



Multi-class Classification of Lung Conditions from Chest X-ray Images



Νευρωνικά Δίκτυα (8^ο εξάμηνο)

Φοιτητής:

Δαδακίδης Γιώργος (iis22127)

Επιβλέπων Καθηγητής: Πρωτοπαπαδάκης Ευτύχιος

-Ιούνιος 2025-

Περίληψη

Η παρούσα εργασία επικεντρώνεται στην ανάπτυξη και αξιολόγηση μοντέλων βαθιάς μάθησης (Deep Learning) για την πολυκατηγορική ταξινόμηση ακτινογραφιών θώρακα, με στόχο τη διάγνωση ασθενειών όπως COVID-19, Viral Pneumonia, Lung Opacity και Normal. Το σύνολο δεδομένων που χρησιμοποιήθηκε ήταν το ευρέως διαδεδομένο COVID-19 Radiography Database, το οποίο περιέχει περισσότερες από 21.000 εικόνες. Αρχικά κατασκευάστηκε από το μηδέν ένα μοντέλο CNN και στη συνέχεια συγκρίθηκε η απόδοσή του με τέσσερις τεχνικές μεταφοράς μάθησης (Transfer Learning): MobileNetV2, DenseNet121, EfficientNetB0 και VGG19.

Η πειραματική διαδικασία περιλάμβανε τεχνικές προεπεξεργασίας, data augmentation, χρήση 4-Fold Stratified Cross Validation και αξιολόγηση μετρικών απόδοσης όπως Accuracy, Precision, Recall, F1-score και ROC-AUC. Τα αποτελέσματα έδειξαν ότι το μοντέλο VGG19 υπερείχε σε ταχύτητα και απόδοση, ενώ το CNN από το μηδέν εμφάνισε αξιοσημείωτα αποτελέσματα δεδομένης της απλότητάς του. Η εργασία καταλήγει σε προτάσεις για περαιτέρω βελτιστοποίηση των μοντέλων και διερεύνηση σε πιο σύνθετα σύνολα δεδομένων.

Περιεχόμενα

1. Εισαγωγή.....	5
2. Σχετική βιβλιογραφία	6
3. Προτεινόμενη μεθοδολογία	8
3.1 Απόκτηση Δεδομένων.....	9
3.2 Προεπεξεργασία Δεδομένων.....	9
3.2.1 Έλεγχος και Αφαίρεση Κατεστραμμένων Εικόνων	9
3.2.2 Δημιουργία Δομής Δεδομένων.....	10
3.2.3 Ομογενοποίηση Διαστάσεων και One Hot Encoding	10
3.3 Διερεύνηση Dataset.....	10
3.4 Data Augmentation.....	12
3.5 Αρχιτεκτονική CNN.....	13
3.6 Εφαρμογή Transfer Learning	14
3.7 Εκπαίδευση και Αξιολόγηση Μοντέλων.....	15
4. Πειραματικά αποτελέσματα	20
4.1 Απόδοση Ανά Μοντέλο.....	20
4.2 Συγκριτική Ανάλυση.....	32
5. Συμπεράσματα.....	35
6. Βιβλιογραφία	36

Λίστα Εικόνων

Figure 1: Κατανομή πλήθους εικόνων ανά κλάση.....	11
Figure 2: Εικόνες από κάθε κλάση του Dataset.....	12
Figure 3: Κατανομή δεδομένων μετά τη διαδικασία διάσπασης – Fold 1.....	16
Figure 4: Κατανομή δεδομένων μετά τη διαδικασία διάσπασης – Fold 2.....	16
Figure 5: Κατανομή δεδομένων μετά τη διαδικασία διάσπασης – Fold 3.....	16
Figure 6: Κατανομή δεδομένων μετά τη διαδικασία διάσπασης – Fold 4.....	17
Figure 7: Καμπύλη Train/Validation Accuracy για το CNN - Fold 1	21
Figure 8: Καμπύλη Train/Validation Loss για το CNN - Fold 1	22
Figure 9: CNN Confusion Matrix - Fold 1.....	22
Figure 10: CNN Roc Curve για κάθε κλάση	23
Figure 11: Καμπύλη Train/Validation Accuracy για το MobileNetV2 - Fold 1	24
Figure 12: Καμπύλη Train/Validation Loss για το MobileNetV2 - Fold 1.....	24
Figure 13: MobileNetV2 Confusion Matrix - Fold 1	25
Figure 14: MobileNetV2 Roc Curve για κάθε κλάση.....	26
Figure 15: Καμπύλες Train/Validation Accuracy και Loss για το DenseNet121 – Fold 1.....	26
Figure 16: DenseNet121 Confusion Matrix - Fold 1.....	27
Figure 17: DenseNet121 Roc Curve για κάθε κλάση	27
Figure 18: Καμπύλες Train/Validation Accuracy και Loss για το EfficientNetB0 – Fold 1	28
Figure 19: EfficientNetB0 Confusion Matrix - Fold1	29
Figure 20: EfficientNetB0 Roc Curve για κάθε κλάση.....	29
Figure 21: Καμπύλες Train/Validation Accuracy και Loss για το VGG19 – Fold 1	30
Figure 22: VGG19 Confusion Matrix - Fold 1	31
Figure 23: VGG19 Roc Curve για κάθε κλάση	31
Figure 24: Συγκριτική Απεικόνιση Μετρικών Απόδοσης για Κάθε Μοντέλο	32
Figure 25: Χρόνος εκπαίδευσης και αξιολόγησης ανά μοντέλο	33
Figure 26: Απεικόνιση Απόδοσης VGG19 ανά κλάση και μετρική	34

Λίστα Πινάκων

Table 1: Αριθμός εικόνων ανα κατηγορία πριν και μετά τον έλεγχο ακεραιότητας.....	10
Table 2: CNN Hyperparameters	14
Table 3: Μετρικές απόδοσης για κάθε μοντέλο	32
Table 4: Μετρικές απόδοσης για κάθε κλάση (VGG19)	34

1. Εισαγωγή

Η παγκόσμια εξάπλωση της νόσου COVID-19 στα τέλη του 2019 άσκησε τεράστια πίεση στα συστήματα υγειονομικής περίθαλψης παγκοσμίως, προκαλώντας την επιτακτική ανάγκη για την ανάπτυξη ταχέων και αξιόπιστων διαγνωστικών εργαλείων. Αν και πολλοί από τους ασθενείς με COVID-19 παρέμειναν ασυμπτωματικοί ή εμφάνισαν ήπια συμπτώματα, ένα σημαντικό ποσοστό εμφάνισε σοβαρές αναπνευστικές επιπλοκές, οι οποίες σε ορισμένες περιπτώσεις οδήγησαν σε θάνατο (Abul Bashar et al., 2021).

Οι παραδοσιακές μέθοδοι όπως η PCR, παρότι είναι ακριβείς, απαιτούν χρόνο και σημαντικούς πόρους. Αντιθέτως, η ιατρική απεικόνιση, και ειδικότερα οι ακτινογραφίες θώρακα (Chest X-rays – CXR), προσφέρουν μια πιο προσβάσιμη και οικονομικά αποδοτική μέθοδο προληπτικού ελέγχου. Ωστόσο, η οπτική ομοιότητα της COVID-19 με άλλες πνευμονικές παθήσεις, όπως η ιογενής πνευμονία ή η αδιαφάνεια των πνευμόνων, καθιστά τη διάγνωση ιδιαίτερα απαιτητική ακόμη και για έμπειρους ακτινολόγους (Lamouadene et al., 2021).

Η Τεχνητή Νοημοσύνη (Artificial Intelligence – AI) και ειδικότερα οι μέθοδοι βαθιάς μάθησης (Deep Learning – DL), όπως τα συνελκτικά νευρωνικά δίκτυα (Convolutional Neural Networks – CNNs), έχουν δείξει σημαντικές δυνατότητες στην υποστήριξη διαγνωστικών διαδικασιών μέσω αυτοματοποιημένης ανάλυσης ιατρικών εικόνων. Οι μέθοδοι αυτές έχουν αποδειχθεί ιδιαίτερα χρήσιμες στην ταξινόμηση εικόνων θώρακα από CT και ακτινογραφίες, βοηθώντας στον έγκαιρο εντοπισμό ενδείξεων της COVID-19 και στην αξιολόγηση της σοβαρότητας της νόσου (Abu Sayeed et al., 2023). Τα DL συστήματα είναι ικανά να επεξεργαστούν μεγάλους όγκους δεδομένων και να επιτύχουν ταχύτητα και ακρίβεια πρόβλεψης που ξεπερνούν τις ανθρώπινες δυνατότητες. Παρ' όλα αυτά, υπάρχουν ακόμα περιορισμοί στις προηγούμενες μελέτες — όπως προκαταλήψεις στα datasets, ανισορροπία μεταξύ των κλάσεων και η δυσκολία διάκρισης αλληλεπικαλυπτόμενων παθήσεων — που ενισχύουν την ανάγκη για συνεχιζόμενη έρευνα ως προς τη γενικευσιμότητα και την ανθεκτικότητα των μοντέλων (Lamouadene et al., 2021).

Στην παρούσα εργασία, στόχος μας είναι η ανάπτυξη και σύγκριση διαφόρων μεθόδων βαθιάς μάθησης για την πολυκατηγορική ταξινόμηση ακτινογραφιών θώρακα σε τέσσερις διαγνωστικές κατηγορίες: Normal, COVID-19, Lung Opacity και Viral Pneumonia. Το dataset που χρησιμοποιήθηκε είναι το COVID-19 Radiography Database, μια δημόσια διαθέσιμη συλλογή με πάνω από 21.000 CXR εικόνες, η οποία δημοσιεύτηκε τον Μάρτιο του 2021 και έχει καθιερωθεί ως σημείο αναφοράς για πλήθος σχετικών μελετών.

Η προσέγγισή μας περιλαμβάνει δύο βασικά σκέλη:

- Ένα CNN μοντέλο εκπαιδευμένο εξ ολοκλήρου από την αρχή, με χρήση 4-Fold Stratified Cross-Validation για την εκπαίδευση και επικύρωση.
- Μια σειρά από Transfer Learning (TL) μοντέλα, όπου έγινε fine-tuning σε προεκπαιδευμένες συνελκτικές αρχιτεκτονικές όπως οι MobileNetV2, VGG19,

EfficientNetB0 και DenseNet121, με κοινές συνθήκες αξιολόγησης ως προς το baseline CNN.

Για την αξιολόγηση των μοντέλων χρησιμοποιείται ένα σύνολο μετρικών απόδοσης: Accuracy, Precision, Recall, F1 Score και ROC-AUC. Επιπρόσθετα, οπτικοποιούνται οι confusion matrices και οι ROC curves, προκειμένου να αναλυθεί η συμπεριφορά του κάθε μοντέλου ανά κατηγορία και να εντοπιστούν πιθανά μοτίβα σφαλμάτων.

Ο συνολικός στόχος της μελέτης είναι η σύγκριση της αποτελεσματικότητας της εκπαίδευσης CNNs από το μηδέν έναντι της εφαρμογής transfer learning με προεκπαιδευμένα μοντέλα, για την ταξινόμηση πνευμονικών παθήσεων από ακτινογραφίες, και η ανάδειξη των αρχιτεκτονικών που προσφέρουν τη μεγαλύτερη ακρίβεια και αποδοτικότητα σε ρεαλιστικές συνθήκες.

2. Σχετική βιβλιογραφία

Η πανδημία COVID-19 προκάλεσε παγκόσμια κρίση στην υγειονομική περίθαλψη, αναδεικνύοντας την ανάγκη για ταχεία, ακριβή και προσβάσιμη διάγνωση. Οι X-ray εικόνες θώρακα έχουν χρησιμοποιηθεί εκτενώς για τον εντοπισμό αναπνευστικών λοιμώξεων λόγω της χαμηλού κόστους και ευκολίας απόκτησής τους σε σχέση με πιο ακριβές τεχνικές όπως η CT. Η Deep Learning τεχνολογία, και ιδιαίτερα τα Convolutional Neural Networks (CNNs), έχουν φανεί υποσχόμενα για την αυτόματη ανίχνευση νόσων σε ιατρικές εικόνες.

Η πανδημία COVID-19 προκάλεσε παγκόσμια κρίση στην υγειονομική περίθαλψη, αναδεικνύοντας την ανάγκη για ταχεία, ακριβή και προσβάσιμη διάγνωση. Η χρήση αλγορίθμων τεχνητής νοημοσύνης, και συγκεκριμένα deep learning τεχνικών όπως τα Convolutional Neural Networks (CNNs), έχει αποδειχθεί ιδιαίτερα αποτελεσματική στον τομέα της ιατρικής απεικόνισης. Ιδιαίτερα σε περιόδους πανδημιών, όπως αυτή του COVID-19, τα CNNs έχουν αξιοποιηθεί ευρέως για την ανάλυση εικόνων ακτινογραφιών θώρακος (Chest X-rays), με στόχο τη γρήγορη και αυτοματοποιημένη διάγνωση μεταξύ κατηγοριών όπως: COVID-19, Lung Opacity, Viral Pneumonia και φυσιολογικές περιπτώσεις (Normal).

Στην εργασία του Rahman et al. (2021), πραγματοποιείται σύγκριση 315 deep learning μοντέλων για την ταξινόμηση COVID-19, Normal και Pneumonia από X-ray εικόνες. Η μελέτη καταδεικνύει ότι το DenseNet201 σε συνδυασμό με SVM προσφέρει accuracy 98.16%, recall 98.93% υπογραμμίζοντας την αποτελεσματικότητα των CNN-based models για το έργο αυτό.

Η μελέτη των Lamouadene et al. (2025) αξιολόγησε την αποτελεσματικότητα διαφόρων μοντέλων μηχανικής μάθησης στην ταξινόμηση ακτινογραφιών θώρακα, χρησιμοποιώντας το COVID-19 Radiography Database, το οποίο περιλαμβάνει εικόνες από τέσσερις κατηγορίες: COVID-19, Lung Opacity, Viral Pneumonia και Normal. Στην αρχική προσέγγιση χρησιμοποιήθηκε ένα Support Vector Machine (SVM), το οποίο πέτυχε accuracy 68%, υποδεικνύοντας ότι απαιτούνται πιο σύνθετες αρχιτεκτονικές για την αποτελεσματική διάκριση των κλάσεων.

Η ανάπτυξη ενός προσαρμοσμένου CNN με συνδυασμό τριών συνελκτικών επιπέδων, ενός επιπέδου pooling και πέντε πλήρως συνδεδεμένων επιπέδων είχε ως αποτέλεσμα accuracy 83% και f1-score 83%. Ωστόσο, τα σημαντικότερα αποτελέσματα επιτεύχθηκαν μέσω Transfer Learning, όπου το μοντέλο EfficientNet-CNN με Adam optimizer πέτυχε accuracy 99.20% για την πολυκατηγορική ταξινόμηση των τεσσάρων παραπάνω κατηγοριών. Οι συγγραφείς αναγνωρίζουν ότι η επίδοση του μοντέλου επηρεάζεται από τη μη ισορροπημένη κατανομή των δεδομένων, καθώς και από τη διαφορετικότητα των απεικονίσεων, υπογραμμίζοντας την ανάγκη για περισσότερα ομοιογενή και ισορροπημένα σύνολα δεδομένων σε μελλοντικές μελέτες.

