



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ  
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ  
ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

**«Τεχνικές μηχανικής μάθησης για τη διάγνωση ατόμων  
που πάσχουν από τον ιό SARS-CoV-2 με χρήση ηχητικών  
καταγραφών βήχα»**

**ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

Τσιάκα Ιωάννα

**Επιβλέπων :** Κωνσταντίνα Σ. Νικήτα

Καθ. Ε.Μ.Π.

Αθήνα, Σεπτέμβριος, 2021





ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ

ΜΗΧΑΝΙΚΩΝ

ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ

ΠΛΗΡΟΦΟΡΙΑΣ

ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

**«Τεχνικές μηχανικής μάθησης για τη διάγνωση ατόμων  
που πάσχουν από τον ιό  
SARS-CoV-2 με χρήση ηχητικών καταγραφών βήχα»**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Τσιάκα Ιωάννα

**Επιβλέπων :** Κωνσταντίνα Σ. Νικήτα

Καθ. Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 04/11/2021.

.....  
Κωνσταντίνα Σ. Νικήτα  
Καθ. Ε.Μ.Π.

.....

.....

Αθήνα, Σεπτέμβριος, 2021

.....  
Τσιάκα Ιωάννα

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

Copyright © Τσιάκα Ιωάννα, 2021.

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.



# Περίληψη

Η COVID-19 είναι μια μεταδοτική αναπνευστική ασθένεια που εξαπλώθηκε σε όλο τον κόσμο το 2020, επιφέροντας καταστροφικές συνέπειες. Η πανδημία COVID-19 έχει οδηγήσει σε δραματική απώλεια ανθρώπινων ζώων παγκοσμίως και αποτελεί μια άνευ προηγουμένου πρόκληση για τη δημόσια υγεία, την ψυχική υγεία και τον κόσμο της εργασίας. Η οικονομική και κοινωνική αναταραχή που προκαλείται από την πανδημία έφερε δεκάδες εκατομμύρια ανθρώπους αντιμέτωπους με τον κίνδυνο της φτώχειας. Επομένως, απαιτείται μια χαμηλού κόστους, γρήγορη και εύκολα διαθέσιμη λύση για την παροχή διάγνωσης COVID-19, με στόχο τον περιορισμό της. Σύμφωνα με πρόσφατες μελέτες, ένα από τα κύρια συμπτώματα της COVID-19 είναι ο βήχας. Στόχος, λοιπόν, της παρούσας διπλωματικής εργασίας αποτελεί η ανάπτυξη μεθόδου που βασίζεται σε τεχνικές μηχανικής μάθησης για την αυτόματη διάγνωση της COVID-19 μέσω ηχητικών καταγραφών βήχα.

Για το σκοπό αυτό αξιοποιήθηκαν βάσεις δεδομένων που περιλαμβάνουν άτομα από μεγάλο σύνολο χωρών και περιέχουν τόσο αναγκαστικό όσο και φυσικό βήχα. Το δημοσίως διαθέσιμο σύνολο δεδομένων Coswara περιέχει 341 θετικά για την COVID-19 και 1195 υγιή άτομα, ενώ το δεύτερο μικρότερο σύνολο δεδομένων συλλέχτηκε κυρίως στο Cambridge και περιέχει 61 θετικά για COVID-19 και 269 αρνητικά για τον COVID-19. Και στα δύο σύνολα δεδομένων παρατηρείται ανισορροπία ανάμεσα στα δείγματα ήχου υγιών ατόμων και ατόμων που πάσχουν από COVID-19 της τάξεως του 20% - 30%. Η ανισορροπία του συνόλου δεδομένων αντιμετωπίστηκε με εφαρμογή κατάλληλων τεχνικών για την εξισορρόπηση των κλάσεων. Για το σκοπό της αναζήτησης κατάλληλων υπερπαραμέτρων των μοντέλων και της επιλογής χαρακτηριστικών, χρησιμοποιήθηκε τυχαίος διαχωρισμός δεδομένων εκπαίδευσης – ελέγχου για επτά μοντέλα ταξινομητών: τυχαίο δάσος (RF), XGBoost (XGB), ταξινομητής ενίσχυσης κλίσης (GBC), k-πλησιέστερος γείτονας (KNN), μηχανή υποστήριξης διανυσμάτων (SVM), μακροχρόνια βραχυπρόθεσμη μνήμη (LSTM) και αμφίδρομη μακροπρόθεσμη μνήμη (Bi-LSTM). Όσον αφορά την επικύρωση της απόδοσης των μοντέλων, εφαρμόστηκαν επαναλαμβανόμενες τεχνικές k-fold και leave-one-out.

Τα αποτελέσματα δείχνουν ότι παρόλο που όλοι οι ταξινομητές μπόρεσαν να αναγνωρίσουν τον βήχα COVID-19, την καλύτερη απόδοση παρουσίασε ο ταξινομητής Bi-LSTM, ο οποίος ήταν σε θέση να διακρίνει μεταξύ του θετικού για την COVID-19 και του υγιούς βήχα με περιοχή κάτω από τη ROC καμπύλη (AUC) 0.70. Ένας ταξινομητής LSTM μπόρεσε να διακρίνει καλύτερα τον βήχα θετικό για την COVID-19 και τον αρνητικό με COVID-19, με AUC 0.68 μετά την επιλογή των καλύτερων 11 χαρακτηριστικών με την βοήθεια της feature importance συνάρτησης των ντετερμινιστικών μοντέλων μηχανικής μάθησης.

Λέξεις κλειδιά: Ανίχνευση, Βήχας, Ταξινόμηση, Covid-19, Μηχανική Μάθηση, Deep Learning, random forest (RF), XGBoost (XGB), Gradient Boosting Classifier (GBC), k-nearest neighbor (KNN), support vector machine (SVM), long short-term memory (LSTM) and Bidirectional long short-term memory (Bi-LSTM)

# Abstract

COVID-19 is a contagious respiratory disease that spread worldwide in 2020, with devastating consequences. The COVID-19 pandemic has led to dramatic loss of human life worldwide and poses an unprecedented challenge to public health, mental health, and social structure. The economic and social upheaval caused by the pandemic has put tens of millions of people at risk of poverty. Therefore, a low-cost, fast, and readily available solution is needed to provide diagnosis of COVID-19, with the aim of alleviating the pandemic consequences. According to recent studies, one of the main symptoms of COVID-19 is cough. The aim of this dissertation is to develop a method based on machine learning techniques for the automatic diagnosis of COVID-19 through cough audio recordings.

The classifiers presented here can discriminate COVID-19 positive coughs from both COVID-19 negative and healthy coughs recorded on a smartphone. This type of screening is non-contact, easy to apply, and can reduce the workload in testing centers as well as limit transmission by recommending early self-isolation to those who have a cough suggestive of COVID-19. The datasets used in this study include subjects from all six continents and contain both forced and natural coughs, indicating that the approach is widely applicable. The publicly available Coswara dataset contains 341 COVID-19 positive and 1195 healthy subjects, while the second smaller dataset was collected mostly in Cambridge and contains 61 COVID-19 positive and 269 COVID-19 negative. Both datasets indicate that COVID-19 positive coughs are 20%-30% shorter than non-COVID coughs. Dataset skew was addressed by applying techniques for balancing the datasets. A random train – test split was used to train and tune the parameters of seven machine learning classifiers: random forest (RF), XGBoost (XGB), Gradient Boosting Classifier (GBC), k-nearest neighbor (KNN), support vector machine (SVM), long short-term memory (LSTM) and Bidirectional long short-term memory (Bi-LSTM). As for the validation of models' performance, repeated k-fold and leave-one-out techniques applied.

The results show that although all classifiers were able to identify COVID-19 coughs, the best performance was exhibited by the Bi-LSTM classifier, which was best able to discriminate between the COVID-19 positive and the healthy coughs with an area under the ROC curve (AUC) of 0.70. An LSTM classifier was best able to discriminate between the COVID-19 positive and COVID-19 negative coughs, with an AUC of 0.80 after selecting the best 11 features from feature importance selection. Since this type of cough audio classification is cost-effective and easy to deploy, it is potentially a useful and viable means of non-contact COVID-19 screening.

Key words: Detection, Cough, Classification, Covid-19, LSTM, Machine Learning, Deep Learning, Random forest (RF), XGBoost (XGB), Gradient Boosting Classifier (GBC), k-nearest neighbor (KNN), support vector machine (SVM), long short-term memory (LSTM) and Bidirectional long short-term memory (Bi-LSTM)

# ΠΕΡΙΕΧΟΜΕΝΑ

<b>Περίληψη.....</b>	<b>6</b>
<b>Abstract.....</b>	<b>7</b>
<b>Εισαγωγή.....</b>	<b>10</b>
<b>Κεφάλαιο 1<sup>ο</sup>: Κορονοϊός.....</b>	<b>11</b>
1.1 Εισαγωγή.....	11
1.2 Εξάπλωση.....	12
1.3 Δημογραφικά στοιχεία.....	13
1.4 Θνησιμότητα – Καταπολέμηση.....	14
1.5 Συμπτώματα.....	15
1.6 Κινητή υγεία και Covid-19.....	16
1.6.1 Εφαρμογές για συμπτώματα κορονοϊού.....	16
1.6.2 Εφαρμογές ιχνηλάτισης επαφών για τον κορονοϊό.....	17
1.6.3 Εφαρμογές επιδημιολογίας του κορονοϊού.....	17
1.6.4 Ερευνητικές εφαρμογές για τον κορονοϊό.....	18
1.6.5 Εφαρμογές κοινωνικής απόστασης.....	18
1.6.6 Εφαρμογές τηλεϊατρικής.....	18
<b>Κεφάλαιο 2<sup>ο</sup>: Μοντέλα.....</b>	<b>19</b>
2.1 Δέντρα Αποφάσεων.....	19
2.1.1 Random Forest Classifier.....	20
2.1.2 Gradient Boosting Classifier - eXtreme Gradient Boosting Classifier.....	20
2.1.3 Πλεονεκτήματα [17].....	20
2.1.4 Μειονεκτήματα [17].....	20
2.2 Μηχανή διανυσματικής υποστήριξης.....	21
2.3 K Κοντινότεροι γείτονες.....	21
2.4 Επαναλαμβανόμενο Νευρωνικό Δίκτυο.....	21
2.4.1 Long Short Term Memory (LSTM).....	22
2.4.1.1 Βασική ιδέα.....	22
2.4.1.2 Forget gate.....	23
2.4.1.3 Πύλη εισόδου.....	23
2.4.1.4 Κατάσταση κελιού.....	23
2.4.1.5 Πύλη εξόδου.....	23
2.4.2 Bidirectional Long Short Term Memory.....	23
2.5 Πλεονεκτήματα (Bi)LSTM.....	24
2.6 Μειονεκτήματα (Bi-)LSTM.....	25
<b>Κεφάλαιο 3<sup>ο</sup>: Ανάλυση - Επεξεργασία Δεδομένων – Μεθοδολογία.....</b>	<b>26</b>
3.1 Δεδομένα.....	26
3.1.1 Coswara δεδομένα.....	26
3.1.2 Cambridge δεδομένα.....	26
3.2 Εξαγωγή Χαρακτηριστικών.....	26
3.2.1 Mel-frequency cepstral coefficients (MFCCs).....	27
3.2.2 Μηδενικός ρυθμός διασταύρωσης.....	27



3.2.3 Επίπεδο ενέργειας.....	27
3.2.4 Φασματικό κέντρο.....	27
3.2.5 Φασματική μετατόπιση.....	28
3.2.6 Φασματικό εύρος ζώνης.....	28
3.2.7 Φασματική επιπεδότητα.....	28
3.3 Οπτικοποίηση Δεδομένων.....	28
3.4 Μη ισορροπημένα δεδομένα.....	32
3.4.1 Υπερδειγματοληψία.....	33
3.4.1.1 Adasyn.....	33
3.4.1.2 Smote.....	34
3.4.1.3 SMOTE + Tomek Links.....	35
3.4.1.4 SMOTE + ENN.....	35
3.4.2 Υποδειγματοληψία.....	35
3.4.2.1 Τυχαία Υποδειγματοληψία.....	35
3.4.2.2 Near Miss.....	36
3.5 Κανονικοποίηση Δεδομένων.....	36
3.5.1 Τυποποίηση.....	36
3.5.2 Ομαλοποίηση.....	36
3.5.3 Επιλογή κανονικοποίησης.....	37
3.6 Επικύρωση των μοντέλων.....	37
3.6.1 K-fold Cross Validation.....	38
3.6.2 Ensemble Learning.....	38
3.6.3 Leave – one – out.....	38
<b>Κεφάλαιο 4<sup>ο</sup>: Αποτελέσματα.....</b>	<b>39</b>
Κεφάλαιο 4.1 Κανονικοποίηση Δεδομένων.....	39
Κεφάλαιο 4.2 Βελτιστοποίηση Παραμέτρων.....	40
Κεφάλαιο 4.3 Εξισορρόπηση Δεδομένων.....	41
Κεφάλαιο 4.4 Επιλογή Καλύτερων Χαρακτηριστικών.....	43
Κεφάλαιο 4.5 Επικύρωση Αποτελεσμάτων.....	44
Κεφάλαιο 4.6 Σύγκριση με υπάρχουσα βιβλιογραφία.....	48
Κεφάλαιο 4.7 Γενίκευση σε άλλη βάση δεδομένων.....	49
<b>Κεφάλαιο 5<sup>ο</sup>: Συζήτηση αποτελεσμάτων και συμπεράσματα.....</b>	<b>50</b>
<b>Κεφάλαιο 6<sup>ο</sup>: Επίλογος-Προτάσεις για μελλοντική έρευνα.....</b>	<b>54</b>
6.1 Επίλογος.....	54
6.2 Προβληματισμοί.....	55
6.3 Μελλοντικά Βήματα.....	56
<b>Βιβλιογραφία.....</b>	<b>57</b>

## Εισαγωγή

Στο πλαίσιο της παρούσας διπλωματικής εργασίας προτείνεται η χρήση νευρωνικών δικτύων αξιοποιώντας ηχητικές καταγραφές βήχα με σκοπό την ανίχνευση παρουσίας του ιού COVID-19. Ο ιός COVID-19 έχει λάβει διαστάσεις πανδημίας, πλήττοντας ένα μεγάλο πλήθος ανθρώπων παγκοσμίως. Μεταδίδεται από άνθρωπο σε άνθρωπο κυρίως μέσω του αναπνευστικού με σταγονίδια από το φτέρνισμα, το βήχα ή την ομιλία. Επίσης, μπορεί να μεταδοθεί μέσω επαφής με επιφάνειες πρόσφατα μολυσμένες με σταγονίδια, αν δεν τηρούνται τα μέτρα υγιεινής των χεριών. Λόγω των καταστροφικών επιπτώσεων του COVID-19 και της τραγικής απώλειας ζωών, είναι εξαιρετικά επείγον να αναπτυχθούν μέθοδοι για την έγκαιρη ανίχνευση της νόσου που μπορεί να συμβάλουν στον περιορισμό της εξάπλωσής της καθώς και στην ανάπτυξη στοχευμένων λύσεων. Ο βήχας και άλλοι φωνητικοί ήχοι περιέχουν πληροφορίες για την πνευμονική υγεία που μπορούν να χρησιμοποιηθούν για διαγνωστικούς σκοπούς.

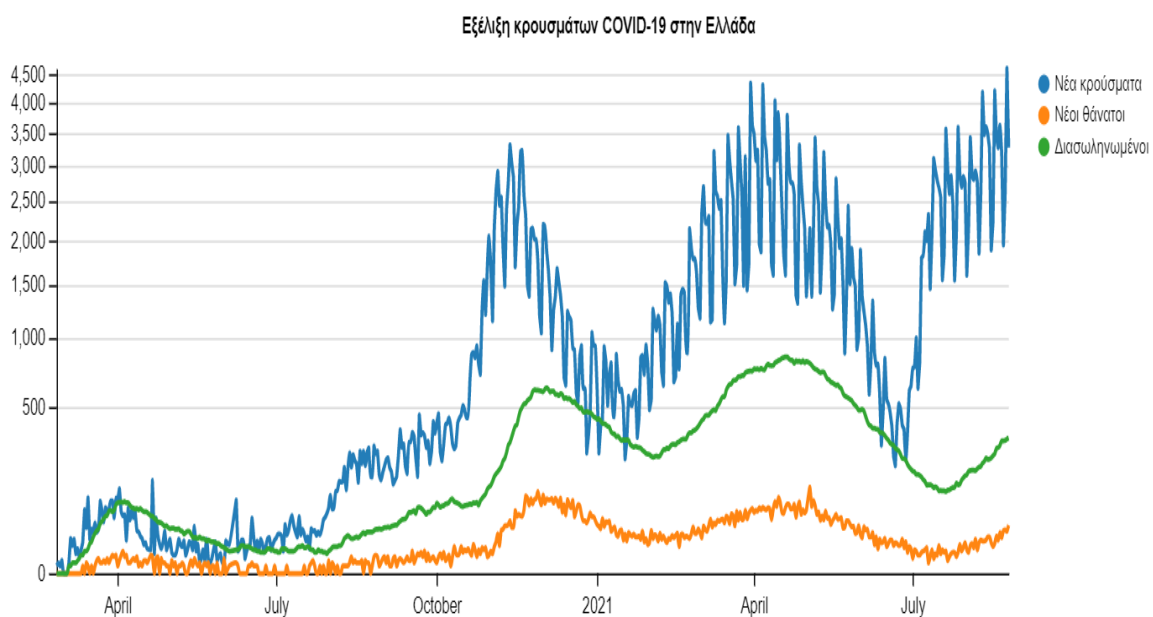
Σκοπός της εργασίας είναι η σχεδίαση και η υλοποίηση ενός νευρωνικού μοντέλου, ικανού να διαχωρίσει θετικούς στον ιό από τους αρνητικούς. Κατά την ανάπτυξη του νευρωνικού μοντέλου, αντιμετωπίστηκαν προκλήσεις όπως η έλλειψη δεδομένων, η ανισορροπία τους ακόμα και ο τρόπος συλλογής τους. Το γεγονός, όμως, ότι πρόκειται για ένα φλέγον ζήτημα συντέλεσε στην άμεση συλλογή δεδομένων, εξαλείφοντας με αυτόν τον τρόπο την πρώτη από τις προκλήσεις. Όσον αφορά τα μη ισορροπημένα δεδομένα, κρίθηκε απαραίτητη η επεξεργασία τους και η εφαρμογή κατάλληλων μεθόδων εξισορρόπησης. Οι μέθοδοι αυτοί θα αναλυθούν και θα επεξηγηθούν πλήρως στο κεφάλαιο 3.

Αναφορικά με τη δομή της εργασίας: αρχικά παραδίνεται μια συνοπτική περιγραφή του κορονοϊού που αφορά γενικά στοιχεία και βασικές πληροφορίες για τον ιό. Στην συνέχεια, παρουσιάζεται η αξιοποίηση της κινητής υγείας στην καταπολέμηση των επιπτώσεων του ιού, με παραδείγματα εφαρμογών και αναφορά στον σκοπό δημιουργίας τους. Ακολούθως, δίνεται μια ενδελεχή ανάλυση των μοντέλων που συντέλεσαν στην επίτευξη του στόχου. Η περιγραφή της σχεδίασης του μοντέλου ολοκληρώνεται με την παρουσίαση τεχνικών που αξιοποιήθηκαν για την επεξεργασία των δεδομένων, την εξαγωγή κατάλληλων χαρακτηριστικών, την επιλογή κυριότερων χαρακτηριστικών και την επικύρωση των μοντέλων. Τέλος, παρατίθενται πίνακες με τα αποτελέσματα που αποτελούν την βάση για την σύγκριση των μεθόδων που θα ακολουθήσει.

# Κεφάλαιο 1<sup>ο</sup>: Κορονοϊός

## 1.1 Εισαγωγή

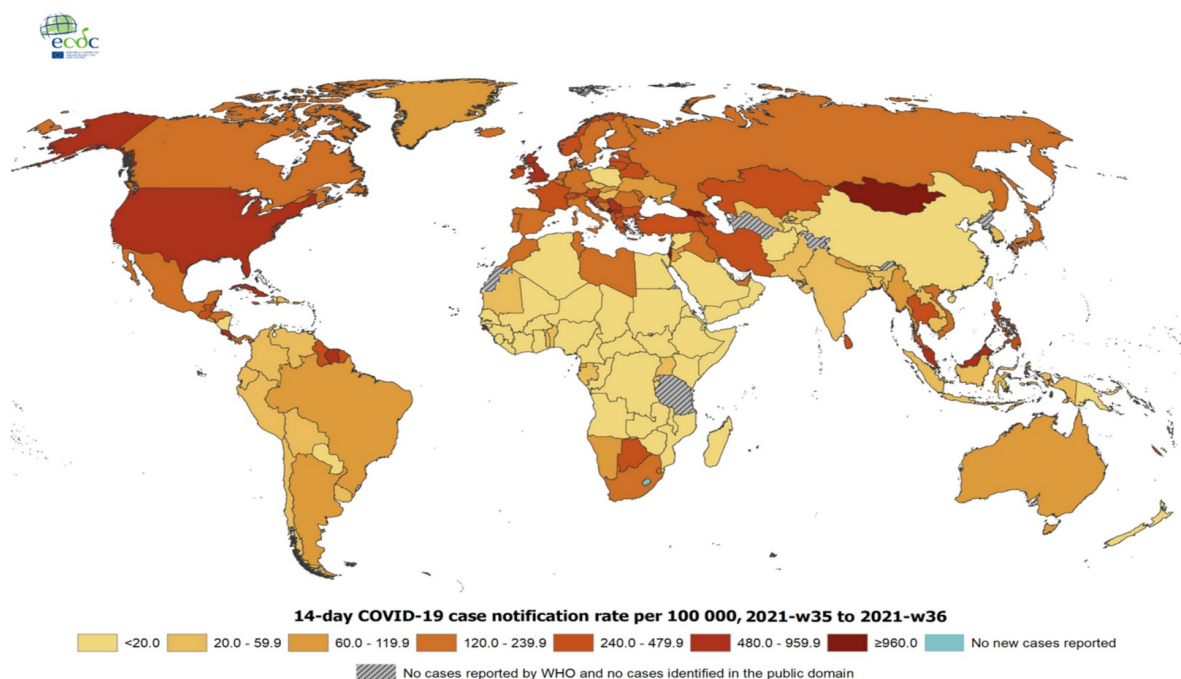
Οι κορονοϊοί είναι μια μεγάλη οικογένεια μονόκλωνων ιών RNA που μολύνουν τους ανθρώπους, αλλά και ένα ευρύ φάσμα ζώων. Διακρίνονται σε τέσσερις υποοικογένειες, συγκεκριμένα οι κορονοϊοί άλφα, βήτα, γάμμα και δέλτα. Οι δύο πρώτοι προέρχονται προφανώς από θηλαστικά, ιδίως από νυχτερίδες, ενώ οι ιοί γάμμα και δέλτα προέρχονται από χοίρους και πτηνά. Μεταξύ των υποτύπων των κορονοϊών που μπορούν να μολύνουν τους ανθρώπους, ο παράγοντας του κινδύνου ποικίλλει, καθώς επιφέρουν λοιμώξεις του αναπνευστικού συστήματος που κυμαίνονται από ήπιες έως θανατηφόρες [1]. Οι ήπιες ασθένειες στον άνθρωπο περιλαμβάνουν ορισμένες περιπτώσεις κοινού κρυολογήματος, ενώ οι θανατηφόρες περιπτώσεις προκαλούνται από τις λοιμώξεις σοβαρού οξέος αναπνευστικού συνδρόμου – SARS (ΣΟΑΣ), η οποία άρχισε το 2002 στην Κίνα και τον αναπνευστικό σύνδρομο της Μέσης Ανατολής το 2012 – MERS (ΑΣΜΑ) και μπορούν να σκοτώσουν περισσότερο από το 30% των ανθρώπων που έχουν μολυνθεί. Συγκεκριμένα, η πανδημία της νόσου του κορονοϊού 2019 (COVID-19) είναι μια τρέχουσα πανδημία που προκλήθηκε από κορονοϊό σοβαρού οξέος αναπνευστικού συνδρόμου τύπου 2 (SARS-CoV-2). Το SARS-CoV-2 αποτελεί ένα νέο στέλεχος κορονοϊού το οποίο δεν είχε ταυτοποιηθεί σε ανθρώπους μέχρι τον Δεκέμβριο του 2019, που ανιχνεύθηκε κατά την πρόσφατη επιδημία στην πόλη Ουχάν, πρωτεύουσα της επαρχίας Χουμπέι της Λαϊκής Δημοκρατίας της Κίνας. Στις 27 Δεκεμβρίου 2019, η επιδημία του νέου κορονοϊού εντοπίστηκε ως ένα αναδυόμενο σύμπλεγμα ανθρώπων με πνευμονία άγνωστης αιτιολογίας [2]. Μέσα σε μια εβδομάδα έγινε σαφές ότι οι πρώτες περιπτώσεις σχετίζονται με την αγορά θαλασσιών της Ουχάν, όπου πωλούνταν επίσης ζωντανά πουλερικά και άγρια ζώα και προφανώς ο SARS - CoV - 2 κατάφερε να κάνει τη μετάβασή του από ζώα σε ανθρώπους. Στις 8 Ιανουαρίου, αναγνωρίστηκε ένας νέος βήτα - κορονοϊός ως η αιτία της πνευμονίας και μέχρι τις 29 Ιανουαρίου, ο ιός είχε εξαπλωθεί σε όλες τις επαρχίες της ηπειρωτικής Κίνας [3] Σφάλμα: Δεν βρέθηκε η πηγή παραπομπής. Στις 30 Ιανουαρίου 2020, η Επιτροπή Έκτακτης Ανάγκης του ΠΟΥ κήρυξε παγκόσμια κατάσταση έκτακτης ανάγκης για την υγεία με βάση τα αυξανόμενα ποσοστά ειδοποίησης περιστατικών σε κινεζικές και διεθνείς τοποθεσίες.



