγμα ανήκει σε μία και μόνο μία από πολλές προκαθορισμένες κλάσεις. Σε αντίθεση με την δυαδική ταξινόμηση, όπου υπάρχουν μόνο δύο δυνατές κλάσεις (π.χ. θετικό ή αρνητικό), η πολυκλασική ταξινόμηση μπορεί να έχει πολλές κατηγορίες. Για παράδειγμα, στο σύνολο δεδομένων NSL-KDD, μπορούμε να κατηγοριοποιήσουμε διαφορετικούς τύπους επιθέσεων δικτύου σε τέσσερις κλάσεις: DoS (Επίθεση Άρνησης Υπηρεσίας), U2R (Επίθεση Χρήστη σε Ρίζα), R2L (Επίθεση από Ρίζα σε Τοπικό) και Probing Attack (Επίθεση Αναγνώρισης). Κάθε παράδειγμα ταξινομείται σε μία από αυτές τις κλάσεις. Σημαντικό είναι ότι η πολυκλασική ταξινόμηση δεν υποστηρίζει την ανάθεση πολλών ετικετών σε ένα παράδειγμα - κάθε παράδειγμα ανήκει μόνο σε μία κλάση.

Ταξινόμηση πολλαπλών ετικετών:

Η ταξινόμηση πολλαπλών ετικετών (multilabel classification) στη μηχανική μάθηση είναι μια σημαντική προσέγγιση όπου ένα παράδειγμα μπορεί να συνδέεται με πολλές κλάσεις ή ετικέτες ταυτόχρονα. Αυτό αποτελεί μια γενίκευση της πολυκλασικής ταξινόμησης, όπου οι κλάσεις είναι ιεραρχικά δομημένες και κάθε παράδειγμα μπορεί να ανήκει σε περισσότερες από μία κλάσεις σε κάθε επίπεδο της ιεραρχίας, όπως συμβαίνει στην πολυεπίπεδη ταξινόμηση κειμένου.

Για παράδειγμα, οι ειδήσεις της Google μπορούν να ταξινομούνται ταυτόχρονα σε κατηγορίες όπως "όνομα πόλης", "τεχνολογία" και "τελευταίες ειδήσεις". Η ταξινόμηση πολλαπλών ετικετών χρησιμοποιεί προηγμένους αλγορίθμους μηχανικής μάθησης που μπορούν να προβλέπουν πολλές μη αποκλειόμενες ετικέτες για κάθε παράδειγμα, σε αντίθεση με τις παραδοσιακές εργασίες ταξινόμησης όπου κάθε παράδειγμα ανήκει μόνο σε μία κλάση.

Τεχνητό Νευρωνικό Δίκτυο και Βαθιά Μάθηση(Deep Learning):

Η βαθιά μάθηση (Deep Learning - DL) είναι μια υποκατηγορία της μηχανικής μάθησης που βασίζεται σε NN. Τα NN είναι μοντέλα υπολογιστικής μάθησης που χρησιμοποιούνται για την επίλυση προβλημάτων όπως η αναγνώριση προτύπων και η πρόβλεψη μελλοντικών τάσεων. Η αναγνώριση προτύπων μπορεί να περιλαμβάνει την ικανότητα ενός συστήματος να εντοπίζει και να ταξινομεί εικόνες, ήχους ή κείμενο, ενώ η πρόβλεψη μελλοντικών τάσεων αναφέρεται σε προβλήματα όπως είναι οι αλλαγές σε χρηματιστηριακές τιμές ή τα δεδομένα πωλήσεων.

Τα NN είναι μια τεχνολογία εμπνευσμένη από την ανθρώπινη βιολογία, ιδιαίτερα από τη λειτουργία του εγκεφάλου και του νευρικού συστήματος. Αυτά τα δίκτυα προσπαθούν να μιμηθούν τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου, χρησιμοποιώντας στοιχεία επεξεργασίας που ονομάζονται νευρώνες ή Perceptrons. Στη βιολογία, οι νευρώνες συνδέονται μεταξύ τους μέσω συνάψεων, που είναι ειδικές διασυνδέσεις που επιτρέπουν τη μεταφορά σημάτων από έναν νευρώνα σε έναν άλλο. Για παράδειγμα, όταν το χέρι μας αγγίξει ένα ζεστό αντικείμενο, οι αισθητήριοι νευρώνες στέλνουν σήματα μέσω συνάψεων σε ενδιάμεσους νευρώνες και τελικά στους κινητικούς νευρώνες, οι οποίοι

ενεργοποιούν τους μύες για να απομακρύνουν το χέρι. Αυτή η διαδικασία περιλαμβάνει πολλούς νευρώνες που συνεργάζονται, μεταφέροντας και ενισχύοντας τα σήματα κατά μήκος των συνάψεων.

Ομοίως, σε ένα NN, οι τεχνητοί "νευρώνες" συνδέονται και επικοινωνούν μεταξύ τους μέσω "συνδετικών βαρών", που καθορίζουν τη δύναμη ή την επιρροή των συνδέσεων μεταξύ των νευρώνων. Αυτά τα βάρη λειτουργούν όπως οι συνάψεις στον εγκέφαλο, καθορίζοντας τον βαθμό στον οποίο το σήμα από έναν νευρώνα θα συμβάλλει στην ενεργοποίηση του επόμενου νευρώνα.

Η μάθηση στα NN επιτυγχάνεται μέσω της προσαρμογής αυτών των βαρών. Κατά τη διάρκεια της εκπαίδευσης, το δίκτυο συγκρίνει τις προβλεπόμενες εξόδους του με τις πραγματικές τιμές ή στόχους και υπολογίζει το σφάλμα. Στη συνέχεια, χρησιμοποιεί αυτό το σφάλμα για να προσαρμόσει τα βάρη, επιτρέποντας στο δίκτυο να "μαθαίνει" να αντιδρά με συγκεκριμένο τρόπο σε διάφορες εισόδους. Αυτή η διαδικασία προσαρμογής είναι παρόμοια με το πώς το βιολογικό νευρωνικό δίκτυο μαθαίνει και προσαρμόζεται με βάση τις εμπειρίες. Όταν ο εγκέφαλος αντιμετωπίζει νέες καταστάσεις ή πληροφορίες, οι συνάψεις μεταξύ των νευρώνων ενισχύονται ή αποδυναμώνονται, βελτιώνοντας έτσι την ικανότητα του εγκεφάλου να αναγνωρίζει πρότυπα ή να προβλέπει αποτελέσματα με βάση τις παρελθοντικές εμπειρίες.

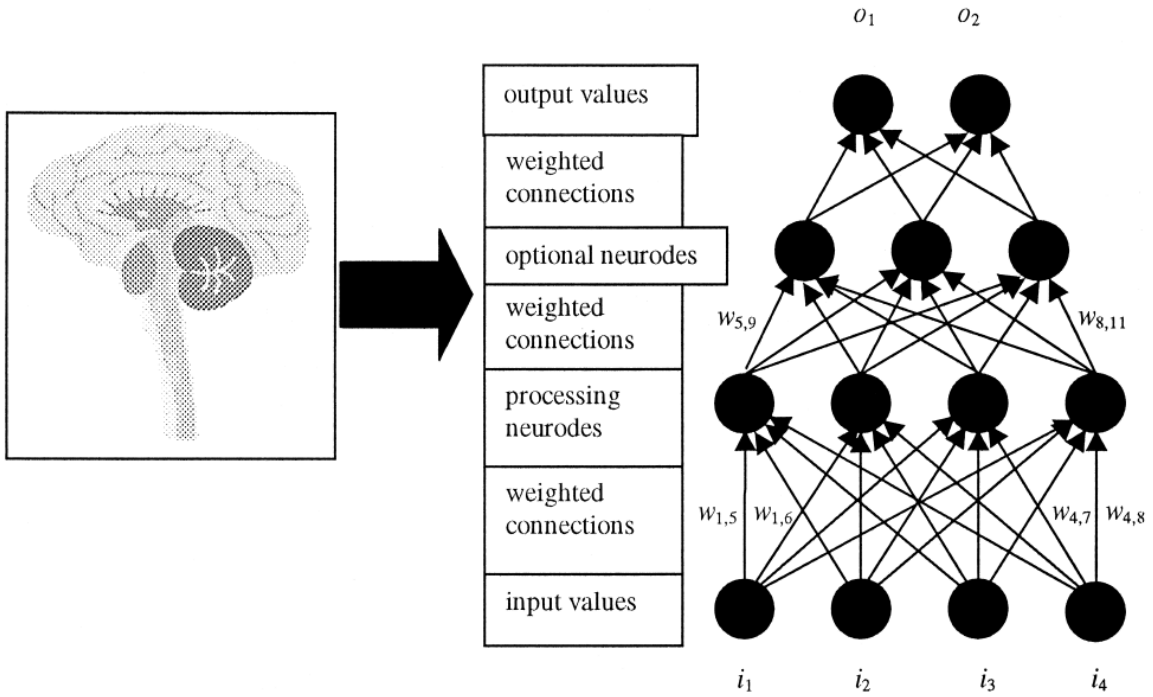


FIGURE 1 Sample artificial neural network architecture (not all weights are shown).

Εικόνα 2.10: Σχηματική αναπαράσταση της γενικής τοπολογίας ενός τεχνητού νευρωνικού δικτύου.

Η εικόνα 2.10 προσφέρει μια οπτική αναπαράσταση του πώς λειτουργούν τα ΝΝ και πώς αυτά μιμούνται τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Δείχνει ένα δείγμα αρχιτεκτονικής ενός ΝΝ, αναδεικνύοντας τη διαδικασία μετάδοσης των πληροφοριών και την επεξεργασία τους, παρόμοια με τον τρόπο που οι νευρώνες και οι συνάψεις συνεργάζονται στον εγκέφαλο.

Ξεκινώντας από τις τιμές εισόδου στην κάτω πλευρά της εικόνας, αυτές αντιπροσωπεύουν τα δεδομένα που εισάγονται στο δίκτυο, είτε πρόκειται για εικόνες, κείμενα ή άλλες μορφές πληροφορίας. Οι είσοδοι αυτές περνούν μέσω των "συνδετικών βαρών" που φαίνονται ως γραμμές σύνδεσης μεταξύ των επιπέδων. Αυτά τα βάρη καθορίζουν πόσο σημαντική είναι κάθε είσοδος για την ενεργοποίηση των επόμενων νευρώνων, παρόμοια με τις συνάψεις στον εγκέφαλο που επηρεάζουν τη μεταφορά των σημάτων μεταξύ των βιολογικών νευρώνων.

Προχωρώντας προς τα πάνω, οι "νευρώνες επεξεργασίας" επεξεργάζονται τις σταθμισμένες εισόδους, υπολογίζοντας μια συνολική τιμή εισόδου. Αυτή η τιμή, στη συνέχεια, περνά μέσω μιας συνάρτησης ενεργοποίησης, η οποία καθορίζει την έξοδο του νευρώνα. Η εικόνα ολοκληρώνεται με τις τιμές εξόδου στην κορυφή, που είναι τα τελικά αποτελέσματα του δικτύου μετά από όλη την επεξεργασία. Αυτές οι έξοδοι μπορεί να είναι σε μορφή κατηγοριοποίησης, προβλέψεων ή άλλων μορφών ανάλυσης, ανάλογα με την εφαρμογή του δικτύου.

Συνολικά, η εικόνα εξηγεί τη διαδικασία επεξεργασίας των πληροφοριών σε ένα ΝΝ, από την είσοδο των δεδομένων μέχρι την έξοδο. Αυτή η αρχιτεκτονική και η μεθοδολογία αποτελούν βασικό συστατικό της βαθιάς μάθησης, επιτρέποντας στα συστήματα να μαθαίνουν από δεδομένα και να βελτιώνουν συνεχώς τις επιδόσεις τους σε ένα ευρύ φάσμα εφαρμογών.Το κύριο πλεονέκτημα της DL σε σχέση με τις παραδοσιακές μεθόδους μηχανικής μάθησης είναι η ανώτερη απόδοσή της, ιδιαίτερα όταν εκπαιδεύεται με μεγάλα σύνολα δεδομένων. Αυτή η δυνατότητα της επιτρέπει να αναγνωρίζει σύνθετα μοτίβα και να βελτιώνει συνεχώς την ακρίβεια των προβλέψεών της.

Εικόνα 2.11: Γραφική απεικόνση της απόδοσης σε συνάρτηση με τον αριθμό των δεδομένων. Παρατηρείται ότι η βαθειά μάθηση (deep learning) δεν παρουσιάζει το ίδιο φαινόμενο κορεσμού όπως η κλασσικέ2ς τεχνικές μηχανικής μάθησης.

Η εικόνα 2.11 δείχνει τη γενική απόδοση της DL σε σχέση με τη ML, λαμβάνοντας υπόψη την αυξανόμενη ποσότητα δεδομένων. Ωστόσο, αυτό μπορεί να διαφέρει ανάλογα με τα χαρακτηριστικά των δεδομένων και την πειραματική ρύθμιση.

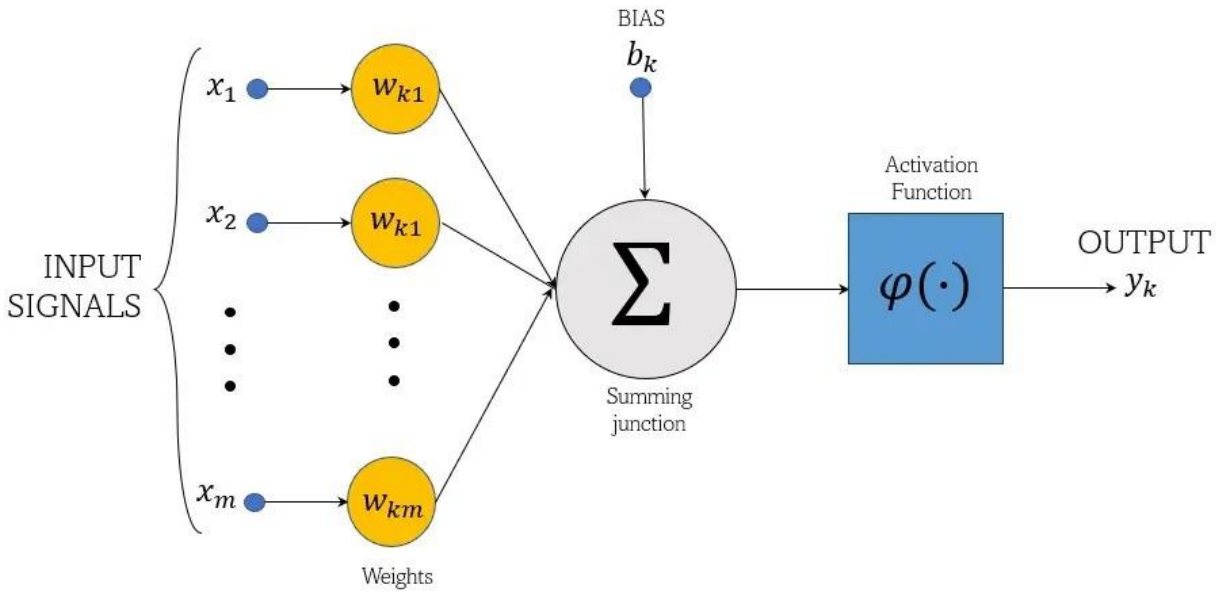
Η DL περιλαμβάνει διάφορες αρχιτεκτονικές, όπως τα πολυεπίπεδα perceptrons (MLP), τα συνελικτικά νευρωνικά δίκτυα (CNN), και τα αναδρομικά νευρωνικά δίκτυα (RNN), με ειδικές παραλλαγές όπως τα Long Short-Term Memory (LSTM). [22]

Multi-layer Perceptron:

Ένα από τα πιο δημοφιλή NN είναι το Multi-Layer Perceptron (MLP). Το MLP αποτελείται από πολλαπλά επίπεδα νευρώνων που διασυνδέονται μεταξύ τους.

Για να κατανοηθεί το MLP, παρέχεται μια σύντομη εισαγωγή στον απλό νευρώνα (one neuron Perceptron), γνωστό και ως νευρώνας με ένα επίπεδο (single layer perceptron). Αυτό αντιπροσωπεύει το απλούστερο μοντέλο ενός NN και περιλαμβάνει μόνο έναν νευρώνα. Αυτός ο νευρώνας έχει μία έξοδο στην οποία είναι συνδεδεμένες όλες οι είσοδοι. Δεδομένων των i = 0,1, ..., m όπου m είναι ο αριθμός των εισόδων, οι ποσότητες {wi} είναι τα βάρη του νευρώνα. Οι είσοδοι {xi} αντιστοιχούν σε χαρακτηριστικά ή μεταβλητές, και η έξοδος y δείχνει σε ποια από τις δύο κατηγορίες ανήκουν τα δεδομένα εισόδου, δηλαδή κάνει μια δυαδική πρόβλεψη. Η διαδικασία υπολογισμού ξεκινά με τον πολλαπλασιασμό της τιμής κάθε χαρακτηριστικού εισόδου xi με το αντίστοιχο wki . Στη συνέχεια, αυτά τα αποτελέσματα

προστίθενται μαζί x0bk+x1wk1 +x2wk2 +⋯+xmwkm. Το τελικό βήμα είναι η εφαρμογή μιας συνάρτησης ενεργοποίησης φ στο άθροισμα, για να παραχθεί μια έξοδος yk .



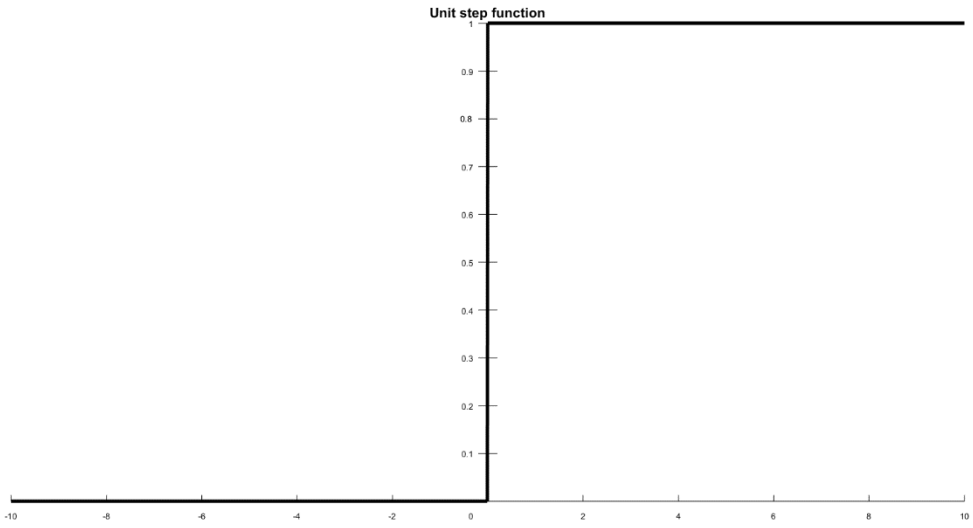
Εικόνα 2.12: Σχηματική αναπαράσταση ενός νευρώνα Perceptron. Κάθε δεδομένο εισόδου πολλαπλασιάζεται με ένα διαφορετικό βάρος (w) και τα επιμέρους γινόμενα αθροίζονται. Το τελικό άθροισμα αποτελεί την είσοδο της μη-γραμμικής συνάρτησης ενεργοποίησης (activation function) και η έξοδος του συγκεκριμένου νευρώνα είναι το αποτέλεσμα της τελευταίας συνάρτησης.

Άρα, η έξοδος yk Ισούται με φ(x0w0+x1wk1 +x2wk2 +⋯+xmwkm) όπου x0=1, w0 κατώφλι ή προκατάληψη(bias) κα yk η έξοδος

Οι συναρτήσεις ενεργοποίησης είναι κρίσιμες στα MLP γιατί επιτρέπουν στο μοντέλο να χειρίζεται μη γραμμικά φαινόμενα, εισάγοντας μη γραμμικότητα στη διαδικασία λήψης αποφάσεων. Με αυτόν τον τρόπο, το perceptron μπορεί να διαχωρίζει δεδομένα που δεν είναι γραμμικά διαχωρίσιμα, επιτρέποντάς του να ορίζει πιο σύνθετα όρια και να κατηγοριοποιεί ή να προβλέπει με μεγαλύτερη ακρίβεια. Χωρίς τις συναρτήσεις ενεργοποίησης, το perceptron θα ήταν περιορισμένο σε απλές γραμμικές διαχωριστικές λειτουργίες. Οι πιο συχνά χρησιμοποιούμενες συναρτήσεις ενεργοποίησης στα perceptrons είναι η Unit step (Heaviside), η Linear και η Logistic (sigmoid).

Unit step (Heaviside):

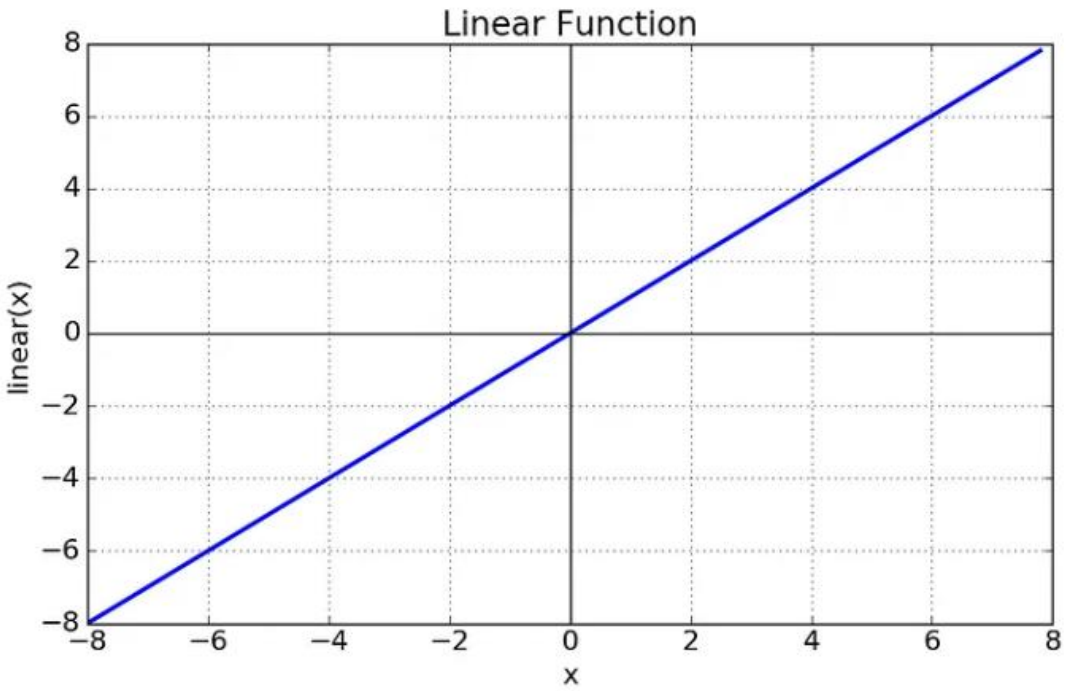
Αυτή η συνάρτηση ενεργοποίησης παράγει έξοδο 0 αν η είσοδος είναι αρνητική και 1 αν η είσοδος είναι θετική ή μηδενική. Είναι χρήσιμη για δυαδική κατηγοριοποίηση, όπου το perceptron πρέπει να αποφασίσει αν ανήκει σε μία από δύο κατηγορίες.



Εικόνα 2.13: Βηματική συνάρτηση (unit step - heavyside). Όλες οι αρνητικές τιμές απεικονίζονται στο 0 ενώ οι μη αρνητικές στο 1.

Linear:

Η γραμμική συνάρτηση ενεργοποίησης παράγει έξοδο που είναι άμεσα αναλογική της εισόδου. Είναι απλή και κατάλληλη όταν οι προβλέψεις χρειάζονται να είναι συνεχείς και αναλογικές των εισόδων. Ωστόσο, δεν μπορεί να εισαγάγει μη γραμμικότητα στο μοντέλο, δηλαδή την ικανότητα να αντιλαμβάνεται και να εκφράζει πιο σύνθετες σχέσεις μεταξύ των εισόδων και της εξόδου. Αυτό περιορίζει την ικανότητά του να χειρίζεται δεδομένα όπου οι σχέσεις δεν είναι ευθείες γραμμές ή απλές αναλογίες, κάτι που συχνά απαιτείται σε πραγματικές εφαρμογές.



Εικόνα 2.14: Γραμμική συνάρτηση. Στο συγκεκριμένο παράδειγμα η συνάρτηση είναι η y = x αλλά τονίζεται ότι μπορεί να χρησιμοποιηθεί κάθε είδους γραμμική σχέση του τύπου y = αx + b

Logistic (sigmoid):

Αυτή η συνάρτηση ενεργοποίησης μετατρέπει την είσοδο σε μια έξοδο που κυμαίνεται μεταξύ 0 και 1, ακολουθώντας μια σιγμοειδή καμπύλη. Είναι ιδιαίτερα χρήσιμη όταν οι έξοδοι πρέπει να ερμηνεύονται ως πιθανότητες ή για προβλήματα κατηγοριοποίησης όπου θέλουμε να καθορίσουμε την πιθανότητα να ανήκει ένα δείγμα σε μια συγκεκριμένη κατηγορία.

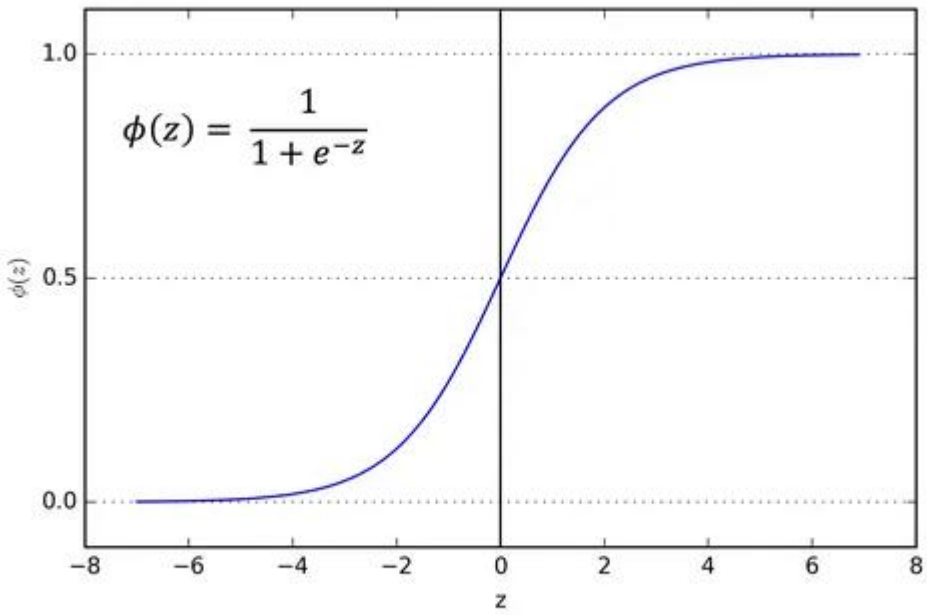


Fig: Sigmoid Function

Σχημα 2.15: Σιγμοειδης (Sigmoid) συνάρτηση. Χρησιμοποιείται για να προσεγγισθούν τατανομές πιθανοτήτων καθώς είναι φραγμένη στο [-1,1]

Ένα perceptron μπορεί να μάθει να διαχωρίζει μόνο γραμμικά διαχωρίσιμες συναρτήσεις, δηλαδή προβλήματα όπου τα δεδομένα μπορούν να χωριστούν σε δύο κατηγορίες χρησιμοποιώντας μια ευθεία γραμμή ή ένα επίπεδο. Για παράδειγμα, μια γραμμική συνάρτηση της μορφής x0w0+x1wk1 +x2wk2 +⋯+xmwkm μπορεί να χρησιμοποιηθεί για να καθορίσει το όριο μεταξύ των δύο κατηγοριών.

Σε έναν δισδιάστατο χώρο, αυτό το όριο εμφανίζεται ως γραμμή που διαχωρίζει τα δεδομένα σε δύο ομάδες. Αντίστοιχα, σε έναν τρισδιάστατο χώρο, η γραμμική συνάρτηση διαμορφώνει ένα επίπεδο που διαχωρίζει τις κατηγορίες. Όταν υπάρχουν περισσότερα χαρακτηριστικά, το όριο διαχωρισμού γίνεται ένα υπερεπίπεδο, το οποίο είναι η γενίκευση της έννοιας της γραμμής και του επιπέδου σε υψηλότερες διαστάσεις. Αυτό το υπερεπίπεδο επιτρέπει στο perceptron να διαχωρίζει τις κατηγορίες δεδομένων με βάση τις τιμές των χαρακτηριστικών τους, διασφαλίζοντας ότι κάθε κατηγορία τοποθετείται σε διαφορετική πλευρά του υπερεπίπεδου.

Κατά τη διαδικασία εκπαίδευσης ενός perceptron, χρησιμοποιούνται δεδομένα εισόδου με γνωστές αποκρίσεις, γνωστά ως δεδομένα εκπαίδευσης, για να "μάθει" το μοντέλο. Η μάθηση περιλαμβάνει τη ρύθμιση των βαρών του δικτύου έτσι ώστε να ελαχιστοποιηθεί η διαφορά μεταξύ των προβλέψεων του μοντέλου και των πραγματικών τιμών, δηλαδή το σφάλμα. Συνήθως, αυτή η διαφορά μετριέται ως το τετραγωνικό σφάλμα, που είναι το τετράγωνο της διαφοράς μεταξύ της προβλεπόμενης και της πραγματικής απόκρισης.

Για να επιτευχθεί η βέλτιστη ρύθμιση των βαρών, χρησιμοποιείται η τεχνική της καθόδου κλίσης (gradient descent). Αυτή η τεχνική αναλύει το σφάλμα και προσαρμόζει τα βάρη έτσι ώστε το σφάλμα να ελαχιστοποιείται σταδιακά. Ουσιαστικά, ο αλγόριθμος "μαθαίνει" ποια βάρη οδηγούν στις καλύτερες προβλέψεις, και προσαρμόζει αυτά τα βάρη μέχρι να φτάσει σε ένα σημείο όπου το σφάλμα δεν μπορεί να μειωθεί περαιτέρω, δηλαδή το μοντέλο έχει φτάσει σε μια λειτουργική διαμόρφωση.

Μετά την εκπαίδευση, είναι σημαντικό να ελεγχθεί πώς το μοντέλο ανταποκρίνεται σε νέα, άγνωστα κατά την διάρκεια εκπαίδευσης δεδομένα. Αυτό γίνεται μέσω της επικύρωσης, χρησιμοποιώντας ένα σύνολο δεδομένων που δεν έχει χρησιμοποιηθεί κατά την εκπαίδευση. Αυτή η διαδικασία βοηθάει να διαπιστωθεί αν το μοντέλο μπορεί να γενικεύσει τις προβλέψεις του και σε άλλες περιπτώσεις πέρα από αυτές που "είδε" κατά την εκπαίδευση. Αν το μοντέλο τα πάει καλά με τα νέα δεδομένα, τότε θεωρείται επιτυχές και ικανό να εφαρμοστεί σε πραγματικές καταστάσεις.

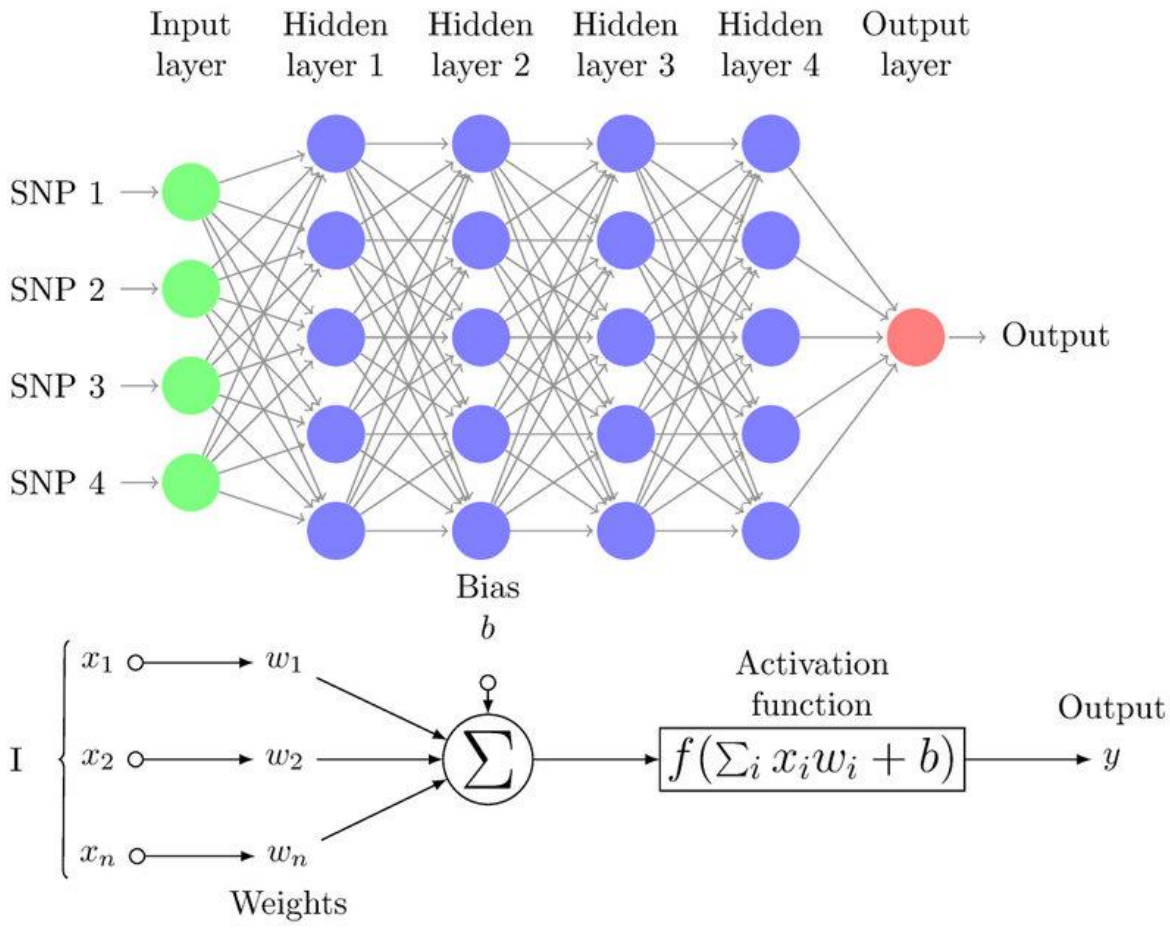
Η παράλληλη σύνδεση πολλών perceptrons δημιουργεί μια αρχιτεκτονική με ένα επίπεδο perceptron (Single Layer Perceptron - SLP), η οποία χρησιμοποιείται στην περίπτωση πολλαπλών εξόδων. Για παράδειγμα, ας υποθέσουμε ότι έχουμε ένα σύστημα που λαμβάνει ως είσοδο τρεις παραμέτρους: την ηλικία, το εισόδημα και την εκπαίδευση ενός ατόμου. Το SLP μπορεί να χρησιμοποιηθεί για να ταξινομήσει αυτά τα άτομα σε τρεις διαφορετικές κατηγορίες: αν είναι πιθανό να αγοράσουν ένα προϊόν, αν θα εγγραφούν σε μια υπηρεσία και αν θα συμμετάσχουν σε μια έρευνα.

Σε αυτό το παράδειγμα, οι είσοδοι (ηλικία, εισόδημα, εκπαίδευση) τροφοδοτούνται στο SLP. Κάθε perceptron στο μοναδικό κρυφό επίπεδο επεξεργάζεται αυτά τα δεδομένα και παράγει μια έξοδο που αντιπροσωπεύει την πιθανότητα του ατόμου να ανήκει σε μία από τις κατηγορίες. Έτσι, το σύστημα μπορεί να έχει τρεις εξόδους, όπου κάθε έξοδος δείχνει την απόφαση για μια συγκεκριμένη κατηγορία.

Ωστόσο, το απλό perceptron και το Single Layer Perceptron (SLP) έχουν ένα σημαντικό περιορισμό: δεν μπορούν να επιλύσουν προβλήματα όπου τα δεδομένα δεν μπορούν να διαχωριστούν με μια ευθεία γραμμή ή ένα επίπεδο. Αυτό σημαίνει ότι αν τα δεδομένα σου ανήκουν σε δύο κατηγορίες που δεν διαχωρίζονται γραμμικά, το μοντέλο δεν θα καταφέρει να βρει ένα σωστό όριο διαχωρισμού.

Για παράδειγμα, μπορεί να εξεταστεί η περίπτωση όπου πρέπει να διαχωριστούν δύο κατηγορίες σημείων σε έναν δισδιάστατο χώρο, με τα σημεία της κάθε κατηγορίας να είναι κατανεμημένα κυκλικά. Ειδικότερα, τα σημεία της πρώτης κατηγορίας βρίσκονται μέσα σε έναν κύκλο, ενώ τα σημεία της δεύτερης κατηγορίας είναι τοποθετημένα έξω από αυτόν. Σε αυτή την περίπτωση, δεν είναι εφικτή η χρήση μιας ευθείας γραμμής για τον διαχωρισμό των σημείων, καθώς ο διαχωρισμός αυτός δεν είναι γραμμικός.

Για την αντιμετώπιση αυτού του περιορισμού, χρησιμοποιείται η αρχιτεκτονική Multilayer Perceptron (MLP), η οποία προσθέτει περισσότερα επίπεδα στο δίκτυο.



Εικόνα 2.16: Πολυστρωματικό δίκτυο Perceptron (Multi-Layer Perceptron - MLP) και ανάλυση του κάθε νευρώνα στο κάτω μέρος της εικόνας.

Το MLP είναι ένας τύπος NN με πολλαπλά επίπεδα, όπου η πληροφορία προωθείται από το ένα επίπεδο στο άλλο με μία κατεύθυνση, από την είσοδο στην έξοδο. Αυτή η δομή εξασφαλίζει ότι οι εισαγωγικές πληροφορίες επεξεργάζονται διαδοχικά μέσα από διάφορα επίπεδα κρυφών νευρώνων, που λειτουργούν ως ενδιάμεσοι επεξεργαστές δεδομένων.

Κάθε σύνδεση μεταξύ των νευρώνων σε ένα MLP έχει ένα συγκεκριμένο βάρος, που καθορίζει τη σημασία του σήματος που μεταφέρεται από έναν νευρώνα στον επόμενο. Αυτά τα βάρη προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης του δικτύου για να βελτιώσουν την απόδοση του μοντέλου στις προβλέψεις.

Σε κάθε επίπεδο του MLP, οι νευρώνες χρησιμοποιούν την ίδια συνάρτηση ενεργοποίησης, η οποία καθορίζει πώς θα επεξεργαστούν τα σήματα. Συνήθως, τα κρυφά επίπεδα χρησιμοποιούν τη σιγμοειδή συνάρτηση ενεργοποίησης, που συμπιέζει τα σήματα σε μια κλίμακα μεταξύ 0 και 1, επιτρέποντας στο δίκτυο να εισάγει μη γραμμικότητα και να αναγνωρίζει πιο σύνθετα πρότυπα στα δεδομένα.

Το επίπεδο εξόδου μπορεί να χρησιμοποιεί είτε σιγμοειδή συνάρτηση είτε γραμμική, ανάλογα με την εφαρμογή. Για παράδειγμα, σε προβλήματα ταξινόμησης, η σιγμοειδής συνάρτηση μπορεί να χρησιμοποιηθεί για να δώσει πιθανότητες κατηγοριοποίησης, ενώ σε προβλήματα πρόβλεψης συνεχών τιμών, όπως η πρόβλεψη τιμής μετοχών, μπορεί να χρησιμοποιηθεί μια γραμμική συνάρτηση για να παράγει μια ευρύτερη κλίμακα εξόδων. Αυτή η ευελιξία επιτρέπει στο MLP να προσαρμόζεται σε μια ποικιλία εφαρμογών, από αναγνώριση φωνής και εικόνας μέχρι χρηματοοικονομικές προβλέψεις.

Ο αλγόριθμος back-propagation είναι ίσως ο πιο γνωστός και ευρέως χρησιμοποιούμενος αλγόριθμος μάθησης για τα MLP. Αυτός ο αλγόριθμος επιτρέπει την εκπαίδευση των NN διορθώνοντας τα βάρη τους ώστε να μειωθούν τα σφάλματα στις προβλέψεις. Η διαδικασία βασίζεται στην αρχή των Ελαχίστων Τετραγωνικών, η οποία επιδιώκει να ελαχιστοποιήσει το τετραγωνικό σφάλμα μεταξύ της πραγματικής και της προβλεπόμενης εξόδου.

Το χαρακτηριστικό που διακρίνει τον αλγόριθμο back-propagation είναι η μέθοδος που χρησιμοποιεί για να ενημερώσει τα βάρη του δικτύου. Ξεκινά με τον υπολογισμό του σφάλματος στο επίπεδο εξόδου, δηλαδή τη διαφορά μεταξύ της πραγματικής εξόδου και της εξόδου που προβλέπει το δίκτυο. Στη συνέχεια, αυτό το σφάλμα "διαδίδεται" πίσω στα προηγούμενα επίπεδα του δικτύου, διορθώνοντας τα βάρη σε κάθε επίπεδο με τρόπο που να μειώνει το συνολικό σφάλμα.

Η διαδικασία αυτή γίνεται σε κάθε επίπεδο, από την έξοδο προς την είσοδο, και είναι αυτή η αντίστροφη διάδοση των σφαλμάτων που δίνει στο back-propagation το όνομά του. Η προσέγγιση αυτή επιτρέπει στο MLP να "μαθαίνει" από τα σφάλματά του και να βελτιώνει τις προβλέψεις του, προσαρμόζοντας τα βάρη του για να ταιριάζουν καλύτερα στα δεδομένα εκπαίδευσης. Αυτή η ικανότητα προσαρμογής καθιστά τον back-propagation βασικό αλγόριθμο για την εκπαίδευση των MLP.

Το MLP είναι ένα ισχυρό εργαλείο για την ανάλυση και πρόβλεψη σύνθετων σχέσεων, αλλά η απόδοσή του εξαρτάται από διάφορους παράγοντες. Οι βασικοί παράγοντες περιλαμβάνουν την επιλογή των κατάλληλων μεταβλητών, τον αριθμό των κρυφών επιπέδων και των νευρώνων (κόμβων) σε αυτά τα επίπεδα, καθώς και την ποιότητα των δεδομένων εκπαίδευσης. Επιπλέον, οι παράμετροι εκπαίδευσης όπως ο ρυθμός μάθησης και ο συντελεστής ορμής παίζουν κρίσιμο ρόλο στη διαδικασία μάθησης. Ο ρυθμός μάθησης καθορίζει το πόσο γρήγορα το μοντέλο προσαρμόζει τα βάρη του, ενώ ο συντελεστής ορμής βοηθά στην εξομάλυνση αυτών των αλλαγών, αποτρέποντας μεγάλες και απότομες προσαρμογές.

Συνοψίζοντας, τα MLP αποτελούν θεμελιώδη εργαλεία στην τεχνητή νοημοσύνη και τη μηχανική μάθηση, επιτρέποντας την ανάλυση και πρόβλεψη σύνθετων και μη γραμμικών σχέσεων. Η ευελιξία τους να διαχειρίζονται διάφορες εισόδους και να προσαρμόζουν τα βάρη μέσω διαδικασιών όπως το back-propagation, τα καθιστά εξαιρετικά για πλήθος εφαρμογών, από τη μοντελοποίηση αλλαγών γης μέχρι την αναγνώριση φωνής και εικόνας. Παρά τους κινδύνους υπερπροσαρμογής (overfitting) και την εξάρτηση από τη σωστή επιλογή παραμέτρων όπως ο ρυθμός μάθησης και ο συντελεστής ορμής, τα MLP παραμένουν ανεκτίμητα για την ικανότητά τους να γενικεύουν και να προσφέρουν ακριβείς προβλέψεις σε νέα δεδομένα.[23]

Συνελικτικό Νευρωνικό Δίκτυο (Convolutional Neural Network - CNN):

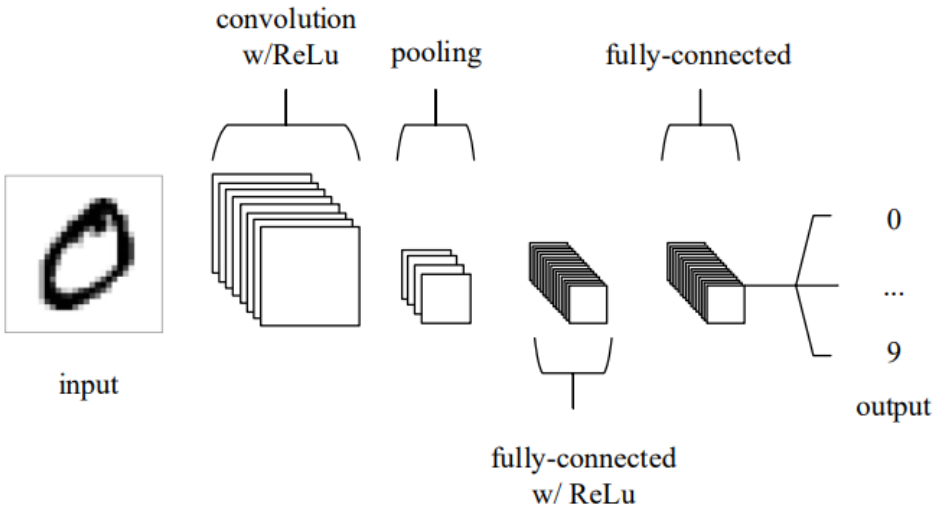
Μια από τις πιο εντυπωσιακές μορφές αρχιτεκτονικής NN είναι το Συνελικτικό Νευρωνικό Δίκτυο (Convolutional Neural Network - CNN). Τα CNNs χρησιμοποιούνται κυρίως για την επίλυση προβλημάτων αναγνώρισης προτύπων που βασίζονται σε εικόνες. Διαθέτουν μια απλή και ακριβή αρχιτεκτονική. Παρόμοια με τα παραδοσιακά NN, τα CNNs αποτελούνται από νευρώνες που βελτιστοποιούνται μέσω της μάθησης. Κάθε νευρώνας λαμβάνει είσοδο και εκτελεί μια λειτουργία (όπως το εσωτερικό γινόμενο ακολουθούμενο από μια μη γραμμική συνάρτηση), κάτι που αποτελεί τη βάση για πολλά NN.