Στην έρευνα της Ahuja et al. (2020), προτείνεται μοντέλο ResNet50 με transfer learning για ανίχνευση COVID-19, υποστηρίζοντας την ιδέα ότι μοντέλα προεκπαιδευμένα σε ImageNet μπορούν να επιτύχουν εξαιρετική απόδοση ακόμη και με περιορισμένα ιατρικά δεδομένα, εφόσον γίνει σωστή προεπεξεργασία και ενίσχυση δεδομένων (data augmentation).

Η μελέτη των Bunyodbek Ibrokhimov et al. (20223) αξιολόγησε τις αρχιτεκτονικές VGG19 και ResNet50 για τη διάγνωση COVID-19 και άλλων τύπων πνευμονίας μέσω X-ray εικόνων, χρησιμοποιώντας transfer learning και τεχνικές parent-child parallel computing για επιτάχυνση της εκπαίδευσης. Το VGG19 πέτυχε accuracy 96.6%, ROC-AUC 0.98 και precision 0.99 στην κατηγορία COVID-19, αποδεικνύοντας ανώτερη απόδοση συγκριτικά με το ResNet50, σε ένα εκτεταμένο dataset 33.920 δειγμάτων (COVID-QU-Ex).

Τέλος, η εργασία Abul Bashaar et al. (2020) παρουσιάζεται βελτιστοποιημένη προσέγγιση βαθιάς μάθησης, όπου εφαρμόστηκε transfer learning σε πέντε CNN (AlexNet, GoogleNet, VGG16, VGG19 και DenseNet) στο dataset COVID-19 Radiography. Το καλύτερο αποτέλεσμα πέτυχε το VGG16, με accuracy 95,63 %, precision 99,18 %, recall 98,78 % και F1-score 98,98 %, υπερέχοντας των υπόλοιπων μοντέλων. Οι συγγραφείς επισημαίνουν ότι η επίδοση μπορεί να βελτιωθεί περαιτέρω εφόσον αντιμετωπιστεί η ανισορροπία των τάξεων και η ετερογένεια του συνόλου δεδομένων.

Από τη σύνθεση της σχετικής βιβλιογραφίας, προκύπτουν τα εξής συμπεράσματα:

- Τα CNNs, σε συνδυασμό με τεχνικές Transfer Learning, έχουν αποδειχθεί ιδιαίτερα αποτελεσματικά για την αυτόματη διάγνωση της νόσου COVID-19 μέσω ακτινογραφιών θώρακα, προσφέροντας υψηλά ποσοστά ακρίβειας και αξιοπιστίας.
- Η ενσωμάτωση προηγμένων αρχιτεκτονικών, όπως οι ResNet, DenseNet, EfficientNet και VGG (ιδίως VGG16 και VGG19), συμβάλλει σημαντικά στη βελτίωση της απόδοσης των διαγνωστικών μοντέλων, ιδίως όταν αυτές συνδυάζονται με ταξινομητές όπως το Support Vector Machine (SVM).
- Παρά τη σημαντική πρόοδο, εξακολουθούν να υπάρχουν προκλήσεις, όπως η ανισορροπία μεταξύ των κατηγοριών (class imbalance), η μορφολογική ομοιότητα του COVID-19 με άλλες αναπνευστικές λοιμώξεις (όπως η πνευμονία), καθώς και ο

περιορισμένος όγκος ομοιογενών και ποιοτικά ελεγχόμενων δεδομένων, γεγονός που επιδρά αρνητικά στη γενικευσιμότητα των μοντέλων.

3. Προτεινόμενη μεθοδολογία

Η παρούσα εργασία εστιάζει στην ανάλυση ακτινογραφιών θώρακα για την ανίχνευση και ταξινόμηση παθήσεων του αναπνευστικού. Για τον σκοπό αυτό χρησιμοποιείται το dataset *COVID-19 Radiography Database* ως multi-class classification πρόβλημα με κλάσεις Normal, COVID-19, Viral Pneumonia, Lung Opacity. Το dataset αποτελείται από 21165 εικόνες απο φωτογραφίες θώρακα.

Όλες οι δοκιμές πραγματοποιήθηκαν στο περιβάλλον Google Colab με χρήση T4 GPU αξιοποιώντας ένα ευρύ φάσμα βιβλιοθηκών της Python για την υποστήριξη των πειραμάτων. Στόχος του κεφαλαίου είναι η ενίσχυση της διαφάνειας και της δυνατότητας αναπαραγωγής των αποτελεσμάτων.

Οι βασικές βιβλιοθήκες της Python που αξιοποιήθηκαν στο πλαίσιο της μελέτης είναι οι εξής:

- **TensorFlow** και **Keras**: για την υλοποίηση, εκπαίδευση και αξιολόγηση νευρωνικών δικτύων (CNN, transfer learning).
- **Scikit-learn**: για διαχωρισμό δεδομένων (train-test split, StratifiedKFold), υπολογισμό μετρικών (accuracy, precision, recall, F1-score) και δημιουργία confusion matrices.
- **Pandas** και **NumPy**: για διαχείριση και επεξεργασία δεδομένων.
- **Matplotlib** και **Seaborn**: για οπτικοποίηση αποτελεσμάτων και δημιουργία γραφημάτων.
- **pickle**: για αποθήκευση και ανάκτηση ιστορικού εκπαίδευσης (training history).
- **os, shutil, pathlib**: για διαχείριση αρχείων και καταλόγων στο Google Drive και τοπική μνήμη του Colab.

Η μελέτη προτείνει μια μεθοδολογία η οποία συγκρίνει ένα συνελικτικό νευρωνικό δίκτυο (CNN) κατασκευασμένο εξ ολοκλήρου από την αρχή με τέσσερις τεχνικές transfer learning, αξιοποιώντας τα προεκπαιδευμένα μοντέλα: MobileNetV2, VGG19, EfficientNetB0 και DenseNet121. Ο στόχος είναι η εκτίμηση της απόδοσης κάθε μοντέλου ως προς τη δυνατότητα ακριβούς ταξινόμησης των εικόνων.

Η προτεινόμενη διαδικασία περιλαμβάνει την:

- Προεπεξεργασία των δεδομένων
- Διερεύνηση του dataset
- Εφαρμογή data augmentation για τη βελτίωση της γενίκευσης των μοντέλων
- Κατασκευή αρχιτεκτονικής CNN
- Εφαρμογή Transfer Learning με προεκπαιδευμένα μοντέλα

- Εκπαίδευση και αξιολόγηση κάθε μοντέλου με βάση μετρικές απόδοσης όπως accuracy, precision, recall, F1-score και ROC-AUC
- καθώς και τη συγκριτική ανάλυση των αποτελεσμάτων με στόχο την ανάδειξη της πιο αποτελεσματικής προσέγγισης.

3.1 Απόκτηση Δεδομένων

Για την παρούσα μελέτη χρησιμοποιήθηκε το δημόσια διαθέσιμο σύνολο δεδομένων *Covid-19 Radiography Database* το οποίο είναι αναρτημένο στην πλατφόρμα [Kaggle](#). Το συγκεκριμένο σύνολο δεδομένων επιλέχθηκε λόγω της πληρότητας του, της ευρείας χρήσης του στη σχετική βιβλιογραφία και της δυνατότητας του να υποστηρίξει πολυκατηγορική ταξινόμηση ακτινογραφιών θώρακα.

Για την αξιοποίηση των δεδομένων στο περιβάλλον Google Colab, το dataset αποθηκεύτηκε αρχικά στο Google Drive και στη συνέχεια μεταφέρθηκε στην τοπική μνήμη του Colab για ταχύτερη πρόσβαση, χρησιμοποιώντας τη βιβλιοθήκη shutil. Με αυτόν τον τρόπο επιτυγχάνεται αποτελεσματική διαχείριση αρχείων και μειώνεται ο χρόνος ανάγνωσης κατά τη διάρκεια της εκπαίδευσης των μοντέλων.

3.2 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία δεδομένων αποτελεί ένα κρίσιμο στάδιο σε προβλήματα μηχανικής μάθησης και ειδικά στην ταξινόμηση ιατρικών εικόνων, καθώς διασφαλίζει την ποιότητα, συνέπεια και αξιοπιστία των δεδομένων εισόδου. Το σύνολο δεδομένων *COVID-19 Radiography Database* περιλαμβάνει χιλιάδες ακτινογραφίες θώρακα από διαφορετικές πηγές, με πιθανές αποκλίσεις ως προς τη μορφή, το μέγεθος ή την ακεραιότητα των αρχείων εικόνας. Η σωστή προεπεξεργασία εξασφαλίζει την απομάκρυνση ακατάλληλων εικόνων, τη μετατροπή των δεδομένων σε ενιαία μορφή, και την αποφυγή σφαλμάτων κατά τη φάση της εκπαίδευσης του μοντέλου, συμβάλλοντας έτσι στην γενικότερη ακρίβεια και σταθερότητα της μεθόδου.

3.2.1 Έλεγχος και Αφαίρεση Κατεστραμμένων Εικόνων

Κατά τη φάση κατασκευής του dataset, πραγματοποιήθηκε έλεγχος ακεραιότητας αρχείων προκειμένου να διασφαλιστεί ότι όλες οι εικόνες μπορούν να διαβαστούν και να επεξεργαστούν κανονικά από τον αλγόριθμο. Χρησιμοποιήθηκαν κατάλληλοι μηχανισμοί διαχείρισης εξαιρέσεων (try-except) ώστε να παρακαμφθούν εικόνες που ήταν κατεστραμμένες ή αδυνατούσαν να αποκωδικοποιηθούν από το σύστημα. Ως αποτέλεσμα, από τις 21.165 αρχικές εικόνες, 48 εικόνες απορρίφθηκαν ως μη αναγνώσιμες, με αποτέλεσμα να απομείνουν 21.117 έγκυρες εικόνες, οι οποίες και χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση των μοντέλων. Αυτή η διαδικασία βοήθησε στην αποτροπή σφαλμάτων κατά την εκτέλεση και βελτίωσε τη συνολική ποιότητα του dataset.

Στον Πίνακα 1 παρουσιάζεται η κατανομή των εικόνων ανά κατηγορία πριν και μετά την απομάκρυνση κατεστραμμένων ή μη αναγνώσιμων αρχείων, καθώς και ο συνολικός αριθμός των εικόνων που διαγράφηκαν από κάθε κλάση.

Table 1: Αριθμός εικόνων ανά κατηγορία πριν και μετά τον έλεγχο ακεραιότητας.

Class	Count(Before)	Count(After)	Deleted
Normal	10192	10159	33
Lung Opacity	6012	6012	0
COVID-19	3616	3601	15
Viral Pneumonia	1345	1345	0
Total	21165	21117	48

3.2.2 Δημιουργία Δομής Δεδομένων

Μετά τη φόρτωση των εικόνων, ακολουθεί η δημιουργία ενός DataFrame, ο οποίος περιλαμβάνει τα πλήρη μονοπάτια (file paths) των εικόνων και τις αντίστοιχες ετικέτες κατηγορίας (labels). Κατά τη διάρκεια αυτής της διαδικασίας, αγνοήθηκαν φάκελοι που περιείχαν επιπλέον πληροφορίες, όπως οι μάσκες (masks), οι οποίοι δεν είναι απαραίτητοι για την παρούσα ανάλυση και θα μπορούσαν να επηρεάσουν την εκπαίδευση των μοντέλων.

3.2.3 Ομογενοποίηση Διαστάσεων και One Hot Encoding

Όλες οι εικόνες μετατράπηκαν σε κοινές διαστάσεις 224×224 pixels με σκοπό τη συμβατότητα με τα Convolutional Neural Networks που χρησιμοποιήθηκαν στην παρούσα μελέτη (είτε από το μηδέν είτε μέσω transfer learning). Παράλληλα, οι ετικέτες των εικόνων, που αντιστοιχούν στις 4 κλάσεις, μετατράπηκαν σε μορφή one-hot encoding μέσω της παραμέτρου `class_mode='categorical'` των ImageDataGenerators.

Αυτή η μετατροπή έχει ως αποτέλεσμα κάθε ετικέτα να αναπαρίσταται ως διανυσματική δυαδική ακολουθία (π.χ., [0, 1, 0, 0]), όπου κάθε θέση αντιστοιχεί σε μία κατηγορία. Η προσέγγιση αυτή είναι απαραίτητη για τη χρήση της συνάρτησης κόστους `categorical_crossentropy`, που απαιτεί ως είσοδο one-hot αναπαραστάσεις, και ταιριάζει με την έξοδο των μοντέλων που διαθέτει softmax ενεργοποίηση στο τελευταίο επίπεδο.

Η ομογενοποίηση διαστάσεων και η μετατροπή ετικετών αποτελούν κρίσιμα βήματα για τη σταθερότητα, τη συγκρισιμότητα και την αποτελεσματική εκπαίδευση των μοντέλων βαθιάς μάθησης.

3.3 Διερεύνηση Dataset

Η διερεύνηση του συνόλου δεδομένων αποτελεί βασικό πυλώνα για την κατανόηση της φύσης και της κατανομής των δειγμάτων μεταξύ των κατηγοριών. Η άνιση κατανομή των δεδομένων μπορεί να οδηγήσει σε προκατάληψη εκπαίδευσης (bias), επηρεάζοντας αρνητικά

την ικανότητα γενίκευσης του μοντέλου. Για τον λόγο αυτό, πραγματοποιήθηκε απεικόνιση της συχνότητας εμφάνισης των εικόνων ανά κατηγορία, όπως φαίνεται στο Figure 1.

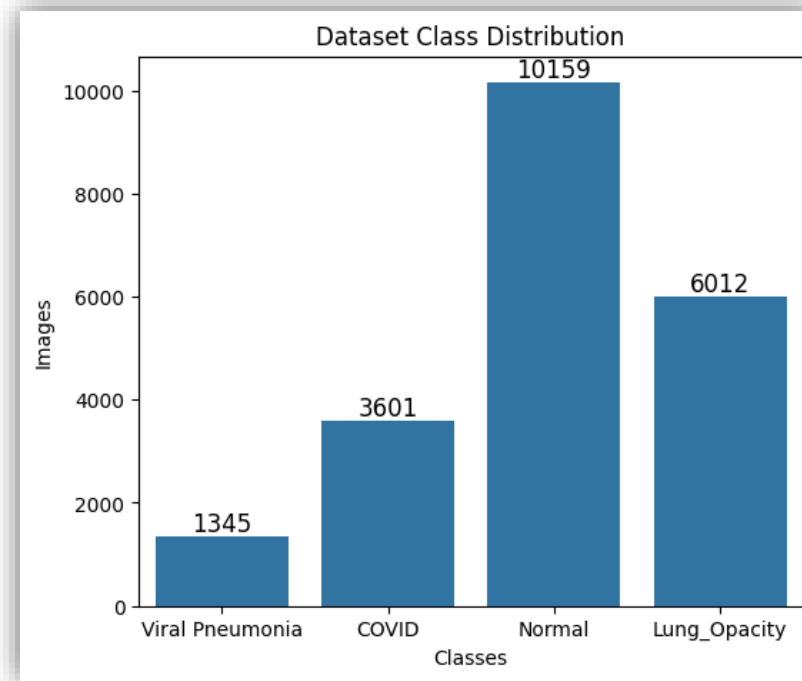


Figure 1: Κατανομή πλήθους εικόνων ανά κλάση

Το dataset περιλαμβάνει τέσσερις κατηγορίες: Normal, COVID-19, Lung Opacity, και Viral Pneumonia. Όπως παρατηρείται, υπάρχει σημαντική ανισορροπία υπέρ της κατηγορίας Normal (10.159 εικόνες), ενώ η κατηγορία Viral Pneumonia διαθέτει τον μικρότερο αριθμό δειγμάτων (1.345 εικόνες). Η κατηγορία COVID-19 περιέχει 3.601 εικόνες και η Lung Opacity 6.012. Αυτή η ανομοιογένεια στη διανομή των κλάσεων καθιστά αναγκαία τη χρήση κατάλληλων τεχνικών όπως data augmentation, ώστε να ελαχιστοποιηθούν τα φαινόμενα υπερπροσαρμογής (overfitting) προς τις πλειοψηφικές κατηγορίες.