Εικόνα 1: Εξέλιξη κρουσμάτων COVID-19 στην Ελλάδα [1]

## 1.2 Εξάπλωση

Κατά τα πρώτα στάδια, ο αριθμός των περιστατικών διπλασιαζόταν συνήθως κάθε 7,5 ημέρες. Πρόσφατες έρευνες σε 425 επιβεβαιωμένα κρούσματα καταδεικνύουν ότι η τρέχουσα επιδημία μπορεί να διπλασιάζεται στον αριθμό των προσβεβλημένων ατόμων κάθε επτά ημέρες και ότι κάθε ασθενής μεταδίδει λοίμωξη σε 2,2 άλλα άτομα κατά μέσο όρο. Στις αρχές και στα μέσα Ιανουαρίου του 2020, ο ιός εξαπλώθηκε σε διάφορες κινεζικές επαρχίες, με τη βοήθεια των μετακινήσεων της Κινεζικής Πρωτοχρονιάς. Ο ΠΟΥ, μάλιστα, είχε εντοπίσει 13 χώρες κορυφαίας προτεραιότητας (Αλγερία, Αγκόλα, Ακτή Ελεφαντοστού, Λαϊκή Δημοκρατία του Κονγκό, Αιθιοπία, Γκάνα, Κένυα, Μαυρίκιος, Νιγηρία, Νότια Αφρική, Τανζανία, Ουγκάντα, Ζάμπια) οι οποίες είτε διατηρούν άμεσες συνδέσεις με Κίνα ή πραγματοποιούν μεγάλο όγκο ταξιδιών στην Κίνα [1]. Οι πυκνές κοινότητες εξαρχής κινδύνευαν περισσότερο και η πιο ευάλωτη περιοχή ήταν σίγουρα η Αφρική, λόγω της πυκνής κίνησης μεταξύ Κίνας και Αφρικής. Πράγματι, ο ιός επηρέασε σύντομα την Νότια Αφρική. Πολύ λίγες αφρικανικές χώρες έχουν επαρκείς και κατάλληλες διαγνωστικές ικανότητες και υπάρχουν προφανείς προκλήσεις για τον χειρισμό τέτοιων εστιών. Μέσα σε λιγότερο από ένα μήνα η COVID-19 είχε εξαπλωθεί εκτός από όλη την Κίνα και τις γειτονικές της χώρες, ακόμη και στις ΗΠΑ και την Ευρώπη. Έγινε, λοιπόν, σαφές ότι ο νέος ιός ήταν πολύ μεταδοτικός από άτομο σε άτομο αλλά ήταν σημαντικά λιγότερο λοιμογόνος, με λιγότερο από το 20% των περιπτώσεων να χαρακτηρίζονται ως σοβαροί. Η ετοιμότητα αντιμετώπισης των κρουσμάτων είναι αδύναμη σε πολλές χώρες. Από τις 45 χώρες χαμηλού εισοδήματος που έχουν πραγματοποιήσει εθνική αξιολόγηση ετοιμότητας, καμία δεν κρίθηκε έτοιμη να ανταποκριθεί, καθιστώντας τις ιδιαίτερα ευάλωτες σε επιδημίες. Υπάρχουν πολλοί λόγοι για αυτό, συμπεριλαμβανομένης της κακής υγείας και διατροφής, που επιδεινώνονται από τα υψηλά ποσοστά ταυτόχρονης ανθρώπινης ανοσοανεπάρκειας στον ιό. Δεδομένων των εύθραυστων συστημάτων υγείας στις περισσότερες αναπτυσσόμενες χώρες, νέες και αναδυόμενες εστίες ασθενειών, όπως η τρέχουσα επιδημία COVID-19, μπορούν να παραλύσουν τα συστήματα υγείας εις βάρος των πρωτογενών απαιτήσεων υγειονομικής περίθαλψης [1].

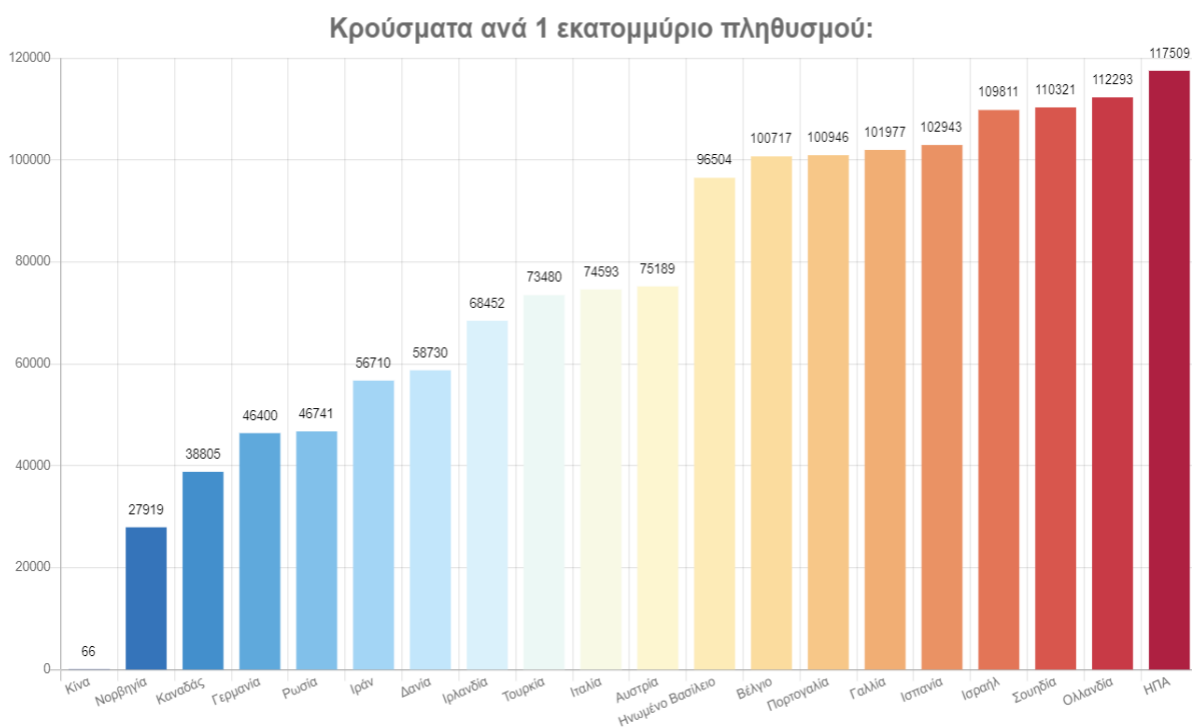


Administrative boundaries: © EuroGeographics © UN-FAO © Turkstat. The boundaries and names shown on this map do not imply official endorsement or acceptance by the European Union. Date of production: 15/09/2021

Εικόνα 2: Γεωγραφική εξάπλωση κρουσμάτων COVID-19

### 1.3 Δημογραφικά στοιχεία

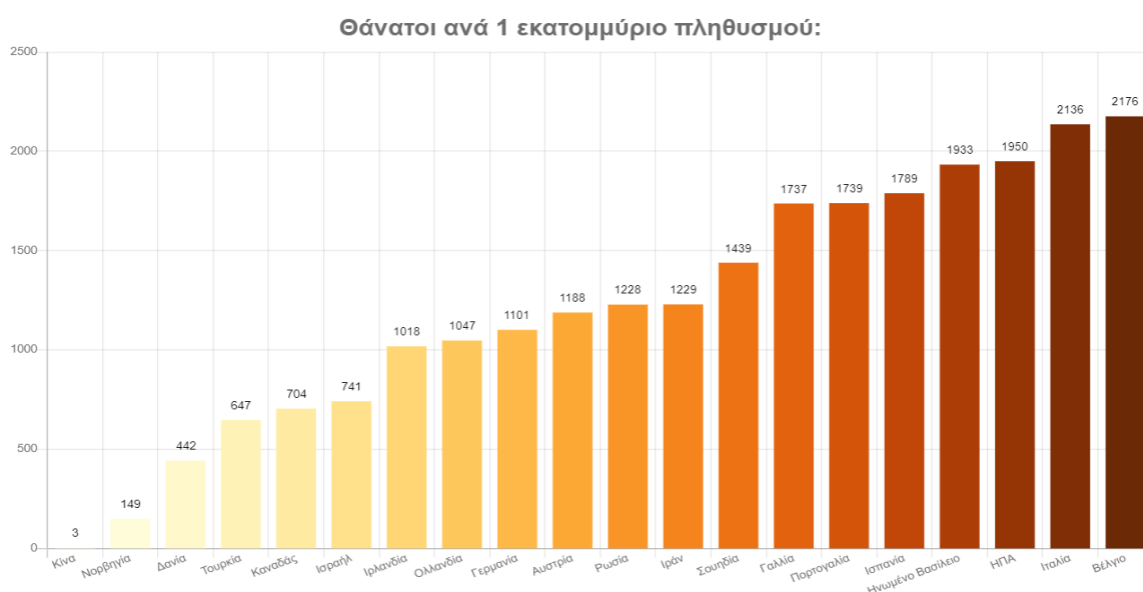
Η πανδημία COVID-19 άλλαξε την Ευρώπη και τον κόσμο εν ριπή οφθαλμού. Ως την 12<sup>η</sup> Αυγούστου 2021, περισσότερα από 205 εκατομμύρια κρούσματα έχουν επιβεβαιωθεί σε όλο τον κόσμο. Θεωρείται, βέβαια, ότι υπάρχει μια ελλιπής αναφορά των κρουσμάτων, ιδιαίτερα σε περιπτώσεις με ηπιότερα συμπτώματα. Το ποσοστό ανίχνευσης περιπτώσεων αλλάζει καθημερινά και μπορεί να παρακολουθείται σχεδόν σε πραγματικό χρόνο στον ιστότοπο που παρέχεται από το Πανεπιστήμιο Johns Hopkins [4], που έχει δημιουργήσει και ανανεώνει καθημερινά ένα ανοικτό αποθετήριο δεδομένων με διεθνή αναλυτικά στοιχεία σχετικά με την πανδημία του SARS-CoV-2. Το αποθετήριο αυτό αποτελεί την κύρια πηγή στατιστικών στοιχείων για όλες τις αναλύσεις σχετικά με την πανδημία που πραγματοποιούνται παγκοσμίως. Μια εναλλακτική πηγή διεθνών στατιστικών για επιβεβαιωμένα κρούσματα, βαριές περιπτώσεις (νοσηλίες σε Μ.Ε.Θ.), αναρρώσεις και θανάτους, η οποία μάλιστα ανανεώνεται σε συνεχή χρόνο, είναι αυτή του ιστότοπου **Worldometer** [5]. Ως σήμερα, έχουν επιβεβαιωθεί κρούσματα σε 215 χώρες και περιοχές, έχουν σημειωθεί περισσότεροι από 2.51 εκατομμύρια θάνατοι που οφείλονται στη νόσο και έχουν ανακάμψει περισσότεροι από 89 εκατομμύρια άνθρωποι. Ενώ, στη χώρα μας τα κρούσματα έως σήμερα ανέρχονται σε περίπου 520 χιλιάδες και οι θάνατοι ξεπέρασαν τους 13 χιλιάδες.



Εικόνα 3: Κρούσματα ανά 1 εκατομμύριο πληθυσμού [5]

## 1.4 Θνησιμότητα – Καταπολέμηση

Κάθε χρόνο, η παρακολούθηση των νέων τάσεων υγείας μπορεί να είναι δύσκολη. Αλλά φέτος, η δυναμική φύση της νόσου του κορονοϊού 2019 (COVID-19) έχει καταστήσει σχεδόν αδύνατη τη σαφήνεια. Οι αθροιστικές μολύνσεις ανήλθαν σε περισσότερες από 36 εκατομμύρια στις Ηνωμένες Πολιτείες, ενώ οι θάνατοι ξεπερνούν τους 618 χιλιάδες. Εν τω μεταξύ, τα μέσα ενημέρωσης αναφέρουν τακτικά μια πλημμύρα στατιστικών που μπορεί να είναι δύσκολο να πλαισιωθεί. Η ανάλυση των SH Woolf θέτει αυτούς τους αριθμούς σε προοπτική. Κατατάσσοντας την άλλοτε άγνωστη απειλή της COVID-19 για πιο οικείες ασθένειες, συγκρίνουν τη θνησιμότητα των ΗΠΑ από την COVID-19 (Μάρτιος-Οκτώβριος 2020) με τις κύριες αιτίες θανάτου 2 χρόνια πριν από την πανδημία (Μάρτιος-Οκτώβριος 2018). Η ανάλυση, σε συνδυασμό με την προηγούμενη έρευνά τους για τις τάσεις του θανάτου και του προσδόκιμου ζωής, προτείνουν ορισμένα σημαντικά συμπεράσματα. Αρχικά, ο COVID-19 κατατάσσεται ως η κύρια αιτία θανάτου. Σε σύγκριση με τις κύριες αιτίες θανάτου από την ίδια περίοδο το 2018, τη νέα COVID-19 ήταν η τρίτη κύρια αιτία θανάτου για παιδιά και ενήλικες, κατατάσσοντας τον κορονοϊό μόνο πίσω από τις καρδιακές παθήσεις και τον καρκίνο. Δεν έχει γλιτώσει καμία ηλικιακή ομάδα, αν και η COVID-19 δεν ήταν η κύρια αιτία θανάτου για τις νεότερες ηλικιακές ομάδες. Αυτά τα στοιχεία πιθανώς υποτιμούν την πραγματική υπερβολική θνησιμότητα κατά τουλάχιστον 20%, λόγω, εν μέρει, των έμμεσων επιπτώσεων της πανδημίας σε θανάτους από COVID-19, συμπεριλαμβανομένου του κινδύνου θανάτου που θέτουν σε άλλους οι θανατηφόρες μεταδοτικές ασθένειες. Ακόμη, ο κίνδυνος θανάτου από την COVID-19 είναι υψηλότερος μεταξύ των παλαιότερων και χαμηλότερος μεταξύ των νεότερων πληθυσμών. Η ηλικία κάνει τη σημαντική διαφορά στον κίνδυνο θνησιμότητας. Σε σύγκριση με άτομα ηλικίας 18 έως 29 ετών, τα άτομα ηλικίας μεταξύ 75 και 84 ετών και εκείνα των 85 ετών και άνω έχουν 200 φορές και 630 φορές μεγαλύτερο μέσο όρο θανάτων, αντίστοιχα. Αντίθετα, για άτομα κάτω των 35 ετών, ο Woolf διαπίστωσε ότι οι θάνατοι από υπερβολική δόση ναρκωτικών, ατυχήματα μεταφοράς (π.χ. θάνατοι από αυτοκινητιστικά οχήματα) και αυτοκτονία ξεπέρασαν τους θανάτους από COVID-19. Για τα παιδιά και τους νέους, ο σχετικός κίνδυνος θανάτου από COVID-19 είναι ακόμα χαμηλότερος [6].



Εικόνα 4: Θάνατοι ανά 1 εκατομμύριο πληθυσμού [5]

## 1.5 Συμπτώματα

Ο 2019-nCoV μεταδίδεται κυρίως μέσω της επαφής με ασθενή, μέσω σταγονιδίων τα οποία παράγονται όταν ο ασθενής βήχει ή φτερνίζεται ή μέσω σταγονιδίων από τη σίελο ή από τη ρινική κοιλότητα. Το αρχικό κλινικό σημάδι της νόσου που σχετίζεται με τον SARS - CoV - 2, το οποίο επέτρεψε την ανίχνευση περιστατικών ήταν η πνευμονία. Έχει, λοιπόν, τα κλινικά χαρακτηριστικά μιας άτυπης πνευμονίας με πυρετό, ξηρό βήχα, κόπωση, δύσπνοια και μυαλγία και είναι πιο συχνά σοβαρή σε άτομα με συννοσηρότητες και ηλικιωμένους. Πιο πρόσφατες αναφορές περιγράφουν επίσης γαστρεντερικά συμπτώματα και ασυμπτωματικές λοιμώξεις, ιδιαίτερα σε μικρά παιδιά. Ο χρόνος μεταξύ της έκθεσης και της εμφάνισης συμπτωμάτων είναι συνήθως από 2 έως 14 ημέρες. Η αναλογία ατόμων που έχουν μολυνθεί από τον SARS - CoV - 2 και παραμένουν ασυμπτωματικά καθ' όλη τη διάρκεια της λοίμωξης δεν έχει ακόμη αξιολογηθεί οριστικά. Σε συμπτωματικούς ασθενείς, οι κλινικές εκδηλώσεις της νόσου ξεκινούν συνήθως μετά από λιγότερο από μία εβδομάδα, που περιλαμβάνουν πυρετό, βήχα, ρινική συμφόρηση, κόπωση και άλλα σημάδια λοιμώξεων του ανώτερου αναπνευστικού συστήματος [7]. Η μόλυνση μπορεί να εξελιχθεί σε σοβαρή νόσο με δύσπνοια και σοβαρά συμπτώματα στο στήθος που αντιστοιχούν σε πνευμονία σε περίπου 75% των ασθενών. Η πνευμονία εμφανίζεται κυρίως τη δεύτερη ή την τρίτη εβδομάδα συμπτωματικής λοίμωξης. Τα κυριότερα σημάδια της ιογενούς πνευμονίας περιλαμβάνουν μειωμένο κορεσμό οξυγόνου, αποκλίσεις αερίων αίματος, αλλαγές ορατές μέσω των ακτινογραφιών θώρακα και άλλες τεχνικές απεικόνισης, αποσπασματική συμπύκνωση, κυψελιδικά εξιδρώματα και μεσοκοιλιακή προσβολή, υποδηλώνοντας τελικά επιδείνωση. Η λεμφοπενία φαίνεται να είναι συχνή και οι φλεγμονώδεις δείκτες (C -αντιδρώσα πρωτεΐνη και προφλεγμονώδεις κυτοκίνες) είναι αυξημένοι. Μελέτες δείχνουν ότι οι ασθενείς ηλικίας  $\geq 60$  ετών διατρέχουν μεγαλύτερο κίνδυνο από τα παιδιά που ενδέχεται να έχουν λιγότερες πιθανότητες να μολυνθούν ή, εάν ναι, μπορεί να εμφανίσουν πιο ήπια συμπτώματα ή ακόμη και ασυμπτωματική λοίμωξη [7].

Ορισμένοι μολυσμένοι ασθενείς εμφάνισαν μόνο νευρολογικά συμπτώματα ως αρχικά συμπτώματα, όπως τα ακόλουθα: πονοκέφαλος, ατονία, ασταθές περπάτημα και κακουχία, που μπορεί να οφείλονται σε μη ειδικές εκδηλώσεις που προκαλούνται από τον COVID-19 (το ποσοστό μη ειδικών εκδηλώσεων καθώς τα πρώτα συμπτώματα πρέπει να διερευνηθούν περαιτέρω). εγκεφαλική αιμορραγία. εγκεφαλικό έμφραγμα. και άλλες νευρολογικές παθήσεις. Σε πρόσφατη μελέτη 214 ασθενών με COVID-19, 78 (36,4%) ασθενείς είχαν νευρολογικές εκδηλώσεις, όπως πονοκέφαλο, ζάλη, οξείες εγκεφαλοαγγειακές παθήσεις και διαταραχή της συνείδησης. Από αυτούς τους 214 ασθενείς, 40 (18,7%) ασθενείς χρειάστηκαν εντατική θεραπεία παρεμβάσεις στη μονάδα φροντίδας (MEΘ) για τη σοβαρή νευρολογική εμπλοκή τους. Δεδομένου ότι η έκφραση του ACE2 σε ασθενείς με υπέρταση είναι ήδη χαμηλή, η μόλυνση από SARS-CoV-2 είναι πιο πιθανό να προκαλέσει εγκεφαλική αιμορραγία σε τέτοιους ασθενείς. Ως δεύτερη σειρά αποδεικτικών στοιχείων που υποδηλώνουν ότι η μόλυνση SARS-CoV-2 μπορεί να προκαλέσει εγκεφαλική αιμορραγία, οι ασθενείς με COVID-19 υποφέρουν συχνά από πήξη και παρατεταμένο χρόνο προθρομβίνης, 7,8 εκ των οποίων και οι δύο είναι παράγοντες που συμβάλλουν στη δευτερογενή εγκεφαλική αιμορραγία. Αντίθετα, δεν έχουν αναφερθεί περιπτώσεις δευτερογενούς εγκεφαλικού εμφράγματος σε ασθενείς με COVID-19. Ωστόσο, η COVID-19 μπορεί να προκαλέσει αύξηση των D-διμερών, που οδηγούν εύκολα σε θρομβωτικά αγγειακά επεισόδια. Προηγούμενες μελέτες είχαν αναφέρει περιπτώσεις δευτερογενούς εγκεφαλικού εμφράγματος στο SARS. Ως εκ τούτου, εικάζουμε ότι η COVID-19 έχει επίσης τη δυνατότητα να προκαλέσει εγκεφαλικά φλεβικά ή/και αρτηριακά εμφράγματα. Ωστόσο, λίγες μελέτες έχουν αναφέρει περιπτώσεις νευρολογικών βλαβών που σχετίζονται με την COVID-19 [7].

Τέλος, μια πρόσφατη έκθεση δείχνει ότι ο SARS-CoV-2 μπορεί επίσης να επιτεθεί και να βλάψει το νευρικό σύστημα, με ανίχνευση του RNA SARS-CoV-2 στο εγκεφαλονωτιαίο υγρό [8]. Με το ξέσπασμα της COVID-19, θα πρέπει όλοι να είναι σε εγρήγορση για την παρουσία νευρολογικών συμπτωμάτων παρόμοιων με εκείνων που αναφέρθηκαν για λοιμώξεις από αυτούς τους προηγούμενους μολυσματικούς κορονοϊούς. Λόγω έλλειψης σαφών και συγκεκριμένων κλινικών συμπτωμάτων στους ασθενείς που αναφέρθηκαν παραπάνω, η διάγνωση της COVID-19 είναι ιδιαίτερα δύσκολη και μπορεί να οδηγήσει σε χαμένες ή λανθασμένες διαγνώσεις που μπορεί να αυξήσουν την πιθανότητα μετάδοσης της λοίμωξης. Επιπλέον, λόγω απουσίας πυρετού και αναπνευστικών συμπτωμάτων, αυτοί οι ασθενείς με COVID-19 μπορεί να αγνοήσουν ή να μην γνωρίζουν την ασθένειά τους [7].

## 1.6 Κινητή υγεία και Covid-19

Η κινητή υγεία είναι ένας ταχέως αναπτυσσόμενος τομέας της ψηφιακής υγείας που παρέχει υποστήριξη, παράδοση και παρέμβαση υγειονομικής περίθαλψης μέσω τεχνολογιών κινητής τηλεφωνίας, όπως smartphone, tablet. Ενώ η κινητή υγεία αναφέρεται σε όλες τις κινητές συσκευές που μπορούν να μεταδώσουν δεδομένα, τα κινητά τηλέφωνα είναι αυτή τη στιγμή η πιο δημοφιλής πλατφόρμα για παρεμβάσεις στην υγεία. Δεδομένου ότι οι αναπτυσσόμενες χώρες αντιμετωπίζουν σταθερή ανάπτυξη στον επιπολασμό χρόνιων ασθενειών, καθώς και συνεχή επιβάρυνση από μεταδοτικές ασθένειες, η κινητή υγεία προσφέρει υποσχέσεις για την ανταπόκριση και στους δύο τύπους επιβαρύνσεων ασθενειών. Οι κινητές τεχνολογίες είναι ευρέως διαθέσιμες και μπορούν να διαδραματίσουν σημαντικό ρόλο στην υγειονομική περίθαλψη σε περιφερειακό, κοινοτικό και ατομικό επίπεδο [9].

