



ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ

**Σχεδιασμός και υλοποίηση
εφαρμογών σε ετερογενή
υπολογιστικά συστήματα**

ΘΕΟΔΩΡΑ Ε. ΣΑΝΙΔΑ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ

ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ

ΥΠΟΛΟΓΙΣΤΩΝ

ΚΟΖΑΝΗ, 2025

ΕΠΙΒΛΕΠΩΝ

Μηνάς Δασυγένης, Αναπληρωτής Καθηγητής Π.Δ.Μ.

ΤΡΙΜΕΛΗΣ**ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ**

Μηνάς Δασυγένης, Αναπληρωτής Καθηγητής Π.Δ.Μ.

Δημήτριος Σούντρης, Καθηγητής Ε.Μ.Π.

Σπυρίδων Νικολαΐδης, Καθηγητής Α.Π.Θ.

ΕΠΤΑΜΕΛΗΣ**ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ**

Μηνάς Δασυγένης, Αναπληρωτής Καθηγητής Π.Δ.Μ.

Δημήτριος Σούντρης, Καθηγητής Ε.Μ.Π.

Σπυρίδων Νικολαΐδης, Καθηγητής Α.Π.Θ.

Μαλαματή Λούτα, Καθηγήτρια Π.Δ.Μ.

Νικόλαος Πλόσκας, Αναπληρωτής Καθηγητής Π.Δ.Μ.

Μιχαήλ Δόσης, Καθηγητής Π.Δ.Μ.

Δημήτριος Καραμπατζάκης, Επίκουρος Καθηγητής ΔΙ.ΠΑ.Ε.

Η έγκριση της παρούσας διδακτορικής διατριβής από το τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας δεν σημαίνει αποδοχή των γνώμων του συγγραφέα.

*Στην οικογένεια μου που ήταν δίπλα μου
όλον αυτόν τον καιρό...*

Περίληψη

Τα τελευταία χρόνια έχει παρατηρηθεί αξιοσημείωτη αύξηση στη σχεδίαση και υλοποίηση εφαρμογών που διαφοροποιούνται από την κλασική αρχιτεκτονική των προσωπικών υπολογιστών, διακομιστών και φορητών υπολογιστών. Οι αρχιτεκτονικές αυτές αντιστοιχούν στα ετερογενή υπολογιστικά συστήματα (Heterogeneous Computing Systems, HCS), τα οποία ενσωματώνουν και συνδυάζουν διαφορετικούς τύπους επεξεργαστικών μονάδων με σκοπό την επιτάχυνση απαιτητικών εφαρμογών υψηλής υπολογιστικής πολυπλοκότητας, προσφέροντας ταυτόχρονα βελτιωμένη ενεργειακή απόδοση και αποδοτικότερη αξιοποίηση των διαθέσιμων υπολογιστικών πόρων.

Εφαρμογές, οι οποίες χαρακτηρίζονται από υψηλή υπολογιστική πολυπλοκότητα, αποτελούν βασικό πεδίο της βαθιάς μάθησης (Deep Learning, DL), η οποία έχει επιφέρει σημαντικές εξελίξεις και έχει εισαγάγει νέες δυνατότητες στην έρευνα της τεχνητής νοημοσύνης (Artificial Intelligence, AI). Η DL έχει αναδειχθεί ως μια από τις πιο δυναμικές και αποδοτικές προσεγγίσεις για την επίλυση σύνθετων προβλημάτων, αξιοποιώντας την ισχύ των νευρωνικών δικτύων για να μάθουν από τεράστια σύνολα δεδομένων και να πραγματοποιούν προβλέψεις με υψηλή ακρίβεια. Ιδιαίτερη έμφαση δίνεται στις εφαρμογές της όρασης υπολογιστών (Computer Vision, CV), οι οποίες έχουν ωφεληθεί σημαντικά από τις προόδους στις τεχνικές DL. Οι επιτυχίες σε αυτόν τον τομέα έχουν καθοδηγήσει την ερευνητική κοινότητα προς την υιοθέτηση βαθύτερων και πιο περίπλοκων μοντέλων, ενισχύοντας την ικανότητα των συστημάτων να αναγνωρίζουν και να αναλύουν οπτικά δεδομένα με υψηλή ακρίβεια. Οι εφαρμογές αυτές περιλαμβάνουν, μεταξύ άλλων, την αναγνώριση προτύπων, την ανάλυση εικόνας και την κατανόηση βίντεο, προσφέροντας σημαντικές βελτιώσεις σε τομείς όπως η ιατρική διάγνωση, η αυτόνομη οδήγηση και η παρακολούθηση ασφαλείας.

Η παρούσα διατριβή επικεντρώνεται στη σχεδίαση, υλοποίηση και επιτάχυνση μοντέλων υψηλής πολυπλοκότητας DL, με έμφαση στα συνελκτικά νευρωνικά δίκτυα (Convolutional Neural Networks, CNNs), σε ετερογενείς αρχιτεκτονικές.

Για την επίτευξη των παραπάνω, προτείνουμε μια μεθοδολογία για την ανάπτυξη επιταχυνόμενων και εξειδικευμένων αρχιτεκτονικών, κατάλληλων για εφαρμογές σε HCS. Η έρευνα στοχεύει στη σχεδίαση και βελτιστοποίηση αρχιτεκτονικών που αξιοποιούν πλήρως τις δυνατότητες προηγμένων HCS, όπως η προγραμματιζόμενη λογική επί του πεδίου με ενσωματωμένο σύστημα στο ίδιο τσιπ (Field-Programmable Gate Array – System on Chip, FPGA-SoC) και τα ενσωματωμένα συστήματα σε μονάδα (System on Module, SoM). Ο κύριος στόχος είναι η αύξηση της υπολογιστικής αποδοτικότητας μέσω της επιτάχυνσης των νευρωνικών μοντέλων, εξασφαλίζοντας παράλληλα την προσαρμοστικότητα των συστημάτων σε ένα ευρύ φάσμα εφαρμογών. Με την αξιοποίηση της παράλληλης επεξεργασίας, της ευελιξίας προγραμματισμού και της εξειδικευμένης αρχιτεκτονικής που προσφέρουν τα FPGA-SoC και τα SoM, τα συστήματα αυτά μπορούν να ανταποκριθούν στις αυξανόμενες απαιτήσεις των εφαρμογών της DL. Αυτή η στρατηγική προσφέρει ταυτόχρονα λύσεις υψηλής απόδοσης και μειωμένο περιβαλλοντικό αποτύπωμα, καθιστώντας τα HCS ιδανικά για εφαρμογές που απαιτούν ισορροπία μεταξύ ισχύος, ακρίβειας και ενεργειακής αποδοτικότητας.

Η αξιολόγηση των προτεινόμενων αρχιτεκτονικών, από τη χρήση της μεθοδολογίας που προτείνουμε, διεξάγεται σε πραγματικά σύνολα δεδομένων, εστιάζοντας σε δύο κρίσιμες εφαρμογές: την ιατρική διάγνωση και τον αγροτικό τομέα. Και οι δύο αυτοί τομείς χαρακτηρίζονται από υψηλές απαιτήσεις επεξεργασίας και ανάγκες για ακριβή και έγκαιρη ανάλυση δεδομένων. Στην ιατρική διάγνωση, η δυνατότητα ταχείας επεξεργασίας μεγάλων συνόλων δεδομένων, όπως ακτινογραφιών, παρέχει στους επαγγελματίες υγείας τα μέσα για την έγκαιρη ανίχνευση παθήσεων και τη λήψη βελτιωμένων θεραπευτικών αποφάσεων. Στον αγροτικό τομέα, η ανάλυση εικόνων από καλλιέργειες, συμβάλλει στην αποτελεσματική διαχείριση των γεωργικών πόρων και την πρόληψη ασθενειών.

Τα αποτελέσματα της παρούσας διατριβής καταδεικνύουν την αποδοτικότητα της μεθοδολογίας και των παραγόμενων αρχιτεκτονικών, υπογραμμίζοντας την ευελιξία τους να προσαρμόζονται στις διαφορετικές ανάγκες των εφαρμογών. Η αξιοποίηση των HCS επιτρέπει την ευέλικτη κατανομή των υπολογιστικών πόρων ανάλογα με την πολυπλοκότητα και τον όγκο των δεδομένων, διασφαλίζοντας υψηλή απόδοση, μειωμένη καθυστέρηση και ενεργειακή αποδοτικότητα. Οι προτεινόμενες αρχιτεκτονικές συμβάλλουν σημαντικά στην επίτευξη λύσεων υψηλής απόδοσης για εφαρμογές στους τομείς της υγείας, της γεωργίας και άλλων βιομηχανικών εφαρμογών με απώτερο στόχο την επίτευξη ισορροπίας μεταξύ της ακρίβειας, ταχύτητας και ενεργειακού αποτυπώματος.

Λέξεις Κλειδιά: Ετερογενή Υπολογιστικά Συστήματα, Επιτάχυνση Υλικού, Βελτιστοποίηση Υλικού, Τεχνητή Νοημοσύνη, Οραση Υπολογιστών, Βαθιά Νευρωνικά Δίκτυα.

Abstract

In recent years, there has been a significant growth in the design and implementation of applications that differ from the traditional architecture of personal computers, servers, and laptops. These architectures correspond to heterogeneous computing systems (HCS), which integrate and combine different types of processing units in order to accelerate demanding applications with high computational complexity, while also offering improved energy efficiency and more efficient utilization of available computing resources.

Applications characterized by high computational complexity constitute a primary domain of deep learning (DL), which has brought significant advancements and introduced new capabilities in artificial intelligence (AI) research. DL has emerged as one of the most dynamic and effective approaches for solving complex problems by leveraging the power of neural networks to learn from massive datasets and generate highly accurate predictions. Particular emphasis has been placed on applications in Computer Vision (CV), which have greatly benefited from advances in DL techniques. Breakthroughs in this area have led the research community to adopt deeper and more complex models, enhancing the ability of systems to recognize and interpret visual data with high precision. These applications include, among others, pattern recognition, image analysis, and video understanding, contributing substantial improvements in domains such as medical diagnosis, autonomous driving, and security monitoring.

This dissertation focuses on the design, implementation, and acceleration of high-complexity DL models, with emphasis on Convolutional Neural Networks (CNNs), on heterogeneous architectures. To achieve these goals, we propose a methodology for the development of accelerated and specialized architectures suitable for applications on HCS. The research aims to design and optimize architectures that fully exploit the capabilities of advanced HCS, such as Field-Programmable Gate Array – System on Chip (FPGA-SoC) and System on Module (SoM) devices. The main objective is to enhance computational efficiency through neural model acceleration, while ensuring the adaptability of the systems to a broad range of applications.

By leveraging parallel processing, programmability, and architectural specialization offered by FPGA-SoCs and SoMs, such systems can meet the increasing demands of DL applications. This strategy enables high-performance solutions with reduced environmental footprint, making HCS ideal for scenarios requiring a balance between computational power, accuracy, and energy efficiency.