Για την ποιοτική κατανόηση του dataset, στο Σχήμα 2 παρουσιάζεται ένα ενδεικτικό δείγμα εικόνας για κάθε μία από τις τέσσερις κατηγορίες του συνόλου δεδομένων. Όλες οι εικόνες είναι ακτινογραφίες θώρακα (Chest X-rays), όπου παρατηρούνται οπτικές διαφορές μεταξύ των παθολογικών καταστάσεων και των φυσιολογικών πνευμόνων. Αυτή η απεικόνιση προσφέρει καλύτερη εποπτεία των δεδομένων που εισάγονται στο μοντέλο και αναδεικνύει τις πιθανές προκλήσεις της ταξινόμησης, καθώς οι ακτινογραφίες με παρόμοια χαρακτηριστικά μπορεί να εμφανίζουν δυσδιάκριτες διαφορές.

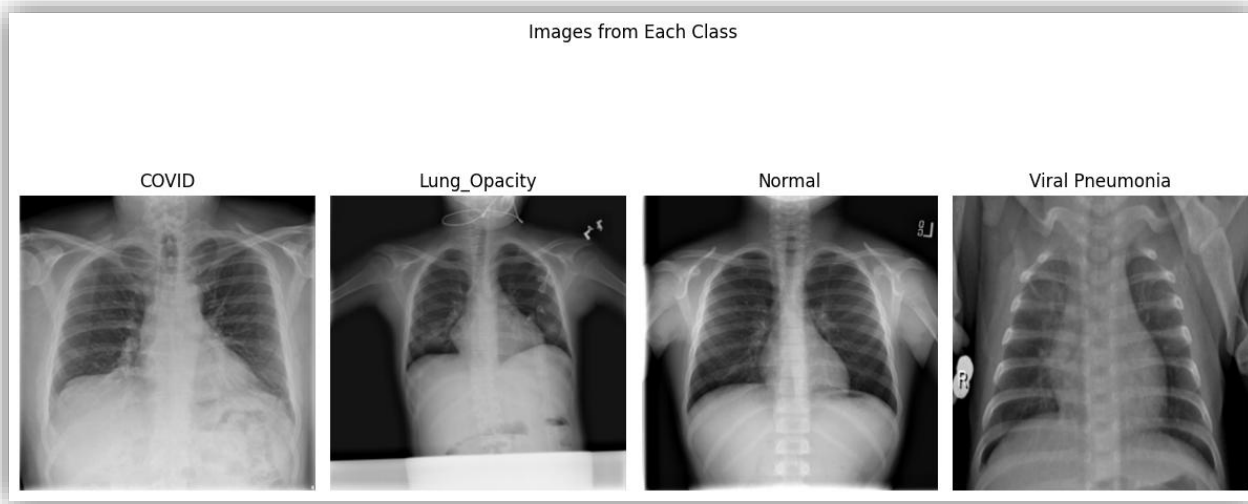


Figure 2: Εικόνες από κάθε κλάση του Dataset

3.4 Data Augmentation

Η ενίσχυση δεδομένων (Data Augmentation) αποτελεί βασική τεχνική στη μηχανική μάθηση και ειδικά στα προβλήματα ταξινόμησης εικόνων, καθώς συμβάλλει στη βελτίωση της γενίκευσης των μοντέλων. Στόχος της είναι η δημιουργία νέων, παραλλαγμένων εκδόσεων των αρχικών εικόνων, ώστε το μοντέλο να εκτεθεί σε μεγαλύτερη ποικιλομορφία εισόδων κατά την εκπαίδευση, χωρίς να αυξάνεται πραγματικά το μέγεθος του dataset. Η πρακτική αυτή μειώνει σημαντικά την πιθανότητα υπερπροσαρμογής (overfitting), ειδικά σε περιπτώσεις όπου παρατηρείται ανισορροπία μεταξύ των κλάσεων ή περιορισμένος αριθμός δειγμάτων για ορισμένες κατηγορίες.

Στην παρούσα εργασία, η ενίσχυση εφαρμόστηκε μόνο στο training set, μέσω της κλάσης ImageDataGenerator της βιβλιοθήκης TensorFlow. Πιο συγκεκριμένα, εφαρμόστηκαν οι εξής τεχνικές παραμόρφωσης εικόνας:

- Κανονικοποίηση τιμών pixel στην κλίμακα [0, 1] (rescale=1./255)
- Τυχαία περιστροφή εικόνων έως και 10 μοίρες (rotation_range=10)
- Τυχαία ζουμ έως και 10% (zoom_range=0.1)
- Οριζόντια μετατόπιση κατά $\pm 10\%$ (width_shift_range=0.1)
- Κατακόρυφη μετατόπιση κατά $\pm 10\%$ (height_shift_range=0.1)
- Οριζόντια ανάκλαση εικόνων (horizontal_flip=True)

Αντιθέτως, το validation και το test set υποβλήθηκαν μόνο σε κανονικοποίηση (rescale=1./255), ώστε να διασφαλιστεί η αντικειμενική αξιολόγηση της απόδοσης των

μοντέλων σε μη αλλοιωμένα δεδομένα. Η εφαρμογή της ενίσχυσης δεδομένων βοήθησε σημαντικά στην επίτευξη καλύτερων αποτελεσμάτων κατά την εκπαίδευση και συνέβαλε στην αύξηση της ικανότητας γενίκευσης των μοντέλων στις κατηγορίες COVID-19, Lung Opacity, Viral Pneumonia και Normal.

3.5 Αρχιτεκτονική CNN

Για τη σύγκριση με προεκπαιδευμένα μοντέλα μεταφοράς μάθησης (transfer learning), αναπτύχθηκε μία απλή αλλά αποτελεσματική συνελκτική αρχιτεκτονική νευρωνικού δικτύου (CNN). Η αρχιτεκτονική σχεδιάστηκε με έμφαση στη χαμηλή υπολογιστική πολυπλοκότητα, αλλά και τη δυνατότητα να εξάγει ουσιαστικά χαρακτηριστικά από τις ακτινογραφίες θώρακα.

Το μοντέλο αποτελείται από τρία συνελκτικά επίπεδα (Conv2D) με φίλτρα αυξανόμενης πολυπλοκότητας (32, 64, 128), το καθένα ακολουθούμενο από MaxPooling2D για μείωση χωρικής διάστασης και επιτάχυνση υπολογισμών. Η συνελκτική δομή καταλήγει σε πλήρως συνδεδεμένο επίπεδο (Dense) με 128 μονάδες και ενεργοποίηση ReLU, ακολουθούμενο από Dropout (0.2) για αποφυγή overfitting. Το τελικό επίπεδο περιλαμβάνει 4 μονάδες εξόδου με ενεργοποίηση softmax, κατάλληλη για πολυκατηγορική ταξινόμηση.

Η είσοδος ορίστηκε σε $224 \times 224 \times 3$, για συμβατότητα τόσο με προεκπαιδευμένα μοντέλα όσο και με τα τυπικά πρότυπα CNNs. Ο ρυθμός μάθησης (learning rate) επιλέχθηκε μετά από πειραματική αξιολόγηση τιμών 0.001, 0.0005 και 0.0001, όπου η τιμή 0.0005 απέδωσε τη βέλτιστη ισορροπία σύγκλισης χωρίς απότομες μεταβολές στο validation loss. Το batch size ορίστηκε σε 256, εξισορροπώντας ταυτόχρονα την ταχύτητα εκπαίδευσης με τη διαθεσιμότητα μνήμης στο περιβάλλον Google Colab.

Η ενεργοποίηση ReLU εφαρμόστηκε σε όλα τα κρυφά επίπεδα λόγω της ικανότητάς της να αποφεύγει το πρόβλημα vanishing gradient και να ενισχύει τη μη γραμμικότητα του μοντέλου. Για τη βελτιστοποίηση, επιλέχθηκε ο Adam optimizer, ο οποίος προσαρμόζει αυτόματα τον ρυθμό μάθησης για κάθε βάρος ξεχωριστά. Σύμφωνα με τους Lamouadene et al. (2021), ο Adam υπερέχει άλλων αλγορίθμων καθώς απαιτεί λιγότερη μνήμη και προσαρμόζεται δυναμικά, γεγονός που τον καθιστά κατάλληλο για πολύπλοκα προβλήματα όπως η διάγνωση από ακτινογραφίες θώρακα.

Συνοπτικά, τα βασικά χαρακτηριστικά και παραμέτροι της αρχιτεκτονικής του CNN παρατίθεται στον Table 2:

Table 2: CNN Hyperparameters

Hyperparameter	Value
Number of layers	5
Units per layer	[32, 64, 128, 128, 4]
Dropout	0.2
Input units	224×224×3
Learning rate	0.0005
Batch size	256
Epochs	40

3.6 Εφαρμογή Transfer Learning

Η μεταφορά μάθησης (Transfer Learning) αποτελεί μία από τις πιο αποτελεσματικές τεχνικές στη σύγχρονη βαθιά μάθηση, ιδίως όταν το διαθέσιμο σύνολο δεδομένων δεν είναι εξαιρετικά μεγάλο ή ποικίλο. Η βασική ιδέα είναι η αξιοποίηση γνώσης που έχει ήδη αποκτηθεί από ένα μεγάλο, γενικό σύνολο δεδομένων και η προσαρμογή της σε ένα πιο εξειδικευμένο πρόβλημα, όπως η ταξινόμηση ακτινογραφιών θώρακα. Με αυτόν τον τρόπο, μειώνεται δραστικά ο χρόνος εκπαίδευσης και αυξάνεται η ακρίβεια του τελικού μοντέλου.

Στο πλαίσιο της παρούσας μελέτης, αξιοποιήθηκαν τέσσερα δημοφιλή προεκπαιδευμένα μοντέλα: MobileNetV2, DenseNet121, EfficientNetB0 και VGG19. Η επιλογή αυτών των αρχιτεκτονικών βασίστηκε αφενός στη θεωρητική τους επάρκεια και επιδόσεις σε πληθώρα vision tasks, και αφετέρου στα αποτελέσματα προηγούμενων ερευνών που αναλύθηκαν στο κεφάλαιο της βιβλιογραφικής ανασκόπησης. Ιδιαίτερα, το EfficientNet και το DenseNet έχουν επιδείξει υψηλή ικανότητα στη γενίκευση χαρακτηριστικών από ιατρικές εικόνες, ενώ το MobileNetV2 αποτελεί ιδανική επιλογή για περιβάλλοντα με περιορισμένους πόρους, όπως το Google Colab με GPU T4.

Η στρατηγική εκπαίδευσης κάθε προεκπαιδευμένου μοντέλου οργανώθηκε σε δύο βασικές φάσεις:

- **Φάση 1 –Head Training:** Στην αρχική φάση, τα προεκπαιδευμένα βάρη του κυρίως μοντέλου διατηρήθηκαν παγωμένα (non-trainable), και εκπαιδεύτηκε μόνο το classification head που προστέθηκε στην έξοδο του δικτύου. Αυτή η φάση επέτρεψε στο νέο head να προσαρμοστεί στα χαρακτηριστικά του dataset χωρίς να διαταραχθεί η γενικευμένη γνώση της βάσης.
- **Φάση 2 – Fine-tuning:** Σε δεύτερο στάδιο, επιλεκτικά ξεκλειδώθηκε (unfroze) ένα υποσύνολο από τα τελευταία convolutional layers του βασικού μοντέλου για να επιτραπεί περαιτέρω εξειδίκευση. Ο αριθμός των επιπέδων που έγινε trainable διαφοροποιήθηκε ανά μοντέλο:

1. MobileNetV2: Unfrozen τα τελευταία 30 layers, αγνοώντας τα BatchNormalization layers.
 2. DenseNet121: Unfrozen τα τελευταία 100 layers, για μεγαλύτερη ευελιξία προσαρμογής, λόγω της μεγαλύτερης βάθους του δικτύου.
 3. EfficientNetB0: Unfrozen επίσης 30 layers, ισορροπώντας απόδοση και σταθερότητα.
 4. VGG19: Unfrozen μόνο τα 8 τελευταία convolutional layers, λόγω του υψηλού αριθμού παραμέτρων και του κινδύνου overfitting.
- **Mixed Precision Training:** Για τη μείωση της κατανάλωσης μνήμης GPU και την επιτάχυνση της εκπαίδευσης, χρησιμοποιήθηκε mixed precision. Συγκεκριμένα, οι υπολογισμοί πραγματοποιήθηκαν σε float16 όπου είναι δυνατόν, ενώ οι κρίσιμες αριθμητικές λειτουργίες (π.χ. softmax, loss) διατηρήθηκαν σε float32 για λόγους ακρίβειας.

Σε όλα τα μοντέλα, κατά τη δεύτερη φάση χρησιμοποιήθηκε μειωμένο learning rate ($1e-5$) για σταδιακή προσαρμογή, και εφαρμόστηκε callback ReduceLROnPlateau για περαιτέρω μείωση του learning rate σε περίπτωση στασιμότητας στο validation loss. Αυτό επιτρέπει ομαλότερη σύγκλιση και αποφυγή τοπικών ελαχίστων.

Η στρατηγική αυτή επέτρεψε τη μέγιστη αξιοποίηση της προεκπαιδευμένης γνώσης, ενώ παράλληλα προσαρμόστηκε επιλεκτικά στα χαρακτηριστικά του dataset. Τα αποτελέσματα κάθε αρχιτεκτονικής καταγράφονται και συγκρίνονται αναλυτικά στο επόμενο κεφάλαιο.

3.7 Εκπαίδευση και Αξιολόγηση Μοντέλων

Η εκπαίδευση και αξιολόγηση των μοντέλων πραγματοποιήθηκε με χρήση 4-Fold Stratified Cross-Validation, ώστε να διασφαλιστεί ισόρροπη κατανομή των δειγμάτων κάθε κλάσης σε όλα τα folds. Για κάθε fold, το σύνολο δεδομένων χωρίστηκε σε train και test υποσύνολο. Επιπλέον από το train set διαχωρίστηκε και ένα υποσύνολο 10% για validation, με στόχο την ορθή αξιολόγηση της γενικευσιμότητας του κάθε μοντέλου.