Η κύρια διαφορά μεταξύ CNNs και παραδοσιακών NN είναι η εξειδίκευση των CNNs στην ανάλυση και αναγνώριση προτύπων μέσα από εικόνες. Τα CNNs χρησιμοποιούν διάφορα επίπεδα, όπως τα συνελικτικά επίπεδα και τα επίπεδα δειγματοληψίας (γνωστά και ως pooling layers), τα οποία συμβάλλουν στην εξαγωγή και την ενίσχυση σημαντικών χαρακτηριστικών από τις εικόνες. Αυτό τα καθιστά ιδανικά για εφαρμογές όπως η αναγνώριση αντικειμένων και προσώπων, η ταξινόμηση εικόνων και άλλες σχετικές λειτουργίες.

Εν συντομία, ενώ τα παραδοσιακά NN μπορούν να εφαρμοστούν σε διάφορους τομείς, τα CNNs προσφέρουν μια εξειδικευμένη προσέγγιση για εργασίες που απαιτούν επεξεργασία εικόνας, κάνοντάς τα αναπόσπαστο μέρος της βαθιάς μάθησης.

Γενική Αρχιτεκτονική των CNNs:

Η γενική αρχιτεκτονική των CNNs αποτελείται από τρία κύρια είδη επιπέδων: τα συνελικτικά επίπεδα(convolutional layers), που εξάγουν χαρακτηριστικά από την είσοδο, τα επίπεδα δειγματοληψίας (pooling layers), που μειώνουν τη διάσταση των δεδομένων και διατηρούν τα σημαντικά χαρακτηριστικά, και τα πλήρως συνδεδεμένα επίπεδα, που παράγουν τις τελικές προβλέψεις για τις κατηγορίες. Αυτά τα επίπεδα συνδυάζονται για να σχηματίσουν τη δομή των CNNs, η οποία είναι ιδιαίτερα αποτελεσματική στην επεξεργασία και αναγνώριση εικόνων.

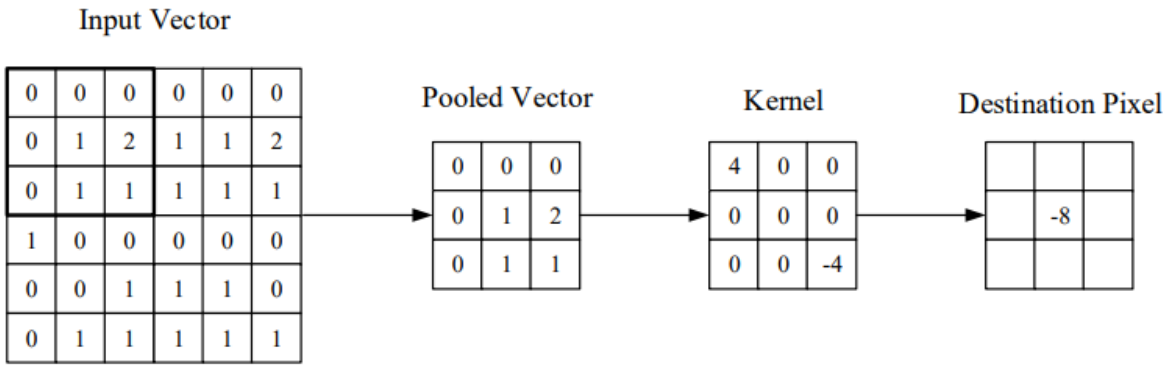


Εικόνα 2.17: Σχηματική απεικόνιση ενός συνελικτικού δικτύου και των επιμέρους στοιχείων που το αποτελούν. Για αρχή, παράγονται τα συνελικτικά χαρακτηριστικά μέσω συνέλιξης της εικόνας (ή άλλων χαρακτηριστικών), γίνεται υποδειγματοληψία μέσω του μηχανισμού pooling και τέλος, χρησιμοποιείται ένα Multi-Layer Perceptron με την τελική στοιβάδα εξόδου να έχει ίδιο αριθμό νευρώνων με τον αριθμό των κλάσεων ενδιαφέροντος

Convolutional_layer:

Το συνελικτικό επίπεδο (convolutional layer) στα CNNs είναι ένα από τα πιο κρίσιμα στοιχεία της αρχιτεκτονικής τους, καθώς είναι υπεύθυνο για την εξαγωγή χαρακτηριστικών από την είσοδο. Το κεντρικό στοιχείο του συνελικτικού επιπέδου είναι τα φίλτρα (kernels), τα οποία είναι μικρά σε μέγεθος όσον αφορά τις χωρικές διαστάσεις (όπως το ύψος και το πλάτος), αλλά εκτείνονται σε όλο το βάθος της εισόδου.

Καθώς τα δεδομένα περνούν από το συνελικτικό επίπεδο, κάθε φίλτρο εξετάζει διαφορετικά τμήματα της εικόνας.



Εικόνα 2.18: Η διαδιακσία min pooling με παράθυρο 3x3. Όπως παρατηρείται από το διάνυσμα εισόδου (input vector), εφαρμόζεται ένα παράθυρο και επιλέγεται η ελάχιστη τιμή του παραθύρου. Έπειτα παρουσιάζεται η συνέλιξη, όπου κάθε τιμή του δειγματοληπτημένου διανύσματος (pooled vector) πολλαπλασιάζεται με την τιμή του πίνακα kernel αντίστοιχης θέσης. Έπειτα, τα γινόμενα αθροίζονται και το τελικό αποτέλεσμα αποθηκεύται σαν μια νέα τιμή του διανύσματος χαρακτηριστικών εξόδου.

Η εικόνα 2.18 δείχνει την εφαρμογή ενός φίλτρου σε ένα συγκεκριμένο σημείο της εικόνας. Το φίλτρο, που φαίνεται ως ένας μικρός πίνακας αριθμών, εφαρμόζεται σε μια τοπική περιοχή της εικόνας, όπως αυτή που φαίνεται στον "Input Vector". Κατά την εφαρμογή, υπολογίζεται το εσωτερικό γινόμενο μεταξύ των τιμών των pixels στην περιοχή αυτή και των αντίστοιχων βαρών του φίλτρου. Αυτός ο υπολογισμός δημιουργεί έναν "Destination Pixel" που αντικατοπτρίζει την απόκριση του φίλτρου σε αυτή την περιοχή. Αυτό το βήμα επαναλαμβάνεται καθώς το φίλτρο μετακινείται σε όλη την εικόνα, εφαρμόζοντας τη διαδικασία σε κάθε σημείο. Η συνεχής μετακίνηση και εφαρμογή του φίλτρου δημιουργεί έναν χάρτη ενεργοποίησης (activation map).

Ο χάρτης ενεργοποίησης αντιπροσωπεύει τα σημεία της εικόνας όπου το φίλτρο "ενεργοποιείται" περισσότερο, δηλαδή τα σημεία όπου ανιχνεύονται τα χαρακτηριστικά για τα οποία το φίλτρο είναι προσαρμοσμένο. Για παράδειγμα, ένα φίλτρο σχεδιασμένο να ανιχνεύει κάθετες γραμμές θα έχει υψηλές τιμές στα σημεία όπου υπάρχουν κάθετες γραμμές στην εικόνα. Κάθε φίλτρο στο συνελικτικό επίπεδο παράγει τον δικό του χάρτη ενεργοποίησης, και όλοι αυτοί οι χάρτες μαζί συνθέτουν την έξοδο του συνελικτικού επιπέδου.

Ένα συνελικτικό επίπεδο περιλαμβάνει συνήθως αρκετά φίλτρα, και κάθε ένα από αυτά τα φίλτρα έχει σχεδιαστεί για να ανιχνεύει διαφορετικά χαρακτηριστικά της εικόνας. Καθώς το κάθε φίλτρο εφαρμόζεται διαδοχικά σε όλη την εικόνα, παράγεται ένας χάρτης ενεργοποίησης για κάθε φίλτρο. Αυτοί οι χάρτες ενεργοποίησης καταγράφουν τα σημεία της εικόνας όπου το φίλτρο ανιχνεύει συγκεκριμένα χαρακτηριστικά, όπως άκρα, γωνίες ή άλλα μοτίβα.

Ο συνδυασμός αυτών των χαρτών ενεργοποίησης από όλα τα φίλτρα συνθέτει την πλήρη έξοδο του συνελικτικού επιπέδου. Κάθε χάρτης ενεργοποίησης παρέχει μια ξεχωριστή διάσταση δεδομένων που περιέχει πληροφορίες σχετικά με την απόκριση του φίλτρου σε διαφορετικά σημεία της εικόνας. Έτσι, το συνελικτικό επίπεδο παράγει ένα "σύνολο εξόδων" που περιλαμβάνει πολλαπλούς χάρτες ενεργοποίησης, έναν για κάθε φίλτρο, οι οποίοι παρέχουν μια λεπτομερή αναπαράσταση των χαρακτηριστικών της εικόνας που ανιχνεύθηκαν.

Αυτό το σύνολο εξόδων αποτελεί το αποτέλεσμα της επεξεργασίας του συνελικτικού επιπέδου, παρέχοντας πλούσιες πληροφορίες για τη δομή και τα μοτίβα της εικόνας, οι οποίες χρησιμοποιούνται στα επόμενα επίπεδα του CNN για περαιτέρω επεξεργασία και ανάλυση.

Επίπεδο Δειγματοληψίας(Pooling Layer):

Το επίπεδο δειγματοληψίας έχει σχεδιαστεί για να μειώνει τις διαστάσεις των δεδομένων που περνούν από το δίκτυο. Λειτουργεί πάνω από κάθε χάρτη ενεργοποίησης που παράγεται από τα προηγούμενα συνελικτικά επίπεδα και μειώνει τις διαστάσεις τους χρησιμοποιώντας τη λειτουργία "MAX".

Στην περίπτωση του max-pooling, το οποίο είναι η πιο κοινή μορφή pooling, χρησιμοποιούνται μικρά φίλτρα συνήθως διαστάσεων 2 × 2. Αυτά τα φίλτρα κινούνται κατά μήκος της εικόνας με βήμα (stride) 2, μειώνοντας το μέγεθος του χάρτη ενεργοποίησης στο 25% του αρχικού. Αν και αυτή η διαδικασία είναι καταστροφική με την έννοια ότι μειώνει την ποσότητα των δεδομένων, διατηρεί τα σημαντικότερα χαρακτηριστικά και τις πληροφορίες που χρειάζονται για την αναγνώριση και κατηγοριοποίηση.

Εκτός από το max-pooling, τα CNNs μπορεί να περιλαμβάνουν και άλλες μορφές δειγματοληψίας, γνωστές ως γενική δειγματοληψία (general pooling). Αυτές περιλαμβάνουν τεχνικές όπως η κανονικοποίηση L1/L2 και το average pooling, που λαμβάνουν τον μέσο όρο των τιμών των pixels αντί να επιλέγουν την μέγιστη τιμή. Ωστόσο, το max-pooling είναι η πιο διαδεδομένη τεχνική.

Συνοψίζοντας, το επίπεδο δειγματοληψίας εφαρμόζεται στα CNNs για να επιτύχει δύο κύριους στόχους. Πρώτον, μειώνει τη χωρική διάσταση των δεδομένων που εισάγονται στο δίκτυο, μειώνοντας έτσι τον αριθμό των παραμέτρων και την υπολογιστική πολυπλοκότητα του μοντέλου. Αυτό καθιστά το δίκτυο πιο αποδοτικό και ταχύτερο στην επεξεργασία δεδομένων, κάτι που είναι ιδιαίτερα σημαντικό όταν το μοντέλο πρέπει να χειριστεί μεγάλους όγκους δεδομένων, όπως μεγάλες εικόνες ή βίντεο. Δεύτερον, το επίπεδο δειγματοληψίας βοηθά στη μείωση της ευαισθησίας του δικτύου σε μικρές μετατοπίσεις ή παραμορφώσεις των χαρακτηριστικών της εικόνας. Με την εξαγωγή των πιο σημαντικών χαρακτηριστικών και την απομάκρυνση των λιγότερο σημαντικών πληροφοριών, το δίκτυο μπορεί να γενικεύσει καλύτερα, αναγνωρίζοντας πρότυπα ανεξάρτητα από τις μικρές διαφοροποιήσεις στις εικόνες εισόδου. Αυτή η διαδικασία συμβάλλει στην αύξηση της ακρίβειας και της σταθερότητας του μοντέλου, κάνοντάς το πιο ικανό να αναγνωρίζει και να κατηγοριοποιεί χαρακτηριστικά σε νέα δεδομένα.

Πλήρως Συνδεδεμένο Επίπεδο (Fully Connected Layer):

Το πλήρως συνδεδεμένο επίπεδο, γνωστό και ως fully-connected layer ή dense layer, αποτελεί ένα από τα τελευταία στάδια σε ένα CNN κι είναι ένας ακόμα τρόπος να περιγραφεί ένα MLP. Σε αυτό το επίπεδο, κάθε νευρώνας συνδέεται με κάθε νευρώνα του προηγούμενου και του επόμενου επιπέδου, δημιουργώντας πλήρη διασύνδεση μεταξύ των επιπέδων. Αυτός ο τρόπος σύνδεσης είναι παρόμοιος με τις παραδοσιακές μορφές τεχνητών NN, όπου κάθε επίπεδο είναι πλήρως συνδεδεμένο με το επόμενο.

Η κύρια λειτουργία του πλήρως συνδεδεμένου επιπέδου είναι να συνδυάζει τα χαρακτηριστικά που εξήχθησαν από τα προηγούμενα επίπεδα και να τα χρησιμοποιεί για να παράγει την τελική πρόβλεψη ή κατηγοριοποίηση. Καθώς αυτό το επίπεδο συνδέει όλους τους νευρώνες μεταξύ των επιπέδων, επιτρέπει τη μάθηση από συνδυασμούς χαρακτηριστικών, παρέχοντας έτσι στο μοντέλο τη δυνατότητα να αναγνωρίζει πιο σύνθετα μοτίβα και συσχετισμούς στα δεδομένα.

Σε ένα CNN, το πλήρως συνδεδεμένο επίπεδο συχνά τοποθετείται μετά από τα συνελικτικά και τα επίπεδα δειγματοληψίας, που έχουν ήδη εξάγει και επιλέξει τα πιο σημαντικά χαρακτηριστικά της εισόδου. Το πλήρως συνδεδεμένο επίπεδο αξιοποιεί αυτές τις πληροφορίες για να κάνει τελικές προβλέψεις, όπως την κατηγοριοποίηση μιας εικόνας σε διαφορετικές κατηγορίες.

Συνοψίζοντας, τα CNNs παρά το γεγονός ότι μοιράζονται ομοιότητες με τα παραδοσιακά NN, παρουσιάζουν σημαντικές διαφορές στην αρχιτεκτονική τους και στις εφαρμογές τους. Αυτές οι διαφορές καθιστούν τα CNNs ιδιαίτερα κατάλληλα για την επεξεργασία και ανάλυση εικόνων, λόγω της ικανότητάς τους να ανιχνεύουν και να αναγνωρίζουν περίπλοκα χαρακτηριστικά στα δεδομένα εικόνας. Για όσους ενδιαφέρονται να κατανοήσουν τη βέλτιστη δομή και τις πρακτικές εφαρμογές των CNNs, συνιστάται να ανατρέξουν στο παρακάτω άρθρο, το οποίο παρέχει εκτενείς οδηγίες και παραδείγματα για τη δημιουργία και βελτιστοποίηση αυτών των ισχυρών αλγορίθμων.[24]

2.2.5 Υπερπαράμετροι

Οι υπερπαράμετροι αποτελούν ένα σύνολο παραμέτρων που καθορίζουν τη διαδικασία εκπαίδευσης ενός μοντέλου μηχανικής μάθησης. Σε αντίθεση με τις παραμέτρους του μοντέλου που προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης, οι υπερπαράμετροι καθορίζονται πριν την έναρξη της εκπαίδευσης και παραμένουν σταθερές καθ' όλη τη διάρκειά της. Η ρύθμιση των υπερπαραμέτρων είναι κρίσιμη, τόσο για μοντέλα με περίπλοκη δομή που απαιτούν πολλές υπερπαραμέτρους, όσο και για μοντέλα με προσεκτικά σχεδιασμένη δομή, όπου κάθε υπερπαράμετρος πρέπει να ρυθμιστεί με ακρίβεια για να επιτευχθεί η επιθυμητή απόδοση.

Η χειροκίνητη ρύθμιση των υπερπαραμέτρων είναι εφικτή και συχνά βασίζεται στην εμπειρία και στη γνώση από προηγούμενες εργασίες. Ωστόσο, για μεγαλύτερα ή πιο πρόσφατα δημοσιευμένα μοντέλα, η ποικιλία των επιλογών υπερπαραμέτρων απαιτεί σημαντική προσπάθεια, χρόνο και υπολογιστικούς πόρους από τους ερευνητές, καθώς και πολλές δοκιμές και σφάλματα. Ακόμα και για έμπειρους

ερευνητές, η αναζήτηση και εφαρμογή των κατάλληλων υπερπαραμέτρων για τη δημιουργία ενός αποδοτικού μοντέλου δεν είναι εύκολη υπόθεση.

Οι χρήστες με λιγότερη εμπειρία έχουν μεγάλη ανάγκη από προτεινόμενες υπερπαραμέτρους και έτοιμα προς χρήση εργαλεία αυτόματης βελτιστοποίησης υπερπαραμέτρων (Automatic Hyperparameter Optimization - AHPO). Για να μειωθούν τα τεχνικά εμπόδια, η AHPO έχει γίνει ένα δημοφιλές θέμα τόσο στον ακαδημαϊκό όσο και στον βιομηχανικό τομέα. Οι αυτοματοποιημένες υπηρεσίες και τα εργαλεία ρύθμισης υπερπαραμέτρων που έχουν αναπτυχθεί από διάφορους ερευνητές και εταιρείες, προσφέρουν λύσεις που μειώνουν την ανάγκη για χειροκίνητο σχεδιασμό και διευκολύνουν την εφαρμογή της βαθιάς μάθησης.

Οι υπερπαράμετροι μπορούν να κατηγοριοποιηθούν σε δύο ομάδες: αυτές που χρησιμοποιούνται για την εκπαίδευση και αυτές που χρησιμοποιούνται για τον σχεδιασμό του μοντέλου. Η κατάλληλη επιλογή υπερπαραμέτρων που σχετίζονται με την εκπαίδευση του μοντέλου επιτρέπει στα NN να μαθαίνουν γρηγορότερα και να επιτυγχάνουν καλύτερη απόδοση.

Εκτός από την επιλογή του βελτιστοποιητή, οι αντίστοιχες υπερπαράμετροι (π.χ., ορμή) είναι κρίσιμες για ορισμένα δίκτυα. Κατά τη διάρκεια της διαδικασίας εκπαίδευσης, το μέγεθος της παρτίδας και ο ρυθμός μάθησης (learning rate - LR) είναι ιδιαίτερα σημαντικοί, καθώς καθορίζουν την ταχύτητα σύγκλισης και πρέπει να ρυθμίζονται προσεκτικά.

Οι υπερπαράμετροι για τον σχεδιασμό του μοντέλου σχετίζονται περισσότερο με τη δομή των NN. Το πιο τυπικό παράδειγμα είναι ο αριθμός των κρυφών επιπέδων και το πλάτος των επιπέδων. Αυτές οι υπερπαράμετροι καθορίζουν την ικανότητα μάθησης του μοντέλου, η οποία εξαρτάται από την πολυπλοκότητα της λειτουργίας που μπορεί να μάθει το μοντέλο.

Η προσεκτική ρύθμιση τόσο των υπερπαραμέτρων εκπαίδευσης όσο και των υπερπαραμέτρων σχεδιασμού είναι κρίσιμη για την επίτευξη της επιθυμητής απόδοσης και της βέλτιστης μάθησης στα νευρωνικά δίκτυα.

Αυτή η ενότητα παρέχει μια εις βάθος συζήτηση για τις υπερπαραμέτρους μεγάλης σημασίας για τις δομές και την εκπαίδευση των μοντέλων, καθώς και εισάγει την επίδρασή τους στα μοντέλα και παρέχει προτεινόμενες τιμές ή προγράμματα.

Βελτιστοποιητές (Optimizers)

Οι βελτιστοποιητές (optimizers) είναι αλγόριθμοι που χρησιμοποιούνται κατά την εκπαίδευση ενός νευρωνικού δικτύου για να προσαρμόζουν τα βάρη του μοντέλου, με σκοπό τη μείωση της συνάρτησης κόστους και τη βελτίωση της απόδοσης του μοντέλου. Η συνάρτηση κόστους (cost function) είναι ένα μαθηματικό μέτρο που αξιολογεί πόσο καλά το μοντέλο προβλέπει τα δεδομένα εκπαίδευσης. Όσο μικρότερη είναι η τιμή της συνάρτησης κόστους, τόσο καλύτερη είναι η απόδοση του μοντέλου.

Οι βελτιστοποιητές ρυθμίζουν τις παραμέτρους του μοντέλου, όπως τα βάρη και τις πολώσεις (biases). Τα βάρη είναι οι παράμετροι που πολλαπλασιάζονται με τις εισόδους του μοντέλου για να

παράγουν τις προβλέψεις, ενώ οι προκαταλήψεις είναι σταθερές τιμές που προστίθενται σε κάθε νευρώνα για να βοηθήσουν το μοντέλο να ταιριάξει καλύτερα τα δεδομένα.

Οι υπερπαράμετροι που σχετίζονται με τους βελτιστοποιητές περιλαμβάνουν την επιλογή του ίδιου του βελτιστοποιητή (όπως SGD, Adam κ.λπ.), το μέγεθος του mini-batch (ο αριθμός των παραδειγμάτων που χρησιμοποιούνται για την ενημέρωση των παραμέτρων σε κάθε βήμα εκπαίδευσης), την ορμή (momentum), η οποία βοηθά στην επιτάχυνση της εκπαίδευσης και τις παραμέτρους beta, που χρησιμοποιούνται σε συγκεκριμένους βελτιστοποιητές για να ρυθμίσουν την προσαρμοστικότητα του ρυθμού μάθησης.

Ορμή (Momentum):

Η ορμή (momentum) είναι μια τεχνική που χρησιμοποιείται για να αποφευχθούν τα τοπικά ελάχιστα κατά την εκπαίδευση του μοντέλου και να μειωθούν οι ταλαντώσεις. Οι ταλαντώσεις είναι ανεπιθύμητες διακυμάνσεις στη διαδικασία εκπαίδευσης, όπου οι ενημερώσεις των βαρών κινούνται προς διάφορες κατευθύνσεις χωρίς σταθερότητα, καθιστώντας τη σύγκλιση πιο αργή και λιγότερο αποδοτική. Τα τοπικά ελάχιστα είναι σημεία όπου η συνάρτηση κόστους φαίνεται να έχει ελάχιστη τιμή, αλλά δεν είναι το πραγματικό ελάχιστο που αναζητάμε.

Η ορμή προσθέτει ένα ποσοστό της προηγούμενης ενημέρωσης βαρών στην τρέχουσα ενημέρωση, βοηθώντας έτσι να διατηρηθεί η κατεύθυνση της βελτιστοποίησης και να αποφευχθεί η παγίδευση σε αυτά τα τοπικά ελάχιστα. Με αυτόν τον τρόπο, η ορμή επιτρέπει στο μοντέλο να κινείται πιο ομαλά και σταθερά προς την πραγματική ελάχιστη τιμή της συνάρτησης κόστους.

Παράμετροι beta (Beta parameters):

Οι παράμετροι beta χρησιμοποιούνται σε αλγόριθμους όπως ο Adam για να ελέγξουν τον ρυθμό φθοράς των εκθετικά σταθμισμένων μέσων τιμών των πρώτων και δεύτερων ροπών των βαθμίδων. Οι τιμές beta1 και beta2 καθορίζουν πόσο βάρος δίνεται στις πρόσφατες βαθμίδες σε σύγκριση με τις παλαιότερες βαθμίδες κατά τον υπολογισμό αυτών των μέσων τιμών.

Η επιλογή του κατάλληλου βελτιστοποιητή είναι μια δύσκολη διαδικασία. Αυτή η ενότητα συζητά τους πιο διαδεδομένους βελτιστοποιητές (mini-batch gradient descent, RMSprop, και Adam), τις σχετικές υπερπαραμέτρους και τις προτεινόμενες τιμές.

Mini-Batch Gradient Descent:

Ο στόχος του mini-batch gradient descent είναι να επιλύσει δύο βασικά προβλήματα: Πρώτον, σε σύγκριση με την απλή κάθοδο κλισης (vanilla gradient descent), το mini-batch gradient descent επιταχύνει τη διαδικασία εκπαίδευσης, ειδικά όταν δουλεύουμε με μεγάλα σύνολα δεδομένων. Δεύτερον, σε σύγκριση με το στοχαστικό gradient descent (SGD) με μέγεθος batch 1, το mini-batch GD μειώνει τον θόρυβο και αυξάνει την πιθανότητα σύγκλισης.

Το μέγεθος του mini-batch είναι μια υπερπαράμετρος και η τιμή του σχετίζεται άμεσα με τη μνήμη της υπολογιστικής μονάδας. Συνιστάται το μέγεθος του mini-batch να είναι δύναμη του 2, καθώς αυτό επιταχύνει την πρόσβαση στη μνήμη της CPU/GPU. Ένα τυπικό προτεινόμενο μέγεθος είναι 32.

SGD με ορμή (Momentum):

Μια βελτιωμένη μέθοδος για την επίλυση του προβλήματος των ταλαντώσεων και της ταχύτητας σύγκλισης είναι η SGD με momentum. Η μέθοδος αυτή επιταχύνει την τυπική SGD υπολογίζοντας τους εκθετικά σταθμισμένους μέσους όρους για τις βαθμίδες. Επιπλέον, βοηθά τη συνάρτηση κόστους να κινηθεί στη σωστή κατεύθυνση προσθέτοντας ένα κλάσμα beta του διανύσματος ενημέρωσης. Η τυπική τιμή για την ορμή (momentum) είναι 0.9, ή 0.99 ή 0.999 αν χρειαστεί.

Adam:

Η προσαρμοστική εκτίμηση ορμής (Adam) επιτυγχάνει επίσης καλά αποτελέσματα γρήγορα σε περισσότερες αρχιτεκτονικές NN. Οι παράμετροι β1 και β2 ελέγχουν τον ρυθμό φθοράς. Στην πράξη, ο Adam συνιστάται ως ο προεπιλεγμένος αλγόριθμος βελτιστοποίησης για την εκπαίδευση βαθιάς μάθησης.

Γενικά, οι αλγόριθμοι βελτιστοποίησης πρέπει να ρυθμίζονται μαζί με το πρόγραμμα του ρυθμού μάθησης. Οι περισσότεροι υπερπαράμετροι για την εκπαίδευση νευρωνικών δικτύων σχετίζονται άμεσα με τους βελτιστοποιητές και τον ρυθμό μάθησης.

Ρυθμος Μάθησης (Learning Rate- LR):

Ο ρυθμός μάθησης (Learning Rate - LR) είναι μια θετική σταθερά που καθορίζει το μέγεθος του βήματος που λαμβάνεται κατά τη διάρκεια της στοχαστικής καθόδου βαθμίδων (SGD). Με απλά λόγια, ο ρυθμός μάθησης καθορίζει πόσο γρήγορα ή αργά το μοντέλο προσαρμόζει τα βάρη του με βάση τα δεδομένα εκπαίδευσης. Σε πολλές περιπτώσεις, ο ρυθμός μάθησης πρέπει να ρυθμίζεται χειροκίνητα κατά τη διάρκεια της εκπαίδευσης του μοντέλου, και αυτή η ρύθμιση είναι συχνά απαραίτητη για την αύξηση της ακρίβειας.

Μια εναλλακτική προσέγγιση στον σταθερό ρυθμό μάθησης είναι η χρήση ενός μεταβλητού ρυθμού μάθησης κατά τη διάρκεια της διαδικασίας εκπαίδευσης, μια μέθοδος γνωστή ως προσαρμοστικός ρυθμός μάθησης ή φθορά του ρυθμού μάθησης. Ένας προσαρμοστικός ρυθμός μάθησης μπορεί να ρυθμίζεται ανάλογα με την απόδοση ή τη δομή του μοντέλου. Αυτή η προσέγγιση μπορεί να βοηθήσει στη βελτίωση της ακρίβειας και της αποδοτικότητας της εκπαίδευσης.

Ο σταθερός ρυθμός μάθησης συχνά ορίζεται ως προεπιλογή από τα πλαίσια βαθιάς μάθησης. Ωστόσο, ο καθορισμός μιας κατάλληλης τιμής για τον ρυθμό μάθησης είναι ένα σημαντικό και δύσκολο έργο. Με έναν κατάλληλο σταθερό ρυθμό μάθησης, το δίκτυο μπορεί να εκπαιδευτεί σε μια αποδεκτή,

αλλά όχι βέλτιστη ακρίβεια, καθώς η αρχική τιμή μπορεί να είναι υπερβολικά μεγάλη, ειδικά στα τελικά βήματα της εκπαίδευσης.

Μια μικρή βελτίωση σε σχέση με τη σταθερή τιμή είναι να οριστεί ένας αρχικός ρυθμός μάθησης, π.χ., 0.1, και στη συνέχεια να μειωθεί σταδιακά όταν η ακρίβεια αρχίσει να σταθεροποιείται. Για παράδειγμα, μπορούμε να ρυθμίσουμε τον ρυθμό μάθησης σε 0.01 όταν η ακρίβεια είναι κορεσμένη, και αν χρειαστεί, να τον μειώσουμε περαιτέρω σε 0.001. Αυτή η προσέγγιση βοηθά στη βελτίωση της ακρίβειας του μοντέλου κατά τη διάρκεια της εκπαίδευσης.

Η γραμμική μείωση του ρυθμού μάθησης είναι μια κοινή επιλογή για τους ερευνητές που επιθυμούν να ορίσουν ένα πρόγραμμα μείωσης του ρυθμού μάθησης. Αυτή η μέθοδος αλλάζει σταδιακά τον ρυθμό μάθησης κατά τη διάρκεια της διαδικασίας εκπαίδευσης, βασιζόμενη στον χρόνο ή το βήμα. Η βασική μαθηματική συνάρτηση της γραμμικής φθοράς είναι: lr = lr0 / (1 + kt), όπου lr0 είναι ο αρχικός ρυθμός μάθησης και k είναι ο ρυθμός φθοράς, με t να αντιπροσωπεύει τον χρόνο ή τον αριθμό των επαναλήψεων.

Εάν το t είναι ο χρόνος εκπαίδευσης, τότε ο ρυθμός μάθησης μειώνεται συνεχώς κατά τη διάρκεια της εκπαίδευσης. Αν το t αντιπροσωπεύει τον αριθμό των επαναλήψεων, ο ρυθμός μάθησης μειώνεται κάθε λίγες επαναλήψεις. Μια τυπική προσέγγιση είναι η μείωση του ρυθμού μάθησης κατά το ήμισυ κάθε 10 εποχές (ο αριθμός των πλήρων διελεύσεων του συνόλου δεδομένων εκπαίδευσης μέσω του αλγορίθμου μάθησης).

Η εκθετική φθορά είναι μία άλλη μέθοδος. Σε σύγκριση με τη γραμμική φθορά, η εκθετική φθορά μειώνει τον ρυθμό μάθησης πιο δραστικά στην αρχή και πιο ήπια καθώς πλησιάζει η σύγκλιση, επιτρέποντας στο μοντέλο να προσαρμόζεται γρηγορότερα στις αρχικές φάσεις της εκπαίδευσης και πιο σταθερά στις τελικές φάσεις.

Η επιλογή του κατάλληλου ρυθμού μάθησης ή του βέλτιστου προγράμματος είναι μια πρόκληση. Ένας μικρός ρυθμός μάθησης οδηγεί σε αργή σύγκλιση, ενώ ένας μεγάλος ρυθμός μάθησης μπορεί να αποτρέψει το μοντέλο από το να συγκλίνει. Η επιλογή του ρυθμού μάθησης πρέπει να ρυθμιστεί ανάλογα με τον αλγόριθμο βελτιστοποίησης. Αν το πρόγραμμα του ρυθμού μάθησης πρέπει να καθοριστεί εκ των προτέρων, συνιστάται να ρυθμιστούν ταυτόχρονα οι υπερπαράμετροι του προγράμματος του ρυθμού μάθησης και οι αντίστοιχοι βελτιστοποιητές.

Η επιλογή του ρυθμού μάθησης διαφέρει ανάλογα με τις συγκεκριμένες εργασίες, αλλά γενικά υπάρχουν κάποιες συμβουλές που βασίζονται στην εμπειρία μας και μπορούν να θεωρηθούν ως γενικός κανόνας για τη ρύθμιση των υπερπαραμέτρων.

Υπερπαράμετροι Σχεδιασμού Μοντέλου:

Παράλληλα με τις προηγούμενες τεχνικές, η επιλογή της αρχιτεκτονικής του μοντέλου και οι λεπτομέρειες για την υλοποίησή του, αποτελούν σημαντικές υπερπαραμέτρους που επηρεάζουν την απόδοση και τη συμπεριφορά του μοντέλου.

Ο αριθμός των κρυφών επιπέδων (d) είναι μια κρίσιμη παράμετρος για τον καθορισμό της συνολικής δομής των NN και έχει άμεση επίδραση στο τελικό αποτέλεσμα. Τα δίκτυα βαθιάς μάθησης με περισσότερα επίπεδα μπορούν να μάθουν πιο σύνθετα χαρακτηριστικά και να επιτύχουν υψηλότερη ακρίβεια. Η προσθήκη περισσότερων επιπέδων σε ένα NN είναι μια συνηθισμένη μέθοδος για την επίτευξη καλύτερων αποτελεσμάτων, καθώς αυξάνει το πεδίο αντίληψης του δικτύου.

Για παράδειγμα, οι χρήστες μπορούν να επιλέξουν από το ResNet-18 έως το ResNet-200, ανάλογα με τις ανάγκες τους για ακρίβεια. Αυτή η μελέτη δημιούργησε μια σειρά NN με διαφορετικούς συνδυασμούς βάθους και πλάτους, επιτυγχάνοντας υψηλή ακρίβεια με χαμηλούς υπολογιστικούς πόρους και περιορισμένο αριθμό παραμέτρων.

Με αυτόν τον τρόπο, μια βασική δομή επαναλαμβάνεται για την αύξηση του πεδίου αντίληψης, επιτρέποντας στους χρήστες να προσαρμόζουν το μοντέλο τους για βέλτιστη απόδοση σύμφωνα με τις συγκεκριμένες απαιτήσεις τους.

Ο αριθμός των νευρώνων σε κάθε επίπεδο (w) πρέπει επίσης να εξεταστεί προσεκτικά. Πολύ λίγοι νευρώνες στα κρυφά επίπεδα μπορεί να προκαλέσουν υποπροσαρμογή(underfitting) επειδή το μοντέλο στερείται πολυπλοκότητας. Η υποπροσαρμογή συμβαίνει όταν ένα μοντέλο είναι πολύ απλό για να κατανοήσει την υποκείμενη δομή των δεδομένων. Αυτό σημαίνει ότι το μοντέλο δεν μπορεί να μάθει αρκετά καλά από τα δεδομένα εκπαίδευσης, με αποτέλεσμα να έχει χαμηλή απόδοση τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα δοκιμών. Τα κύρια χαρακτηριστικά της υποπροσαρμογή περιλαμβάνουν χαμηλή ακρίβεια τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα δοκιμών, το μοντέλο αποτυγχάνει να καταγράψει τα μοτίβα των δεδομένων εκπαίδευσης και έχει υψηλό σφάλμα εκπαίδευσης.

Αντίθετα, πάρα πολλοί νευρώνες μπορεί να οδηγήσουν σε υπερπροσαρμογή (overfitting) και να αυξήσουν τον χρόνο εκπαίδευσης. Η υπερπροσαρμογή συμβαίνει όταν ένα μοντέλο είναι πολύ περίπλοκο και μαθαίνει όχι μόνο τα μοτίβα των δεδομένων εκπαίδευσης αλλά και τον "θόρυβο" και τις τυχαίες διακυμάνσεις των δεδομένων. Αυτό σημαίνει ότι το μοντέλο έχει πολύ καλή απόδοση στα δεδομένα εκπαίδευσης αλλά αποτυγχάνει να γενικεύσει καλά στα νέα δεδομένα δοκιμών. Τα κύρια χαρακτηριστικά της υπερπροσαρμογής περιλαμβάνουν πολύ υψηλή ακρίβεια στα δεδομένα εκπαίδευσης, αλλά χαμηλή ακρίβεια στα δεδομένα δοκιμών, το μοντέλο απομνημονεύει τα δεδομένα εκπαίδευσης αντί να μαθαίνει τα υποκείμενα μοτίβα και έχει χαμηλό σφάλμα εκπαίδευσης αλλά υψηλό σφάλμα δοκιμών.

Σε αντίθεση με την προσθήκη επιπέδων ή νευρώνων σε κάθε επίπεδο, η κανονικοποίηση (regularization) εφαρμόζεται για να μειώσει την πολυπλοκότητα των νευρωνικών δικτύων, ειδικά όταν υπάρχουν ανεπαρκή δεδομένα εκπαίδευσης. Η υπερπροσαρμογή συμβαίνει συνήθως σε σύνθετα NN με πολλά επίπεδα και νευρώνες. Για να μειωθεί η πολυπλοκότητα και να αποφευχθεί η υπερβολική εκπαίδευση, χρησιμοποιούνται μέθοδοι κανονικοποίησης.

Αυτές οι μέθοδοι προσθέτουν έναν όρο κανονικοποίησης στη συνάρτηση κόστους, που είναι μια επιπλέον συνιστώσα στην εξίσωση που το μοντέλο προσπαθεί να ελαχιστοποιήσει κατά την εκπαίδευση.

Αυτός ο όρος κανονικοποίησης τιμωρεί μεγάλα βάρη, αποθαρρύνοντας το μοντέλο από το να προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης και να απομνημονεύει θόρυβο ή τυχαία διακυμάνσεις. Με αυτόν τον τρόπο, το μοντέλο μπορεί να γενικεύσει καλύτερα, δηλαδή να αποδίδει πιο αξιόπιστα σε νέα, αόρατα δεδομένα. Επιπλέον, η κανονικοποίηση βοηθά στη διαδικασία επιλογής χαρακτηριστικών, επιτρέποντας στο μοντέλο να εστιάσει στα πιο σημαντικά χαρακτηριστικά των δεδομένων.

Η γενική έκφραση του όρου κανονικοποίησης είναι:

cost = lossfunction + regularizationterm	2.6

Για να κατανοήσουμε καλύτερα τις διαφορές και τις ιδιαιτερότητες των μεθόδων κανονικοποίησης, ας δούμε τον πίνακα, ο οποίος συγκρίνει τις δύο πιο κοινές μεθόδους κανονικοποίησης, L1 και L2:

Χαρακτηριστικό	**L1**	**L2**
Τιμωρεί	Το άθροισμα των απόλυτων τιμών των βαρών	Το άθροισμα των τετραγωνικών τιμών των βαρών
Λύση	Σπάνια (Sparse)	Μη σπάνια (Nonsparse)
Πλήθος Λύσεων	Πολλαπλές Λύσεις	Μία Λύση
Επιλογή Χαρακτηριστικών	Ενσωματωμένη	Χωρίς επιλογή χαρακτηριστικών
Ανθεκτικότητα σε Ακραίες Τιμές	Εύρωστη(Robust)	Μη Εύρωστη(Not robust)
Ικανότητα Εκμάθησης Πολύπλοκων Προτύπων	Αδύναμη	Ισχυρή

Πίνακας 2.1: Πίνακα με συγκριτική ανάλυση των πιο πολυχρησιμοποιημένων συναρτήσεων κανονικοποίησης.

Ο πίνακας 2.1 βοηθά να κατανοηθούν οι διαφορές μεταξύ L1 και L2 regularization:

Η L1 τιμωρεί το άθροισμα των απόλυτων τιμών των βαρών. Αυτό σημαίνει ότι προωθεί τη σπανιότητα στα βάρη, δηλαδή πολλοί συντελεστές γίνονται μηδενικοί. Ο όρος L1 χρησιμοποιεί το άθροισμα των απόλυτων τιμών των βαρών. Το μοντέλο παλινδρόμησης που χρησιμοποιείται για την L1 ονομάζεται επίσης Lasso Regression ή L1 norm:

$$w* = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_i j W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

Παράγει σπάνιες λύσεις, όπου πολλοί από τους συντελεστές είναι μηδενικοί, κάτι που βοηθά στην απλοποίηση του μοντέλου. Μπορεί να παράγει πολλαπλές λύσεις λόγω της σπανιότητας. Ενσωματώνει την επιλογή χαρακτηριστικών, καθώς μπορεί να εξαλείψει εντελώς ορισμένα χαρακτηριστικά κάνοντάς τα μηδενικά. Είναι ανθεκτική σε ακραίες τιμές, καθώς τιμωρεί τα βάρη με έναν γραμμικό τρόπο. Δεν μπορεί να μάθει πολύπλοκα πρότυπα όσο καλά όσο η L2 λόγω της προώθησης της σπανιότητας.

Η L2 τιμωρεί το άθροισμα των τετραγωνικών τιμών των βαρών. Αυτό σημαίνει ότι τα βάρη μειώνονται αλλά δεν γίνονται μηδενικά. Ο όρος L2 χρησιμοποιεί το άθροισμα των τετραγωνικών τιμών των βαρών. Το μοντέλο παλινδρόμησης που χρησιμοποιείται για την L2 ονομάζεται επίσης Ridge Regression ή L2 norm:

$$w* = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_i j W_j)^2 + \lambda \sum_{j=0}^{M} W_j^2$$

Παράγει μη σπάνιες λύσεις, όπου οι περισσότεροι συντελεστές δεν είναι μηδενικοί. Τείνει να παράγει μία μοναδική λύση, καθώς μειώνει ομοιόμορφα τα βάρη. Δεν κάνει επιλογή χαρακτηριστικών, καθώς απλώς μειώνει τα βάρη χωρίς να τα μηδενίζει. Δεν είναι ανθεκτική σε ακραίες τιμές, καθώς οι ακραίες τιμές μπορούν να έχουν υπερβολική επιρροή λόγω του τετραγώνου των βαρών. Είναι ικανή να μάθει πιο πολύπλοκα πρότυπα, καθώς διατηρεί όλα τα χαρακτηριστικά με μη μηδενικά βάρη.

Ένα πολύ μεγάλο λ θα απλοποιήσει υπερβολικά τη δομή του δικτύου βαθιάς μάθησης επειδή κάποια βάρη θα είναι κοντά στο μηδέν, ενώ ένα πολύ μικρό λ δεν είναι αρκετά ισχυρό για να μειώσει τα βάρη. Στην πράξη, η κανονικοποίηση L2 χρησιμοποιείται περισσότερο, κυρίως λόγω της υπολογιστικής της αποδοτικότητας. Ωστόσο, η κανονικοποίηση L1 υπερέχει στην ιδιότητα σπανιότητας. Τα μοντέλα που δημιουργούνται με το L1 norm είναι πιο απλά και ευανάγνωστα, ενώ το L2 παρέχει ανώτερες προβλέψεις.

Εκτός από τις παραπάνω μεθόδους, η προσαύξηση δεδομένων (data augmentation) και το dropout είναι συνήθεις τεχνικές κανονικοποίησης. Η προσαύξηση δεδομένων δημιουργεί και προσθέτει ψευδή δεδομένα στο σύνολο δεδομένων εκπαίδευσης για να αποφευχθεί η υπερβολική εκπαίδευση. Αυτές οι τεχνικές περιλαμβάνουν προσθήκη θορύβου, που εισάγει τυχαίο θόρυβο στο ηχητικό σήμα για να γίνει το μοντέλο πιο ανθεκτικό σε πραγματικές συνθήκες, μετατόπιση συχνότητας, που αλλάζει τη συχνότητα του ηχητικού σήματος για να προσομοιωθεί η μεταβολή τόνου, αλλαγή ταχύτητας, που μεταβάλλει την ταχύτητα αναπαραγωγής του ήχου χωρίς αλλαγή της συχνότητας για να διαφοροποιηθεί η διάρκεια των ηχητικών σημάτων, και αλλαγή ύψους, που μεταβάλλει τον τόνο του ηχητικού σήματος. Επιπλέον, η διαμόρφωση της ηχητικής έντασης μπορεί να χρησιμοποιηθεί για να διαφοροποιήσει την ένταση του ήχου. Η επιλογή και οι ρυθμίσεις αυτών των μεθόδων ενίσχυσης δεδομένων (όπως το είδος και η ένταση των μετασχηματισμών) αποτελούν υπερπαραμέτρους, καθώς καθορίζονται πριν από την εκπαίδευση και επηρεάζουν την απόδοση του μοντέλου.