Οι εφαρμογές για κινητά για τον κορονοϊό είναι μέθοδοι που χρησιμοποιούν οι κυβερνήσεις, τα νοσοκομεία, τα κολέγια, τα πανεπιστήμια και άλλες ομάδες για να βοηθήσουν τη δημόσια ανταπόκριση στην COVID-19. Αυτές οι εφαρμογές και οι πίνακες ελέγχου μπορούν να παρακολουθούν τα συμπτώματα, να παρέχουν τα πιο πρόσφατα δεδομένα σχετικά με την εξάπλωση του ιού, να βοηθήσουν να περιοριστούν οι επαφές και πολλά άλλα. Τα προγράμματα βοηθούν να βρεθεί όσα πρέπει να γνωρίζει κάποιος για τον ιό με διάφορους τρόπους [10].

### 1.6.1 Εφαρμογές για συμπτώματα κορονοϊού

Οι εφαρμογές μπορούν να είναι ένα σημαντικό εργαλείο για την παρακολούθηση της υγείας. Το εργαλείο ελέγχου COVID-19 από το CDC και την Apple καθοδηγεί σε μια σειρά ερωτήσεων σχετικά με την υγεία και την έκθεσή στον κορονοϊό. Στη συνέχεια, συμβουλεύει κατάλληλα εάν κάποιος πρέπει να δει έναν γιατρό. Μοιράζεται επίσης συμβουλές για κοινωνική αποστασιοποίηση, απομόνωση και παρακολούθηση συμπτωμάτων. Διατίθεται ως εφαρμογή για iPhone και στον ιστό [11]. Το WebMD διαθέτει επίσης ένα εργαλείο ελέγχου. Άλλες εφαρμογές συμπτωμάτων περιλαμβάνουν:

- COVID Symptom Study
- COVID Symptom Tracker



### 1.6.2 Εφαρμογές ιχνηλάτισης επαφών για τον κορονοϊό

Τα τμήματα υγείας χρησιμοποιούν εφαρμογές εντοπισμού επαφών για να ανιχνεύσουν άτομα που μπορεί να έχουν έρθει σε επαφή με κάποιον με COVID-19. Οι εφαρμογές βοηθούν στη λήψη δεδομένων και παρακολουθούν την κίνηση των ανθρώπων για να κάνουν τη διαδικασία γρηγορότερη και πιο αξιόπιστη. Δίνει την δυνατότητα να εντοπιστούν άτομα που μπορεί να είχαν εκτεθεί στον ιό, ώστε να το γνωρίζουν, να απομονωθούν και να προσέχουν τα συμπτώματα. Οι τεχνολογικοί γίγαντες Apple και Google συνεργάστηκαν σε μια πλατφόρμα που χρησιμοποιεί Bluetooth και το λειτουργικό σύστημα ενός τηλεφώνου για τον εντοπισμό επαφών. Μερικοί άνθρωποι είναι επιφυλακτικοί με μια εφαρμογή που παρακολουθεί πού πηγαίνουν και ποιους συναντούν. Αντί να αποθηκεύουν δεδομένα σε έναν κεντρικό διακομιστή που μπορεί να είναι ευάλωτα σε χάκερ, η Google και η Apple λένε ότι οι εφαρμογές τους δεν θα μπορούν να διαβάζουν οι ίδιες τα ακατέργαστα δεδομένα. Αντί αυτού, οι πληροφορίες θα είναι διαθέσιμες μόνο στους οργανισμούς υγείας μέσω αυτού που ονομάζεται διεπαφή προγραμματισμού εφαρμογών (API). Οι εφαρμογές εντοπισμού επαφών στις ΗΠΑ περιλαμβάνουν [12]:

- Crush COVID RI (for Rhode Island)
- Care19 Diary (for Wyoming, South Dakota, and North Dakota)
- GuideSafe (for Alabama)
- COVIDSafe (Washington)
- COVID Defense (Louisiana)
- COVIDWISE (Virginia)
- COVIDAlert (Connecticut, Delaware, Michigan, New Jersey, New York, Pennsylvania)
- Pathcheck (Massachusetts)
- Protect Texas Together (Texas)
- PunchAlert (Georgia)
- SlowCOVIDNC (North Carolina)

### 1.6.3 Εφαρμογές επιδημιολογίας του κορονοϊού

Οι εφαρμογές και οι πίνακες εργαλείων μπορούν να παρέχουν τις πιο πρόσφατες πληροφορίες για τον ιό, καθώς και για τους πόρους υγείας και ασφάλειας. Η επίσημη εφαρμογή του CDC παρέχει ενημερωμένες ειδήσεις για την υγεία και την COVID-19. Ο πίνακας ελέγχου του Παγκόσμιου Οργανισμού Υγείας (ΠΟΥ) ακολουθεί τον αριθμό των επιβεβαιωμένων κρουσμάτων και θανάτων ανά παγκόσμια περιοχή και χώρες που έχουν πληγεί περισσότερο. Δεν μπορεί να εγκατασταθεί σε φορητή συσκευή αυτή η δυνατότητα, αλλά έχει σχεδιαστεί για εύκολη προβολή σε κινητή συσκευή [12].

#### **1.6.4 Ερευνητικές εφαρμογές για τον κορονοϊό**

Ορισμένα κολέγια και πανεπιστήμια έχουν αναπτύξει εφαρμογές για κινητά για τον εντοπισμό και τη μελέτη της εξάπλωσης της COVID-19 και των επιπτώσεών του. Οι χρήστες υποβάλλουν δεδομένα σχετικά με την υγεία και τα συμπτώματά τους για να βοηθήσουν τους ερευνητές να βρουν τρόπους για να αποτρέψουν μελλοντικές εστίες. Η εφαρμογή COVID Control από το Πανεπιστήμιο Johns Hopkins επιτρέπει να εισαχθεί η θερμοκρασία του σώματός καθημερινά για να βοηθήσει τους ερευνητές να πάρουν μια ιδέα για το πού μπορεί να είναι το επόμενο «καυτό σημείο» για τη νόσο [12].

#### **1.6.5 Εφαρμογές κοινωνικής απόστασης**

Υπάρχουν πολλοί τρόποι χρήσης εφαρμογών για να μας βοηθήσουν να μειωθεί η επαφή με άλλους. Για παράδειγμα, τα εστιατόρια μπορούν να χρησιμοποιήσουν τις κάτωθι εφαρμογές για να ειδοποιήσουν τους πελάτες όταν το τραπέζι τους είναι έτοιμο για να αποφύγουν τους πολυσύχναστους χώρους αναμονής. Βοηθούν επίσης τους οδηγούς παράδοσης να επιβεβαιώσουν τις ανέπαφες παραδόσεις. Ορισμένες εφαρμογές μπορούν να βοηθήσουν να ελεγχθεί το διάστημα μεταξύ ενός ατόμου και ενός άλλου για να τους βοηθήσει να διατηρήσουν μια ασφαλή απόσταση μεταξύ τους. Οι εφαρμογές κοινωνικής απόστασης περιλαμβάνουν [12].

- 1point5
- mContain

#### **1.6.6 Εφαρμογές τηλεϊατρικής**

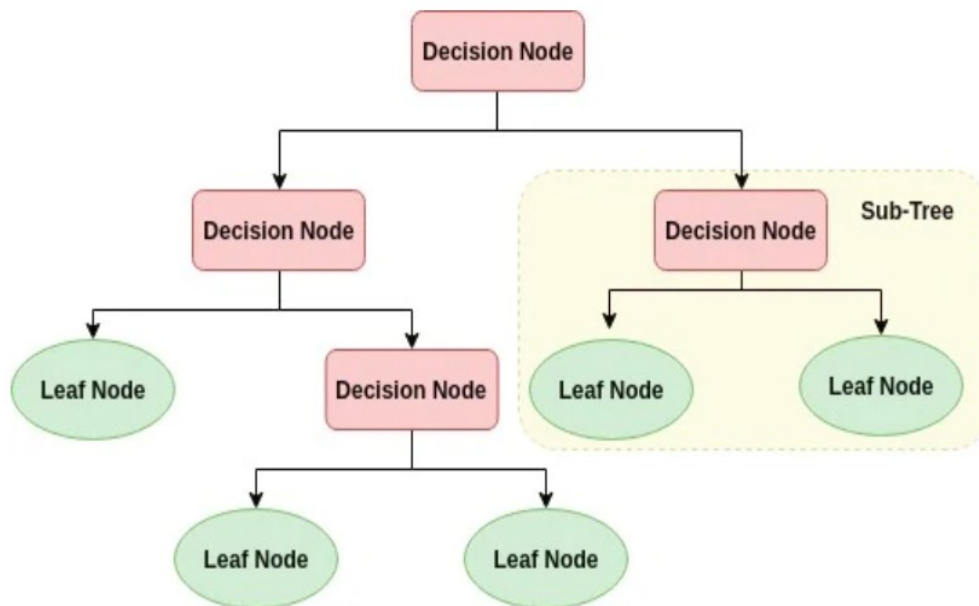
Ορισμένοι πάροχοι υγειονομικής περίθαλψης έχουν αναβάλει τις προσωπικές επισκέψεις κατά τη διάρκεια της πανδημίας. Ο γιατρός πλέον μπορεί να δει τον ασθενή σχεδόν μέσω τηλεθεραπείας. Κάνει εφικτά πράγματα όπως προγραμματισμός ραντεβού, μεταφόρτωση πληροφοριών όπως τα επίπεδα σακχάρου στο αίμα και συνομιλία μέσω βίντεο με το γιατρό. Υπάρχουν δεκάδες εφαρμογές τηλεϊατρικής, μεταξύ των οποίων [12].

- MDLive
- Teladoc
- Doctor On Demand
- Amwell
- MyOnCallDoc

## Κεφάλαιο 2<sup>ο</sup>: Μοντέλα

### 2.1 Δέντρα Αποφάσεων

Τα δέντρα αποφάσεων είναι ένας από τους πιο δημοφιλείς αλγόριθμους ταξινόμησης που χρησιμοποιούνται στην εξόρυξη δεδομένων και την μηχανική μάθηση. Η τεχνική εξόρυξης δεδομένων έχει γίνει δημοφιλές ερευνητικό εργαλείο για ιατρικούς ερευνητές που επιδιώκουν να εντοπίσουν και να εκμεταλλευτούν μοτίβα και σχέσεις μεταξύ μεγάλου αριθμού μεταβλητών και να είναι σε θέση να προβλέψουν την έκβαση μιας ασθένειας χρησιμοποιώντας τις ιστορικές περιπτώσεις που αποθηκεύονται σε σύνολα δεδομένων [13]. Για πολλά προβλήματα ταξινόμησης όπου χρησιμοποιούνται μεγάλα σύνολα δεδομένων και οι πληροφορίες που περιέχονται είναι πολύπλοκες και ενδέχεται να περιέχουν σφάλματα, τα δέντρα αποφάσεων παρέχουν μια χρήσιμη λύση. Όσον αφορά την ικανότητα, τα δέντρα αποφάσεων είναι μια γρήγορη και αποτελεσματική μέθοδος ταξινόμησης καταχωρήσεων συνόλων δεδομένων και μπορούν να παρέχουν καλές δυνατότητες υποστήριξης αποφάσεων [13]. Οι ταξινομητές δέντρων αποφάσεων μπορούν να εκτελέσουν αυτόματη επιλογή χαρακτηριστικών και μείωση της πολυπλοκότητας και η δομή δέντρων τους παρέχει εύκολα κατανοητές και ερμηνεύσιμες πληροφορίες σχετικά με την ικανότητα πρόβλεψης ή δημιουργίας της ταξινόμησης. Το δέντρο αποφάσεων στη συνέχεια κατασκευάζεται με αναδρομική διαίρεση ενός συνόλου δεδομένων σε πιο καθαρά, πιο ομοιογενή υποσύνολα με βάση ένα σύνολο δοκιμών που εφαρμόζονται σε μία ή περισσότερες τιμές χαρακτηριστικών σε κάθε κλάδο ή κόμβο στο δέντρο [13].



Εικόνα 5: Οπτικοποίηση του μοντέλου Decision Tree [13].

### 2.1.1 Random Forest Classifier

Ένα “τυχαίο δέντρο” είναι ένα δέντρο που κατασκευάστηκε τυχαία από ένα σύνολο από πιθανά δέντρα με  $K$  τυχαία χαρακτηριστικά σε κάθε κόμβο. Τυχαιότητα, στο πλαίσιο της εκπαίδευσης, σημαίνει ότι κάθε δέντρο έχει ίσες πιθανότητες να δειγματοληπτηθεί, δηλαδή έχει ομοιόμορφη κατανομή. Ο συνδυασμός μεγάλων συνόλων τυχαίων δέντρων οδηγεί σε ακριβή μοντέλα [14].

### 2.1.2 Gradient Boosting Classifier - eXtreme Gradient Boosting Classifier

Πολλά δέντρα αποφάσεων κατασκευάζονται όπου η εκπαίδευση κάθε δέντρου εξαρτάται από προηγούμενως εκπαιδευμένα δέντρα. Στην πράξη, ξεκινά με την εκπαίδευση ενός δέντρου αποφάσεων στο οποίο κάθε παρατήρηση έχει ίσο βάρος. Αφού αξιολογηθεί το πρώτο δέντρο, αυξάνονται τα βάρη εκείνων των παρατηρήσεων που είναι δύσκολο να ταξινομηθούν και μειώνονται τα βάρη για εκείνες που είναι εύκολο να ταξινομηθούν. Το δεύτερο δέντρο λοιπόν αναπτύσσεται σε αυτά τα σταθμισμένα δεδομένα [15]. Κάθε δέντρο εκπαιδεύεται να προβλέπει ψευδό-υπολείμματα των προηγούμενων δέντρων δεδομένης μιας προκαθορισμένης συνάρτησης (objective function). Η αναφορά σε νέα παραδείγματα εφαρμόζεται προσθετικά στο υπόλοιπο αθροιστικό αποτέλεσμα. Ιδιαίτερα το XGBoost διαθέτει μηχανισμούς (παραμέτρους) στοχευμένους στην μείωση του overfitting [16].

### 2.1.3 Πλεονεκτήματα [17]

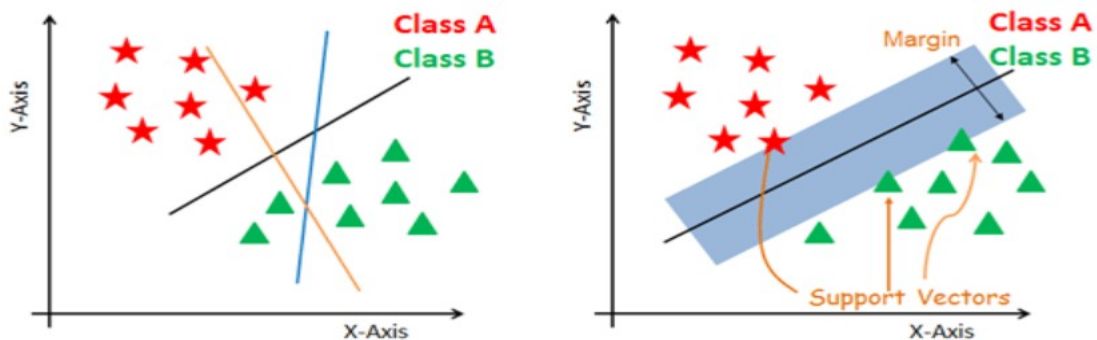
- Τα δέντρα αποφάσεων είναι εύκολο να ερμηνευτούν και να απεικονιστούν.
- Μπορούν εύκολα να καταγράψουν μη γραμμικά μοτίβα.
- Απαιτούν λιγότερη προ-επεξεργασία δεδομένων από τον χρήστη, για παράδειγμα, δεν χρειάζεται να κανονικοποιηθούν οι στήλες.
- Μπορούν να χρησιμοποιηθούν για κατασκευή χαρακτηριστικών, όπως πρόβλεψη τιμών που λείπουν, κατάλληλες για επιλογή μεταβλητών.
- Το δέντρο αποφάσεων δεν έχει υποθέσεις σχετικά με την κατανομή λόγω της μη παραμετρικής φύσης του αλγορίθμου.

### 2.1.4 Μειονεκτήματα [17]

- Ευαίσθητο σε θορυβώδη δεδομένα.
- Μπορεί να υπερκαλύψει θορυβώδη δεδομένα.
- Η μικρή διακύμανση στα δεδομένα μπορεί να οδηγήσει σε διαφορετικό δέντρο αποφάσεων. Αυτό μπορεί να μειωθεί με την αποθήκευση και την ενίσχυση αλγορίθμων.
- Τα δέντρα αποφάσεων είναι προκατειλημμένα με βάση δεδομένων ανισορροπίας, επομένως συνιστάται η εξισορρόπηση του συνόλου δεδομένων πριν από τη δημιουργία του δέντρου απόφασης.

## 2.2 Μηχανή διανυσματικής υποστήριξης

Ο στόχος του SVM είναι να βρει μια “γραμμή” που μεγιστοποιεί το περιθώριο μεταξύ δύο κλάσεων. Όταν όμως αντιμετωπίζουν μη γραμμικά προβλήματα, τα SVM δημιουργούν μια αντιστοίχιση μεταξύ ενός συνόλου τιμών εισόδου (παραδείγματα) και ενός χώρου χαρακτηριστικών, στο οποίο αυτές οι αρχικές μη γραμμικές οριακές κλάσεις γίνονται γραμμικά διαχωρίσιμες από έναν μετασχηματισμό (ή χαρτογράφηση) του χώρου των χαρακτηριστικών. Αυτή η χαρτογράφηση γίνεται από ένα σύνολο μαθηματικών συναρτήσεων που ονομάζονται πυρήνες. Μετά την εκτέλεση αυτής της αντιστοίχισης, οι SVM χρησιμοποιούν έναν επαναληπτικό αλγόριθμο εκπαίδευσης για να ελαχιστοποιήσουν μια λειτουργία σφάλματος [18].



Εικόνα 6: Οπτικοποίηση του μοντέλου SVM [6]

## 2.3 K Κοντινότεροι γείτονες

Ο ταξινομητής K κοντινότεροι γείτονες αναθέτει μια νέα παρατήρηση σε μια κλάση με πλειοψηφία από K πλησιέστερους γείτονες. Το μέτρο ανομοιότητας στον αλγόριθμο KNN ορίζεται συνήθως με βάση την απόσταση Minkowski ως εξής:

$$d(x_a, x_b) = \left( \sum_{j=1}^p (x_{aj} - x_{bj})^q \right)^{1/q}$$

όπου  $p$  : το πλήθος των χαρακτηριστικών

$q$  : θετική σταθερά (συνήθως 1 ή 2)

$d(x_a, x_b)$  : η απόσταση μεταξύ των σημείων  $a$  και  $b$

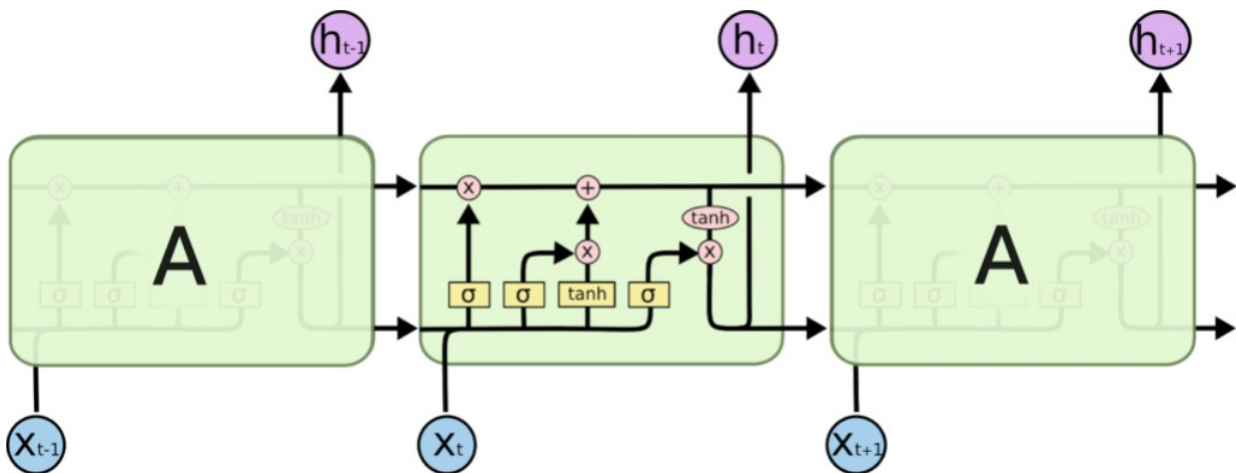
Η καλύτερη τιμή για το  $k$  επιλέγεται μετά από cross validation [19].

## 2.4 Επαναλαμβανόμενο Νευρωνικό Δίκτυο

Ένα επαναλαμβανόμενο νευρωνικό δίκτυο είναι μια κατηγορία τεχνητών νευρωνικών δικτύων όπου οι συνδέσεις μεταξύ των κόμβων σχηματίζουν ένα κατευθυνόμενο γράφημα κατά μήκος μιας χρονικής ακολουθίας. Αυτό του επιτρέπει να επιδεικνύει χρονική δυναμική συμπεριφορά. Προερχόμενα από νευρωνικά δίκτυα τροφοδοσίας, τα επαναλαμβανόμενα νευρωνικά δίκτυα μπορούν να χρησιμοποιήσουν την εσωτερική τους κατάσταση (μνήμη) για την επεξεργασία ακολουθιών μεταβλητού μήκους εισόδων [20].

### 2.4.1 Long Short Term Memory (LSTM)

Το LSTM έχει αποδεχθεί σταθερό και ισχυρό εργαλείο για τη μοντελοποίηση εξαρτήσεων μεγάλης εμβέλειας [21]. Λειτουργεί με παρόμοιο τρόπο με τα RNN, αφού επεξεργάζεται δεδομένα που μεταδίδουν πληροφορίες καθώς διαδίδονται προς τα εμπρός. Η διαφορά έγκειται στις λειτουργίες που μπορούν να εκτελεστούν στα κελιά του. Αυτές οι λειτουργίες χρησιμοποιούνται για να επιτρέψουν στο LSTM να διατηρήσει ή να ξεχάσει πληροφορίες [20].



Εικόνα 7: Οπτικοποίηση του μοντέλου LSTM [6]

$$\begin{aligned}
 i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\
 f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\
 c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
 o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co} \circ c_t + b_o) \\
 h_t &= o_t \circ \tanh(c_t)
 \end{aligned}$$

#### 2.4.1.1 Βασική ιδέα

Η κατάσταση του κελιού λειτουργεί ως αυτοκινητόδρομος μεταφοράς που διαδίδει σχετικές πληροφορίες σε όλη την κατεύθυνση της αλυσίδας ακολουθιών. Παραπέμπει σαν ένα νευρωνικό ανθρώπινο σύστημα όπου τα κελιά είναι οι νευρώνες και μεταφέρουν σχετικές πληροφορίες καθ' όλη τη διάρκεια της επεξεργασίας της αλληλουχίας [22]. Έτσι, ακόμη και οι πληροφορίες από τα προηγούμενα χρονικά βήματα μπορούν να οδηγήσουν σε μεταγενέστερα χρονικά βήματα, μειώνοντας τις επιπτώσεις της βραχυπρόθεσμης μνήμης. Καθώς η κατάσταση του κελιού συνεχίζει το ταξίδι της, οι πληροφορίες προστίθενται ή αφαιρούνται στην κατάσταση του κελιού μέσω πυλών. Οι πύλες είναι διαφορετικά νευρωνικά δίκτυα που αποφασίζουν ποιες πληροφορίες επιτρέπονται. Οι πύλες μπορούν να μάθουν ποιες πληροφορίες είναι σχετικές για να διατηρηθούν ή να ξεχαστούν κατά τη διάρκεια της προπόνησης [22].

#### 2.4.1.2 Forget gate

Αυτή η πύλη αποφασίζει ποιες πληροφορίες πρέπει να πετιούνται ή να φυλάσσονται. Οι πληροφορίες από την προηγούμενη κρυφή κατάσταση και οι πληροφορίες από την τρέχουσα είσοδο περνούν μέσω της σιγμοειδούς συνάρτησης. Οι τιμές βγαίνουν μεταξύ 0 και 1. Όσο πιο κοντά στο 0 σημαίνει ξεχνάς, όσο πιο κοντά στο 1 σημαίνει κρατάς [22].

#### 2.4.1.3 Πύλη εισόδου

Για να ενημερωθεί η κατάσταση κελιού υπάρχει η πύλη εισόδου. Πρώτον, περνάει η προηγούμενη κρυφή κατάσταση και η τρέχουσα είσοδος σε μια συνάρτηση ενεργοποίησης, συνήθως σιγμοειδής. Αυτό αποφασίζει ποιες τιμές θα ενημερώνονται επηρεαζόμενη από την συνάρτηση ενεργοποίησης. Μπορεί επίσης να περάσει η κρυφή κατάσταση και η τρέχουσα είσοδος στη συνάρτηση  $\tanh$  για να περιοριστούν οι τιμές μεταξύ -1 και 1 για την ρύθμιση του δικτύου. Στη συνέχεια, πολλαπλασιάζεται η έξοδος  $\tanh$  με την έξοδο της σιγμοειδούς. Η έξοδος σιγμοειδούς θα καθορίσει ποιες πληροφορίες είναι σημαντικές για να κρατηθούν από την έξοδο  $\tanh$  [22].