Η δημιουργία των datasets έγινε με τη βοήθεια ImageDataGenerator, όπου εφαρμόστηκαν augmentations μόνο στα δεδομένα εκπαίδευσης. Το CNN εκπαιδεύτηκε για 40 εποχές (epochs) ενώ τα προεκπαιδευμένα μοντέλα εκπαιδεύτηκαν με συνολική διάρκεια 30 εποχών για τα περισσότερα μοντέλα (10 εποχές με παγωμένα τα προεκπαιδευμένα βάρη και 20 επιπλέον εποχές με σταδιακά ξεκλείδωμα συγκεκριμένων επιπέδων του backbone).

Οι παρακάτω εικόνες παρουσιάζουν τη διαδικασία διαχωρισμού και δημιουργίας των folds, καθώς και την αναλογία των δεδομένων σε κάθε σύνολο:

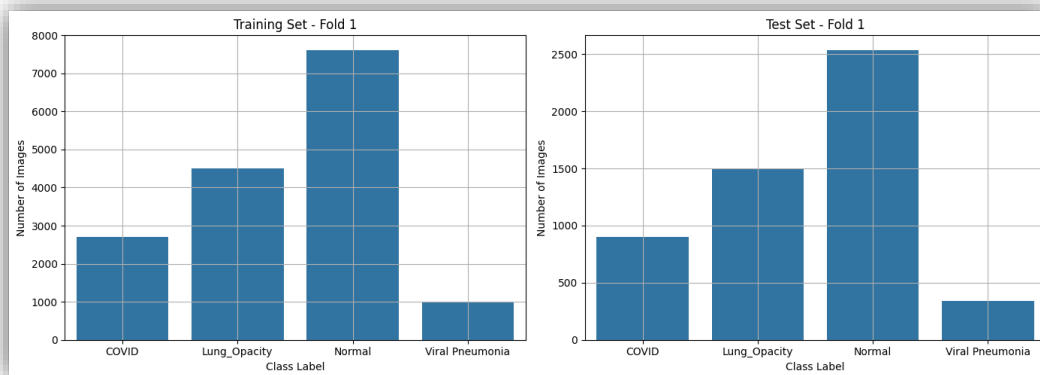


Figure 3: Κατανομή δεδομένων μετά τη διαδικασία διάσπασης – Fold 1

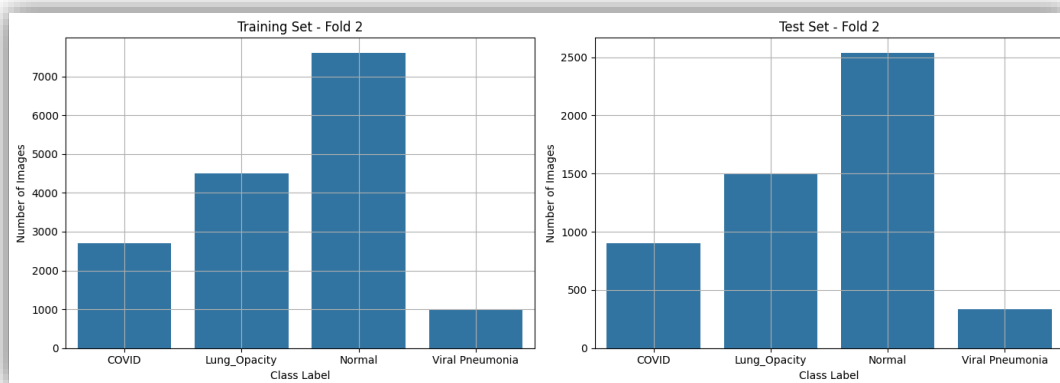


Figure 4: Κατανομή δεδομένων μετά τη διαδικασία διάσπασης – Fold 2

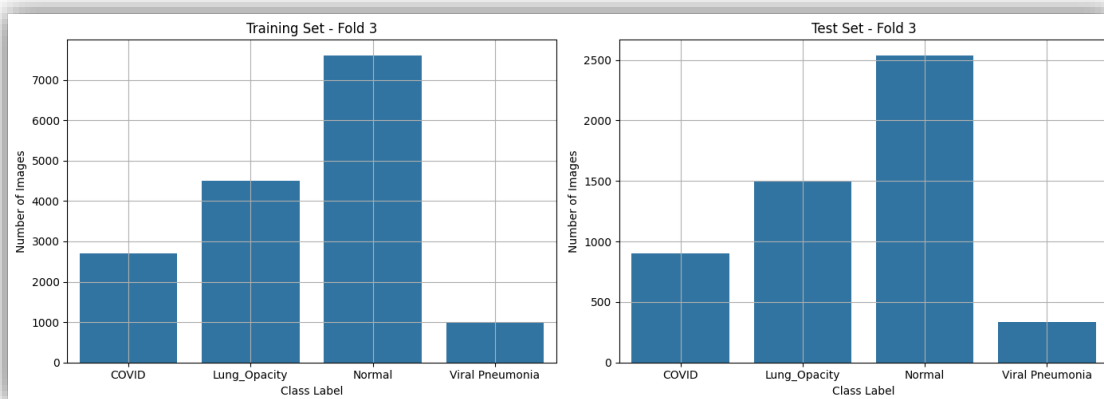


Figure 5: Κατανομή δεδομένων μετά τη διαδικασία διάσπασης – Fold 3

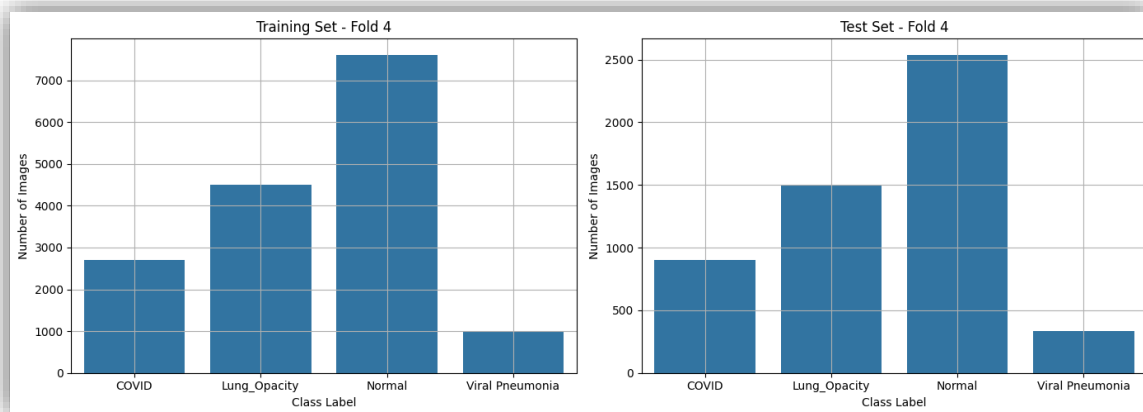


Figure 6: Κατανομή δεδομένων μετά τη διαδικασία διάσπασης – Fold 4

Για την παρακολούθηση της προόδου και την αποτροπή overfitting, χρησιμοποιήθηκε ο μηχανισμός EarlyStopping με στόχο τον τερματισμό της εκπαίδευσης όταν η απόδοση στο validation loss δεν βελτιωνόταν για 5 συνεχόμενες εποχές. Παράλληλα, η τεχνική ModelCheckpoint αποθήκευε τα βέλτιστα βάρη του μοντέλου (με βάση τη χαμηλότερη τιμή validation loss), τα οποία χρησιμοποιήθηκαν τελικά για τις τελικές προβλέψεις.

Η απόδοση κάθε μοντέλου αξιολογήθηκε με βάση τις μετρικές accuracy, precision, recall και F1-score, ενώ ο συνολικός χρόνος εκπαίδευσης καταγράφηκε για κάθε fold. Για την καλύτερη κατανόηση της εκπαίδευσης και της γενικευσιμότητας των μοντέλων, παρουσιάζονται τα αντίστοιχα διαγράμματα train/validation accuracy καθώς και train/validation loss για κάθε fold, επιτρέποντας τον εντοπισμό φαινομένων όπως underfitting ή overfitting και τη σύγκριση της συμπεριφοράς των μοντέλων καθ' όλη τη διάρκεια της εκπαίδευσης.

Παρακάτω αναλύονται οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση των μοντέλων:

Accuracy

Η ακρίβεια (accuracy) εκφράζει το ποσοστό των σωστά προβλεφθέντων παρατηρήσεων (ανεξαρτήτως κατηγορίας) ως προς το σύνολο των προβλέψεων. Υπολογίζεται με βάση τον τύπο:

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

Όπου:

- **TP (True Positives):** ακτινογραφίες που ανήκουν σωστά στην κατηγορία που προέβλεψε το μοντέλο

- **TN (True Negatives):** ακτινογραφίες που δεν ανήκουν σε μια συγκεκριμένη κατηγορία και το μοντέλο σωστά δεν τις κατέταξε εκεί
- **FP (False Positives):** ακτινογραφίες που το μοντέλο κατέταξε λανθασμένα σε μια κατηγορία
- **FN (False Negatives):** ακτινογραφίες που ανήκουν σε μια κατηγορία αλλά το μοντέλο τις κατέταξε αλλού

Παρόλο που η ακρίβεια αποτελεί μια βασική μετρική αξιολόγησης, μπορεί να είναι παραπλανητική σε datasets με ανισοκατανομή κατηγοριών. Στο παρόν σύνολο δεδομένων, η κατηγορία "Normal" είναι η πολυπληθέστερη, γεγονός που ενέχει τον κίνδυνο ένα μοντέλο να αποδίδει υψηλή ακρίβεια απλώς επειδή προβλέπει συχνότερα την επικρατούσα κατηγορία, χωρίς να αναγνωρίζει επαρκώς τις υπόλοιπες. Για τον λόγο αυτό, η συνολική αξιολόγηση των μοντέλων βασίστηκε επίσης στις μετρικές precision, recall και F1-score, οι οποίες προσφέρουν πιο ισορροπημένη εικόνα της απόδοσης ανά κλάση.

Precision

Το precision μετράει το ποσοστό των σωστών θετικών προβλέψεων, δηλαδή πόσες από τις ακτινογραφίες που το μοντέλο ταξινόμησε σε μία συγκεκριμένη κλάση ήταν πράγματι σωστές. Υπολογίζεται ως ο λόγος των True Positives προς το σύνολο των δειγμάτων που προβλέφθηκαν ως θετικά:

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

Το precision είναι ιδιαίτερα σημαντικό σε περιπτώσεις όπου το κόστος μιας λανθασμένης θετικής πρόβλεψης (false positive), όπως η εσφαλμένη διάγνωση μιας παθολογίας, είναι υψηλό.

Recall

Το recall αξιολογεί την ικανότητα του μοντέλου να εντοπίζει σωστά όλες τις θετικές περιπτώσεις μιας κατηγορίας, δηλαδή πόσες από τις ακτινογραφίες που ανήκουν πράγματι σε μια κλάση εντοπίστηκαν σωστά από το μοντέλο. Υπολογίζεται ως:

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

Ένα υψηλό recall υποδεικνύει ότι το μοντέλο καταφέρνει να ανιχνεύσει τη μεγάλη πλειοψηφία των πραγματικών θετικών δειγμάτων, ακόμα κι αν ενδέχεται να περιλαμβάνει κάποια false positives. Σε ιατρικές εφαρμογές, όπως η διάγνωση μέσω ακτινογραφιών, το recall είναι συχνά

πιο κρίσιμη από το precision, καθώς η αποτυχία εντοπισμού μιας παθολογικής περίπτωσης μπορεί να έχει σοβαρές συνέπειες στην υγεία του ασθενούς.

F1 Score

Το F1 score είναι ο αρμονικός μέσος της precision και recall. Συνδυάζει τις δύο αυτές μετρικές σε μία ενιαία τιμή και είναι ιδιαίτερα χρήσιμο σε προβλήματα με ανισόρροπη κατανομή κλάσεων. Ο τύπος υπολογισμού είναι:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Ένα υψηλό F1 score υποδηλώνει ότι το μοντέλο επιτυγχάνει καλές επιδόσεις τόσο στο recall όσο και στο precision, γεγονός που το καθιστά αξιόπιστη ένδειξη της συνολικής ποιότητας ταξινόμησης. Στην παρούσα μελέτη, το F1 χρησιμοποιήθηκε για να εκτιμηθεί η ισορροπία μεταξύ των εσφαλμένων και σωστών ταξινομήσεων σε όλες τις κατηγορίες.

ROC-AUC (Receiver Operating Characteristic – Area Under Curve)

Η καμπύλη ROC απεικονίζει τη σχέση μεταξύ του ποσοστού αληθώς θετικών (True Positive Rate - Recall) και του ποσοστού ψευδώς θετικών (False Positive Rate), μεταβάλλοντας το κατώφλι απόφασης του ταξινομητή. Η επιφάνεια κάτω από την καμπύλη (AUC) δίνει ένα συνολικό μέτρο της διακριτικής ικανότητας του μοντέλου:

$$AUC = \int_0^1 TPR(FPR), dFPR$$

Μία τιμή AUC ίση με 0.5 υποδηλώνει ότι το μοντέλο δεν έχει καμία διακριτική ικανότητα (τυχαία πρόβλεψη), ενώ τιμή ίση με 1.0 αντιστοιχεί σε τέλεια ταξινόμηση. Η μέτρηση AUC είναι εξαιρετικά χρήσιμη για τη σύγκριση μοντέλων ως προς την ικανότητά τους να διαχωρίζουν σωστά τα δείγματα, ανεξάρτητα από το συγκεκριμένο κατώφλι που επιλέγεται για την τελική απόφαση.

Accuracy and Loss Curves

Οι καμπύλες accuracy και loss χρησιμοποιούνται ευρέως για την αξιολόγηση της πορείας εκπαίδευσης ενός νευρωνικού δικτύου και παρέχουν κρίσιμη οπτική πληροφόρηση σχετικά με τη συμπεριφορά του μοντέλου ανά εποχή. Η accuracy αποτυπώνει την ικανότητα του μοντέλου να ταξινομεί σωστά τα δείγματα, ενώ το loss μετράει το πόσο αποκλίνουν οι προβλέψεις του μοντέλου από τις πραγματικές τιμές, βάσει της επιλεγμένης συνάρτησης κόστους. Η ταυτόχρονη ανάλυση των καμπυλών αυτών επιτρέπει την ανίχνευση προβλημάτων όπως overfitting ή underfitting, καθώς και τη βελτιστοποίηση των υπερπαραμέτρων. Επιπλέον, αποτελούν βασικό εργαλείο για την κατανόηση της γενικευσιμότητας του μοντέλου σε μη ορατά δεδομένα, μέσω της σύγκρισης των επιδόσεων στο training και validation σύνολο.