The evaluation of the proposed architectures, as implemented through the methodology we propose, is conducted on real-world datasets, focusing on two critical applications: medical diagnosis and the agricultural sector. Both of these domains are characterized by high processing demands and the need for accurate and timely data analysis. In medical diagnostics, the rapid processing of large-scale datasets such as X-ray images enables healthcare professionals to detect diseases early and make improved clinical decisions. In agriculture, image analysis of crops supports efficient resource management and early disease prevention.

The results of the present dissertation demonstrate the efficiency of the proposed methodology and the developed architectures, highlighting their flexibility to adapt to the diverse requirements of various applications. The exploitation of HCS allows for flexible allocation of computational resources based on data complexity and volume, ensuring high throughput, reduced latency, and energy-efficient processing. The proposed architectures significantly contribute to the realization of high-performance solutions in domains such as healthcare, agriculture, and other industrial applications, ultimately aiming to achieve a well-balanced trade-off among accuracy, speed, and energy footprint.

Keywords: Heterogeneous Computing Systems, Hardware Acceleration, Hardware Optimization, Artificial Intelligence, Computer Vision, Deep Neural Networks.

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες σε όλους εκείνους που συνέβαλαν, με άμεσο ή έμμεσο τρόπο, στην επιτυχή ολοκλήρωση της παρούσας διδακτορικής διατριβής. Η πορεία αυτή δεν θα ήταν εφικτή χωρίς την πολύτιμη καθοδήγηση, την ακαδημαϊκή υποστήριξη, καθώς και την ηθική ενίσχυση ανθρώπων που στάθηκαν δίπλα μου όλα αυτά τα χρόνια. Η συμβολή τους υπήρξε καθοριστική καθ' όλη τη διάρκεια αυτής της ερευνητικής αλλά και δημιουργικής πορείας.

Πρώτα απο όλα θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον αναπληρωτή καθηγητή του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας, Δασυγένη Μηνά, για την καθοδήγηση, την αμέριστη συμπαράσταση, την πίστη και την εμπιστοσύνη που έδειξε στο πρόσωπό μου καθ' όλη τη διάρκεια αυτής της έρευνας. Η συνεισφορά του υπήρξε ανεκτίμητη και καθοριστική για την ολοκλήρωση αυτής της διατριβής.

Επιπλέον, θα ήθελα να ευχαριστήσω τα μέλη της διδακτορικής μου συμβουλευτικής επιτροπής, τους καθηγητές Σούντρη Δημήτριο και Νικολαΐδη Σπυρίδωνα, για τα διορατικά σχόλια, τις συστάσεις και την τεχνογνωσία τους. Η πολύτιμη συμβολή τους βελτίωσε τη συνολική ποιότητα αυτής της διατριβής και παρείχε σημαντικές κατευθύνσεις για την έρευνά μου.

Ακόμη, θα ήθελα να ευχαριστήσω όλους τους συν-συγγραφείς των ερευνητικών μας δημοσιεύσεων. Η εποικοδομητική συνεργασία μας, οι ανταλλαγές ιδεών, η ουσιαστική βοήθεια και η καθοδήγησή τους συνέβαλαν σημαντικά στην εξέλιξη και την επιτυχία των ερευνητικών μας προσπαθειών.

Τέλος, το σημαντικότερο απ' όλα, εκφράζω την ευγνωμοσύνη μου στην οικογένειά μου, για την αδιάκοπη συμπαράσταση και κατανόηση που έδειξαν καθ' όλη τη διάρκεια αυτής της διαδικασίας. Η αφοσίωσή τους, όλα αυτά τα χρόνια, υπήρξε το θεμέλιο πάνω στο οποίο μπόρεσα να οικοδομήσω την ακαδημαϊκή μου πορεία.

Δημοσιεύσεις

Μέρος του υλικού που παρουσιάζεται στην παρούσα διδακτορική διατριβή έχει ήδη δημοσιευθεί σε επιστημονικά περιοδικά και πρακτικά διεθνών συνεδρίων με κρίση. Οι δημοσιεύσεις αυτές περιλαμβάνουν αναλύσεις, πειραματικά αποτελέσματα και τεχνικές που συνεισφέρουν στη συνολική έρευνα και ανάπτυξη της διατριβής.

Διεθνή Περιοδικά (με κρίση)

1. Sanida, T., & Dasygenis, M. (2024). MiniTomatoNet: A lightweight CNN for tomato leaf disease recognition on heterogeneous FPGA-SoC. The Journal of Supercomputing, 80(15), 21837-21866. **(Scimago Q2)**
2. Sanida, T., & Dasygenis, M. (2024). A novel lightweight CNN for chest X-ray-based lung disease identification on heterogeneous embedded system. Applied Intelligence, 54(6), 4756-4780. **(Scimago Q2)**
3. Sanida, T., Sanida, M. V., Sideris, A., & Dasygenis, M. (2024). Optimizing Lung Condition Categorization through a Deep Learning Approach to Chest X-ray Image Analysis. BioMedInformatics, 4(3), 2002-2021. **(Scimago Q2)**
4. Sanida, T., Sanida, M. V., Sideris, A., & Dasygenis, M. (2024). Enhancing Pulmonary Diagnosis in Chest X-rays through Generative AI Techniques. J, 7(3), 302-318.
5. Sanida, T., Sideris, A., Sanida, M. V., & Dasygenis, M. (2023). Tomato leaf disease identification via two-stage transfer learning approach. Smart Agric. Technol. 5, 100275 (2023). **(Scimago Q1)**
6. Sanida, T., Tabakis, I. M., Sanida, M. V., Sideris, A., & Dasygenis, M. (2023). A robust hybrid deep convolutional neural network for COVID-19 disease identification from chest X-ray images. Information, 14(6), 310. **(Scimago Q2)**

7. Sanida, T., Tsiktisiris, D., Sideris, A., & Dasygenis, M. (2022). A heterogeneous implementation for plant disease identification using deep learning. Multimedia Tools and Applications, 81(11), 15041-15059. **(Scimago Q1)**
8. Sanida, T., Sideris, A., Tsiktisiris, D., & Dasygenis, M. (2022). Lightweight neural network for COVID-19 detection from chest X-ray images implemented on an embedded system. Technologies, 10(2), 37.

Πρακτικά Διεθνών Συνεδρίων (με κρίση)

1. Sanida, T., Sideris, A., Sanida, M. V., Dossis, M., & Dasygenis, M. (2024, September). Accelerating CNNs for Pneumonia Disease Diagnosis via Heterogeneous FPGA Systems. In 2024 9th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 159-162). IEEE.
2. Sanida, T., Sideris, A., Sanida, M. V., Dossis, M., & Dasygenis, M. (2023, November). An Efficiency CNN Solution for Olive Disease Management Through FPGA. In 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
3. Sanida, T., Sanida, M. V., Sideris, A., Dossis, M., & Dasygenis, M. (2023, November). Efficient categorization of pneumonia diagnosis using low-power embedded devices. In 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
4. Sanida, T., Sideris, A., Sanida, M. V., Dossis, M., & Dasygenis, M. (2023, November). Acceleration of GANs for Potato Crop Disease Identification via FPGA. In 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
5. Sanida, T., Sanida, M. V., Sideris, A., & Dasygenis, M. (2023, June). A lightweight CNN model for tomato crop diseases on heterogeneous embedded system. In 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCASST) (pp. 1-4). IEEE.

6. Sanida, T., Sideris, A., Chatzisavvas, A., Dossis, M., & Dasygenis, M. (2022, September). Radiography images with transfer learning on embedded system. In 2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
7. Sanida, T., Tsiktisiris, D., Sideris, A., & Dasygenis, M. (2021, July). A heterogeneous lightweight network for plant disease classification. In 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST) (pp. 1-4). IEEE.

Περιεχόμενα

Περίληψη	vi
Abstract	x
Ευχαριστίες	xiii
Δημοσιεύσεις	xv
Κατάλογος Σχημάτων	xxiii
Κατάλογος Πινάκων	xxvi
Κατάλογος Αλγορίθμων	xxix
Συντομογραφίες - Ακρωνύμια	xxx
Ελληνικές Ορολογίες	xxxiii
1 Εισαγωγή	1
1.1 Ορισμός του Ερευνητικού Προβλήματος	1
1.2 Κίνητρο Διατριβής	3
1.3 Ερευνητικά Κενά και Προκλήσεις	5
1.4 Ερευνητικοί Στόχοι και Συνεισφορές	6
1.5 Διάρθρωση Διατριβής	8
2 Θεωρητικό Υπόβαθρο	10
2.1 Ετερογενή Υπολογιστικά Συστήματα	10
2.2 Τεχνητή Νοημοσύνη	13
2.3 Μηχανική Μάθηση	16
2.4 Βαθιά Μάθηση	18
2.5 Βασικά στοιχεία των CNNs	21
2.5.1 Συνελικτικό Επίπεδο	22
2.5.2 Επίπεδο Συγκέντρωσης	23
2.5.3 Συναρτήσεις Ενεργοποίησης	24

2.5.4	Κανονικοποίηση Παρτίδας	27
2.5.5	Απόσυρση	27
2.5.6	Πλήρως Συνδεδεμένο Επίπεδο	28
2.6	Επιτάχυνση Υλικού στην AI	29
2.6.1	Ετερογενείς Αρχιτεκτονικές Επιτάχυνσης	29
2.6.2	Τεχνικές Επιτάχυνσης Υλικού και Βελτιστοποίηση	31
2.6.3	Αλληλεπίδραση Επιτάχυνσης Υλικού και Αποδοτικότητας σε HCS	33
2.7	Σύνοψη Κεφαλαίου	34
3	Μεθοδολογία Σχεδίασης και Ανάπτυξης σε HCS	36
3.1	Μεθοδολογική Προσέγγιση	36
3.2	Μετρικές Σχεδιασμού	39
3.2.1	Αρχιτεκτονική Μοντέλου	39
3.2.1.1	Εξισορρόπηση Βάθους και Εύρους	39
3.2.1.2	Χαμηλής Πολυπλοκότητας Δομοστοιχεία	40
3.2.1.3	Συναρτήσεις Ενεργοποίησης Φιλικές προς το Υλικό	44
3.2.2	Τεχνικές Βελτιστοποίησης	48
3.2.2.1	Κβαντοποίηση	48
3.2.2.2	Κλάδεμα	51
3.2.2.3	Διαχείριση Ανισορροπίας Κλάσεων	53
3.2.3	Συμβιβασμοί Επιδόσεων	54
3.2.4	Ενεργειακός Σχεδιασμός	56
3.3	Μετρικές Ποιότητας	57
3.4	Μετρικές Υπολογιστικού Αποτυπώματος	59
3.4.1	Φάση Εκπαίδευσης	60
3.4.2	Φάση Εξαγωγής Συμπερασμάτων	64
3.4.3	Φάση Εκπαίδευσης και Εξαγωγής Συμπερασμάτων	68
3.5	Συνολική Αξιολόγηση Καταλληλότητας Μοντέλων ως προς τις Μετρικές Αποδοτικότητας	72
3.6	Σύνοψη Κεφαλαίου	73
4	Σχεδιασμός και Υλοποίηση Επιτάχυνσης σε FPGA–SoC	74
4.1	Περίληψη	74
4.2	Εισαγωγή	76
4.3	Επισκόπηση Επιστημονικού Πεδίου	80
4.4	Υλικά και Μέθοδοι	85
4.4.1	Σύνολο Δεδομένων Εικόνας	85
4.4.2	Αρχιτεκτονική Δικτύου MiniTomatoNet	87
4.4.3	Συνάρτηση Επανασταθμισμένης Εστιακής Απώλειας	90
4.5	Επισκόπηση Αρχιτεκτονικής Συστήματος	94
4.5.1	Μεθοδολογία Σχεδιασμού και Αρχιτεκτονική Υλοποίησης	94
4.5.2	Πλακέτα PYNQ–Z1	95
4.5.3	Αρχιτεκτονική Υλικού	96