#### 2.4.1.4 Κατάσταση κελιού

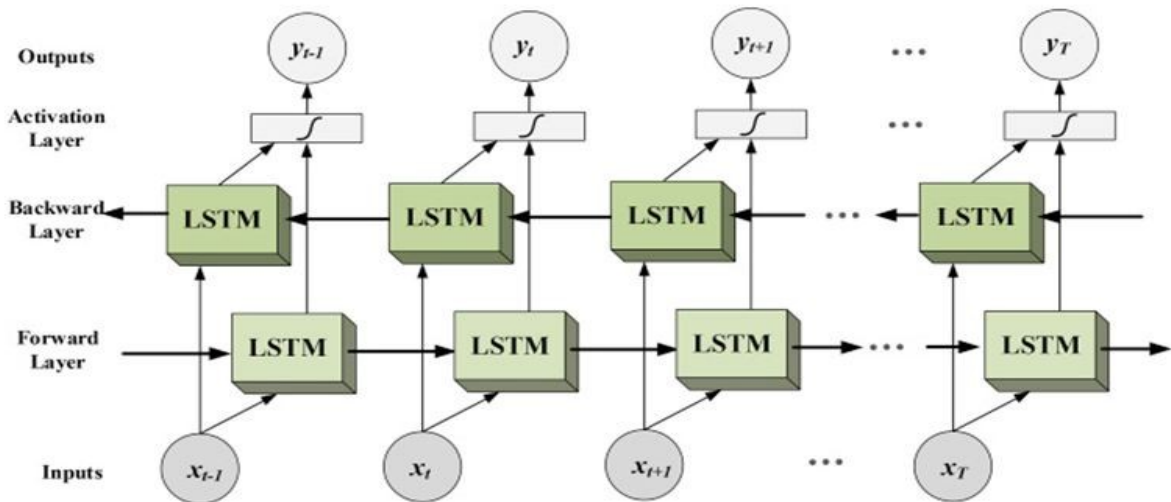
Πρώτον, η κατάσταση του κελιού πολλαπλασιάζεται κατά τομή με το διάνυσμα λήψης. Αυτό έχει τη δυνατότητα να ρίξει τιμές στην κατάσταση του κελιού αν πολλαπλασιαστεί με τιμές κοντά στο 0. Στη συνέχεια, η έξοδος από την πύλη εισόδου ενημερώνει την κατάσταση του κελιού σε νέες τιμές που το νευρωνικό δίκτυο θεωρεί συναφείς. Αυτό δίνει τη νέα κατάσταση του κελιού [22].

#### 2.4.1.5 Πύλη εξόδου

Η πύλη εξόδου αποφασίζει ποια πρέπει να είναι η επόμενη κρυφή κατάσταση, η οποία χρησιμοποιείται επίσης για προβλέψεις. Πρώτα, περνάει η προηγούμενη κρυφή κατάσταση και η τρέχουσα είσοδος σε μια σιγμοειδή συνάρτηση. Στη συνέχεια, περνάει η πρόσφατα τροποποιημένη κατάσταση του κελιού στη συνάρτηση  $\tanh$  και η έξοδος πολλαπλασιάζεται με την έξοδο σιγμοειδούς για να αποφασιστεί ποιες πληροφορίες πρέπει να φέρει η κρυφή κατάσταση. Η έξοδος είναι η κρυφή κατάσταση. Η νέα κατάσταση κελιού και η νέα κρυφή μεταφέρονται στη συνέχεια στο επόμενο χρονικό βήμα [22].

### 2.4.2 Bidirectional Long Short Term Memory

Η αμφίδρομη μακροπρόθεσμη μνήμη (bi-lstm) είναι η διαδικασία δημιουργίας οποιουδήποτε νευρωνικού δικτύου που έχει πληροφορίες αλληλουχίας και προς τις δύο κατευθύνσεις προς τα πίσω ή προς τα εμπρός. Σε αμφίδρομη κατεύθυνση, η εισροή ρέει προς δύο κατευθύνσεις, κάνοντας το bi-lstm διαφορετικό από το κανονικό LSTM. Με το κανονικό LSTM, είναι εφικτή η ροή εισόδου προς μία μόνο κατεύθυνση, είτε προς τα πίσω είτε προς τα εμπρός. Με την αμφίδρομη κατεύθυνση διατηρείται λοιπόν το μέλλον και οι προηγούμενες πληροφορίες [21].



Εικόνα 8: Οπτικοποίηση του μοντέλου Bi-LSTM [6]

## 2.5 Πλεονεκτήματα (Bi)LSTM

- Οι αρχιτεκτονικές RNN όπως το LSTM και το Bi-LSTM χρησιμοποιούνται σε περιπτώσεις όπου το μαθησιακό πρόβλημα είναι σειριακό, όπως είναι και η περίπτωση των δεδομένων ήχου
- Οι LSTM και οι αμφίδρομες παραλλαγές τους είναι δημοφιλείς επειδή προσπαθούν να μάθουν πώς και πότε να ξεχνούν και πότε να μην χρησιμοποιούν πύλες στην αρχιτεκτονική τους.
- Το LSTM έχει αποδεχθεί σταθερό και ισχυρό εργαλείο για τη μοντελοποίηση εξαρτήσεων μεγάλης εμβέλειας [22].
- Έχουν μια επαναλαμβανόμενη κατάσταση η οποία ενημερώνεται κάθε φορά που τροφοδοτούνται νέα δεδομένα μέσω του δικτύου. Με αυτόν τον τρόπο, το LSTM έχει μνήμη [23].
- Οι μακροχρόνιες καθυστερήσεις σε ορισμένα προβλήματα γεφυρώνονται χρησιμοποιώντας τα LSTM όπου χειρίζονται επίσης θόρυβο, κατανεμημένες αναπαραστάσεις και συνεχείς τιμές [24].
- Τα LSTM παρέχουν ένα ευρύ φάσμα παραμέτρων όπως τα ποσοστά εκμάθησης και οι προκαταλήψεις εισόδου και εξόδου. Ως εκ τούτου, δεν χρειάζεται ιδιαίτερες προσαρμογές. Η πολυπλοκότητα για την ενημέρωση κάθε βάρους μειώνεται στο  $O(1)$  [24].



## 2.6 Μειονεκτήματα (Bi-)LSTM

- Τα LSTM έγιναν δημοφιλή επειδή μπορούσαν να λύσουν το πρόβλημα των εξαφανισμένων διανυσμάτων. Εν τέλει, δεν καταφέρνουν να το αφαιρέσουν εντελώς. Το πρόβλημα έγκειται στο γεγονός ότι τα δεδομένα πρέπει ακόμα να μετακινηθούν από κύτταρο σε κύτταρο για την αξιολόγησή τους. Επιπλέον, το κελί έχει γίνει αρκετά πολύπλοκο τώρα με τα πρόσθετα χαρακτηριστικά (όπως *forget gates*) [24].
- Απαιτούν πολλούς πόρους και χρόνο για να εκπαιδευτούν και να γίνουν έτοιμα για εφαρμογή. Από τεχνική άποψη, χρειάζονται υψηλό εύρος ζώνης μνήμης λόγω των γραμμικών στρωμάτων που υπάρχουν σε κάθε κελί, τα οποία το σύστημα συνήθως αποτυγχάνει να παρέχει. Έτσι, από άποψη υλικού, τα LSTM γίνονται αρκετά αναποτελεσματικά [24].
- Με την άνοδο της εξόρυξης δεδομένων, οι προγραμματιστές αναζητούν ένα μοντέλο που μπορεί να θυμάται παλιότερες πληροφορίες για μεγαλύτερο χρονικό διάστημα από τα LSTM. Η πηγή έμπνευσης για ένα τέτοιο μοντέλο είναι η ανθρώπινη συνήθεια να χωρίζει μια δεδομένη πληροφορία σε μικρά μέρη για εύκολη ανάμνηση [24].
- Τα LSTM είναι επιρρεπή σε υπερπροσαρμογή και είναι δύσκολο να εφαρμοστεί ο αλγόριθμος 'εγκατάλειψης' για να περιοριστεί αυτό το ζήτημα. Το Dropout είναι μια μέθοδος κανονικοποίησης όπου οι εισόδους και οι επαναλαμβανόμενες συνδέσεις με μονάδες LSTM αποκλείονται πιθανώς από την ενεργοποίηση και τις ενημερώσεις βάρους κατά την εκπαίδευση ενός δικτύου [24].

## **Κεφάλαιο 3<sup>ο</sup>: Ανάλυση - Επεξεργασία Δεδομένων – Μεθοδολογία**

### **3.1 Δεδομένα**

Στην παρούσα διπλωματική θα τεθούν υπό πειραματική επεξεργασία και ανάλυση δύο βάσεις δεδομένων : Coswara dataset και Cambridge dataset

#### **3.1.1 Coswara δεδομένα**

Το έργο Coswara στοχεύει στην ανάπτυξη ενός διαγνωστικού εργαλείου για τον COVID-19 με βάση το αναπνευστικό, το βήχα και την ομιλία [25]. Οι συμμετέχοντες κλήθηκαν να συνεισφέρουν καταγραφές βήχα μέσω μιας διαδικτυακής πλατφόρμας συλλογής δεδομένων χρησιμοποιώντας τα smartphone τους [26]. Τα συλλεγμένα δεδομένα ήχου περιλαμβάνουν γρήγορη και αργή αναπνοή, βαθύ και ρηχό βήχα, ομιλία παρατεταμένων φωνηέντων και προφορικά ψηφία. Ηλικία, φύλο, γεωγραφική θέση, η τρέχουσα κατάσταση της υγείας και οι προϋπάρχουσες ιατρικές καταστάσεις έχουν επίσης καταγραφεί. Η κατάσταση της υγείας διακρίνεται «υγιή», «εκτεθειμένη», «θεραπευμένη» ή «μολυσμένη». Ηχογραφήσεις ηχογραφήθηκαν στα 44,1 KHz και οι συμμετέχοντες προέρχονται από όλες τις ηπείρους εκτός από την Αφρική. Σε αυτή τη μελέτη, χρησιμοποιήθηκε η ακατέργαστη ηχογράφηση [25].

#### **3.1.2 Cambridge δεδομένα**

Τα δεδομένα προέρχονται από διαδικτυακές πλατφόρμες αλλά και από εφαρμογές σε περιβάλλον Android. Οι συμμετέχοντες κλήθηκαν να εισαγάγουν την ηλικία και το φύλο τους, ένα σύντομο ιατρικό ιστορικό καθώς και αν νοσηλεύονται. Στη συνέχεια, οι χρήστες εισάγουν τα συμπτώματά τους (εάν υπάρχουν) και ξεκινούν την καταγραφή. Τους ζητείται να βήξουν τρεις φορές και να αναπνεύσουν βαθιά από το στόμα τους τρεις έως πέντε φορές. Τέλος, οι χρήστες καλούνται να απαντήσουν εάν έχουν διαγνωστεί με COVID-19 ύστερα από εργαστηριακό έλεγχο. Οι ροές δεδομένων κρυπτογραφημένες και τα δεδομένα αποθηκεύονται με ασφάλεια στους διακομιστές. Τα δεδομένα διαβιβάζονται από το τηλέφωνο όταν ο χρήστης είναι συνδεδεμένος σε WiFi και αποθηκεύεται τοπικά μέχρι τότε. Εάν γίνει επιτυχής μετάδοση, τα δεδομένα αφαιρούνται από τη συσκευή. Δεν συλλέγονται ηλεκτρονικές διευθύνσεις χρηστών ή προσωπικά αναγνωριστικά. Οι εφαρμογές εμφανίζουν ένα μοναδικό αναγνωριστικό στο τέλος της έρευνας για να επιτρέψει στους χρήστες να επικοινωνήσουν και να ζητήσουν την διαγραφή των δεδομένων τους. Τα δεδομένα είναι διαθέσιμα στην ιστοσελίδα [27]. Λόγω ευαίσθητων δεδομένων, πρέπει να εκχωρηθεί άδεια για την χρήση τους [28].

### **3.2 Εξαγωγή Χαρακτηριστικών**

Είναι η διαδικασία ανάδειξης των βασικών χαρακτηριστικών ενός σήματος. Ένα κατάλληλο χαρακτηριστικό μιμείται τις ιδιότητες ενός σήματος με πολύ συμπαγή τρόπο [29]. Για τον προσδιορισμό ηχητικών δεδομένων υπάρχει ένα σύνολο παραμέτρων που ονομάζονται χαρακτηριστικά ομιλίας. Ο προσδιορισμός αυτών των χαρακτηριστικών είναι ένα σημαντικό βήμα που επηρεάζει την ακρίβεια του συστήματος. Οι αλγόριθμοι επεξεργασίας ηχητικού σήματος γενικά περιλαμβάνουν ανάλυση σήματος, εξαγωγή των ιδιοτήτων του, πρόβλεψη της συμπεριφοράς του, αναγνώριση εάν υπάρχει κάποιο μοτίβο στο σήμα και πώς συσχετίζεται ένα συγκεκριμένο σήμα με άλλο παρόμοιο σήμα.

Τα χαρακτηριστικά υπολογίστηκαν μέσω της βιβλιοθήκη Python librosa. Αξίζει να σημειωθεί ότι κάθε χαρακτηριστικό είναι μια χρονική σειρά μήκους συμβατή με το μήκος του αρχικού κύματος βήχα [29].

### 3.2.1 Mel-frequency cepstral coefficients (MFCCs)

Η κλίμακα Mel συσχετίζει την αντιλαμβανόμενη συχνότητα ή τον τόνο ενός καθαρού τόνου με την πραγματική μετρημένη συχνότητά της. Οι άνθρωποι είναι πολύ καλύτεροι στο να διακρίνουν μικρές αλλαγές στο ύψος σε χαμηλές συχνότητες από ό,τι στις υψηλές συχνότητες. Η ενσωμάτωση αυτής της κλίμακας κάνει τα χαρακτηριστικά μας να ταιριάζουν περισσότερο με αυτά που ακούνε οι άνθρωποι [30].

Ο τύπος για τη μετατροπή από συχνότητα σε κλίμακα Mel είναι:

$$M(f) = 1125 \ln(1+f/700)$$

### 3.2.2 Μηδενικός ρυθμός διασταύρωσης

Το ZCR είναι ένας πολύ αποτελεσματικός τρόπος για την ανίχνευση της φωνητικής δραστηριότητας που καθορίζει εάν ένα πλαίσιο ομιλίας εκφράζεται, δεν ακούγεται ή είναι αθόρυβο. Το ZCR είναι υψηλότερο για τα μη φωνεμένα τμήματα της ομιλίας σε σύγκριση με το φωνητικό τμήμα της ομιλίας. Είναι σαφές ότι το ZCR για τμήματα χωρίς φωνή είναι πολύ υψηλό από ό,τι για τα τμήματα με φωνή. Φυσικά, σε ιδανικές συνθήκες το τμήμα σιωπής σε μια καθαρή ομιλία το ZCR πρέπει να είναι ίσο με το μηδέν [30].

### 3.2.3 Επίπεδο ενέργειας

Η ενέργεια ενός σήματος αντιστοιχεί στο συνολικό μέγεθος του σήματος. Για τα ηχητικά σήματα, αυτό αντιστοιχεί περίπου στο πόσο δυνατό είναι το σήμα [29]. Η ενέργεια σε ένα σήμα ορίζεται ως:

$$\sum_n |x(n)|^2$$

### 3.2.4 Φασματικό κέντρο

Το φασματικό κεντροειδές είναι ένα μέτρο που χρησιμοποιείται στην ψηφιακή επεξεργασία σήματος για τον χαρακτηρισμό ενός φάσματος. Υποδεικνύει πού βρίσκεται το κέντρο μάζας του φάσματος. Αντιληπτικά, έχει μια ισχυρή σύνδεση με την εντύπωση της φωτεινότητας ενός ήχου.

Υπολογίζεται ως ο σταθμισμένος μέσος όρος των συχνοτήτων που υπάρχουν στο σήμα, που προσδιορίζεται με τη χρήση μετασχηματισμού Fourier, με τα μεγέθη τους ως βάρη [29]:

$$\text{Centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

όπου  $x(n)$  : αντιπροσωπεύει τη σταθμισμένη τιμή συχνότητας, ή μέγεθος, του αριθμού κάδου  $n$

$f(n)$  : αντιπροσωπεύει την κεντρική συχνότητα αυτού του κάδου.

### 3.2.5 Φασματική μετατόπιση

Η φασματική μετάβαση ορίζεται ως η συχνότητα όπου το 85% της ενέργειας στο φάσμα είναι κάτω από αυτό το σημείο. Συχνά χρησιμοποιείται ως δείκτης της κλίσης των συχνοτήτων που υπάρχουν σε ένα παράθυρο [29].

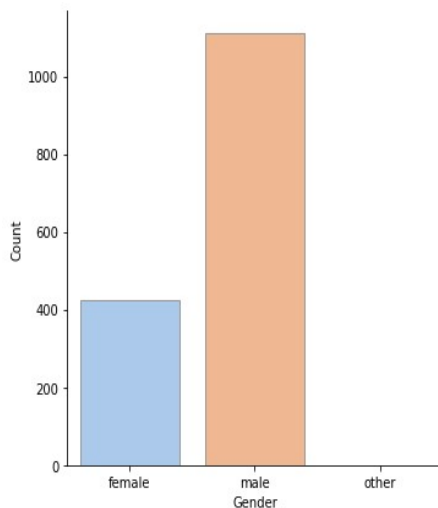
### 3.2.6 Φασματικό εύρος ζώνης

Το φασματικό εύρος ζώνης ή η φασματική εξάπλωση προέρχεται από το φασματικό κεντροειδές. Είναι το φασματικό εύρος ενδιαφέροντος γύρω από το κεντροειδές, δηλαδή η απόκλιση από το φασματικό κεντροειδές. Το εύρος ζώνης είναι ευθέως ανάλογο με την ενέργεια που κατανέμεται στις ζώνες συχνοτήτων. Μαθηματικά, είναι ο σταθμισμένος μέσος όρος των αποστάσεων των ζωνών συχνοτήτων από το φασματικό κεντροειδές [30].

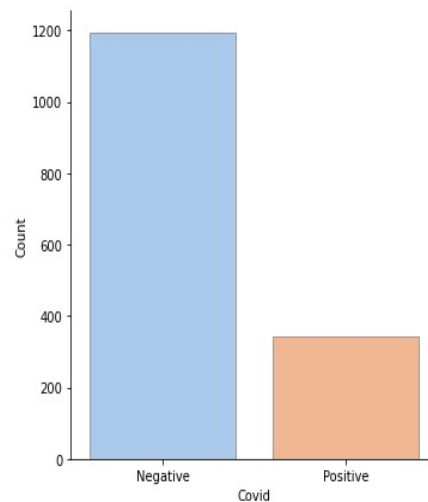
### 3.2.7 Φασματική επιπεδότητα

Η φασματική επιπεδότητα είναι ένα χαρακτηριστικό των ακουστικών σημάτων που ήταν χρήσιμο σε πολλές εφαρμογές επεξεργασίας ηχητικού σήματος. Ο παραδοσιακός ορισμός της φασματικής επιπεδότητας είναι ο λόγος του γεωμετρικού μέσου όρου προς τον αριθμητικό μέσο του φάσματος μεγέθους του σήματος, όπως λαμβάνεται από το DFT [31].

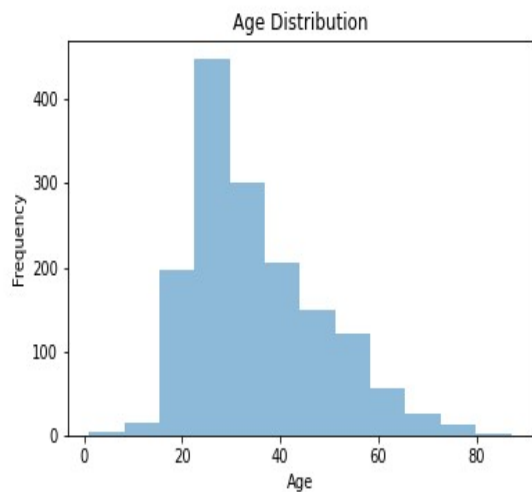
## 3.3 Οπτικοποίηση Δεδομένων



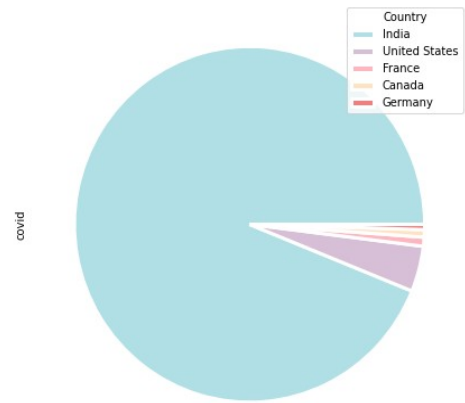
Εικόνα 9: Πλήθος των θηλυκών / αρσενικών



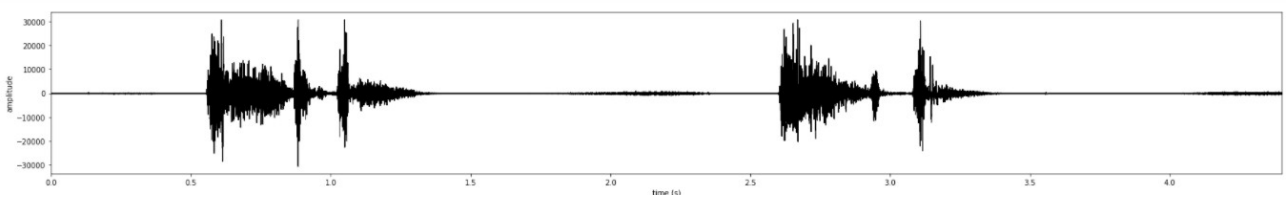
Εικόνα 10: Πλήθος θετικών / αρνητικών σε Covid-19



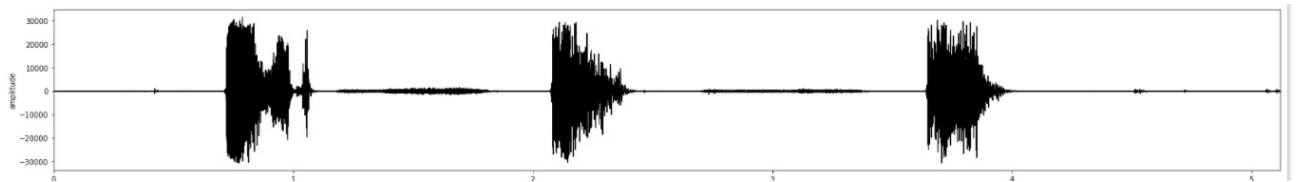
Εικόνα 11: Ηλικιακή διασπορά του δείγματος



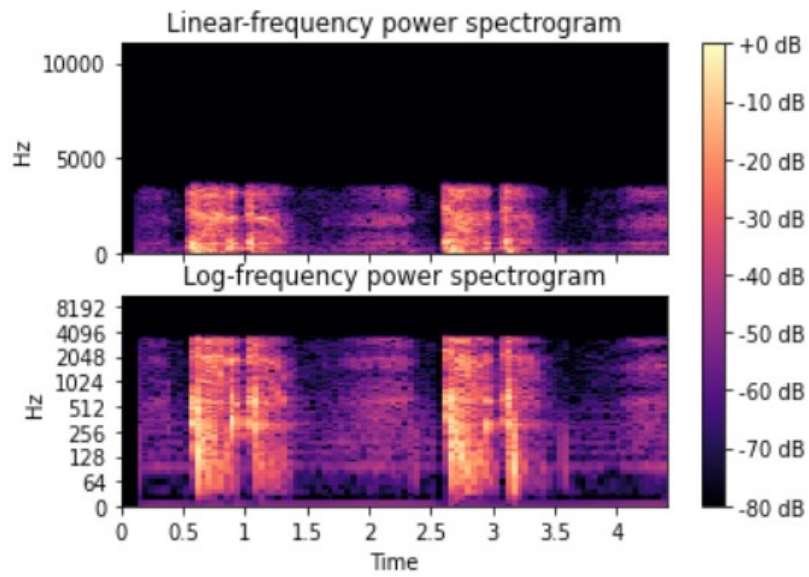
Εικόνα 12: Γεωγραφική διασπορά του δείγματος (top 5)