4. Πειραματικά αποτελέσματα

Μετά την υλοποίηση και εκπαίδευση των διαφόρων αρχιτεκτονικών που περιγράφηκαν στο προηγούμενο κεφάλαιο, το παρόν κεφάλαιο επικεντρώνεται στην αναλυτική παρουσίαση και σύγκριση των αποτελεσμάτων. Η αξιολόγηση των μοντέλων πραγματοποιείται τόσο ποσοτικά, μέσω των βασικών μετρικών ταξινόμησης, όσο και οπτικά, με τη βοήθεια γραφημάτων και καμπυλών που τεκμηριώνουν την απόδοση και τη γενικευσιμότητά τους.

Ειδικότερα, παρουσιάζονται οι καμπύλες ROC για κάθε κλάση και κάθε μοντέλο, πίνακες σύγχυσης (confusion matrices) για τον εντοπισμό σφαλμάτων ανά κατηγορία, καθώς και ραβδογράμματα που συγκρίνουν τις αρχιτεκτονικές με βάση τις μετρικές ακρίβειας (accuracy), ευαισθησίας (recall), ειδικότητας (precision) και F1-score. Επιπλέον, περιλαμβάνονται οι καμπύλες εκπαίδευσης (training/validation loss και accuracy) για κάθε μοντέλο, οι οποίες παρέχουν σημαντική πληροφόρηση για την πορεία της εκπαίδευσης και φαινόμενα όπως overfitting ή underfitting.

Η σύνθεση όλων αυτών των στοιχείων καθιστά δυνατή την εξαγωγή συμπερασμάτων σχετικά με την αποδοτικότητα, τη σταθερότητα και τη γενικευσιμότητα κάθε προσέγγισης, παρέχοντας έτσι τη βάση για μια τεκμηριωμένη σύγκριση των διαφορετικών μοντέλων.

4.1 Απόδοση Ανά Μοντέλο

Σε αυτή την ενότητα παρουσιάζεται η απόδοση κάθε μοντέλου ξεχωριστά. Για κάθε αρχιτεκτονική που εφαρμόστηκε (CNN και μοντέλα Transfer Learning), περιλαμβάνονται οι καμπύλες ROC ανά κλάση, οι Confusion Matrices καθώς και τα διαγράμματα Loss/Accuracy ανά fold. Ο στόχος είναι η οπτικοποίηση της συμπεριφοράς κάθε μοντέλου και η αξιολόγηση της καταλληλότητάς του στην ταξινόμηση των τεσσάρων κατηγοριών ακτινογραφιών θώρακα.

CNN

Το πρώτο μοντέλο που αξιολογήθηκε ήταν το CNN που δημιουργήθηκε με 5 layers το οποίο εκπαιδεύτηκε για 40 epochs και αξιολογήθηκε εφαρμόζοντας 4-fold cross-validation. Ο συνολικός χρόνος εκπαίδευσης και αξιολόγησης για όλα τα folds ανήλθε σε 9 ώρες και 13 λεπτά, καθιστώντας το μοντέλο το πιο άργο από όλες τις τεχνικές transfer learning.

Οι καμπύλες accuracy και Loss προσφέρουν χρήσιμες ενδείξεις για την συμπεριφορά του μοντέλου κατά τη διάρκεια εκπαίδευσης.

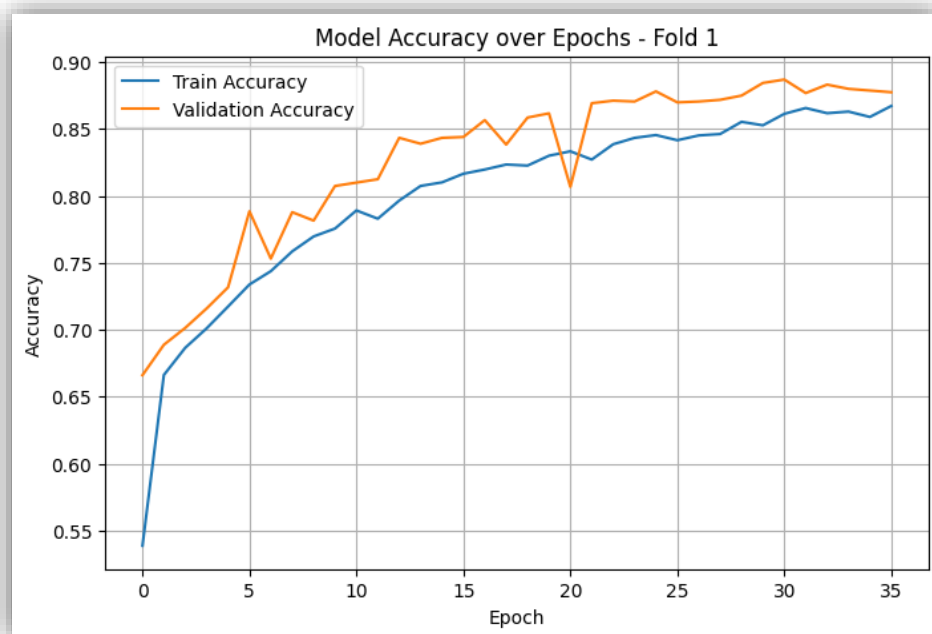


Figure 7: Καμπύλη Train/Validation Accuracy για το CNN - Fold 1

Όπως παρατηρείται από τις καμπύλες εκπαίδευσης, το CNN μοντέλο παρουσίασε σταθερή βελτίωση τόσο στο accuracy όσο και στη μείωση του loss για τα δεδομένα εκπαίδευσης και επικύρωσης. Οι δύο καμπύλες ακρίβειας (training και validation accuracy) κινούνται σχεδόν παράλληλα καθ' όλη τη διάρκεια των εποχών, χωρίς να αποκλίνουν αισθητά μεταξύ τους. Αυτό αποτελεί ισχυρή ένδειξη ότι το μοντέλο δεν υπερεκπαιδεύεται (overfitting) αλλά μαθαίνει σταδιακά με τρόπο που γενικεύει καλά και σε μη ορατά δεδομένα.

Το γεγονός ότι η καμπύλη του validation accuracy υπερβαίνει αυτή της εκπαίδευσης ενισχύει την επίδραση της τεχνικής data augmentation, η οποία εφαρμόστηκε αποκλειστικά στο training set και ενισχύει την ποικιλία των δειγμάτων. Παράλληλα, η επιλογή ενός σχετικά μεγάλου batch size (256) φαίνεται να βοήθησε στην επίτευξη σταθερότερων ενημερώσεων των βαρών, χωρίς απότομες διακυμάνσεις (spikes) ανά εποχή, βελτιώνοντας τη συνολική σταθερότητα της εκπαίδευσης.

Η εκπαίδευση του μοντέλου τερματίστηκε αυτόματα στην 35η εποχή λόγω του μηχανισμού EarlyStopping, καθώς δεν παρατηρήθηκε βελτίωση στο validation loss για 5 συνεχόμενες εποχές. Η παρέμβαση αυτή απέτρεψε την περαιτέρω εκπαίδευση του μοντέλου, προστατεύοντάς το από πιθανή υπερπροσαρμογή (overfitting) και περιορίζοντας το training time.

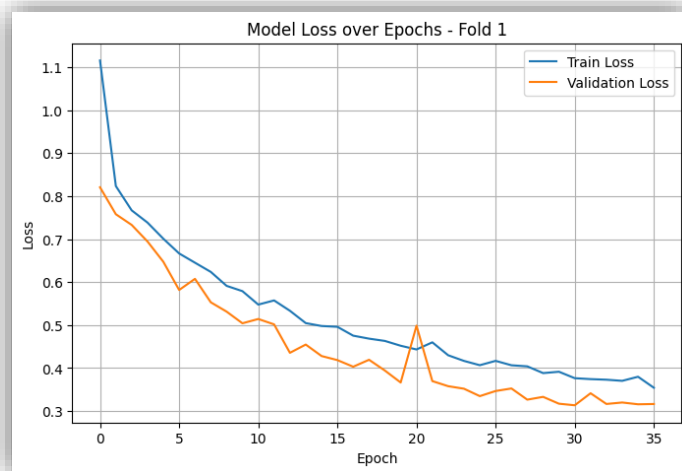


Figure 8: Καμπύλη Train/Validation Loss για το CNN - Fold 1

Αντίστοιχη εικόνα εντοπίζεται και στις καμπύλες απώλειας, όπου τόσο το training loss όσο και το validation loss παρουσιάζουν συνεχή και σταθερή πτωτική τάση. Το validation loss μάλιστα είναι σε ελαφρώς χαμηλότερα επίπεδα από το training loss, κάτι που σε συνδυασμό με τις παράλληλες καμπύλες δηλώνει καλό generalization.

Συνοπτικά, οι καμπύλες επιβεβαιώνουν τη σταθερή μάθηση του CNN μοντέλου και την αποτελεσματικότητα των υπερπαραμέτρων και τεχνικών regularization που εφαρμόστηκαν (όπως dropout και data augmentation).

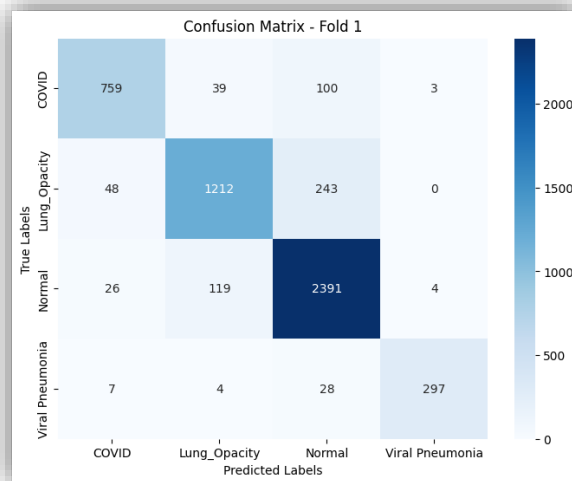


Figure 9: CNN Confusion Matrix - Fold 1

Ο πίνακας σύγχυσης του CNN από το πρώτο fold αποτυπώνει την ικανότητα του μοντέλου να διακρίνει σωστά μεταξύ των τεσσάρων κατηγοριών. Παρατηρούμε ότι η κλάση Normal είχε τη μεγαλύτερη ακρίβεια ταξινόμησης (2391 σωστές προβλέψεις από 2540 δείγματα), ενώ

ακολουθεί η Viral Pneumonia με 297 σωστές προβλέψεις από 336 δείγματα. Ωστόσο, εμφανίζονται σημαντικές συγχύσεις μεταξύ Lung Opacity και Normal, καθώς 243 περιπτώσεις Lung Opacity προβλέφθηκαν ως Normal, και αντίστοιχα 119 Normal ως Lung Opacity. Αυτή η αλληλοκάλυψη υποδεικνύει τη μορφολογική ομοιότητα μεταξύ των δύο παθήσεων στις ακτινογραφίες.

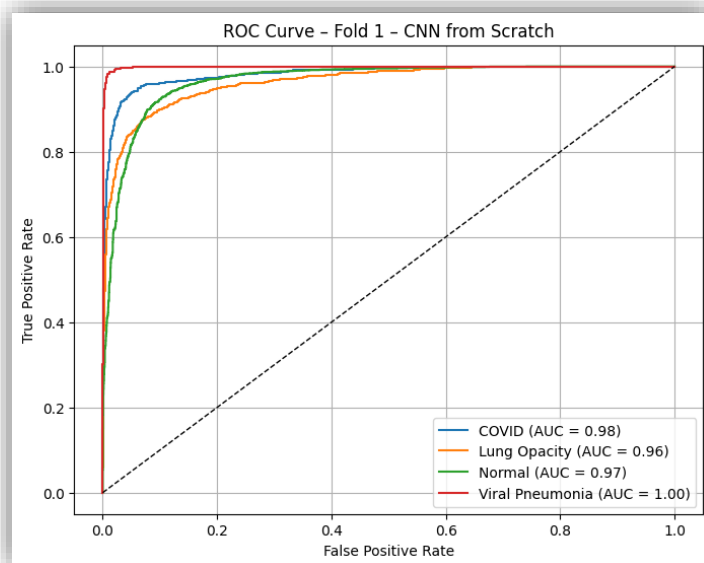


Figure 10: CNN Roc Curve για κάθε κλάση

Η καμπύλη ROC για το Fold 1 του CNN καταδεικνύει ότι το μοντέλο έχει υψηλή ικανότητα διάκρισης μεταξύ των τεσσάρων κλάσεων. Οι καμπύλες και των τεσσάρων κατηγοριών βρίσκονται πολύ κοντά στο άνω αριστερό άκρο του γραφήματος, γεγονός που φανερώνει ότι το μοντέλο πετυχαίνει υψηλά true positive rates διατηρώντας χαμηλά false positives. Αυτό επιβεβαιώνει την ικανότητα του μοντέλου να αναγνωρίζει αποτελεσματικά ακόμα και τις πιο δύσκολες κατηγορίες, όπως η "COVID-19" και η "Lung Opacity", οι οποίες συχνά παρουσιάζουν οπτική ομοιότητα σε ακτινογραφίες θώρακα.

Σημαντικό είναι να τονιστεί ότι τα παραπάνω αποτελέσματα προήλθαν από ένα απλό CNN μοντέλο, το οποίο εκπαιδεύτηκε εξ' ολοκλήρου από την αρχή (χωρίς χρήση προεκπαιδευμένων βαρών) και χωρίς ιδιαίτερα βαθιά ή πολύπλοκη αρχιτεκτονική. Το γεγονός ότι επιτυγχάνονται τόσο αξιόλογες επιδόσεις, ενισχύει την αποτελεσματικότητα της συνολικής διαδικασίας προεπεξεργασίας, της επιλογής υπερπαραμέτρων και της σωστής διαχείρισης του συνόλου δεδομένων.

MobileNetV2

Το MobileNetV2 χρησιμοποιήθηκε ως μία από τις προσεγγίσεις transfer learning στην παρούσα μελέτη, με στόχο την ταξινόμηση ακτινογραφιών θώρακα σε τέσσερις κατηγορίες. Αν και πολλά άρθρα της βιβλιογραφίας αξιοποιούν αρχιτεκτονικές όπως το ResNet50, στην παρούσα εργασία προτιμήθηκε το MobileNetV2 λόγω της χαμηλής του υπολογιστικής πολυπλοκότητας, της ταχύτερης εκπαίδευσης (6 ώρες αντί για 9) και της εξαιρετικής του επίδοσης σε περιβάλλοντα με περιορισμένους πόρους, όπως το Google Colab. Το MobileNetV2 έχει σχεδιαστεί ειδικά για αποδοτικότητα σε κινητές και embedded συσκευές, καθιστώντας το ιδανικό για προβλήματα όπου απαιτείται ταχύτητα χωρίς σημαντική απώλεια ακρίβειας. Το μοντέλο φορτώθηκε με προεκπαιδευμένα βάρη από το ImageNet dataset, γεγονός που του επέτρεψε να ξεκινήσει την εκπαίδευση με ήδη ισχυρές γενικές απεικονιστικές δυνατότητες.