Το dropout είναι μια τεχνική όπου επιλέγονται τυχαία νευρώνες με μια δεδομένη πιθανότητα να μην χρησιμοποιηθούν κατά την εκπαίδευση, καθιστώντας το δίκτυο λιγότερο ευαίσθητο στα συγκεκριμένα βάρη των νευρώνων. Η πιθανότητα dropout (για παράδειγμα, 0.5) είναι μια υπερπαράμετρος, καθώς καθορίζεται πριν από την εκπαίδευση και επηρεάζει την ικανότητα του μοντέλου να αποφεύγει την υπερβολική εκπαίδευση, εξισορροπώντας την απόδοση και τη γενίκευση του μοντέλου.

Οι συναρτήσεις ενεργοποίησης είναι κρίσιμες στη βαθιά μάθηση, καθώς εισάγουν μη γραμμικές ιδιότητες στην έξοδο των νευρώνων. Αυτό σημαίνει ότι επιτρέπουν στο NN να μάθει και να αναπαραστήσει πολύπλοκα πρότυπα και σχέσεις στα δεδομένα. Χωρίς συνάρτηση ενεργοποίησης, κάθε επίπεδο του δικτύου θα εκτελούσε απλά γραμμικούς υπολογισμούς. Στην περίπτωση αυτή, ανεξάρτητα από τον αριθμό των επιπέδων, το τελικό μοντέλο θα ήταν ισοδύναμο με ένα απλό γραμμικό μοντέλο παλινδρόμησης.

Ένα γραμμικό μοντέλο παλινδρόμησης μπορεί μόνο να αναπαραστήσει γραμμικές σχέσεις, δηλαδή σχέσεις όπου η αλλαγή στην είσοδο οδηγεί σε ανάλογη και άμεση αλλαγή στην έξοδο. Ωστόσο, τα περισσότερα πραγματικά προβλήματα περιλαμβάνουν πολύπλοκες, μη γραμμικές σχέσεις. Για παράδειγμα, στην αναγνώριση εικόνων, τα δεδομένα συχνά περιέχουν πολύπλοκες δομές και μοτίβα που δεν μπορούν να αποτυπωθούν με γραμμικούς υπολογισμούς.

Οι συναρτήσεις ενεργοποίησης επιτρέπουν στο NN να μαθαίνει αυτές τις μη γραμμικές σχέσεις, καθιστώντας το ικανό να αναγνωρίζει και να αναπαραγάγει πιο περίπλοκα χαρακτηριστικά των δεδομένων. Δημοφιλείς συναρτήσεις ενεργοποίησης όπως η sigmoid, η υπερβολική εφαπτομένη (tanh), οι διορθωμένες γραμμικές μονάδες (ReLU), η Maxout και η Swish χρησιμοποιούνται ευρέως, καθώς καθεμία έχει συγκεκριμένες ιδιότητες που βοηθούν στην εκπαίδευση και την απόδοση των νευρωνικών δικτύων σε διαφορετικά προβλήματα.

Η επιλογή της συνάρτησης ενεργοποίησης είναι μια υπερπαράμετρος, διότι καθορίζεται πριν από την εκπαίδευση και δεν προσαρμόζεται κατά τη διάρκεια της εκπαίδευσης. Η συγκεκριμένη επιλογή επηρεάζει σημαντικά την ικανότητα του νευρωνικού δικτύου να μάθει και να αναπαραστήσει πολύπλοκα πρότυπα. Κάθε συνάρτηση ενεργοποίησης έχει τα δικά της πλεονεκτήματα και μειονεκτήματα, και η σωστή επιλογή μπορεί να βελτιώσει την απόδοση του μοντέλου. Για παράδειγμα, η ReLU είναι γνωστή για την ικανότητά της να επιταχύνει τη σύγκλιση της εκπαίδευσης, ενώ η sigmoid και η tanh χρησιμοποιούνται συχνά σε περιπτώσεις όπου απαιτείται ομαλή κλιμάκωση των εξόδων. Η Maxout και η Swish είναι πιο πρόσφατες προτάσεις που μπορούν να προσφέρουν βελτιώσεις σε συγκεκριμένα σενάρια.[25]

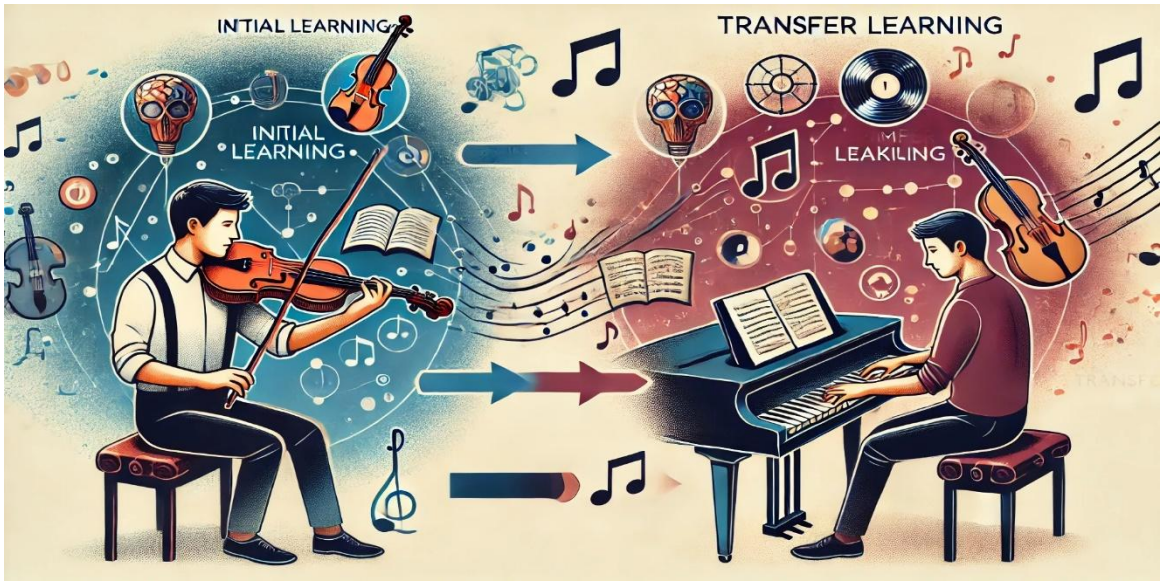
2.2.6 Μεταφορά Μάθησης (Transfer Learning)

Παρόλο που η παραδοσιακή τεχνολογία ML έχει σημειώσει μεγάλη επιτυχία και έχει εφαρμοστεί σε πολλές πρακτικές εφαρμογές, εξακολουθεί να έχει ορισμένους περιορισμούς για ορισμένα σενάρια του

πραγματικού κόσμου. Το ιδανικό σενάριο της ML είναι να υπάρχουν άφθονα επισημασμένα παραδείγματα εκπαίδευσης, τα οποία έχουν την ίδια κατανομή με τα δεδομένα δοκιμής. Ωστόσο, η συλλογή επαρκών δεδομένων εκπαίδευσης είναι συχνά δαπανηρή, χρονοβόρα ή ακόμη και μη ρεαλιστική σε πολλές περιπτώσεις.

Η μεταφορά μάθησης (Transfer Learning -), η οποία επικεντρώνεται στη μεταφορά γνώσης μεταξύ διαφορετικών πεδίων, είναι μια πολλά υποσχόμενη μεθοδολογία μηχανικής μάθησης για την επίλυση του προβλήματος της έλλειψης δεδομένων. Η έννοια της TL μπορεί αρχικά να προήλθε από την εκπαιδευτική ψυχολογία. Σύμφωνα με τη θεωρία της γενίκευσης της μεταφοράς, όπως προτάθηκε από τον ψυχολόγο C.H. Judd, η μάθηση της μεταφοράς είναι το αποτέλεσμα της γενίκευσης της εμπειρίας. Δηλαδή, είναι δυνατόν να πραγματοποιηθεί η μεταφορά γνώσης από μια κατάσταση σε μια άλλη, εφόσον ένα άτομο μπορεί να γενικεύσει την εμπειρία του.

Σύμφωνα με αυτή τη θεωρία, η προϋπόθεση για τη μεταφορά είναι να υπάρχει σύνδεση μεταξύ δύο μαθησιακών δραστηριοτήτων. Για παράδειγμα, ένα άτομο που έχει μάθει να παίζει βιολί μπορεί να μάθει να παίζει πιάνο πιο γρήγορα από άλλους, καθώς τόσο το βιολί όσο και το πιάνο είναι μουσικά όργανα και μοιράζονται κάποια κοινή γνώση. Αυτή η γενίκευση της εμπειρίας επιτρέπει τη μεταφορά της μάθησης από το ένα πεδίο στο άλλο.



Εικόνα 2.19: Απεικόνιση της διαδικασίας μεταφοράς γνώσης υπό την σκοπιά της ψυχολογίας. Όταν ένα άτομο έχει αναπτύξει δεξιότητες σε έναν τομέα είναι πολύ πιο εύκολο να προσαρμοστεί σε έναν νέο αλλά παρεμφερή τομέα. Στο σχήμα παρουσιάζεται η μεταφορά γνώσης μεταξύ μουσικών δεξιοτήτων για διαφορετικά όργανα.

Εμπνευσμένη από την ικανότητα των ανθρώπων να μεταφέρουν γνώση μεταξύ διαφορετικών πεδίων, η μεταφορά μάθησης στοχεύει στην αξιοποίηση της γνώσης από ένα σχετικό πεδίο (που ονομάζεται πηγή πεδίο) για να βελτιώσει την απόδοση της μάθησης ή να μειώσει τον αριθμό των επισημασμένων παραδειγμάτων που απαιτούνται σε ένα πεδίο στόχου. Ωστόσο, αξίζει να σημειωθεί ότι η μεταφερόμενη γνώση δεν έχει πάντα θετικό αντίκτυπο σε νέες εργασίες. Αν υπάρχει μικρή ομοιότητα μεταξύ των πεδίων, η μεταφορά γνώσης μπορεί να είναι ανεπιτυχής. Για παράδειγμα, η μάθηση ποδηλασίας δεν μπορεί να βοηθήσει στην ταχύτερη εκμάθηση του πιάνου.

Επιπλέον, οι ομοιότητες μεταξύ των πεδίων δεν διευκολύνουν πάντα τη μάθηση, καθώς μπορεί να είναι παραπλανητικές. Για παράδειγμα, παρόλο που τα Ισπανικά και τα Γαλλικά είναι στενά συνδεδεμένες γλώσσες, οι μαθητές των Ισπανικών μπορεί να αντιμετωπίσουν δυσκολίες στην εκμάθηση Γαλλικών, όπως η χρήση λανθασμένου λεξιλογίου ή κλίσης. Αυτό συμβαίνει επειδή η προηγούμενη εμπειρία και γνώση των Ισπανικών μπορεί να παρεμβαίνει στη μάθηση και το σχηματισμό λέξεων, της χρήσης, της προφοράς και της κλίσης στα Γαλλικά. Στην ψυχολογία, αυτό το φαινόμενο, όπου η προηγούμενη εμπειρία έχει αρνητική επίδραση στη μάθηση νέων εργασιών, ονομάζεται αρνητική μεταφορά.

Παρομοίως, στον τομέα της TL, αν η μεταφερόμενη γνώση έχει αρνητική επίδραση στον μαθητή στόχο, το φαινόμενο αυτό ονομάζεται επίσης αρνητική μεταφορά. Το αν θα συμβεί αρνητική μεταφορά μπορεί να εξαρτηθεί από διάφορους παράγοντες, όπως η συνάφεια μεταξύ της πηγής και του στόχου και η ικανότητα του μαθητή να εντοπίσει και να αξιοποιήσει το ωφέλιμο μέρος της γνώσης μεταξύ των πεδίων.

Σε γενικές γραμμές, η TL μπορεί να χωριστεί σε δύο κατηγορίες, ανάλογα με τη διαφορά μεταξύ των πεδίων: ομοιογενή και ετερογενή μεταφορά μάθησης. Οι ομοιογενείς προσεγγίσεις μεταφοράς μάθησης αναπτύσσονται για περιπτώσεις όπου τα πεδία έχουν τον ίδιο χώρο χαρακτηριστικών.

Σε αυτή την κατηγορία, ορισμένες μελέτες υποθέτουν ότι τα πεδία διαφέρουν μόνο στις οριακές κατανομές. Αυτό σημαίνει ότι η διαφορά μεταξύ των πεδίων οφείλεται στην κατανομή των δεδομένων και όχι στα ίδια τα χαρακτηριστικά. Η οριακή κατανομή αναφέρεται στη συνολική κατανομή των χαρακτηριστικών χωρίς να λαμβάνεται υπόψη η σχέση τους με τις ετικέτες ή τα αποτελέσματα. Για παράδειγμα, μπορεί να υπάρχουν διαφορετικές κατανομές ηλικιών σε δύο σύνολα δεδομένων, αλλά τα ίδια χαρακτηριστικά ηλικίας υπάρχουν και στα δύο.

Για να αντιμετωπιστούν αυτές οι διαφορές, προσαρμόζονται τα πεδία διορθώνοντας την προκατάληψη επιλογής δείγματος ή την αλλαγή συνθηκών. Η προκατάληψη επιλογής δείγματος αναφέρεται στην ανισορροπία που μπορεί να προκύψει όταν τα δείγματα που επιλέγονται για εκπαίδευση δεν αντιπροσωπεύουν πλήρως τον πληθυσμό. Αυτό μπορεί να διορθωθεί με τεχνικές όπως η αναπροσαρμογή των βαρών των δειγμάτων ή η επιλογή περισσότερων αντιπροσωπευτικών δειγμάτων.

Ωστόσο, αυτή η υπόθεση δεν ισχύει πάντα. Για παράδειγμα, στο πρόβλημα ταξινόμησης συναισθημάτων, μια λέξη μπορεί να έχει διαφορετικές σημασίες σε διαφορετικά πεδία, κάτι που

ονομάζεται προκατάληψη χαρακτηριστικών περιεχομένου. Για την επίλυση αυτού του προβλήματος, ορισμένες μελέτες προσαρμόζουν περαιτέρω τις συνθήκες κατανομών.

Η ετερογενής TL αναφέρεται στη διαδικασία μεταφοράς γνώσης σε περιπτώσεις όπου τα πεδία έχουν διαφορετικούς χώρους χαρακτηριστικών. Εκτός από την προσαρμογή των κατανομών, η ετερογενής μεταφορά μάθησης απαιτεί επίσης την προσαρμογή του χώρου χαρακτηριστικών, κάτι που την καθιστά πιο περίπλοκη από την ομοιογενή μεταφορά μάθησης. [26]

Κατηγοριοποιήση τεχνικών μεταφοράς μάθησης

Στην έρευνα σχετικά με τη TL, αναγνωρίζονται τρία κύρια ερευνητικά θέματα: Ποιοι είναι οι κατάλληλοι παράγοντες ή γνώσεις προς μεταφορά, ποια είναι η κατάλληλη μέθοδος ή διαδικασία για τη μεταφορά αυτών των παραγόντων, και ποια είναι η βέλτιστη χρονική στιγμή για την εκτέλεση της μεταφοράς.

Το πρώτο θέμα αναφέρεται στο ποιο μέρος της γνώσης μπορεί να μεταφερθεί μεταξύ διαφορετικών πεδίων ή εργασιών. Κάποιες γνώσεις είναι συγκεκριμένες για μεμονωμένα πεδία ή εργασίες, ενώ άλλες μπορεί να είναι κοινές μεταξύ διαφορετικών πεδίων, έτσι ώστε να μπορούν να βοηθήσουν στη βελτίωση της απόδοσης στο πεδίο ή την εργασία στόχο. Μετά την ανακάλυψη ποια γνώση μπορεί να μεταφερθεί, πρέπει να αναπτυχθούν αλγόριθμοι μάθησης για τη μεταφορά αυτής της γνώσης, γεγονός που συνδέεται με το δεύτερο ερευνητικό θέμα, δηλαδή ποια είναι η κατάλληλη μέθοδος ή διαδικασία για τη μεταφορά.

Η χρονική στιγμή της μεταφοράς είναι κρίσιμη και αφορά την αξιολόγηση των καταστάσεων στις οποίες η μεταφορά γνώσης πρέπει να εφαρμοστεί. Είναι ουσιώδες να αναγνωρίζονται οι συνθήκες όπου η μεταφορά γνώσης δεν είναι επιθυμητή. Σε ορισμένες περιπτώσεις, όταν το πεδίο πηγής και το πεδίο στόχου δεν έχουν επαρκή συσχέτιση, η επιβολή της μεταφοράς μπορεί να αποτύχει. Στη χειρότερη εκδοχή, αυτό μπορεί να επιφέρει αρνητικές επιπτώσεις στην απόδοση της μάθησης στο πεδίο στόχου, φαινόμενο γνωστό ως αρνητική μεταφορά.

Οι περισσότερες τρέχουσες εργασίες στη μεταφορά μάθησης επικεντρώνονται στο "Τι να μεταφέρουμε" και "Πώς να μεταφέρουμε", υποθέτοντας σιωπηρά ότι τα πεδία πηγής και στόχου σχετίζονται μεταξύ τους. Ωστόσο, το πώς να αποφευχθεί η αρνητική μεταφορά είναι ένα σημαντικό ανοιχτό ζήτημα που προσελκύει όλο και περισσότερη προσοχή στο μέλλον.

Βάσει του ορισμού της μεταφοράς μάθησης, αυτή κατηγοριοποιείται σε τρεις υποκατηγορίες: επαγωγική μεταφορά μάθησης(inductive transfer leanring), μεταδοτική μεταφορά μάθησης (transductive transfer learning) και μη επιβλεπόμενη μεταφορά μάθησης (unsupervised transfer learning), ανάλογα με τις διαφορετικές καταστάσεις μεταξύ των πεδίων και των εργασιών πηγής και στόχου.

Στην επαγωγική μεταφοράς μάθησης, η εργασία στόχος (το πρόβλημα που θέλουμε να λύσουμε) είναι διαφορετική από την εργασία πηγής (το αρχικό πρόβλημα από το οποίο αντλούμε γνώση), ανεξάρτητα από το αν τα πεδία πηγής και στόχου (τα δεδομένα που χρησιμοποιούμε) είναι τα ίδια ή όχι. Σε αυτή την

περίπτωση, απαιτούνται επισημασμένα δεδομένα στον τομέα στόχο για την ανάπτυξη ενός προβλεπτικού μοντέλου για χρήση στον τομέα στόχο.

Ανάλογα με τη διαθεσιμότητα των επισημασμένων και μη επισημασμένων δεδομένων στον τομέα πηγής, η επαγωγική ρύθμιση μεταφοράς μάθησης μπορεί να κατηγοριοποιηθεί περαιτέρω σε δύο περιπτώσεις:

Στην πρώτη περίπτωση, υπάρχουν πολλά επισημασμένα δεδομένα στον τομέα πηγής. Αυτή η κατάσταση μοιάζει με την multi-task learning μάθηση, με τη διαφορά ότι η επαγωγική μεταφορά μάθησης στοχεύει αποκλειστικά στη βελτίωση της απόδοσης στην εργασία στόχο, χρησιμοποιώντας γνώση από την εργασία πηγής. Αντίθετα, η multi-task learning μάθηση προσπαθεί να μάθει ταυτόχρονα τόσο την εργασία στόχο όσο και την εργασία πηγής.

Στη δεύτερη περίπτωση, δεν υπάρχουν επισημασμένα δεδομένα στον τομέα πηγής. Αυτή η κατάσταση μοιάζει με την self-taugh learning μάθηση, στην οποία οι χώροι ετικετών μεταξύ των πεδίων πηγής και στόχου μπορεί να είναι διαφορετικοί, πράγμα που σημαίνει ότι οι πληροφορίες από τον τομέα πηγής δεν μπορούν να χρησιμοποιηθούν άμεσα. Έτσι, αυτή η περίπτωση είναι παρόμοια με τη ρύθμιση επαγωγικής μεταφοράς μάθησης όπου τα επισημασμένα δεδομένα στον τομέα πηγής δεν είναι διαθέσιμα.

Στη μεταδοτική μεταφορά μάθησης, οι εργασίες πηγής και στόχου είναι οι ίδιες, αλλά τα πεδία πηγής και στόχου είναι διαφορετικά. Αυτό σημαίνει ότι το πρόβλημα που προσπαθούμε να λύσουμε (η εργασία) είναι το ίδιο και στα δύο πεδία. Για παράδειγμα, μπορεί να θέλουμε να ταξινομήσουμε τα ίδια είδη δεδομένων, όπως emails σε κατηγορίες "spam" και "not spam", τόσο στο πεδίο πηγής όσο και στο πεδίο στόχου.

Όταν λέμε ότι τα πεδία πηγής και στόχου είναι διαφορετικά, εννοούμε ότι τα δεδομένα που προέρχονται από κάθε πεδίο έχουν διαφορετικές ιδιότητες ή χαρακτηριστικά. Για παράδειγμα, τα emails στον τομέα πηγής μπορεί να προέρχονται από έναν συγκεκριμένο πάροχο υπηρεσιών email, ενώ τα emails στον τομέα στόχου μπορεί να προέρχονται από έναν άλλο πάροχο υπηρεσιών email. Αν και η εργασία (ταξινόμηση emails σε spam και not spam) είναι η ίδια, οι ιδιότητες ή τα χαρακτηριστικά των δεδομένων μπορεί να διαφέρουν μεταξύ των δύο πεδίων λόγω της διαφορετικής προέλευσης ή του τρόπου συλλογής τους.

Σε αυτή την περίπτωση, δεν υπάρχουν επισημασμένα δεδομένα στον τομέα στόχο, ενώ υπάρχουν πολλά επισημασμένα δεδομένα στον τομέα πηγής. Ανάλογα με τις διαφορές μεταξύ των πεδίων πηγής και στόχου, η μεταδοτική μεταφορά μάθησης μπορεί να διακριθεί περαιτέρω σε δύο περιπτώσεις.

Πρώτον, όταν οι χώροι χαρακτηριστικών μεταξύ των πεδίων πηγής και στόχου είναι διαφορετικοί. Αυτό σημαίνει ότι τα χαρακτηριστικά που χρησιμοποιούνται για την περιγραφή των δεδομένων στις δύο περιοχές είναι διαφορετικά. Για παράδειγμα, τα χαρακτηριστικά που περιγράφουν τα emails από τον έναν πάροχο μπορεί να διαφέρουν από τα χαρακτηριστικά που περιγράφουν τα emails από τον άλλον πάροχο.

Δεύτερον, όταν οι χώροι χαρακτηριστικών είναι οι ίδιοι, αλλά οι οριακές κατανομές των δεδομένων εισόδου είναι διαφορετικές. Αυτό σημαίνει ότι, ενώ τα ίδια χαρακτηριστικά χρησιμοποιούνται για την περιγραφή των δεδομένων, η συχνότητα ή η κατανομή των χαρακτηριστικών διαφέρει μεταξύ των δύο πεδίων. Για παράδειγμα, οι συχνότητες των λέξεων στα emails μπορεί να είναι διαφορετικές μεταξύ των δύο παρόχων.

Η τελευταία περίπτωση σχετίζεται με την προσαρμογή πεδίου (domain adaptation) για τη μεταφορά γνώσης. Αυτό μπορεί να περιλαμβάνει τεχνικές για τη μεταφορά γνώσης στην ταξινόμηση κειμένων και την αντιμετώπιση της προκατάληψης επιλογής δείγματος ή της μετατόπισης μεταβλητής, όπου οι υποθέσεις είναι παρόμοιες. Με αυτές τις τεχνικές, μπορούμε να προσαρμόσουμε το μοντέλο ώστε να λειτουργεί αποτελεσματικά στο πεδίο στόχου, παρά τις διαφορές στα δεδομένα μεταξύ των δύο πεδίων.

Τέλος, στη μη επιβλεπόμενη μεταφορά μάθησης, η εργασία στόχος είναι διαφορετική από την εργασία πηγής αλλά έχει κάποια σχέση με αυτήν. Αυτό σημαίνει ότι, παρόλο που οι συγκεκριμένες εργασίες διαφέρουν, υπάρχει κάποια σύνδεση ή συσχέτιση που επιτρέπει τη μεταφορά γνώσης από την εργασία πηγής στην εργασία στόχο. Η μη επιβλεπόμενη μεταφορά μάθησης επικεντρώνεται στην επίλυση μη επιβλεπόμενων εργασιών μάθησης στον τομέα στόχο, όπως η ομαδοποίηση και η μείωση διαστάσεων. Σε αυτή την περίπτωση, δεν υπάρχουν επισημασμένα δεδομένα διαθέσιμα τόσο στον τομέα πηγής όσο και στον τομέα στόχο κατά την εκπαίδευση. [27] Παρουσιαζονται πανω απο σαραντα προσεγγισεις, μηχανισμοι, στρατηγικες μεταφορας μάθησης στο ακόλουθο άρθρο. [26]

2.2.7 Μετρικές Αξιολόγησης Απόδοσης

Μετά την εφαρμογή των αλγορίθμων μηχανικής μάθησης, είναι απαραίτητο να χρησιμοποιηθούν εργαλεία για την αξιολόγηση της απόδοσής τους. Αυτά τα εργαλεία είναι γνωστά ως μετρικές αξιολόγησης απόδοσης. Ένας σημαντικός αριθμός μετρικών έχει εισαχθεί σε μελέτες, και κάθε μία εξετάζει συγκεκριμένες πτυχές της απόδοσης ενός αλγορίθμου. Επομένως, για κάθε πρόβλημα μηχανικής μάθησης απαιτείται ένα κατάλληλο σύνολο μετρικών για την αξιολόγηση της απόδοσης.

Μερικές από τις πιο κοινές μετρικές για την αξιολόγηση της απόδοσης περιλαμβάνουν το precision, την ανάκληση (recall), το f1-score, την ακρίβεια (accuracy), τον πίνακα σύγχυσης (confusion matrix) και το ROC-AUC score. Κάθε μία από αυτές τις μετρικές παρέχει διαφορετικές πληροφορίες για την απόδοση του αλγορίθμου, βοηθώντας τους ερευνητές και τους μηχανικούς να κατανοήσουν καλύτερα πώς λειτουργεί το μοντέλο τους και πού μπορεί να χρειάζεται βελτίωση.

Πριν από την ανάλυση των μετρικών, είναι απαραίτητο να εξηγηθούν οι παρακάτω όροι: True Positive (TP), False Positive (FP), True Negative (TN) και False Negative (FN). Έστω ένα πρόβλημα δυαδικής ταξινόμησης, όπου το σύνολο των δεδομένων αποτελείται από δύο κλάσεις, τη θετική και την αρνητική. Οι όροι αυτοί χρησιμοποιούνται για να περιγράψουν τις προβλέψεις του μοντέλου σε σχέση με τις πραγματικές τιμές. Το TP σημαίνει ότι το μοντέλο προβλέπει πως ένα δείγμα ανήκει στην θετική κλάση

και στην πραγματικότητα ανήκει όντως στην θετική κλάση. Το FP σημαίνει ότι το μοντέλο προβλέπει πως ένα δείγμα ανήκει στην θετική κλάση, αλλά στην πραγματικότητα ανήκει στην αρνητική κλάση. Το TN σημαίνει ότι το μοντέλο πως ένα δείγμα ανήκει στην αρνητική κλάση και στην πραγματικότητα ανήκει όντως στην αρνητική. Το FN σημαίνει ότι το μοντέλο προβλέπει πως ένα δείγμα ανήκει στην αρνητική κλάση, αλλά στην πραγματικότητα ανήκει στην θετική.

Precision:

Το precision δείχνει πόσες από τις παρατηρήσεις που ο αλγόριθμος έχει προβλέψει ως θετικές είναι πραγματικά θετικές. Σύμφωνα με τον τύπο, το precision ισούται με τον αριθμό των πραγματικών θετικών διαιρεμένο με το άθροισμα των πραγματικών θετικών και των ψευδών θετικών.

Precision = TP/(TP+FP)	2.6

Είναι ιδιαίτερα χρήσιμο όταν η εσφαλμένη πρόβλεψη μιας περίπτωσης ως θετικής μπορεί να έχει αρνητικές συνέπειες, δηλαδή όταν το κόστος ενός λάθους είναι μεγάλο. Σε αυτές τις περιπτώσεις, θέλουμε να είμαστε πολύ σίγουροι ότι όταν το μοντέλο μας προβλέπει κάτι ως θετικό, είναι όντως θετικό, ακόμη και αν αυτό σημαίνει ότι μπορεί να μην ανιχνεύουμε όλες τις θετικές περιπτώσεις.

Ωστόσο, το precision έχει ένα σημαντικό μειονέκτημα, δεν λαμβάνει υπόψη τις περιπτώσεις όπου το μοντέλο αποτυγχάνει να εντοπίσει τις θετικές περιπτώσεις, τα FN δηλαδη. Αυτό σημαίνει ότι το recismετρά μόνο πόσο καλά το μοντέλο αναγνωρίζει τις θετικές προβλέψεις, αλλά δεν δείχνει πόσες πραγματικές θετικές περιπτώσεις έχουν χαθεί. Ως αποτέλεσμα, αν βασιζόμαστε μόνο στο precision για να αξιολογήσουμε την απόδοση του μοντέλου, μπορεί να μην έχουμε πλήρη εικόνα της απόδοσής του. Αυτό είναι ιδιαίτερα σημαντικό σε καταστάσεις όπου είναι κρίσιμο να εντοπίζονται όλες οι θετικές περιπτώσεις, όπως στην ιατρική διάγνωση όπου το να χάσουμε μια θετική περίπτωση μπορεί να έχει σοβαρές συνέπειες.

Ανάκληση(Recall):

Το recallδείχνει πόσες από τις παρατηρήσεις που είναι πραγματικά θετικές έχουν προβλεφθεί σωστά από τον αλγόριθμο. Ο τύπος για το recall είναι ο αριθμός των TP διαιρεμένος με το άθροισμα των TP και των FN.

Recall = TP/(TP+FN)	2.7

Παρά τη χρησιμότητά της, η ανάκληση δεν είναι πάντα η πιο κατάλληλη μετρική απόδοσης σε ορισμένες περιπτώσεις. Ειδικά όταν η ανάκληση χρησιμοποιείται μόνη της, μπορεί να δώσει παραπλανητικά αποτελέσματα σε καταστάσεις όπου οι FP είναι εξίσου σημαντικές.

F1-score:

Αυτή η μετρική, επίσης γνωστή ως f-score ή f-measure, λαμβάνει υπόψη τόσο το precision όσο και την ανάκληση για να υπολογίσει την απόδοση ενός αλγορίθμου. Μαθηματικά, είναι ο αρμονικός μέσος του precision και recallκαι υπολογίζεται από τον τύπο:

F1-score = 2 X (Precision X Recall)/(Precision + Recall)	2.7

Το F1-score είναι ιδιαίτερα χρήσιμο όταν υπάρχει ανισορροπία μεταξύ των θετικών και αρνητικών παρατηρήσεων στα δεδομένα. Συνδυάζοντας το precision και recall, το F1-score παρέχει μια πιο ισορροπημένη μέτρηση της απόδοσης του αλγορίθμου, καθώς λαμβάνει υπόψη τόσο την ικανότητα του αλγορίθμου να εντοπίζει σωστά τις θετικές περιπτώσεις (recall) όσο και την ακρίβεια με την οποία προβλέπει θετικές περιπτώσεις (precision).

Ακρίβεια(Accuracy):

Είναι η πιο χρησιμοποιούμενη μετρική και ίσως η πρώτη επιλογή για την αξιολόγηση της απόδοσης ενός αλγορίθμου σε προβλήματα ταξινόμησης. Μπορεί να οριστεί ως ο λόγος των σωστά ταξινομημένων δεδομένων προς το συνολικό αριθμό των παρατηρήσεων.

Accuracy = (TP + TN)/ (TP+TN+FP+FN)	2.8

Με άλλα λόγια, δείχνει το ποσοστό των προβλέψεων του αλγορίθμου που είναι σωστές, είτε πρόκειται για θετικές είτε για αρνητικές περιπτώσεις. Δεν είναι πάντα η καλύτερη επιλογή, ειδικά όταν τα δεδομένα είναι μη σταθμισμένα. Σε σύνολα δεδομένων όπου οι θετικές περιπτώσεις είναι πολύ λιγότερες από τις αρνητικές (ή το αντίστροφο), το accuracy μπορεί να δώσει μια παραπλανητική εικόνα της πραγματικής απόδοσης του αλγορίθμου. Για παράδειγμα, σε ένα πρόβλημα όπου το 20% των περιπτώσεων είναι θετικές και το 80% είναι αρνητικές, ένας αλγόριθμος που προβλέπει πάντα αρνητικές περιπτώσεις θα έχει accuracy 80%, παρόλο που δεν εντοπίζει καθόλου τις θετικές περιπτώσεις. Σε τέτοιες περιπτώσεις, είναι σημαντικό να χρησιμοποιούνται επιπλέον μετρικές, όπως το precision, το recall και το F1-score, για να έχουμε μια πληρέστερη εικόνα της απόδοσης του αλγορίθμου.

Πίνακας Σύγχυσης(Confusion Matrix):

Ο πίνακας αυτός αποτελεί μία από τις πιο σημαντικές και περιγραφικές μετρικές για την αξιολόγηση της απόδοσης ενός αλγορίθμου μηχανικής μάθησης, ειδικά σε προβλήματα ταξινόμησης. Πρόκειται για έναν δισδιάστατο πίνακα, στον οποίο η μία διάσταση αντιπροσωπεύει τις πραγματικές κλάσεις των αντικειμένων και η άλλη τις κλάσεις που προβλέπει ο αλγόριθμος ταξινόμησης. Ο παρακάτω πίνακας Πίνακας παρουσιάζει ένα παράδειγμα πίνακα σύγχυσης για ένα πρόβλημα ταξινόμησης τριών κλάσεων, με τις κλάσεις A, B και C.

	ASSIGNED CLASS

		A	B	C
ACTUAL CLASSS	A	10	2	1
	B	0	6	1
	C	0	3	8

Πίνακας 2.2: Παράδειγμα πίνακα σύγχυσης για πρόβλημα ταξινόμησης.

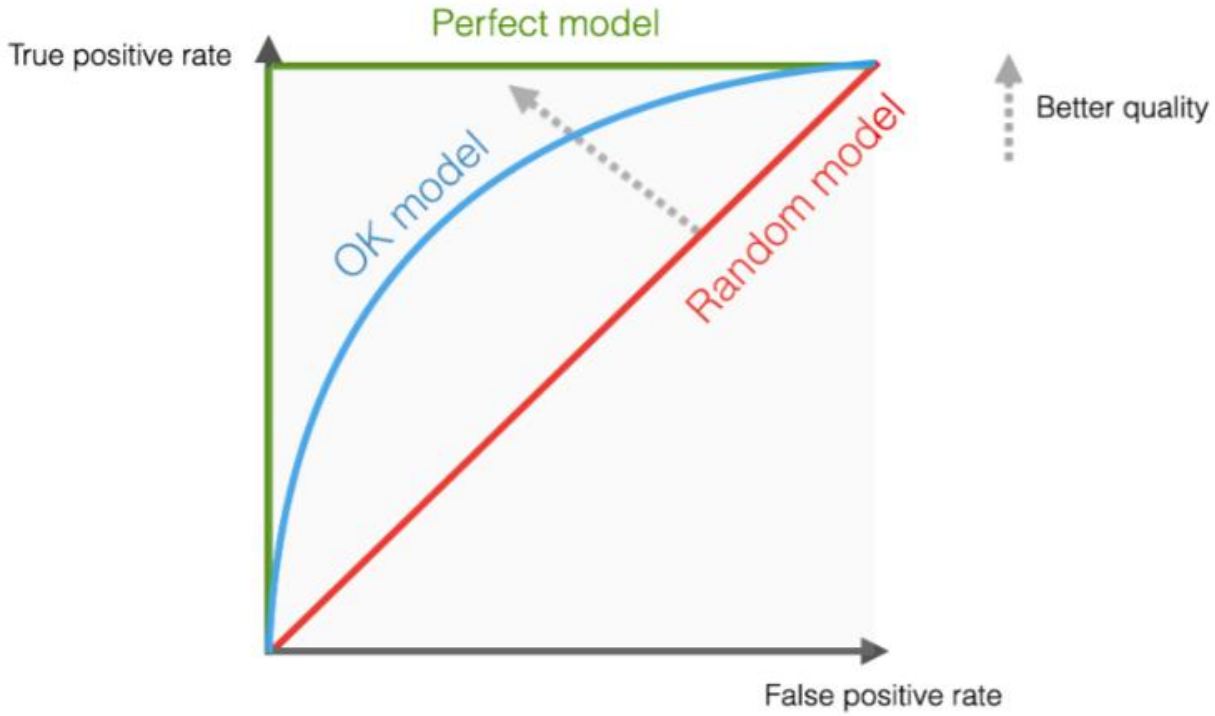
Η πρώτη γραμμή του πίνακα υποδεικνύει ότι 13 αντικείμενα ανήκουν στην κλάση A και ότι 10 ταξινομήθηκαν σωστά ως ανήκοντα στην A, δύο ταξινομήθηκαν εσφαλμένα ως ανήκοντα στην B και ένα ως ανήκον στην C. Αντίστοιχα τα ίδια ισχύουν και για τις επόμενες γραμμές του πίνακα.

Μια ειδική περίπτωση του πίνακα σύγχυσης χρησιμοποιείται συχνά με δύο κλάσεις, μία που ορίζεται ως θετική κλάση και η άλλη ως αρνητική κλάση. Σε αυτό το πλαίσιο, τα τέσσερα κελιά του πίνακα ορίζονται ως αληθώς θετικά (TP), ψευδώς θετικά (FP), αληθώς αρνητικά (TN) και ψευδώς αρνητικά (FN) όπως υποδεικνύεται στον Πίνακα.

		ASSIGNED CLASS	
		Positive	Negative
ACTUAL CLASS	Positive	TP	FN
	Negative	FP	TN

Πίνακας 2.3: Συσχέτιση του πίνακα σύγχυσης και των True Positive (TP), False Positive (FP), True Negative(TN), False Negative(FN)

Ένας αριθμός μετρήσεων της απόδοσης ταξινόμησης ορίζονται με όρους αυτών των τεσσάρων αποτελεσμάτων ταξινόμησης. Το True Negative Rate ορίζεται ως TN/(TN+FP) ενώ το False Positive Rate ως FP/(TN+FP).



Εικόνα 2.20: Γραφική παράσταση για την κατανόηση της μετρικής Receiver Operating Characteristic - Area Under the Curve (ROC-AUC). Οι άξονες x,y συμβολίζουν τα False Positive rate και True Positive rate αντίστοιχα. Για να σχηματιστεί η συγκεκριμένη γραφική παράσταση γίνεται χρήση διαφορετικών κατωφλίων (threshold) για την μετατροπή της πιθανότητας να είναι αληθής η κάθε κλάση. Ένα τέλειος μοντέλο θα έχει έξοδο ίση με 1 για αληθείς κλάσεις και 0 για ψευδείς και επομένως, η γραφική παράσταση θα είναι η πράσινη. Οποιοδήποτε μοντέλο προσεγγίζει την y=x (κόκκινη ευθεία) τείνει να είναι τυχαίο και κάθε ενδιάμεση κατάσταση είναι ανεκτό αποτέλεσμα. Για την ποσοτικοποίηση της συγκεκριμένης μετρικής γίνεται υπολογισμός του εμβαδού κάτω από την αντίστοιχη καμπύλη όπου η ιδανική περίπτωση έχει εμβαδόν 1 και η τυχαία 0.5.

3. Βιβλιογραφική ανασκόπηση

3.1 Σχετικές εργασίες

Στο πλαίσιο της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε ανασκόπηση της βιβλιογραφίας σχετικά με τεχνικές ανίχνευσης της COVID-19 από ηχητικές καταγραφές. Συγκεκριμένα, εξετάστηκαν διάφορες μελέτες και ερευνητικές εργασίες που προτείνουν μεθόδους και αλγορίθμους για την ανίχνευση της ασθένειας μέσω αναλύσεων ήχου. Η ανασκόπηση αυτή είχε ως στόχο να εντοπίσει τις πιο πρόσφατες και καινοτόμες προσεγγίσεις, να αξιολογήσει την αποτελεσματικότητά τους και να αναδείξει τις προκλήσεις και τις δυνατότητες βελτίωσης στον τομέα αυτό.

Voice Features of Sustained Phoneme as COVID-19 Biomarker

Η εργασία αυτή στοχεύει στη διερεύνηση της αποτελεσματικότητας διαφορετικών φωνημάτων και χαρακτηριστικών φωνής στην ανίχνευση ατόμων που έχουν μολυνθεί από τον COVID-19, ειδικά σε ασθενείς με συμπτώματα στο αναπνευστικό σύστημα. Ένα από τα βασικά ζητήματα που επισημαίνονται στην έρευνα είναι ότι η χρήση χαρακτηριστικών φωνής για την ταυτοποίηση ασθενειών μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα λόγω δημογραφικών, υποκειμενικών και ακουστικών παραγόντων προκατάληψης. Αυτές οι προκαταλήψεις μπορεί να περιλαμβάνουν τη διακύμανση της φωνής ανάλογα με την ηλικία, το φύλο, το υγιές ή παθολογικό ιστορικό του ατόμου, καθώς και θορύβους στο περιβάλλον όπου γίνονται οι ηχογραφήσεις. Για να περιορίσουν αυτές τις προκαταλήψεις, οι συγγραφείς αποφάσισαν να χρησιμοποιήσουν και να εξάγουν χαρακτηριστικά μόνο από παρατεταμένα φωνήματα, ώστε να εξασφαλίσουν σταθερά δεδομένα.

Στην εργασία, δόθηκε έμφαση σε έξι συγκεκριμένα φωνήματα: /a/, /e/, /i/, /o/, /u/, και /m/, που επιλέχθηκαν για να μελετηθούν διαφορετικές πτυχές της παραγωγής φωνής. Οι συμμετέχοντες κλήθηκαν να παράγουν αυτά τα φωνήματα για όσο διάστημα ήταν άνετο, με φυσικό τόνο και ένταση, κρατώντας τη φωνή τους όσο το δυνατόν πιο σταθερή. Η συλλογή δεδομένων πραγματοποιήθηκε μέσω μιας ειδικά σχεδιασμένης εφαρμογής Android, η οποία χρησιμοποιούσε μικρόφωνα χαμηλής συχνότητας δειγματοληψίας (8 kHz), ώστε να είναι λειτουργική ακόμα και σε λιγότερο ανεπτυγμένες περιοχές. Η διάρκεια των ηχογραφήσεων κυμαινόταν από 3 έως 15 δευτερόλεπτα.

Η μελέτη περιλάμβανε συνολικά 40 ασθενείς με COVID-19 και 48 υγιείς συμμετέχοντες, οι οποίοι ηχογράφησαν τα φωνήματά τους κατά τη διάρκεια της νοσηλείας τους, μερικές φορές καθημερινά, ανάλογα με την κατάσταση της υγείας τους. Συλλέχθηκαν 1 έως 3 ημέρες ηχογραφήσεων ανά ασθενή, παρέχοντας δεδομένα για τη στατιστική ανάλυση και τον ταξινομητή.

Για την ανάλυση, εξήχθησαν συνολικά 34 χαρακτηριστικά φωνής από τις ηχογραφήσεις, τα οποία περιλάμβαναν παραμέτρους όπως Jitter, Shimmer, formants, MFCCs και Intensity-SD. Στη συνέχεια, έγινε στατιστική ανάλυση για να εντοπιστούν τα χαρακτηριστικά που ήταν πιο αποτελεσματικά για την

ανίχνευση του COVID-19, ανάλογα με το φώνημα που εξετάστηκε. Μετά τη στατιστική ανάλυση, τα χαρακτηριστικά αυτά δοκιμάστηκαν σε έναν ταξινομητή SVM (Support Vector Machine) για την αξιολόγηση της δυνατότητάς τους να διαφοροποιήσουν τους ασθενείς με COVID-19 από τους υγιείς συμμετέχοντες.

Τα αποτελέσματα έδειξαν ότι τα χαρακτηριστικά που σχετίζονται με τη διαμόρφωση του φωνητικού σωλήνα (όπως τα MFCCs, οι formants και το Vocal Tract Length - VTL) και τη σταθερότητα των αναπνευστικών μυών και του όγκου των πνευμόνων (όπως το Intensity-SD) ήταν τα πιο ευαίσθητα στις αλλαγές στη φωνή που προκάλεσε ο COVID-19. Σύμφωνα με τη στατιστική ανάλυση, τα χαρακτηριστικά που εξάγονται από το φώνημα /i/ κατά τις πρώτες τρεις ημέρες μετά τη διάγνωση του ασθενή ήταν τα πιο αποτελεσματικά για τη διαφοροποίηση των ασθενών από τους υγιείς συμμετέχοντες. Ειδικότερα, το μοντέλο SVM με χρήση 18 χαρακτηριστικών από το φώνημα /i/ έφτασε σε accuracy 93.5% και F1 score 94.3%. Αυτό υποδεικνύει ότι το φώνημα /i/ είναι το πιο αξιόπιστο για την ανίχνευση του COVID-19.

Ένα άλλο σημαντικό εύρημα ήταν ότι ενώ τα φωνήματα /a/ και /u/ εμφάνισαν επίσης διαφορές, το φώνημα /i/ απέδωσε τις καλύτερες επιδόσεις. Το /i/ είναι ένα φώνημα που απαιτεί ακριβή έλεγχο της γλώσσας και του φωνητικού σωλήνα, κάτι που το καθιστά πιο ευαίσθητο στις αλλαγές που προκαλούνται από τη νόσο, όπως η φλεγμονή ή η πίεση στους μύες του φωνητικού συστήματος. Αντίθετα, το φώνημα /a/, το οποίο χρησιμοποιείται ευρέως σε άλλες μελέτες, απαιτεί λιγότερο ακριβή έλεγχο και δεν είναι τόσο ευαίσθητο στις αλλαγές αυτές.