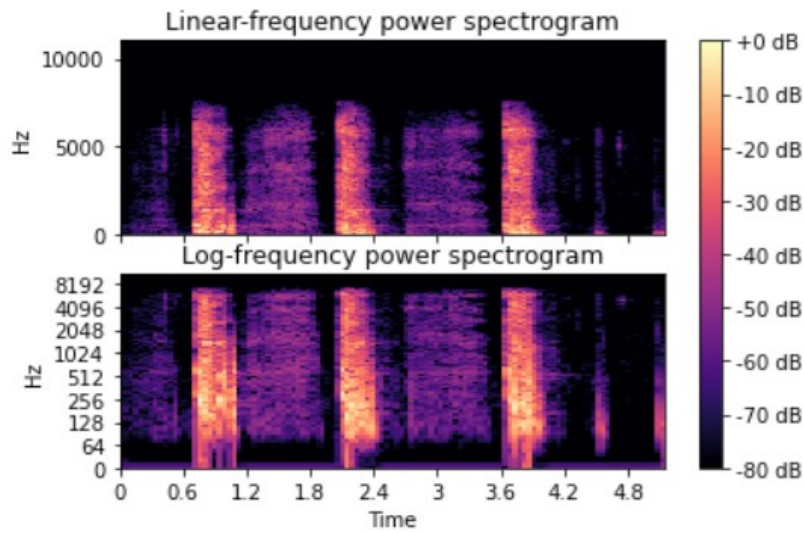
Εικόνα 13: Ηχητική καταγραφή βήχα αρνητικού στον κορονοϊό



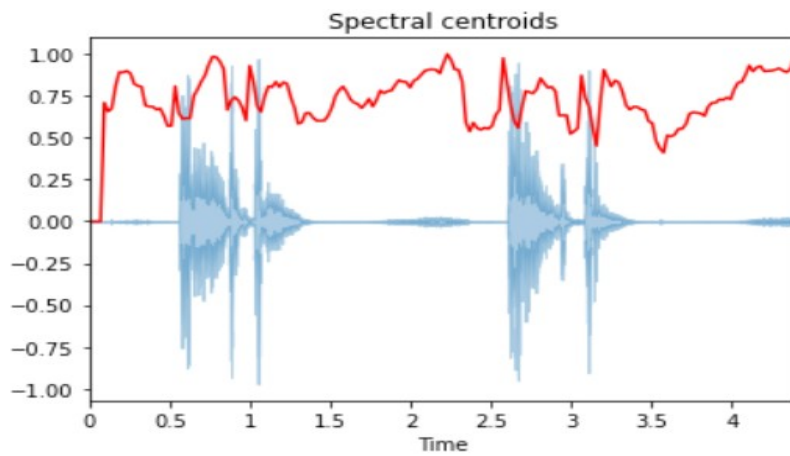
Εικόνα 14: Ηχητική καταγραφή βήχα θετικού στον κορονοϊό



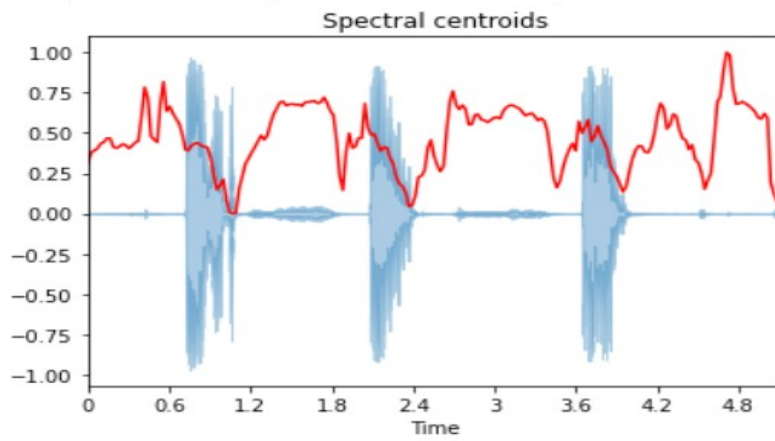
Εικόνα 15: Spectrogram βήχα αρνητικού στον κορονοϊό



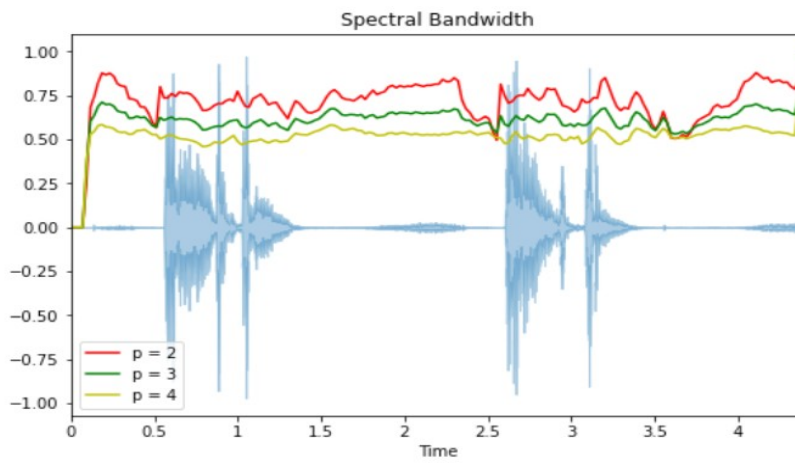
Εικόνα 16: Spectrogram βήχα θετικού στον κορονοϊό



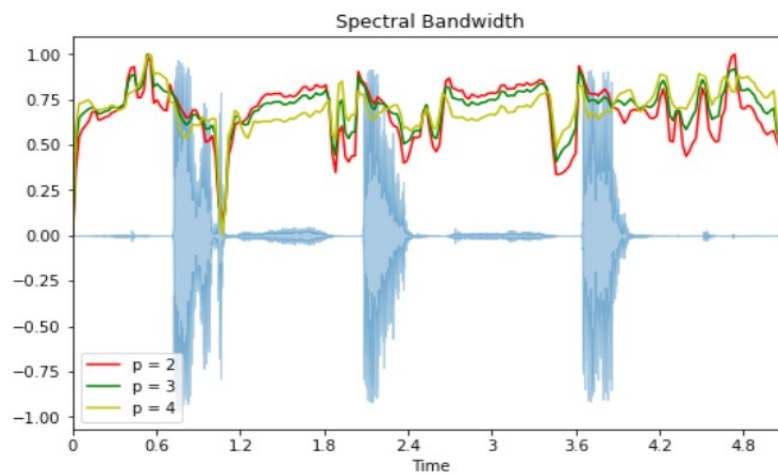
Εικόνα 17: Spectral centroids βήχα αρνητικού στον κορονοϊό



Εικόνα 18: Spectral centroids βήχα θετικού στον κορονοϊό



Εικόνα 19: Spectral bandwidth βήχα αρνητικού στον κορονοϊό



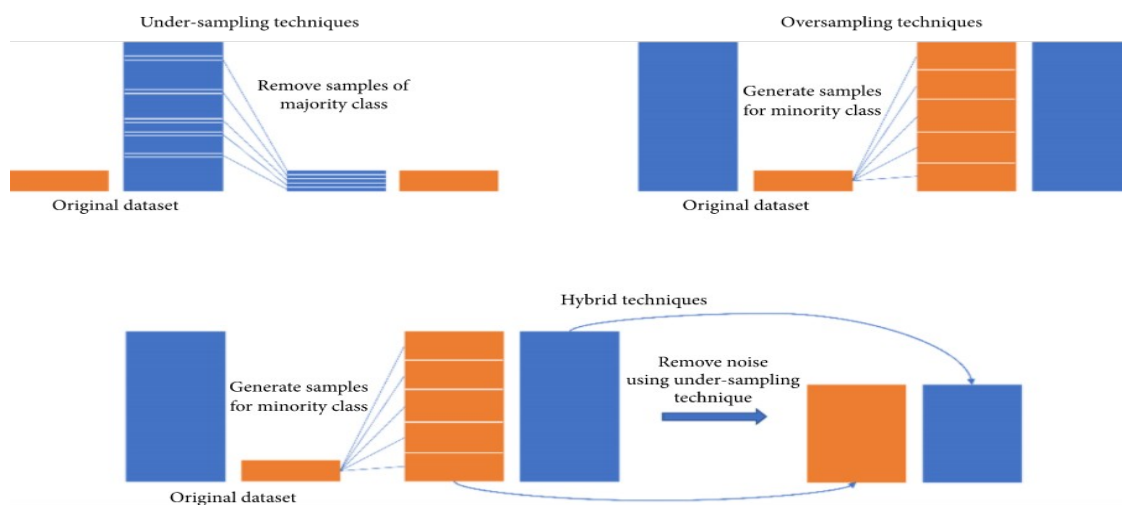
Εικόνα 20: Spectral bandwidth βήχα αρνητικού στον κορονοϊό

### 3.4 Μη ισορροπημένα δεδομένα

Το πρόβλημα ανισορροπίας κλάσης αντιστοιχεί σε τομείς για τους οποίους μία κλάση αντιπροσωπεύεται από μεγάλο αριθμό παραδειγμάτων ενώ το άλλο αντιπροσωπεύεται από λίγα μόνο. Έχει αναγνωρισθεί σε πολλές εφαρμογές και αποτελεί ένα εξελισσόμενο θέμα έρευνας μηχανικής μάθησης.

Ένα από τα βασικά προβλήματα κατά τη μάθηση με μη ισορροπημένα σύνολα δεδομένων είναι η έλλειψη δεδομένων όπου ο αριθμός των δειγμάτων είναι μικρός ή δεν υπάρχει διαθέσιμο δείγμα για μια συγκεκριμένη τάξη. Εάν υπάρχει έλλειψη δεδομένων, το εκτιμώμενο όριο απόφασης μπορεί να απέχει πολύ από το πραγματικό όριο. Ο Jarokowicz και ο Stephen [32] ανέφεραν ότι για απλά σύνολα δεδομένων που ήταν γραμμικά διαχωρίσιμα, οι επιδόσεις των ταξινομητών δεν ήταν επιρρεπείς σε οποιαδήποτε ανισορροπία. Πράγματι, καθώς ο βαθμός πολυπλοκότητας των δεδομένων αυξανόταν, ο παράγοντας ανισορροπίας της κλάσης άρχισε να επηρεάζει την ικανότητα γενίκευσης των ταξινομητών. Οι περισσότεροι αλγόριθμοι που βασίζονται στην ακρίβεια είναι προκατειλημμένοι προς την επικρατούσα τάξη. Οι προσεγγίσεις μηχανικής μάθησης βελτίωσαν τη συνολική ακρίβεια αναθέτοντας την επικαλυπτόμενη περιοχή στην τάξη της πλειοψηφίας και αγνόησαν ή αντιμετώπισαν τη μειοψηφία ως θόρυβο.

Πάρα πολλοί αλγόριθμοι έχουν σχεδιαστεί για την αντιμετώπιση των μη ισορροπημένων δεδομένων. Από την πλευρά της στρατηγικής, οι μέθοδοι έχουν κατηγοριοποιηθεί σε δύο κύριους τύπους. Σε αλγοριθμικό επίπεδο, η προσαρμογή των συντελεστών των σφαλμάτων στη συνάρτηση απώλειας, που ονομάζεται επίσης μάθηση ευαίσθητη στη συνάρτηση κόστους. Ο πίνακας κόστους που μετρά διαφορετικές “ποινές” για εσφαλμένη ταξινόμηση είναι κρίσιμος για τη βελτίωση της απόδοσης. Σε επίπεδο δειγματοληψίας, ο ευκολότερος τρόπος είναι η τυχαία δειγματοληψία των δεδομένων εκπαίδευσης από ολόκληρο το σύνολο των δεδομένων εκπαίδευσης με τέτοιο τρόπο ώστε να εξισορροπηθούν οι αναλογίες στις τάξεις. Η τυχαία δειγματοληψία τείνει να οδηγήσει σε overfitting, εάν οι δείκτες δειγματοληψίας δεν είχαν διαμορφωθεί σωστά. Η επαναλαμβανόμενη δειγματοληψία ήταν ευκολότερη στην εφαρμογή, αλλά δύσκολη να εφαρμοστεί αποτελεσματικά. Μια άλλη δημοφιλής μέθοδος είναι η τεχνική συνθετικής μειονότητας υπερδειγματοληψίας (SMOTE) που προτάθηκε από τους Chawla et al. [33] και η βασική ιδέα ήταν η κατασκευή των συνθετικών δειγμάτων μειοψηφίας μέσω της παρεμβολής μεταξύ των δεδομένων εκπαίδευσης μειονοτήτων και των πλησιέστερων γειτονιών της.



Εικόνα 21: Τεχνική undersampling & oversampling



### 3.4.1 Υπερδειγματοληψία

Οι μέθοδοι υπερδειγματοληψίας αντιγράφουν ή δημιουργούν νέα συνθετικά παραδείγματα στην τάξη μειοψηφίας. Περιγράφονται παρακάτω μερικά παραδείγματα τέτοιων μεθόδων [34].

#### 3.4.1.1 Adasyn

Η βασική ιδέα του Adasyn είναι να χρησιμοποιήσει μια σταθμισμένη κατανομή για διαφορετικά παραδείγματα τάξεων μειονοτήτων ανάλογα με το επίπεδο δυσκολίας τους στη μάθηση, όπου παράγονται περισσότερα συνθετικά δεδομένα για παραδείγματα τάξεων μειονοτήτων που είναι πιο δύσκολο να μάθουν σε σύγκριση με εκείνα τα παραδείγματα μειονοτήτων που είναι πιο εύκολο να μάθουν [35], [36]. Ως αποτέλεσμα, η προσέγγιση Adasyn βελτιώνει τη μάθηση σε σχέση με τις κατανομές δεδομένων με δύο τρόπους:

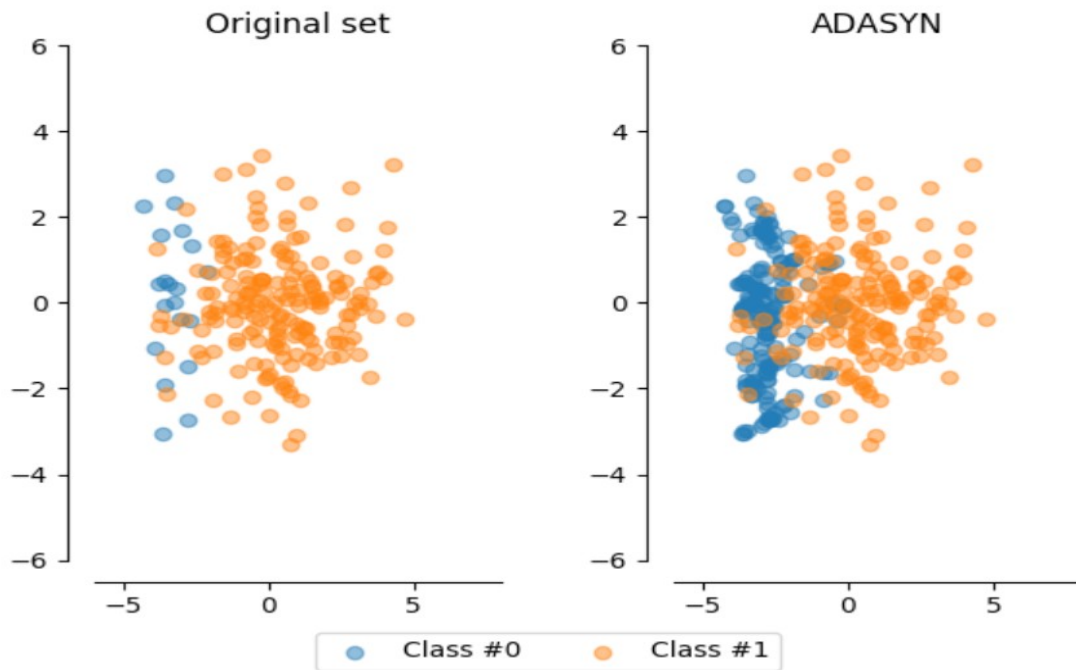
- μείωση της προκατάληψης που εισάγεται από την ανισορροπία της τάξης και
- προσαρμοστική προσαρμογή του ορίου της απόφασης ταξινόμησης στα δύσκολα παραδείγματα.

Διαδικασία [37]:

1. Υπολογισμός του βαθμού ανισορροπίας,  $d = \frac{m_r}{m_x}$
2. Εάν  $d < d_x$  (όπου  $d_x$  είναι το προκαθορισμένο όριο για μέγιστη ανεκτή ανισορροπία) τότε:
  - Υπολογισμός του αριθμού των συνθετικών δειγμάτων που θα δημιουργηθούν από την μειονοτική τάξη:  $G = (m_x - m_r) \times \beta$ ,  $\beta$  είναι το επίπεδο ισορροπίας των συνθετικών δειγμάτων που παράγονται. Όταν  $\beta = 1$  σημαίνει ότι υπάρχει μια συνολική ισορροπία μεταξύ των δύο τάξεων.
  - Για κάθε δείγμα μειονότητας  $x_i$ , εύρεση των  $k$ -πλησιέστερων γειτόνων με βάση την Ευκλείδεια απόσταση και υπολογισμός της αναλογίας  $r_i$ ,  $r_i = \frac{\Delta_i}{K}$
  - Κανονικοποίηση του  $r_x \leftarrow r_i / \sum r_i$ , έτσι ώστε το  $r_x$  να είναι πλέον κατανομή πυκνότητας.
  - Υπολογισμός συνθετικού δείγματος που παράγεται για κάθε σημείο δεδομένων μειοψηφίας  $g_i = r_x \times G$ , όπου  $G$  είναι ο συνολικός αριθμός παραδειγμάτων συνθετικών δεδομένων που πρέπει να δημιουργηθούν για την τάξη μειοψηφίας όπως ορίζεται στην προαναφερθείσα εξίσωση.
  - Για κάθε παράδειγμα δεδομένων μειονότητας κατηγορίας  $x_i$ , δημιουργία παραδειγμάτων συνθετικών δεδομένων  $g_i$  σύμφωνα με τα ακόλουθα βήματα:

Κάντε το βρόχο από το 1 στο  $g_i$ :

- i. Επιλογή τυχαίου παραδείγματος από τα δεδομένα μειοψηφίας,  $x_u$ , από τους  $K$  πλησιέστερους γείτονες για δεδομένα  $x_i$ .
- ii. Δημιουργία του παραδείγματος συνθετικών δεδομένων:  $s_i = x_i + (x_u - x_i)$  λόπου  $(x_u - x_i)$  είναι το διάνυσμα διαφοράς σε  $n$ -διαστάσεις χώρους, και  $\lambda$  είναι ένας τυχαίος αριθμός:  $\lambda \in [0, 1]$ .



Εικόνα 22: Πριν και μετά την υπερδעיματοληψία

### 3.4.1.2 Smote

Το SMOTE είναι μια τεχνική υπερδעיματοληψίας, με την οποία νέα δείγματα δημιουργούνται για την τάξη μειοψηφίας. Αυτός ο αλγόριθμος βοηθά στην υπέρβαση του προβλήματος υπερπροσαρμογής που δημιουργείται από την τυχαία υπερδעיματοληψία. Επικεντρώνεται στο χώρο των χαρακτηριστικών για τη δημιουργία νέων εμφανίσεων με τη βοήθεια παρεμβολής μεταξύ των θετικών παρουσιών που βρίσκονται μαζί [35].

Το SMOTE μπορεί να είναι προβληματικό για τους ταξινομητές που συμπεραίνουν την ανεξαρτησία μεταξύ των δειγμάτων, όπως για παράδειγμα penalized logistic regression ή μέθοδοι ανάλυσης διακρίσεων. Επίσης, η εκτέλεση επιλογής μεταβλητών μετά τη χρήση του SMOTE πρέπει να γίνεται με προσοχή, επειδή οι περισσότερες μέθοδοι επιλογής μεταβλητών υποθέτουν ότι τα δείγματα είναι ανεξάρτητα [38].

Το SMOTE είναι μια τεχνική υπερδעיματοληψίας που δημιουργεί συνθετικά δείγματα από την τάξη μειονοτήτων χρησιμοποιώντας τις πληροφορίες που είναι διαθέσιμες στα δεδομένα. Για κάθε δείγμα από την κατηγορία μειοψηφίας  $x$ , 5 (ή  $n_{min}-1$  εάν  $n_{min} \leq 5$ ) δείγματα από την τάξη μειονοτήτων με τη μικρότερη ευκλείδεια απόσταση προσδιορίστηκαν από το αρχικό δείγμα (ως πλησιέστεροι γείτονες) και ένα από αυτά επιλέγεται τυχαία ( $x^R$ ). Το νέο συνθετικό δείγμα SMOTE ορίστηκε ως :

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

όπου  $u$  επιλέχθηκε τυχαία από το  $U(0, 1)$  και είναι ίδιο για όλες τις μεταβλητές, αλλά διαφέρει για κάθε δείγμα SMOTE. Αυτή η επιλογή εγγυάται ότι το δείγμα SMOTE βρίσκεται στη γραμμή που ενώνει τα δύο αρχικά δείγματα που χρησιμοποιήθηκαν για τη δημιουργία του [38].

### 3.4.1.3 SMOTE + Tomek Links

Αποτελεί μια υβριδική τεχνική που στοχεύει στον “καθαρισμό” αλληλεπικαλυπτόμενων δεδομένων για κάθε μία από τις κλάσεις που κατανέμονται στο χώρο του δείγματος. Αφού πραγματοποιηθεί υπερδειγματοληψία με βάση το SMOTE, οι κλάσεις μπορεί να υπερκαλύπτονται οδηγώντας σε overfitting. Για τον λόγο αυτό, δημιουργούνται τα σύνδεσμοι Tomek, τα αντίθετα δηλαδή ζεύγη δειγμάτων κλάσης που είναι οι πιο κοντινοί γείτονες μεταξύ τους. Επομένως, η πλειοψηφία των παρατηρήσεων κλάσης από αυτούς τους συνδέσμους αφαιρείται καθώς πιστεύεται ότι αυξάνει τον διαχωρισμό τάξης κοντά στα όρια απόφασης. Για καλύτερα αποτελέσματα, οι σύνδεσμοι Tomek εφαρμόζονται σε δείγματα της κλάσης μειοψηφίας που υπερανεδείχθηκαν από το SMOTE. Έτσι, αντί να αφαιρούνται παρατηρήσεις μόνο από την πλειοψηφία, αφαιρούνται γενικά και από τις δύο κλάσεις παρατηρήσεις από τους συνδέσμους Tomek [36].

### 3.4.1.4 SMOTE + ENN

Αποτελεί μια υβριδική τεχνική όπου περισσότερες παρατηρήσεις αφαιρούνται από τον χώρο του δείγματος. Εδώ, το ENN είναι μια άλλη τεχνική υποδειγματοληψίας όπου υπολογίζονται οι πλησιέστεροι γείτονες καθενός από την τάξη πλειοψηφίας. Εάν οι πλησιέστεροι γείτονες κατατάξουν λανθασμένα τη συγκεκριμένη περίπτωση της τάξης της πλειοψηφίας, τότε αυτή η περίπτωση διαγράφεται. Η ενσωμάτωση αυτής της τεχνικής με υπερμετρημένα δεδομένα που πραγματοποιούνται από την SMOTE βοηθά στον εκτεταμένο καθαρισμό δεδομένων. Αυτό έχει ως αποτέλεσμα έναν πιο σαφή και συνοπτικό διαχωρισμό τάξης [36].

## 3.4.2 Υποδειγματοληψία

Η υποδειγματοληψία αναφέρεται σε μια ομάδα τεχνικών που έχουν σχεδιαστεί για να εξισορροπήσουν την κατανομή κλάσης για ένα σύνολο δεδομένων ταξινόμησης που έχει μια στρεβλή κατανομή κλάσης. Οι τεχνικές υποδειγματοληψίας αφαιρούν παραδείγματα από το σύνολο δεδομένων εκπαίδευσης που ανήκουν στην τάξη της πλειοψηφίας, προκειμένου να εξισορροπηθεί καλύτερα η κατανομή των τάξεων [34].

### 3.4.2.1 Τυχαία Υποδειγματοληψία

Η απλούστερη τεχνική υποδείγματος περιλαμβάνει τυχαία επιλογή παραδειγμάτων από την τάξη πλειοψηφίας και διαγραφή τους από το σύνολο δεδομένων εκπαίδευσης. Αυτό αναφέρεται ως τυχαία υποδειγματοληψία. Αν και απλή και αποτελεσματική, ένας περιορισμός αυτής της τεχνικής είναι ότι τα παραδείγματα αφαιρούνται χωρίς καμία ανησυχία για το πόσο χρήσιμα ή σημαντικά μπορεί να είναι για τον καθορισμό του ορίου απόφασης μεταξύ των τάξεων. Αυτό σημαίνει ότι είναι πιθανό, ή ακόμη και πιθανό, να διαγραφούν χρήσιμες πληροφορίες [34].

### 3.4.2.2 Near Miss

Το Near Miss αναφέρεται σε μια συλλογή μεθόδων υποδειγματοληψίας που επιλέγουν παραδείγματα με βάση την απόσταση των παραδειγμάτων της τάξης της πλειοψηφίας από τα παραδείγματα της τάξης των μειονοτήτων [39].