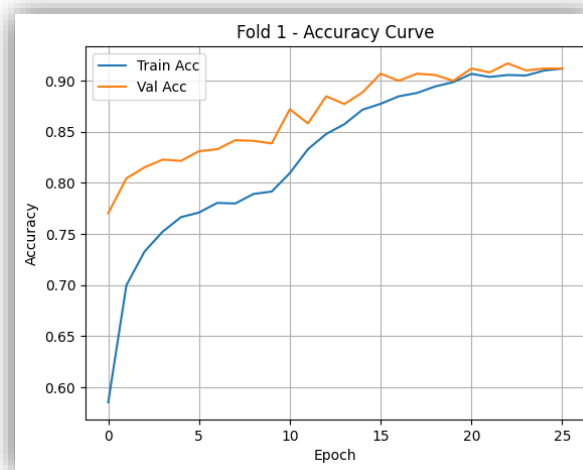


Figure 11: Καμπύλη Train/Validation Accuracy για το MobileNetV2 - Fold 1

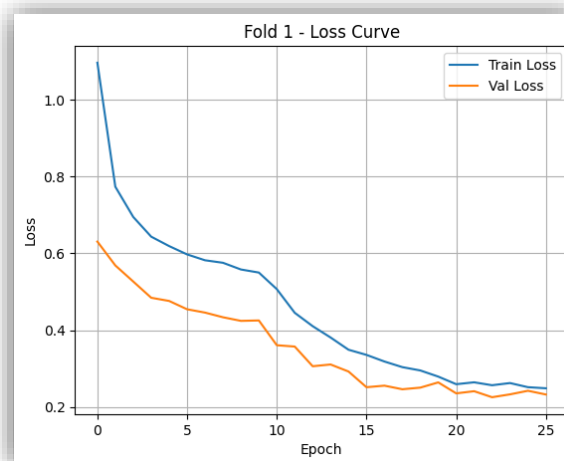


Figure 12: Καμπύλη Train/Validation Loss για το MobileNetV2 - Fold 1

Όπως διακρίνεται από τις καμπύλες Accuracy και Loss για το MobileNetV2 παρουσίασε ομαλή και συνεπή βελτίωση τόσο στο training όσο και στο validation set κατά τη διάρκεια των εποχών. Η καμπύλη accuracy δείχνει ότι το validation accuracy αυξάνεται σταθερά, ξεκινώντας από περίπου 78% και φτάνοντας πάνω από 91%, ενώ η αντίστοιχη καμπύλη απώλειας μειώνεται σταθερά, χωρίς έντονα σημάδια overfitting.

Η επιλογή μικρότερου batch size σε σύγκριση με το CNN (64 αντί 256) σχετίζεται με το αυξημένο μέγεθος και βάθος του προεκπαιδευμένου δικτύου, ώστε να διασφαλιστεί η απρόσκοπτη εκπαίδευση στο περιβάλλον Google Colab χωρίς εξάντληση VRAM. Επιπλέον, η εφαρμογή της τεχνικής ReduceLROnPlateau συνέβαλε περαιτέρω στη δυναμική προσαρμογή του ρυθμού μάθησης όταν το validation loss σταθεροποιούνταν, ενισχύοντας την αποτελεσματικότητα της μάθησης στις τελευταίες εποχές.

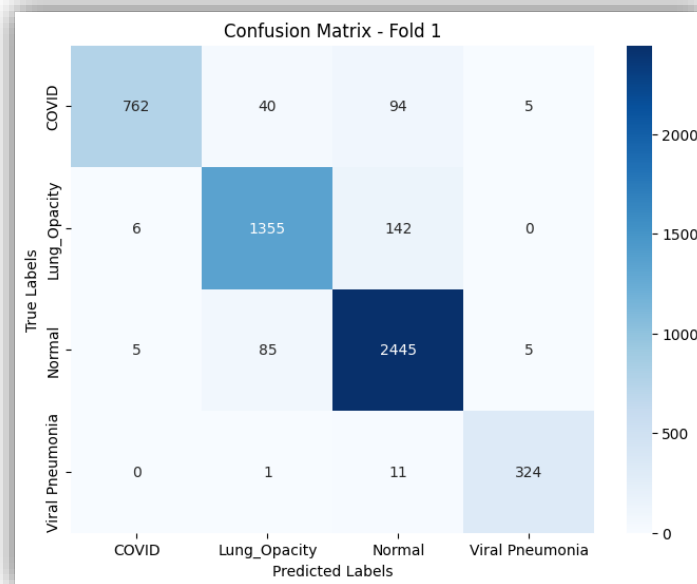


Figure 13: MobileNetV2 Confusion Matrix - Fold 1

Σύμφωνα με το Confusion Matrix του MobileNetV2 για το Fold 1, παρατηρείται υψηλότερη ακρίβεια ταξινόμησης για όλες τις κλάσεις με πολύ περιορισμένα λάθη. Επιπλέον, οι καμπύλες ROC επιβεβαιώνουν την υψηλή απόδοση του μοντέλου καθώς όλες οι κλάσεις παρουσιάζουν AUC άνω του 0.98 με τις κλάσεις COVID-19 και Viral Pneumonia σχεδόν τέλεια (πολύ χαμηλός αριθμός False Positive). Τα αποτελέσματα αυτά υποδεικνύουν ότι το MobileNetV2 προσφέρει εξαιρετική διαχωριστική ικανότητα, ειδικά λαμβάνοντας υπόψη την υπολογιστική του αποδοτικότητα.

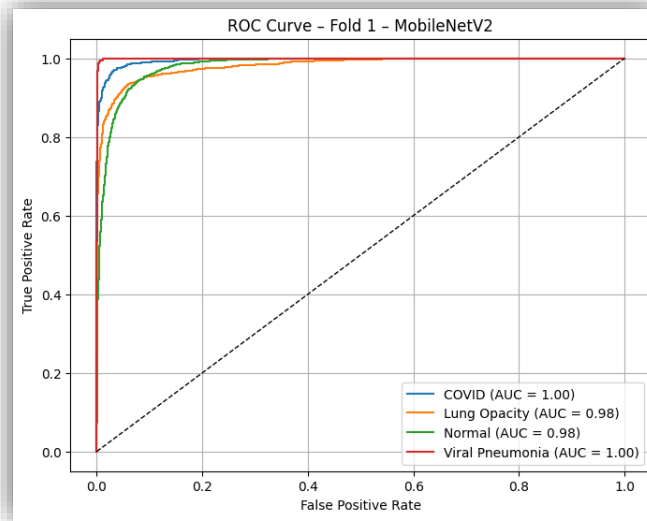


Figure 14: MobileNetV2 Roc Curve για κάθε κλάση

DenseNet121

Το προεκπαιδευμένο μοντέλο DenseNet121 επιλέχθηκε για αξιολόγηση λόγω της αποτελεσματικής χρήσης παραμέτρων και της υψηλής δυνατότητας επαναχρησιμοποίησης χαρακτηριστικών μέσω των πυκνών διασυνδέσεων που το χαρακτηρίζουν. Το συγκεκριμένο μοντέλο θεωρείται ιδιαίτερα κατάλληλο για εφαρμογές που απαιτούν εκμετάλλευση λεπτομερών πληροφοριών από ιατρικές εικόνες, καθώς έχει αποδείξει εξαιρετική απόδοση σε ανάλογες μελέτες στον τομέα της ακτινογραφίας θώρακα. Ωστόσο, σε σχέση με τα υπόλοιπα transfer learning μοντέλα της μελέτης, παρουσίασε τον μεγαλύτερο χρόνο εκπαίδευσης καθώς χρειάστηκε 8 ώρες συνολικά για την εκπαίδευση και αξιολόγηση όλων των folds.

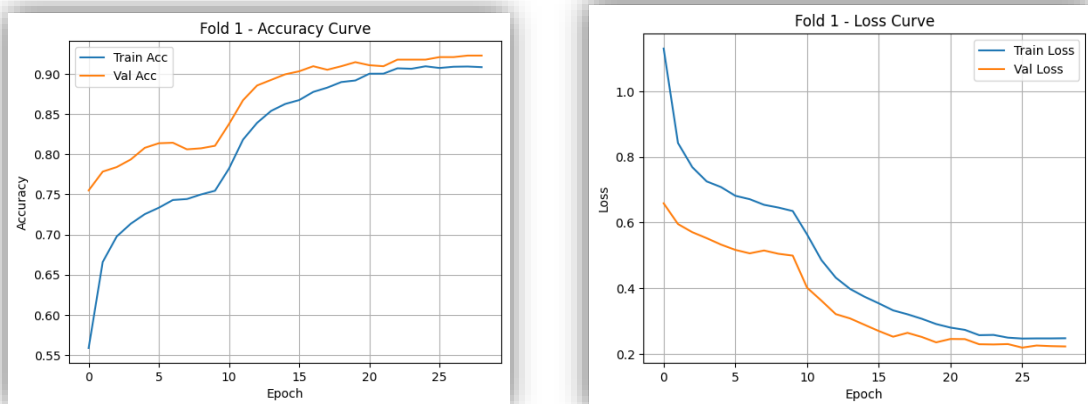


Figure 15: Καμπύλες Train/Validation Accuracy και Loss για το DenseNet121 – Fold 1

Οι καμπύλες accuracy και loss του DenseNet121 υποδεικνύουν σταθερή και προβλέψιμη συμπεριφορά κατά την εκπαίδευση, χωρίς ενδείξεις υπερεκπαίδευσης. Η βελτίωση της απόδοσης είναι σταδιακή και ομαλή, ενώ η ευθυγράμμιση μεταξύ training και validation curve επιβεβαιώνει την καλή γενίκευση του μοντέλου. Αξίζει να σημειωθεί ότι η τελική απόδοση του DenseNet121 είναι παρόμοια με εκείνη του MobileNetV2, με ελάχιστες διαφορές στις επιμέρους μετρικές, γεγονός που θέτει ερωτήματα για την αναλογία cost/benefit δεδομένου του σημαντικά μεγαλύτερου χρόνου εκπαίδευσης που απαιτεί.

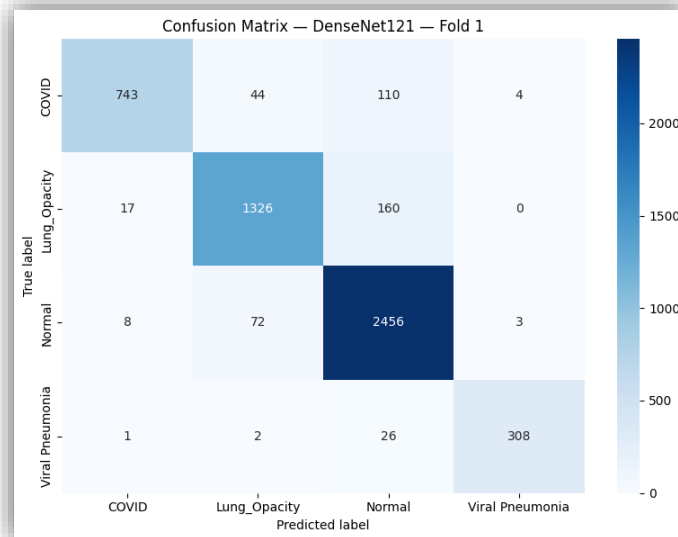


Figure 16: DenseNet121 Confusion Matrix - Fold 1

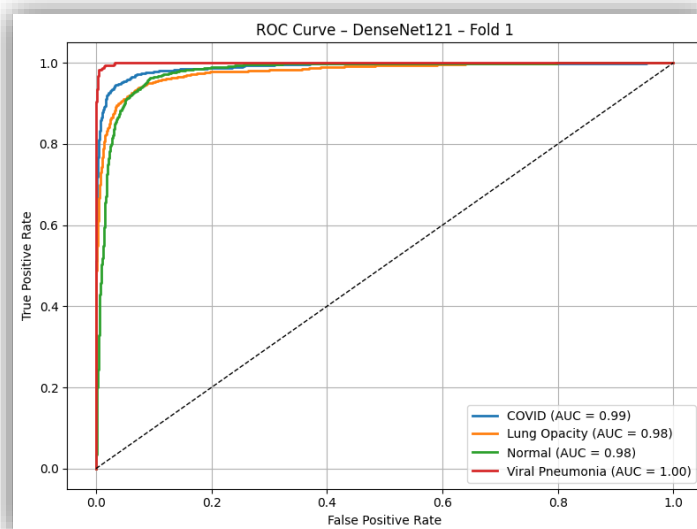


Figure 17: DenseNet121 Roc Curve για κάθε κλάση

Η ανάλυση των αποτελεσμάτων του DenseNet121 για το Fold 1 παρουσιάζει εξαιρετικά παρόμοια απόδοση με εκείνη του MobileNetV2. Η καμπύλη ROC υποδεικνύει υψηλές τιμές AUC και για τις τέσσερις κλάσεις, υποδηλώνοντας ικανότητα διάκρισης μεταξύ των κατηγοριών με μεγάλη ακρίβεια. Παράλληλα, ο confusion matrix επιβεβαιώνει την αξιοπιστία του μοντέλου, παρουσιάζοντας υψηλά ποσοστά σωστής ταξινόμησης, ειδικά για τις κατηγορίες "Normal" και "Viral Pneumonia", με μικρό αριθμό λανθασμένων προβλέψεων μεταξύ COVID και Lung Opacity.

Παρότι και τα δύο μοντέλα εκπαιδεύτηκαν σε δύο φάσεις (10 εποχές frozen, 20 fine-tune), στο DenseNet121 στα τελευταία 100 layers πραγματοποιήθηκε unfreeze, έναντι μόλις 30 layers του MobileNetV2, επιτρέποντας πιο εντατική προσαρμογή του δικτύου στο εκάστοτε fold του dataset. Ωστόσο, λόγω της μεγαλύτερης πολυπλοκότητας του DenseNet121 (περισσότερα από 100 layers συνολικά), απαιτήθηκε χαμηλότερο batch size (48) και περισσότεροι υπολογιστικοί πόροι, οδηγώντας σε μεγαλύτερο χρόνο εκπαίδευσης.