Επιπλέον, τα ευρήματα της μελέτης έδειξαν ότι τα φωνήματα που καταγράφηκαν τις πρώτες ημέρες μετά τη διάγνωση του COVID-19 ήταν τα πιο ενδεικτικά για την ανίχνευση της νόσου. Παρόλο που η στατιστική ανάλυση έδειξε ότι τα χαρακτηριστικά των ηχογραφήσεων από την 4η έως την 6η ημέρα ήταν τα πιο σημαντικά, το SVM ταξινομητικό μοντέλο έδειξε καλύτερη απόδοση με τα δεδομένα των πρώτων τριών ημερών. Αυτό δείχνει ότι οι πρώτες ημέρες μετά τη μόλυνση είναι πιθανώς οι πιο κρίσιμες για την εκδήλωση των συμπτωμάτων που επηρεάζουν τη φωνή.

Συμπερασματικά, η μελέτη υποστηρίζει ότι τα χαρακτηριστικά φωνής, ιδιαίτερα από το φώνημα /i/, μπορούν να χρησιμοποιηθούν ως αξιόπιστοι βιοδείκτες για τον COVID-19, παρέχοντας μια πολλά υποσχόμενη προσέγγιση για τη γρήγορη, μη επεμβατική και αυτοδιαχειριζόμενη ανίχνευση της νόσου. Αυτά τα αποτελέσματα μπορούν να αξιοποιηθούν για την ανάπτυξη συσκευών ή διαδικασιών ανίχνευσης μέσω φωνητικής ανάλυσης, κάτι που μπορεί να είναι ιδιαίτερα χρήσιμο σε περιβάλλοντα με περιορισμένη πρόσβαση σε παραδοσιακές μεθόδους διάγνωσης. Ωστόσο, οι συγγραφείς αναγνωρίζουν την ανάγκη για περαιτέρω έρευνα, με μεγαλύτερο αριθμό συμμετεχόντων και πιο ελεγχόμενες συνθήκες, για να επιβεβαιωθούν τα ευρήματα και να αναπτυχθούν εργαλεία που θα μπορούσαν να χρησιμοποιηθούν σε πραγματικές συνθήκες.

Παρά τη σημασία των αποτελεσμάτων, η μελέτη είχε και ορισμένα μειονεκτήματα. Πρώτον, ο αριθμός των συμμετεχόντων ήταν σχετικά μικρός, γεγονός που ενδέχεται να επηρεάζει τη γενίκευση των αποτελεσμάτων σε μεγαλύτερο πληθυσμό. Δεύτερον, οι ηχογραφήσεις έγιναν σε νοσοκομειακό περιβάλλον με ποικίλα επίπεδα θορύβου, κάτι που θα μπορούσε να επηρεάσει την ποιότητα των

δεδομένων. Επίσης, λόγω της κατάστασης της υγείας των ασθενών, δεν ήταν πάντα εφικτό να πραγματοποιούνται καθημερινές ηχογραφήσεις. Τέλος, οι ηχογραφήσεις συλλέχθηκαν αφού οι ασθενείς είχαν ήδη διαγνωστεί θετικοί στον COVID-19 μέσω RT-PCR, οπότε δεν μπορούμε να γνωρίζουμε αν τα χαρακτηριστικά της φωνής θα μπορούσαν να ανιχνεύσουν τη μόλυνση πριν από την εμφάνιση συμπτωμάτων. [29]

A Comprehensive Review on COVID-19 Cough Audio Classification through Deep Learning

Η έρευνα που παρουσιάζεται επιχειρεί να διερευνήσει εκτενώς τη χρήση τεχνικών βαθιάς μάθησης για την ταξινόμηση του ήχου βήχα που σχετίζεται με την COVID-19. Διερευνά συστηματικά τις διάφορες μεθοδολογίες που χρησιμοποιούνται για την εξαγωγή χαρακτηριστικών και την εκπαίδευση των μοντέλων, ρίχνοντας φως στα πλεονεκτήματα και τις αδυναμίες αυτών των προσεγγίσεων.

Το σύνολο δεδομένων που χρησιμοποιείται, γνωστό ως CoughVid, περιλαμβάνει περισσότερες από 25.000 ηχογραφήσεις βήχα που συγκεντρώθηκαν από διάφορες πηγές. Αυτές οι ηχογραφήσεις καλύπτουν ένα ευρύ φάσμα δημογραφικών χαρακτηριστικών συμμετεχόντων, όπως ηλικία, φύλο, γεωγραφική τοποθεσία και κατάσταση COVID-19. Για να ενισχυθεί η αξιοπιστία του συνόλου δεδομένων, ένα υποσύνολο 2.800 ηχογραφήσεων επισημάνθηκε προσεκτικά από τέσσερις έμπειρους ιατρούς, αποτελώντας ένα από τα πιο εκτεταμένα σύνολα δεδομένων βήχα με ετικέτες από ειδικούς που είναι διαθέσιμο.

Στην προεπεξεργασία των ηχητικών σημάτων, πραγματοποιούνται κοπή και αφαίρεση θορύβου για να διασφαλιστεί η ομοιομορφία και η διαχειρισιμότητα στα επόμενα στάδια επεξεργασίας, καθώς και για τη βελτίωση της ποιότητας των ηχογραφήσεων βήχα. Η διαδικασία κοπής περιλαμβάνει την απομάκρυνση ανεπιθύμητων τμημάτων του σήματος, όπως σιωπές στην αρχή και στο τέλος της εγγραφής. Παράλληλα, για τη βελτίωση της ποιότητας των ηχογραφήσεων βήχα, εφαρμόζεται αφαίρεση θορύβου, η οποία περιλαμβάνει τη χρήση τεχνικών όπως το φιλτράρισμα συχνοτήτων και την προσαρμοστική αφαίρεση θορύβου, που στοχεύουν στη μείωση των εξωτερικών παρεμβολών και την αύξηση του λόγου σήματος προς θόρυβο.

Από τα προεπεξεργασμένα ηχητικά σήματα εξάγονται σημαντικά ακουστικά χαρακτηριστικά, όπως τονικό ύψος και ένταση, καθώς και χαρακτηριστικά φασματογραφήματος για την καταγραφή των συχνοτικών και χρονικών ιδιοτήτων των σημάτων βήχα. Τα Mel-spectrogram χρησιμοποιούνται για να τονιστούν οι σχετικές ζώνες συχνοτήτων που ευθυγραμμίζονται με την ανθρώπινη αντίληψη.

Χρησιμοποιήθηκαν διάφορα μοντέλα μεταφοράς μάθησης για την ταξινόμηση του ήχου βήχα, όπως AlexNet, VggNet, ResNet και EfficientNet. Η ανάλυση αυτών των μοντέλων κατέληξε στα ακόλουθα συμπεράσματα.

Η ανάλυση κατέληξε στο συμπέρασμα ότι τα ακουστικά χαρακτηριστικά είναι αποτελεσματικά στην καταγραφή βασικών ιδιοτήτων των ηχητικών σημάτων και χρησιμοποιούνται ευρέως στην παραδοσιακή επεξεργασία ήχου. Ωστόσο, έχουν περιορισμούς στην ανίχνευση πιο σύνθετων μοτίβων που απαιτούν πιο προχωρημένες μεθόδους ανάλυσης. Τα φασματογραφήματα προσφέρουν καλύτερη σύλληψη χρονικών και φασματικών πληροφοριών, παρέχοντας μια ολοκληρωμένη αναπαράσταση του ήχου. Ωστόσο, μπορεί να μην καταγράφουν λεπτομέρειες επειδή η ανάλυσή τους δεν είναι αρκετά ευέλικτη για να προσαρμόζεται στις διάφορες λεπτές αλλαγές στον ήχο. Τα Mel-spectrogram δίνουν έμφαση στις σχετικές ζώνες συχνότητας ευθυγραμμισμένες με την ανθρώπινη αντίληψη, καθιστώντας τα πιο κατάλληλα για την ανίχνευση και κατηγοριοποίηση των ήχων βήχα, παρότι υπάρχει απώλεια λεπτομερών συχνοτήτων και περιορισμοί στη σταθερή ανάλυση χρόνου και συχνότητας.

Το AlexNet, αν και αρχικά σχεδιασμένο για εξαγωγή χαρακτηριστικών από εικόνες, μπορεί να προσαρμοστεί για χρήση με ηχητικά δεδομένα μέσω της χρήσης φασματογραφημάτων. Τα φασματογραφήματα είναι οπτικές αναπαραστάσεις των ηχητικών σημάτων που μετατρέπουν τα δεδομένα ήχου σε εικόνες, επιτρέποντας στο AlexNet να εφαρμόσει τις δυνατότητές του στην ανάλυση αυτών των εικόνων. Έτσι, το AlexNet μπορεί να εκπαιδευτεί σε φασματογράμματα που προέρχονται από ήχους βήχα για να ανιχνεύσει χαρακτηριστικά που σχετίζονται με τον COVID-19. Παρόλο που είναι αποτελεσματικό στην εξαγωγή χαρακτηριστικών από αυτά τα φασματογράμματα, το AlexNet παραμένει σχετικά μεγάλο και υπολογιστικά δαπανηρό, ενώ μπορεί να είναι επιρρεπές σε υπερεκπαίδευση όταν χρησιμοποιείται με μικρότερα σύνολα δεδομένων.

Το VGG διαθέτει μια απλοποιημένη αρχιτεκτονική με ομοιόμορφα μεγέθη φίλτρων. Αυτό σημαίνει ότι τα φίλτρα που χρησιμοποιούνται σε κάθε layer του δικτύου έχουν το ίδιο μέγεθος, κάτι που απλοποιεί την κατασκευή και τη συντήρηση του μοντέλου. Η ομοιομορφία αυτή καθιστά την αρχιτεκτονική πιο εύκολη στην κατανόηση και στην υλοποίηση, διότι οι κανόνες που εφαρμόζονται σε κάθε layer παραμένουν συνεπείς σε όλο το δίκτυο. Ως αποτέλεσμα, το VGG είναι ευκολότερο να σχεδιαστεί και να εκπαιδευτεί σε σχέση με πιο πολύπλοκες αρχιτεκτονικές, διατηρώντας παράλληλα την ικανότητά του να εκτελεί αποτελεσματική εξαγωγή χαρακτηριστικών από τα δεδομένα εισόδου. Ωστόσο, η βαθύτερη αρχιτεκτονική του οδηγεί σε αυξημένη υπολογιστική πολυπλοκότητα και απαιτήσεις μνήμης.

Το ResNet είναι αποτελεσματικό στην ανίχνευση λεπτών αποχρώσεων στα ηχητικά σήματα, αλλά οι πολύπλοκες αρχιτεκτονικές του μπορεί να οδηγήσουν σε υπερεκπαίδευση και αυξημένες υπολογιστικές απαιτήσεις. Το EfficientNet επιτυγχάνει υψηλή ακρίβεια με λιγότερες παραμέτρους και προσφέρει αποτελεσματική κλιμάκωση, αλλά απαιτεί προσεκτική ισορροπία των συντελεστών κλιμάκωσης και μπορεί να μην αποδίδει εξίσου καλά σε εξειδικευμένες εργασίες.

Συνοψίζοντας, η εργασία συμπεραίνει ότι το Mel-Spectrogram είναι ίσως το πιο ελπιδοφόρο χαρακτηριστικό για να αξιοποιηθεί από τα μοντέλα, καθώς δίνει έμφαση στις σχετικές ζώνες συχνοτήτων που ευθυγραμμίζονται με την ανθρώπινη αντίληψη, παρέχοντας έτσι μια πιο ακριβή και αποτελεσματική ανάλυση του ήχου βήχα.[30]

MSCCov19Net: multi-branch deep learning model for COVID-19 detection from cough sounds

Η εργασία που παρουσιάζεται στο συγκεκριμένο paper επιδιώκει την ανάπτυξη ενός εργαλείου ανίχνευσης COVID-19 βασισμένου στη βαθιά μάθηση, χρησιμοποιώντας ηχητικά σήματα βήχα. Πιο συγκεκριμένα, αναπτύχθηκαν τέσσερα εναλλακτικά μοντέλα ανίχνευσης: στο πρώτο μοντέλο χρησιμοποιούνται ως είσοδοι τα χαρακτηριστικά MFCC, στο δεύτερο τα χαρακτηριστικά φασματογράφημα, στο τρίτο τα χαρακτηριστικά Chromagram, και στο τέταρτο χρησιμοποιούνται συνδυαστικά όλα τα προηγούμενα χαρακτηριστικά. Το συνδυαστικό μοντέλο ονομάζεται MSCCov19Net.

Για το μοντέλο MFCC, εξήχθησαν 39 συντελεστές MFC από κάθε ηχητικό σήμα βήχα και χρησιμοποιήθηκαν ως είσοδος σε ένα μικρό MLP με τέσσερα επίπεδα. Το μοντέλο φασματογραφήματος εξήγαγε χαρακτηριστικά από φασματογραφήματα, από τα οποία στη συνέχεια προέκυψαν τα MFCC. Αυτά στη συνέχεια χρησιμοποιήθηκαν ως είσοδος σε ένα μικρό CNN εμπνευσμένο από το δίκτυο VGG. Το μοντέλο Chromagram εξήγαγε chroma-based χαρακτηριστικά από τα ηχητικά σήματα και χρησιμοποιήθηκε μια παρόμοια αρχιτεκτονική MLP με το μοντέλο MFCC. Τέλος, το συνδυαστικό μοντέλο MSCCov19Net συνδύασε χαρακτηριστικά από όλους τους προηγούμενους τομείς και χρησιμοποίησε μια πολυκλαδική αρχιτεκτονική CNN.

Για την εκπαίδευση και την αξιολόγηση των μοντέλων, χρησιμοποιήθηκαν τα σύνολα δεδομένων Coswara και Coughvid. Τα δεδομένα χωρίστηκαν σε ομάδες εκπαίδευσης, επαλήθευσης και δοκιμής με αναλογία 80%-10%-10%. Επιπλέον, τα σύνολα δεδομένων Virufy και NoCoCoDa χρησιμοποιήθηκαν μόνο στη φάση δοκιμής και επαλήθευσης, χωρίς να χρησιμοποιηθούν κατά τη φάση εκπαίδευσης.

Οι μετρικές που χρησιμοποιήθηκαν για τη σύγκριση των μοντέλων περιλαμβάνουν την accuracy και την ROC-AUC. Το μοντέλο MSCCov19Net παρουσίασε την καλύτερη απόδοση, με αύξηση 4.5% στο accuracy ταξινόμησης και 2.9% στην βαθμολογία ROC-AUC σε σύγκριση με τη δεύτερη καλύτερη προσέγγιση (βασισμένη σε MFCC).

Το καλύτερο από τα προηγούμενα μοντέλα (MSCCov19Net) συγκρίθηκε στη συνέχεια με διάφορες αρχιτεκτονικές DL όπως το βασικό μοντέλο CNN, το ResNet50, το EfficientNetB0, το MobileNetV2 και το Xception. Τα αποτελέσματα έδειξαν ότι το MSCCov19Net επιτυγχάνει τις υψηλότερες βαθμολογίες ακρίβειας και ROC-AUC μεταξύ όλων των άλλων προσεγγίσεων.

Συνολικά, η εργασία αυτή παρουσιάζει μια ανθεκτική και αποτελεσματική προσέγγιση για την ανίχνευση COVID-19 μέσω ανάλυσης ηχητικών σημάτων βήχα, προσφέροντας βελτιωμένες επιδόσεις σε σύγκριση με τα μοντέλα που αναφέρθηκαν προηγουμένως στην εργασία. Για περισσότερες πληροφορίες, οι ενδιαφερόμενοι μπορούν να ανατρέξουν στο αντίστοιχο άρθρο. [31]

COVID-19 Detection Model with Acoustic Features from Cough Sound and Its Application

Η συγκεκριμένη εργασία προσπαθεί να αναπτύξει ένα διαγνωστικό μοντέλο για τον COVID-19 χρησιμοποιώντας ήχους βήχα. Βασιζόμενη σε προηγούμενες μελέτες που χρησιμοποίησαν ξεχωριστά τα σύνολα δεδομένων του Cambridge, Coswara, Virufy και COUGHVID, αυτή η εργασία επιδιώκει να ξεπεράσει τους περιορισμούς που προκύπτουν από την περιορισμένη ποσότητα δεδομένων. Για τον λόγο αυτό, χρησιμοποιήθηκαν όλα τα σύνολα δεδομένων Cambridge, Coswara και COUGHVID, και κατασκευάστηκε ένα υψηλής ποιότητας σύνολο δεδομένων μέσω προσεκτικής προεπεξεργασίας.

Κατά την προεπεξεργασία των δεδομένων, αφαιρέθηκαν δεδομένα σε περιπτώσεις όπου ο θόρυβος υπερίσχυε του ήχου του βήχα, η ποιότητα της εγγραφής ήταν πολύ κακή, ο θόρυβος στο παρασκήνιο (όπως συνομιλία) που αναμειγνυόταν με τον ήχο του βήχα ή ήταν δύσκολο να αναγνωριστεί ο ήχος του βήχα. Μετά από την επιθεώρηση, από τα 4200 αρχεία ήχου, επιλέχθηκαν τελικά 2049 αρχεία βήχα, αποτελούμενα από 1106 αρχεία βήχα από συμμετέχοντες θετικούς στον COVID-19, 530 από υγιείς και 413 από άτομα με συμπτώματα. Επίσης, πραγματοποιήθηκε επεξεργασία των δεδομένων ώστε να αφαιρεθούν περιττές φωνές και θόρυβοι, εξασφαλίζοντας ότι τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση του μοντέλου ήταν καθαρά.

Το σύνολο των χαρακτηριστικών που χρησιμοποιήθηκε περιλάμβανε MFCC, Δ-MFCC, Δ2-MFCC, spectral contrast, chroma, spectral flatness, spectral bandwidth, spectral roll-off, RMS energy, spectral centroid, zero-crossing rate και onset. Τα χαρακτηριστικά αυτά επιλέχθηκαν με βάση την απόσταση Bhattacharyya, η οποία μετρά πόσο καλά μπορούν να διαχωριστούν τα χαρακτηριστικά μεταξύ διαφορετικών κατηγοριών. Τα κορυφαία χαρακτηριστικά ήταν το MFCC, Δ2-MFCC, Δ-MFCC και spectral contrast.

Το προτεινόμενο μοντέλο συνδυάζει το ResNet-50 και ένα DNN. Το ResNet-50 εκπαιδεύτηκε με την εικόνα του Mel-spectrogram, ενώ το DNN εκπαιδεύτηκε με το σύνολο των χαρακτηριστικών που προαναφέρθηκαν. Οι έξοδοι των δύο κλάδων συνδέθηκαν και έγιναν η είσοδος σε ένα SLP όπου πραγματοποιήθηκε κανονικοποίηση παρτίδας και dropout, και τελικά η συνάρτηση sigmoid χρησιμοποιήθηκε για να καθορίσει αν ο ήχος του βήχα ανήκει σε άτομο θετικό ή αρνητικό στον COVID-19.

Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου περιλάμβαναν την ακρίβεια, την ευαισθησία, την εξειδίκευση και την ακρίβεια πρόβλεψης. Τα αποτελέσματα της μελέτης έδειξαν ότι το προτεινόμενο μοντέλο είχε accuracy 0.96, sensitivity 0.95, specificity 0.96 και F1-score 0.95, επιδεικνύοντας καλύτερη απόδοση σε σύγκριση με προηγούμενες μελέτες. Για περισσότερες πληροφορίες, οι ενδιαφερόμενοι μπορούν να ανατρέξουν στο αντίστοιχο άρθρο.[32]

A Cough-based deep learning framework for detecting COVID-19

Η παρούσα εργασία, εμπνευσμένη από τα challenges DiCOVA, στοχεύει να συμβάλει στη διάγνωση της COVID-19 μέσω ακουστικών σημάτων.

Τα challenges DiCOVA έχουν σχεδιαστεί για να παρέχουν σύνολα δεδομένων με ήχους βήχα, ομιλίας και αναπνοής από άτομα θετικά και αρνητικά στον COVID-19, με στόχο να αξιοποιηθούν αυτά τα δεδομένα σε μελέτες που προτείνουν συστήματα ανίχνευσης της COVID-19. Στο πλαίσιο αυτό, πολλές μελέτες έχουν εστιάσει στην ανάπτυξη συστημάτων που μπορούν να εντοπίσουν τη νόσο μέσω ακουστικών σημάτων.

Συγκεκριμένα, η παρούσα εργασία στοχεύει να αξιοποιήσει τους ήχους βήχα και να προτείνει ένα πλαίσιο για την ανίχνευση της COVID-19, χρησιμοποιώντας τα δεδομένα της Track-2 που δόθηκαν στη Δεύτερη Πρόκληση DiCOVA, η οποία έλαβε χώρα το 2021. Το σύνολο δεδομένων της Track-2 περιλαμβάνει ένα σύνολο Development με 965 ηχογραφήσεις και ένα σύνολο Blind Test με 471 ηχογραφήσεις. Όλες οι ηχογραφήσεις δεν είναι μικρότερες από 500 χιλιοστά του δευτερολέπτου και έχουν ηχογραφηθεί με διάφορους ρυθμούς δειγματοληψίας. Το σύνολο Development χρησιμοποιείται για εκπαίδευση και την επίτευξη του καλύτερου μοντέλου, ενώ το σύνολο Blind Test χρησιμοποιείται για την αξιολόγηση και σύγκριση της απόδοσης των υποβληθέντων συστημάτων.

Για την αξιολόγηση των συστημάτων, η κύρια μετρική που χρησιμοποιείται είναι η ROC-AUC, ενώ οι δευτερεύουσες μετρικές αξιολόγησης περιλαμβάνουν την Sensitivity και την Specificity.

Η μεθοδολογία που ακολουθήθηκε περιλαμβάνει την εξαγωγή handcrafted χαρακτηριστικά και embeddedχαρακτηριστικά από τις ηχογραφήσεις, τα οποία στη συνέχεια συνδυάζονται για να ληφθούν τα συνδυασμένα χαρακτηριστικά. Αυτά τα συνδυασμένα χαρακτηριστικά τροφοδοτούνται σε διαφορετικά μοντέλα ταξινόμησης για την ανίχνευση θετικών κρουσμάτων COVID-19. Τα handcrafted χαρακτηριστικά που χρησιμοποιούνται είναι τα: 64 Mel-frequency cepstral coefficients (MFCCs), 12 Chromatic (Chroma), 128 Mel Spectrogram (Mel), 1 Zero-Crossing rate, 1 Gender, and 1 Duration. Αυτά τα χαρακτηριστικά επιλέχθηκαν λόγω της ευρείας αποδοχής και ανθεκτικότητάς τους στην επεξεργασία ομιλίας.

Τα embedded χαρακτηριστικά εξάγονται από μια ευρεία γκάμα προεκπαιδευμένων μοντέλων, όπως YAMNet, Wave2Vec, TRILL και τα σύνολα χαρακτηριστικών COMPARE 2016. Για να προσδιοριστεί ο καλύτερος συνδυασμός χαρακτηριστικών, πραγματοποιήθηκαν εκτενείς πειραματισμοί και αξιολογήσεις. Συγκεκριμένα, χρησιμοποιήθηκε το σύνολο δεδομένων ανάπτυξης (test set) από το Track-2 της Second 2021 DiCOVA Challenge, το οποίο περιλαμβάνει ηχογραφήσεις βήχα.

Η διαδικασία περιελάμβανε την εξαγωγή και τον συνδυασμό διαφορετικών χαρακτηριστικών, τα οποία στη συνέχεια τροφοδοτήθηκαν στο μοντέλο LightGBM.

Αρχικά, κάθε τύπος χαρακτηριστικού (π.χ., μόνο τα χειροποίητα χαρακτηριστικά, μόνο τα χαρακτηριστικά από το YAMNet, κ.λπ.) αξιολογήθηκε ξεχωριστά. Στη συνέχεια, συνδυασμοί των χειροποίητων χαρακτηριστικών με χαρακτηριστικά ενσωμάτωσης (π.χ., χειροποίητα χαρακτηριστικά και

YAMNet, χειροποίητα χαρακτηριστικά και TRILL) αξιολογήθηκαν για να εντοπιστεί ο συνδυασμός που προσέφερε την καλύτερη απόδοση.

Τα αποτελέσματα των πειραμάτων έδειξαν ότι ο συνδυασμός των χειροποίητων χαρακτηριστικών και των χαρακτηριστικών ενσωμάτωσης από το προεκπαιδευμένο μοντέλο TRILL είχε την υψηλότερη απόδοση. Αυτός ο συνδυασμός κατέγραψε το υψηλότερο ROC-AUC και την καλύτερη ισορροπία μεταξύ ευαισθησίας και ειδικότητας, καθιστώντας τον τον πιο αποτελεσματικό για την ανίχνευση θετικών κρουσμάτων COVID-19.

Εφόσον βρέθηκε ο καλύτερος συνδυασμός χαρακτηριστικών, αξιολογήθηκαν τα διάφορα μοντέλα ταξινομήσης, το LightGBM, το Support Vector Machine (SVM), το Random Forest (RF), το Extra Trees Classifier (ETC) και το Multi-layer Perceptron (MLP).

Το LightGBM μοντέλο απέδωσε καλύτερα, πετυχαίνοντας υψηλότερα σκορ απόδοσης. Συγκεκριμένα, συμμετείχαν στη Track-2 της Δεύτερης Πρόκλησης DiCOVA, όπου υπέβαλαν το σύστημά τους και πέτυχαν το δεύτερο υψηλότερο σκορ ROC-AUC με 81.21, μόλις πίσω από το κορυφαίο σκορ ROC-AUC των 81.97. Επιπλέον, πέτυχαν σκορ Sensitivity 48.33, Accuracy 59.18 και F1-score 53.21, παρουσιάζοντας σημαντικές βελτιώσεις από τα βασικά συστήματα.

Για περισσότερες πληροφορίες, οι ενδιαφερόμενοι μπορούν να ανατρέξουν στο αντίστοιχο άρθρο.[33]

Machine learning approach for detecting Covid-19 from speech signal using Mel frequency magnitude coefficient

Οι λογοθεραπευτές έχουν ανακαλύψει συσχέτιση μεταξύ συγκεκριμένων ιατρικών καταστάσεων και των προκλήσεων που αντιμετωπίζουν τα άτομα με τις φωνές τους. Έχουν δείξει ότι ορισμένες ασθένειες, όπως ο καρκίνος, το άσθμα, η νόσος του Alzheimer, η νόσος του Parkinson, η κατάθλιψη, η σχιζοφρένεια προκαλούν φωνητικές διαταραχές. Η ανάλυση της ομιλίας με την τεχνητή νοημοσύνη δημιουργεί νέες προοπτικές στη βιομηχανία της υγειονομικής περίθαλψης. Ο Covid-19 είναι μια αναπνευστική ασθένεια που επηρεάζει την αναπνοή και τη φωνή, δημιουργώντας συμπτώματα όπως ξηρός βήχας, πονόλαιμος και ασυνήθιστα πρότυπα αναπνοής. Με βάση αυτά, η εργασία στοχεύει στη δημιουργία ενός μοντέλου ταξινόμησης που βασίζεται στη μηχανική μάθηση και έχει την ικανότητα να αναγνωρίζει αξιόπιστα αν ένα άτομο έχει Covid ή όχι, βασισμένο στην ομιλία του.

Για την εκπαίδευση και τον έλεγχο του μοντέλου, χρησιμοποιήθηκε η βάση δεδομένων Coswara, η οποία δημιουργήθηκε στο Indian Institute of Science (IISc) Bangalore. Η βάση δεδομένων περιλαμβάνει ηχογραφήσεις βήχα, αναπνοής και φωνητικών ήχων των συμμετεχόντων, καθώς και πληροφορίες για τη γενική υγεία, το φύλο, την ηλικία και προϋπάρχουσες χρόνιες ασθένειες, όπως ο διαβήτης και το άσθμα. Για την εκπαίδευση του μοντέλου χρησιμοποιήθηκαν δεδομένα αποτελούμενα από 2089 τυχαία δείγματα από κάθε κατηγορία. Για τον έλεγχο του μοντέλου, χρησιμοποιήθηκαν τα υπόλοιπα 486 δείγματα υγιών ατόμων και 168 δείγματα ατόμων με θετικά συμπτώματα Covid-19.

Η προτεινόμενη προσέγγιση για την ανάλυση και την κατηγοριοποίηση Covid και μη-Covid από ηχητικά σήματα περιλαμβάνει τρία στάδια: την προεπεξεργασία του ηχητικού σήματος, την εξαγωγή χαρακτηριστικών και την ταξινόμηση. Στην προεπεξεργασία, οι ηχογραφήσεις κανονικοποιούνται, αναγνωρίζονται οι ηχητικές περιοχές και διαχωρίζονται από τις μη ηχητικές, προετοιμάζοντας έτσι τα δεδομένα για τα επόμενα στάδια ανάλυσης.

Για την εξαγωγή χαρακτηριστικών, χρησιμοποιούνται δύο κύρια χαρακτηριστικά: οι συντελεστές Mel Frequency Magnitude Coefficients (MFMC) και MFCC. Οι MFMC χρησιμοποιούν το μέγεθος του γρήγορου μετασχηματισμού Fourier (FFT) αντί του τετραγώνου του μεγέθους και δεν εφαρμόζουν τον διακριτό συνημιτονικό μετασχηματισμό (DCT). Αυτή η προσέγγιση παρέχει καλύτερη ανάλυση συχνότητας και λιγότερο συνολικό θόρυβο. Οι MFCC προκύπτουν μέσω μιας διαδικασίας που περιλαμβάνει διακριτούς μετασχηματισμούς Fourier, υπολογισμό φάσματος ισχύος, μετασχηματισμό κλίμακας Mel και DCT της λογαριθμικής ενέργειας.

Για την εκτέλεση της ταξινόμησης, χρησιμοποιήθηκαν τέσσερις διαφορετικοί ταξινομητές: οι Decision Tree, Random Forest, K-Nearest Neighbors και Support Vector Machine. Από αυτούς, οι K-Nearest Neighbors και Random Forest παρουσίασαν τις καλύτερες επιδόσεις.

Η απόδοση των ταξινομητών μετρήθηκε χρησιμοποιώντας την Sensitivity, την Specificity και ROC-AUC για να αξιολογηθεί η ικανότητά τους να διακρίνουν μεταξύ ασθενών με Covid και μη-Covid συμμετεχόντων.

Σύμφωνα με την εργασία, οι MFMC αποδίδουν καλύτερα από τους MFCC στην ανίχνευση του Covid-19. Για περισσότερες πληροφορίες, οι ενδιαφερόμενοι μπορούν να ανατρέξουν στο αντίστοιχο άρθρο. [34]

Rapid and Scalable COVID-19 Screening using Speech, Breath, and Cough Recordings

Προηγούμενες μελέτες έχουν δείξει ελπιδοφόρα αποτελέσματα στην ταυτοποίηση ασθενών θετικών στον COVID-19 μέσω ανάλυσης ήχου βήχα. Ωστόσο, δεν έχουν αντιμετωπίσει διάφορα ζητήματα και παραμένουν αναπάντητα ερωτήματα σχετικά με τη διαφοροποίηση μεταξύ βήχα για ασθενείς με χρόνιες ασθένειες, άλλους τύπους λοιμώξεων ή ασυμπτωματικούς ασθενείς. Επιπλέον, τα συστήματα ανίχνευσης COVID που απαιτούν αναγκαστικό βήχα από τους χρήστες παρέχουν σημαντικές ανησυχίες για την υγιεινή, καθώς αποτελούν σημαντικό φορέα μετάδοσης αναπνευστικών ασθενειών, καθιστώντας τα ακατάλληλα για χρήση σε δημόσιους χώρους.

Η ομιλία και η εκπνοή, αν και επηρεάζονται το ίδιο όπως ο βήχας από άλλες ασθένειες, παρέχουν σημαντικά λιγότερες ανησυχίες για την υγιεινή και είναι πολύ πιο φυσικοί ήχοι για παραγωγή και παρακολούθηση σε δημόσιους χώρους, ενεργά και παθητικά. Λαμβάνοντας όλα αυτά υπόψιν, ο στόχος της εργασίας είναι να να συγκρίνει την απόδοση της ομιλίας, της αναπνοής και του βήχα στην ανίχνευση του COVID-19.

Στην ανασκόπηση αυτή χρησιμοποιήθηκε η βάση δεδομένων Coswara η οποία περιλαμβάνει ηχογραφήσεις από άτομα θετικά ή αρνητικά στον COVID-19. Ξεκίνησε τον Απρίλιο του 2020 και επιτρέπει στα άτομα να συνεισφέρουν ηχογραφήσεις ήχου χρησιμοποιώντας τα smartphones ή τους υπολογιστές τους. Η βάση δεδομένων περιλαμβάνει συλλογές ήχου από γρήγορη και αργή αναπνοή, βαθύ και ρηχό βήχα, προφορά παρατεταμένων φωνηέντων και προφορά αριθμών σε δύο ταχύτητες, κανονική και γρήγορη. Επίσης, συλλέγονται πληροφορίες όπως η ηλικία, το φύλο, η γεωγραφική τοποθεσία, η τρέχουσα κατάσταση υγείας και προϋπάρχουσες ιατρικές καταστάσεις για να συμπληρώσουν τις ηχογραφήσεις.

Στην εργασία χρησιμοποίηθηκε ένα υποσύνολο της βάσης δεδομένων Coswara με δύο ομάδες. Η πρώτη ομάδα αποτελούνταν από ηχογραφήσεις ήχων βήχα, ενώ η δεύτερη ομάδα περιλάμβανε ηχογραφήσεις βαθιάς αναπνοής και προφοράς αριθμών. Η πρώτη ομάδα αποτελούνταν από 1040 συνολικά υποκείμενα, εκ των οποίων μόνο 75 ήταν θετικά στον COVID-19. Η δεύτερη ομάδα αποτελούνταν από 1199 συνολικά υποκείμενα, εκ των οποίων μόνο 80 ήταν θετικά στον COVID-19.

Διάφορες λειτουργίες εφαρμόστηκαν για την προετοιμασία των σημάτων για ανάλυση και ταξινόμηση. Εφαρμόστηκε κανονικοποίηση για να περιοριστούν όλες οι τιμές των δειγμάτων στο εύρος [-1,1]. Λόγω της σχεδόν σταθερής φύσης των ηχογραφήσεων ομιλίας, αναπνοής και βήχα, απαιτήθηκε ανάλυση μικρών χρονικών παραθύρων διάρκειας 25ms. Τέλος, έγινε ανίχνευση δραστηριότητας ομιλίας με τη χρήση κατωφλίων ενέργειας για την απομάκρυνση των σιωπηλών καρέ και τη διατήρηση αυτών με επαρκή ενέργεια.

Για την εκπαίδευση των ταξινομητών, χρησιμοποιήθηκαν τα ακόλουθα χαρακτηριστικά: οι συντελεστές MFCCs, οι μεταβολές των MFCC (MFCC deltas) και η μέθοδος RelAtive SpecTrA Perceptual Linear Prediction (RASTA-PLP). Αυτά τα χαρακτηριστικά επιλέχθηκαν λόγω της ευρείας χρήσης τους στην επεξεργασία ομιλίας και της επιτυχίας τους στην ανάλυση ήχων βήχα για την ανίχνευση του COVID-19.

Για την ταξινόμηση των ηχογραφήσεων φωνής ατόμων με COVID-19 και μη-COVID, χρησιμοποιήθηκαν δύο δημοφιλείς τεχνικές ταξινόμησης, Random Forest και τα DNN.

Για την αξιολόγηση της επιτυχίας των ταξινομητών στη διάκριση των ατόμων με COVID από τα άτομα χωρίς COVID, οι μετρικές απόδοσης που χρησιμοποιήθηκαν ήταν το ROC-AUC.

Από την ανάλυση των αποτελεσμάτων προέκυψε ότι ο συνδυασμός χαρακτηριστικών MFCC, MFCC deltas και RASTA-PLP είχε την καλύτερη απόδοση, σε σύγκριση με την αξιοποίηση μεμονωμένων χαρακτηριστικων, για την ομιλία και την αναπνοή, ανεξάρτητα από τον τύπο ταξινομητή που χρησιμοποιήθηκε. Η ομιλία είχε την υψηλότερη απόδοση σε σύγκριση με την αναπνοή και τον βήχα. Επιπλέον, οι ταξινομητές Random Forest (RF) και Deep Neural Network (DNN) παρουσίασαν παρόμοιες αποδόσεις, με τους RF να έχουν ελαφρώς καλύτερη απόδοση σε ορισμένες περιπτώσεις.

Εφόσον προέκυψαν τα παραπάνω αποτελέσματα, η εργασία προχωρά στη σύγκριση της προτεινόμενης μεθόδου, δηλαδή της ανίχνευσης COVID-19 μέσω ηχογραφήσεων ομιλίας με συνδυασμένα χαρακτηριστικά (MFCC, MFCC deltas και RASTA-PLP), με άλλες υπάρχουσες μεθόδους από

τη βιβλιογραφία. Η σύγκριση δείχνει ότι η προτεινόμενη μέθοδος χρησιμοποιώντας τον ταξινομητή Random Forest, είχε καλύτερη απόδοση από τις άλλες μεθόδους που περιγράφηκαν στην ανασκόπηση.

Για περισσότερες πληροφορίες, οι ενδιαφερόμενοι μπορούν να ανατρέξουν στο αντίστοιχο άρθρο. [35]

Exploring the Use of Artificial Intelligence Techniques to Detect the Presence of Coronavirus Covid-19 Through Speech and Voice Analysis

Ο στόχος της εργασίας είναι να εντοπιστεί η πιο αξιόπιστη τεχνική μηχανικής μάθησης (ML) για την ανίχνευση αλλαγών στη φωνή λόγω Covid-19 και να ενσωματωθεί αυτή η τεχνική σε μια έξυπνη λύση για κινητές συσκευές στον τομέα της υγείας, προκειμένου να διακριθούν με ακρίβεια άτομα που πάσχουν από Covid από υγιή άτομα, δίνοντας έμφαση στους ασυμπτωματικούς φορείς του ιού.

Η μεθοδολογία της εργασίας περιλαμβάνει την επιλογή κατάλληλων δειγμάτων φωνής από τη βάση δεδομένων Coswara. Για κάθε ηχητική καταγραφή, συλλέχθηκαν τα φωνήεντα /a/, /e/ και /o/, τα οποία στη συνέχεια επεξεργάστηκαν για την εξαγωγή κατάλληλων χαρακτηριστικών. Αυτά τα χαρακτηριστικά χρησιμοποιήθηκαν ως είσοδος στους αλγορίθμους μηχανικής μάθησης που αναλύθηκαν. Επιπλέον, δοκιμάστηκε και ένας συνδυασμός των χαρακτηριστικών που εξήχθησαν από τα τρία φωνήεντα, ο οποίος πέτυχε καλύτερα αποτελέσματα όσον αφορά τη σωστή ταξινόμηση μεταξύ υγιών και ασθενών ατόμων.

Για την ανάλυση από αλγορίθμους μηχανικής μάθησης, χρησιμοποιήθηκαν τα εξής χαρακτηριστικά: Fundamental Frequency (F0), jitter, shimmer, Harmonic to Noise Ratio (HNR), MFCC, οι πρώτες και δεύτερες παράγωγες των συντελεστών cepstral, το Spectral Centroid (SC) και το Spectral Roll-off (SR).

Για την ανάλυση χρησιμοποιήθηκαν οι ακόλουθοι αλγόριθμοι μηχανικής μάθησης: από την κατηγορία Bayes, οι Naive Bayes (NB) και BayesNet (BN), από την κατηγορία Functions, οι Support Vector Machine (SVM) και Stochastic Gradient Descent (SGD), από την κατηγορία Lazy, οι k-nearest neighbor (Ibk) και Locally Weighted Learning (LWL). Επιπλέον, από την κατηγορία Meta, οι Adaboost και Bagging, από την κατηγορία Rules, οι One-R και Decision Table (DT), και από την κατηγορία Trees, οι C4.5 decision tree (J48) και REPTree.

Οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν για να μετρήσουν την απόδοση των διάφορων αλγορίθμων μηχανικής μάθησης ήταν η accuracy, η specificity, το F1-score, η recall, το precision και ROC-A.

Τα αποτελέσματα και τα συμπεράσματα της μελέτης δείχνουν ότι οι αλγόριθμοι SVM και SGD πέτυχαν τα καλύτερα αποτελέσματα ταξινόμησης τόσο στα σύνολα εκπαίδευσης και επαλήθευσης, με την SVM να διακρίνει τα υγιή και ασθενή άτομα με accuracy περίπου 97%, specificity και recall περίπου 97% και 93% αντίστοιχα. Το αποτέλεσμα του F1-score, περίπου 82%, επιβεβαιώνει την αξιοπιστία του αλγορίθμου, ειδικά λόγω της μεγάλης ανισορροπίας μεταξύ των υγιών και των ασθενών δειγμάτων στο δεδομένο σύνολο.

Για περισσότερες πληροφορίες, οι ενδιαφερόμενοι μπορούν να ανατρέξουν στο αντίστοιχο άρθρο. [36]

Machine Learning based COVID-19 Detection from Smartphone Recordings: Cough, Breath and Speech

Ο στόχος της εργασίας είναι να διερευνήσει αν η ανάλυση των ηχητικών δεδομένων από τον βήχα, την αναπνοή και την ομιλία μπορεί να χρησιμοποιηθεί αποτελεσματικά για την ανίχνευση της COVID-19.

Το Coswara dataset αναπτύχθηκε ειδικά για τη δοκιμή αλγορίθμων ταξινόμησης για την ανίχνευση της COVID-19. Η συλλογή δεδομένων πραγματοποιείται μέσω διαδικτύου και οι συμμετέχοντες συνεισφέρουν χρησιμοποιώντας τα smartphones τους για να ηχογραφήσουν τον βήχα, την αναπνοή και την ομιλία τους. Συγκεκριμένα, οι ηχογραφήσεις περιλαμβάνουν μέτρημα από το ένα ως το είκοσι σε κανονικό και γρήγορο ρυθμό, καθώς και την εκφορά αγγλικών φωνηέντων. Επειδή έχουμε πολλά δείγματα από την αρνητική κλάση της COVID-19, χρησιμοποιείται η τεχνική εξισορρόπησης δεδομένων SMOTE, η οποία δημιουργεί συνθετικά δείγματα για να αυξήσει τα δείγματα της θετικής τάξης.

Το ComParE dataset χρησιμοποιείται ως μέρος της πρόκλησης Interspeech Computational Paralinguistics Challenge (ComParE) του 2021. Αυτό το σύνολο δεδομένων περιλαμβάνει ηχογραφήσεις βήχα και ομιλίας, με τη δεύτερη να περιλαμβάνει την εκφορά της φράσης «I hope my data can help to manage the virus pandemic» στη γλώσσα του ομιλητή. Οι συμμετέχοντες στο ComParE dataset συνεισφέρουν σε περισσότερες από τρεις διαφορετικές γλώσσες.

Στην εργασία αυτή, χρησιμοποιείται ένας συνδυασμός των Coswara και ComParE datasets. Η διαδικασία εκπαίδευσης και βελτιστοποίησης των ταξινομητών γίνεται με τη συγχώνευση των συνόλων δεδομένων εκπαίδευσης και αξιολόγησης, εφαρμόζοντας τη μέθοδο leave-p-out nested cross-validation. Η χρήση του SMOTE για το Coswara dataset είναι απαραίτητη λόγω της παρουσίας πολλών δειγμάτων από την αρνητική τάξη της COVID-19, γεγονός που μπορεί να επηρεάσει αρνητικά την απόδοση των νευρωνικών δικτύων.

Από τα ηχητικά σήματα εξάχθηκαν χαρακτηριστικά όπως οι MFCCs, οι log energies of linearly spaced filters, η velocity και η acceleration, καθώς και το ZCR και το kurtosis. Τα χαρακτηριστικά αυτά είναι χρήσιμα τόσο στην αυτόματη αναγνώριση ομιλίας όσο και στην ταξινόμηση ήχων βήχα.

Για την εξαγωγή των χαρακτηριστικών χρησιμοποιήθηκαν συγκεκριμένες παράμετροι, γνωστές ως feature extraction hyperparameters. Αυτές περιλαμβάνουν το μήκος των frames (F), τον αριθμό των segments (S), τον αριθμό των MFCCs (M) και τον αριθμό των linearly spaced filters (B). Συγκεκριμένα, χρησιμοποιήθηκαν μεταξύ 13 και 65 MFCCs και μεταξύ 40 και 200 linearly spaced filters για την εξαγωγή των log energies από το ηχητικό σήμα. Το μήκος του παραθύρου (δ) υπολογίζεται ως $\delta = \Lambda/S$, όπου Λ είναι το μήκος του ηχητικού σήματος σε δείγματα. Ο πίνακας χαρακτηριστικών που εισάγεται στους ταξινομητές έχει διαστάσεις (3M + 2, S), περιλαμβάνοντας τα MFCCs, τις παραμέτρους velocity και acceleration. Αυτή η διαδικασία εξαγωγής χαρακτηριστικών εξασφαλίζει ότι ολόκληρη η ηχογράφηση

αποτυπώνεται σε έναν σταθερό αριθμό frames, επιτρέποντας στα deep neural networks (DNN) να ανακαλύψουν πιο χρήσιμα χρονικά μοτίβα και να παρέχουν καλύτερη απόδοση στην ταξινόμηση.