Για να αποφευχθεί το πρόβλημα της απώλειας πληροφοριών στις περισσότερες τεχνικές υποδειγματοληψίας, χρησιμοποιούνται ευρέως μέθοδοι πλησιέστερων γειτόνων. Η βασική διαίσθηση σχετικά με τη λειτουργία μεθόδων πλησίον γειτόνων έχει ως εξής:

1. Η μέθοδος εντοπίζει πρώτα τις αποστάσεις μεταξύ όλων των περιπτώσεων της τάξης της πλειοψηφίας και των περιπτώσεων της τάξης της μειονότητας. Εδώ, η πλειοψηφική τάξη πρόκειται να υποδειγματιστεί.
2. Στη συνέχεια, επιλέγονται η περιπτώσεις της τάξης της πλειοψηφίας που έχουν τις μικρότερες αποστάσεις από αυτές της τάξης της μειονότητας.
3. Εάν υπάρχουν  $k$  περιπτώσεις στην κατηγορία μειονότητας, η πλησιέστερη μέθοδος θα οδηγήσει σε  $k \cdot n$  περιπτώσεις της πλειοψηφίας [34].

Υπάρχουν τρεις εκδόσεις της τεχνικής, που ονομάζονται NearMiss-1, NearMiss-2 και NearMiss-3 :

- Το NearMiss-1 επιλέγει παραδείγματα από την πλειοψηφική τάξη που έχουν τη μικρότερη μέση απόσταση από τα τρία πλησιέστερα παραδείγματα από την τάξη μειοψηφίας [34].
- Το NearMiss-2 επιλέγει παραδείγματα από την πλειοψηφική τάξη που έχουν τη μικρότερη μέση απόσταση από τα τρία απομακρυσμένα παραδείγματα από την τάξη μειοψηφίας [34]. Σφάλμα: Δεν βρέθηκε η πηγή παραπομπής.
- Το NearMiss-3 περιλαμβάνει την επιλογή ενός δεδομένου αριθμού παραδειγμάτων κλάσης πλειοψηφίας για κάθε παράδειγμα της τάξης μειονότητας που είναι πιο κοντά [34].

## 3.5 Κανονικοποίηση Δεδομένων

Οι αλγόριθμοι μηχανικής μάθησης, όπως η γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση, το νευρωνικό δίκτυο κ.λπ. που χρησιμοποιούν την τεχνική της διανυσματικής κλίσης (gradient descent) ως τεχνική βελτιστοποίησης απαιτούν κλιμάκωση (scalling) δεδομένων. Η ύπαρξη χαρακτηριστικών σε παρόμοια κλίμακα μπορεί να βοηθήσει την τεχνική της διανυσματικής κλίσης να συγκλίνει πιο γρήγορα προς τα ελάχιστα [40].

### 3.5.1 Τυποποίηση

Η τυποποίηση είναι μια τεχνική κλιμάκωσης όπου οι τιμές επικεντρώνονται γύρω από το μέσο όρο με τυπική απόκλιση μονάδας. Αυτό σημαίνει ότι ο μέσος όρος του χαρακτηριστικού γίνεται μηδέν και η προκύπτουσα κατανομή έχει τυπική απόκλιση μονάδας. Ακολουθεί τον παρακάτω τύπο [40]:

$$X' = \frac{X - \mu}{\sigma}$$

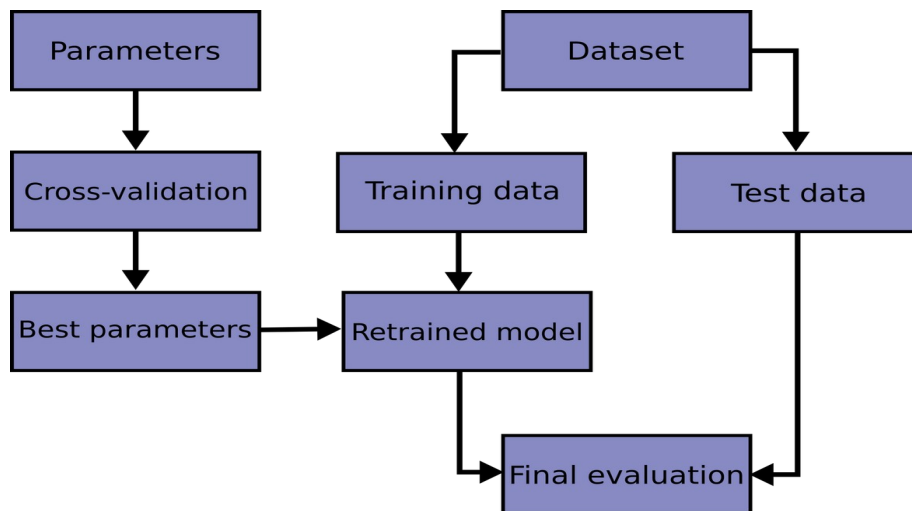
### 3.5.2 Ομαλοποίηση

Η ομαλοποίηση είναι μια τεχνική κλιμάκωσης στην οποία οι τιμές μετατοπίζονται και κλιμακώνονται έτσι ώστε να καταλήγουν να κυμαίνονται μεταξύ 0 και 1. Είναι επίσης γνωστή ως κλιμάκωση Min-Max [40].

### 3.5.3 Επιλογή κανονικοποίησης

- Η ομαλοποίηση είναι καλό να χρησιμοποιείται όταν είναι γνωστό ότι η διανομή των δεδομένων δεν ακολουθεί μια διανομή Gauss. Αυτό μπορεί να είναι χρήσιμο σε αλγόριθμους που δεν υποθέτουν καμία διανομή δεδομένων, όπως K-Nearest Neighbours και Neural Networks [40].
- Η τυποποίηση, από την άλλη πλευρά, μπορεί να είναι χρήσιμη σε περιπτώσεις όπου τα δεδομένα ακολουθούν μια κατανομή Gauss. Ωστόσο, αυτό δεν χρειάζεται να είναι απαραίτητα αληθινό. Επίσης, σε αντίθεση με την ομαλοποίηση, η τυποποίηση δεν έχει όριο. Έτσι, ακόμη και αν έχετε υπερβολικά υψηλά στα δεδομένα σας, δεν θα επηρεαστούν από την τυποποίηση [40].

## 3.6 Επικύρωση των μοντέλων



Εικόνα 23: Πριν και μετά την υπερδευματοληγία

Η εκμάθηση των παραμέτρων μιας συνάρτησης πρόβλεψης και η δοκιμή της στα ίδια δεδομένα είναι ένα μεθοδολογικό λάθος: ένα μοντέλο που απλώς θα επαναλάμβανε τις ετικέτες των δειγμάτων που μόλις είδε θα είχε μια τέλεια βαθμολογία αλλά δεν θα μπορούσε να προβλέψει κάτι χρήσιμο σε δεδομένα που δεν έχει δει. Αυτή η κατάσταση ονομάζεται υπερπροσαρμογή (overfitting). Για να αποφευχθεί, είναι συνηθισμένη πρακτική κατά την εκτέλεση ενός (υπό επίβλεψη) πειράματος μηχανικής εκμάθησης να διατηρείτε μέρος των διαθέσιμων δεδομένων ως σύνολο δοκιμών  $X_{test}$ ,  $y_{test}$  [41].

Κατά την αξιολόγηση διαφορετικών παραμέτρων ("υπερπαραμέτρους") για εκτιμητές, εξακολουθεί να υπάρχει κίνδυνος υπερβολικής προσαρμογής στο σύνολο δοκιμών, επειδή οι παράμετροι μπορούν να τροποποιηθούν έως ότου ο εκτιμητής αποδίδει βέλτιστα. Με αυτόν τον τρόπο, η γνώση σχετικά με το σύνολο δοκιμών μπορεί να "διαρρέυσει" στο μοντέλο και οι μετρήσεις αξιολόγησης δεν αναφέρουν πλέον την απόδοση γενίκευσης. Για την επίλυση αυτού του προβλήματος, ένα ακόμη μέρος του συνόλου δεδομένων μπορεί να χρησιμοποιηθεί ως το λεγόμενο «σύνολο επικύρωσης» (validation set): η εκπαίδευση προχωρά στο σετ εκπαίδευσης, μετά την οποία γίνεται αξιολόγηση στο σύνολο επικύρωσης

και όταν το πείραμα φαίνεται να είναι επιτυχές, η τελική αξιολόγηση μπορεί να γίνει στο σετ δοκιμών (test set). Ωστόσο, χωρίζοντας τα διαθέσιμα δεδομένα σε τρία σύνολα, μειώνουμε δραστικά τον αριθμό των δειγμάτων που μπορούν να χρησιμοποιηθούν για την εκμάθηση του μοντέλου και τα αποτελέσματα μπορεί να εξαρτώνται από μια συγκεκριμένη τυχαία επιλογή για το ζευγάρι (εκπαίδευση, επικύρωση) συνόλων [41].

### 3.6.1 K-fold Cross Validation

Μια λύση στο πρόβλημα με την τυχαία επιλογή των συνόλων εκπαίδευσης, επικύρωσης είναι μια διαδικασία που ονομάζεται cross validation (συντομία CV). Ένα τεστ θα πρέπει να διατηρηθεί για την τελική αξιολόγηση, για αυτό το λόγο το σετ επικύρωσης δεν χρειάζεται πλέον όταν εφαρμόζεται cross validation. Στη βασική προσέγγιση, που ονομάζεται k-fold CV, το σετ εκπαίδευσης χωρίζεται σε k μικρότερα σύνολα [41].

Ακολουθείται η ακόλουθη διαδικασία για κάθε μία από τα σπασίματα:

- Ένα μοντέλο εκπαιδεύεται χρησιμοποιώντας (k-1) σπασίματα ως δεδομένα εκπαίδευσης.
- Το προκύπτον μοντέλο επικυρώνεται στο υπόλοιπο μέρος των δεδομένων.

Το μέτρο απόδοσης που αναφέρεται στο cross validation είναι τότε ο μέσος όρος των τιμών που υπολογίζονται στον βρόχο. Αυτή η προσέγγιση μπορεί να είναι υπολογιστικά δαπανηρή, αλλά δεν σπαταλάει πάρα πολλά δεδομένα (όπως συμβαίνει κατά τον καθορισμό ενός συνόλου αυθαίρετων επικυρώσεων), το οποίο είναι ένα σημαντικό πλεονέκτημα σε προβλήματα όπως η αντίστροφη συμπεράσματα όπου ο αριθμός των δειγμάτων είναι πολύ μικρός [41].

### 3.6.2 Ensemble Learning

Οι μέθοδοι υποδειγματοληψίας υποδηλώνουν ότι τα δείγματα της πλειοψηφικής κλάσης χάνονται κατά τη διαδικασία εξισορρόπησης. Οι μέθοδοι Ensemble προσφέρουν μια εναλλακτική λύση για να χρησιμοποιηθούν περισσότερα δείγματα. Στην πραγματικότητα, ένα σύνολο ισορροπημένων συνόλων δημιουργείται και χρησιμοποιείται για να εκπαιδευτεί αργότερα οποιοδήποτε ταξινομητής [36].

### 3.6.3 Leave – one – out

Η διαδικασία leave one out, ή LOOCV, αξιοποιείται για την εκτίμηση της απόδοσης των αλγορίθμων μηχανικής μάθησης όσον αφορά την πρόβλεψη δεδομένων που δεν χρησιμοποιούνται για την εκπαίδευση του μοντέλου [42].

Είναι μια υπολογιστικά δαπανηρή διαδικασία για εκτέλεση, αν και έχει ως αποτέλεσμα μια αξιόπιστη και αμερόληπτη εκτίμηση της απόδοσης του μοντέλου. Αν και απλό στη χρήση και χωρίς ρύθμιση παραμέτρων, υπάρχουν στιγμές που η διαδικασία δεν πρέπει να χρησιμοποιείται, όπως όταν διατίθεται ένα πολύ μεγάλο σύνολο δεδομένων ή ένα υπολογιστικά ακριβό μοντέλο για αξιολόγηση [42].

Η leave one out τεχνική είναι μια ακραία έκδοση K-fold Cross Validation τεχνικής που έχει το μέγιστο υπολογιστικό κόστος. Απαιτείται να δημιουργηθεί και να αξιολογηθεί ένα μοντέλο για κάθε παράδειγμα στο σύνολο δεδομένων εκπαίδευσης.

Το όφελος από τόσα πολλά προσαρμοσμένα και αξιολογημένα μοντέλα είναι μια πιο ισχυρή εκτίμηση της απόδοσης του μοντέλου, καθώς σε κάθε σειρά δεδομένων δίνεται η ευκαιρία να αντιπροσωπεύει το σύνολο του συνόλου δεδομένων δοκιμής [42].

## Κεφάλαιο 4<sup>ο</sup>: Αποτελέσματα

### Κεφάλαιο 4.1 Κανονικοποίηση Δεδομένων

Παρατίθενται δύο πίνακες με τις μετρικές των μοντέλων, που χρησιμοποιήθηκαν, συγκριτικά για δύο μεθόδους κανονικοποίησης (τυποποίηση και ομαλοποίηση αντίστοιχα) :

#### Τυποποίηση

Μοντέλα	Accuracy	Precision	Recall	AUC
Random Forest	0.64	0.69	0.64	0.60
GBC	0.66	0.72	0.66	0.63
XGBoost	0.64	0.70	0.64	0.60
SVM	0.70	0.74	0.70	0.67
KNN	0.57	0.69	0.57	0.59
LSTM	0.68	0.72	0.68	0.65
Bi-LSTM	0.69	0.74	0.69	0.66

Πίνακας 24: Τυποποίηση

#### Ομαλοποίηση

Μοντέλα	Accuracy	Precision	Recall	AUC
Random Forest	0.64	0.69	0.64	0.60
GBC	0.66	0.72	0.66	0.63
XGBoost	0.64	0.70	0.64	0.60
SVM	0.70	0.74	0.70	0.67
KNN	0.58	0.69	0.58	0.59
LSTM	0.75	0.72	0.75	0.62
Bi-LSTM	0.75	0.56	0.75	0.50

Πίνακας 25: Ομαλοποίηση

## Κεφάλαιο 4.2 Βελτιστοποίηση Παραμέτρων

Παρατίθενται πίνακες με τις μετρικές των μοντέλων, που χρησιμοποιήθηκαν με την πρόοδο της βελτιστοποίησης των παραμέτρων:

### Χρήση προκαθορισμένων παραμέτρων

Μοντέλα	Accuracy	AUC
Random Forest	0.61	0.58
GBC	0.67	0.66
XGBoost	0.61	0.65
SVM	0.58	0.57
KNN	0.60	0.60
LSTM	0.57	0.45
Bi-LSTM	0.58	0.46

Πίνακας 26: Χρήση προκαθορισμένων παραμέτρων

### Βελτιστοποίηση με Εξονυχιστική Αναζήτηση

Μοντέλα	Accuracy	AUC
Random Forest	0.68	0.67
GBC	0.64	0.63
XGBoost	0.63	0.61
SVM	0.65	0.63
KNN	0.63	0.65

Πίνακας 27: Βελτιστοποίηση με Εξονυχιστική Αναζήτηση

### Βελτιστοποίηση αυξάνοντας τις διαστάσεις

Μοντέλα	Accuracy	AUC
LSTM	0.69	0.68
Bi-LSTM	0.75	0.59

Πίνακας 28: Βελτιστοποίηση αυξάνοντας τις διαστάσεις

### Βελτιστοποίηση αλλάζοντας τη συνάρτηση ενεργοποίησης στο 2ο layer σε relu

Μοντέλα	Accuracy	AUC
LSTM	0.75	0.69
Bi-LSTM	0.75	0.66

Πίνακας 29: Βελτιστοποίηση αλλάζοντας τη συνάρτηση ενεργοποίησης στο 2ο layer σε relu



### Βελτιστοποίηση προσθέτονας 3o layer

Μοντέλα	Accuracy	AUC
LSTM	0.74	0.50
Bi-LSTM	0.75	0.50

Πίνακας 30: Βελτιστοποίηση προσθέτονας 3o layer

### Βελτιστοποίηση προσθέτονας Dropout

Μοντέλα	Accuracy	AUC
LSTM	0.68	0.50
Bi-LSTM	0.75	0.50

Πίνακας 31: Βελτιστοποίηση προσθέτονας Dropout

### Βελτιστοποίηση προσθέτονας 2° Dropout

Μοντέλα	Accuracy	AUC
LSTM	0.74	0.50
Bi-LSTM	0.75	0.50

Πίνακας 32: Βελτιστοποίηση προσθέτονας 2° Dropout

## Κεφάλαιο 4.3 Εξισορρόπηση Δεδομένων

Παρατίθεται μια διερεύνηση κατάλληλης τεχνικής για την εξισορρόπηση των δεδομένων:

### NearMiss

Μοντέλα	Accuracy	Precision	Recall	AUC
Random Forest	0.50	0.67	0.50	0.56
GBC	0.48	0.67	0.48	0.55
XGBoost	0.52	0.67	0.52	0.56
SVM	0.45	0.66	0.45	0.54
KNN	0.69	0.72	0.69	0.63
LSTM	0.59	0.68	0.59	0.58
Bi-LSTM	0.54	0.68	0.54	0.58

Πίνακας 33: Undersampling with NearMiss

### Random Undersampling και SMOTE

Μοντέλα	Accuracy	Precision	Recall	AUC
Random Forest	0.68	0.68	0.68	0.67
GBC	0.64	0.64	0.64	0.63
XGBoost	0.63	0.62	0.63	0.61
SVM	0.65	0.65	0.65	0.63
KNN	0.63	0.66	0.63	0.65
LSTM	0.65	0.65	0.65	0.64
Bi-LSTM	0.66	0.65	0.66	0.63

Πίνακας 34: Random Undersampling and SMOTE

### Random Undersampling και Adasyn

Μοντέλα	Accuracy	Precision	Recall	AUC
Random Forest	0.60	0.59	0.60	0.58
GBC	0.60	0.60	0.60	0.59
XGBoost	0.57	0.57	0.57	0.56
SVM	0.64	0.64	0.64	0.63
KNN	0.64	0.68	0.64	0.66
LSTM	0.68	0.68	0.68	0.65
Bi-LSTM	0.69	0.69	0.69	0.66

Πίνακας 35: Random Undersampling and Adasyn

### Random Undersampling και SMOTETomek

Μοντέλα	Accuracy	Precision	Recall	AUC
Random Forest	0.64	0.66	0.64	0.65
GBC	0.62	0.62	0.62	0.61
XGBoost	0.55	0.56	0.55	0.54
SVM	0.67	0.67	0.67	0.65
KNN	0.62	0.65	0.62	0.64
LSTM	0.60	0.60	0.60	0.58
Bi-LSTM	0.66	0.66	0.66	0.64

Πίνακας 36: Random Undersampling and SMOTETomek

### Random Undersampling και SMOTEENN

Μοντέλα	Accuracy	Precision	Recall	AUC
Random Forest	0.50	0.67	0.5	0.56
GBC	0.58	0.64	0.58	0.61
XGBoost	0.57	0.64	0.57	0.60
SVM	0.59	0.63	0.59	0.61
KNN	0.58	0.66	0.58	0.62
LSTM	0.62	0.64	0.62	0.63
Bi-LSTM	0.61	0.62	0.61	0.61

Πίνακας 37: Random Undersampling and SMOTEENN

### Κεφάλαιο 4.4 Επιλογή Καλύτερων Χαρακτηριστικών

Παρατίθεται ο πίνακας με τις μετρικές των μοντέλων, που χρησιμοποιήθηκαν μετά την επιλογή των έντεκα καλύτερων χαρακτηριστικών:

#### Καλύτερα 11 χαρακτηριστικά

Μοντέλα	Accuracy	Precision	Recall	AUC
Random Forest	0.69	0.73	0.69	0.65
GBC	0.67	0.72	0.66	0.64
XGBoost	0.71	0.73	0.71	0.66
SVM	0.70	0.74	0.70	0.67
KNN	0.61	0.70	0.61	0.61
LSTM	0.81	0.80	0.81	0.68
Bi-LSTM	0.77	0.77	0.77	0.70

Πίνακας 38: Με το top-11 καλύτερα χαρακτηριστικά

## Κεφάλαιο 4.5 Επικύρωση Αποτελεσμάτων

Παρατίθενται οι κατά μέσο όρο μετρικές της ‘ακρίβειας’ και της ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ με την διακύμανση τους για την επικύρωση των μοντέλων όπως προέκυψαν από την τεχνική Repeated k-fold για τεχνική εξισορρόπησης random undersampling και smote:

### Επαναλαμβανόμενη Επικύρωση (Ακρίβεια)– random undersampling & smote

Μοντέλα	Accuracy	STD Accuracy
Random Forest	0.70	0.04
GBC	0.72	0.04
XGBoost	0.68	0.04
SVM	0.77	0.04
KNN	0.61	0.04
LSTM	0.75	0.05
Bi-LSTM	0.77	0.05

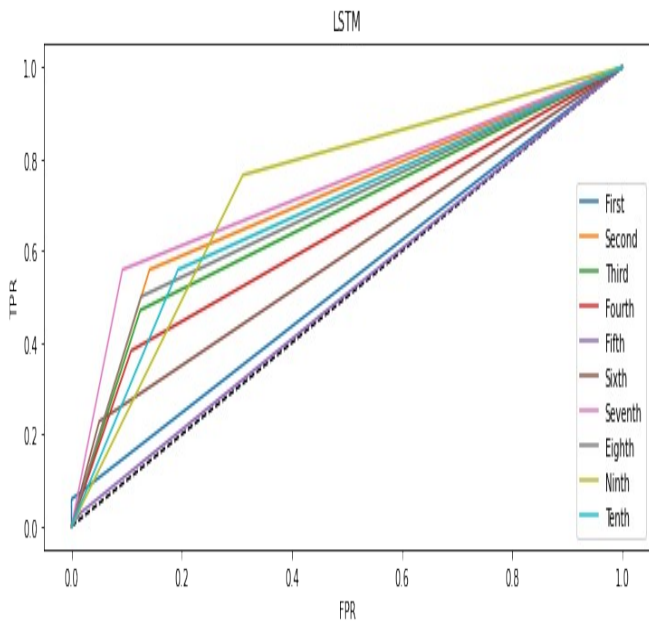
Πίνακας 39: Repeated k-fold (ακρίβεια)– random undersampling & smote

### Επαναλαμβανόμενη Επικύρωση (roc)– random undersampling & smote

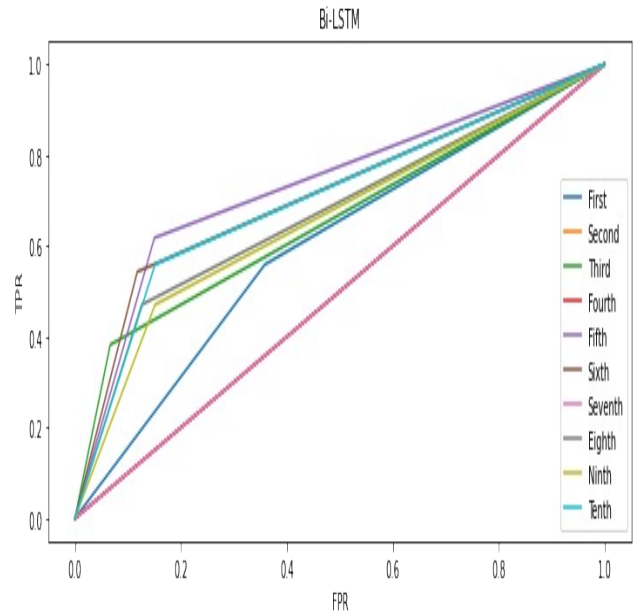
Μοντέλα	AUC	STD AUC
Random Forest	0.79	0.03
GBC	0.77	0.02
XGBoost	0.75	0.03
SVM	0.80	0.03
KNN	0.62	0.04
LSTM	0.67	0.05
Bi-LSTM	0.67	0.04

Πίνακας 40: Repeated k-fold (roc)– random undersampling & smote

Συγκεκριμένα για τα μοντέλα LSTM και Bi-LSTM παρουσιάζονται οι γραφικές της μετρικής ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ :



Εικόνα 41: Roc – random undersampling & smote



Εικόνα 42: Roc – random undersampling & smote

Παρατίθενται οι κατά μέσο όρο μετρικές της ‘ακρίβειας’ και της ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ με την διακύμανση τους για την επικύρωση των μοντέλων όπως προέκυψαν από την τεχνική Repeated k-fold για τεχνική εξισορρόπησης near miss:

#### Επαναλαμβανόμενη Επικύρωση (ακρίβεια) – near miss

Μοντέλα	Accuracy	STD Accuracy
Random Forest	0.57	0.06
GBC	0.61	0.04
XGBoost	0.60	0.05
SVM	0.66	0.05
KNN	0.62	0.04
LSTM	0.67	0.05
Bi-LSTM	0.73	0.04

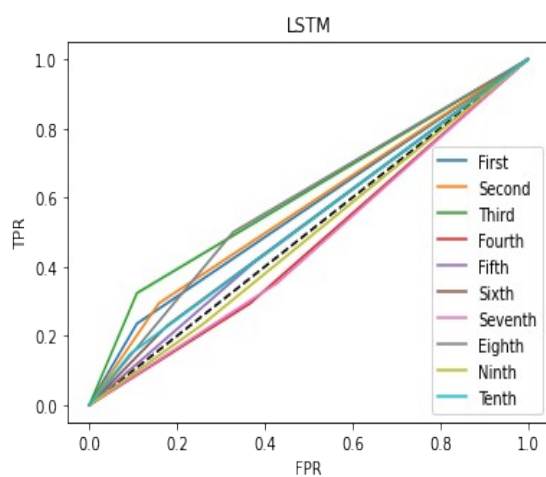
Πίνακας 43: Repeated k-fold (ακρίβεια) – near miss