4.5.4	Σχεδίαση Αρθρωτών Κυκλωμάτων του MiniTomatoNet	97
4.5.5	Τεχνικές Βελτιστοποίησης HLS του MiniTomatoNet σε Ετερογενή FPGA—SoC	98
4.5.5.1	Ξετύλιγμα Βρόχων	100
4.5.5.2	Διασωλήνωση	102
4.5.5.3	Κατατμήση Πινάκων	102
4.5.5.4	Επιπέδωση Βρόχων	103
4.6	Πειραματικά Αποτελέσματα	103
4.6.1	Φάση Εκπαίδευσης και Επικύρωσης	103
4.6.1.1	Παράμετροι Εκπαίδευσης	103
4.6.1.2	Καμπύλες Ακρίβειας/Απώλειας Εκπαίδευσης και Επικύρωσης	105
4.6.2	Φάση Ελέγχου	106
4.6.2.1	Αξιολόγηση με Μήτρες Σύγχυσης	106
4.6.2.2	Σύγκριση Μετρικών Ποιότητας	106
4.6.3	Επιλογή Μεθόδου Συνελκτικής Επεξεργασίας	108
4.6.4	Υπολογιστικό Κόστος των Τύπων Δεδομένων	112
4.6.5	Υπολογιστικό Κόστος των Τεχνικών Βελτιστοποίησης	114
4.6.6	Συγκριτική Ανάλυση σε Διαφορετικές Υπολογιστικές Συσκευές	116
4.7	Συζήτηση Αποτελεσμάτων	118
4.8	Συμπεράσματα και Μελλοντικές Επεκτάσεις	120
5	Σχεδιασμός και Υλοποίηση Επιτάχυνσης σε SoM	123
5.1	Περίληψη	123
5.2	Εισαγωγή	124
5.3	Επισκόπηση Επιστημονικού Πεδίου	131
5.4	Υλικά και Δεδομένα	138
5.4.1	Σύνολο Δεδομένων Εικόνας	139
5.4.2	Προεπεξεργασία Δεδομένων	140
5.4.3	Διαχωρισμός του Συνόλου Δεδομένων	142
5.5	Μεθοδολογία Σχεδιασμού και Υλοποίησης του Συστήματος	144
5.5.1	Συνάρτηση Εστιακής Απώλειας	144
5.5.2	Σχεδιασμός CNN Χαμηλής Πολυπλοκότητας	146
5.5.3	Σχεδιασμός Προεκπαιδευμένων CNNs	148
5.6	Πειραματικά Αποτελέσματα	153
5.6.1	Ετερογενές Σύστημα Επιτάχυνσης NVIDIA Jetson AGX Xavier	153
5.6.2	Μετρικές Ποιότητας	155
5.6.3	Φάση Εκπαίδευσης και Επικύρωσης	156
5.6.3.1	Χρόνος Εκπαίδευσης και Ελέγχου	157
5.6.3.2	Καμπύλες Μέσης Ακρίβειας Εκπαίδευσης και Επικύρωσης	159
5.6.3.3	Καμπύλες Μέσης Απώλειας Εκπαίδευσης και Επικύρωσης	160
5.6.4	Φάση Ελέγχου	161

5.6.4.1	Αξιολόγηση με Μήτρες Σύγχυσης	161
5.6.4.2	Διακριτική Ικανότητα μέσω Καμπυλών ROC	163
5.6.4.3	Σύγκριση Μετρικών Ποιότητας	165
5.6.4.4	Ανάλυση Χρόνου Εκτέλεσης Εξαγωγής Συμπερασμάτων	170
5.7	Συζήτηση Ευρημάτων	171
5.8	Συμπεράσματα και Μελλοντικές Επεκτάσεις	175
6	Συμπεράσματα	178
6.1	Κύριες Συνεισφορές της Διατριβής	179
6.2	Περίληψη Ευρημάτων	179
6.3	Προτάσεις για Μελλοντική Έρευνα	182
A'	Δημοσιεύσεις	227
A'.1	Περιοδικά	227
A'.2	Συνέδρια	228

Κατάλογος Σχημάτων

2.1	Η ιεραρχική σχέση μεταξύ AI, ML και DL.	21
3.1	Διαγραμματική απεικόνιση της πολυκριτηριακής μεθοδολογίας αξιολόγησης καταλληλότητας CNNs σε HCS.	38
3.2	Μεθοδολογική προσέγγιση σχεδίασης και ανάπτυξης των CNNs σε HCS.	39
4.1	Διάγραμμα ροής εργασίας που απεικονίζει τα βήματα για την επίτευξη των στόχων αναγνώρισης ασθενειών στα φύλλα τομάτας. . .	79
4.2	Συγκριτική απεικόνιση της συνολικής ακρίβειας που επιτεύχθηκε από διάφορα αρχιτεκτονικά σχήματα CNNs, όπως έχουν παρουσιαστεί στη διεθνή βιβλιογραφία, για την αναγνώριση ασθενειών τομάτας.	85
4.3	Ενδεικτικά δείγματα διαφορετικών φύλλων τομάτας. Ένα φύλλο για κάθε ασθένεια και ένα υγιές φύλλο τομάτας.	88
4.4	Συχνότητα κάθε κλάσης στο σύνολο δεδομένων τομάτας. Η σημαντική διαφοροποίηση στο πλήθος δειγμάτων μεταξύ των κλάσεων αναδεικνύει την ύπαρξη ανισορροπίας στο σύνολο δεδομένων. .	88
4.5	Δομική απεικόνιση του MiniTomatoNet για την αναγνώριση ασθενειών της τομάτας.	91
4.6	Αρχιτεκτονική επισκόπηση υλικού της πλατφόρμας PYNQ–Z1 που χρησιμοποιείται για την αναγνώριση ασθενειών της τομάτας.	97
4.7	Μπλοκ διάγραμμα υλικού με τους IP πυρήνες της αρχιτεκτονικής MiniTomatoNet.	99
4.8	Συγκριτική απεικόνιση των καμπυλών ακρίβειας/απώλειας εκπαίδευσης και επικύρωσης, με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.	105
4.9	Μήτρες σύγχυσης με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.	107
4.10	Συγκριτική απεικόνιση των μέσων όρων των μετρικών ποιότητας, με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας. .	109
4.11	Συγκριτική απεικόνιση των διαφορετικών μεθόδων συνελκτικής επεξεργασίας στο μοντέλο MiniTomatoNet.	112

4.12	Συγκριτική απεικόνιση του υπολογιστικού κόστους υλικού των διαφορετικών τύπων δεδομένων (FP32 και FP16) στο μοντέλο MiniTomatoNet.	113
4.13	Συγκριτική απεικόνιση του υπολογιστικού κόστους υλικού των τεχνικών βελτιστοποίησης (ξετύλιγμα βρόχων, διασωλήνωση, κατάτμηση πινάκων και επιπέδωση βρόχων) στο μοντέλο MiniTomatoNet.	115
5.1	Συγκριτική απεικόνιση της συνολικής ακρίβειας που επιτεύχθηκε από διάφορα αρχιτεκτονικά σχήματα CNNs, όπως έχουν παρουσιαστεί στη διεθνή βιβλιογραφία, για την αναγνώριση των πνευμονικών παθήσεων.	135
5.2	Δομή του συστήματος για την αναγνώριση των πνευμονικών παθήσεων με πενταπλή διασταυρούμενη επικύρωση από το ενοποιημένο σύνολο δεδομένων ακτινογραφιών θώρακα.	139
5.3	Ενδεικτικά δείγματα ακτινογραφιών θώρακα από το ενοποιημένο σύνολο δεδομένων. Κάθε δείγμα αντιστοιχεί σε μία από τις επτά κλάσεις.	141
5.4	Κατανομή του πλήθους εικόνων ακτινογραφιών θώρακα ανά κλάση στο ενοποιημένο σύνολο δεδομένων. Η σημαντική διαφοροποίηση στο πλήθος δειγμάτων μεταξύ των κλάσεων αναδεικνύει την ύπαρξη ανισορροπίας στο σύνολο δεδομένων.	141
5.5	Δομική απεικόνιση του CNN χαμηλής πολυπλοκότητας που αναπτύχθηκε σύμφωνα με τη μεθοδολογία, για την κατηγοριοποίηση πνευμονικών παθήσεων από ακτινογραφίες θώρακα.	149
5.6	Δομική απεικόνιση της αρχιτεκτονικής των προεκπαιδευμένων CNNs για την ταξινόμηση των πνευμονικών παθήσεων από ακτινογραφίες θώρακα.	152
5.7	Συγκριτική γραφική απεικόνιση της μέσης συνολικής διάρκειας εκπαίδευσης (σε λεπτά) και του χρόνου ελέγχου (σε δευτερόλεπτα) για κάθε ένα από τα εξεταζόμενα CNNs.	159
5.8	Συγκριτική απεικόνιση της μέσης καμπύλης ακρίβειας εκπαίδευσης και επικύρωσης για τα εξεταζόμενα CNNs, μετά τη διαδικασία της πενταπλής διασταυρούμενης επικύρωσης, με εφαρμογή της συνάρτησης διασταυρούμενης εντροπίας και της εστιακής απώλειας.	160
5.9	Συγκριτική απεικόνιση της μέσης καμπύλης απώλειας εκπαίδευσης και επικύρωσης για τα εξεταζόμενα CNNs, μετά τη διαδικασία της πενταπλής διασταυρούμενης επικύρωσης, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.	162
5.10	Μήτρες σύγχυσης για κάθε εξεταζόμενο CNN μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης διασταυρούμενης εντροπίας.	163
5.11	Μήτρες σύγχυσης για κάθε εξεταζόμενο CNN μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της εστιακής συνάρτησης απώλειας.	164

- 5.12 Συγκριτική απεικόνιση του αριθμού λανθασμένων ταξινομήσεων για καθένα από τα επτά CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας. 165
- 5.13 Καμπύλες ROC για τα επτά εξεταζόμενα CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης απώλειας διασταυρούμενης εντροπίας. 166
- 5.14 Καμπύλες ROC για τα επτά εξεταζόμενα CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της εστιακής συνάρτησης απώλειας. 167
- 5.15 Συγκριτική απεικόνιση των τιμών AUC για κάθε μοντέλο, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας. 167
- 5.16 Συγκριτική απεικόνιση των μέσων όρων των μετρικών ποιότητας, για τα εξεταζόμενα CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας. 168
- 5.17 Σύγκριση του μέσου χρόνου εκτέλεσης εξαγωγής συμπερασμάτων για κάθε CNN, τόσο σε GPU όσο και σε CPU, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας. 172
- 5.18 Σύγκριση των καρτέ ανά δευτερόλεπτο για τα εξεταζόμενα CNN, κατά την εκτέλεση εξαγωγής συμπεράσματος σε GPU και CPU, τόσο σε GPU όσο και σε CPU, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας. 172