Στην εργασία αυτή αξιολογήθηκαν επτά ταξινομητές: logistic regression (LR), support vector machines (SVM), multilayer perceptrons (MLP), k-nearest neighbour (KNN), convolutional neural networks (CNN), long short-term memory (LSTM) recurrent neural networks και ένα residual based network (Resnet50). Ο LR εφαρμόστηκε τόσο στο Coswara όσο και στο ComParE dataset για να παρέχει μια βασική γραμμή απόδοσης. Για το ComParE dataset, τα αποτελέσματα αναφέρονται μόνο για τους SVM, KNN και MLP ταξινομητές, καθώς οι deep neural networks (DNN) απέδωσαν ανεπαρκώς. Αυτό οφείλεται στον περιορισμένο αριθμό συμμετεχόντων σε αυτό το dataset, γεγονός που επηρεάζει αρνητικά την απόδοση των DNN. Από την άλλη πλευρά, για το Coswara dataset, αναφέρονται μόνο οι αποδόσεις των DNN, καθώς απέδωσαν σημαντικά καλύτερα από τους απλούστερους ταξινομητές.

Για την αξιολόγηση των ταξινομητών χρησιμοποιήθηκε η μετρική απόδοσης ROC-AUC.

Η logistic regression (LR) παρουσίασε παρόμοια απόδοση για την αναπνοή και την ομιλία στο Coswara dataset, με ROC-AUC μεταξύ 0.69 και 0.74. Στο ComParE dataset, η απόδοση της LR ήταν παρόμοια για τον βήχα και την ομιλία, με ROC-AUC μεταξύ 0.77 και 0.78.

Στην ταξινόμηση της ομιλίας, το Coswara dataset εμφάνισε AUC 0.85 για την κανονική και την γρήγορη ομιλία, που επιτεύχθηκε με το Resnet50 και το LSTM αντίστοιχα. Στο ComParE dataset, το MLP πέτυχε ROC-AUC 0.89.

Η ταξινόμηση της αναπνοής στο Coswara dataset παρουσίασε την καλύτερη απόδοση με την αρχιτεκτονική Resnet50, επιτυγχάνοντας ROC- AUC 0.92.

Για την ταξινόμηση του βήχα στο ComParE dataset, ο KNN ταξινομητής πέτυχε μέση ROC- AUC 0.85 χρησιμοποιώντας linearly spaced filters για την εξαγωγή log energy. Με την εφαρμογή του sequential forward selection (SFS) σε αυτόν τον ταξινομητή για την εύρεση των πιο σημαντικών χαρακτηριστικών, η ROC- AUC βελτιώθηκε σε 0.93 με την εξαγωγή των καλύτερων 12 χαρακτηριστικών.

Για το ComParE speech signal, η καλύτερη απόδοση (ROC- AUC 0.89) επιτεύχθηκε από τον MLP ταξινομητή χρησιμοποιώντας MFCC χαρακτηριστικά. Όταν εφαρμόστηκε το SFS σε αυτόν τον ταξινομητή, η AUC βελτιώθηκε σε 0.91 για τα καλύτερα 23 χαρακτηριστικά.

Μια άτυπη αξιολόγηση των dataset Coswara και ComParE δείχνει ότι το Coswara έχει μεγαλύτερη διακύμανση και περισσότερο θόρυβο σε σύγκριση με το ComParE. Αυτό αντικατοπτρίζεται στην απόδοση των ταξινομητών, όπου παρατηρείται μεγαλύτερη διακύμανση στην απόδοση για το Coswara. Είναι ενδιαφέρον ότι τα MFCCs είναι πάντα τα προτιμώμενα χαρακτηριστικά για το θορυβώδες Coswara, ενώ οι λογαριθμικές ενέργειες των γραμμικών φίλτρων συχνά προτιμώνται για τα λιγότερο θορυβώδη δεδομένα του ComParE. Μια παρόμοια παρατήρηση είχε γίνει και σε προηγούμενη μελέτη όπου οι βήχες καταγράφηκαν σε ελεγχόμενο περιβάλλον με ελάχιστο θόρυβο. Επιπλέον, η χρήση μεγαλύτερου αριθμού τμημάτων (segments) γενικά οδηγεί σε καλύτερη απόδοση, καθώς επιτρέπει στον ταξινομητή να αναγνωρίσει μοτίβα σε μικρότερα τμήματα του ηχητικού σήματος.

Συμπερασματικά, η μελέτη μας επιβεβαιώνει ότι η ανίχνευση COVID-19 μέσω ηχητικών εγγραφών βήχα, αναπνοής και ομιλίας από smartphones είναι εφικτή με υψηλή ακρίβεια με τον βήχα να φέρει τις ισχυρότερες ενδείξεις COVID-19, ακολουθούμενος από την αναπνοή και την ομιλία. Τα ποσοστά ROC-AUC για την ανίχνευση COVID-19 ήταν 0.93 για τον βήχα, 0.92 για την αναπνοή και 0.91 για την ομιλία.

Για περισσότερες πληροφορίες, οι ενδιαφερόμενοι μπορούν να ανατρέξουν στο αντίστοιχο άρθρο. [37]

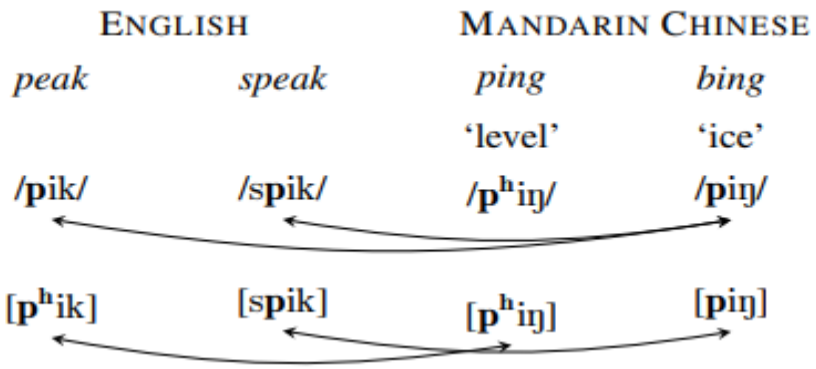
Αναλύοντας τις διάφορες μελέτες και ερευνητικές εργασίες που εξετάστηκαν στην παρούσα διπλωματική εργασία, γίνεται σαφές ότι η ανίχνευση της COVID-19 μέσω ηχητικών καταγραφών αποτελεί ένα πολλά υποσχόμενο πεδίο έρευνας με σημαντικές δυνατότητες εφαρμογής. Οι διάφορες τεχνικές που αναπτύχθηκαν, όπως η εξαγωγή χαρακτηριστικών από φασματογραφήματα, MFCC και MFMC, καθώς και η χρήση προηγμένων αλγορίθμων μηχανικής μάθησης και βαθιάς μάθησης, όπως τα CNN, ResNet, και LightGBM, αποδεικνύονται αποτελεσματικές στην ανίχνευση των χαρακτηριστικών της νόσου. Η συγκριτική αξιολόγηση των διαφορετικών προσεγγίσεων δείχνει ότι η συνδυαστική χρήση χαρακτηριστικών, όπως τα MFCC, Mel Spectrogram, και Chromagram, επιτρέπει την αξιοποίηση των πλεονεκτημάτων κάθε μεμονωμένου χαρακτηριστικού, ενώ η εφαρμογή τεχνικών μεταφοράς μάθησης ενισχύει την ικανότητα των μοντέλων να γενικεύουν και να αναγνωρίζουν μοτίβα σε νέα δεδομένα. Αυτή η πολυδιάστατη προσέγγιση μπορεί να οδηγήσει σε σημαντικές βελτιώσεις στην απόδοση των συστημάτων ανίχνευσης, καθιστώντας τα πιο ακριβή και αξιόπιστα.

3.2 Allosaurus

Υπάρχει αυξανόμενο ενδιαφέρον για την ανάπτυξη εργαλείων ομιλίας που ωφελούν τις γλώσσες με περιορισμένους πόρους, δηλαδή γλώσσες για τις οποίες δεν υπάρχουν πολλά διαθέσιμα δεδομένα ομιλίας και φωνητικής επισημείωσης (transcriptions), δηλαδή γραπτές αναπαραστάσεις της ομιλίας σε φωνήματα (phonemes). Συγκεκριμένα, οι ερευνητές επικεντρώνονται σε πολυγλωσσικά μοντέλα που μπορούν να βελτιώσουν την αναγνώριση ομιλίας σε αυτές τις γλώσσες, χρησιμοποιώντας τους πλούσιους πόρους που είναι διαθέσιμοι σε άλλες γλώσσες, όπως τα Αγγλικά και τα Μανδαρινικά.

Ένα τυπικό εργαλείο για την αναγνώριση σε γλώσσες με περιορισμένους πόρους είναι η πολυγλωσσική ακουστική μοντελοποίηση. Τα ακουστικά μοντέλα εκπαιδεύονται συνήθως σε παράλληλα δεδομένα κυμάτων ομιλίας και transcription επιτρέποντας έτσι τη χρήση γνώσης από γλώσσες με πλούσιους πόρους για τη βελτίωση της αναγνώρισης σε γλώσσες με περιορισμένους πόρους.

Είναι σημαντικό να σημειωθεί ότι τα φωνήματα είναι αντιληπτικές μονάδες ήχου που συσχετίζονται στενά με αλλά δεν αντιστοιχούν ακριβώς στους πραγματικούς ήχους που ακούγονται, δηλαδη τα phones. Ένα παράδειγμα αυτού φαίνεται στο Σχήμα, το οποίο δείχνει δύο αγγλικές λέξεις που μοιράζονται το ίδιο φώνημα /p/, αλλά διαφέρουν στις πραγματικές φωνητικές εκφορές [p] και [ph].



Εικόνα 3.1: Σχηματική απεικόνιση της αντιστοιχίας των φωνημάτων σε διαφορετικές γλώσσες.

Τα αλλόφωνα (allophones), δηλαδή οι διαφορετικές εκδοχές phones ενός συγκεκριμένου φωνήματος, είναι γλωσσικά εξειδικευμένα. Αυτό σημαίνει ότι οι διαφορές που είναι σημαντικές σε ορισμένες γλώσσες μπορεί να μην έχουν καμία σημασία σε άλλες.

Τα περισσότερα πολυγλωσσικά ακουστικά μοντέλα χρησιμοποιούν τα υπάρχοντα transcriptions όπως είναι, συνδυάζοντας τα σύνολα φωνημάτων για να δημιουργήσουν ένα κοινό σύνολο για όλες τις γλώσσες εκπαίδευσης. Αυτή η προσέγγιση είναι λογική υπό ορισμένες συνθήκες, επειδή τα ονόματα των φωνημάτων συνήθως σχετίζονται με το πιο κοινό ή το λιγότερο διαφοροποιημένο αλλόφωνό τους.

Ωστόσο, αυτή η άποψη είναι υπερβολικά απλοϊκή. Για παράδειγμα, στα Αγγλικά, τα phones [p] και [ph] αντιστοιχούν στο φώνημα /p/, κάτι που λειτουργεί καλά για την αγγλική γλώσσα. Όμως, αυτή η προσέγγιση μπορεί να δημιουργήσει προβλήματα όταν αναγνωρίζουμε τα Μανδαρινικά, όπου τα phones [p] και [ph] αντιστοιχούν σε δύο διαφορετικά φωνήματα /p/ και /ph/. Επομένως, η απλή ένωση των συνόλων φωνημάτων μπορεί να οδηγήσει σε ανακρίβειες στην αναγνώριση γλωσσών που διαχωρίζουν λεπτές φωνητικές διαφορές.

Η παρούσα εργασία θα επικεντρωθεί στα παρακάτω φωνήνετα με σύμβολα: [ɔ, i, u, ɛ, j,ο,α] τα οποία βασίζονται στον Παγκόσμιο Αλφάβητο Φωνημάτων. Γίνεται περιληπτική, επιγραμματική παρουσίαση του ήχου που συμβολίζεται από τα παραπάνω αλλά ο ενδιαφερόμενος αναγνώστης μπορεί να επισκεφθεί ιστοσελίδες που έχουν και ηχογραφήσεις (https://www.internationalphoneticalphabet.org/ipa-sounds/ipa-chart-with-sounds/).

1. ɔ: ήχος όμικρον με μερικώς ανοιχτό το στόμα και έμφαση στο πίσω μέρος της στοματικής κοιλότητας
2. i: ήχος 'γιώτα' με την γλωττίδα να ακουμπάει τον ουρανίσκο, το στόμα αρκετά κλειστό και έμφαση στο έμπροσθεν μέρος της στοματικής κοιλότητας.
3. u: ήχος του δίφθογγου 'ου' με την γλωττίδα να ακουμπάει τον ουρανίσκο, αρκετά κλειστό το στόμα και έμφαση στο πίσω μέρος της στοματικής κοιλότητας

4. ε: ήχος του ‘έψιλον’ με μερικώς ανοιχτό το στόμα, την γλωττίδα τοποθετημένη στην κάτω γνάθο και έμφαση στο μπροστινό αλλά λίγο πιο πίσω από την οδοντοστοιχία .
5. j: ήχος που μπορεί να συνοψιστεί από το γράμμα ‘γιώτα’ και τον δίφθογγο ‘ου’. Ένα παράδειγμα είναι το ‘γιούρια’. Είναι ήχος που παράγεται με κλειστό το στόμα, την γλωττίδα στο ενδιάμεσο μεταξύ άνω και κάτω γνάθου και έμφαση το κέντρο της στοματικής κοιλότητας.
6. ο: ήχος του ‘όμικρον’ με ανοιχτό το στόμα, την γλωττίδα τοποθετημένη στην κάτω γνάθο και έμφαση στο πίσω μέρος της στοματικής κοιλότητας
7. α: ήχος του ‘άλφα’ με ανοιχτό το στόμα, την γλωττίδα τοποθετημένη στην κάτω γνάθο και έμφαση στο πίσω μέρος της στοματικής κοιλότητας

Μια μέθοδος για πολυγλωσσική αναγνώριση βασισμένη σε φωνητική ανάλυση αντιμετωπίζει αυτό το πρόβλημα: το Allosaurus (Αλλόφωνο Σύστημα Αυτόματης Αναγνώρισης για Καθολική Ομιλία). Η μέθοδος αυτή ενσωματώνει τη γνώση της φωνολογίας στο πολύγλωσσο μοντέλο μέσω ενός αλλόφωνου επιπέδου, το οποίο συνδέει ένα καθολικό σύνολο phones με τα φωνήματα που εμφανίζονται στο transcription κάθε γλώσσας. Με άλλα λόγια, χρησιμοποιεί ένα κοινό σύνολο phones και το προσαρμόζει για να αντιστοιχεί στα φωνήματα κάθε γλώσσας. Έτσι, τα ίδια phones μπορούν να χρησιμοποιηθούν για διαφορετικές γλώσσες, λαμβάνοντας υπόψη τις φωνητικές διαφορές τους, βελτιώνοντας έτσι την ακρίβεια της αναγνώρισης ομιλίας.

Το μοντέλο του Allosaurus υπολογίζει πρώτα την κατανομή των phones χρησιμοποιώντας έναν τυπικό κωδικοποιητή ASRAutomatic Speech Recognition). Στη συνέχεια, το αλλόφωνο επίπεδο αντιστοιχίζει την κατανομή των phones στην κατανομή των φωνημάτων για κάθε γλώσσα. Αυτό σημαίνει ότι το μοντέλο αρχικά αναγνωρίζει και διανέμει τα phones από το ηχητικό σήμα και έπειτα τα συνδέει με τα αντίστοιχα φωνήματα για την κάθε γλώσσα, λαμβάνοντας υπόψη τις γλωσσικές ιδιαιτερότητες. Το μοντέλο του Allosaurus μπορεί να εκπαιδευτεί από άκρο σε άκρο, χρησιμοποιώντας μόνο φωνημικά transcriptions και έναν κατάλογο αλλόφωνων που δημιουργείται από ειδικούς.

Επιπλέον, η αρχιτεκτονική του Allosaurus καθιστά δυνατή την καθολική αναγνώριση phones. Προηγούμενες προσεγγίσεις δεν μπορούσαν να πραγματοποιήσουν αναγνώριση phones με καθολικό τρόπο, καθώς εξαρτώνται από τα φωνήματα της γλώσσας. Για παράδειγμα, στα Αγγλικά, τα phones [p] και [ph] δεν διακρίνονται όπως απαιτείται στα Μανδαρινικά. Αντιθέτως, αυτή η προσέγγιση επιτρέπει την αναγνώριση phones απευθείας, ανεξάρτητα από τη γλώσσα.

Οι δημιουργοί του μοντέλου του Allosaurus εκμεταλλεύτηκαν αυτό το γεγονός και ενσωμάτωσαν μια μεγάλη βάση δεδομένων phones, την PHOIBLE, η οποία έχει συλλεχθεί από γλωσσολόγους. Απέδειξαν ότι το μοντέλο αυτό μπορεί να προσαρμοστεί ώστε να αναγνωρίζει πάνω από 2000 γλώσσες, χωρίς να χρειάζεται δεδομένα εκπαίδευσης στις ίδιες τις γλώσσες. Αξιολόγησαν λοιπόν τον αναγνωριστή με γλώσσες που δεν είχαν χρησιμοποιηθεί καθόλου στην εκπαίδευση και διαπίστωσαν ότι επιτυγχάνει 17% καλύτερη απόδοση σε σύγκριση με την παραδοσιακή προσέγγιση.

Πώς Λειτουργεί το Μοντέλο του Allosaurus: Από τη Ανάλυση Phones στη Φωνημική Αναγνώριση

Ας υποθέσουμε ότι υπάρχουν |L| γλώσσες εκπαίδευσης και κάθε γλώσσα Li έχει το δικό της απόθεμα φωνημάτων Qi. Οι περισσότερες παραδοσιακές πολύγλωσσες προσεγγίσεις δουλεύουν με αυτά τα φωνήματα και δημιουργούν ένα κοινό σύνολο φωνημάτων Qsha ενώνοντας τα σύνολα φωνημάτων από όλες τις γλώσσες εκπαίδευσης.

Σε αντίθεση, η μέθοδος που ακολουθήθηκε για την υλοποίηση του μοντέλου του Allosaurus διαχωρίζει τα φωνήματα από τις φωνητικές τους υλοποιήσεις (phones). Οι γλωσσολόγοι αντιστοιχούν σε κάθε φώνημα (q) το αντίστοιχο σύνολο αλλοφώνων (Pi_q), όπου κάθε phone (p) που ανήκει στο (Pi_q) είναι μια συγκεκριμένη υλοποίηση του φωνήματος (q) στη γλώσσα (Li).

Για παράδειγμα, το φώνημα /p/ στα Αγγλικά μπορεί να έχει δύο αλλόφωνα, το [p] και το [ph]. Αυτά τα διαφορετικά phones αντιστοιχούν σε διαφορετικές φωνητικές υλοποιήσεις του ίδιου φωνήματος. Άρα, το φώνημα /p/ αντιστοιχεί στο σύνολο των αλλοφώνων του, δηλαδή [p] και [ph]. Αν το φώνημα /p/ έχει και άλλα phones σε άλλες γλώσσες, όπως [p'] ή [b], τότε το σύνολο που αντιστοιχίζεται σε αυτό το φώνημα θα περιλαμβάνει τα [p], [ph], [p'] και [b].

Συγχωνεύοντας αυτά τα σύνολα από όλες τις γλώσσες, δημιουργείται ένα καθολικό απόθεμα phones (Puni). Έτσι, καλύπτονται όλες οι φωνητικές υλοποιήσεις που εμφανίζονται σε οποιαδήποτε από τις γλώσσες εκπαίδευσης.

Ο γλωσσικά ανεξάρτητος κωδικοποιητής (language-independent encoder) του Allosaurus παράγει μια κατανομή πιθανότητας, η οποία ονομάζεται h, πάνω από το καθολικό απόθεμα των phones (Puni). Αυτό σημαίνει ότι ο κωδικοποιητής παίρνει την είσοδο ομιλίας και υπολογίζει τις πιθανότητες για κάθε phone στο καθολικό απόθεμα, ανεξάρτητα από τη γλώσσα στην οποία ανήκει η ομιλία. Με άλλα λόγια, ο κωδικοποιητής δεν ενδιαφέρεται για το ποια γλώσσα χρησιμοποιείται, αλλά απλώς για το ποια phones μπορεί να υπάρχουν στην ομιλία.

Για παράδειγμα, αν ένας ήχος που ακούγεται θα μπορούσε να είναι το phone [p], ο κωδικοποιητής θα υπολογίσει μια πιθανότητα για αυτό το phone και θα το ενσωματώσει στην κατανομή h. Η πιθανότητα αυτή εκφράζει την "βεβαιότητα" του μοντέλου ότι αυτός ο ήχος ακούγεται στην ομιλία. Μια υψηλή πιθανότητα σημαίνει ότι το μοντέλο είναι αρκετά σίγουρο ότι το [p] είναι παρόν στην ηχητική είσοδο, ενώ μια χαμηλή πιθανότητα σημαίνει ότι είναι λιγότερο πιθανό να περιέχει αυτόν τον ήχο. Επομένως, η συνολική κατανομή πιθανοτήτων h αντικατοπτρίζει την εκτίμηση του μοντέλου για το ποια phones μπορεί να υπάρχουν στην ηχητική είσοδο και σε ποιο βαθμό το κάθε ένα από αυτά είναι πιθανό να ακούγεται.

Η σημασία αυτής της κατανομής h έγκειται στη δυνατότητα του μοντέλου να τη μετασχηματίζει σε μια κατανομή φωνημάτων για κάθε συγκεκριμένη γλώσσα. Το αλλόφωνο επίπεδο (allophone layer), που αποτελεί μέρος του μοντέλου, χρησιμοποιεί την κατανομή h ως είσοδο για να δημιουργήσει την κατανομή φωνημάτων gi για κάθε γλώσσα (Li). Αυτό σημαίνει ότι το αλλόφωνο επίπεδο παίρνει τις

γενικές πληροφορίες σχετικά με τα phones που υπάρχουν στην ηχητική είσοδο και τις προσαρμόζει ώστε να αναγνωρίσει τα φωνήματα μιας συγκεκριμένης γλώσσας.

Για να το επιτύχει αυτό, το αλλόφωνο επίπεδο χρησιμοποιεί έναν εκπαιδεύσιμο πίνακα αλλοφώνων (Wi). Αυτός ο πίνακας είναι ουσιαστικά ένας πίνακας που περιέχει πληροφορίες για το πώς τα phones σχετίζονται με τα φωνήματα σε κάθε γλώσσα. Στην αρχή της εκπαίδευσης, ο πίνακας αρχικοποιείται με βάση έναν κατάλογο αλλοφώνων που έχουν δημιουργηθεί από γλωσσολόγους για κάθε γλώσσα. Αυτό σημαίνει ότι, στην αρχή, περιέχει τις γλωσσολογικές πληροφορίες για το ποια phones αντιστοιχούν σε ποια φωνήματα, όπως έχει καθοριστεί από τους γλωσσολόγους.

Κατά τη διάρκεια της εκπαίδευσης του μοντέλου, ο πίνακας Wi προσαρμόζεται περαιτέρω για να βελτιωθεί η ακρίβεια του μοντέλου. Αυτή η διαδικασία προσαρμογής, γνωστή και ως βελτιστοποίηση, επιτρέπει στο μοντέλο να μάθει ακριβέστερα τις σχέσεις μεταξύ των phones και των φωνημάτων για κάθε γλώσσα. Η προσαρμογή αυτή πραγματοποιείται κατά τη διάρκεια της εκπαίδευσης με τη βοήθεια δεδομένων εκπαίδευσης, δηλαδή πραγματικών δειγμάτων φωνητικής εισόδου και των αντίστοιχων μεταγραφών.

Ο πίνακας Wi επιτρέπει στο αλλόφωνο επίπεδο να μετασχηματίζει την κατανομή phones h που παράγεται από τον κωδικοποιητή, σε μια κατανομή φωνημάτων gi για κάθε γλώσσα. Με άλλα λόγια, χρησιμοποιώντας την αρχική γλωσσολογική γνώση και προσαρμόζοντας τον πίνακα Wi κατά την εκπαίδευση, το μοντέλο μπορεί να κατανοεί και να αναγνωρίζει τα φωνήματα σε κάθε γλώσσα με μεγαλύτερη ακρίβεια.

Εκμεταλλευόμενοι αυτή την ιδιότητα του κωδικοποιητή, οι δημιουργοί του Allosaurus ενσωμάτωσαν μια μεγάλη βάση δεδομένων phones, την PHOIBLE, η οποία έχει συλλεχθεί από γλωσσολόγους και περιλαμβάνει πληροφορίες για τις φωνητικές υλοποιήσεις πολλών γλωσσών. Η βάση δεδομένων PHOIBLE χρησιμοποιείται για την αρχικοποίηση της μήτρας Wi με τις γλωσσολογικές πληροφορίες. Αυτό διευκολύνει την προσαρμογή του μοντέλου για την αναγνώριση πάνω από 2000 γλώσσες, ακόμη και χωρίς δεδομένα εκπαίδευσης στις ίδιες τις γλώσσες.

Αξιολόγηση και Απόδοση του Μοντέλου Allosaurus στην Πολυγλωσσική Αναγνώριση Ομιλίας

Το μοντέλο Allosaurus αξιολογήθηκε χρησιμοποιώντας το σύνολο δεδομένων CommonVoice (CV) της Mozilla, που περιέχει διάφορες εκφωνήσεις από πολλούς ομιλητές σε πολλές γλώσσες. Εστίασαν σε 6 γλώσσες με περιορισμένους πόρους και 2 γλώσσες με πολλούς πόρους. Για κάθε γλώσσα, το σύνολο δεδομένων χωρίστηκε σε τμήματα για εκπαίδευση, επικύρωση και δοκιμές. Η απόδοση αναφέρθηκε με βάση τον ρυθμό σφάλματος φωνημάτων (PER). Το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας τον Adam με αρχικό ρυθμό μάθησης 0.001, εφαρμόζοντας dropout και data augm για να αυξηθεί η ποικιλία των δειγμάτων εκπαίδευσης.

Το μοντέλο Allosaurus αξιολογήθηκε και συγκρίθηκε με άλλα μοντέλα αναγνώρισης ομιλίας. Συγκεκριμένα, η αξιολόγηση έγινε σε έξι γλώσσες με περιορισμένους πόρους, δηλαδή γλώσσες για τις οποίες υπάρχουν λίγα διαθέσιμα δεδομένα εκπαίδευσης. Τα αποτελέσματα έδειξαν ότι το Allosaurus σημείωσε σημαντική βελτίωση στον ρυθμό σφάλματος φωνημάτων (PER) σε σύγκριση με τα προηγούμενα μοντέλα. Ο ρυθμός σφάλματος φωνημάτων (PER) είναι ένα μέτρο που δείχνει το ποσοστό των φωνημάτων που αναγνωρίστηκαν εσφαλμένα από το μοντέλο.

Σε αυτές τις έξι γλώσσες, το Allosaurus πέτυχε μείωση του σφάλματος κατά 2-5% σε σχέση με τα συμβατικά πολυγλωσσικά μοντέλα. Αυτό σημαίνει ότι το Allosaurus ήταν πιο ακριβές στην αναγνώριση φωνημάτων σε αυτές τις γλώσσες. Τα αποτελέσματα αυτά αποδεικνύουν την αποτελεσματικότητα του Allosaurus στην αναγνώριση ομιλίας σε γλώσσες με περιορισμένους πόρους. Η βελτιωμένη απόδοση του μοντέλου οφείλεται στη χρήση της γνώσης που αποκτήθηκε από γλώσσες με πολλούς πόρους, όπως τα Αγγλικά και τα Μανδαρινικά, και στην προσαρμογή αυτής της γνώσης στις γλώσσες με περιορισμένους πόρους.

Το Allosaurus επίσης αξιολογήθηκε για τη δυνατότητά του να αναγνωρίζει phones ανεξάρτητα από τη γλώσσα. Χρησιμοποιώντας την καθολική βάση δεδομένων phones (PHOIBLE), το μοντέλο μπορούσε να εντοπίσει phones σε νέες, αδόκιμες γλώσσες χωρίς να έχει εκπαιδευτεί σε αυτές. Οι δοκιμές πραγματοποιήθηκαν σε γλώσσες όπως τα Μπουρμέζικα, Βιετναμέζικα, και Ταϊλανδέζικα, που δεν υπήρχαν στο σύνολο εκπαίδευσης, και τα αποτελέσματα έδειξαν ότι το Allosaurus πέτυχε αναγνώριση με υψηλή ακρίβεια, αποδεικνύοντας την ικανότητά του να γενικεύει τη μάθηση των phones σε διαφορετικές γλώσσες.

Το Allosaurus προσφέρει μια καινοτόμο προσέγγιση στην αναγνώριση ομιλίας, συνδυάζοντας τη γνώση της φωνολογίας με πολυγλωσσική ακουστική μοντελοποίηση. Η χρήση ενός ανεξάρτητου από τη γλώσσα κωδικοποιητή, σε συνδυασμό με το αλλόφωνο επίπεδο και τη βάση δεδομένων PHOIBLE, επιτρέπει την αποτελεσματική αναγνώριση ομιλίας σε γλώσσες με περιορισμένους πόρους. Τα αποτελέσματα των δοκιμών δείχνουν ότι το μοντέλο πετυχαίνει σημαντικές βελτιώσεις στην αναγνώριση φωνημάτων και τη γενίκευση της αναγνώρισης phones σε νέες γλώσσες. Αυτές οι δυνατότητες καθιστούν το Allosaurus ένα πολύτιμο εργαλείο για την ενίσχυση της αναγνώρισης ομιλίας σε παγκόσμια κλίμακα. [38]

3.3 VGG

Η Εξέλιξη και η Επίδραση του VGG Μοντέλου στα Συνελικτικά Νευρωνικά Δίκτυα (CNNs)

Το VGG μοντέλο, που αναπτύχθηκε από τους Karen Simonyan και Andrew Zisserman του Visual Geometry Group (VGG) στο Πανεπιστήμιο της Οξφόρδης, είναι ένα από τα πιο γνωστά και επιτυχημένα

μοντέλα CNNs. Η αρχιτεκτονική του VGG παρουσιάστηκε στο paper "Very Deep Convolutional Networks for Large-Scale Image Recognition", το οποίο υποβλήθηκε στο ILSVRC-2014 (ImageNet Large Scale Visual Recognition Challenge).

Το ILSVRC (ImageNet Large Scale Visual Recognition Challenge) είναι ένας ετήσιος διαγωνισμός αναγνώρισης εικόνων μεγάλης κλίμακας, που ξεκίνησε το 2010. Ο στόχος του διαγωνισμού είναι η ανάπτυξη αλγορίθμων που μπορούν να ταξινομήσουν και να ανιχνεύσουν αντικείμενα σε εικόνες με ακρίβεια. Το ILSVRC χρησιμοποιεί το σύνολο δεδομένων ImageNet, το οποίο περιλαμβάνει εκατομμύρια εικόνες με χιλιάδες κατηγορίες αντικειμένων.

Κατά τη διάρκεια του διαγωνισμού, τα μοντέλα αξιολογούνται σε ένα σύνολο δεδομένων που δεν έχουν ξαναδεί, και η απόδοσή τους μετριέται χρησιμοποιώντας διάφορες μετρικές, όπως το σφάλμα top-1 και το σφάλμα top-5(γίνεται ανάλυση στις επόμενες παραγράφους). Ο διαγωνισμός αυτός έχει συμβάλει σημαντικά στην πρόοδο της έρευνας στην υπολογιστική όραση, καθώς παρέχει μια κοινή βάση σύγκρισης για την αξιολόγηση των αλγορίθμων.

Γενική Περιγραφή της Αρχιτεκτονικής VGG:

Κατά την εκπαίδευση του μοντέλου VGG, χρησιμοποιούνται εικόνες RGB με σταθερό μέγεθος 224x224 pixels. Αυτό σημαίνει ότι όλες οι εικόνες που εισάγονται στο δίκτυο πρέπει να έχουν αυτό το συγκεκριμένο μέγεθος. Για την προεπεξεργασία των εικόνων, αφαιρείται η μέση τιμή της έντασης των χρωμάτων RGB. Η μέση τιμή είναι η μέση ένταση των χρωμάτων κόκκινου, πράσινου και μπλε για κάθε pixel, η οποία υπολογίζεται από όλες τις εικόνες του συνόλου εκπαίδευσης. Συγκεκριμένα, υπολογίζεται ο μέσος όρος της τιμής κάθε χρώματος (κόκκινου, πράσινου και μπλε) για όλα τα pixels σε όλες τις εικόνες του συνόλου εκπαίδευσης και αυτή η μέση τιμή αφαιρείται από την τιμή κάθε χρώματος σε κάθε pixel της εικόνας. Αυτή η διαδικασία βοηθά στη σταθεροποίηση της εκπαίδευσης και στη βελτίωση της απόδοσης του μοντέλου, καθιστώντας το πιο ομοιόμορφο και βοηθώντας στην ταχύτερη σύγκλιση των παραμέτρων του μοντέλου κατά τη διάρκεια της εκπαίδευσης.

Η εικόνα στη συνέχεια περνά μέσα από μια σειρά από συνελικτικά επίπεδα, όπου χρησιμοποιούνται φίλτρα με μέγεθος 3x3. Αυτά τα φίλτρα είναι πολύ μικρά, αλλά επαρκή για να καταγράψουν βασικά χαρακτηριστικά της εικόνας, όπως άκρες και γωνίες. Σε μία από τις διαμορφώσεις του μοντέλου, χρησιμοποιούνται επίσης φίλτρα συνελίξεως 1x1, τα οποία λειτουργούν ως γραμμική μετατροπή των εισόδων των χρωματικών καναλιών (κόκκινο, πράσινο, μπλε) και ακολουθούνται από μια μη-γραμμική λειτουργία, όπως η ReLU. Με άλλα λόγια, τα φίλτρα 1x1 αλλάζουν την τιμή κάθε pixel βάσει γραμμικών συνδυασμών των εισερχόμενων καναλιών χρωμάτων και στη συνέχεια εφαρμόζουν μια μη-γραμμική λειτουργία για να εισαγάγουν πολυπλοκότητα και να βελτιώσουν τη δυνατότητα του δικτύου να κατανοεί πολύπλοκα χαρακτηριστικά. Το βήμα της συνελίξεως είναι σταθερό σε 1 pixel, το οποίο σημαίνει ότι το φίλτρο μετακινείται κατά ένα pixel κάθε φορά κατά τη διαδικασία της συνελίξεως.

Για να διασφαλιστεί ότι οι διαστάσεις της εικόνας παραμένουν αμετάβλητες μετά τη συνελίξη, χρησιμοποιείται συμπλήρωση (padding) 1 pixel γύρω από την εικόνα όταν εφαρμόζονται φίλτρα 3x3 στα

επίπεδα του νευρωνικού δικτύου. Αυτό σημαίνει ότι προστίθεται ένα πλαίσιο ενός pixel γύρω από την εικόνα, ώστε η έξοδος του συνελικτικού επιπέδου να έχει το ίδιο μέγεθος με την είσοδο. Με άλλα λόγια, η προσθήκη αυτού του πλαισίου εξασφαλίζει ότι η ανάλυση, δηλαδή το ύψος και το πλάτος της εικόνας, δεν μειώνονται μετά την εφαρμογή των φίλτρων 3x3, διατηρώντας έτσι τη χωρική διάσταση της εικόνας σταθερή.

Η διαδικασία max-pooling πραγματοποιείται σε πέντε επίπεδα του δικτύου. Το max-pooling χρησιμοποιεί ένα παράθυρο διαστάσεων 2x2 pixels και το βήμα του είναι 2, που σημαίνει ότι το παράθυρο μετακινείται κατά δύο pixels κάθε φορά. Αυτό το παράθυρο ελέγχει κάθε υποπεριοχή της εικόνας και επιλέγει την υψηλότερη τιμή (μέγιστη) από αυτές τις περιοχές, αντικαθιστώντας την υποπεριοχή με αυτήν την μέγιστη τιμή. Το max-pooling μειώνει τη χωρική διάσταση της εικόνας, δηλαδή το ύψος και το πλάτος της, διατηρώντας τα σημαντικότερα χαρακτηριστικά της. Αυτή η μείωση της διάστασης βοηθά στη μείωση της υπολογιστικής πολυπλοκότητας του μοντέλου, καθώς μικρότερες εικόνες απαιτούν λιγότερους υπολογισμούς. Επιπλέον, το max-pooling συμβάλλει στη μείωση του overfitting, καθώς βοηθά στη γενίκευση του μοντέλου, καθιστώντας το πιο ικανό να αναγνωρίζει σημαντικά χαρακτηριστικά από τις εικόνες χωρίς να εξαρτάται από συγκεκριμένες λεπτομέρειες ή θόρυβο.

Αυτή η προσεκτικά σχεδιασμένη αρχιτεκτονική βοηθά το μοντέλο VGG να καταγράφει και να επεξεργάζεται λεπτομέρειες σε πολλαπλά επίπεδα, δημιουργώντας μια πλούσια αναπαράσταση χαρακτηριστικών επιτρέπει την ακριβή αναγνώριση και ταξινόμηση εικόνων.

Μετά από μια σειρά συνελικτικών επιπέδων, τα οποία μπορεί να έχουν διαφορετικό αριθμό επιπέδων ανάλογα με την αρχιτεκτονική του δικτύου, το μοντέλο VGG ακολουθείται από MLP τριών επιπέδων. Αυτά τα επίπεδα είναι σχεδιασμένα για να συγκεντρώνουν και να επεξεργάζονται τα χαρακτηριστικά που εξάγονται από τα προηγούμενα συνελικτικά επίπεδα.

Το πρώτο και το δεύτερο πλήρως συνδεδεμένο επίπεδο περιέχουν από 4096 κανάλια η καθεμία. Αυτό σημαίνει ότι κάθε επίπεδο έχει 4096 νευρώνες που λαμβάνουν σήματα από όλους τους νευρώνες του προηγούμενου επιπέδου, επιτρέποντας την επεξεργασία και την κατηγοριοποίηση των εξαγόμενων χαρακτηριστικών σε πιο περίπλοκα μοτίβα.

Το τρίτο πλήρως συνδεδεμένο επίπεδο είναι υπεύθυνο για την τελική ταξινόμηση των δεδομένων σε 1000 κατηγορίες, όπως απαιτείται από τον διαγωνισμό ILSVRC. Περιέχει 1000 κανάλια, με κάθε κανάλι να αντιπροσωπεύει μία από τις κατηγορίες ταξινόμησης. Κάθε κανάλι στο τρίτο επίπεδο αντιστοιχεί σε μία συγκεκριμένη κατηγορία αντικειμένων (π.χ. γάτες, σκύλοι, αυτοκίνητα, κ.λπ.) και η τιμή του καναλιού αυτού δείχνει την πιθανότητα το εισαγόμενο δεδομένο να ανήκει σε αυτή την κατηγορία. Για παράδειγμα, αν το εισαγόμενο δεδομένο είναι μια εικόνα σκύλου, το κανάλι που αντιστοιχεί στην κατηγορία "σκύλος" θα έχει υψηλή τιμή, ενώ τα κανάλια που αντιστοιχούν σε άλλες κατηγορίες θα έχουν χαμηλότερες τιμές.

Το τελευταίο επίπεδο του μοντέλου είναι το επίπεδο soft-max, το οποίο μετατρέπει τα αποτελέσματα του τρίτου πλήρους συνδεδεμένου επιπέδου σε πιθανότητες. Κάθε πιθανότητα δείχνει την πιθανότητα

το δεδομένο εισόδου να ανήκει σε μία από τις 1000 κατηγορίες. Το επίπεδο soft-max εξασφαλίζει ότι το άθροισμα των πιθανοτήτων για όλες τις κατηγορίες είναι ίσο με 1, επιτρέποντας έτσι την ερμηνεία των εξόδων ως κανονισμένες πιθανότητες.

Η διαμόρφωση των πλήρως συνδεδεμένων επιπέδων παραμένει σταθερή σε όλες τις παραλλαγές του δικτύου VGG, διασφαλίζοντας συνεπή επεξεργασία και ταξινόμηση των χαρακτηριστικών ανεξαρτήτως του βάθους των συνελικτικών επιπέδων.

Διαδικασία Εκπαίδευσης και Προεπεξεργασίας Εικόνων για τα VGG Μοντέλα

Η εκπαίδευση των VGG μοντέλων ακολουθεί μια μεθοδολογία παρόμοια με εκείνη που περιγράφεται από τους Krizhevsky et al. (2012), με ορισμένες προσαρμογές. Το μέγεθος του batch για την εκπαίδευση ορίστηκε στα 256, ενώ το momentum ήταν 0.9. Το momentum, με τιμή 0.9, βοηθά στην επιτάχυνση της σύγκλισης και στην αποφυγή τοπικών ελαχίστων. Για να βελτιωθεί η γενίκευση του μοντέλου και να αποφευχθεί το overfitting, εφαρμόστηκε κανονικοποίηση με weight decay, με τον πολλαπλασιαστή της ποινής L2 να ορίζεται στο 5 × 10^−4, και κανονικοποίηση dropout στα δύο πρώτα πλήρως συνδεδεμένα επίπεδαμε λόγο 0.5.

Το learning rate αρχικά ορίστηκε στο 0.01 (10^−2) και μειώθηκε κατά έναν παράγοντα 10 όταν η ακρίβεια του συνόλου επικύρωσης σταμάτησε να βελτιώνεται. Συνολικά, η ταχύτητα εκμάθησης μειώθηκε 3 φορές κατά τη διάρκεια της εκπαίδευσης, και η διαδικασία εκπαίδευσης σταμάτησε μετά από 370.000 επαναλήψεις, που αντιστοιχούν σε 74 εποχές. Παρά τον μεγαλύτερο αριθμό παραμέτρων και το μεγαλύτερο βάθος των δικτύων VGG σε σύγκριση με το αρχικό μοντέλο των Krizhevsky et al., τα δίκτυα VGG απαιτούσαν λιγότερες εποχές για να συγκλίνουν. Αυτό μπορεί να αποδοθεί στην έμμεση κανονικοποίηση που επιβάλλεται από το μεγαλύτερο βάθος και τα μικρότερα μεγέθη φίλτρων συνελικτικών επιπέδων καθώς και στην προκαταρκτική αρχικοποίηση ορισμένων επιπέδων.

Η αρχικοποίηση των βαρών ενός NN είναι κρίσιμης σημασίας, καθώς μια κακή αρχικοποίηση μπορεί να εμποδίσει τη μάθηση λόγω της αστάθειας των βαθμίδων (gradients) σε βαθιά δίκτυα. Για να αποφύγουν αυτό το πρόβλημα, οι ερευνητές άρχισαν την εκπαίδευση με μια αρχιτεκτονική που ήταν αρκετά ρηχή ώστε να μπορεί να εκπαιδευτεί με τυχαία αρχικοποίηση των βαρών. Αυτή η αρχιτεκτονική περιλάμβανε λίγα συνελικτικά και πλήρως συνδεδεμένα επίπεδα, επιτρέποντας την αποτελεσματική εκπαίδευση ακόμα και με τυχαία αρχικοποίηση.

Όταν προχώρησαν στην εκπαίδευση βαθύτερων αρχιτεκτονικών, οι ερευνητές χρησιμοποίησαν τα εκπαιδευμένα βάρη από την αρχική ρηχή αρχιτεκτονική για τα πρώτα τέσσερα συνελικτικά επίπεδα και τα τελευταία τρία πλήρως συνδεδεμένα επίπεδα των βαθύτερων δικτύων. Αυτό σημαίνει ότι αντί να αρχικοποιηθούν τυχαία, τα επίπεδα αυτές ξεκίνησαν την εκπαίδευσή τους από τα ήδη εκπαιδευμένα βάρη της ρηχής αρχιτεκτονικής. Τα ενδιάμεσα επίπεδαπου προστέθηκαν στις βαθύτερες αρχιτεκτονικές αρχικοποιήθηκαν τυχαία.

Κατά την εκπαίδευση των βαθύτερων δικτύων, η ταχύτητα εκμάθησης δεν μειώθηκε για τα προκαταρκτικά αρχικοποιημένα επίπεδα επιτρέποντάς τους να προσαρμόζονται και να αλλάζουν κατά τη διάρκεια της μάθησης, βελτιώνοντας την απόδοσή τους.

Για την τυχαία αρχικοποίηση των βαρών, χρησιμοποιήθηκε μια κανονική κατανομή με μηδενική μέση τιμή και διασπορά 10−2, ενώ οι μεροληψίες (biases) αρχικοποιήθηκαν με μηδενική τιμή, διευκολύνοντας την έναρξη της μάθησης.

Για να διασφαλιστεί ότι όλες οι εικόνες εισόδου έχουν το ίδιο μέγεθος και είναι κατάλληλες για επεξεργασία από το δίκτυο, εφαρμόζεται η ακόλουθη διαδικασία προεπεξεργασίας. Αρχικά, κάθε εικόνα κλιμακώνεται έτσι ώστε η μικρότερη διάστασή της να είναι 256 pixels. Αυτό σημαίνει ότι αν μια εικόνα είναι οριζόντια (landscape) ή κάθετη (portrait), η μικρότερη πλευρά της (ύψος ή πλάτος) θα προσαρμοστεί στα 256 pixels, διατηρώντας την αναλογία διαστάσεων της αρχικής εικόνας.

Αφού οι εικόνες κλιμακωθούν έτσι ώστε η μικρότερη διάστασή τους να είναι 256 pixels, στη συνέχεια λαμβάνονται τυχαία αποκόμματα (crops) μεγέθους 224x224 pixels από αυτές τις κλιμακωμένες εικόνες. Αυτή η διαδικασία διασφαλίζει ότι όλες οι εικόνες εισόδου στο δίκτυο έχουν σταθερό μέγεθος, επιτρέποντας τη συνεπή επεξεργασία και ανάλυση από το μοντέλο. Η τυχαία αποκοπή βοηθά στο να δημιουργηθεί ποικιλία στις εικόνες εισόδου, καθώς το μοντέλο δεν θα βλέπει πάντα το ίδιο τμήμα της εικόνας σε κάθε επανάληψη της εκπαίδευσης. Κατά τη διάρκεια κάθε επανάληψης του mini-batch gradient descent επιλέγεται τυχαία ένα τμήμα από την ανακλιμακωμένη εικόνα για να χρησιμοποιηθεί ως είσοδος στο δίκτυο.