### Επαναλαμβανόμενη Επικύρωση (roc) – near miss

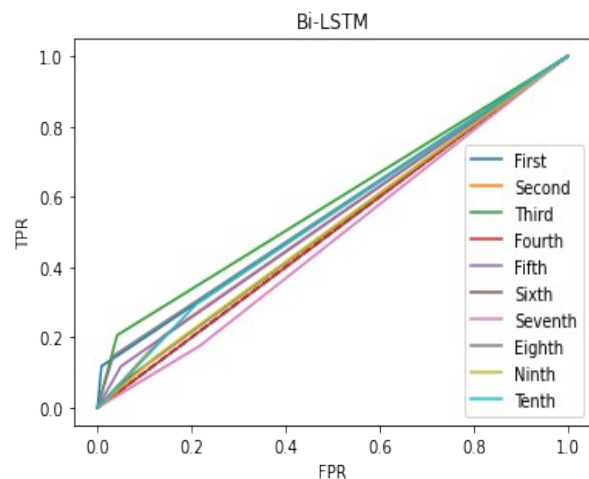
Μοντέλα	AUC	STD AUC
Random Forest	0.58	0.05
GBC	0.56	0.04
XGBoost	0.58	0.04
SVM	0.69	0.04
KNN	0.64	0.03
LSTM	0.54	0.04
Bi-LSTM	0.54	0.03

Πίνακας 44: Επαναλαμβανόμενη Επικύρωση (roc) – near miss

Συγκεκριμένα για τα μοντέλα LSTM και Bi-LSTM παρουσιάζονται οι γραφικές της μετρικής ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ :



Εικόνα 45: Roc - Near Miss



Εικόνα 46: Roc - Near Miss

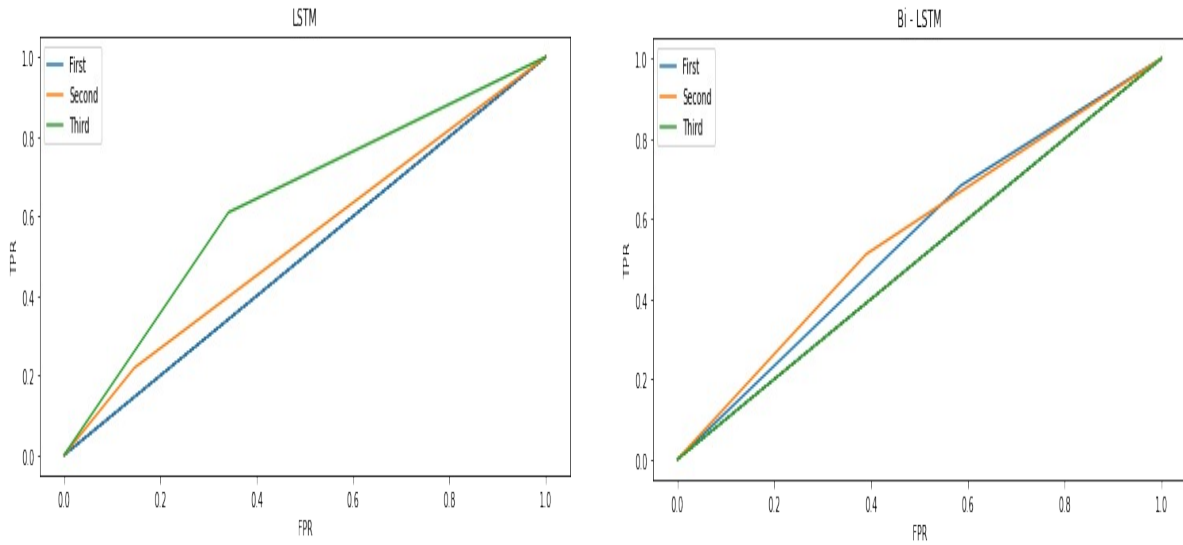
Παρατίθενται οι κατά μέσο όρο μετρικές της ‘ακρίβειας’ και της ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ με την διακύμανση τους για την επικύρωση των μοντέλων όπως προέκυψαν από την τεχνική Ensemble Learning:

### Ensemble Learning

Μοντέλα	Accuracy	Precision	Recall	AUC
LSTM	0.59	0.59	0.59	0.59
Bi-LSTM	0.56	0.56	0.56	0.56

Πίνακας 47: Ensemble Learning

Συγκεκριμένα για τα μοντέλα LSTM και Bi-LSTM παρουσιάζονται οι γραφικές της μετρικής ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ :



Εικόνα 48: Roc – Ensemble Learning

Εικόνα 49: Roc - Ensemble Learning

Παρατίθενται οι μετρικές των μοντέλων LSTM και Bi-LSTM για την επικύρωση των μοντέλων όπως προέκυψαν από την τεχνική Leave-one-out:

#### Leave-one-out

Μοντέλα	Accuracy	Precision	Recall	ROC
LSTM	0.74	0.76	0.74	0.72
Bi-LSTM	0.73	0.77	0.73	0.70

Πίνακας 50: Leave-one-out

## Κεφάλαιο 4.6 Σύγκριση με υπάρχουσα βιβλιογραφία

Παρατίθενται οι κατά μέσο όρο μετρικές της ‘ακρίβειας’ και της ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ με την διακύμανση τους για την επικύρωση των μοντέλων όπως προέκυψαν από την τεχνική Repeated k-fold για τεχνική εξισορρόπησης smote:

### Επαναλαμβανόμενη Επικύρωση (ακρίβεια) – Smote

Μοντέλα	Accuracy	STD Accuracy
Random Forest	0.69	0.04
GBC	0.73	0.05
XGBoost	0.69	0.04
SVM	0.75	0.05
KNN	0.66	0.06
LSTM	0.77	0.03
Bi-LSTM	0.77	0.05

Πίνακας 51: Repeated k-fold (ακρίβεια) – Smote

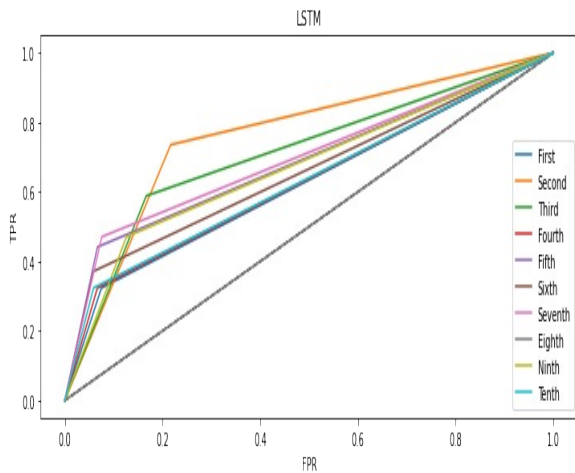
### Επαναλαμβανόμενη Επικύρωση (roc) – Smote

Μοντέλα	AUC	STD AUC
Random Forest	0.79	0.03
GBC	0.77	0.03
XGBoost	0.76	0.03
SVM	0.75	0.04
KNN	0.70	0.04
LSTM	0.68	0.06
Bi-LSTM	0.69	0.06

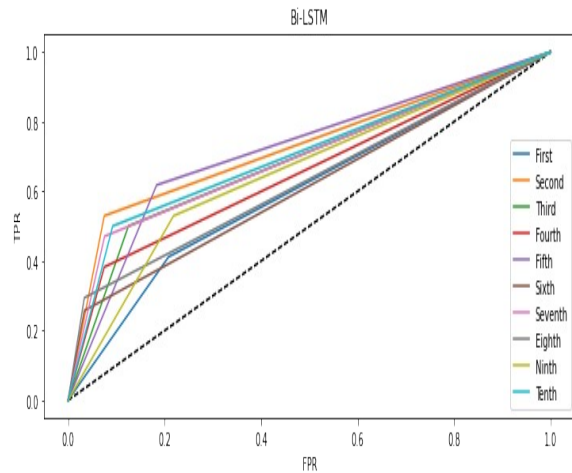
Πίνακας 52: Repeated k-fold (roc)– Smote



Συγκεκριμένα για τα μοντέλα LSTM και Bi-LSTM παρουσιάζονται οι γραφικές της μετρικής ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ :



Εικόνα 53: Roc - Smote



Εικόνα 54: Roc –Smote

## Κεφάλαιο 4.7 Γενίκευση σε άλλη βάση δεδομένων

Παρατίθενται οι μετρικές των μοντέλων LSTM και Bi-LSTM θεωρώντας ως δεδομένα εκπαίδευσης την βάση δεδομένων από το έργο Coswara και ως δεδομένα ελέγχου την βάση δεδομένων από το πανεπιστήμιο Cambridge αξιοποιώντας την τεχνική SMOTE:

### Γενίκευση των μοντέλων σε άλλη βάση δεδομένων με χρήση SMOTE

Μοντέλα	Accuracy	Precision	Recall	AUC
LSTM	0.71	0.67	0.71	0.47
Bi-LSTM	0.71	0.65	0.71	0.49

Πίνακας 55: Γενίκευση των μοντέλων σε άλλη βάση - SMOTE

Παρατίθενται οι μετρικές των μοντέλων LSTM και Bi-LSTM θεωρώντας ως δεδομένα εκπαίδευσης την βάση δεδομένων από το έργο Coswara και ως δεδομένα ελέγχου την βάση δεδομένων από το πανεπιστήμιο Cambridge αξιοποιώντας την τεχνική Random Undersampling & SMOTE:

### Γενίκευση των μοντέλων σε άλλη βάση δεδομένων - Random Undersampling & SMOTE

Μοντέλα	Accuracy	Precision	Recall	AUC
LSTM	0.73	0.69	0.73	0.47
Bi-LSTM	0.69	0.65	0.69	0.51

Πίνακας 56: Γενίκευση των μοντέλων σε άλλη βάση - Random Undersampling & SMOTE

## Κεφάλαιο 5<sup>ο</sup>: Συζήτηση αποτελεσμάτων και συμπεράσματα

Ολοκληρώνοντας την παρουσίαση των αποτελεσμάτων, έπεται η συζήτηση αναφορικά με αυτά. Συνοπτικά, η διαδικασία που ακολουθήθηκε αποτελούταν αρχικά από την επιλογή ντετερμινιστικών μοντέλων επιτηρούμενης μηχανικής μάθησης. Τα ντετερμινιστικά μοντέλα δέντρων απόφασης, μηχανής διανυσματικής υποστήριξης και κοντινότερων γειτόνων συντέλεσαν στην ανάλυση χαρακτηριστικών και στην εύρεση των καλύτερων τεχνικών εξισορρόπησης δεδομένων. Με αυτόν τον τρόπο, επιτεύχθηκε η βέλτιστη αξιοποίηση των υπολογιστικών πόρων, δηλαδή, αύξηση της ταχύτητας και μείωση της μνήμης. Ταυτόχρονα, επιλέχθηκαν δύο επιπλέον μοντέλα, LSTM και Bi-LSTM, που θεωρήθηκαν κατάλληλα για την ανάλυση μας. Το LSTM βασίζεται πάνω στην νοοτροπία του επαναλαμβανόμενου νευρωνικού δικτύου, γεγονός που το κατατάσσει στην κατηγορία των μοντέλων μηχανικής μάθησης που διαχειρίζονται διαδοχικά δεδομένα. Το Bi-LSTM αποτελεί επέκταση του LSTM. Χρησιμοποιώντας Bi-LSTM, τροφοδοτείται ο αλγόριθμος εκμάθησης με τα αρχικά δεδομένα μία φορά από την αρχή μέχρι το τέλος, όπως και στο LSTM, αλλά και ακόμη μία φορά από το τέλος στην αρχή. Το βασικό χαρακτηριστικό είναι ότι αυτά τα δίκτυα μπορούν να αποθηκεύουν πληροφορίες που μπορούν να χρησιμοποιηθούν για μελλοντική επεξεργασία κυψελών.

Στη συνέχεια, επιλέχθηκαν χαρακτηριστικά από την βιβλιοθήκη της python librosa που κατορθώνουν να περιγράψουν τις ιδιότητες του ήχου (βήχα), με σκοπό τον διαχωρισμό του βήχα ενός θετικού από έναν αρνητικό σε COVID-19. Τα χαρακτηριστικά αυτά αναλύθηκαν πλήρως στο τρίτο κεφάλαιο και επιγραμματικά είναι τα τριάντα πρώτα mfccs, μηδενικός ρυθμός διασταύρωσης, φασματικό κέντρο, φασματική μετατόπιση, φασματικό εύρος ζώνης, επίπεδο ενέργειας, φασματική επιπεδότητα και διακύμανση. Στην ανάλυση θα αξιοποιηθούν όλα τα προαναφερθέντα χαρακτηριστικά μέχρι την επιλογή των καλύτερων.

Την εύρεση των χαρακτηριστικών ακολουθεί η επιλογή κατάλληλης μεθόδου κανονικοποίησης. Στην παρούσα διπλωματική θα τεθούν υπό εξέταση δύο μέθοδοι κανονικοποίησης, η τυποποίηση και η ομαλοποίηση. Η κανονικοποίηση κρίθηκε αναγκαία γιατί τα χαρακτηριστικά μετρούνται σε διαφορετικές κλίμακες, με συνέπεια να μην συνεισφέρουν εξίσου στην ανάλυση και μπορεί να καταλήξουν να δημιουργήσουν μια μεροληψία. Σε αυτό το στάδιο δεν μας αφορά η ανάλυση των μοντέλων, για αυτό τον λόγο χρησιμοποιήθηκαν οι αρχικές παράμετροι που ορίζει η βιβλιοθήκη της python. Τα δεδομένα έχουν σπάσει με τυχαίο τρόπο θέτοντας τα δεδομένα εκπαίδευσης στο 70% και τα δεδομένα ελέγχου στο 30%. Το σπάσιμο είναι ίδιο και για τις δύο μεθόδους για την καλύτερη σύγκρισή τους. Κατά βάση, τα συμπεράσματα ήταν τα αναμενόμενα. Πιο συγκεκριμένα, τα μοντέλα που ακολουθούν την νοοτροπία των δέντρων απόφασης δεν επηρεάζονται από τεχνικές κανονικοποίησης. Όπως φαίνεται από σύγκριση των Πινάκων 1 και 2 στα μοντέλα random forest, extreme gradient boosting, gradient boost classifier, οι μετρικές δεν αλλάζουν σχεδόν καθόλου. Το μοντέλο support vector machine δεν μεταβάλλει την συμπεριφορά του, παρόλο που είθισται να έχει καλύτερη απόδοση με την μέθοδο της τυποποίησης. Στον αντίποδα έρχονται τα υπόλοιπα τρία μοντέλα που αποδίδουν καλύτερα με την μέθοδο της ομαλοποίησης. Με εξαίρεση το μοντέλο k nearest neighbors που η

επίδοση παρουσιάζει μια μικρή βελτίωση, τα άλλα δύο μοντέλα (lstm και bi-lstm) κατορθώνουν να βελτιωθούν έως και 7%.

Έχοντας καταλήξει σε μέθοδο κανονικοποίησης, κρίθηκε αναγκαία η βελτιστοποίηση των παραμέτρων των μοντέλων. Η βελτιστοποίηση έγινε με κριτήριο τις μετρικές της ακρίβειας και της περιοχής κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη. Για τα ντετερμινιστικά μοντέλα (random forest, extreme gradient boosting, gradient boost classifier, support vector machine, k nearest neighbors) υλοποιήθηκε αναζήτηση πλέγματος. Η αναζήτηση πλέγματος μπορεί να θεωρηθεί ως μια εξαντλητική αναζήτηση για την επιλογή ενός μοντέλου. Στην αναζήτηση πλέγματος, δημιουργείτε ένα πλέγμα τιμών υπερπαραμέτρων και για κάθε συνδυασμό, εκπαιδεύεται ένα μοντέλο που βαθμολογεί τα δεδομένα δοκιμών. Σε αυτή την προσέγγιση, δοκιμάζεται κάθε συνδυασμός τιμών υπερπαραμέτρων που μπορεί να είναι πολύ αναποτελεσματικός. Με εξαίρεση το μοντέλο gradient boost classifier του οποίου η απόδοση ελαφρώς μειώθηκε, στα υπόλοιπα μοντέλα η αναζήτηση πλέγματος επέφερε καλύτερα αποτελέσματα. Συνεπώς, θα διατηρηθούν οι τιμές που προέκυψαν από την βελτιστοποίηση για την υπόλοιπη ανάλυση. Όσον αφορά τα νευρωνικά μοντέλα ακολουθήθηκε η νοοτροπία δοκιμή και λάθος. Έγιναν αρκετές δοκιμές και επιλέχθηκαν οι παράμετροι ανάλογα με το αποτέλεσμα των μετρικών. Για λόγους ταχύτητας τα δεδομένα έχουν σπάσει με τυχαίο τρόπο, θέτοντας τα δεδομένα εκπαίδευσης στο 70% και τα δεδομένα ελέγχου στο 30%. Το σπάσιμο είναι ίδιο και για όλες τις δοκιμές. Αρχικά επιλέχθηκαν ως κριτήριο το binary crossentropy και ως βελτιστοποιητής ορίστηκε ο adam, αφού αποτελούν τις καταλληλότερες επιλογές για δυαδική κατηγοριοποίηση. Παράλληλα, αξιοποιήθηκε η τεχνική του early stopping με σκοπό την αποφυγή υπερεκπαίδευσης των μοντέλων. Στην συνέχεια, δοκιμάστηκε ένα νευρωνικό μοντέλο με δύο επίπεδα με χαμηλό πλήθος νευρώνων. Οι μετρικές του μοντέλου, που παρουσιάζονται στον Πίνακα 3, είναι σχετικά χαμηλές και σίγουρα επιδέχονται βελτίωση. Η αύξηση, λοιπόν, του πλήθους των νευρώνων επέφερε βελτίωση στην επίδοση των μοντέλων κατά περίπου 12% όπως παρατηρείται μετά την σύγκριση των Πινάκων 3 και 5. Δοκιμάστηκε περαιτέρω βελτίωση των μοντέλων αλλάζοντας τις συναρτήσεις ενεργοποίησης του δεύτερου επιπέδου σε relu. Η δοκιμή αυτή οδήγησε σε αύξηση της επίδοσης κατά μέσο όρο κατά 3.5%. Η προσθήκη ενός επιπλέον επιπέδου με μικρό πλήθος νευρώνων δεν βελτίωσε το μοντέλο. Αντιθέτως, ο Πίνακας 7 δείχνει ότι η μετρική της 'περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη' μειώθηκε. Το ίδιο αποτέλεσμα επέφερε και η προσθήκη ενός Dropout επιπέδου. Οι καταλληλότερες παράμετροι θεωρήθηκαν αυτές που επέφεραν τα αποτελέσματα του πίνακα 5, για αυτό και στην υπόλοιπη ανάλυση θα χρησιμοποιηθούν αυτές.

Με το πέρας της βελτιστοποίησης των παραμέτρων, έπεται η εξισορρόπηση των δεδομένων. Γίνεται, λοιπόν, μια διερεύνηση για την εύρεση της καταλληλότερης τεχνικής για εξισορρόπηση των κλάσεων. Η διερεύνηση αυτή κρίνεται αναγκαία για την αποφυγή της μεροληψίας των μοντέλων υπέρ της μιας κλάσης. Αρχικά χρησιμοποιήθηκε η τεχνική Near Miss, η οποία μαζί με την random undersampling και smoteen τεχνική επέφεραν τα χειρότερα αποτελέσματα συγκριτικά με τις άλλες τρεις μεθόδους. Με την αναφορά 'χειρότερα' εννοείται χαμηλότερα ποσοστά σε ακρίβεια και σε περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη. Λαμβάνονται υπόψιν φυσικά και οι

μετρικές precision και recall για την τελική απόφαση. Για την περίπτωση των δύο τεχνικών αυτών δίνεται η εντύπωση ‘τυχειότητας’ αφού τα ποσοστά τους σε πολλές περιπτώσεις αγγίζουν το 50%. Εξάιρεση αποτελεί το μοντέλο k nearest neighbors, το οποίο επιτυγχάνει την καλύτερη επίδοση του μοντέλου σε σχέση με όλες τις τεχνικές εξισορρόπησης, αγγίζοντας ακόμα και το 69%. Εν συνεχεία, οι δύο επόμενες υπό μελέτη τεχνικές random undersampling και smote, random undersampling και adasyn, παρουσιάζουν πολύ κοντινά αποτελέσματα και την καλύτερη επίδοση κατά μέσο όρο σε όλα τα μοντέλα. Πιο συγκεκριμένα, η τεχνική random undersampling και smote έχει μια γενικότερα καλή επίδοση σε όλα τα μοντέλα με τα ποσοστά της να κυμαίνονται μεταξύ του 63% και 68%. Από την άλλη πλευρά η τεχνική random undersampling και adasyn κατορθώνει καλύτερα αποτελέσματα στα βασικά μοντέλα της ανάλυσης LSTM και Bi-LSTM, με τα καλύτερα ποσοστά για όλες τις μετρικές. Ικανοποιητικά, βέβαια θεωρούνται τα αποτελέσματα και για την τεχνική random undersampling και smote to mek.

Ολοκληρώνοντας την διερεύνηση κατάλληλης τεχνικής εξισορρόπησης κλάσεων, ακολουθεί η επιλογή των καλύτερων χαρακτηριστικών για την εκπαίδευση των μοντέλων. Η επιλογή βασίστηκε στην συνάρτηση feature importance των ντετερμινιστικών μοντέλων. Στην αρχή, χρησιμοποιήθηκαν τα δέκα καλύτερα χαρακτηριστικά του κάθε μοντέλου ξεχωριστά. Τα καλύτερα, όμως, αποτελέσματα που παρατηρήθηκαν προέκυψαν από την επιρροή όλων των μοντέλων για την επιλογή των καταλληλότερων. Δηλαδή, για κάθε μοντέλο προέκυψε μια λίστα με τα δέκα καλύτερα χαρακτηριστικά, συνοδευόμενα από ένα ποσοστό επιρροής. Οι επιμέρους λίστες προστέθηκαν σε μία τελική που περιλάμβανε τα χαρακτηριστικά αυτά ταξινομένα χωρίς διπλότυπα. Μετά από μερικές δοκιμές για το προτιμότερο πλήθος των χαρακτηριστικών επιλέχθηκαν να είναι τα έντεκα με το μεγαλύτερο ποσοστό. Η επιρροή της επιλογής των καταλληλότερων χαρακτηριστικών οδήγησε στα καλύτερα αποτελέσματα μέχρι στιγμής, δίνοντας ακόμα και 77% ακρίβεια με τις υπόλοιπες μετρικές να είναι από 70% και πάνω όπως διακρίνεται στον Πίνακα 15.

Έχοντας τελειώσει με την επιλογή των μεθόδων, των καλύτερων χαρακτηριστικών και την βελτιστοποίηση των μοντέλων, μένει το τελικό στάδιο της επικύρωσης των αποτελεσμάτων. Η επικύρωση των αποτελεσμάτων έγινε με τρεις τεχνικές : repeated k-fold, ensemble learning, leave – one – out. Όσον αφορά με την τεχνική repeated k-fold, τα αποτελέσματα προέκυψαν ύστερα από δέκα τυχαία ‘σπασίματα’ όπου το τελικό ποσοστό που αναφέρετε στους πίνακες αποτελεί τον μέσο όρο αυτών των δέκα ‘σπασμάτων’. Σε κάθε σπάσιμο υλοποιούνται δύο τεχνικές εξισορρόπησης δεδομένων. Χρησιμοποιήθηκε μία από τις τεχνικές με τα καλύτερα αποτελέσματα όπως προέκυψε από την προηγούμενη ανάλυση και μία από τις τεχνικές με τα χειρότερα. Αυτό εφαρμόστηκε με σκοπό να ελεγχθεί αν η προηγούμενη ανάλυση επεκτείνεται πέρα του αρχικού τυχαίου σπασίματος. Η τεχνική αυτή οδήγησε, όντως, στα ίδια αποτελέσματα με την πρότερη ανάλυση. Δηλαδή, όταν συνδυάζεται με την τεχνική random undersampling και smote εξακολουθεί να δίνει συγκριτικά για κάθε μοντέλο και μετρική ξεχωριστά καλύτερα ποσοστά. Την καλύτερη επίδοση έχει το SVM, με ποσοστά στο 77% και 80% σε ακρίβεια και σε περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη αντίστοιχα. Ακολουθούν το Bi-LSTM και το LSTM υστερώντας ελαφρώς από άποψη περιοχής κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη, δίνοντας ένα ποσοστό κοντά στο 70%. Παρατηρώντας

βέβαια τις εικόνες 23 και 24 φαίνεται για το Bi-LSTM να συγκλίνουν σε μία τιμή οι γραφικές της μετρικής ‘περιοχή κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ με μία μόνο εξαίρεση που αγγίζει το 50% και αύξησε την διακύμανση. Συνεπώς το Bi-LSTM είναι ελαφρώς καλύτερα από το LSTM.