Κατάλογος Πινάκων

2.1	Κατηγοριοποίηση των πεδίων εφαρμογής της ΑΙ.	14
2.2	Συγκριτική αποτίμηση πλεονεκτημάτων και μειονεκτημάτων των βασικών κατηγοριών ML.	19
2.3	Συγκριτική αποτίμηση βασικών συναρτήσεων ενεργοποίησης ως προς τα πλεονεκτήματα και τα μειονεκτήματά τους.	26
2.4	Σύγκριση ετερογενών αρχιτεκτονικών υλικού για επιτάχυνση ΑΙ. . . .	31
3.1	Συγκριτική ανάλυση χαμηλής πολυπλοκότητας δομοστοιχείων CNNs. . . .	45
3.2	Συγκριτική ανάλυση χαμηλής πολυπλοκότητας συναρτήσεων ενεργοποίησης φιλικών προς το υλικό.	49
3.3	Σύγκριση αριθμητικών προτύπων FP32, FP16 και INT8 σε εφαρμογές CNN.	50
3.4	Κατηγοριοποίηση τεχνικών διαχείρισης ανισορροπίας κλάσεων. . . .	55
3.5	Σύνοψη βασικών μετρικών ποιότητας CNN.	59
3.6	Συγκριτική ανάλυση μετρικών εκπαίδευσης για CNNs σε HCS.	65
3.7	Συγκριτική ανάλυση μετρικών εξαγωγής συμπερασμάτων για CNNs σε HCS.	69
3.8	Συγκριτική ανάλυση μετρικών εκπαίδευσης και εξαγωγής συμπερασμάτων για CNNs σε HCS.	71
4.1	Σύνοψη της ερευνητικής βιβλιογραφίας σε προσεγγίσεις CNNs για την αναγνώριση των ασθενειών τομάτας.	86
4.2	Αναλυτική κατανομή του αριθμού εικόνων που χρησιμοποιήθηκαν για εκπαίδευση και έλεγχο κατά τη διαδικασία διαχωρισμού βάσει της μεθόδου συγκράτησης.	89
4.3	Οι προδιαγραφές της αρχιτεκτονικής MiniTomatoNet για την αναγνώριση ασθενειών της τομάτας.	91
4.4	Καθορισμός των υπερπαραμέτρων που χρησιμοποιήθηκαν για την εκπαίδευση του MiniTomatoNet.	104
4.5	Μέσοι όροι των μετρικών ποιότητας του MiniTomatoNet στα δεδομένα ελέγχου, με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.	109
4.6	Ακρίβεια των θετικών προβλέψεων ανά κατηγορία για το MiniTomatoNet με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.	110

4.7	Ανάκληση ανά κατηγορία για το MiniTomatoNet, με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.	110
4.8	Βαθμολογία F1 ανά κατηγορία για το MiniTomatoNet με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.	110
4.9	Σύγκριση διαφορετικών μεθόδων συνελκτικής επεξεργασίας στο μοντέλο MiniTomatoNet.	112
4.10	Υπολογιστικό κόστος υλικού του MiniTomatoNet για διαφορετικούς τύπους δεδομένων (FP32 και FP16).	113
4.11	Υπολογιστικό κόστος και καθυστέρηση του MiniTomatoNet των τεχνικών βελτιστοποίησης (ξετύλιγμα βρόχων, διασωλήνωση, κατάτμηση πινάκων και επιπέδωση βρόχων).	115
4.12	Επίδραση της υλοποίησης MiniTomatoNet σε διαφορετικές υπολογιστικές συσκευές.	118
5.1	Σύνοψη της σύγχρονης βιβλιογραφίας με τη χρήση CNNs για την ανίχνευση πνευμονικών παθήσεων μέσω ακτινογραφιών θώρακα. . . .	136
5.2	Κατανομή του συνολικού αριθμού των ακτινογραφιών θώρακα ανά κλάση στο τελικό ενοποιημένο σύνολο δεδομένων.	140
5.3	Αναλυτική κατανομή του αριθμού ακτινογραφιών θώρακα που χρησιμοποιήθηκαν για εκπαίδευση (Tr) και έλεγχο (Ts) σε κάθε μία από τις πέντε πτυχές της διαδικασίας πενταπλής διασταυρούμενης επικύρωσης.	143
5.4	Οι προδιαγραφές της αρχιτεκτονικής του CNN χαμηλής πολυπλοκότητας για την αναγνώριση πνευμονικών παθήσεων από ακτινογραφίες θώρακα.	150
5.5	Συγκριτική παρουσίαση προδιαγραφών των CNNs για την ταξινόμηση πνευμονικών παθήσεων από ακτινογραφίες θώρακα. . . .	153
5.6	Καθορισμός των υπερπαραμέτρων που χρησιμοποιήθηκαν για την εκπαίδευση των CNNs.	157
5.7	Αναλυτική παρουσίαση της μέσης διάρκειας εκπαίδευσης (σε λεπτά) και χρόνου ελέγχου (σε δευτερόλεπτα) για καθένα από τα εξεταζόμενα CNNs.	158
5.8	Μέσοι όροι των μετρικών ποιότητας, για τα εξεταζόμενα CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας. 168	
5.9	Μέσοι όροι ακρίβειας των θετικών προβλέψεων για τα εξεταζόμενα CNNs, με συνάρτηση απώλειας διασταυρούμενης εντροπίας και πενταπλή διασταυρούμενη επικύρωση.	169
5.10	Μέσοι όροι ακρίβειας των θετικών προβλέψεων για τα εξεταζόμενα CNNs, με χρήση της εστιακής συνάρτησης απώλειας και πενταπλή διασταυρούμενη επικύρωση.	169
5.11	Μέσοι όροι ανάκλησης για τα εξεταζόμενα CNNs, με συνάρτηση απώλειας διασταυρούμενης εντροπίας και πενταπλή διασταυρούμενη επικύρωση.	169

5.12	Μέσοι όροι ανάκλησης για τα εξεταζόμενα CNNs, με χρήση της εστιακής συνάρτησης απώλειας και πενταπλή διασταυρούμενη επικύρωση.	170
5.13	Μέσοι όροι βαθμολογίας F1 για τα εξεταζόμενα CNNs, με συνάρτηση απώλειας διασταυρούμενης εντροπίας και πενταπλή διασταυρούμενη επικύρωση.	170
5.14	Μέσοι όροι βαθμολογίας F1 για τα εξεταζόμενα CNNs, με χρήση της εστιακής συνάρτησης απώλειας και πενταπλή διασταυρούμενη επικύρωση.	170
5.15	Σύγκριση του μέσου χρόνου εκτέλεσης εξαγωγής συμπερασμάτων και των καρτέ ανά δευτερόλεπτο για κάθε CNN, σε GPU και CPU, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.	172

Κατάλογος Αλγορίθμων

4.5.1 Depthwise convolution layer — Επίπεδο συνέλιξης κατά βάθος.	99
4.5.2 Pointwise convolution layer — Επίπεδο σημειακής συνέλιξης.	100
4.5.3 Max—pool layer — Επίπεδο μέγιστης συγκέντρωσης.	101
4.5.4 Fully connected layer — Πλήρως συνδεδεμένο επίπεδο.	101

Συντομογραφίες - Ακρωνύμια

AI	Artificial Intelligence
ALU	Arithmetic Logic Unit
ARM	Advanced RISC Machine
ASIC	Application-Specific Integrated Circuit
AUC	Area Under the Curve
CAD	Computer-Aided Diagnosis
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
CV	Computer Vision
DCNN	Deep Convolutional Neural Network
DDR	Double Data Rate
DL	Deep Learning
DLA	Deep Learning Accelerator
DMA	Direct Memory Access
DRAM	Dynamic Random-Access Memory
DSP	Digital Signal Processing
FC	Fully Connected
FLOP	Floating-point Operations
FLOPs	Floating-point Operations Per Second
FPS	Frames Per Second
FPGA	Field Programmable Gate Array
GAN	Generative Adversarial Network

GB	Gigabyte
GPU	Graphics Processing Unit
HCS	Heterogeneous Computing System
HDL	Hardware Description Language
HDMI	High—Definition Multimedia Interface
HLS	High—Level Synthesis
HPC	High—Performance Computing systems
IoT	Internet of Things
ISA	Instruction Set Architecture
LR	Learning Rate
LPDDR	Low Power Double Data Rate
LUT	Look—up Table
ML	Machine Learning
NLP	Natural Processing Language
NPU	Neural Processing Unit
NVDLA	NVIDIA Deep Learning Accelerator
OS	Operating System
PCA	Principal Component Analysis
PCIe	Peripheral Component Interconnect Express
PE	Processing Element
PL	Programmable Logic
PTQ	Post—Training Quantization
QAT	Quantization—Aware Training
RAM	Random Access Memory
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SGD	Stochastic Gradient Descent
SoC	System—on—Chip
SoM	System—on—Module
TPU	Tensor Processing Unit
USB	Universal Serial Bus

ViT	Vision Transformer
VHDL	VHSIC Hardware Description Language
VHSIC	Very High Speed Integrated Circuit

Ελληνικές Ορολογίες

Accuracy	Ακρίβεια
Activation Function	Συνάρτηση Ενεργοποίησης
Artificial Intelligence	Τεχνητή Νοημοσύνη
AUC	Εμβαδόν Κάτω από την Καμπύλη ROC
Batch	Παρτίδα
Batch Normalization	Κανονικοποίηση Παρτίδας
Bias	Πόλωση
Capacity	Χωρητικότητα
Class Imbalance	Ανισορροπία Κατηγοριών
Classification	Κατηγοριοποίηση — Ταξινόμηση
Computer—Aided Diagnosis	Υποβοηθούμενη Διάγνωση από τον Υπολογιστή
Computer Vision	Όραση Υπολογιστών
Convolutional Layer	Συνελικτικό Επίπεδο
Convolutional Neural Networks	Συνελικτικά Νευρωνικά Δίκτυα
Cross—Entropy	Διασταυρούμενη Εντροπία
Data Augmentation	Ενίσχυση Δεδομένων
Data Imbalance	Ανισορροπία Δεδομένων
Deep Learning	Βαθιά Μηχανική Μάθηση
Dropout	Απόσυρση
Embedded System	Ενσωματωμένο Σύστημα
Epoch	Εποχή
F1—score	Σταθμισμένος Μέσος Ακρίβειας και Ανάκλησης
Feature	Χαρακτηριστικό