Συνοψίζοντας, η διαδικασία προεπεξεργασίας των εικόνων εκπαίδευσης περιλαμβάνει την κλιμάκωση, την τυχαία αποκοπή, την οριζόντια αναστροφή και τη μετατόπιση χρώματος, με σκοπό να δημιουργηθούν ποικιλόμορφες και σταθερού μεγέθους εικόνες εισόδου για την αποτελεσματική εκπαίδευση των μοντέλων.

Το σύνολο δεδομένων ILSVRC-2012, που χρησιμοποιήθηκε στους διαγωνισμούς ILSVRC από το 2012 έως το 2014, περιλαμβάνει εικόνες από 1000 διαφορετικές κατηγορίες και είναι χωρισμένο σε τρία μέρη: το σύνολο εκπαίδευσης, το οποίο περιέχει 1.3 εκατομμύρια εικόνες, το σύνολο επικύρωσης με 50.000 εικόνες και το σύνολο δοκιμής με 100.000 εικόνες των οποίων οι ετικέτες κατηγοριών είναι αποκρυμμένες.

Η απόδοση των μοντέλων αξιολογείται με δύο βασικές μετρήσεις: το σφάλμα top-1 και το σφάλμα top-5. Το σφάλμα top-1 αντιπροσωπεύει το ποσοστό των εικόνων που ταξινομήθηκαν λανθασμένα στην κορυφαία κατηγορία. Αυτό σημαίνει ότι για κάθε εικόνα, εάν η κατηγορία που προβλέφθηκε ως η πιο πιθανή δεν είναι η σωστή, τότε η πρόβλεψη θεωρείται λανθασμένη. Για παράδειγμα, αν μια εικόνα ενός σκύλου ταξινομηθεί ως γάτα, αυτό μετριέται ως σφάλμα top-1.

Το σφάλμα top-5 είναι το κύριο κριτήριο αξιολόγησης που χρησιμοποιείται στους διαγωνισμούς ILSVRC. Μετρά το ποσοστό των εικόνων για τις οποίες η σωστή κατηγορία δεν περιλαμβάνεται στις πέντε κορυφαίες προβλεπόμενες κατηγορίες. Αυτό σημαίνει ότι για κάθε εικόνα, αν η σωστή κατηγορία (π.χ., "σκύλος") δεν βρίσκεται μέσα στις πέντε πιο πιθανές κατηγορίες που προβλέφθηκαν από το μοντέλο,

τότε καταγράφεται ένα σφάλμα top-5. Η μέτρηση αυτή παρέχει μια πιο ευέλικτη αξιολόγηση της απόδοσης του μοντέλου, δεδομένου ότι η σωστή κατηγορία μπορεί να μην είναι η πρώτη επιλογή αλλά να περιλαμβάνεται στις πέντε πρώτες προβλέψεις.

Για τα περισσότερα πειράματα, οι ερευνητές χρησιμοποίησαν το σύνολο επικύρωσης για να αξιολογήσουν την απόδοση των μοντέλων τους, επιτρέποντάς τους να μετρήσουν με ακρίβεια την αποτελεσματικότητα των μοντέλων σε δεδομένα που δεν είχαν δει κατά την εκπαίδευση. Τα αποτελέσματα αυτών των πειραμάτων υποβλήθηκαν στον επίσημο διακομιστή του ILSVRC ως συμμετοχή της ομάδας "VGG" στον διαγωνισμό ILSVRC-2014. Αυτή η διαδικασία επέτρεψε την επίσημη αξιολόγηση των μοντέλων σε ένα πραγματικό διαγωνιστικό περιβάλλον.

Αξιολόγηση της Απόδοσης Μοντέλων Στον Διαγωνισμό ILSVRC-2014: Σφάλμα Top-1 και Top-5:

Στη μελέτη της απόδοσης των μοντέλων ConvNet, εξετάστηκαν διαφορετικές αρχιτεκτονικές με διάφορο βάθος (δηλαδή αριθμό επιπέδων. Ένα από τα κύρια ευρήματα ήταν ότι η χρήση τοπικής κανονικοποίησης απόκρισης (Local Response Normalisation - LRN) δεν βελτίωσε την απόδοση των μοντέλων. Συγκεκριμένα, το μοντέλο με 11 επίπεδα(8 συνελικτικά και 3 πλήρως συνδεδεμένα) είχε παρόμοιο σφάλμα ταξινόμησης top-1 με και χωρίς LRN, επιβεβαιώνοντας ότι η κανονικοποίηση δεν ήταν ωφέλιμη.

Όσον αφορά το βάθος του δικτύου, παρατηρήθηκε ότι όσο αυξανόταν ο αριθμός των επιπέδων, μειωνόταν το σφάλμα ταξινόμησης. Για παράδειγμα, το μοντέλο με 11 επίπεδα (8 συνελικτικά και 3 πλήρως συνδεδεμένα) είχε υψηλότερο σφάλμα ταξινόμησης από το μοντέλο με 19 επίπεδα (16 συνελικτικά και 3 πλήρως συνδεδεμένα). Συγκεκριμένα, το σφάλμα top-1 μειώθηκε από 28.5% σε 24.8%, δείχνοντας σαφή βελτίωση με την αύξηση του βάθους.

Επιπλέον, διαπιστώθηκε ότι η χρήση συνελικτικών φίλτρων 3 × 3 σε όλο το δίκτυο ήταν πιο αποτελεσματική από τη χρήση συνελικτικών φίλτρων 1 × 1 σε ορισμένα επίπεδα Για παράδειγμα, το μοντέλο με τρία επίπεδα 1 × 1 φίλτρων είχε σφάλμα top-1 26.4%, ενώ το μοντέλο με 3 × 3 φίλτρα σε όλα τα επίπεδα είχε σφάλμα top-1 24.8%. Αυτό υποδηλώνει ότι η επιπλέον μη γραμμικότητα βοηθά στη βελτίωση της απόδοσης, αλλά είναι επίσης κρίσιμο να καταγράφεται ο χωρικός συσχετισμός χρησιμοποιώντας φίλτρα με μεγαλύτερο πεδίο λήψης.

Επιπλέον, συγκρίνοντας ένα βαθύ δίκτυο με μικρά φίλτρα και ένα ρηχό δίκτυο με μεγαλύτερα φίλτρα, βρέθηκε ότι το βαθύ δίκτυο αποδίδει καλύτερα. Το ρηχό δίκτυο με πέντε επίπεδασυνελικτικών φίλτρων 5 × 5 είχε 7% υψηλότερο σφάλμα top-1 σε σύγκριση με το βαθύ δίκτυο που χρησιμοποιούσε επίπεδασυνελικτικών φίλτρων 3 × 3. Αυτό επιβεβαιώνει ότι τα βαθιά δίκτυα με μικρά φίλτρα είναι πιο αποτελεσματικά.

Τέλος, διαπιστώθηκε ότι η μεταβολή της κλίμακας κατά την εκπαίδευση (scale jittering) οδήγησε σε σημαντικά καλύτερα αποτελέσματα από την εκπαίδευση με σταθερό μέγεθος εικόνας. Συγκεκριμένα, η

εκπαίδευση με κλίμακες εικόνας από 256 έως 512 pixels οδήγησε σε σφάλμα top-1 24.8%, ενώ η εκπαίδευση με σταθερή κλίμακα εικόνας είχε υψηλότερα ποσοστά σφάλματος. Αυτό δείχνει ότι η εμπλουτισμός του συνόλου εκπαίδευσης με μεταβολή της κλίμακας βοηθά στη σύλληψη στατιστικών πολλαπλών κλιμάκων της εικόνας, βελτιώνοντας έτσι την απόδοση του μοντέλου.

Μετά την αξιολόγηση των μεμονωμένων αρχιτεκτονικών των μοντέλων VGG ακολούθησε η συνδυασμένη αξιολόγηση πολλών μοντέλων. Σε αυτή τη φάση, χρησιμοποιήθηκε μια προσέγγιση όπου τα αποτελέσματα πολλών μοντέλων συνδυάζονται, κάνοντας τον μέσο όρο των πιθανοτήτων κλάσεων που προκύπτουν από τη λειτουργία soft-max. Η λειτουργία soft-max δίνει τις πιθανότητες για κάθε κατηγορία, και ο συνδυασμός των αποτελεσμάτων μέσω του μέσου όρου επιτρέπει πιο σταθερές και αξιόπιστες προβλέψεις. Αυτή η τεχνική βελτιώνει την απόδοση του συστήματος λόγω της συμπληρωματικότητας των διαφορετικών μοντέλων, κάτι που έχει αποδειχθεί αποτελεσματικό στις κορυφαίες υποβολές του διαγωνισμού ILSVRC το 2012 και το 2013.

Τα αποτελέσματα αυτού του πειράματος δείχνουν σαφή βελτίωση. Μέχρι την υποβολή στο ILSVRC, είχαν εκπαιδευτεί δίκτυα που λειτουργούσαν σε μία κλίμακα εικόνας, καθώς και ένα πολυκλιμακικό μοντέλο με 16 συνελικτικές επίπεδα και 3 πλήρως συνδεδεμένες επίπεδα. Το σύνολο αυτών των 7 δικτύων πέτυχε σφάλμα δοκιμής 7.3% στο ILSVRC. Μετά την υποβολή, εξετάστηκε η απόδοση ενός συνόλου από τα δύο καλύτερα αποδίδοντα μοντέλα. Το πρώτο μοντέλο είχε 16 συνελικτικά επίπεδα και 3 πλήρως συνδεδεμένες επίπεδα ενώ το δεύτερο μοντέλο είχε 19 συνελικτικά επίπεδα και 3 πλήρως συνδεδεμένες επίπεδα. Αυτός ο συνδυασμός μοντέλων μείωσε το σφάλμα 6.8%.

Για αναφορά, το καλύτερο από τα μεμονωμένα μοντέλα, το οποίο είχε 19 συνελικτικές επίπεδα και 3 πλήρως συνδεδεμένες επίπεδα πέτυχε σφάλμα 7.1%. Αυτό δείχνει ότι η συνδυασμένη προσέγγιση των αποτελεσμάτων πολλών μοντέλων προσφέρει σημαντική βελτίωση στην απόδοση του συστήματος ταξινόμησης εικόνων.

Η ομάδα "VGG" κατάφερε να μειώσει το ποσοστό σφάλματος τους στο 6,8%, χρησιμοποιώντας μόνο δύο μοντέλα, μετά την αρχική υποβολή με 7 μοντέλα και ποσοστό σφάλματος 7,3%. Αυτό το ποσοστό είναι πολύ κοντά στον νικητή της πρόκλησης, το GoogLeNet, που είχε σφάλμα 6,7%. Η ομάδα "VGG" σημείωσε επίσης καλύτερα αποτελέσματα από τον νικητή του ILSVRC-2013, Clarifai, που είχε υψηλότερα ποσοστά σφάλματος. Επιπλέον, το μοντέλο ενός μόνο δικτύου της "VGG" παρουσίασε καλύτερη απόδοση από το μοντέλο ενός μόνο δικτύου του GoogLeNet κατά 0,9%. Είναι αξιοσημείωτο ότι η "VGG" βελτίωσε την κλασική αρχιτεκτονική ConvNet, αυξάνοντας σημαντικά το βάθος των δικτύων.

Τα μοντέλα CNNs αυτής της αρχιτεκτονικής VGG αποδείχθηκαν εξαιρετικά αποτελεσματικά με βάση τα παραπάνω αποτελέσματα. Αποτελούν λοιπόν μια ισχυρή επιλογή για μεγάλες κλίμακες ταξινόμησης εικόνων. [39]

Η Ικανότητα και Αποδοτικότητα των Συνελικτικών Νευρωνικών Δικτύων (CNNs) στην Ανάλυση και Επεξεργασία Ηχητικών Δεδομένων

Το paper "CNN ARCHITECTURES FOR LARGE-SCALE AUDIO CLASSIFICATION" εξετάζει την εφαρμογή διαφόρων CNN για την ταξινόμηση ήχου χρησιμοποιώντας το πολύ μεγάλο σύνολο δεδομένων YouTube-100M. Οι αρχιτεκτονικές CNN όπως τα AlexNet, VGG, Inception και ResNet, οι οποίες είχαν τεράστια επιτυχία στον τομέα της επεξεργασίας εικόνας και βίντεο, αξιολογήθηκαν για την ταξινόμηση των ηχητικών στοιχείων ενός βίντεο. Η ταξινόμηση αυτή αναφέρεται στη διαδικασία κατά την οποία τα ηχητικά στοιχεία που συνοδεύουν τα βίντεο ταξινομούνται αυτόματα σε προκαθορισμένες κατηγορίες ή ετικέτες. Αυτές οι ετικέτες μπορεί να περιλαμβάνουν ήχους όπως "ομιλία", "μουσική", "ήχοι περιβάλλοντος" κ.λπ.

Η απόδοση των αρχιτεκτονικών CNN εξετάστηκε σε σχέση με το μέγεθος του εκπαιδευτικού συνόλου και το λεξιλόγιο ετικετών. Το μέγεθος του εκπαιδευτικού συνόλου αναφέρεται στον αριθμό των βίντεο που χρησιμοποιούνται για την εκπαίδευση των μοντέλων. Ένα μεγαλύτερο εκπαιδευτικό σύνολο προσφέρει περισσότερα δεδομένα για μάθηση και μπορεί να βελτιώσει την ακρίβεια του μοντέλου. Το λεξιλόγιο ετικετών αναφέρεται στον αριθμό και την ποικιλία των κατηγοριών ή ετικετών που χρησιμοποιούνται για την ταξινόμηση των ηχητικών δεδομένων. Ένα μεγαλύτερο λεξιλόγιο ετικετών σημαίνει ότι το μοντέλο πρέπει να μάθει να διακρίνει μεταξύ περισσότερων κατηγοριών, κάτι που μπορεί να είναι πιο απαιτητικό αλλά και πιο ευέλικτο.

Το σύνολο δεδομένων YouTube-100M περιλαμβάνει 100 εκατομμύρια βίντεο από το YouTube, εκ των οποίων 70 εκατομμύρια χρησιμοποιούνται για εκπαίδευση, 10 εκατομμύρια για αξιολόγηση και 20 εκατομμύρια για επικύρωση. Τα βίντεο έχουν μέση διάρκεια 4.6 λεπτών και συνολικά 5.4 εκατομμύρια ώρες εκπαίδευσης. Κάθε βίντεο έχει επισημανθεί με ετικέτες από ένα σύνολο 30,871 θεμάτων. Αυτές οι ετικέτες αποδίδονται αυτόματα, δηλαδή, ένα σύστημα λογισμικού αναλύει τις πληροφορίες όπως τον τίτλο, την περιγραφή, τα σχόλια και το περιεχόμενο του βίντεο και στη συνέχεια αποδίδει τις κατάλληλες ετικέτες χωρίς ανθρώπινη παρέμβαση.

Κατά την εκπαίδευση, ο ήχος χωρίστηκε σε καρέ των 960 ms, δηλαδή σε τμήματα διάρκειας 960 χιλιοστών του δευτερολέπτου. Κάθε καρέ επεξεργάστηκε χρησιμοποιώντας τη μετασχηματισμό Fourier μικρής διάρκειας (STFT), μια τεχνική που επιτρέπει τη μετατροπή του ηχητικού σήματος από τον χρόνο στον συχνότητα. Με αυτόν τον τρόπο, μπορούμε να δούμε ποιες συχνότητες υπάρχουν στο ηχητικό σήμα και πώς αυτές οι συχνότητες αλλάζουν με την πάροδο του χρόνου.

Χρησιμοποιήθηκε ο βελτιστοποιητής Adam και μετά από κάθε συνελικτικό επίπεδο, εφαρμόστηκε κανονικοποίηση παρτίδας (batch normalization). Αυτή η τεχνική βοηθάει στη σταθεροποίηση και επιτάχυνση της εκπαίδευσης, κανονικοποιώντας την έξοδο κάθε επιπέδου, έτσι ώστε να έχει μέση τιμή 0 και διασπορά 1.

Το τελικό επίπεδο του NN χρησιμοποιεί τη συνάρτηση ενεργοποίησης sigmoid. Σε αντίθεση με τη συνάρτηση ενεργοποίησης softmax, που χρησιμοποιείται για την ταξινόμηση ενός μόνο αντικειμένου σε

μία κατηγορία, η συνάρτηση ενεργοποίησης sigmoid επιτρέπει την ταξινόμηση ενός αντικειμένου σε πολλές κατηγορίες ταυτόχρονα. Αυτό είναι σημαντικό όταν ένα ηχητικό κλιπ μπορεί να περιέχει περισσότερα από ένα ηχητικά γεγονότα. Η συνάρτηση κόστουςπου χρησιμοποιήθηκε είναι η cross-entropy, η οποία μετρά τη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών ετικετών.

Η αξιολόγηση της απόδοσης των μοντέλων έγινε χρησιμοποιώντας τρία ισορροπημένα σύνολα δεδομένων που επιλέχθηκαν από το σύνολο των 10 εκατομμυρίων βίντεο αξιολόγησης. Ισορροπημένα σύνολα δεδομένων σημαίνει ότι κάθε σύνολο περιέχει ένα ίσο αριθμό παραδειγμάτων από κάθε κατηγορία που πρόκειται να αναγνωριστεί. Το πρώτο σύνολο περιλάμβανε 1 εκατομμύριο βίντεο και καλύπτει 30.000 διαφορετικές ετικέτες, παρέχοντας ένα ευρύ φάσμα κατηγοριών. Το δεύτερο σύνολο περιλάμβανε 100.000 βίντεο και επικεντρώνεται στις 3.087 πιο συχνές ετικέτες, δηλαδή σε ετικέτες που εμφανίζονται συχνότερα στα δεδομένα. Το τρίτο σύνολο δεδομένων περιλάμβανε 12.000 βίντεο και επικεντρώνεται στις 400 πιο συχνές ετικέτες, παρέχοντας μια πιο εστιασμένη ανάλυση στις πιο κοινές κατηγορίες.

Κατά την αξιολόγηση, τα φασματογραφήματα των 960 ms που δημιουργήθηκαν από τα ηχητικά κλιπ επεξεργάστηκαν από τον ταξινομητή, και τα αποτελέσματα αυτής της επεξεργασίας συνδυάστηκαν για κάθε βίντεο. Δηλαδή, οι προβλέψεις για κάθε καρέ των 960 ms συγχωνεύτηκαν για να δώσουν μια συνολική πρόβλεψη για το σύνολο του βίντεο. Η απόδοση των μοντέλων μετρήθηκε χρησιμοποιώντας δύο κύριους δείκτες: την περιοχή κάτω από την καμπύλη ROC -AUC και τον μέσο όρο ακρίβειας (mAP). Με αυτές τις μετρικές οι ερευνητές μπόρεσαν να αξιολογήσουν και να συγκρίνουν την απόδοση των διαφορετικών μοντέλων CNN στην ταξινόμηση των ηχητικών δεδομένων από τα βίντεο του YouTube-100M.

Η σύγκριση των αρχιτεκτονικών δικτύων έγινε εκπαιδεύοντας όλα τα δίκτυα με ένα σύνολο 3.000 ετικετών και χρησιμοποιώντας 70 εκατομμύρια βίντεο από το σύνολο δεδομένων. Αυτό σημαίνει ότι κάθε δίκτυο έμαθε να αναγνωρίζει και να ταξινομεί ήχους που ανήκουν σε 3.000 διαφορετικές κατηγορίες, χρησιμοποιώντας ένα τεράστιο αριθμό δειγμάτων βίντεο για την εκπαίδευσή του. Η χρήση ενός τόσο μεγάλου συνόλου δεδομένων βοηθάει τα δίκτυα να μάθουν με μεγαλύτερη ακρίβεια και να γενικεύσουν καλύτερα σε νέα δεδομένα.

Το μοντέλο ResNet-50 πέτυχε περιοχή κάτω από την καμπύλη ROC-AUC 0.959 και μέσο όρο ακρίβειας (mAP) 0.327. Το μοντέλο Inception V3 πέτυχε ROC- AUC 0.950 και mAP 0.320. Το μοντέλο VGG πέτυχε ROC-AUC 0.945 και mAP 0.315.

Η μελέτη επίσης εξέτασε πώς το μέγεθος του συνόλου ετικετών επηρεάζει την απόδοση του μοντέλου. Τα αποτελέσματα έδειξαν ότι η απόδοση των μοντέλων βελτιώθηκε ελαφρώς καθώς αυξήθηκε ο αριθμός των ετικετών εκπαίδευσης. Δηλαδή, τα μοντέλα που εκπαιδεύτηκαν με περισσότερες κατηγορίες ετικετών είχαν ελαφρώς καλύτερη απόδοση στην ταξινόμηση των ηχητικών δεδομένων.

Στα τελευταία πειράματα, οι ερευνητές εξέτασαν αν τα εκπαιδευμένα τους δίκτυα μπορούν να χρησιμοποιηθούν για την εξαγωγή χρήσιμων χαρακτηριστικών από διαφορετικά σύνολα δεδομένων.

Χρησιμοποιώντας το σύνολο δεδομένων Audio Set, που περιλαμβάνει 2 εκατομμύρια κλιπ ήχου με διάφορες ετικέτες ήχου, εκπαιδεύτηκε ένα πλήρως συνδεδεμένο δίκτυο με τα χαρακτηριστικά που εξήχθησαν από τα μοντέλα ResNet-50, Inception και VGG. Η εκπαίδευση έγινε σε 70% του συνόλου εκπαίδευσης και τα αποτελέσματα έδειξαν καλή απόδοση, με το μοντέλο να επιτυγχάνει ROC- AUC 0.959 και mAP 0.327, δείχνοντας την αποτελεσματικότητα των εξαγόμενων χαρακτηριστικών για ανάλυση και ταξινόμηση ήχου. Συνοψίζοντας, η παρούσα εργασία εξέτασε την απόδοση διαφόρων αρχιτεκτονικών CNNs στην ταξινόμηση ηχητικών δεδομένων, χρησιμοποιώντας το πολύ μεγάλο σύνολο δεδομένων YouTube-100M. Τα μοντέλα που αναλύθηκαν, όπως τα ResNet-50, Inception και VGG, έχουν γνωρίσει μεγάλη επιτυχία στον τομέα της επεξεργασίας εικόνας και βίντεο, και αποδείχθηκαν εξίσου αποτελεσματικά και στην ανάλυση ήχου. Τα αποτελέσματα υποδεικνύουν ότι τα CNNs είναι πολύ αποτελεσματικά για την ταξινόμηση και ανάλυση ηχητικών δεδομένων, προσφέροντας σημαντικές δυνατότητες για μελλοντική έρευνα και εφαρμογές στον τομέα αυτό. [40]

4. Διαδικασία και Μεθοδολογία Έρευνας

4.1 Σκοπός του Πειράματος

Σκοπός του πειράματος είναι η περαιτέρω μελέτη για τα φωνήεντα μπορούν έχουν μεγαλύτερη προβλεπτική ικανότητα έναντι των ολόκληρων ηχογραφήσεων. Περαιτέρω, θα γίνει εκτίμηση του κατά πόσο μπορεί να βελτιωθεί η επίδοση προσαυξάνοντας τα χαρακτηριστικά VGG με γενικά χαρακτηριστικά ήχου, όπως έχουν παρουσιαστεί στο κεφάλαιο 2.1.4.

4.2 Σύνολο δεδομένων Smart4Covid

Το smarty4covid επιδιώκει τη δημιουργία ενός προηγμένου συστήματος για την εκτίμηση και παρακολούθηση της υγειονομικής κατάστασης σε σχέση με τον COVID-19, χρησιμοποιώντας καινοτόμες τεχνικές βαθιάς μάθησης που εξασφαλίζουν σαφήνεια και ευκολία κατανόησης. Αυτό συνεπάγεται ότι τα αποτελέσματα των αναλύσεων είναι εύκολα κατανοητά και επεξηγήσιμα τόσο από τους χρήστες όσο και από τους ειδικούς. Για την υλοποίηση αυτού του έργου, λήφθηκαν οι απαραίτητες εγκρίσεις από την Επιτροπή Ηθικής της Έρευνας του Εθνικού Μετσόβιου Πολυτεχνείου. Μια διαδικτυακή εφαρμογή (www.smarty4covid.org) δημιουργήθηκε για να επιτρέπει τη συλλογή δεδομένων από τους χρήστες.

Το σύνολο δεδομένων smarty4covid περιλαμβάνει ηχογραφήσεις βήχα, αναπνοής και φωνής από 4.673 χρήστες, οι οποίοι είναι πολίτες της Ελλάδας και της Κύπρου. Για τις ηχητικές καταγραφές φωνής, ζητήθηκε από κάθε συμμετέχοντα να επαναλάβει τρεις φορές τη φράση: "Η φωνή μου έχει δύναμη στην πανδημία." Επιπλέον, συλλέχθηκαν πληροφορίες όπως δημογραφικά στοιχεία, συμπτώματα, υποκείμενες παθήσεις, κατάσταση καπνίσματος, ζωτικά σημεία, κατάσταση εμβολιασμού, νοσηλεία, συναισθηματική κατάσταση, συνθήκες εργασίας και αποτελέσματα διαγνωστικών τεστ για τον COVID-19.

Αφού συλλέχθηκαν, τα δεδομένα υποβλήθηκαν σε επεξεργασία για την αφαίρεση σφαλμάτων και θορύβου, διασφαλίζοντας έτσι την ακρίβεια και την ποιότητά τους. Ένα μέρος του συνόλου δεδομένων επισημάνθηκε από ιατρικούς ειδικούς, προκειμένου να χρησιμοποιηθεί ως πρότυπο αναφοράς για την εκπαίδευση και την αξιολόγηση των μοντέλων.

Η συλλεγείσα πληροφορία οργανώθηκε σε μια βάση γνώσεων οντολογίας (web ontology knowledge - OWL). Αυτό επιτρέπει την κατηγοριοποίηση και την ιεράρχηση των δεδομένων, κάνοντάς τα πιο εύκολα στην αναζήτηση και ανάλυση. Οι ιατρικές έννοιες στη βάση γνώσεων συνδέονται με αναγνωριστικά από το σύστημα SNOMED-CT, ένα διεθνώς αναγνωρισμένο σύστημα κωδικοποίησης ιατρικών όρων, που εξασφαλίζει την τυποποίηση και παγκόσμια αναγνώριση των εννοιών. Η βάση γνώσεων οντολογίας όχι μόνο οργανώνει τις πληροφορίες, αλλά και διευκολύνει την αναζήτηση, ανάλυση και ερμηνεία τους από

ερευνητές, ιατρούς και άλλους επαγγελματίες υγείας. Αυτό συμβάλλει στην αποτελεσματικότερη χρήση των δεδομένων για την υποστήριξη της έρευνας και της λήψης αποφάσεων στην υγεία.

Επιπρόσθετα, το σύνολο δεδομένων smarty4covid έχει δημοσιοποιηθεί, ωστόσο, εξαιρέθηκαν οι ηχητικές καταγραφές που θεωρούνται προσωπικά δεδομένα σύμφωνα με τον κανονισμό GDPR, για την προστασία της ιδιωτικότητας των χρηστών. Η βάση γνώσεων OWL του smarty4covid επιτρέπει την ενοποίηση δεδομένων από διάφορες πηγές σε ένα ενιαίο σύστημα, διευκολύνοντας την ανάλυση και την κατανόηση των δεδομένων.

Το σύνολο δεδομένων smarty4covid έχει αξιοποιηθεί για την ανάπτυξη μοντέλων που μπορούν να ταξινομούν τμήματα ηχητικών σημάτων ως "βήχας", "αναπνοή", "φωνή" και "άλλα". Επιπλέον, τα μοντέλα μπορούν να ανιχνεύουν τμήματα εισπνοής και εκπνοής από ηχογραφήσεις αναπνοής, τα οποία χρησιμοποιούνται για την εξαγωγή κλινικά σχετιζόμενων χαρακτηριστικών, όπως οι ρυθμοί αναπνοής (RR), η αναλογία εισπνοής προς εκπνοή (I/E ratio) και ο κλασματικός χρόνος εισπνοής (FIT).

Μεθοδολογια Δημιουργιας Smart4Covid

Ανάπτυξη και Συλλογή Δεδομένων μέσω της Εφαρμογής smarty4covid

Η συλλογή δεδομένων για το smarty4covid πραγματοποιήθηκε μέσω μιας διαδικασίας που ονομάζεται crowd-sourcing, δηλαδή με τη συμμετοχή πολλών χρηστών που συνεισφέρουν δεδομένα. Αυτή η μέθοδος εγκρίθηκε από την Επιτροπή Ηθικής Έρευνας του Εθνικού Μετσόβιου Πολυτεχνείου και ακολουθήθηκαν όλοι οι σχετικοί ηθικοί κανονισμοί για την προστασία των συμμετεχόντων.

Για τη συλλογή των δεδομένων, δημιουργήθηκε μια ευέλικτη και φιλική προς τον χρήστη διαδικτυακή εφαρμογή (www.smarty4covid.org), η οποία απευθυνόταν σε Έλληνες και Κύπριους πολίτες άνω των 18 ετών. Οι χρήστες καλούνταν να συμπληρώσουν ένα ερωτηματολόγιο που περιλάμβανε διάφορες ενότητες με πληροφορίες όπως τα δημογραφικά τους στοιχεία, η κατάσταση εμβολιασμού κατά του COVID-19, το ιατρικό τους ιστορικό, ζωτικά σημεία που μετρήθηκαν με παλμικό οξύμετρο και πιεσόμετρο, συμπτώματα COVID-19, συνήθειες καπνίσματος, νοσηλεία, συναισθηματική κατάσταση και συνθήκες εργασίας. Επιπλέον, οι χρήστες έπρεπε να εκτελέσουν ηχογραφήσεις τεσσάρων τύπων: τρεις ηχογραφήσεις φωνής όπου διάβαζαν μια συγκεκριμένη πρόταση, πέντε βαθιές αναπνοές, 30 δευτερόλεπτα κανονικής αναπνοής κοντά στο μικρόφωνο της συσκευής και τρεις εκούσιοι βήχες.

Με την προώθηση της εφαρμογής μέσω μιας αποτελεσματικής στρατηγικής μέσων, περισσότερα από 10.000 άτομα παρείχαν τις δημογραφικές και ιατρικές πληροφορίες τους στην εφαρμογή smarty4covid. Ωστόσο, σχεδόν οι μισοί από αυτούς (περίπου 4.679) επέλεξαν να δώσουν την άδεια τους για να πραγματοποιήσουν τις ηχογραφήσεις.

Η διαδικτυακή εφαρμογή κυκλοφόρησε τον Ιανουάριο του 2022, κατά τη διάρκεια της εξάπλωσης της παραλλαγής όμικρον του COVID-19 στην Ελλάδα, γεγονός που οδήγησε σε υψηλή επικράτηση της νόσου. Συγκεκριμένα, το 17,3% των χρηστών της εφαρμογής διαγνώστηκαν θετικοί στον COVID-19.

Διαδικασία Καθαρισμού και Εμπλουτισμού του Συνόλου Δεδομένων smarty4covid με Συμμετοχή Επαγγελματιών Υγείας

Μετά τη συλλογή των δεδομένων από την εφαρμογή smarty4covid, ανακαλύφθηκε ότι ένα μέρος των δεδομένων δεν ήταν έγκυρο λόγω προβληματικών ηχογραφήσεων που υποβλήθηκαν από τους χρήστες, οι οποίες περιλάμβαναν παραμορφώσεις και υψηλό θόρυβο στο υπόβαθρο. Αυτά τα προβλήματα καθιστούσαν τις ηχογραφήσεις δύσκολες στη χρήση για ανάλυση και εξαγωγή χρήσιμων πληροφοριών.

Για να αντιμετωπιστεί αυτό το πρόβλημα, οργανώθηκε μια εκστρατεία καθαρισμού δεδομένων μέσω crowd-sourcing. Σε αυτήν την περίπτωση, χρησιμοποιήθηκε το ανοικτό λογισμικό Label Studio, το οποίο είναι ένα εργαλείο που επιτρέπει την επισήμανση και κατηγοριοποίηση των δεδομένων. Μηχανικοί τεχνητής νοημοσύνης, που προσφέρθηκαν εθελοντικά να βοηθήσουν σε αυτήν τη διαδικασία, ανέλαβαν να ακούσουν τις ηχογραφήσεις και να τις επισημάνουν. Δηλαδή, ανέλαβαν να αποφασίσουν ποιες ηχογραφήσεις ήταν έγκυρες και ποιες όχι, με βάση την ποιότητα του ήχου και την παρουσία θορύβου ή παραμορφώσεων.

Πριν ξεκινήσουν τη δουλειά τους, οι εθελοντές υπέγραψαν Συμφωνία Εμπιστευτικότητας (NDA) για να διασφαλιστεί ότι τα δεδομένα θα παραμείνουν ασφαλή και ιδιωτικά. Επίσης, τους παρασχέθηκαν οι απαραίτητες άδειες πρόσβασης, ώστε να μπορούν να εργαστούν με τα δεδομένα και να ολοκληρώσουν την επισήμανση.

Ένα φιλικό προς τον χρήστη περιβάλλον δημιουργήθηκε για τους σχολιαστές, το οποίο τους επέτρεπε να ακούν τα ηχητικά σήματα και να απαντούν σε ερωτήσεις σχετικά με την εγκυρότητά τους (ναι/όχι) και την ποιότητά τους (Καλή, Αποδεκτή, Κακή) με βάση τον θόρυβο υποβάθρου και τις παραμορφώσεις. Για να διασφαλιστεί η ποιότητα των σχολιασμών, ένα τυχαίο δείγμα από 1.389 ηχογραφήσεις εξετάστηκε πολλές φορές (έως 5) στη διαδικασία επισήμανσης, και παρατηρήθηκε υψηλή συνέπεια (92,5%) μεταξύ των σχολιαστών, δείχνοντας ότι δεν ήταν απαραίτητο να υπάρχουν πολλοί σχολιαστές για κάθε ηχογράφηση.

Το σύνολο δεδομένων smarty4covid δεν σταμάτησε μόνο στον καθαρισμό και την επισήμανση από μηχανικούς τεχνητής νοημοσύνης. Για να αυξηθεί η ακρίβεια και η χρησιμότητα των δεδομένων, εμπλουτίστηκε με ετικέτες που επισημάνθηκαν από επαγγελματίες υγείας. Αυτοί οι επαγγελματίες περιλάμβαναν πνευμονολόγους, αναισθησιολόγους και παθολόγους, οι οποίοι προσφέρθηκαν εθελοντικά να αξιολογήσουν τις ηχογραφήσεις.

Οι ειδικοί αυτοί είχαν την ευθύνη να ακούσουν τις ηχογραφήσεις και να εντοπίσουν τυχόν ακουστικές ανωμαλίες. Αυτές οι ανωμαλίες μπορεί να περιλαμβάνουν παράγοντες όπως ασυνήθιστοι ήχοι στην αναπνοή ή στο βήχα που θα μπορούσαν να υποδεικνύουν προβλήματα υγείας. Αφού αξιολόγησαν τις ηχογραφήσεις, παρείχαν εξατομικευμένες συστάσεις σχετικά με το αν οι χρήστες χρειάζονταν να ζητήσουν ιατρική συμβουλή.

Για αυτήν τη διαδικασία, πραγματοποιήθηκαν τέσσερις εθελοντικές αξιολογήσεις από ειδικούς. Τρεις από αυτές επικεντρώθηκαν στις ηχογραφήσεις αναπνοής, φωνής και βήχα αντίστοιχα. Οι επαγγελματίες

υγείας, κατά τη διάρκεια αυτών των αξιολογήσεων, αξιολογούσαν την παρουσία ακουστικών ανωμαλιών επιλέγοντας μία ή περισσότερες επιλογές από τις διαθέσιμες ετικέτες που τους δίνονταν. Για παράδειγμα, αν μια ηχογράφηση περιείχε ήχους που θα μπορούσαν να υποδηλώνουν ασθένεια, αυτό σημειωνόταν ως ακουστική ανωμαλία.

Στην τέταρτη αξιολόγηση, οι επαγγελματίες υγείας είχαν πρόσβαση σε όλες τις διαθέσιμες πληροφορίες για τον χρήστη, εκτός από τα ζωτικά σημεία (όπως κορεσμός οξυγόνου, παλμοί ανά λεπτό, διαστολική/συστολική πίεση). Ο λόγος που τα ζωτικά σημεία εξαιρέθηκαν ήταν για να αποφευχθεί οποιαδήποτε προκατάληψη στην αξιολόγηση της υγείας του χρήστη. Με αυτές τις πληροφορίες, οι επαγγελματίες υγείας μπορούσαν να εκτιμήσουν τον κίνδυνο επιδείνωσης της υγείας του χρήστη και να προτείνουν τα επόμενα βήματα που πρέπει να ακολουθηθούν. Οι προτεινόμενες ενέργειες περιλάμβαναν την αναζήτηση ιατρικής συμβουλής αν η κατάσταση του χρήστη θεωρούνταν ανησυχητική, την επανάληψη του τεστ Smarty4Covid σε 24 ώρες για παρακολούθηση της κατάστασης, και την επανάληψη του τεστ σε περίπτωση που παρατηρηθούν αλλαγές στην υγεία του χρήστη.

Επιπλέον, ζητήθηκε από τους επαγγελματίες υγείας να καθορίσουν τον βαθμό βεβαιότητας για την αξιολόγησή τους, σε μια κλίμακα από 1 έως 10. Αυτό σημαίνει ότι, μετά την ολοκλήρωση της αξιολόγησής τους, οι επαγγελματίες έπρεπε να δηλώσουν πόσο σίγουροι ήταν για την ακρίβεια και την ορθότητα των συστάσεών τους. Αυτή η πληροφορία ήταν σημαντική για πολλούς λόγους.

Πρώτον, βοήθησε να εκτιμηθεί η αξιοπιστία των συστάσεων που έδιναν οι επαγγελματίες υγείας. Εάν ένας επαγγελματίας υγείας δήλωνε υψηλό βαθμό βεβαιότητας (κοντά στο 10), αυτό σήμαινε ότι ήταν πολύ σίγουρος για την αξιολόγησή του και τις συστάσεις του. Αντίθετα, ένας χαμηλός βαθμός βεβαιότητας (κοντά στο 1) σήμαινε ότι υπήρχε αβεβαιότητα και ενδεχομένως η ανάγκη για περαιτέρω διερεύνηση ή επιβεβαίωση.

Δεύτερον, η καταγραφή του βαθμού βεβαιότητας πρόσθεσε μια επιπλέον διάσταση αξιοπιστίας στην όλη διαδικασία αξιολόγησης. Έτσι, όσοι χρησιμοποιούσαν τα αποτελέσματα των αξιολογήσεων μπορούσαν να λάβουν υπόψη την βεβαιότητα των επαγγελματιών υγείας, προκειμένου να αποφασίσουν πώς να προχωρήσουν με βάση τις συστάσεις.

Εξαγωγή Αναπνευστικών Χαρακτηριστικών από Καταγραφές Ηχητικών Σημάτων

Αφού ολοκληρώθηκαν τα προηγούμενα βήματα της συλλογής και επεξεργασίας των δεδομένων, καθώς και της επισήμανσης από ειδικούς, οι ερευνητές προχώρησαν στην εξαγωγή χαρακτηριστικών από τις καταγραφές της αναπνοής.

Οι δείκτες που χρησιμοποιούνται για την περιγραφή ενός αναπνευστικού προτύπου περιλαμβάνουν τον αριθμό αναπνοών ανά λεπτό (RR), τον λόγο εισπνοής προς εκπνοή (I/E ratio) και τον κύκλο λειτουργίας της εισπνοής (FIT). Ο RR είναι ο αριθμός αναπνοών ανά λεπτό, που κανονικά είναι 16-20 αναπνοές/λεπτό. Μπορεί να επηρεαστεί από διάφορους παράγοντες, όπως η θερμοκρασία, η ισορροπία οξέων-βάσεων, η μεταβολική κατάσταση και οι ασθένειες. Ο I/E ratio είναι ο λόγος μεταξύ του χρόνου εισπνοής (Ti) και του χρόνου εκπνοής (Te) και μπορεί να υποδηλώνει διαταραχή ροής στην αναπνευστική

οδό. Η κανονική αναπνοή συνήθως παρουσιάζει λόγο I/E 1:2 ή 1:3 σε κατάσταση ηρεμίας. Το FIT, επίσης γνωστό ως "κύκλος λειτουργίας" της εισπνοής, είναι ο λόγος μεταξύ του Ti και της διάρκειας ενός συνολικού αναπνευστικού κύκλου (Ttot), παρέχοντας μια πρόχειρη μέτρηση της απόφραξης των αεραγωγών και του άγχους στους αναπνευστικούς μυς.

Για την εξαγωγή αυτών των δεικτών από τα ηχητικά σήματα αναπνοής, αναπτύχθηκε μια προσέγγιση δύο βημάτων. Πρώτον, εντοπίστηκαν τα τμήματα του ηχητικού σήματος που περιέχουν αναπνοή χρησιμοποιώντας ένα μοντέλο βασισμένο στην AI. Στη συνέχεια, αυτά τα τμήματα διαχωρίστηκαν σε μη σιωπηλά διαστήματα. Το δεύτερο βήμα περιλάμβανε την ανίχνευση των τμημάτων εισπνοής και εκπνοής. Για να αντιμετωπιστούν οι προκλήσεις της παραμόρφωσης και του θορύβου υποβάθρου, αναπτύχθηκε μια μη εποπτευόμενη μέθοδος που χρησιμοποιεί έναν αλγόριθμο ομαδοποίησης σε επίπεδο συχνοτήτων. Για τον σκοπό αυτό, το mel-φασματογράφημα (MFCC-128) του ηχητικού σήματος εξήχθη και μετασχηματίστηκε σε διάνυσμα 128 συχνοτήτων, καθεμία από τις οποίες αντιστοιχεί στο άθροισμα των αντίστοιχων συχνοτήτων με την πάροδο του χρόνου. Αυτή η μέθοδος δεν απαιτεί ανθρώπινα ετικετοποιημένα δεδομένα για εκπαίδευση και προσθέτει ανθεκτικότητα έναντι της παραμόρφωσης και του θορύβου υποβάθρου.

Η προτεινόμενη μέθοδος πέτυχε χαμηλές τιμές Τετραγωνικής Μέσης Ρίζας Σφάλματος (RMSE) για τους δείκτες RR, FIT και I/E ratio, υποδεικνύοντας την ακρίβεια και αξιοπιστία της μεθόδου.

Δημόσια Διάθεση και Οργάνωση Δεδομένων

Αφού ολοκληρώθηκαν τα βήματα της συλλογής, επεξεργασίας και εξαγωγής χαρακτηριστικών από τα δεδομένα αναπνοής, το επόμενο βήμα ήταν η οργάνωση και δημόσια διάθεση των δεδομένων για ευρύτερη χρήση από την επιστημονική κοινότητα.

Μέρος του συνόλου δεδομένων smarty4covid που συλλέχθηκε (4.303 υποβολές) οργανώθηκε σε αρχεία δεδομένων προκειμένου να είναι δημόσια διαθέσιμα. Τα αρχεία δεδομένων κατατέθηκαν στο Zenodo Repository (DOI: 10.5281/zenodo.7760170). Κάθε κατάλογος περιέχει τις υποβολές ενός συγκεκριμένου χρήστη. Ο κατάλογος του χρήστη ονομάζεται σύμφωνα με το αναγνωριστικό του, που δημιουργείται σύμφωνα με το πρωτόκολλο UUID V4.

Εκτός από τις υποβολές, περιλαμβάνεται επίσης ένα αρχείο json (“demographics_underlying_conditions.json”) με πληροφορίες σχετικά με τα δημογραφικά στοιχεία (π.χ. BMI, ηλικιακή ομάδα, φύλο) και πιθανές υποκείμενες παθήσεις. Κάθε υποβολή αντιστοιχεί σε έναν ξεχωριστό υποκατάλογο που ονομάζεται σύμφωνα με το μοναδικό αναγνωριστικό της υποβολής και περιέχει:

έγκυρες ηχογραφήσεις βήχα (“audio.cough.mp3”), βαθιάς αναπνοής (“audio.breath_deep.mp3”) και κανονικής αναπνοής (“audio.breath_regular.mp3”). Κάθε ηχογράφηση έχει ρυθμό δειγματοληψίας 48 kHz και ρυθμό μετάδοσης bit 64 kb/s.

ένα αρχείο json (“main_questionnaire.json”) με πληροφορίες σχετικές με το τεστ COVID-19 (αποτέλεσμα, τύπος και ημερομηνία), την κατάσταση εμβολιασμού κατά του COVID-19, τα σχετικά συμπτώματα, τα ζωτικά σημεία και άλλα.

Main questionnaire json file description (Part 1/3: COVID-19 related information)

Field name	Description	Type	Values
participantid	Participant’s identification number.	String	UUID
submissionid	Questionnaire’s Identificiation number.	String	UUID
Covid_status	Tested for COVID-19.	String	“positive”: Positive, “Negative”: Negative, “no”: Not tested
pcr_test	Tested with PCR.	Bool	
rapid_test	Tested with a Rapid Antigen test.	Bool	
self_test	Tested with a Rapid Antigen Self test	Bool	
test_last_3_days	Tested in the last 3 days.	Bool	
last_negative_test_Date	Date of the last negative test.	String	“yyyy-mm-dd"
first_positive_test_Date	Date of the first negative test.	String	“yyyy-mm-dd"
vaccination_status	COVID-19 vaccination status.	String	"no": No, "partially": One of two shots, "fully": Fully, "booster1": Fully and Booster dose, "booster2": Fully and two Booster doses
latest_vaccination_date	Date of the last vaccination dose.	String	"yyyy-mm-dd"
hospitalization	Whether the user was hospitalised for COVID-19.	String	"0": "No", "1": "I am currently hospitalized", "2":

			"Yes, discharged a week ago", "3": "Yes, discharged more than a month ago"
exposure_to_someone_with_covid	Whether the user was exposed to a confirmed COVID-19 case.	String	"No" / "Maybe" / "Yes"
travelled_abroad	Whether the user has travelled abroad the last 14 days.	String	"0": No, "1": Yes
submission_timestamp	Timestamp when the submission was received	String	

Πίνακας 3.1: Αρχείο json (“Main questionnaire.json”).