Παράλληλα, εφαρμόστηκε η τεχνική ensemble learning σε μια προσπάθεια να αυξηθεί η ακρίβεια. Στην προκειμένη τεχνική διατηρήθηκαν σε ξεχωριστό πίνακα μεγέθους 82 τα δεδομένα ελέγχου. Τα δεδομένα αυτά ήταν πλήρως ισορροπημένα, διατηρώντας 41 καταγραφές από κάθε κλάση με τυχαίο τρόπο. Τα υπόλοιπα δεδομένα της κλάσης πλειοψηφίας χωρίστηκαν σε 3 και καθένα από αυτά συνδυασμένα με τα υπόλοιπα δεδομένα της κλάσης μειοψηφίας και δόθηκαν στα μοντέλα LSTM και Bi-LSTM για εκπαίδευση. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 22. Όπως φαίνεται, η επίδοση των μοντέλων δεν είναι επαρκής καθιστώντας αδύνατο να χρησιμοποιηθεί αυτή η τεχνική για πρόβλεψη. Αντιθέτως, η τεχνική leave – one – out κατόρθωσε να βελτιώσει την επίδοση του μοντέλου, με τις μετρικές να έχουν τιμές μεγαλύτερες του 70%. Ακόμα και η μετρική της περιοχής κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη ανέβηκε για πρώτη φορά στο 70% για τα μοντέλα LSTM και Bi-LSTM. Η τεχνική αυτή έχει συνδυαστεί με την τεχνική random undersampling και smote για κάθε σπάσιμο, με σκοπό την εξισορρόπηση των κλάσεων.

Εφόσον επαληθεύτηκε η ακρίβεια των μοντέλων, ακολουθεί η σύγκριση με ήδη υπάρχουσα βιβλιογραφία. Οι μετρικές στον Πίνακα 27 προέκυψαν από έρευνα για την κατηγοριοποίηση του βήχα μέσω μηχανικής μάθησης [25]. Οι μετρικές της έρευνας είναι σαφώς καλύτερες από αυτές της παραπάνω ανάλυσης. Η σύγκριση γίνεται με τους Πίνακες 20 και 21, αφού χρησιμοποιήθηκε η ίδια τεχνική για επικύρωση και εξισορρόπηση των κλάσεων. Να επισημανθεί ότι χρησιμοποιήθηκε η ίδια βάση δεδομένων, απλώς λόγω διαφορετικής χρονικής περιόδου αξιοποίησης της ήταν λιγότερα τα δεδομένα στην περίπτωση της έρευνας. Όσον αφορά την κλάση μειοψηφίας, στην έρευνα τα δεδομένα ήταν υποτριπλάσια, γεγονός που μπορεί να εξηγεί αυτή την τεράστια διαφορά της τάξης του 20%. Επιπλέον, τα μοντέλα έχουν εκπαιδευτεί με διαφορετικά χαρακτηριστικά. Η έρευνα αξιοποίησε τη μέθοδο Sequential Feature Selector από την βιβλιοθήκη της python και κατέληξε σε 13 καλύτερα χαρακτηριστικά.

<b>Μοντέλα</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>AUC</b>
LSTM (1)	0.97	0.91	0.94	0.94
LSTM (2)	0.97	0.90	0.94	0.93

Πίνακας 27: Αποτελέσματα έρευνας [25]

Τέλος, κρίθηκε απαραίτητος ο έλεγχος αν μπορούν τα μοντέλα να γενικεύσουν την σχετικά καλή επίδοση τους και σε άλλη βάση. Η βάση που αξιοποιήθηκε για τον σκοπό αυτό είναι τα δεδομένα Cambridge. Η βάση αυτή περιλαμβάνει 269 υγιείς και 61 COVID-19 ασθενείς. Δεν χρησιμοποιήθηκε για εκπαίδευση λόγω του μικρού πλήθους δεδομένων. Τα αποτελέσματα παρατίθενται στους Πίνακες 25 και 26. Παρατηρώντας την μετρική της ‘περιοχής κάτω από τη χαρακτηριστική καμπύλη λειτουργίας του δέκτη’ φαίνεται ότι τα νευρωνικά δίκτυα αδυνατούν να γενικεύσουν σε άλλη βάση.

## Κεφάλαιο 6<sup>ο</sup>: Επίλογος-Προτάσεις για μελλοντική έρευνα

### 6.1 Επίλογος

Το ξέσπασμα του COVID-19 είχε βαθιά επίδραση στην ευημερία των ανθρώπων παγκοσμίως και στον αριθμό των θανάτων που σχετίζονται με αυτών συνεχίζουν να αυξάνονται παγκοσμίως. Παρόλο που η τεχνολογία έχει εισχωρήσει στην καθημερινότητά μας με μεγάλη επιτυχία, ειδικά σε μηχανική μάθηση και βαθιά μάθηση. Η τεχνητή νοημοσύνη, γενικότερα, συνέβαλε στην υποστήριξη αυτής της δύσκολης μάχης κατά του COVID-19.

Η παρούσα διπλωματική εργασία αναφέρεται σε μια παρουσίαση ενός ταξινομητή βήχα COVID-19 που βασίζεται στη μηχανική μάθηση, ο οποίος μπορεί να διακρίνει τον θετικό COVID-19 βήχα τόσο από τον αρνητικό COVID-19 όσο και από τον υγιή βήχα που καταγράφεται σε κινητά. Για την εκπαίδευση και την αξιολόγηση των ταξινομητών, εμείς έχουμε χρησιμοποιήσει δύο σύνολα δεδομένων. Το πρώτο, μεγαλύτερο, σύνολο δεδομένων είναι δημόσια διαθέσιμο και περιέχει δεδομένα από το έργο Coswara που περιλαμβάνει 341 θετικά για τον COVID-19 και 1195 υγιή άτομα. Ενώ το δεύτερο μικρότερο σύνολο δεδομένων συλλέχθηκε κυρίως στο Cambridge και περιέχει 61 θετικά για COVID-19 και 269 αρνητικά για τον COVID-19. Οι βάσεις δεδομένων που χρησιμοποιούνται σε αυτή τη μελέτη περιλαμβάνουν άτομα και από τις έξι ηπείρους και περιέχουν τόσο αναγκαστικό όσο και φυσικό βήχα, υποδεικνύοντας ότι η προσέγγιση είναι ευρέως εφαρμόσιμη. Και τα δύο σύνολα δεδομένων υποδεικνύουν ότι όσοι διαγνώστηκαν θετικοί σε COVID-19 είναι 20%-30% μικρότεροι από όσους έχουν διαγνωστεί αρνητικοί σε COVID-19. Η ανισορροπία του συνόλου δεδομένων αντιμετωπίστηκε με εφαρμογή τεχνικών για την εξισορρόπηση των κλάσεων. Χρησιμοποιήθηκε ένας τυχαίος διαχωρισμός δεδομένων εκπαίδευσης (70%) – ελέγχου (30%) για τον συντονισμό των παραμέτρων και την εύρεση της καλύτερης τεχνικής εξισορρόπησης των κλάσεων επτά ταξινομητών μηχανικής μάθησης: τυχαίο δάσος (RF), XGBoost (XGB), ταξινομητής ενίσχυσης κλίσης (GBC), k-πλησιέστερος γείτονας (KNN), μηχανή υποστήριξης διανυσμάτων (SVM), μακροχρόνια βραχυπρόθεσμη μνήμη (LSTM) και αμφίδρομη μακροπρόθεσμη μνήμη (Bi-LSTM). Με βάση την συνάρτηση των ντετερμινιστικών ταξινομητών feature importance καθορίστηκαν τα καλύτερα χαρακτηριστικά για την εκπαίδευση των μοντέλων μειώνοντας την μνήμη και τον χρόνο εκτέλεσης. Όσον αφορά την επικύρωση της απόδοσης των μοντέλων, εφαρμόστηκαν επαναλαμβανόμενες τεχνικές k-fold και leave-one-out, συνδυασμένες με τα αποτελέσματα της προηγούμενης ανάλυσης για τον έλεγχο της μέχρι τώρα διαδικασίας. Τα αποτελέσματα δείχνουν ότι παρόλο που όλοι οι ταξινομητές μπόρεσαν να αναγνωρίσουν τον βήχα COVID-19, την καλύτερη απόδοση παρουσίασε ο ταξινομητής Bi-LSTM, ο οποίος ήταν καλύτερος σε θέση να διακρίνει μεταξύ του θετικού για τον COVID-19 και του υγιούς βήχα με περιοχή κάτω από το ROC καμπύλη (AUC) 0.70. Ένας ταξινομητής LSTM μπόρεσε να διακρίνει καλύτερα τον βήχα θετικό για τον COVID-19 και τον αρνητικό COVID-19, με AUC 0.68 μετά την επιλογή των καλύτερων 11 χαρακτηριστικών με την βοήθεια της feature importance συνάρτησης των ντετερμινιστικών μοντέλων μηχανικής μάθησης. Δεδομένου ότι αυτός ο τύπος προληπτικού ελέγχου χωρίς επαφή, είναι εύκολος στην εφαρμογή και μπορεί να μειώσει τον φόρτο εργασίας στα κέντρα δοκιμών καθώς και να περιορίσει τη μετάδοση συνιστώντας έγκαιρη αυτο-απομόνωση σε όσους έχουν διαγνωστεί με COVID-19.

## 6.2 Προβληματισμοί

Μια ανησυχία είναι ότι οι αλγόριθμοι μηχανικής μάθησης μπορούν απλώς να διακρίνουν μεταξύ υγιών ατόμων και ατόμων που δεν είναι γενικά καλά, αντί να ανιχνεύουν τον ίδιο τον COVID-19. Η διάκριση μεταξύ υγιών και άρρωστων ατόμων είναι μια πολύ πιο εύκολη αλλά λιγότερο χρήσιμη εργασία. Αν και ορισμένοι ερευνητές έχουν διερευνήσει αυτό το ζήτημα, για παράδειγμα, κατασκευάζοντας εργασίες στις οποίες εκπαιδευμένα μοντέλα ταξινομούν άτομα με άλλες ασθένειες που σχετίζονται με το αναπνευστικό και άτομα με COVID-19, καμία μελέτη δεν μπόρεσε να καταλήξει στο συμπέρασμα ότι ο ίδιος ο COVID-19 είναι αυτός που ταυτοποιείται.

Η περιβαλλοντική διαφθορά είναι μια άλλη ανησυχία. Όταν εγγράφεται ένα ηχητικό συμβάν, όπως ένας βήχας, στο αρχείο ήχου περιλαμβάνονται επίσης πληροφορίες για το ακουστικό περιβάλλον. Ο θόρυβος του περιβάλλοντος εισάγει μια πιθανότητα μεροληψίας στο σύνολο δεδομένων ήχου. Η μεροληψία από ακούσιες προσθήκες ήχου μπορεί να υπάρχει με πολλές μορφές. Ένα παράδειγμα θα μπορούσε να είναι ότι τα θετικά στον COVID-19 άτομα είναι πιο πιθανό να βρίσκονται σε εσωτερικούς χώρους, ίσως σε ιατρικό περιβάλλον τη στιγμή της καταγραφής, από εκείνα χωρίς COVID-19, ενώ τα άτομα που είναι αρνητικά στον COVID-19 είναι πιο πιθανό να είναι έξω ή ίσως σε περιβάλλον εργασίας από εκείνους με COVID-19. Κάθε ρύθμιση έχει τη δική της μοναδική, αναγνωρίσιμη λειτουργία περιβαλλοντικού ήχου. Οποιαδήποτε τέτοια συσχέτιση μπορεί να αναγνωριστεί από τα μοντέλα μηχανικής μάθησης, καθώς προσπαθεί να κάνει μια πρόβλεψη για τον COVID-19. Αυτή η μεροληψία μπορεί να είναι λεπτή και, όταν υπάρχει στο σύνολο δεδομένων, είναι πολύ δύσκολο να εντοπιστεί και να αφαιρεθεί. Αυτή η λεπτότητα θα μπορούσε να εξηγήσει γιατί δεν έχει αντιμετωπιστεί ως περιορισμός σε καμία από τις μέχρι τώρα μελέτες. Η μελλοντική έρευνα θα πρέπει να προσπαθήσει να ελέγξει το περιβάλλον ήχου τη στιγμή της εγγραφής για να αντιμετωπίσει αυτό το ζήτημα.

Ένα άλλο ζήτημα είναι η εγκυρότητα των συνόλων δεδομένων. Οι εποπτευόμενες μέθοδοι μηχανικής εκμάθησης βασίζονται στην ακριβή επισήμανση αληθείας των περιπτώσεων. Το προκύπτον εργαλείο μηχανικής εκμάθησης μπορεί να είναι τόσο ακριβές όσο αυτές οι ετικέτες. Ωστόσο, η εγκυρότητα των ετικετών που χρησιμοποιούνται στα σύνολα δεδομένων ήχου COVID-19 που έχουν συλλεχθεί μέχρι στιγμής είναι αμφίβολη, επειδή τα περισσότερα σύνολα δεδομένων επιτρέπουν στους συμμετέχοντες να αναφέρουν μόνοι τους την κατάστασή τους COVID-19 και αποτυγχάνουν να καταγράψουν τους συμμετέχοντες που βασίζουν την καταγραφή τους στην αλυσιδωτή αντίδραση πολυμέρισης (PCR) ή, της λιγότερο ακριβής, πλευρικής ροής έλεγχος. Αν και ορισμένες μελέτες απαιτούν μια επικυρωμένη δοκιμή PCR, τα σύνολα δεδομένων είναι μικρά και κανένα δεν έχει δημοσιοποιηθεί τη στιγμή της σύνταξης. Για να επισημανθεί η σοβαρότητα αυτού του προβλήματος, σημειώνεται ότι ορισμένα σύνολα δεδομένων έχουν αποδεχθεί την αυτοαξιολόγηση ως μέσο για την επισήμανση του συνόλου δεδομένων, και άλλα απέτυχαν να διευκρινίσουν ακόμη και πώς καθορίστηκε η κατάσταση του COVID-19.

### 6.3 Μελλοντικά Βήματα

Στο μέλλον, προτείνεται να γίνει περισσότερη δουλειά για να ξεπεραστούν οι περιορισμοί της τρέχουσας μεθόδου. Πρώτον, πρέπει να συλλεχθούν περισσότερα δεδομένα για την εκπαίδευση του μοντέλου. Δεν είναι μόνο σημαντικό να αυξηθεί ο αριθμός των συλλεγόμενων εγγραφών, αλλά υπάρχει επίσης ανάγκη να συλλεχθούν δείγματα από διαφορετικές χώρες, περιοχές, ηλικίες, φύλα και εθνότητες για να βελτιωθεί η γενίκευση του μοντέλου και να βελτιωθεί περαιτέρω η ακρίβεια μέσω της εκπαίδευσης το μοντέλο με περισσότερα δεδομένα. Δεύτερον, χρησιμοποιήθηκαν χαρακτηριστικά αποκλειστικά σχετιζόμενα με τις ιδιότητες του ήχου για την εκπαίδευση του μοντέλου. Θα μπορούσαν να δοκιμαστεί η εισαγωγή συμπτωμάτων ως χαρακτηριστικά του μοντέλου. Με την αξιοποίηση της τεχνικής one hot encoding συνδυασμένη με PCA θα αποτελέσει μια ελπιδοφόρα λύση για την αύξηση της επίδοσης του μοντέλου. Επιπλέον, θα μπορούσε να εφαρμοστεί κάποια μέθοδος αποθορυβοποίησης. Ένα φίλτρο κατάλληλο να διατηρεί την βασική πληροφορία του βήχα θα συντελέσει σε πιο ακριβή δεδομένα για την εκπαίδευση του μοντέλου. Τέλος, χρήσιμο θα ήταν να χρησιμοποιηθεί για την εκπαίδευση βάση δεδομένων με καταχώρηση από γιατρούς και ειδικούς για το αν ο ασθενής πάσχει ή όχι από COVID-19. Αφήνοντας την καταχώρηση και την εκπόνηση της διαδικασίας στον ασθενή αυξάνονται οι πιθανότητες για λάθη που οδηγούν με την σειρά τους σε λανθασμένη εκπαίδευση του μοντέλου.



## Βιβλιογραφία

- [1] T. P. Velavan and C. G. Meyer, “The COVID-19 epidemic,” *Trop. Med. Int. Heal.*, vol. 25, no. 3, pp. 278–280, Mar. 2020, doi: 10.1111/TMI.13383.
- [2] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth, “How will country-based mitigation measures influence the course of the COVID-19 epidemic?,” *Lancet*, vol. 395, no. 10228, pp. 931–934, Mar. 2020, doi: 10.1016/S0140-6736(20)30567-5.
- [3] J. Whitworth, “COVID-19: a fast evolving pandemic,” *Trans. R. Soc. Trop. Med. Hyg.*, vol. 114, no. 4, p. 241, Apr. 2020, doi: 10.1093/TRSTMH/TRAA025.
- [4] “Home - Johns Hopkins Coronavirus Resource Center.” <https://coronavirus.jhu.edu/> (accessed Oct. 27, 2021).
- [5] “COVID Live Update: 245,437,543 Cases and 4,982,178 Deaths from the Coronavirus - Worldometer.” <https://www.worldometers.info/coronavirus/#countries> (accessed Oct. 27, 2021).
- [6] J. Y. Wick, “COVID-19,” *Pharmacother. A Multimed. Learn. Resour.*, May 2017, doi: 10.21019/PHARMACOTHERAPYFIRST.COVID19\_OVERVIEW.
- [7] H.-Y. Wang, X.-L. Li, Z.-R. Yan, X.-P. Sun, J. Han, and B.-W. Zhang, “Potential neurological symptoms of COVID-19:,” <https://doi.org/10.1177/1756286420917830>, vol. 13, Mar. 2020, doi: 10.1177/1756286420917830.
- [8] “Beijing hospital confirms nervous system infections by novel coronavirus-China.org.cn.” [http://www.china.org.cn/china/2020-03/05/content\\_75777888.htm](http://www.china.org.cn/china/2020-03/05/content_75777888.htm) (accessed Oct. 27, 2021).
- [9] “What Is... Mobile Health? - SSA.” <https://www.addiction-ssa.org/knowledge-hub/what-is-mobile-health/> (accessed Oct. 27, 2021).
- [10] L. Chiau *et al.*, “Mobile Health Apps on COVID-19 Launched in the Early Days of the Pandemic: Content Analysis and Review,” *JMIR Mhealth Uhealth* 2020;8(9)e19796 <https://mhealth.jmir.org/2020/9/e19796>, vol. 8, no. 9, p. e19796, Sep. 2020, doi: 10.2196/19796.
- [11] “Coronavirus Self-Checker | CDC.” <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/coronavirus-self-checker.html> (accessed Oct. 27, 2021).
- [12] “Mobile Apps for Coronavirus (COVID-19): See the List.” <https://www.webmd.com/lung/coronavirus-apps#2> (accessed Oct. 27, 2021).
- [13] A. T. Azar and S. M. El-Metwally, “Decision tree classifiers for automated medical diagnosis,” *Neural Comput. Appl.*, vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013, doi: 10.1007/S00521-012-1196-7.
- [14] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random Forests and Decision Trees,” 2012, Accessed: Oct. 27, 2021. [Online]. Available: [www.IJCSI.org](http://www.IJCSI.org).
- [15] “Understanding Gradient Boosting Machines | by Harshdeep Singh | Towards Data Science.” <https://towardsdatascience.com/understanding-gradient-boosting-machines->

9be756fe76ab (accessed Nov. 01, 2021).

- [16] O. Sagi and L. Rokach, “Approximating XGBoost with an interpretable decision tree,” *Inf. Sci. (Ny)*, vol. 572, pp. 522–542, Sep. 2021, doi: 10.1016/J.INS.2021.05.055.
- [17] “Python Decision Tree Classification with Scikit-Learn DecisionTreeClassifier - DataCamp.” <https://www.datacamp.com/community/tutorials/decision-tree-classification-python> (accessed Oct. 27, 2021).
- [18] “Support Vector Machine - an overview | ScienceDirect Topics.” <https://www.sciencedirect.com/topics/computer-science/support-vector-machine> (accessed Oct. 27, 2021).
- [19] H. Raeisi Shahraki, S. Pourahmad, and N. Zare, “K Important Neighbors: A Novel Approach to Binary Classification in High Dimensional Data,” *Biomed Res. Int.*, vol. 2017, 2017, doi: 10.1155/2017/7560807.
- [20] A. Hassan, I. Shahin, and M. B. Alsabek, “COVID-19 Detection System using Recurrent Neural Networks,” *Proc. 2020 IEEE Int. Conf. Commun. Comput. Cybersecurity, Informatics, CCCI 2020*, Nov. 2020, doi: 10.1109/CCCI49893.2020.9256562.
- [21] “Complete Guide To Bidirectional LSTM (With Python Codes).” <https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/> (accessed Oct. 27, 2021).
- [22] “Illustrated Guide to LSTM’s and GRU’s: A step by step explanation | by Michael Phi | Towards Data Science.” <https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21> (accessed Oct. 27, 2021).
- [23] “Neural Networks for Real-Time Audio: Stateful LSTM | by Keith Bloemer | Towards Data Science.” <https://towardsdatascience.com/neural-networks-for-real-time-audio-stateful-lstm-b534babeae5d> (accessed Oct. 27, 2021).
- [24] “Understanding of LSTM Networks - GeeksforGeeks.” <https://www.geeksforgeeks.org/understanding-of-lstm-networks/> (accessed Oct. 27, 2021).
- [25] M. Pahar, M. Klopper, R. Warren, and T. Niesler, “COVID-19 Cough Classification using Machine Learning and Global Smartphone Recordings,” Accessed: Oct. 27, 2021. [Online]. Available: <https://coswara.iisc.ac.in>.
- [26] “Project Coswara | IISc.” <https://coswara.iisc.ac.in/> (accessed Oct. 27, 2021).
- [27] “COVID-19 Sounds App - University of Cambridge.” <https://www.covid-19-sounds.org/el/> (accessed Oct. 27, 2021).
- [28] C. Brown *et al.*, “Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 3474–3484, Jun. 2020, doi: 10.1145/3394486.3412865.
- [29] B. Lange, D. Li, E. Nehoran, E. Tuzhilina, and M. Lu, “Early Detection of COVID-19 from Cough Sounds, Symptoms, and Context Machine Learning / Signal Processing Sub-Team CS 472: Data science and AI for COVID-19.”
- [30] G. Sharma, K. Umapathy, and S. Krishnan, “Trends in audio signal feature extraction methods,” *Appl. Acoust.*, vol. 158, p. 107020, Jan. 2020, doi:

10.1016/J.APACOUST.2019.107020.

- [31] N. Madhu, “Note on measures for spectral flatness,” *Electron. Lett.*, vol. 45, no. 23, pp. 1195–1196, 2009, doi: 10.1049/EL.2009.1977.
- [32] N. Japkowicz and S. Stephen, “The class imbalance problem: A systematic study,” *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, 2002, doi: 10.3233/IDA-2002-6504.
- [33] Q. Wang, Z. H. Luo, J. C. Huang, Y. H. Feng, and Z. Liu, “A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM,” *Comput. Intell. Neurosci.*, vol. 2017, 2017, doi: 10.1155/2017/1827016.
- [34] “Undersampling Algorithms for Imbalanced Classification.” <https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/> (accessed Oct. 27, 2021).
- [35] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” *Proc. Int. Jt. Conf. Neural Networks*, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.
- [36] I. D. S. Miranda, A. H. Diacon, and T. R. Niesler, “A Comparative Study of Features for Acoustic Cough Detection Using Deep Architectures\*.”
- [37] “ADASYN: Adaptive Synthetic Sampling Method for Imbalanced Data | by Kushal Vala | Towards Data Science.” <https://towardsdatascience.com/adasyn-adaptive-synthetic-sampling-method-for-imbalanced-data-602a3673ba16> (accessed Oct. 27, 2021).
- [38] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinforma. 2013 141*, vol. 14, no. 1, pp. 1–16, Mar. 2013, doi: 10.1186/1471-2105-14-106.
- [39] “ML | Handling Imbalanced Data with SMOTE and Near Miss Algorithm in Python - GeeksforGeeks.” <https://www.geeksforgeeks.org/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/> (accessed Oct. 27, 2021).
- [40] “Feature Scaling | Standardization Vs Normalization.” <https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/> (accessed Oct. 28, 2021).
- [41] “3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.0.1 documentation.” [https://scikit-learn.org/stable/modules/cross\\_validation.html](https://scikit-learn.org/stable/modules/cross_validation.html) (accessed Oct. 27, 2021).
- [42] “LOOCV for Evaluating Machine Learning Algorithms.” <https://machinelearningmastery.com/loocv-for-evaluating-machine-learning-algorithms/> (accessed Oct. 27, 2021).