EfficientNetB0

Το EfficientNetB0 αποτελεί ένα από τα μοντέρνα lightweight transfer learning μοντέλα, σχεδιασμένο για μέγιστη ακρίβεια με ελάχιστο αριθμό παραμέτρων και υπολογιστικό κόστος. Παρά το γεγονός ότι σε πολλές περιπτώσεις έχει δείξει εξαιρετικές επιδόσεις, στην παρούσα μελέτη το μοντέλο χρειάστηκε περίπου 6 ώρες και 20 λεπτά για την εκπαίδευση και αξιολόγηση σε όλα τα folds. Επιλέχθηκε για να συγκριθεί η απόδοση ενός πιο "optimized" transfer learning μοντέλου έναντι πιο βαθιών αρχιτεκτονικών όπως το DenseNet121

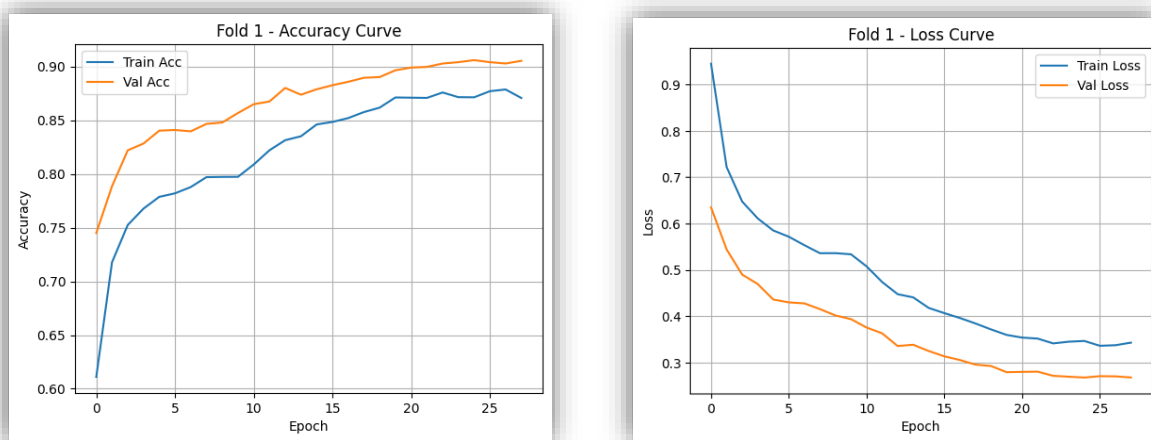


Figure 18: Καμπύλες Train/Validation Accuracy και Loss για το EfficientNetB0 – Fold 1

Από την καμπύλη accuracy παρατηρείται ανοδική πορεία τόσο στο training όσο και στο validation set, ωστόσο με μικρότερη τελική επίδοση σε σύγκριση με τα άλλα μοντέλα transfer learning (0.87% validation accuracy έναντι 0.89%). Ομοίως, η καμπύλη loss δείχνει συνεχή μείωση της τιμής του loss, χωρίς απότομες αυξομειώσεις, γεγονός που υποδηλώνει σταθερή μάθηση. Παρόλα αυτά, οι τελικές τιμές σε accuracy και loss είναι σαφώς κατώτερες σε σχέση

με τα MobileNetV2 και DenseNet121, και παρουσιάζουν αποτελέσματα που πλησιάζουν εκείνα του CNN, γεγονός που ενδεχομένως σχετίζεται με την αρχιτεκτονική βάσης του EfficientNetB0 που είναι πιο περιορισμένη σε βάθος και χωρητικότητα.

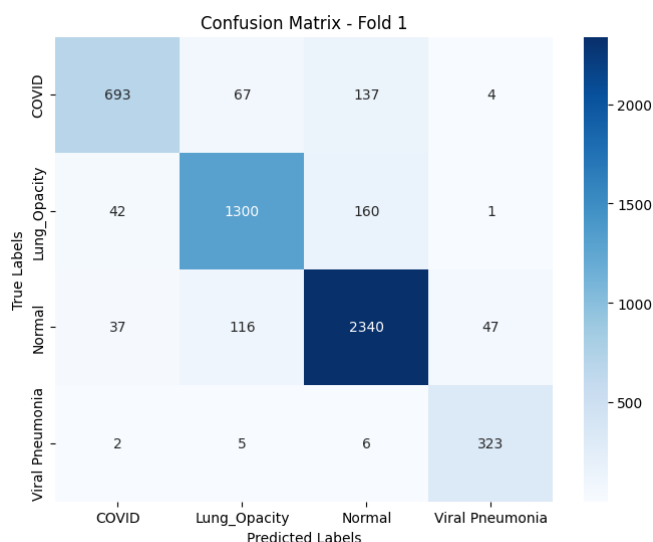


Figure 19:: EfficientNetB0 Confusion Matrix - Fold1

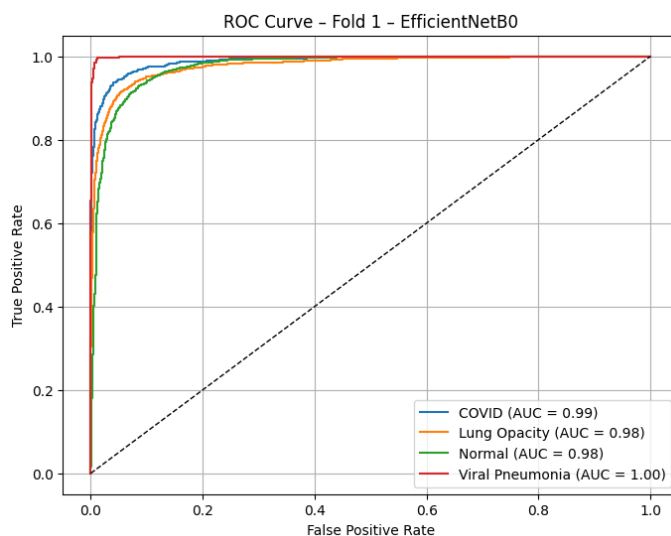


Figure 20: EfficientNetB0 Roc Curve για κάθε κλάση

Από τα διαγράμματα Confusion Matrix και Roc Curves επιβεβαιώνονται τα προηγούμενα συμπεράσματα. Το μοντέλο εμφάνισε γενικά χαμηλότερα απόδοση με τα άλλα transfer learning μοντέλα. Παρόλο που οι τιμές AUC παραμένουν υψηλές από το Confusion Matrix παρατηρούνται περισσότερα σφάλματα ταξινόμησης σε σύγκριση με τα υπόλοιπα μοντέλα, ιδιαίτερα στις κλάσεις COVID και Normal όπου καταγράφηκαν οι περισσότερες εσφαλμένες προβλέψεις. Η συνολική συμπεριφορά ταυτίζεται περισσότερο με το CNN και με γνώμονα την διάρκεια εκπαίδευσης κατατάσσεται ως το λιγότερο αποδοτικό σε σχέση με χρόνο/απόδοση.

VGG 19

Το VGG-19, παρότι πιο παλιό αρχιτεκτονικά σε σχέση με άλλα μοντέλα όπως τα EfficientNet και DenseNet, επέδειξε εξαιρετική απόδοση στο παρόν πρόβλημα ταξινόμησης ακτινογραφιών θώρακα. Το μοντέλο επιλέχθηκε βάσει της αξιοπιστίας του σε προηγούμενες μελέτες και της απλής αλλά βαθιάς δομής του (με 19 επίπεδα), που το καθιστά κατάλληλο για λεπτομερή εξαγωγή χαρακτηριστικών. Παρά τον μεγάλο αριθμό παραμέτρων, η εκπαίδευση του VGG-19 ήταν ταχύτερη σε σύγκριση με τα άλλα transfer learning μοντέλα, ολοκληρώνοντας την εκπαίδευση και αξιολόγηση σε περίπου 5 ώρες και 35 λεπτά. Αυτή η ταχύτητα οφείλεται κυρίως στο γεγονός ότι η διαδικασία early stopping ενεργοποιήθηκε νωρίτερα από τα άλλα μοντέλα, συγκεκριμένα στο 18ο epoch, όταν το validation loss σταθεροποιήθηκε χωρίς περαιτέρω βελτίωση.

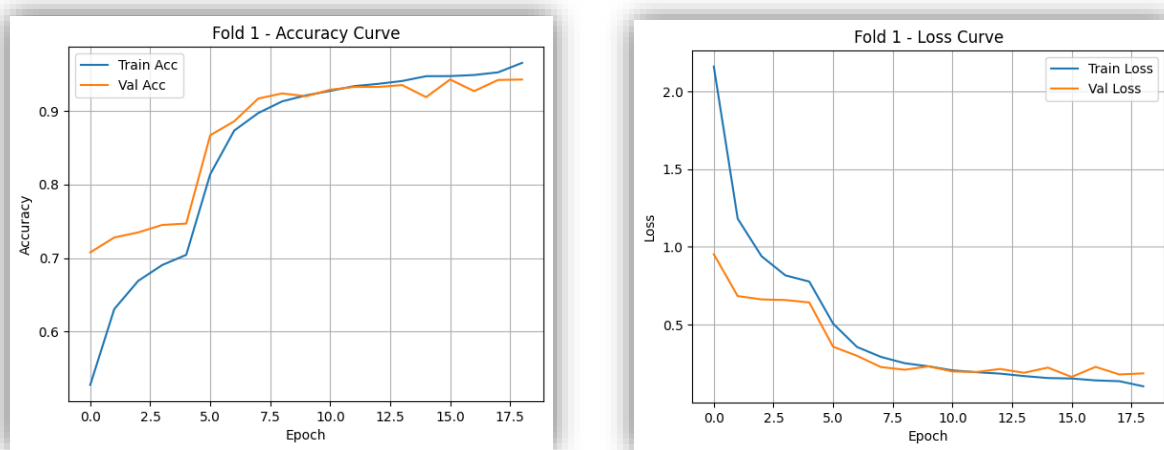


Figure 21: Καμπύλες Train/Validation Accuracy και Loss για το VGG19 – Fold 1

Οι καμπύλες accuracy και loss 1 επιβεβαιώνουν τη σταθερή και γρήγορη σύγκλιση του μοντέλου. Η απόδοση στο validation set ήταν υψηλή ήδη από τις πρώτες εποχές, με την καμπύλη validation accuracy να ξεπερνά τη train accuracy σε αρκετά σημεία, γεγονός που ενδέχεται να οφείλεται στην επίδραση του data augmentation στο training set. Παράλληλα, η απουσία έντονων διακυμάνσεων ή απότομων αυξομειώσεων στην καμπύλη loss δείχνει σταθερότητα και καλή γενίκευση, ενώ η πολύ χαμηλή τελική τιμή του validation loss καθιστά το VGG-19 το πιο αποδοτικό μοντέλο μεταξύ των προσεγγίσεων που δοκιμάστηκαν.

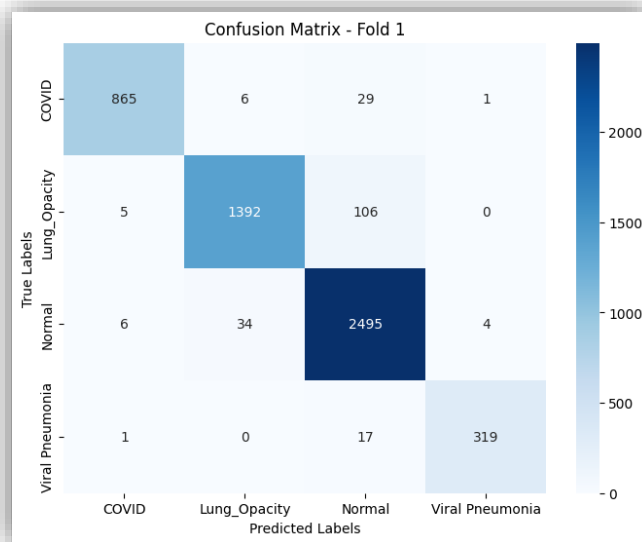


Figure 22: VGG19 Confusion Matrix - Fold 1

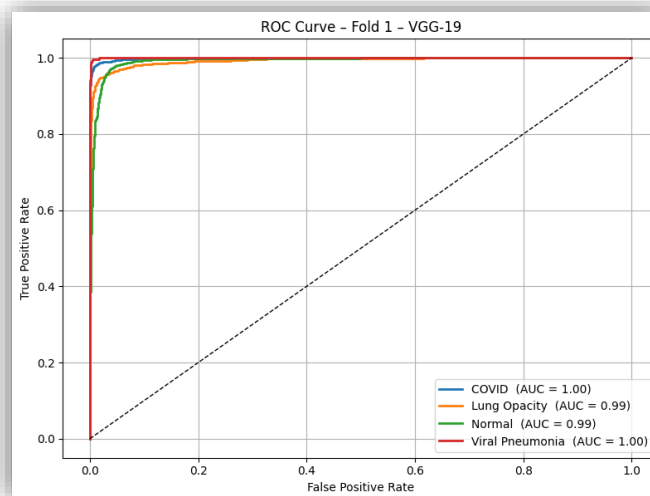


Figure 23: VGG19 Roc Curve για κάθε κλάση

Η Confusion Matrix του μοντέλου VGG-19 δείχνει εξαιρετική ακρίβεια στην ταξινόμηση των περισσότερων κατηγοριών, με ελάχιστα σφάλματα μεταξύ παρόμοιων παθολογιών. Ειδικά για τις κατηγορίες COVID και Viral Pneumonia, το μοντέλο παρουσιάζει σχεδόν τέλεια πρόβλεψη, γεγονός που επιβεβαιώνεται και από τις καμπύλες ROC όπου οι αντίστοιχοι δείκτες AUC φτάνουν το 1.00. Οι κατηγορίες Lung Opacity και Normal παρουσιάζουν ελαφρώς περισσότερες αστοχίες, όμως το μοντέλο καταφέρνει να διατηρήσει υψηλό διαχωριστικό δυναμικό και σε αυτές τις περιπτώσεις. Συνολικά, το VGG-19 δείχνει εξαιρετική ικανότητα διάκρισης ανάμεσα σε κλάσεις, επιτυγχάνοντας εξαιρετικές επιδόσεις τόσο σε επίπεδο ακρίβειας όσο και ευαισθησίας.

4.2 Συγκριτική Ανάλυση

Μετά την αναλυτική παρουσίαση της απόδοσης κάθε μοντέλου ξεχωριστά, σε αυτή την ενότητα πραγματοποιείται συγκριτική αξιολόγηση μεταξύ όλων των δοκιμασθέντων μοντέλων βάσει των βασικών μετρικών απόδοσης: accuracy, precision, recall και F1-score. Η ανάλυση αυτή αποσκοπεί στην ανάδειξη του μοντέλου με τη βέλτιστη συνολική συμπεριφορά και γενικευσιμότητα, λαμβάνοντας υπόψη και τη σταθερότητα της απόδοσης σε όλες τις κλάσεις.