Field name	Description	Type	Values
RR	Estimated respiratory rate	Float	[0, ∞)
I_E_ratio	Estimated Inhalation/Exhalation Ratio	Float	[0, ∞)
FIT	Estimated Fractional Inspiration Time	Float	[0, ∞)
annotated_inhalation	Manual annotations of inhalation parts. Contains a list of (start, end) pairs indicating the start and the end time (in s) of an inhalation.	List of tuples	[(start1, end1), . . . , (startN, endN)]
annotated_exhalation	Manual annotations of exhalation parts. Contains a list of (start, end) pairs	List of tuples	[(start1, end1), . . . , (startN, endN)]

	indicating the start and the end time (in s) of an exhalation.		

Πίνακας 3.2: Αρχεία json (“experts.breath.json”, “experts.cough.json”, “experts.medical_advice.json”, “experts.speech.json”) που περιλαμβάνουν τις εισροές/ετικέτες (χαρακτηρισμοί, συμβουλές) από τους επαγγελματίες υγείας.

Field name	Description	Type	Values
breath_depth	Depth of breathing	String	“Can breathe deeply enough”/“Cannot breathe deeply enough”
dyspnea_shortness_of_breath	Dyspnea, Shortness of breath	Bool	
stridor	Stridor	Bool	
inspiratory_stridor	Inspiratory stridor	Bool	
expiratory_stridor	Expiratory stridor	Bool	
wheezing	Wheezing	Bool	
respiratory_crackles	Respiratory crackles	Bool	
prolonged_expiration	Prolonged expiration	Bool	
other_and_unspecified_abnormalities_of_breathing	Other and unspecified abnormalities of breathing	Bool	
audible_choking	Audible choking	Bool	
audible_nasal_congestion	Audible nasal congestion	Bool	
no_audible_abnormalities	No audible abnormalities	Bool	

Πίνακας 3.3: Επισημείωση ειδικών επί του βήχα.

Field name	Description	Type	Values
	Experts Impression:		

patient_has	"I think this patient has:"	String	"An upper respiratory tract infection", "A lower respiratory tract infection", "Obstructive lung disease (Asthma, COPD, ...)",
cough_is	"The cough is probably:"	String	"Nothing (healthy cough)", "Pseudo cough/Healthy cough", "Mild (from a sick person)", "Severe (from a sick person)", "Can't tell"
	Type of Cough		
productive	Productive	Bool	
dry	Dry	Bool	
barking_cough	Barking cough	Bool	
hacking_cough	Hacking cough	Bool	
croupy_cough	Croupy cough	Bool	
other_specified_cough	Other specified cough	Bool	
can't_tell	Can't tell	Bool	
	Audible Abnormalities		
audible_dyspnea	Audible dyspnea	Bool	
audible_wheezing	Audible wheezing	Bool	
audible_stridor	Audible stridor	Bool	
audible_choking	Audible choking	Bool	
audible_nasal_congestion	Audible nasal congestion	Bool	
nothing_specific	Nothing specific	Bool	

Πίνακας 3.4: Επισημείωση των ειδικών σχετικά με τον βήχα.

Field name	Description	Type	Values
completion	Expert's answer whether the user completed the phrase	String	"Yes", "No"
hoarseness	Hoarseness of voice	String	"Yes", "No"
volume	Vocal intensity	String	"Low (whisper)", "Normal"
	Audible Abnormalities		
audible_dyspnea	Audible dyspnea	Bool	
audible_wheezing	Audible wheezing	Bool	
audible_stridor	Audible stridor	Bool	
audible_choking	Audible choking	Bool	
audible_nasal_congestion	Audible nasal congestion	Bool	
no_audible_abnormalities	No Audible Abnormalities	Bool	

Πίνακας 3.5: Επισημείωση ειδικών για τις ηχογραφήσεις.

Ανάπτυξη Βάσης Γνώσεων smarty4covid

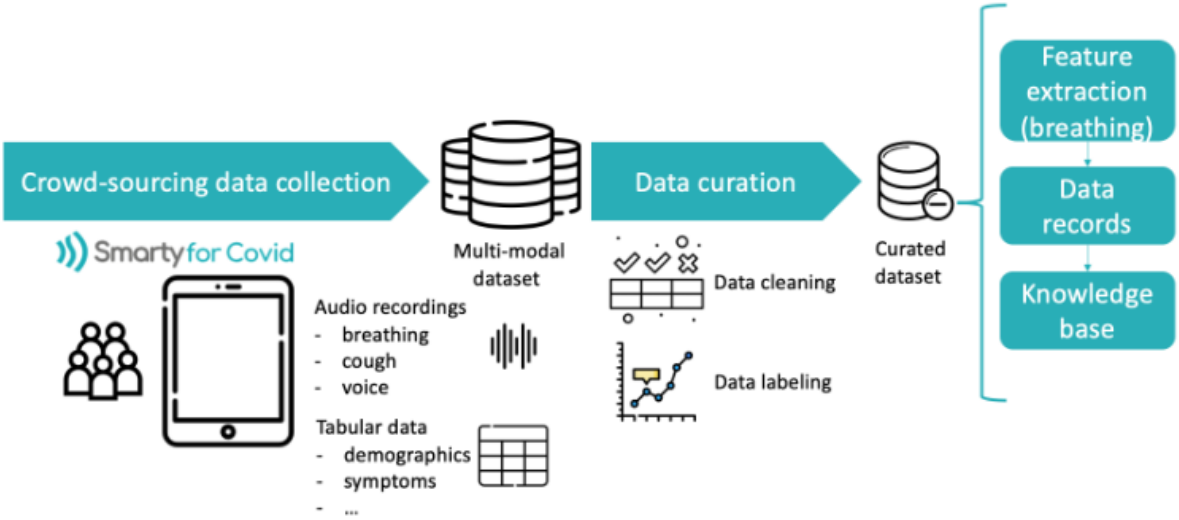
Αφού ολοκληρώθηκαν τα βήματα της συλλογής, επεξεργασίας, εξαγωγής χαρακτηριστικών από τις ηχητικές καταγραφές αναπνοής και της οργάνωσης και δημόσιας διάθεσης των δεδομένων, το επόμενο βήμα ήταν η ανάπτυξη μιας βάσης γνώσεων. Ο σκοπός αυτής της βάσης είναι να ενοποιήσει τα δεδομένα και να διευκολύνει την ανάλυσή τους, επιτρέποντας την εκτέλεση σύνθετων ερωτημάτων και την ανίχνευση χρηστών με συγκεκριμένα χαρακτηριστικά.

Η βάση γνώσεων smarty4covid OWL περιλαμβάνει όλες τις πληροφορίες που συλλέχθηκαν, καθαρίστηκαν και επισημάνθηκαν κατά τη διάρκεια του έργου. Περιέχει δεδομένα σχετικά με τους συμμετέχοντες, τα ερωτηματολόγια, τα αρχεία ήχου και τους επαγγελματίες υγείας που συμμετείχαν στη διαδικασία επισήμανσης. Αυτά τα δεδομένα συνδέονται μέσω καθορισμένων ρόλων, οι οποίοι ορίζουν τις σχέσεις μεταξύ τους.

Η βάση γνώσεων οργανώνει τα δεδομένα σε ιεραρχίες εννοιών, όπως οι τύποι ηχογραφήσεων, τα τεστ COVID-19, οι προϋπάρχουσες καταστάσεις και τα συμπτώματα. Αυτή η οργάνωση διευκολύνει την αναζήτηση και ανάλυση των δεδομένων, επιτρέποντας στους ερευνητές να εκτελούν σύνθετα ερωτήματα και να συνδυάζουν δεδομένα από διαφορετικές πηγές.

Με τη χρήση αυτής της βάσης γνώσεων, οι ερευνητές μπορούν να βελτιώσουν την ανάλυση και την κατανόηση των δεδομένων που σχετίζονται με τον COVID-19, καθιστώντας τα πιο χρήσιμα για ερευνητικούς και κλινικούς σκοπούς.

Η παραπάνω μεθοδολογία παρουσιάζεται στο παρακάτω Σχήμα:



Εικόνα 4.1: Απεικόνιση του συστήματος καταγραφής δεδομένων για το σύνολο δεδομένων Smarty for Covid. Κάθε χρήστης είχε την ικανότητα ελεύθερα να συμμετάσχει και να ηχογραφήσει τον εαυτό του να βήχει, να λέει μια πρόταση και να εισαγάγει σχετικές πληροφορίες. Έπειτα, τα δεδομένα επεξεργάστηκαν και καθαρίστηκαν από ειδικούς.

Στατιστική Ανάλυση και Συμπεράσματα

Στατιστική Ανάλυση και Συμπεράσματα Παρατηρήθηκε ότι υπήρχε υψηλότερο ποσοστό ανδρών (61,0%) σε σχέση με γυναίκες χρήστες. Ωστόσο, υπήρχε ένα ευρύ φάσμα ηλικιών, με την πλειοψηφία των χρηστών να είναι μεταξύ 30 και 59 ετών, μια ηλικιακή ομάδα που χαρακτηρίζεται από αυξημένη εξοικείωση με κινητές συσκευές.

Ένα σημαντικό ποσοστό των χρηστών (79,5%) παρείχε αποτελέσματα τεστ COVID-19 από διάφορους τύπους, όπως PCR και Rapid Antigen. Όσον αφορά τις υποκείμενες ιατρικές καταστάσεις, περισσότεροι από το 27% των χρηστών ανέφεραν τουλάχιστον μία, με την υπέρταση να είναι η πιο συχνά αναφερόμενη. Η κατανομή αυτών των καταστάσεων ήταν παρόμοια με τα δεδομένα της Eurostat για τον γενικό πληθυσμό στην Ελλάδα.

Σχετικά με τα συμπτώματα του COVID-19, περισσότεροι από τους μισούς χρήστες ανέφεραν τουλάχιστον ένα σύμπτωμα. Οι χρήστες που είχαν λάβει δόση ενίσχυσης παρουσίασαν λιγότερα

συμπτώματα από αυτούς που δεν είχαν εμβολιαστεί. Επίσης, η επικράτηση του COVID-19 ήταν χαμηλότερη στους χρήστες που είχαν εμβολιαστεί με δόση ενίσχυσης.

Το σύνολο δεδομένων smarty4covid περιλάμβανε επίσης ζωτικά σημεία όπως ο κορεσμός οξυγόνου και οι παλμοί ανά λεπτό, καθώς και αυτοαναφερόμενα επίπεδα άγχους που σχετίζονται με τον COVID-19. Η ανάλυση έδειξε ότι ο κορεσμός οξυγόνου μειώνεται με την ηλικία, και ότι τα υψηλότερα επίπεδα άγχους σχετίζονται με υψηλότερα ποσοστά εμβολιασμού με δόση ενίσχυσης.

Από τους 4.673 χρήστες, το 17,3% (περίπου 809 άτομα) είχε διαγνωστεί θετικό για COVID-19, όπως αναφέρθηκε κατά την περίοδο εξάπλωσης της παραλλαγής Όμικρον στην Ελλάδα. Αυτά τα συμπεράσματα προσφέρουν πολύτιμες πληροφορίες για την κατανόηση της επίδρασης του COVID-19 και των εμβολιαστικών προγραμμάτων στον πληθυσμό. [41]

4.3 Μεθοδολογία

Αρχικά, όλες οι ηχογραφήσεις βρίσκονταν σε μορφή MP3. Επειδή το μοντέλο Allosaurus δέχεται ως είσοδο μόνο αρχεία WAV, πραγματοποιήθηκε η μετατροπή όλων των αρχείων MP3 σε WAV χρησιμοποιώντας κατάλληλες βιβλιοθήκες μετατροπής ήχου. Αυτή η διαδικασία εξασφαλίζει τη συμβατότητα των αρχείων με το μοντέλο αναγνώρισης φωνημάτων.

Αρχικά, φορτώθηκαν όλα τα αρχεία WAV από τον σχετικό φάκελο και δημιουργήθηκε μια λίστα με τις διαδρομές τους. Στη συνέχεια, χρησιμοποιήθηκε το εργαλείο Allosaurus για την αναγνώριση φωνημάτων από τα ηχητικά αρχεία, εστιάζοντας μόνο στα φωνήεντα και αγνοώντας τα σύμφωνα. Τα αποτελέσματα αυτής της ανάλυσης, τα οποία περιλαμβάνουν το φώνημα και τον χρόνο έναρξής του, αποθηκεύτηκαν σε ένα αρχείο CSV για περαιτέρω επεξεργασία και ανάλυση.

Για την επόμενη φάση, εξήχθησαν πληροφορίες από αρχεία συμμετεχόντων και ερωτηματολογίων. Τα αρχεία αυτά περιλαμβάνουν πληροφορίες όπως το ύψος των συμμετεχόντων και την κλινική στην οποία νοσηλεύτηκαν, ενώ επίσης περιέχουν δεδομένα σχετικά με την κατάσταση του συμμετέχοντα σε σχέση με τον κορωνοϊό. Ακολούθως, τα αποτελέσματα της ανάλυσης φωνημάτων συνδυάστηκαν με τις πληροφορίες από τα ερωτηματολόγια. Σε κάθε συμμετέχοντα προστέθηκε η επιπλέον πληροφορία για το εάν είχε διαγνωστεί θετικός ή αρνητικός στον κορωνοϊό.

Με την ολοκλήρωση αυτής της διαδικασίας, τα δεδομένα συγχωνεύτηκαν σε ένα ενιαίο αρχείο, το οποίο περιλαμβάνει όλες τις σχετικές πληροφορίες για τα φωνήματα και τις ιατρικές καταστάσεις των συμμετεχόντων. Επομένως, δημιουργήθηκε ένα τελικό αρχείο CSV που περιλαμβάνει όλες αυτές τις πληροφορίες, παρέχοντας μια ολοκληρωμένη εικόνα των ηχητικών καταγραφών σε σχέση με τις ιατρικές πληροφορίες, διευκολύνοντας περαιτέρω αναλύσεις και μελέτες.

	patient	questionnaire_id	has_covid	phone	start_time
0	1a1decfc-e047-41eb-abca-306f11f4ad22	a6652b0c-8a32-4817-8fb2-67acf50e7d86	False	i	2.31
1	1a1decfc-e047-41eb-abca-306f11f4ad22	a6652b0c-8a32-4817-8fb2-67acf50e7d86	False	ɔ	2.49
2	1a1decfc-e047-41eb-abca-306f11f4ad22	a6652b0c-8a32-4817-8fb2-67acf50e7d86	False	i	2.67
3	1a1decfc-e047-41eb-abca-306f11f4ad22	a6652b0c-8a32-4817-8fb2-67acf50e7d86	False	u	2.85
4	1a1decfc-e047-41eb-abca-306f11f4ad22	a6652b0c-8a32-4817-8fb2-67acf50e7d86	False	ɛ	3.03
...	...	...	...	...	...
148129	9a7a1563-9e89-4990-9d92-674c47860718	c00ca518-31b7-4228-993f-c024cb28518b	False	o	9.54
148130	9a7a1563-9e89-4990-9d92-674c47860718	c00ca518-31b7-4228-993f-c024cb28518b	False	ɛ	9.66
148131	9a7a1563-9e89-4990-9d92-674c47860718	c00ca518-31b7-4228-993f-c024cb28518b	False	o	9.87
148132	9a7a1563-9e89-4990-9d92-674c47860718	c00ca518-31b7-4228-993f-c024cb28518b	False	o	10.05
148133	9a7a1563-9e89-4990-9d92-674c47860718	c00ca518-31b7-4228-993f-c024cb28518b	False	j	10.23

Εικόνα 4.2: Απεικόνιση του τελικού σχήματος που χρησιμοποιήθηκε για τα πειράματα. Φαίνεται το αποκλειστικό όνομα ανά χρήστη, ανά ερωτηματολόγιο σε περίπτωση που κάποιος χρήστης είχε παραπάνω από μια καταγραφή, το αν έχει covid ή όχι, το φώνημα που εκτιμήθηκε από το Allosaurus καθώς και ο χρόνος όπου το κάθε φώνημα ξεκίνησε σύμφωνα με το προηγούμενο μοντέλο.

Αφού δημιουργήθηκε η παραπάνω δομή CSV, επόμενος στόχος είναι η αξιοποίηση της πληροφορίας που περιέχει. Ο στόχος είναι να εξαχθούν χαρακτηριστικά από κάθε τμήμα ηχητικής καταγραφής που έχει προσδιοριστεί (το οποίο καθορίζεται από την ώρα έναρξης του φωνήματος και εκτείνεται μέχρι μια προκαθορισμένη διάρκεια). Αυτά τα χαρακτηριστικά προκύπτουν μέσω του υπολογισμού του log_mel_spectrogram, το οποίο στη συνέχεια τροφοδοτείται στο μοντέλο VGGish, που είναι προεκπαιδευμένο για να εξάγει χαρακτηριστικά από Spectrograms.

Ωστόσο, πριν από αυτή τη διαδικασία, απαιτείται μια προεπεξεργασία του log_mel_spectrogram, καθώς το μοντέλο VGGish έχει συγκεκριμένες απαιτήσεις για την είσοδό του. Αντλώντας την πληροφορία από το τελικό αρχείο CSV, το οποίο περιλαμβάνει την χρονική στιγμή έναρξης του φωνήματος, μαζί με μια προκαθορισμένη διάρκεια(π.χ. 1 δευτερόλεπτο), προσδιορίζεται το αντίστοιχο κομμάτι της ηχητικής καταγραφής που θα επεξεργαστεί. Υπολογίζεται το log_mel_spectrogram για το επιλεγμένο κομμάτι ηχητικής καταγραφής χρησιμοποιώντας τις απαιτούμενες παραμέτρους για το μοντέλο VGGish.

Αυτές οι παράμετροι περιλαμβάνουν τον ρυθμό δειγματοληψίας του ήχου, που ορίζεται στα 16000 Hz (SAMPLE_RATE), μια μικρή σταθερά 0.01 (LOG_OFFSET) που προστίθεται πριν από τον λογαριθμισμό για να αποφευχθεί η λήψη του λογαρίθμου του μηδενός, τη διάρκεια του παραθύρου για τον υπολογισμό της Σύντομης Μετασχηματισμού Fourier (STFT) που είναι 0.025 δευτερόλεπτα (STFT_WINDOW_LENGTH_SECONDS), τον χρόνο μετατόπισης μεταξύ των παραθύρων STFT που είναι 0.010 δευτερόλεπτα (STFT_HOP_LENGTH_SECONDS), τον αριθμό των φίλτρων Mel που είναι 64 (NUM_MEL_BINS), την κατώτερη συχνότητα για τα φίλτρα Mel που ορίζεται στα 125 Hz (MEL_MIN_HZ), και την ανώτερη συχνότητα για τα φίλτρα Mel που ορίζεται στα 7500 Hz (MEL_MAX_HZ).

Εφόσον παραχθεί το log_mel_spectrogram, το οποίο είναι ένας δισδιάστατος πίνακας όπου οι γραμμές αντιπροσωπεύουν τις διάφορες χρονικές στιγμές και οι στήλες αντιπροσωπεύουν τις τιμές των εντάσεων σε κλίμακα Mel, εφαρμόζεται padding ώστε το μήκος του να είναι πολλαπλάσιο του 96. Αυτό είναι απαραίτητο καθώς το μοντέλο VGGish απαιτεί συγκεκριμένες διαστάσεις εισόδου, συγκεκριμένα πίνακες με σχήμα (96, 64). Για αυτό το λόγο, πρώτα προστίθεται padding στον πίνακα του φασματογραφήματος ώστε το μήκος των χρονικών στιγμών να είναι πολλαπλάσιο του ζητούμενου μήκους. Το padding προσθέτει μηδενικά στο τέλος του log_mel_spectrogram, εξασφαλίζοντας ότι ο πίνακας έχει το σωστό σχήμα. Μετά την προσθήκη του padding, το log_mel_spectrogram χωρίζεται σε τμήματα των 96 χρονικών στιγμών. Κάθε τμήμα έχει σχήμα (96, 64), όπου 64 είναι ο αριθμός των φίλτρων Mel.

Μετά τη διαδικασία padding και διαχωρισμού σε τμήματα, τα τμήματα του log_mel_spectrogram τροφοδοτούνται στο μοντέλο VGGish. Το μοντέλο εξάγει χαρακτηριστικά για κάθε τμήμα του αρχικού Mel Spectrogram. Στη συνέχεια, εφαρμόζεται ένας μέσος όρος στα εξαγόμενα χαρακτηριστικά, παράγοντας έναν πίνακα με 128 αριθμούς που αντιπροσωπεύουν τα τελικά χαρακτηριστικά για την αρχική ηχητική καταγραφή.

Με αυτή τη διαδικασία, χρησιμοποιούνται οι πληροφορίες από το τελικό αρχείο CSV για την εξαγωγή χαρακτηριστικών από τα ηχητικά δεδομένα. Τα χαρακτηριστικά αυτά υπολογίζονται με βάση το log_mel_spectrogram και το μοντέλο VGGish. Η διαδικασία περιλαμβάνει την προεπεξεργασία του log_mel_spectrogram, την προσθήκη padding, τον διαχωρισμό σε τμήματα, την τροφοδότηση στο μοντέλο VGGish και την εξαγωγή των τελικών χαρακτηριστικών.

Αφού περιγράφηκε η διαδικασία εξαγωγής χαρακτηριστικών από τις ηχητικές καταγραφές, το επόμενο βήμα είναι η κατασκευή των συνόλων εκπαίδευσης και δοκιμής. Ο στόχος είναι να δημιουργηθούν αυτά τα σύνολα δεδομένων με τρόπο που να αποφεύγεται το bias, δηλαδή η προκατάληψη στα αποτελέσματα. Αυτό σημαίνει ότι δεν πρέπει να υπάρχουν φωνήματα από τον ίδιο ασθενή τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής, ώστε να διασφαλιστεί η αντικειμενικότητα της αξιολόγησης του μοντέλου.

Αρχικά, επιλέγονται όλες οι γραμμές από το τελικό αρχείο CSV που περιλαμβάνουν το φωνήεν που θα χρησιμοποιηθεί για την ανάλυση. Στη συνέχεια, τα δεδομένα χωρίζονται σε δύο σύνολα: ένα για την εκπαίδευση και ένα για τη δοκιμή. Αυτό επιτυγχάνεται με την τυχαία επιλογή ενός ποσοστού (90%) των

ασθενών για το σύνολο εκπαίδευσης και το υπόλοιπο 10% για το σύνολο δοκιμής, διασφαλίζοντας ότι κάθε ασθενής εμφανίζεται μόνο σε ένα από αυτά τα σύνολα. Με αυτόν τον τρόπο, αποφεύγεται η εισαγωγή bias, καθώς τα φωνήματα του ίδιου ασθενή δεν εμφανίζονται και στα δύο σύνολα.

Για παράδειγμα, η κατανομή των ασθενών μπορεί να είναι ως εξής:

```
Train set
Total: 3368, with Covid: 617, without Covid: 2751 patients
Test
Total: 375, with Covid: 56, without Covid: 319 patients
```

Εικόνα 4.3: Ανάλυση των ασθενών για το σύνολο εκπαίδευσης και επαλήθευσης.

Αφού γίνει ο διαχωρισμός των δεδομένων, δημιουργούνται τα τελικά σύνολα train_df και test_df. Το train_df περιλαμβάνει τα φωνήεντα από τους ασθενείς που βρίσκονται στο σύνολο εκπαίδευσης, χωρισμένα σε δύο κατηγορίες: φωνήεντα από ασθενείς με Covid και φωνήεντα από ασθενείς χωρίς Covid. Αντίστοιχα, το test_df περιλαμβάνει τα φωνήεντα από τους ασθενείς που βρίσκονται στο σύνολο δοκιμής, με τον ίδιο διαχωρισμό: φωνήεντα από ασθενείς με και χωρίς Covid. Αυτός ο διαχωρισμός επιτρέπει την εκπαίδευση του μοντέλου στο train_df και την αξιολόγησή του στο test_df, διασφαλίζοντας ότι το μοντέλο θα μπορεί να γενικεύει τα αποτελέσματά του σε νέους ασθενείς.

Με τα τελικά σύνολα train_df και test_df στη διάθεσή μας, η επόμενη φάση περιλαμβάνει την εξαγωγή χαρακτηριστικών για τα τμήματα των ηχητικών καταγραφών που περιλαμβάνονται σε αυτά τα σύνολα, με βάση την διαδικασία που αναλύθηκε προηγουμένως. Τα τμήματα αυτά καθορίζονται από την ώρα έναρξης (start_time) που έχει εντοπιστεί από το μοντέλο Allosaurus, καθώς και από τη διάρκεια του φωνήματος (phoneme duration) που έχει οριστεί. Μέσω αυτής της διαδικασίας, προκύπτουν τα σύνολα χαρακτηριστικών train_features και test_features, τα οποία θα χρησιμοποιηθούν για την περαιτέρω ανάλυση και εκπαίδευση των μοντέλων.

Με τον υπολογισμό των train_features και test_features, ακολουθεί η εκπαίδευση ενός ταξινομητή MLP (Multi-Layer Perceptron) και η πρόβλεψη με βάση αυτά τα χαρακτηριστικά για το αν ο ασθενής είναι θετικός στον Covid ή όχι. Για την αξιολόγηση της απόδοσης των μοντέλων χρησιμοποιήθηκαν οι μετρικές ακρίβειας (accuracy), precision, ανάκλησης (recall), F1-score, ειδικότητας (specificity), και ROC-AUC score.

Το σύνολο train_df που χρησιμοποιήθηκε είναι μη ισορροπημένο, καθώς περιλαμβάνει λίγα δείγματα ηχητικών καταγραφών από συμμετέχοντες που είναι θετικοί στον κορωνοϊό και πολλά δείγματα από συμμετέχοντες που είναι αρνητικοί. Για να αντιμετωπιστεί αυτό το πρόβλημα, εφαρμόστηκε η τεχνική του cross-validation. Σε αυτή τη διαδικασία, δημιουργήθηκαν υποσύνολα των αρνητικών περιπτώσεων, κάθε ένα από τα οποία είχε τον ίδιο αριθμό δειγμάτων με τις θετικές περιπτώσεις. Με αυτόν τον τρόπο,

διασφαλίστηκε η ισορροπία μεταξύ των θετικών και αρνητικών περιπτώσεων σε κάθε fold της διαδικασίας εκπαίδευσης.

Για παράδειγμα, αν στο σύνολο train_df υπάρχουν 200 φωνήεντα που αντιστοιχούν σε ασθενείς με Covid και 1000 φωνήεντα που αντιστοιχούν σε ασθενείς χωρίς Covid, τότε θα δημιουργηθούν πέντε υποσύνολα, καθένα από τα οποία θα περιλαμβάνει 200 φωνήεντα από ασθενείς με Covid και 200 φωνήεντα από ασθενείς χωρίς Covid. Δηλαδή, το πρώτο υποσύνολο θα έχει 200 θετικά και 200 αρνητικά φωνήεντα, το δεύτερο υποσύνολο θα έχει άλλα 200 αρνητικά φωνήεντα και τα ίδια 200 θετικά, και ούτω καθεξής. Με αυτόν τον τρόπο, κάθε fold της διαδικασίας εκπαίδευσης έχει ισορροπημένα δεδομένα.

Για κάθε fold υποσύνολο που δημιουργήθηκε, πραγματοποιήθηκε αρχικοποίηση και εκπαίδευση ενός MLPClassifier με δύο κρυφά επίπεδα, τα οποία περιλαμβάνουν 128 και 64 νευρώνες αντίστοιχα. Ο ρυθμός εκμάθησης (learning rate) ορίστηκε στο 0.001 και ο μέγιστος αριθμός επαναλήψεων στις 100. Για παράδειγμα, αν το σύνολο εκπαίδευσης περιλαμβάνει 200 ηχητικές καταγραφές από συμμετέχοντες που είναι θετικοί στον κορωνοϊό και 1000 ηχητικές καταγραφές από συμμετέχοντες που είναι αρνητικοί, δημιουργούνται 5 υποσύνολα των 200 αρνητικών περιπτώσεων, καθένα από τα οποία συνδυάζεται με τις 200 θετικές περιπτώσεις. Με αυτόν τον τρόπο, εκπαιδεύονται 5 διαφορετικά μοντέλα MLP classifiers, με σκοπό να ελεγχθεί η απόδοση του Ensemble Learning.

Με τη χρήση της λογικής του ensemble learning, έγινε συνδυασμός των προβλέψεων από τα επιμέρους μοντέλα. Εφαρμόστηκε ένας κανόνας πλειοψηφίας για να καθοριστεί η τελική πρόβλεψη για κάθε δείγμα. Ο κανόνας πλειοψηφίας λειτουργεί ως εξής: για κάθε δείγμα, υπολογίζεται το άθροισμα των προβλέψεων των επιμέρους μοντέλων. Αν το άθροισμα είναι μικρότερο από το μισό του αριθμού των μοντέλων, η τελική πρόβλεψη είναι 0 (αρνητικό), ενώ αν είναι μεγαλύτερο από το μισό, η τελική πρόβλεψη είναι 1 (θετικό). Σε περίπτωση ισοπαλίας (δηλαδή, το άθροισμα είναι ακριβώς ίσο με το μισό των μοντέλων), η τελική πρόβλεψη αποφασίζεται τυχαία.

Με αυτή τη διαδικασία, προέκυψαν οι τελικές τιμές των μετρικών απόδοσης για το σύνολο των μοντέλων (ensemble). Οι τελικές μετρικές παρέχουν μια ολοκληρωμένη εικόνα της απόδοσης του συστήματος και δείχνουν πόσο καλά μπορεί να προβλέψει αν ένας ασθενής είναι θετικός ή αρνητικός στον Covid.

Στη συνέχεια της μεθοδολογίας που περιγράφηκε παραπάνω, ακολουθήθηκε ακριβώς η ίδια διαδικασία και στις επόμενες περιπτώσεις που θα αναφερθούν. Συγκεκριμένα, για όλες τις αναλύσεις που πραγματοποιήθηκαν, διατηρήθηκαν οι ίδιες προσεγγίσεις για τη μετατροπή των αρχείων ήχου, την αναγνώριση των φωνημάτων, την εξαγωγή των log_mel_spectrograms, και την προεπεξεργασία αυτών. Εφαρμόστηκαν οι ίδιες τεχνικές για τη δημιουργία των συνόλων εκπαίδευσης και δοκιμής (train_df και test_df), καθώς και η ίδια προσέγγιση για την αντιμετώπιση του μη ισορροπημένου συνόλου δεδομένων μέσω της τεχνικής του cross-validation. Επίσης, διατηρήθηκε η ίδια διαδικασία εκπαίδευσης του MLP classifier με τη χρήση της λογικής του ensemble learning για τη βελτιστοποίηση της απόδοσης των μοντέλων.

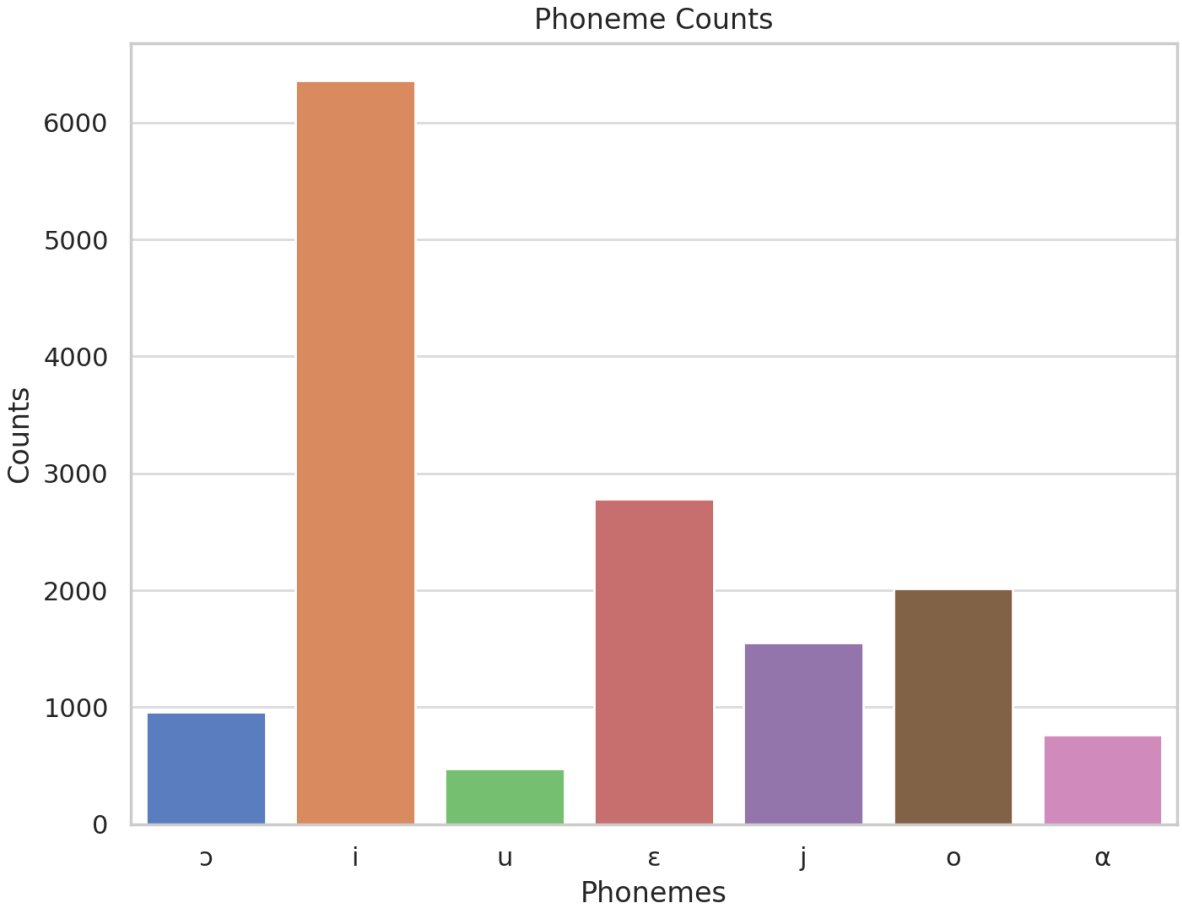
Ειδικότερα, πραγματοποιήθηκαν επιπλέον συνδυασμοί χαρακτηριστικών για την εξαγωγή μετρικών απόδοσης. Συγκεκριμένα, εξήχθησαν τα MFCC , το ZCR, το Spectral Centroid, το Spectral Bandwidth, το Spectral Flatness, το Spectral Rolloff και το Spectral Contrast. Για κάθε έναν από αυτούς τους συνδυασμούς, πραγματοποιήθηκε συνδυασμός των χαρακτηριστικών vgg_features με τα αντίστοιχα χαρακτηριστικά που προαναφέρθηκαν. Επιπλέον, εξετάστηκε και η περίπτωση όπου συνδυάστηκαν όλα τα παραπάνω χαρακτηριστικά. Στη συνέχεια, υπολογίστηκαν οι ίδιες ακριβώς μετρικές απόδοσης που αναφέρθηκαν προηγουμένως, οι μετρικές accuracy, precision, recall, F1-score, ειδικότητας specificity, και ROC-AUC.

Με αυτόν τον τρόπο, διερευνήθηκε η επίδραση των διαφορετικών συνδυασμών χαρακτηριστικών στην απόδοση του μοντέλου MLP και διαπιστώθηκε ποιοι συνδυασμοί παρέχουν τα καλύτερα αποτελέσματα στη διάγνωση του Covid μέσω φωνητικών δεδομένων.

5. Αποτελέσματα

Επαναλαμβάνεται ότι χρησιμοποιήθηκε το μοντέλο Allosaurus όπου γίνεται εξαγωγή των φωνημάτων για το ηχητικό σήμα εισόδου. Επιλέχθηκε να γίνει εκτεταμένη χρήση των φωνηέντων και βάσει αυτού, αποθηκεύτηκε η έξοδος του Allosaurus για τα φωνήματα ‘ɔ','i','u','ɛ', 'j','o' και 'ɑ'.

Τα προηγούμενα έχουν διαφορετική συχνότητα εμφάνισης στον λόγο. Η έξοδος του μοντέλου περιελάμβανε κάθε αναγνωρισμένο φωνήεν, καθώς και τη χρονική στιγμή εμφάνισής του στην ηχητική καταγραφή. Στο παρακάτω Σχήμα 5.1 παρουσιάζεται το πλήθος των φωνηέντων που αναγνώρισε το Allosaurus για κάθε φωνήεν ξεχωριστά (π.χ. "o", "i", "j" κ.λπ.) στο σύνολο ελέγχου.



Εικόνα 5.1: Ραβδόγραμμα που παρουσιάζει την συχνότητα εμφάνισης κάθε φωνήεντος στο σύνολο ελέγχου.

Υπενθυμίζεται ότι λόγω του ότι το Allosaurus δεν εκτιμάει το εύρος εμφάνισης φωνήματος παρά μόνο μια χρονική στιγμή έναρξης, γίνεται χρήση της τελευταίας και ελέγχονται 5 χρονικά διαστήματα (0.1, 0.2, 0.5, 1, 2, 5).

Ως μέτρο σύγκρισης ορίζεται η επίδοση που επιτεύχθηκε χρησιμοποιώντας την μεθοδολογία που περιγράφηκε με ηχητικά δεδομένα μέγιστης διάρκειας, χωρίς δηλαδή να τοποθετηθεί επιπλέον έμφαση σε κάποιο φώνημα. Αυτές οι μετρικές παρουσιάζονται στους πίνακες 5.1. Για λόγους οικονομίας χώρου θα παρουσιάζονται μόνο οι μετρικές για το test set καθώς υπάρχει σύγκλιση κατά την διάρκεια της εκπαίδευσης σε όλα τα μοντέλα.

True Negative	False Positive	False Negative	True Positive
185	161	30	43

Πίνακας 5.1.α

Phoneme	Accuracy	Recall	Precision	F1 score	Specificity	Sensitivity
0.544	0.589	0.211	0.31	0.535	0.589	0.554

Πίνακας 5.1.β

Πίνακας 5.1

Ακολουθούν τα αποτελέσματα όπου τροφοδοτήθηκαν μόνο τα vgg_features για την εκπαίδευση των διαφόρων μοντέλων MLP και παρουσιάζονται στους πίνακες 5.2 έως 5.7. Υπενθυμίζεται ότι στην παρούσα φάση της μεθοδολογίας γίνεται χρήση των φωνημάτων ανεξάρτητα από τον κάθε ασθενή. Λεπτομερέστερα, ο διαχωρισμός σε σύνολα δεδομένων εκπαίδευσης και ελέγχου γίνεται πριν από την εξαγωγή των φωνημάτων και επομένως, μαζεύονται τα φωνήματα για τους χρήστες που ανήκουν στα σύνολα δεδομένων εκπαίδευσης και ελέγχου. Έπειτα, τα φωνήματά τους τροφοδοτούνται ανεξάρτητα ως δεδομένα εκπαίδευσης αλλά και εκτιμάται το αν είναι φώνημα από ασθενή με OVID-19ανεξάρτητα. Για τον παραπάνω λόγο υπάρχει διαφορά στο πλήθος δεδομένων στους πίνακες 5.1 και 5.2 και έπειτα.

Start time of Phoneme + 0.1 sec

Phoneme	True Negative	False Positive	False Negative	True Positive
ɔ	316	442	74	123
i	3720	1510	745	389
u	115	260	27	69
ε	1139	1238	153	248
j	424	830	90	205
o	553	1210	73	201
α	167	434	60	105

Πίνακας 5.2.α

Start time of Phoneme + 0.1 sec							
Phoneme	Accuracy	Recall	Precision	F1 score	Specificity	Sensitivity	Roc-Auc
ɔ	0.459	0.624	0.217	0.322	0.416	0.624	0.529
i	0.645	0.343	0.204	0.256	0.711	0.343	0.54
u	0.39	0.718	0.209	0.324	0.306	0.718	0.52
ε	0.499	0.618	0.166	0.262	0.479	0.618	0.57
j	0.406	0.694	0.198	0.308	0.338	0.694	0.538
o	0.363	0.733	0.142	0.238	0.305	0.733	0.534
α	0.355	0.636	0.194	0.298	0.277	0.636	0.449

Πίνακας 5.2.β

Πίνακας 5.2

Start time of Phoneme + 0.2 sec				
Phoneme	True Negative	False Positive	False Negative	True Positive
ɔ	289	469	77	120
i	3361	1869	667	467
u	175	200	32	64
ε	1181	1196	178	223
j	464	789	102	193
o	1027	716	157	117
α	269	332	70	95

Πίνακας 5.3.α

Start time of Phoneme + 0.2 sec							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.428	0.609	0.203	0.305	0.381	0.609	0.493
i	0.601	0.411	0.199	0.269	0.642	0.411	0.534
u	0.507	0.666	0.242	0.355	0.466	0.666	0.549
ε	0.505	0.556	0.157	0.245	0.496	0.556	0.544
j	0.424	0.654	0.196	0.302	0.370	0.654	0.521
o	0.567	0.427	0.140	0.211	0.589	0.427	0.513
α	0.475	0.575	0.222	0.32	0.447	0.575	0.51

Πίνακας 5.3.β

Πίνακας 5.3

Start time of Phoneme + 0.5 sec				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	317	441	86	111
i	3009	2221	645	489
u	169	206	39	57
ε	964	1413	150	251
j	572	682	121	174
o	1054	689	161	114
α	380	221	82	83

Πίνακας 5.4.α

Start time of Phoneme + 0.5 sec							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.448	0.563	0.201	0.296	0.418	0.563	0.504
i	0.550	0.431	0.18	0.254	0.575	0.431	0.507
u	0.480	0.594	0.217	0.318	0.451	0.594	0.519

ε	0.437	0.626	0.151	0.243	0.406	0.626	0.520
j	0.482	0.590	0.203	0.302	0.456	0.590	0.522
o	0.579	0.415	0.142	0.212	0.605	0.415	0.503
α	0.604	0.503	0.273	0.354	0.632	0.503	0.594

Πίνακας 5.4.β

Πίνακας 5.4

Start time of Phoneme + 1 sec				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	333	425	71	126
i	2218	3012	476	658
u	184	191	36	60
ε	944	1433	127	274
j	525	729	97	198
o	1021	722	156	120
α	346	255	91	74

Πίνακας 5.5.α

Start time of Phoneme + 1 sec							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.48	0.639	0.228	0.336	0.439	0.639	0.541
i	0.451	0.58	0.179	0.273	0.424	0.582	0.50
u	0.518	0.625	0.239	0.345	0.490	0.625	0.567
ε	0.438	0.683	0.16	0.26	0.3971	0.683	0.553
j	0.466	0.671	0.213	0.324	0.418	0.671	0.549
o	0.565	0.434	0.142	0.214	0.5858	0.434	0.525
α	0.548	0.448	0.224	0.299	0.575	0.448	0.515

Πίνακας 5.5.β

Πίνακας 5.5

Start time of Phoneme + 2 sec				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	304	454	68	129
i	2451	2779	481	653
u	186	189	30	66
ε	1009	1368	128	273
j	555	699	122	173
o	977	766	134	143
α	353	248	101	64

Πίνακας 5.6.α

Start time of Phoneme + 2 sec							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.453	0.654	0.221	0.330	0.401	0.654	0.532
i	0.487	0.575	0.190	0.286	0.468	0.575	0.532
u	0.535	0.687	0.258	0.376	0.496	0.687	0.594
ε	0.461	0.680	0.166	0.267	0.424	0.680	0.574
j	0.469	0.586	0.198	0.296	0.442	0.586	0.509
o	0.554	0.516	0.157	0.241	0.560	0.516	0.548
α	0.544	0.387	0.205	0.268	0.587	0.387	0.517

Πίνακας 5.6.β

Πίνακας 5.6

Start time of Phoneme + 5 sec				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	236	522	69	128
i	2434	2796	495	639
u	184	191	45	51
ε	1204	1173	172	229
j	561	693	117	178
o	907	836	115	162

α	338	263	88	77

Πίνακας 5.7.α

Start time of Phoneme + 5 sec							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.381	0.649	0.196	0.302	0.311	0.649	0.466
i	0.482	0.563	0.186	0.279	0.465	0.563	0.52
u	0.498	0.531	0.217	0.301	0.490	0.531	0.543
ɛ	0.515	0.571	0.163	0.254	0.506	0.571	0.555
j	0.477	0.603	0.204	0.305	0.447	0.603	0.542
o	0.529	0.584	0.162	0.254	0.520	0.584	0.571
α	0.541	0.466	0.226	0.304	0.562	0.466	0.524

Πίνακας 5.7.β

Πίνακας 5.7

Ακολουθούν τα βέλτιστα φωνήματα που προέκυψαν για κάθε χρονική διάρκεια χρησιμοποιώντας μόνο τα vgg_features.