Flatten	Επίπεδωση
Five—Fold Cross—Validation	Πενταπλή Διασταυρούμενη Επικύρωση
Focal Loss	Εστιακή Συνάρτηση Απώλειας
Fully Connected Layer	Πλήρως Συνδεδεμένο Επίπεδο
Generalization	Γενίκευση
Global Average Pooling Layer	Επίπεδο Καθολικού Μέσου Συμψηφισμού
Heterogeneous Computing System	Ετερογενές Υπολογιστικό Σύστημα
Hidden Layer	Κρυφό Επίπεδο
Hyperparameter	Υπερπαραμέτρος
Input Layer	Επίπεδο Εισόδου
Inference Time	Χρόνος Πρόβλεψης
Layer	Επίπεδο
Learning Rate	Ρυθμός Μάθησης
Loss Function	Συνάρτηση Κόστους
Machine Learning	Μηχανική Μάθηση
MaxPooling	Μέγιστη Δειγματοληψία
Multiclass Classification	Πολυκατηγορική Ταξινόμηση
Neuron	Νευρώνας
Optimization	Βελτιστοποίηση
Optimizer	Βελτιστοποιητής
Overfitting	Υπερπροσαρμογή
Performance	Επίδοση
Pixel	Εικονοστοιχείο
Point—of—Care	Σημείο Περίθαλψης
Pooling Layer	Επίπεδο Συμψηφισμού
Precision	Ακρίβεια Θετικών Προβλέψεων
Recall	Ανάκληση
Sample	Δείγμα
Self—Supervised Learning	Αυτο—επιβλεπόμενη Μάθηση
Semi—Supervised Learning	Ημι—επιβλεπόμενη Μάθηση
Softmax	Συνάρτηση Εκθετικής Κανονικοποίησης

Supervised Learning	Επιβλεπόμενη Μάθηση
Test Set	Σύνολο Ελέγχου
Training Set	Σύνολο Εκπαίδευσης
Transfer Learning	Μεταφορά Μάθησης
Underfitting	Υποπροσαρμογή
Unrolling	Ξετύλιγμα
Unsupervised Learning	Μη Επιβλεπόμενη Μάθηση

Κεφάλαιο 1

Εισαγωγή

Το παρόν κεφάλαιο αποσκοπεί στην ανάδειξη του ερευνητικού προβλήματος που πραγματεύεται η διατριβή, το οποίο εστιάζει στον αποδοτικό σχεδιασμό, την υλοποίηση και την αξιολόγηση αρχιτεκτονικών για εφαρμογές συνελκτικών νευρωνικών δικτύων (Convolutional Neural Networks, CNNs) σε ετερογενή υπολογιστικά συστήματα (Heterogeneous Computing Systems, HCS). Αρχικά, παρουσιάζονται τα υφιστάμενα ερευνητικά κενά και οι τεχνολογικές προκλήσεις που απορρέουν από την αυξανόμενη πολυπλοκότητα των CNNs, ιδιαίτερα όταν αυτά καλούνται να υλοποιηθούν σε HCS με περιορισμένους πόρους. Η αναγνώριση αυτών των προκλήσεων οδηγεί στη διατύπωση των ερευνητικών στόχων της διατριβής, όπου αναλύεται η αναγκαιότητα ενσωμάτωσης πολυκριτηριακών μετρικών αποδοτικότητας, πέραν των παραδοσιακών μετρικών ποιότητας, με σκοπό τη διασφάλιση της λειτουργικής συμβατότητας, ενεργειακής αποδοτικότητας και υπολογιστικής αξιοπιστίας. Τέλος, το κεφάλαιο ολοκληρώνεται με τη διάρθρωση της διατριβής, παρέχοντας μια συνοπτική επισκόπηση των θεματικών ενοτήτων που θα αναπτυχθούν στα επόμενα κεφάλαια για την αντιμετώπιση των προκλήσεων και την επίτευξη των στόχων της παρούσας διατριβής.

1.1 Ορισμός του Ερευνητικού Προβλήματος

Η ραγδαία ανάπτυξη εφαρμογών που απαιτούν υψηλή υπολογιστική ισχύ καθιστά αναγκαία την υιοθέτηση υπολογιστικών αρχιτεκτονικών πέρα από τα παραδοσιακά πρότυπα των προσωπικών υπολογιστών, διακομιστών και φορητών συσκευών. Τα HCS ενσωματώνουν διαφορετικούς τύπους επεξεργαστικών μονάδων με σκοπό

την επιτάχυνση απαιτητικών εφαρμογών, επιτυγχάνοντας παράλληλα βελτιωμένη ενεργειακή απόδοση και βέλτιστη αξιοποίηση των διαθέσιμων υπολογιστικών πόρων, ώστε να ανταποκρίνονται στις αυξανόμενες ανάγκες των σύγχρονων υπολογιστικών περιβαλλόντων. Παράλληλα, η ταχεία πρόοδος της τεχνητής νοημοσύνης (Artificial Intelligence, AI), και ειδικότερα οι επαναστατικές εξελίξεις στην υπολογιστική όραση (Computer Vision, CV) έχει αναδείξει την επιτακτική ανάγκη για τον σχεδιασμό ευφύων και ενεργειακά αποδοτικών CNNs. Οι αρχιτεκτονικές αυτές οφείλουν να διασφαλίζουν υψηλή αξιοπιστία και υπολογιστική αποδοτικότητα σε HCS με περιορισμένους πόρους σε επίπεδο ενέργειας, μνήμης και επεξεργαστικής ισχύος [1–3].

Η διαρκώς αυξανόμενη πολυπλοκότητα και το μέγεθος των σύγχρονων CNNs συνεπάγονται σημαντική αύξηση των απαιτήσεων σε υπολογιστικούς πόρους, γεγονός που έρχεται σε αντίθεση με τις απαιτήσεις φορητών ή ενσωματωμένων συστημάτων, όπου η αποδοτική αξιοποίηση των διαθέσιμων υποδομών αποτελεί αδιαπραγμάτευτη προτεραιότητα. Ιδιαίτερα σε εφαρμογές που υλοποιούνται σε συσκευές υπολογισμού στο άκρο (Edge Computing Devices), όπως τα συστήματα του διαδικτύου των πραγμάτων (Internet of Things, IoT), οι μονάδες επεξεργασίας εικόνας σε πραγματικό χρόνο ή οι αυτόνομες φορητές οντότητες, οι περιορισμοί σε ενεργειακή αυτονομία, καθυστέρηση απόκρισης και κατανάλωση ενέργειας καθιστούν την αποδοτική σχεδίαση καθοριστικό παράγοντα επιτυχίας [4–6].

Η τάση προς CNNs με ολοένα αυξανόμενο αριθμό επιπέδων, με σκοπό την αύξηση της ακρίβειας συμπερασμού, οδηγεί σε ιδιαίτερα αυξημένες απαιτήσεις σε μνήμη και επεξεργαστική ισχύ. Όταν τέτοιες αρχιτεκτονικές καλούνται να υλοποιηθούν σε συσκευές με περιορισμένους πόρους, προκύπτουν προκλήσεις που άπτονται όχι μόνο της συμβατότητας του μοντέλου με την υποδομή, αλλά και της βιωσιμότητάς του. Η κατανάλωση ενέργειας, καθώς και το περιβαλλοντικό αποτύπωμα της εκπαίδευσης και εξαγωγής συμπερασμάτων σε μεγάλης κλίμακας μοντέλα, έχουν αναδειχθεί ως ζητήματα μείζονος σημασίας σε ερευνητικό και τεχνολογικό επίπεδο [7, 8].

Η αξιολόγηση του σχεδιασμού των αρχιτεκτονικών CNNs οφείλει να υπερβαίνει την παραδοσιακή εστίαση σε μετρικές ποιότητας, όπως η συνολική ακρίβεια, και να ενσωματώνει παραμέτρους υπολογιστικού κόστους, όπως η χρήση μνήμης, η χρονική καθυστέρηση εξαγωγής συμπερασμάτων και η ενεργειακή κατανάλωση. Η παράβλεψη αυτών των μετρικών, ιδίως σε εφαρμογές πραγματικού χρόνου, μπορεί να οδηγήσει σε επιλογή ακατάλληλων αρχιτεκτονικών, ασύμβατων με τις λειτουργικές προδιαγραφές του συστήματος. Επομένως, η σχεδίαση αρχιτεκτονικών

που χαρακτηρίζονται από αποδοτική χρήση πόρων, χωρίς συμβιβασμούς στην ακρίβεια ή την αξιοπιστία, αποτελεί ερευνητική πρόκληση πρώτης γραμμής [9, 10].

Επιπλέον, η επιλογή των κατάλληλων μετρικών αξιολόγησης δεν μπορεί να είναι ανεξάρτητη από το λειτουργικό πλαίσιο στο οποίο πρόκειται να αναπτυχθεί το εκάστοτε μοντέλο CNN. Οι περιορισμοί σε ενέργεια, μνήμη ή χρόνο απόκρισης επιβάλλουν την ενσωμάτωση μετρήσεων αποδοτικότητας ως βασικό κριτήριο επιλογής και όχι ως μεταγενέστερη προσθήκη. Η έλλειψη τυποποιημένων σημείων αναφοράς που να συνεκτιμούν ταυτόχρονα σχεδιαστικές, ποιοτικές και ποσοτικές παραμέτρους συνιστά σημαντικό ερευνητικό κενό, το οποίο περιορίζει τη δυνατότητα γενίκευσης και αναπαραγωγής των μετρήσεων απόδοσης [11–13].

Τέλος, είναι κρίσιμο να τονιστεί ότι η διαδικασία βελτιστοποίησης των αρχιτεκτονικών CNNs αποτελεί μια αμιγώς επαναληπτική διαδικασία. Η συνεχής ανατροφοδότηση από μετρήσεις σε επίπεδο υλοποίησης και η προσαρμογή των σχεδιαστικών αποφάσεων ως προς τις πραγματικές επιδόσεις του μοντέλου, αποτελούν αναπόσπαστο τμήμα ενός επιτυχούς κύκλου ανάπτυξης. Η επίτευξη ισορροπίας ανάμεσα σε μετρικές ακρίβειας και υπολογιστικού αποτυπώματος δεν αποτελεί αμετάβλητη ρύθμιση, αλλά προϊόν σύνθετου συμβιβασμού, ο οποίος καθορίζεται από τις ανάγκες της εκάστοτε εφαρμογής και τους περιορισμούς του εκάστοτε περιβάλλοντος [14–16].

1.2 Κίνητρο Διατριβής

Το κίνητρο της παρούσας διατριβής εδράζεται στη διαρκώς αυξανόμενη ανάγκη για αποδοτική υλοποίηση εφαρμογών AI, ειδικά βαθιάς μάθησης (Deep Learning, DL). Η αποδοτικότητα αυτή αναδύεται ως κρίσιμος παράγοντας σε ένα ευρύ φάσμα τεχνολογικών τομέων και διεπιστημονικών πεδίων. Με την διεύρυνση της χρήσης τεχνολογιών DL σε ποικίλα περιβάλλοντα, καθίσταται επιτακτική η ανάπτυξη αρχιτεκτονικών και μεθοδολογιών που επιτρέπουν την αποτελεσματική ενσωμάτωση και εφαρμογή τέτοιων εφαρμογών σε ετερογενή υπολογιστικά συστήματα [17].