Table 3: Μετρικές απόδοσης για κάθε μοντέλο

	Accuracy	Precision	Recall	F1 Score
CNN from scratch	0.89	0.89	0.89	0.89
MobileNetV2	0.91	0.91	0.91	0.91
EfficientNetB0	0.90	0.90	0.90	0.90
VGG19	0.95	0.95	0.95	0.95
DenseNet121	0.91	0.91	0.91	0.91

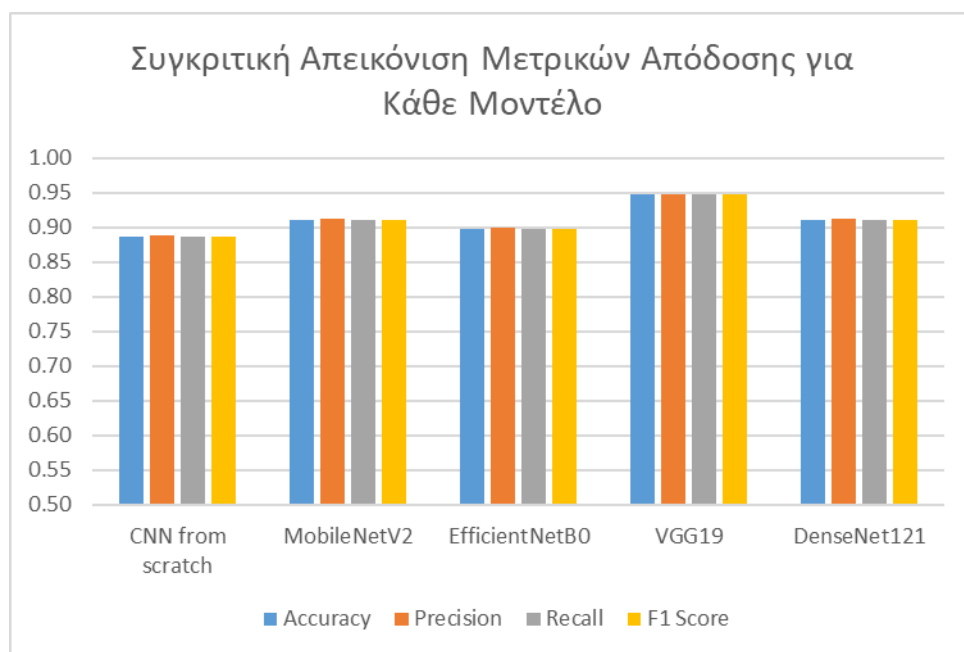


Figure 24: Συγκριτική Απεικόνιση Μετρικών Απόδοσης για Κάθε Μοντέλο

Όπως απεικονίζεται στον πίνακα και στο γράφημα, το μοντέλο VGG-19 ξεχωρίζει εμφανώς, επιτυγχάνοντας τις υψηλότερες τιμές σε όλες τις μετρικές, με accuracy και F1-score που αγγίζουν το 0.95. Τα μοντέλα MobileNetV2 και DenseNet121 ακολουθούν με πολύ κοντινές επιδόσεις γύρω στο 0.91, παρουσιάζοντας αξιόπιστη και ισορροπημένη συμπεριφορά. Το EfficientNetB0, αν και ελαφρώς υποδεέστερο, διατήρησε ικανοποιητική απόδοση παρόμοια με το CNN από το μηδέν, το οποίο παρά την απλότητά του πέτυχε αξιοπρεπείς μετρήσεις κοντά στο 0.89.

Η συνολική εικόνα του διαγράμματος αποτυπώνει καθαρά την υπεροχή των προεκπαιδευμένων μοντέλων και τη σημασία της χρήσης transfer learning σε προβλήματα ταξινόμησης με σύνθετα χαρακτηριστικά όπως οι ακτινογραφίες θώρακα.

Σε προβλήματα με μεγάλο όγκο δεδομένων, ο απαιτούμενος χρόνος εκπαίδευσης και αξιολόγησης κάθε μοντέλου αποτελεί κρίσιμο παράγοντα. Για τον λόγο αυτό, παρουσιάζεται το ακόλουθο διάγραμμα, το οποίο απεικονίζει τη συνολική χρονική διάρκεια (σε λεπτά) που απαιτήθηκε για την εκπαίδευση και αξιολόγηση κάθε μοντέλου στο πλαίσιο της παρούσας μελέτης.

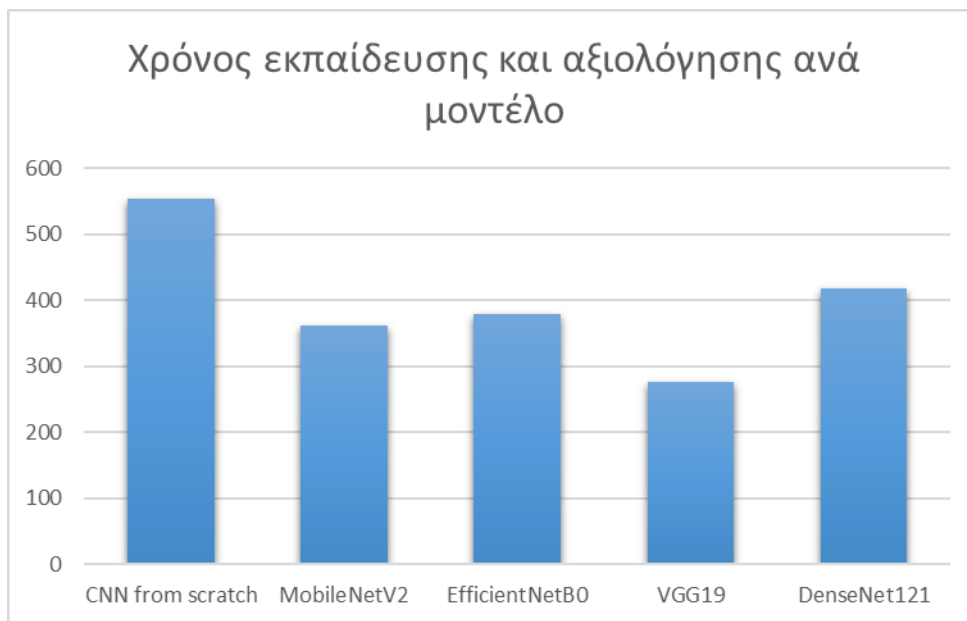


Figure 25: Χρόνος εκπαίδευσης και αξιολόγησης ανά μοντέλο

Παρατηρείται ότι το CNN from scratch απαιτεί σημαντικά περισσότερο χρόνο για εκπαίδευση και αξιολόγηση, ξεπερνώντας τα 540 λεπτά, γεγονός που οφείλεται στην απουσία προεκπαιδευμένων χαρακτηριστικών και στην ανάγκη εκμάθησης από μηδενική βάση. Αντιθέτως, το VGG19 ήταν το ταχύτερο μοντέλο (περίπου 280 λεπτά), κυρίως λόγω της ταχείας ενεργοποίησης του μηχανισμού EarlyStopping. Τα υπόλοιπα μοντέλα (MobileNetV2, EfficientNetB0 και DenseNet121) κυμάνθηκαν μεταξύ 360 και 420 λεπτών, με το DenseNet121 να είναι το πιο αργό μεταξύ των transfer learning μοντέλων, εξαιτίας της μεγαλύτερης πολυπλοκότητας και του αριθμού επιπέδων που ξεκλειδώθηκαν για fine-tuning.

Τέλος, παρουσιάζεται το παρακάτω διάγραμμα απεικονίζει το οποίο την απόδοση του μοντέλου VGG19 ανά διαγνωστική κατηγορία (COVID, Lung Opacity, Normal, Viral Pneumonia) ως προς τέσσερις βασικές μετρικές αξιολόγησης. Το VGG19 επιλέχθηκε ως το πιο αποδοτικό μοντέλο transfer learning βάσει της συνολικής του απόδοσης στο προηγούμενο κεφάλαιο, γι' αυτό και παρουσιάζεται αναλυτικά η συμπεριφορά του ανά κλάση.

Table 4: Μετρικές απόδοσης για κάθε κλάση (VGG19)

VGG19	Accuracy	Precision	Recall	F1-Score
COVID	0.991	0.986	0.960	0.973
Lung Opacity	0.971	0.972	0.926	0.949
Normal	0.963	0.943	0.983	0.962
Viral Pneumonia	0.996	0.985	0.947	0.965

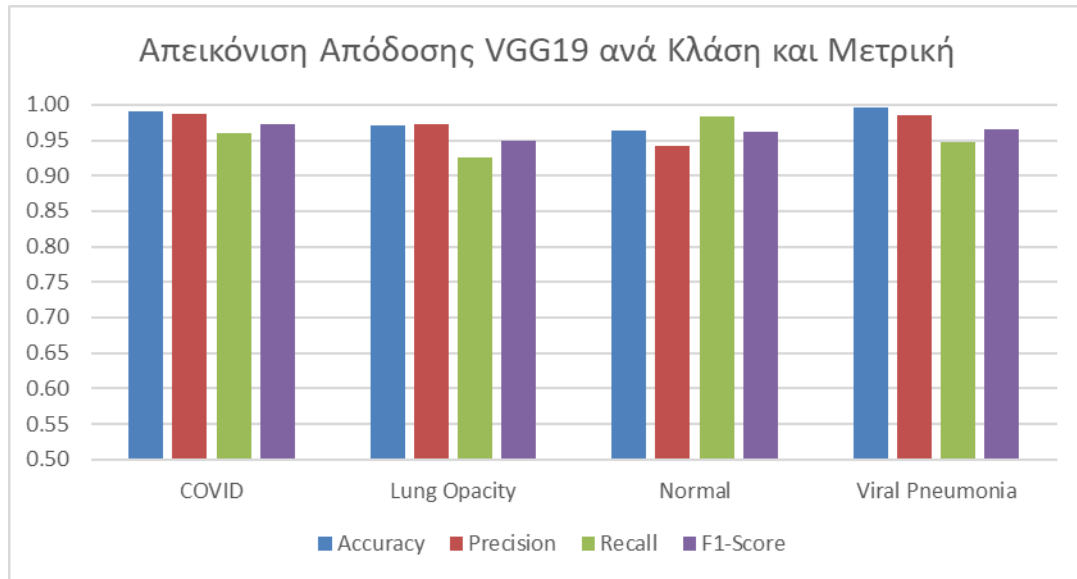


Figure 26: Απεικόνιση Απόδοσης VGG19 ανά κλάση και μετρική

Παρατηρείται ότι οι κλάσεις COVID και Viral Pneumonia επιτυγχάνουν τις υψηλότερες τιμές σε όλες σχεδόν τις μετρικές, γεγονός που υποδηλώνει ιδιαίτερα καλή ικανότητα διάκρισης αυτών των παθήσεων από το μοντέλο. Η κλάση Lung Opacity εμφανίζει υψηλή Precision αλλά σχετικά μειωμένο Recall, στοιχείο που φανερώνει ότι το μοντέλο είναι πιο «συντηρητικό» ως προς την απόδοση αυτής της ετικέτας – δηλαδή κάνει λιγότερα false positives, αλλά ενδέχεται να παραλείψει ορισμένα πραγματικά περιστατικά. Αντίστοιχα, η κλάση Normal παρουσιάζει τη χαμηλότερη Precision, πιθανόν λόγω της οπτικής ομοιότητας με άλλες παθολογικές κατηγορίες.

Συνολικά, οι επιδόσεις του μοντέλου VGG19 παραμένουν πολύ υψηλές και ισορροπημένες σε όλες τις κατηγορίες, με όλες τις μετρικές να κυμαίνονται πάνω από το 0.94, επιβεβαιώνοντας την αξιοπιστία του για πολύ-κλασική ταξινόμηση ακτινογραφιών θώρακα.

5. Συμπεράσματα

Η παρούσα μελέτη ανέδειξε τη δυναμική της βαθιάς μάθησης στην ταξινόμηση ακτινογραφιών θώρακα για τον εντοπισμό πνευμονικών παθήσεων, με ιδιαίτερη έμφαση στον COVID-19. Μέσω συγκριτικής αξιολόγησης μεταξύ ενός CNN μοντέλου κατασκευασμένου εξ αρχής και τεσσάρων προεκπαιδευμένων μοντέλων μεταφοράς μάθησης, επιβεβαιώθηκε η υπεροχή των τελευταίων ως προς την ακρίβεια και την ταχύτητα σύγκλισης.

Συνοψίζοντας τα πειραματικά αποτελέσματα:

- Το VGG19 σημείωσε την υψηλότερη ακρίβεια και τη χαμηλότερη απώλεια, ενώ ήταν και το ταχύτερο σε χρόνο εκπαίδευσης λόγω ενεργοποίησης του early stopping.
- Το MobileNetV2 και το DenseNet121 κατέγραψαν παρόμοιες επιδόσεις, με το DenseNet να απαιτεί σημαντικά μεγαλύτερο χρόνο λόγω της αρχιτεκτονικής του.
- Το EfficientNetB0 παρουσίασε χειρότερη επίδοση μεταξύ των προεκπαιδευμένων μοντέλων, παραπλήσια με το απλό CNN, όμως με μικρότερο χρόνο εκπαίδευσης.
- Το CNN from scratch απέδωσε ικανοποιητικά, παρά την απλότητα και την απουσία προεκπαίδευσης, καταδεικνύοντας τη σημασία της σωστής αρχιτεκτονικής και των τεχνικών regularization.

Μελλοντική Εργασία

Για τη συνέχιση και βελτίωση της παρούσας εργασίας, προτείνονται τα εξής:

- Επέκταση του dataset ή χρήση άλλων βάσεων δεδομένων για μεγαλύτερη γενίκευση των αποτελεσμάτων.
- Πειραματισμός με πιο σύνθετες αρχιτεκτονικές ή συνδυασμό μοντέλων (ensemble methods).
- Διερεύνηση τεχνικών attention ή explainability (π.χ. Grad-CAM) για καλύτερη κατανόηση των αποφάσεων του μοντέλου.
- Χρήση τεχνικών hyperparameter tuning (όπως grid/random search ή Bayesian optimization) για περαιτέρω βελτιστοποίηση.

Η παρούσα εργασία θέτει ένα ισχυρό υπόβαθρο για μελλοντική έρευνα στον τομέα της ιατρικής απεικόνισης μέσω βαθιάς μάθησης, ενισχύοντας την αξιοπιστία των αυτόματων μεθόδων διάγνωσης και υποβοήθησης των επαγγελματιών υγείας.

6. Βιβλιογραφία

1. Lamouadene, H., El Kassaoui, M., El Yadari, M., El Kenz, A., Benyoussef, A., El Moutaouakil, A., & Mounkachi, O. (2025). Detection of COVID-19, lung opacity, and viral pneumonia via X-ray using machine learning and deep learning. *Computers in Biology and Medicine*, 191(11013),
2. Patnaik, V., Mohanty, M., & Subudhi, A. K. (2025). CLASSIFICATION OF COVID-19 FROM CHEST X-RAY IMAGES USING ReSqNet. *Biomedical Engineering: Applications, Basis and Communications*, 2550009.
3. Patnaik, V., Mohanty, M., & Subudhi, A. K. (2024). DETECTION AND CLASSIFICATION OF COVID-19 USING GRAY-LEVEL FEATURES AND ENSEMBLE CLASSIFIER. *Biomedical Engineering: Applications, Basis and Communications*, 36(05), 2450025.
4. Sayeed, A., Khansur, N. O., Srizon, A. Y., Faruk, M. F., Alyami, S. A., Azad, A. K. M., & Moni, M. A. (2024). An effective screening of COVID-19 pneumonia by employing chest X-ray segmentation and attention-based ensembled classification. *IET Image Processing*, 18(9), 2400-2416.
5. Rahman, S., Sarker, S., Miraj, M. A. A., Nihal, R. A., Haque, A. K. M. N., & Noman, A. L. (2021). *Deep learning-driven automated detection of COVID-19 from radiography images: a comparative analysis*. *Cogn Comput*.
6. Ibrokhimov, B., & Kang, J. Y. (2022). Deep learning model for COVID-19-infected pneumonia diagnosis using chest radiography images. *BioMedInformatics*, 2(4), 654-670.
7. Islam, M. K., Habiba, S. U., Khan, T. A., & Tasnim, F. (2022). Cov-radnet: a deep convolutional neural network for automatic detection of Covid-19 from chest X-rays and CT scans. *Computer Methods and Programs in Biomedicine Update*, 2, 100064.
8. Bashar, A., Latif, G., Ben Brahim, G., Mohammad, N., & Alghazo, J. (2021). COVID-19 pneumonia detection using optimized deep learning techniques. *Diagnostics*, 11(11), 1972.