Phoneme/ Features	**Duration**	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.1	0.459	0.624	0.217	0.322	0.416	0.624	0.529
u	0.2	0.507	0.666	0.242	0.355	0.466	0.666	0.549
α	0.5	0.604	0.503	0.273	0.354	0.632	0.503	0.594
u	1	0.518	0.625	0.239	0.345	0.49	0.625	0.56
u	2	0.535	0.687	0.258	0.376	0.496	0.687	0.594
j	5	0.477	0.603	0.204	0.305	0.447	0.603	0.542

Πίνακας 5.8

Για την βελτίωση των παραπάνω αποτελεσμάτων και κατάλληλη χρήση της πληροφορίας φωνημάτων, προσαυξάνονται τα διανύσματα VGG με μια πληθώρα από επιλεγμένα χαρακτηριστικά

ομιλίας και ήχου τα οποία είναι τα Mel Frequency Cepstral Coefficients, Spectral Bandwidth, Spectral Centroid, Spectral Contrast, Spectral Flatness, Spectral Rolloff και Zero-Crossing Rate.

Λόγω του αυξημένου πλήθους των μετρικών και των επιπλέον χαρακτηριστικών που θα προστεθούν (concatenate) στο διάνυσμα χαρακτηριστικών του VGG, προκύπτει μεγάλος όγκος αποτελεσμάτων. Αναφέρονται παρακάτω οι βέλτιστες περιπτώσεις που εντοπίστηκαν με την προσαύξηση χαρακτηριστικών. Στην πρώτη στήλη του πίνακα 5.8 προστίθεται σε παρένθεση το σύνολο των χαρακτηριστικών που έχουν χρησιμοποιηθεί για το εκάστοτε πείραμα. Επισημαίνεται ότι με τη λέξη "all" χρησιμοποιούνται όλα τα χαρακτηριστικά που αναφέρθηκαν στην προηγούμενη παράγραφο, δηλαδή τα χαρακτηριστικά MFCCs, Spectral Bandwidth, Spectral Centroid, Spectral Contrast, Spectral Flatness, Spectral Rolloff και Zero-Crossing Rate, σε συνδυασμό με τα vgg_features. Το άνω βέλος (↑) προστίθεται στις μετρικές όπου υπήρξε βελτίωση σε σχέση με τις αντίστοιχες μετρικές που προέκυψαν με τη χρήση μόνο των vgg_features.

Phoneme/Features	Duration	Accuracy	Recall	Precision	F1 score	Specificity	Sensitivity	Roc-Auc
ɛ (VGG + Spectral Centroid)	0.1	0.546 ↑	0.454	0.149	0.224	0.562 ↑	0.454	0.51
ɔ (All)	0.2	0.413	0.752 ↑	0.156	0.258	0.36	0.752 ↑	0.556 ↑
i (All)	0.5	0.546	0.618 ↑	0.222 ↑	0.327 ↑	0.53	0.618 ↑	0.586 ↑
ɑ (All)	1	0.231	0.976 ↑	0.216	0.353 ↑	0.03	0.976 ↑	0.588 ↑
ɑ (All)	2	0.342	0.861 ↑	0.228 ↑	0.36 ↑	0.2	0.861 ↑	0.542 ↑
o (All)	5	0.402 ↑	0.906 ↑	0.175 ↑	0.294 ↑	0.323	0.906 ↑	0.61 ↑

Πίνακας 5.9

Ενδεικτικά γίνεται παρουσίαση και των άλλων φωνημάτων για τις παραπάνω περιπτώσεις για λόγους επιβεβαίωσης του γεγονότος όπου επιλέχθηκαν αυτά ως τα βέλτιστα. Πλέον στον τίτλο του πάνω μέρους των πινάκων αναγράφεται και ο συνδυασμός των χαρακτηριστικών που επιλέχθηκαν να χρησιμοποιηθούν.

Start time of Phoneme + 0.1 sec (VGG+Spectral Centroid)				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	758	0	197	0
i	0	5230	0	1134
u	0	375	0	96

ε	979	1398	163	238
j	0	1254	1	294
o	1743	0	274	0
α	594	7	163	2

Πίνακας 5.10.α

Start time of Phoneme + 0.1 sec (VGG + Spectral Centroid)							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.794	0.00	0.00	0.00	1.00	0.00	0.50
i	0.178	1.00	0.178	0.302	0.00	1.00	0.525
u	0.204	1.00	0.204	0.339	0.00	1.00	0.505
ε	0.438	0.593	0.145	0.234	0.412	0.594	0.505
j	0.19	0.997	0.19	0.319	0.00	0.997	0.497
o	0.864	0.00	0.00	0.00	1.00	0.000	0.488
α	0.778	0.01	0.222	0.023	0.989	0.012	0.505

Πίνακας 5.10.β

Πίνακας 5.10

Start time of Phoneme + 0.2 sec (All)				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	435	323	86	111
i	3984	1246	783	351
u	272	103	63	33
ε	2250	127	362	39
j	886	368	171	124
o	199	1544	21	253
α	388	213	99	66

Πίνακας 5.11.α

Start time of Phoneme + 0.2 sec (All)							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.572	0.563	0.256	0.352	0.574	0.563	0.572
i	0.681	0.31	0.22	0.257	0.762	0.309	0.556

u	0.648	0.344	0.243	0.284	0.725	0.344	0.552
ε	0.823	0.097	0.235	0.138	0.947	0.097	0.523
j	0.652	0.42	0.252	0.315	0.706	0.42	0.59
o	0.224	0.923	0.141	0.244	0.114	0.923	0.56
α	0.592	0.4	0.237	0.298	0.646	0.4	0.534

Πίνακας 5.11.β

Πίνακας 5.11

Start time of Phoneme + 0.5 sec (All)				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	50	708	16	181
i	2774	2456	433	701
u	373	2	95	1
ε	1175	1202	134	267
j	1137	117	232	63
o	1523	220	215	60
α	337	264	82	83

Πίνακας 5.12.α

Start time of Phoneme + 0.5 sec (All)							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.242	0.919	0.204	0.333	0.066	0.919	0.508
i	0.546	0.618	0.222	0.327	0.53	0.618	0.586
u	0.794	0.01	0.333	0.02	0.995	0.01	0.524
ε	0.519	0.666	0.182	0.286	0.494	0.666	0.614
j	0.775	0.214	0.34	0.265	0.907	0.214	0.573
o	0.784	0.218	0.21	0.216	0.874	0.218	0.575
α	0.5348	0.503	0.219	0.324	0.56	0.503	0.546

Πίνακας 5.12.β

Πίνακας 5.12

Start time of Phoneme + 1 sec (All)				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**

ɔ	607	151	160	37
i	4515	715	892	242
u	41	334	6	90
ε	173	2204	5	396
j	245	1009	25	270
o	133	1610	8	268
α	16	585	4	161

Πίνακας 5.13.α

Start time of Phoneme + 1 sec (All)							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.674	0.188	0.197	0.192	0.8	0.188	0.506
i	0.747	0.213	0.252	0.231	0.863	0.213	0.546
u	0.278	0.938	0.212	0.346	0.109	0.937	0.508
ε	0.205	0.988	0.152	0.264	0.072	0.988	0.593
j	0.332	0.915	0.142	0.249	0.076	0.971	0.585
o	0.199	0.971	0.143	0.249	0.076	0.971	0.585
α	0.2	0.976	0.216	0.343	0.026	0.976	0.588

Πίνακας 5.13.β

Πίνακας 5.13

Start time of Phoneme + 2 sec (All)				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	351	407	77	120
i	3675	1555	650	484
u	41	334	6	90
ε	9	2368	0	401
j	1191	63	263	32
o	0	1743	0	277
α	120	481	23	142

Πίνακας 5.14.α

Start time of Phoneme + 2 sec (All)							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**

ɔ	0.493	0.609	0.228	0.331	0.463	0.609	0.539
i	0.654	0.427	0.237	0.305	0.703	0.427	0.569
u	0.278	0.936	0.212	0.346	0.109	0.938	0.527
ε	0.148	1.000	0.145	0.253	0.003	1.000	0.504
j	0.79	0.108	0.337	0.164	0.95	0.108	0.587
o	0.137	1	0.137	0.241	0.000	1.000	0.539
α	0.342	0.861	0.228	0.361	0.2	0.861	0.542

Πίνακας 5.14.β

Πίνακας 5.14

Start time of Phoneme + 5 sec (All)				
Phoneme	**True Negative**	**False Positive**	**False Negative**	**True Positive**
ɔ	55	703	9	188
i	528	4702	73	1061
u	17	358	2	94
ε	1968	409	289	112
j	1181	73	266	29
o	563	1180	26	251
α	250	351	51	114

Πίνακας 5.15.α

Start time of Phoneme + 5 sec (All)							
Phoneme	**Accuracy**	**Recall**	**Precision**	**F1 score**	**Specificity**	**Sensitivity**	**Roc-Auc**
ɔ	0.255	0.954	0.211	0.346	0.073	0.954	0.508
i	0.25	0.936	0.184	0.308	0.101	0.936	0.552
u	0.236	0.98	0.208	0.343	0.045	0.979	0.557
ε	0.749	0.279	0.215	0,243	0.828	0.279	0.566
j	0.781	0.098	0.284	0.146	0.942	0.098	0.561
o	0.403	0.906	0.175	0.294	0.323	0.906	0.61
α	0.475	0.691	0.245	0.362	0.416	0.691	0.57

Πίνακας 5.15.β

Πίνακας 5.15

Τα προηγούμενα αποτελέσματα αφορούσαν την αξιολόγηση κάθε φωνήματος ξεχωριστά. Ωστόσο, για να προκύψουν μετρικές που είναι συγκρίσιμες με τον πίνακα 5.1, πρέπει για τις βέλτιστες περιπτώσεις που παρουσιάστηκαν προηγουμένως, να υπολογιστεί για κάθε χρήστη αν έχει COVID-19 ή όχι, λαμβάνοντας υπόψη την πλειοψηφία των φωνημάτων. Συγκεκριμένα, αν η πλειοψηφία των φωνημάτων κριθεί ως θετική από τα αντίστοιχα μοντέλα, τότε ο χρήστης θα χαρακτηριστεί ως ασθενής, ενώ σε αντίθετη περίπτωση θα θεωρηθεί υγιής. Το άνω βέλος (↑) συμβολίζει σχετική βελτίωση σε σύγκριση με τον πίνακα 5.1.

Phoneme/Feature	Duration	True Negative	False Positive	False Negative	True Positive
ε (VGG + Spectral Centroid)	0.1	163	171	25	35
ɔ (All)	0.2	148	92	31	24
i (All)	0.5	206	132	29	46
α (All)	1	4	231	0	66
α (All)	2	39	196	6	60
o (All)	5	106	213	10	46

Πίνακας 5.16.α

Phoneme/Features	Duration	Accuracy	Recall	Precision	F1 score	Specificity	Sensitivity	Roc-Auc
ε (VGG + Spectral Centroid)	0.1	0.502	0.583	0.17	0.263	0.488	0.583	0.536
ɔ (All)	0.2	0.583↑	0.436	0.207	0.28	0.617↑	0.436	0.527
i (All)	0.5	0.61↑	0.613↑	0.258↑	0.364↑	0.609↑	0.613↑	0.611↑
α (All)	1	0.233	1.00↑	0.222	0.364↑	0.017	1.00↑	0.509
α (All)	2	0.329	0.909↑	0.234↑	0.373↑	0.166	0.909↑	0.538
o (All)	5	0.405	0.821↑	0.178	0.292	0.332	0.821↑	0.577↑

Πίνακας 5.16.β

Πίνακας 5.16

Για την πληρότητα των πειραμάτων και της σύγκρισης, γίνεται αναφορά και στις βέλτιστες επιδόσεις που επιτεύχθηκαν με χρήση των VGG χαρακτηριστικών αποκλειστικά σε σύγκριση με κάθε ασθενή και όχι ανά φώνημα. Το άνω βέλος (↑) συμβολίζει σχετική βελτίωση σε σύγκριση με τον πίνακα 5.1.

Phoneme/ Feature	Duration	True Negative	False Positive	False Negative	True Positive
ɔ	0.1	102	138	19	36
u	0.2	105	101	17	28
α	0.5	139	96	24	42
u	1	105	101	15	30
u	2	99	107	13	32
o	5	117	152	22	37

Πίνακας 5.17.α

Phoneme/ Features	Duration	Accuracy	Recall	Precision	F1 score	Specificity	Sensitivity	Roc-Auc
ɔ	0.1	0.468	0.655 ↑	0.207	0.314 ↑	0.425	0.655↑	0.54
u	0.2	0.53	0.622 ↑	0.217 ↑	0.322 ↑	0.51	0.622↑	0.566 ↑
α	0.5	0.601 ↑	0.636 ↑	0.304 ↑	0.411 ↑	0.591↑	0.636↑	0.614 ↑
u	1	0.538	0.667 ↑	0.229 ↑	0.341 ↑	0.51	0.667↑	0.588 ↑
u	2	0.522	0.711 ↑	0.23 ↑	0.348 ↑	0.481	0.711↑	0.596 ↑
o	5	0.469	0.627 ↑	0.196	0.298	0.435	0.627↑	0.531

Πίνακας 5.17.β

Πίνακας 5.17

6. Συμπεράσματα

6.1. Περιορισμοί

Πριν παρουσιασθούν τα συμπεράσματα, είναι κρίσιμο να αναφερθούν προβλήματα σχετικά με την παρούσα μεθοδολογία. Πρώτον, η ύπαρξη και ο χρόνος εμφάνισης των φωνημάτων βασίζονται έντονα στις προβλέψεις του Allosaurus. Το μοντέλο αυτό όπως έχει προαναφερθεί, εκτιμά σωστά το φώνημα ενδιαφέροντος και χρησιμοποιεί ένα επιπλέον μοντέλο ώστε να εκτιμηθεί και ο χρόνος εμφάνισης. Επομένως, ένα σοβαρό σημείο που μπορεί να απαιτεί βελτίωση είναι το δευτερεύον μοντέλο εκτίμησης του χρόνου εμφάνισης ενός φωνήματος. Στην παρούσα εργασία, έγινε χειροκίνητη εκτίμηση των αποτελεσμάτων που προέκυψαν για τα φωνήεντα που αναγνωρίστηκαν μαζί με τον αντίστοιχο χρόνο έναρξης τους και διαπιστώθηκε ότι η εκτίμηση του χρόνου ήταν ικανοποιητική για τα περισσότερα φωνήεντα.

Σε επόμενο χρόνο και γνωρίζοντας ότι τα φωνήματα έχουν σχετικά μικρή διάρκεια, είναι απαραίτητο να αναφερθεί ότι οι διάρκειες φωνημάτων 1,2 και 5 δευτερολέπτων είναι σίγουρο ότι περιέχουν και πληροφορία άλλων φωνημάτων (φωνηέντων, συμφώνων και κενών ομιλίας). Επομένως και σε περίπτωση όπου αναφερθεί βελτίωση λόγο χρήσης φωνήματος με διάρκεια πάνω από 0.5 δευτερόλεπτα, απαιτείται επιπλέον μελέτη σχετικά με το αν όντως υπάρχει περισσότερη διάρκεια του φωνήματος ενδιαφέροντος ή τις υπόλοιπης πληροφορίας.

Τρίτον, το μοντέλο VGG έχει εκπαιδευθεί σε πληροφορία ήχου, μη εξειδικευμένη στην ομιλία. Επομένως, τα χαρακτηριστικά που παράγονται μπορεί να μην είναι ικανά έτσι ώστε να ενσωματώσουν με λεπτομέρεια τις πληροφορίες φωνημάτων και ομιλίας, παρά γενικότερα χαρακτηριστικά που μπορεί να είναι χρήσιμα σε εφαρμογές μουσικής και περιβαλλοντικών ήχων. Η επιλογή του VGG έγινε για να είναι συγκρίσιμη η παρούσα εργασία με προηγούμενες που χρησιμοποίησαν το ίδιο σύνολο δεδομένων, έτσι ώστε να γίνει ανάλυση του κατά πόσο απομονωμένα φωνήματα είναι σημαντικά για την εύρεση του Covid μέσω της ομιλίας.

Τέταρτον, τα δεδομένα εκπαίδευσης έχουν παραχθεί από προσωπική ηχογράφηση ομιλίας. Ως αποτέλεσμα, δεν υπάρχει ομοιογένεια στο μικρόφωνο, στην συμπίεση του σήματος, στον θόρυβο που μπορεί να παρουσιασθεί ακόμα και στην συχνότητα δειγματοληψίας. Επομένως, κρίνεται απαραίτητο να αναφερθεί ότι πρέπει το παρών σύνολο δεδομένων να καθαριστεί επιπλέον έτσι ώστε να επιτευχθεί η ομοιογένειά του. Επιπλέον, τα δεδομένα έχουν παραδοθεί σε μορφή MP3, μια μορφή απωλεστικής συμπίεσης που οδηγεί σε περαιτέρω απώλεια ακουστικής πληροφορίας, κυρίως στις υψηλές συχνότητες.

Πέμπτον, είναι αναγκαίο να γίνει περαιτέρω διερεύνηση του κατάλληλου ορίου (threshold) για την εξαγωγή της τελικής απόφασης του ταξινομητή. Υπενθυμίζεται ότι ως θετικό ασθενή (ή φώνημα) κρίνεται αυτό για το οποίο το μοντέλο έχει έξοδο μεγαλύτερη του 0.5. Είναι γνωστό ότι σε μη ισοροπημένα

(unbalanced) σύνολα δεδομένων το παραπάνω μπορεί να μην ισχύει και αφήνεται ως μελλοντική δουλειά.

6.2 Σχολιασμός αποτελεσμάτων

Από τον πίνακα 5.1 είναι εμφανές ότι τα αποτελέσματα του βασικού πειράματος είναι χαμηλά και προσεγγίζουν την τυχαία επιλογή καθώς το ROC-AUC τείνει στο 0.5. Επομένως, η γενίκευση του μοντέλου είναι μη ισχυρή και είναι δύσκολο να εξαχθεί το συμπέρασμα ότι χρησιμοποιώντας όλη την ηχογράφηση υπάρχει δυνατότητα εκτίμησης της πάθησης. Παρ'όλα αυτά, το παραπάνω αποτελεί το πείραμα βάσης.

Η προηγούμενη συμπεριφορά παρατηρείται σε αρκετά από τα επιπλέον πειράματα διερεύνησης του βέλτιστου φωνήματος και διάρκειας φωνήματος, καθώς υπήρξαν και κάποιες ακραίες συμπεριφορές όπου το Precision έγινε βέλτιστο, ρίχνοντας αρκετά όλες τις άλλες επιδόσεις ενώ το ROC-AUC παρέμεινε να προσεγγίζει την τυχαία περίπτωση. Παράλληλα, δεν υπήρχε σχεδόν ποτέ κάποιος πρωταγωνιστής σε οποιοδήποτε από τα πειράματα, καθώς υπήρχαν ταυτόχρονες αυξήσεις, αλλά και μειώσεις, στις μετρικές που υπολογίσθηκαν. Επομένως, ήταν αναγκαίο να γίνει εποπτεία και να επιλεχθούν οι περιπτώσεις όπου υπήρχε μια γενικότερη βελτίωση, χωρίς σύγκριση μεμονομένων μετρικών.

Για τα πειράματα χρησιμοποιώντας μόνο τα δεδομένα του VGG με σκοπό την εύρεση των πιο υποσχόμενων φωνημάτων για κάθε διάρκεια από τις 6 που επιλέχθηκαν για έλεγχο, κρίνεται αναγκαίο να αναφερθεί ότι δεν υπήρχε αδιαφιλονίκητος πρωταγωνιστής, όπως προαναφέρθηκε. Οι τιμές του ROC-AUC κυμαίνονται από 0.44 έως 0.59. Το παραπάνω προσδίδει βαρύτητα στο επιχείρημα ότι το VGG μπορεί να μην είναι το κατάλληλο για να επεξεργαστεί αυτά τα δεδομένα ομιλίας, καθώς σε καμία περίπτωση δεν υπήρχαν ικανοποιητικές αυξήσεις στις μετρικές. Επίσης, δεν υπήρξε κάποιο φώνημα όπου ξεχώρισε, πράγμα το οποίο χρειάζεται περαιτέρω διερεύνηση αλλά μπορεί να αποτελεί ένδειξη ότι χρήση συνόλου δικτύων, εκπαιδευμένων σε διαφορετικές πτυχές της ομιλίας μπορεί να αποτελεί την κατάλληλη μέθοδο.

Ενδεικτικά αναφέρονται τα 2 καλύτερα φωνήματα ανά χρονική περίοδο χρησιμοποιώντας μόνο τα δεδομένα του VGG:

1. 0.1: ε, ɔ με μικρό προβάδισμα του ɔ 4/7 μετρικές αλλά καλύτερες μετρικές ολιστικά από το ε
2. 0.2: u, ɔ: Με σχετική πτώση των μετρικών συνολικά από τις περιπτώσεις για διάρκεια 0.1 δευτερόλεπτα. Στην συγκεκριμένη περίπτωση το u υπερτερεί ως προς την απόδοση του ɔ σχεδόν σε όλες τις μετρικές.
3. 0.5: α, ε: Το α πλέον αποτελεί την ιδανικότερη λύση με ROC-AUC 0.594, όπου πλέον αποτελεί μια σχετικά καλή απόδοση.
4. 1: Τα φωνήματα για διάρκεια 0.1 παραμένουν ως βέλτιστα και η σχέση τους παραμένει η ίδια όπως και στην περίπτωση των 0.2 δευτερολέπτων, με το u να είναι λίγο λιγότερο διακριτό ως το βέλτιστο φώνημα.

5. 2: u,ε: Το u πλέον πετυχαίνει τις καλύτερες επιδόσεις ως προς το Recall (0.687) και F1-Score (0.376) και έχει ίση επίδοση με το α για την διάρκεια 0.5. Επομένως, το u με διάρκεια 2 δευτερολέπτων αποτελεί ακόμα μια επιλογή ως το βέλτιστο φώνημα. Σημειώνεται ξανά ότι φωνήματα με διάρκεια 2 δευτερολέπτων είναι εξαιρετικά σπάνια, επομένως το πόρισμα ότι το συγκεκριμένο φώνημα έχει προβλεπτικές ιδιότητες γίνεται ασθενέστερο και απαιτείται ενδελεχής έλεγχος.
6. 5: o, j: Πλέον τα ROC-AUC είναι χαμηλότερα των βέλτιστων αλλά το Accuracy παρουσιάζει μια μικρή, αλλά αντιληπτή, αύξηση με ταυτόχρονη μείωση του F1-Score.

Στην συνέχεια και λόγω της χαμηλής επίδοσης των περισσότερων μοντέλων, έγινε διερεύνηση ως προς την χρήση επιπλέον χαρακτηριστικών επί των χαρακτηριστικών που εξάγει το VGG. Επιλέχθηκαν γενικά χαρακτηριστικά ήχου που βασίζονται κυρίως στην φασματική πληροφορία. Γενικότερα, δεν υπήρχε κάποια πολύ αισθητή αύξηση στις μετρικές και ταυτόχρονα, σε κάποιες περιπτώσεις όπου παρουσιάζονται στους πίνακες 5.9 - 5.14, υπήρχε και πλήρης αδυναμία των μοντέλων για γενίκευση καθώς είχαν ως έξοδο μόνιμα Positive ή Negative. Το προηγούμενο φαινόμενο μπορεί να οφείλεται στην αύξηση των διαστάσεων του διανύσματος χαρακτηριστικών χωρίς την αντίστοιχη αύξηση του αριθμού των διαθέσιμων δεδομένων.

Παρουσιάζεται μια σύγκριση μεταξύ των βέλτιστων φωνημάτων με την προσαύξηση χαρακτηριστικών και αυτών για τα οποία χρησιμοποιήθηκαν μόνο τα VGG:

1. 0.1: αρκετή πτώση του ε εκτός από το Accuracy και το Specificity. Σε κάθε περίπτωση όμως αποτελεί υποβέλτιστο φώνημα έναντι των ε, ɔ με χρήση των χαρακτηριστικών VGG.
2. 0.2: βέλτιστο ROC-AUC με ταυτόχρονη βελτίωσή του Recall, αλλά μείωση όλων των υπόλοιπων μετρικών
3. 0.5: Στην συγκεκριμένη περίπτωση κρίνεται απαραίτητο να σημειωθεί ότι με την χρήση όλων των χαρακτηριστικών, το φωνημα i πήρε πρωταγωνιστικό ρόλο έναντι της σχεδόν τελευταίας θέσης λόγω κακής επίδοσης με αποκλειστική χρήση των χαρακτηριστικών VGG. Επίσης το ROC-AUC είχε μικρή πτώση από το συνολικά βέλτιστο α διάρκειας 0.5 δευτερολέπτων και αντίστοιχη αύξηση με την περίπτωση των 0.2 δευτερολέπτων για Recall και συγκρίσιμα αποτελέσματα με το προαναφερθέν α για τις προηγούμενες περιπτώσεις.
4. 1: Ιδιάζουσα περίπτωση όπου υπονομεύθηκε η επίδοση για κάθε φώνημα, όπου φαίνεται στο πολύ μικρό Specificity. Παρ'όλα αυτά και για να αποδοθεί παραπάνω βαρύτητα στο επιχείρημα ότι είναι απαραίτητη η χρήση αρκετών μετρικών, παρατηρείται αύξηση σε 4/7 μετρικές έναντι των αντίστοιχων φωνημάτων με αποκλειστική χρήση του VGG.
5. 2 και 5 δευτερόλεπτα: Αντίστοιχη συμπεριφορά με την περίπτωση της προσαύξησης χαρακτηριστικών με διάρκεια 1 δευτερολέπτου και υπέρβαση του βέλτιστου ROC-AUC κατά 0.1 με τελικό βέλτιστο στα 0.61.

Από την προηγούμενη ανάλυση προκύπτει περισσότερη βαρύτητα στην αδυναμία του VGG, καθώς τα βέλτιστα φωνήματα και οι αντίστοιχες μετρικές άλλαξαν. Σε συνολική εικόνα φαίνεται ότι η χρήση

επιπλέον χαρακτηριστικών μπορούν να οδηγήσουν σε σχετική βελτίωση, και κυρίως των μετρικών Recall και ROC-AUC. Παράλληλα επειδή δεν υπάρχει σταθερή και εύρωστη βελτίωση, δεν μπορεί να εξαχθεί το συμπέρασμα ότι η χρήση επιπλέον γενικών χαρακτηριστικών ήχου αποτελεί βελτίωση της μεμονωμένης χρήσης των χαρακτηριστικών VGG.

Για την πλήρη σύγκριση με το πείραμα βάσης, πρέπει να ληφθεί υπόψη ότι όλες οι προηγούμενες περιπτώσεις αφορούσαν πειράματα όπου τα αποτελέσματα προβλέψεων ήταν σε επίπεδο φωνήματος. Ωστόσο, για να καταστούν συγκρίσιμα τα αποτελέσματα των βέλτιστων περιπτώσεων που εντοπίστηκαν με αυτά του πειράματος βάσης, πρέπει να εφαρμοστεί ένας κανόνας πλειοψηφίας για να εξαχθούν αποτελέσματα σε επίπεδο ασθενή. Συγκεκριμένα, για κάθε ασθενή, αν τα περισσότερα από τα φωνήματα που εξέτασε το μοντέλο προβλέπουν ότι έχει Covid, τότε η τελική πρόβλεψη για τον ασθενή θα είναι θετική, αλλιώς αρνητική. Με αυτόν τον τρόπο, οι μετρικές μπορούν να συγκριθούν άμεσα με τα αποτελέσματα του πειράματος βάσης, όπου η πρόβλεψη έγινε σε επίπεδο ασθενή.

Συγκρίνοντας τους πίνακες 5.16.β και 5.17.β με τον πίνακα 5.1, προκύπτει αδιαμφισβήτητα ότι τα φωνήματα α και i, διάρκειας 0.5 δευτερολέπτων, αποτελούν τους κύριους πρωταγωνιστές στην πρόβλεψη του Covid. Και στις δύο περιπτώσεις παρατηρείται σχετικά μικρή αλλά ικανοποιητική βελτίωση σε όλες τις μετρικές. Λαμβάνοντας υπόψη αυτή τη βελτίωση, καθώς και το γεγονός ότι τόσο με τη χρήση μόνο των χαρακτηριστικών VGG όσο και με τον συνδυασμό όλων των χαρακτηριστικών, τα βέλτιστα αποτελέσματα εμφανίστηκαν για διάρκεια 0.5 δευτερολέπτων – που ισούται με τη μέση χρονική διάρκεια ενός φωνήματος – μπορεί να δοθεί βάρος στο επιχείρημα ότι η απομόνωση συγκεκριμένων φωνημάτων αποτελεί αποτελεσματική προσέγγιση έναντι της χρήσης ολόκληρης της ηχογράφησης για την εκτίμηση του Covid.

Στη μελέτη Voice Features of Sustained Phoneme as COVID-19 Biomarker, που παρουσιάστηκε στην παράγραφο Related Work, με στόχο την διερεύνηση διαφόρων χαρακτηριστικών και φωνημάτων για την τελική αναγνώριση του COVID-19, κατέληξαν στο ότι, χρησιμοποιώντας το φώνημα i με ένα σύνολο 18 χαρακτηριστικών, προέκυψαν εξαιρετικά αποτελέσματα. Συγκεκριμένα, το μοντέλο SVM πέτυχε ακρίβεια 93.5%, F1 score 94.3%, ευαισθησία 96.7% και επιλεκτικότητα 89.6%.

Αφενός, στην παρούσα εργασία με τη χρήση μόνο των χαρακτηριστικών VGG, το φώνημα α αναδείχθηκε ως ο κύριος πρωταγωνιστής, ενώ το φώνημα i είχε χαμηλή απόδοση και κατατάχθηκε τελευταίο σε σχέση με τα άλλα φωνήματα. Ωστόσο, με την προσθήκη επιπλέον χαρακτηριστικών που βασίζονται κυρίως σε φασματική πληροφορία, η απόδοση του i βελτιώθηκε σημαντικά στην τελική πρόβλεψη του COVID-19, καθιστώντας το το νέο πρωταγωνιστή.

Αφετέρου, στη μελέτη Voice Features of Sustained Phoneme as COVID-19 Biomarker, παρά τις διαφοροποιήσεις που παρατηρήθηκαν στο φώνημα α, οι συγγραφείς σημείωσαν ότι, αν και το φώνημα αυτό εμφάνισε κάποια ικανότητα διάκρισης μεταξύ ασθενών και υγιών ατόμων, δεν κατάφερε να προσεγγίσει την απόδοση του φωνήματος i. Το φώνημα i αποδείχθηκε ότι πέτυχε τις καλύτερες επιδόσεις χρησιμοποιώντας 18 χαρακτηριστικά, τα οποία περιλάμβαναν MFCCs, formants και Vocal Tract Length (VTL), τα οποία έχουν συσχέτιση με τα χαρακτηριστικά που εξετάστηκαν στην παρούσα εργασία.

Αυτό ίσως να υποδεικνύει ότι χαρακτηριστικά που βασίζονται στη φασματική πληροφορία, τη διαμόρφωση του φωνητικού σωλήνα, καθώς και τη σταθερότητα των αναπνευστικών μυών, είναι κρίσιμα για την ανίχνευση αλλαγών στη φωνή που προκαλεί ο COVID-19.

Επίσης, η παρούσα εργασία καταδεικνύει σημαντική βελτίωση στην ανίχνευση COVID-19, δίνοντας έμφαση στη χρήση απομονωμένων φωνημάτων αντί για ολόκληρες ηχητικές καταγραφές, και η μελέτη Voice Features of Sustained Phoneme as COVID-19 Biomarker παρουσιάζει εξαιρετικές μετρικές για το φώνημα i. Έτσι λοιπόν, μπορεί συνδυαστικά να εξαχθεί το συμπέρασμα ότι η επιλογή των σωστών χαρακτηριστικών και η χρήση απομονωμένων φωνημάτων είναι καθοριστικής σημασίας για τη δημιουργία ενός ταξινομητή που θα προβλέπει τη μόλυνση από COVID-19.

Τέλος, από τη στιγμή που και οι δύο μελέτες έχουν ως κοινό πρωταγωνιστή το φώνημα i, αυτό ενισχύει την υπόθεση ότι σε συνδυασμό με την επιλογή χαρακτηριστικών που σχετίζονται με τη φασματική πληροφορία, τη διαμόρφωση του φωνητικού σωλήνα, καθώς και τη σταθερότητα των αναπνευστικών μυών, το φώνημα αυτό μπορεί να είναι το πιο αξιόπιστο για την ανίχνευση του COVID-19. Αυτό μπορεί να οφείλεται στο γεγονός ότι το φώνημα i απαιτεί ακριβή έλεγχο της γλώσσας και του φωνητικού σωλήνα, γεγονός που το καθιστά πιο ευαίσθητο στις αλλαγές που προκαλούνται από τη νόσο, όπως η φλεγμονή ή η πίεση στους μύες του φωνητικού συστήματος. Ωστόσο, απαιτείται περισσότερη έρευνα τόσο από την παρούσα εργασία όσο και από τη μελέτη για να επιβεβαιωθούν αυτά τα συμπεράσματα και να αναπτυχθούν αξιόπιστες διαγνωστικές μέθοδοι.

6.3 Μελλοντική δουλειά

Κομβική σημασία στην συγκεκριμένη εργασία αποτελούν 3 μοντέλα:

1. Μοντέλο παραγωγής χαρακτηριστικών από δεδομένα ομιλίας
2. Μοντέλο εξαγωγής φωνημάτων
3. Μοντέλο προσδιορισμού χρονικής διάρκειας φωνήματος

Το VGG χρησιμοποιήθηκε για να καλύψει την θέση του πρώτου μοντέλου όμως, είναι αρκετά παλιά αρχιτεκτονική εκπαιδευμένη σε δεδομένα τα οποία δεν εξειδικεύονται στην ομιλία. Επομένως είναι κρίσιμο να επαναληφθεί το πείραμα με την συγκεκριμένη μεθοδολογία με αντικατάσταση του VGG με κάποιο πιο σύγχρονό του, η προσαρμογή του (adaptation) σε δεδομένα ομιλίας και σε δεδομένα όπου υπάρχουν διαθέσιμες επισημειώσεις με φωνήματα και τέλος, ή μετεκπαίδευσή (fine tuning) του σε δεδομένα ομιλίας.

Κομβικής σημασίας είναι η διερεύνηση των βέλτιστων υπερπαραμέτρων του μοντέλου που οδηγεί στην τελική πρόβλεψη, καθώς αυτό μπορεί να βελτιώσει την ακρίβεια και την απόδοση της πρόβλεψης. Παράλληλα, μπορεί να εξεταστεί η επιλογή άλλων, πιο πολύπλοκων μοντέλων, εκτός από ένα απλό Multi Layer Perceptron.

Για τα μοντέλα 2 και 3 χρησιμοποιήθηκε το μοντέλο Allosaurus. Μετά από χειροκίνητο έλεγχο σχετικά με την ποιότητα των αποτελεσμάτων εύρεσης των φωνημάτων, εκτιμήθηκε και η δυνατότητά του να προσδιορίζει την χρονική στιγμή έναρξης του φωνήματος. Και οι δύο κρίθηκαν ικανοποιητικές από τον συγγραφέα. Σε κάθε περίπτωση όμως, δεν υπήρχε προσδιορισμός χρονικής στιγμής λήξης και επομένως, τμηματοποίηση φωνήματος. Επομένως θα ήταν αρκετά χρήσιμο να εκτιμηθεί η ποιότητα εξόδων από μοντέλα τμηματοποίησης ομιλίας έτσι ώστε να υπάρχει συγκεκριμένη διάρκεια και να αποφευχθεί ο τεχνητός περιορισμός της διάρκειας που εισήχθη από τον συγγραφέα στην παρούσα εργασία.

Επιπλέον, θα ήταν ωφέλιμο να χρησιμοποιηθούν περισσότερα από ένα φωνήματα για την τελική πρόβλεψη, καθώς η προσέγγιση της μάθησης ομάδας (ensemble learning) έχει δείξει θετικά αποτελέσματα. Ο συγγραφέας υποθέτει ότι το ίδιο μπορεί να εφαρμοστεί και σε επίπεδο φωνημάτων, δηλαδή η τελική πρόβλεψη να βασίζεται στην πλειοψηφία των προβλέψεων από πολλαπλά φωνήματα.

Αξίζει να εξεταστεί η περίπτωση αξιοποίησης των διαθέσιμων ιατρικών δεδομένων που περιλαμβάνονται στη συλλογή δεδομένων Smart4covid, όπως ηχογραφήσεις αναπνοής και βήχα. Στο σχετικό έργο (related work), υπάρχουν μελέτες που συνδυάζουν ομιλία, αναπνοή και βήχα, καταλήγοντας σε πολύ καλά αποτελέσματα για την τελική πρόβλεψη του COVID-19. Η ενσωμάτωση αυτών των δεδομένων μπορεί να ενισχύσει την ακρίβεια της πρόβλεψης. Ωστόσο, αυτή η προσέγγιση θα περιορίσει τον γενικό χαρακτήρα της μεθοδολογίας, καθώς ο αρχικός στόχος ήταν να αναλυθεί το κατά πόσο απομονωμένα φωνήματα μπορούν να βοηθήσουν στην αναγνώριση του COVID-19. Παρ' όλα αυτά, είναι ενδιαφέρον να μελετηθεί αυτή η εναλλακτική, καθώς μπορεί να οδηγήσει σε σημαντικά βελτιωμένα αποτελέσματα.

Τα δεδομένα που χρησιμοποιούνται είναι αφιλτράριστα και προέρχονται από αρκετά ανομοιογενείς πηγές. Κρίνεται αναγκαίο να γίνει ενδελεχής ακρόαση των παραπάνω και χειροκίνητη αφαίρεση λευκού θορύβου ή κενών ομιλίας και επανάληψη των πειραμάτων.

Λόγω της υψηλής διαφοράς μεταξύ των δειγμάτων με ασθενή και υγιή άτομα, θα μπορούσε το πρόβλημα να αναδιατυπωθεί ως πρόβλημα εκτίμησης ανωμαλίας (anomaly detection). Στην συγκεκριμένη κατηγορία, χρησιμοποιούνται τα υγιή δεδομένα με σκοπό την κατανομή των υγιών ατόμων και έπειτα, αν και εφόσον τα μη υγιή άτομα προβλεφθούν ότι βρίσκονται σε σχετικά μεγάλη απόσταση από την παραπάνω κατανομή, κρίνονται ως ανωμαλία.

Τέλος και λόγων της μη υπάρξης μεγάλου αποθετηρίου δεδομένων ομιλίας συσχετιζόμενων με την COVID, κρίνεται απαραίτητη η διερεύνηση χρήσης μεθόδων τεχνητής προσαύξησης των δεδομένων (data augmentation).

7. Αναφορές

[1] M. Moazzam *et al.*, “Understanding covid-19: From origin to potential therapeutics,” *Int J Environ Res Public Health*, vol. 17, no. 16, pp. 1–22, Aug. 2020, doi: 10.3390/ijerph17165904.

[2] C. O’Callaghan-Gordo and J. M. Antó, “COVID-19: The disease of the anthropocene,” Aug. 01, 2020, *Academic Press Inc.* doi: 10.1016/j.envres.2020.109683.

[3] X. Liu, C. Liu, G. Liu, W. Luo, and N. Xia, “COVID-19: Progress in diagnostics, therapy and vaccination,” 2020, *Ivyspring International Publisher*. doi: 10.7150/thno.47987.

[4] “Tone”.

[5] S. M. Jin, “소리의 정의(Definition of Sound) The Physics of Sound.”

[6] J. Herre and S. Dick, “Psychoacoustic models for perceptual audio coding-A tutorial review,” Jul. 01, 2019, *MDPI AG*. doi: 10.3390/app9142854.

[7] P. Rao, “Audio Signal Processing.”

[8] A. Ehab Sakran, S. Mahdy Abdou, S. Eldeen Hamid, and M. Rashwan, “International Journal of Computer Science and Mobile Computing A Review: Automatic Speech Segmentation,” 2017. [Online]. Available: www.ijcsmc.com

[9] M. Nilsson and M. Ejnarsson, “Speech Recognition using Hidden Markov Model performance evaluation in noisy environment,” 2002.

[10] “Modern Journal of Language Teaching Methods”, [Online]. Available: www.mjltm.com

[11] G. Smith, “The Fast Fourier Transform and its Applications,” 2019. [Online]. Available: https://github.com/gillian-smith/fft-project.

[12] R. Elmasian and M. H. Birnbaum, “A harmonious note on pitch: Scales of pitch derived from subtractive model of comparison agree with the musical scale,” 1984.

[13] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, “Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges,” Jun. 01, 2021, *Elsevier Ltd*. doi: 10.1016/j.eswa.2021.114591.

[14] V. Tiwari, “MFCC and its applications in speaker recognition.”

[15] W. Flores-Fuentes *et al.*, “Combined application of Power Spectrum Centroid and Support Vector Machines for measurement improvement in Optical Scanning Systems,” *Signal Processing*, vol. 98, pp. 37–51, 2014, doi: 10.1016/j.sigpro.2013.11.008.

[16] M. Feldman, “Hilbert transform in vibration analysis,” 2011, *Academic Press*. doi: 10.1016/j.ymssp.2010.07.018.

[17] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, “Music type classification by spectral contrast feature,” in *Proceedings - 2002 IEEE International Conference on Multimedia and Expo, ICME 2002*, Institute of Electrical and Electronics Engineers Inc., 2002, pp. 113–116. doi: 10.1109/ICME.2002.1035731.

[18] S. Dubnov, “Generalization of spectral flatness measure for non-Gaussian linear processes,” *IEEE Signal Process Lett*, vol. 11, no. 8, pp. 698–701, Aug. 2004, doi: 10.1109/LSP.2004.831663.

[19] Y. Yu *et al.*, “Spectral Roll-off Points Variations: Exploring Useful Information in Feature Maps by Its Variations,” Jan. 2021, [Online]. Available: http://arxiv.org/abs/2102.00369

[20] J. Liu *et al.*, “Artificial intelligence in the 21st century,” *IEEE Access*, vol. 6, pp. 34403–34421, Mar. 2018, doi: 10.1109/ACCESS.2018.2819688.

[21] I. El, N. R. Li, and M. J. Murphy, “Theory and Applications Machine Learning in Radiation Oncology.”

[22] S. Walczak and N. Cerpa, “Artificial Neural Networks,” 2001. [Online]. Available: http://www.emsl.pnl.gov:2080/proj/neuron/neural/sys-

[23] M. T. C. Olmedo, M. Paegelow, J. F. Mas, and F. Escobar, “Geomatic Approaches for Modeling Land Change Scenarios. An Introduction,” in *Lecture Notes in Geoinformation and Cartography*, Springer Science and Business Media Deutschland GmbH, 2018, pp. 1–8. doi: 10.1007/978-3-319-60801-3_1.

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

[25] T. Yu and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications," Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.05689

[26] F. Zhuang *et al.*, "A Comprehensive Survey on Transfer Learning," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.02685

[27] S. J. Pan and Q. Yang, "A survey on transfer learning," 2010. doi: 10.1109/TKDE.2009.191.

[28] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification."

[29] N. D. Pah, V. Indrawati, and D. K. Kumar, "Voice Features of Sustained Phoneme as COVID-19 Biomarker," *IEEE J Transl Eng Health Med*, vol. 10, 2022, doi: 10.1109/JTEHM.2022.3208057.

[30] P. Gupta and S. Degadwala, "A Comprehensive Review on COVID-19 Cough Audio Classification through Deep Learning," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 289–294, Nov. 2023, doi: 10.32628/cseit2361049.

[31] S. Ulukaya, A. A. Sarıca, O. Erdem, and A. Karaali, "MSCCov19Net: multi-branch deep learning model for COVID-19 detection from cough sounds," *Med Biol Eng Comput*, vol. 61, no. 7, pp. 1619–1629, Jul. 2023, doi: 10.1007/s11517-023-02803-4.

[32] S. Kim, J. Y. Baek, and S. P. Lee, "COVID-19 Detection Model with Acoustic Features from Cough Sound and Its Application," *Applied Sciences (Switzerland)*, vol. 13, no. 4, Feb. 2023, doi: 10.3390/app13042378.

[33] T. Hoang, L. Pham, D. Ngo, and H. D. Nguyen, "A Cough-based deep learning framework for detecting COVID-19," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 3422–3425. doi: 10.1109/EMBC48229.2022.9871179.

[34] S. S. Nayak, A. D. Darji, and P. K. Shah, "Machine learning approach for detecting Covid-19 from speech signal using Mel frequency magnitude coefficient," *Signal Image Video Process*, vol. 17, no. 6, pp. 3155–3162, Sep. 2023, doi: 10.1007/s11760-023-02537-8.

[35] D. Grant, I. McLane, and J. West, "Rapid and Scalable COVID-19 Screening using Speech, Breath, and Cough Recordings," in *BHI 2021 - 2021 IEEE EMBS International Conference on Biomedical and*

Health Informatics, Proceedings, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/BHI50953.2021.9508482.

[36] L. Verde, G. De Pietro, A. Ghoneim, M. Alrashoud, K. N. Al-Mutib, and G. Sannino, "Exploring the Use of Artificial Intelligence Techniques to Detect the Presence of Coronavirus Covid-19 through Speech and Voice Analysis," *IEEE Access*, vol. 9, pp. 65750–65757, 2021, doi: 10.1109/ACCESS.2021.3075571.

[37] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 Detection in Cough, Breath and Speech using Deep Transfer Learning and Bottleneck Features," Apr. 2021, doi: 10.1016/j.compbiomed.2021.105153.

[38] X. Li *et al.*, "Universal Phone Recognition with a Multilingual Allophone System," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.11800

[39] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[40] S. Hershey *et al.*, "CNN Architectures for Large-Scale Audio Classification," Sep. 2016, [Online]. Available: http://arxiv.org/abs/1609.09430

[41] K. Zarkogianni *et al.*, "The smarty4covid dataset and knowledge base as a framework for interpretable physiological audio data analysis," *Sci Data*, vol. 10, no. 1, Dec. 2023, doi: 10.1038/s41597-023-02646-6.