Αυτό οδήγησε σε έντονο ερευνητικό και τεχνολογικό ενδιαφέρον για την αξιοποίηση των επιταχυντών υλικού σε ετερογενείς αρχιτεκτονικές, όπως οι μονάδες επεξεργασίας γραφικών (Graphics Processing Units, GPUs), οι προγραμματιζόμενες συστοιχίες επί του πεδίου (Field-Programmable Gate Arrays, FPGAs), καθώς και των υπολογιστικών συστημάτων υψηλής απόδοσης (High-Performance

Computing systems, HPC), όπως τα ενσωματωμένα συστήματα τύπου μονάδας (System—on—Module, SoM). Οι εν λόγω επιταχυντές προσφέρουν σημαντικά οφέλη όσον αφορά την υπολογιστική απόδοση και, σε πολλές περιπτώσεις, την ενεργειακή αποδοτικότητα, καθιστώντας τους κατάλληλους για την αποτελεσματική ενσωμάτωση τεχνικών DL σε απαιτητικά λειτουργικά περιβάλλοντα [18–20].

Οι GPUs, χάρη στην υψηλή παράλληλη επεξεργαστική τους ισχύ, έχουν καθιερωθεί ως βασική τεχνολογική υποδομή για την επιτάχυνση των CNNs και άλλων βαθιών αρχιτεκτονικών μάθησης, οι οποίες χαρακτηρίζονται από υψηλό υπολογιστικό κόστος. Από την άλλη πλευρά, τα FPGAs προσφέρουν δυνατότητες υλοποίησης χαμηλής κατανάλωσης και προσαρμοσμένων αρχιτεκτονικών, οι οποίες παραμετροποιούνται κατάλληλα, ώστε να εξυπηρετούν εφαρμογές σε πραγματικό χρόνο, χωρίς την ανάγκη επανασχεδιασμού σε επίπεδο υλικού. Πέραν αυτών, αξιοσημείωτο ρόλο διαδραματίζουν και τα SoMs, τα οποία ενσωματώνουν ετερογενείς υπολογιστικούς επιταχυντές σε ένα ενιαίο, αρθρωτό σύστημα. Τα συστήματα αυτά υποστηρίζουν την αποδοτική εκτέλεση εφαρμογών DL σε περιβάλλοντα με περιορισμένους ενεργειακούς πόρους, ενώ ταυτόχρονα προσφέρουν υψηλή ευελιξία ανάπτυξης και χαμηλό ενεργειακό αποτύπωμα [21–23].

Η παρούσα διατριβή αποσκοπεί στη συστηματική διερεύνηση του τρόπου με τον οποίο η αξιοποίηση εξειδικευμένων επιταχυντών υλικού συνεισφέρει στην αποδοτική εφαρμογή υπολογιστικά απαιτητικών DL αλγορίθμων, ιδίως στο πλαίσιο εφαρμογών με αυξημένο φόρτο επεξεργασίας. Εστιάζει στην ανάπτυξη μιας μεθοδολογίας που επιδιώκει τη βέλτιστη ισορροπία μεταξύ του σχεδιασμού της αρχιτεκτονικής, της ακρίβειας συμπερασμού, της χρονικής απόδοσης (τόσο κατά τη φάση της εκπαίδευσης όσο και κατά τη φάση του συμπερασμού) και της ενεργειακής κατανάλωσης. Η ανάλυση επικεντρώνεται στην αποτίμηση της αποδοτικότητας των αρχιτεκτονικών επιτάχυνσης, με στόχο την επιλογή εκείνων που κρίνονται κατάλληλες για ενσωμάτωση σε HCS όπου η διαθεσιμότητα πόρων είναι περιορισμένη.

Πιο συγκεκριμένα, μελετάται ο τρόπος με τον οποίο οι μετρικές αποδοτικότητας επηρεάζουν τη συνολική καταλληλότητα των αρχιτεκτονικών για την ενσωμάτωση τους σε HCS. Η αξιολόγηση δεν περιορίζεται στις μετρικές ποιότητας, αλλά επεκτείνεται και σε μετρικές που αφορούν το υπολογιστικό κόστος και το σχεδιασμό των CNNs. Η συνδυαστική ανάλυση των ποιοτικών μετρικών, του υπολογιστικού αποτυπώματος και του σχεδιασμού καθίσταται απαραίτητη για την τεκμηριωμένη συγκριτική αξιολόγηση και τελική επιλογή των εκτιμώμενων

υποδομών επιτάχυνσης, ιδίως όταν πρόκειται για υλοποίηση σε συστήματα με ενεργειακούς και υπολογιστικούς περιορισμούς. Η πειραματική αξιολόγηση πραγματοποιήθηκε σε δύο διαφορετικές προσεγγίσεις υλοποίησης HCS, εκ των οποίων η πρώτη βασίζεται σε αρχιτεκτονική τύπου FPGA–SoC, ενώ η δεύτερη σε σύστημα τύπου SoM.

1.3 Ερευνητικά Κενά και Προκλήσεις

Η ραγδαία εξέλιξη της DL, σε συνδυασμό με την ολοένα αυξανόμενη διείσδυσή της σε πληθώρα πεδίων εφαρμογής, συμπεριλαμβανομένων της υγειονομικής περίθαλψης, της αυτόνομης πλοήγησης, των ευφύων συστημάτων ελέγχου και των ρομποτικών συστημάτων αυτοματισμού, καθιστά επιτακτική την ανάγκη σχεδίασης αρχιτεκτονικά αποδοτικών και ενεργειακά βιώσιμων CNNs. Παρά την αξιοσημείωτη πρόοδο που έχει καταγραφεί στο πεδίο των αλγορίθμων CNNs ειδικά στον τομέα της CV, εντούτοις υφίστανται σημαντικά ερευνητικά κενά και τεχνολογικές προκλήσεις που περιορίζουν την αποτελεσματική τους ενσωμάτωση σε HCS με δεδομένους υπολογιστικούς και ενεργειακούς περιορισμούς.

Ένα από τα πλέον σημαντικά ερευνητικά κενά στο πεδίο της ΑΙ σχετίζεται με την έλλειψη ολοκληρωμένων και πολυκριτηριακών μεθοδολογιών για την αξιολόγηση της αποδοτικότητας των αρχιτεκτονικών CNNs. Η επικρατούσα μεθοδολογία αξιολόγησης των αλγορίθμων εστιάζει σχεδόν αποκλειστικά σε μετρικές απόδοσης συμπερασμού, όπως η ακρίβεια (accuracy), η ανάκληση (recall), η ακρίβεια θετικών προβλέψεων (precision), η βαθμολογία F1 και, κατά περίπτωση, η επιφάνεια κάτω από την καμπύλη ROC (AUC).

Στις υπάρχουσες μεθοδολογίες, η σχεδίαση και επιλογή των CNNs πραγματοποιείται με μοναδικό κριτήριο τις προαναφερθείσες μετρικές, χωρίς να συνυπολογίζονται εξίσου σημαντικές παράμετροι όπως το υπολογιστικό κόστος, η καθυστέρηση απόκρισης κατά τον συμπερασμό, η ενεργειακή κατανάλωση κατά τη διάρκεια εκπαίδευσης και συμπερασμού, καθώς και οι απαιτήσεις αποθηκευτικού χώρου. Ιδιαίτερα σε εφαρμογές πραγματικού χρόνου ή σε ενσωματωμένα συστήματα με περιορισμένες υπολογιστικές δυνατότητες, η παράβλεψη των ποσοτικών αυτών παραμέτρων μπορεί να οδηγήσει σε αρχιτεκτονικές μη συμβατές με τις ειδικές απαιτήσεις του εκάστοτε υπολογιστικού περιβάλλοντος.

Μία επιπλέον θεμελιώδης τεχνολογική πρόκληση αποτελεί η εγγενής ετερογένεια των επιταχυντών υλικού που αξιοποιούνται στην υλοποίηση και εκτέλεση αλγορίθμων DL. Η διαφοροποίηση των επιταχυντών, όπως οι GPUs, τα FPGAs και τα SoMs, δημιουργεί ανομοιογένεια όχι μόνο σε επίπεδο αρχιτεκτονικών δυνατοτήτων, αλλά και ως προς την αλληλεπίδραση με τα διαθέσιμα λογισμικά εργαλεία. Καθώς κάθε σύστημα συνοδεύεται από διαφορετικούς περιορισμούς ως προς τη διαχείριση μνήμης και υπολογιστικών πόρων, προκύπτει η αναγκαιότητα εξατομικευμένου σχεδιασμού και προσαρμοσμένης βελτιστοποίησης, γεγονός που αυξάνει την πολυπλοκότητα της διαδικασίας και παρατείνει το χρονικό διάστημα ανάπτυξης.

Η παρούσα διατριβή στοχεύει στην αντιμετώπιση των προαναφερθέντων ερευνητικών κενών και τεχνολογικών προκλήσεων, προτείνοντας μία ολοκληρωμένη μεθοδολογία για την αποδοτική αξιολόγηση και τον βέλτιστο σχεδιασμό εφαρμογών CNNs σε ετερογενή συστήματα. Η μεθοδολογία που αναπτύξαμε εφαρμόζεται και αξιολογείται σε δύο ενδεικτικές υλοποιήσεις HCS, όπου η πρώτη βασίζεται σε αρχιτεκτονική τύπου FPGA—SoC, ενώ η δεύτερη υλοποιείται σε σύστημα τύπου SoM, με στόχο την εξειδικευμένη διαχείριση των περιορισμών που επιβάλλει η κάθε επιταχυντική συσκευή.

1.4 Ερευνητικοί Στόχοι και Συνεισφορές

Η παρούσα διατριβή θέτει ως πρωταρχικό ερευνητικό στόχο τον καθορισμό και την ανάπτυξη μιας ολοκληρωμένης και τεκμηριωμένης μεθοδολογίας για τον αποδοτικό σχεδιασμό και την υλοποίηση αλγορίθμων CNNs σε HCS. Η προτεινόμενη μεθοδολογία επιδιώκει τη γεφύρωση του χάσματος μεταξύ της θεωρητικής σχεδίασης των αρχιτεκτονικών CNNs και της πρακτικής τους εφαρμογής σε περιβάλλοντα με περιορισμένους πόρους, ενσωματώνοντας τρεις κρίσιμες διαστάσεις αξιολόγησης, τις μετρικές σχεδιαστικής αποδοτικότητας, τις μετρικές ποιοτικής αξιολόγησης και τις μετρικές υπολογιστικού αποτυπώματος.

Η εν λόγω συνδυαστική προσέγγιση καθιστά δυνατή τη συνολική βελτιστοποίηση των επιμέρους σταδίων ανάπτυξης, από την αρχική επιλογή ή τον σχεδιασμό της αρχιτεκτονικής έως την τελική υλοποίηση και αξιολόγηση επί του συστήματος—στόχου. Σκοπός είναι η προτεινόμενη μεθοδολογία να ανταποκρίνεται στις αυξημένες απαιτήσεις των σύγχρονων εφαρμογών AI/CV ως προς την

υπολογιστική απόδοση, την ενεργειακή αποδοτικότητα και την αρχιτεκτονική ευελιξία, με ιδιαίτερη έμφαση στην προσαρμοστικότητα σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους.

Παράλληλα, η μεθοδολογία αυτή στοχεύει στην εις βάθος διερεύνηση και συγκριτική αξιολόγηση δύο διακριτών και τεχνολογικά σημαντικών προσεγγίσεων για την υλοποίηση αλγορίθμων CNNs σε HCS. Η πρώτη προσέγγιση εστιάζει σε αρχιτεκτονικές τύπου FPGA–SoC, οι οποίες προσφέρουν υψηλό βαθμό παραμετροποίησης υλικού και δυνατότητες για βελτιστοποιήσεις χαμηλού επιπέδου. Η δεύτερη προσέγγιση αφορά συστήματα τύπου SoM, τα οποία παρέχουν μια πιο ολοκληρωμένη λύση με ισχυρούς επιταχυντές (κυρίως GPUs) και ώριμα οικοσυστήματα λογισμικού. Σκοπός της εν λόγω διάκρισης είναι η αποδοτική διαχείριση, η προσαρμοσμένη βελτιστοποίηση και η εξειδικευμένη αντιμετώπιση των εγγενών χαρακτηριστικών, των αρχιτεκτονικών ιδιοτεροτήτων και των υπολογιστικών περιορισμών που παρουσιάζει το εκάστοτε περιβάλλον, λαμβάνοντας υπόψη τόσο τις επιδόσεις σε πραγματικό χρόνο όσο και την ενεργειακή αποδοτικότητα.

Συνοψίζοντας, οι βασικοί ερευνητικοί στόχοι της παρούσας διατριβής είναι οι ακόλουθοι:

- Ολοκληρωμένη μεθοδολογία βάσει μετρικών σχεδιασμού: Προσδιορισμός, ανάπτυξη και τεκμηρίωση μιας ολοκληρωμένης μεθοδολογίας για τον σχεδιασμό αρχιτεκτονικών CNNs για εφαρμογές CV. Η μεθοδολογία αυτή εστιάζει σε τέσσερις πυλώνες σχεδιασμού και καθοδηγείται από τις απαιτήσεις της εκάστοτε εφαρμογής και τις διαθέσιμες δυνατότητες του συστήματος—στόχου HCS, με κύριο γνώμονα τη βελτιστοποίηση της υπολογιστικής και ενεργειακής αποδοτικότητας, αξιοποιώντας ένα σύνολο καθορισμένων μετρικών σχεδιασμού.
- Συστηματική μεθοδολογία βάσει μετρικών ποιότητας και υπολογιστικού αποτυπώματος: Καθορισμός και ανάπτυξη μιας συστηματικής μεθοδολογικής προσέγγισης για την επιλογή και εφαρμογή του κατάλληλου συνόλου ποιοτικών και ποσοτικών μετρικών αξιολόγησης. Οι μετρικές αυτές ευθυγραμμίζονται με τις λειτουργικές απαιτήσεις των εφαρμογών και τα χαρακτηριστικά του συστήματος—στόχου HCS.
- Επιτάχυνση εφαρμογών σε HCS: Σχεδίαση, υλοποίηση και βελτιστοποίηση αντιπροσωπευτικών εφαρμογών CNNs σε δύο διακριτές κατηγορίες HCS:

αρχιτεκτονικές τύπου FPGA—SoC και συστήματα τύπου SoM. Στόχος είναι η εξειδικευμένη και αποδοτική διαχείριση των ιδιοτήτων, των δυνατοτήτων και των περιορισμών του εκάστοτε υπολογιστικού συστήματος, στο πλήρες σύνολο των μετρικών που ορίζονται από την προτεινόμενη μεθοδολογία.

1.5 Διάρθρωση Διατριβής

Η διάρθρωση της παρούσας διατριβής έχει ως εξής:

Το Κεφάλαιο 2 εστιάζει στην ανασκόπηση του θεωρητικού υποβάθρου της διατριβής. Ειδικότερα, εξετάζονται τα HCS, με έμφαση στις αρχιτεκτονικές επιλογές, τις επιδόσεις και την προσαρμοστικότητά τους σε εφαρμογές υψηλών υπολογιστικών απαιτήσεων. Στη συνέχεια, αναλύεται η ιεραρχική σχέση μεταξύ AI, ML και DL, αναδεικνύοντας τις κατηγορίες αλγορίθμων και τις εφαρμογές τους. Ιδιαίτερη έμφαση δίδεται στα CNNs, μέσω της ανάλυσης της αρχιτεκτονικής τους δομής και των επιμέρους λειτουργικών τους στοιχείων. Τέλος, παρουσιάζονται οι τεχνικές επιτάχυνσης υλικού και οι συναφείς βελτιστοποιήσεις που επιτρέπουν την αποδοτική ενσωμάτωση των CNNs σε HCS, οι οποίες ανταποκρίνονται αποτελεσματικά στους περιορισμούς ενέργειας, μνήμης και χρόνου απόκρισης που χαρακτηρίζουν τις σύγχρονες υπολογιστικές εφαρμογές.

Το Κεφάλαιο 3 παρουσιάζει τη μεθοδολογία που αναπτύξαμε σε αυτήν τη διατριβή για την αποδοτική σχεδίαση και υλοποίηση των CNNs σε HCS. Η μεθοδολογία αυτή οργανώνεται σε τρεις βασικούς άξονες: τις μετρικές σχεδίασης, τις μετρικές ποιότητας και τις μετρικές υπολογιστικού αποτυπώματος. Μέσω αυτών, επιδιώκεται η διαμόρφωση ενός ολοκληρωμένου και καλά δομημένου πλαισίου αναφοράς με στόχο την επίτευξη ισορροπίας ανάμεσα στην απόδοση εξαγωγής συμπερασμάτων, την υπολογιστική αποδοτικότητα και την κατανάλωση ενέργειας, προκειμένου τα CNNs να είναι κατάλληλα για ένα ευρύ φάσμα απαιτητικών εφαρμογών και περιορισμών.

Το Κεφάλαιο 4 εστιάζει στον σχεδιασμό και την υλοποίηση τεχνικών επιτάχυνσης CNN σε FPGA—SoC, χρησιμοποιώντας τη μεθοδολογία που αναπτύξαμε στο Κεφάλαιο 3. Αρχικά, περιγράφεται η αρχιτεκτονική που υιοθετείται για την αποδοτική υλοποίηση του δικτύου, ενώ στη συνέχεια αναλύεται η διαδικασία βελτιστοποίησης μέσω προηγμένων μεθόδων επιτάχυνσης. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση και τη συζήτηση των αποτελεσμάτων για

την αξιολόγηση της προτεινόμενης υλοποίησης μέσω μετρικών απόδοσης και συγκρίσεων με εναλλακτικές υλοποιήσεις σε άλλες συσκευές.

Το Κεφάλαιο 5 εστιάζει στον σχεδιασμό και την υλοποίηση ενός αποδοτικού CNN σε SoM, εφαρμόζοντας τη μεθοδολογία που αναπτύξαμε στο Κεφάλαιο 3. Αρχικά, παρουσιάζεται η αρχιτεκτονική του μοντέλου, καθώς και οι τεχνικές βελτιστοποίησης που εφαρμόστηκαν με στόχο τη χαμηλή πολυπλοκότητα. Στη συνέχεια, περιγράφεται η μεθοδολογία αξιολόγησης του μοντέλου μέσω πειραμάτων και η συγκριτική αξιολόγηση έναντι καθιερωμένων αρχιτεκτονικών CNN. Το κεφάλαιο ολοκληρώνεται με τη συζήτηση των αποτελεσμάτων ως προς την αποδοτικότητα της υλοποίησης σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους.

Το Κεφάλαιο 6 συνοψίζει τα βασικά ευρήματα της διατριβής σχετικά με τον σχεδιασμό, τη βελτιστοποίηση και την υλοποίηση αποδοτικών CNNs σε HCS. Αρχικά, παρουσιάζονται τα κύρια συμπεράσματα που προκύπτουν από τα πειραματικά αποτελέσματα ως προς την υπολογιστική αποδοτικότητα, την ενεργειακή κατανάλωση και την συνολική ακρίβεια των προτεινόμενων υλοποιήσεων σε συσκευές τύπου FPGA—SoC και SoM. Στη συνέχεια, αναλύονται τα ευρήματα όσον αφορά την πρόοδο του τομέα της ΑΙ και την υλοποίηση αλγορίθμων DL σε ενσωματωμένα περιβάλλοντα. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση προτάσεων και κατευθύνσεων για μελλοντική έρευνα, εστιάζοντας σε ανοιχτά ζητήματα και προκλήσεις που αναδείχθηκαν κατά τη διάρκεια της διατριβής.

Τέλος, στο Παράρτημα Α' παρουσιάζεται το σύνολο της δημοσιευμένης έρευνας που προέκυψε κατά τη διάρκεια της παρούσας διατριβής σε επιστημονικά περιοδικά και πρακτικά διεθνών συνεδρίων με κρίση.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Σε αυτό το κεφάλαιο παρουσιάζεται μια σύντομη εισαγωγή στις θεμελιώδεις έννοιες και βασικές τεχνολογίες που αποτελούν τον πυρήνα της παρούσας διατριβής. Αρχικά, γίνεται μια εισαγωγή στα HCS, τα οποία αποτελούν ραγδαία αναπτυσσόμενα συστήματα στην υπολογιστική τεχνολογία, ικανά να ανταποκριθούν στις αυξανόμενες απαιτήσεις απόδοσης, ταχύτητας και ενεργειακής αποδοτικότητας των σύγχρονων εφαρμογών. Στη συνέχεια, παρουσιάζονται οι βασικές αρχές και διαφορές που διέπουν τη σχέση μεταξύ της AI, ML και DL. Ιδιαίτερη έμφαση δίδεται στα CNNs, τόσο ως προς τα επιμέρους δομικά τους στοιχεία όσο και ως προς τη διαδικασία ενσωμάτωσής τους σε HCS. Τέλος, εξετάζονται οι τεχνολογικές λύσεις που καθιστούν εφικτή την αποδοτική τους υλοποίηση, με έμφαση στην επιτάχυνση υλικού και τις σχετικές αρχιτεκτονικές βελτιστοποιήσεις.

2.1 Ετερογενή Υπολογιστικά Συστήματα

Τα ετερογενή υπολογιστικά συστήματα (Heterogeneous Computing Systems, HCS) αντιπροσωπεύουν έναν διαρκώς εξελισσόμενο τομέα στην πληροφορική, προσφέροντας λύσεις που ανταποκρίνονται αποτελεσματικά στις συνεχώς αυξανόμενες ανάγκες των σύγχρονων εφαρμογών για υψηλές υπολογιστικές απαιτήσεις. Σε αντίθεση με τα παραδοσιακά ομοιογενή υπολογιστικά συστήματα, τα οποία βασίζονται σε μία ενιαία αρχιτεκτονική επεξεργασίας, τα HCS ενσωματώνουν και συνδυάζουν διαφορετικούς τύπους επεξεργαστικών μονάδων, όπως CPUs, GPUs και FPGAs [24, 25]. Η κάθε υπολογιστική μονάδα παρουσιάζει ειδικά χαρακτηριστικά: οι CPUs προσφέρουν ευελιξία και επεξεργαστική ισχύ γενικού

σκοπού, κατάλληλη για σειριακού τύπου διεργασίες, οι GPUs είναι εξειδικευμένες για παράλληλη επεξεργασία μεγάλων όγκων δεδομένων, ενώ τα FPGAs παρέχουν δυνατότητες επιτάχυνσης, ιδιαίτερα σε εφαρμογές σε πραγματικές συνθήκες λειτουργίας [26–28]. Το θεμελιώδες πλεονέκτημα της ετερογενούς προσέγγισης έγκειται στην επίτευξη αισθητά υψηλότερης συνολικής υπολογιστικής απόδοσης και βελτιωμένης ενεργειακής αποδοτικότητας, σε σύγκριση με συστήματα μίας αρχιτεκτονικής.

Η προσαρμοστικότητα των HCS στις απαιτήσεις διαφορετικών υπολογιστικών φορτίων τα καθιστά ιδανική επιλογή για περιβάλλοντα που απαιτούν εντατική επεξεργασία δεδομένων και ταυτόχρονη εκτέλεση πολλαπλών διεργασιών. Σε εφαρμογές όπως η CV, που περιλαμβάνει εργασίες όπως ταξινόμηση αντικειμένων, αναγνώριση προτύπων και αυτόματη κατανόηση περιεχομένου εικόνας ή βίντεο, ή η επεξεργασία φυσικής γλώσσας (Natural Language Processing, NLP), που περιλαμβάνει ανάλυση συναισθήματος, αυτόματη μετάφραση και εξαγωγή γνώσης από αδόμητο κείμενο, τα HCS επιτρέπουν την αποδοτική εκτέλεση σύνθετων υπολογιστικών φορτίων με ταχύτητα και ακρίβεια που υπερβαίνουν τις δυνατότητες μιας ενιαίας αρχιτεκτονικής επεξεργασίας [29–31].

Η ετερογένεια στα σύγχρονα υπολογιστικά συστήματα εκδηλώνεται σε τρία επίπεδα οργάνωσης και σχεδίασης. Σε πρώτο επίπεδο, αναφέρεται στη συνύπαρξη επεξεργαστικών πυρήνων που υλοποιούν την ίδια αρχιτεκτονική συνόλου εντολών (Instruction Set Architecture, ISA), αλλά παρουσιάζουν διαφοροποιήσεις σε επίπεδο μικροαρχιτεκτονικής, όπως συμβαίνει στην προσέγγιση big.LITTLE [32], όπου συνδυάζονται υψηλής απόδοσης και ενεργειακά αποδοτικοί πυρήνες, ώστε να επιτυγχάνεται δυναμική προσαρμογή της υπολογιστικής ισχύος και της κατανάλωσης ενέργειας ανάλογα με τις απαιτήσεις του φορτίου εργασίας. Σε δεύτερο επίπεδο, περιλαμβάνεται η συνεργασία υπολογιστικών μονάδων με διαφορετικές αρχιτεκτονικές δυνατότητες, όπως οι CPU και GPU. Τέλος, σε τρίτο επίπεδο, η ετερογένεια σχετίζεται με τη συνύπαρξη υπολογιστικών μονάδων που υλοποιούν διαφορετικά μοντέλα εκτέλεσης, όπως οι CPU και τα FPGA, με τα τελευταία να βασίζονται σε επαναπρογραμματιζόμενη λογική και παραλληλία σε επίπεδο υλικού [33, 34].

Καθώς η τεχνολογία των HCS συνεχίζει να εξελίσσεται, αναμένεται να διαδραματίσει καθοριστικό ρόλο σε ποικίλους τομείς, συμβάλλοντας στην πρόοδο της υπολογιστικής επιστήμης και της μηχανικής. Η σταδιακή επέκταση και η συνεχιζόμενη ανάπτυξη τους δημιουργούν νέες δυνατότητες και άξονες προσέγγισης

για την αντιμετώπιση απαιτητικών προβλημάτων, τα οποία δεν μπορούν να αντιμετωπιστούν αποτελεσματικά μέσω συμβατικών αρχιτεκτονικών [35, 36]. Η εγγενής αρχιτεκτονική ευελιξία των HCS καθιστά δυνατή την ενσωμάτωση προηγμένων τεχνολογιών, όπως αλγορίθμων DL, των οποίων η εκπαίδευση απαιτεί τεράστια υπολογιστική ισχύ. Αυτή η ευελιξία επιτρέπει την αποδοτική υλοποίηση σύνθετων αλγορίθμων, αξιοποιώντας τις εξειδικευμένες δυνατότητες κάθε επεξεργαστικής μονάδας για την επιτάχυνση συγκεκριμένων φάσεων της διαδικασίας εκπαίδευσης και εξαγωγής συμπερασμάτων των μοντέλων DL [37, 38].

Τα HCS προσφέρουν σημαντικές δυνατότητες καινοτομίας στον ερευνητικό τομέα, διευκολύνοντας την ανάπτυξη νέων αλγορίθμων και αρχιτεκτονικών που αξιοποιούν αποδοτικά τις διάφορες επεξεργαστικές μονάδες. Μέσω της χρήσης αυτών των τεχνολογιών, οι ερευνητές αναπτύσσουν αλγορίθμους με μεγαλύτερη προσαρμοστικότητα και αποδοτικότητα, ικανούς να ανταποκρίνονται σε απαιτήσεις πραγματικών συνθηκών. Καθώς η τεχνολογία των HCS γίνεται ολοένα και πιο προσιτή και προσβάσιμη, αναμένεται η ευρύτερη εφαρμογή της σε διάφορους επιστημονικούς τομείς. Η εξέλιξη αυτή δημιουργεί νέες προοπτικές για την υλοποίηση λύσεων υψηλής απόδοσης σε πραγματικά περιβάλλοντα [39–41].

Σήμερα, μία από τις κυριότερες τάσεις στα σύγχρονα HCS, καθίσταται ολοένα και πιο διαδεδομένη η αρχιτεκτονική πρακτική της ενσωμάτωσης διαφορετικών τύπων επεξεργαστών και εξειδικευμένων επιταχυντών εντός του ιδίου ολοκληρωμένου κυκλώματος (System-on-Chip, SoC). Η ενοποίηση αυτή ενισχύει την υπολογιστική απόδοση του συστήματος, καθώς επιτρέπει την ταχύτερη και χαμηλού λανθάνοντος χρόνου ενδοσυστημική επικοινωνία μεταξύ των υποσυστημάτων, μειώνοντας την εξάρτηση από εξωτερικούς διαύλους μεταφοράς και περιορίζοντας τα φαινόμενα συμφόρησης. Παράλληλα, επιτρέπει τον καλύτερο συγχρονισμό των υπολογιστικών ροών, την αποδοτικότερη κατανομή του υπολογιστικού φόρτιου και τη βελτιστοποιημένη αξιοποίηση των διαθέσιμων πόρων [42–45].

Επιλέον, ιδιαίτερη σημασία αποκτούν και τα SoMs, τα οποία βασίζονται σε ετερογενή SoCs και παρέχουν ολοκληρωμένες λύσεις υπολογιστικής υποδομής, συμπεριλαμβάνοντας επιπλέον στοιχεία όπως μνήμη, διασυνδέσεις και συστήματα διαχείρισης ισχύος σε ένα ενιαίο προς ενσωμάτωση υποσύστημα. Τα SoMs διευκολύνουν την ταχεία ανάπτυξη και υλοποίηση εφαρμογών σε τομείς αιχμής, προσφέροντας ισχυρούς και παραμετροποιήσιμους επιταχυντές με χαμηλό ενεργειακό αποτύπωμα [46, 47].

2.2 Τεχνητή Νοημοσύνη

Η τεχνητή νοημοσύνη (Artificial Intelligence, AI) συνιστά έναν ευρύ και πολυεπίπεδο κλάδο της επιστήμης των υπολογιστών, ο οποίος επικεντρώνεται στον θεωρητικό και πρακτικό σχεδιασμό εφαρμογών υψηλής πολυπλοκότητας, ικανών να επιδεικνύουν γνωστικές λειτουργίες αντίστοιχες ή συγκρίσιμες με εκείνες της ανθρώπινης νοημοσύνης. Κεντρικός στόχος της AI παραμένει η ανάπτυξη αλγοριθμικών μηχανισμών που επιτρέπουν στις μηχανές να λαμβάνουν αποφάσεις υπό συνθήκες αβεβαιότητας, να επιλύουν σύνθετα προβλήματα, να προσαρμόζονται σε νέα δεδομένα και να μαθαίνουν μέσω εμπειρικής αλληλεπίδρασης με το περιβάλλον τους [48].

Παράλληλα, η AI έχει διεισδύσει σε διεπιστημονικές και κοινωνικοοικονομικές εφαρμογές, προσδίδοντας αυξημένη αυτονομία, προσαρμοστικότητα και υπολογιστική ευφυΐα σε ποικίλους τομείς. Η ενσωμάτωση ευφύων τεχνικών σε πεδία όπως η υγειονομική περίθαλψη, η βιομηχανική παραγωγή, τα χρηματοοικονομικά, το ηλεκτρονικό εμπόριο και η ασφάλεια αποφέρει σημαντικά οφέλη, ενισχύοντας τόσο τη λειτουργική αποτελεσματικότητα, όσο και την ποιότητα παροχής υπηρεσιών, με άμεσο αντίκτυπο στην αξιοπιστία, την αποδοτικότητα και τη βιωσιμότητα των αντίστοιχων συστημάτων [49]. Ο Πίνακας 2.1 συνοψίζει τις κυριότερες κατηγορίες εφαρμογών της AI στο σύγχρονο τεχνολογικό περιβάλλον.

Η ταχεία πρόοδος της AI και ειδικότερα της DL έχει φέρει επανάσταση στις τεχνολογίες επεξεργασίας δεδομένων, οδηγώντας σε νέες απαιτήσεις για αρχιτεκτονικές υποδομές που μπορούν να ανταποκριθούν σε αυξημένα φορτία και χρονική αμεσότητα. Η ικανότητα των αλγορίθμων DL να αναλύουν μεγάλα σύνολα δεδομένων και να εξάγουν πληροφορίες με υψηλή ακρίβεια έχει καταστήσει απαραίτητη την αξιοποίηση προηγμένων συστημάτων υπολογιστικής επιτάχυνσης με βελτιστοποιημένη απόδοση. Τα HCS, συνδυάζοντας διαφορετικές μονάδες επεξεργασίας, όπως CPUs, GPUs και FPGAs, έχουν αναδειχθεί ως μια ιδανική λύση για την υποστήριξη τέτοιων απαιτήσεων, προσφέροντας την αναγκαία υπολογιστική ισχύ για τη διαχείριση και την ανάλυση δεδομένων μεγάλης κλίμακας, την εκπαίδευση και την υλοποίηση πολύπλοκων συνελκτικών μοντέλων [50, 51].

Η λειτουργικότητα των HCS προέρχεται από την αρχιτεκτονική τους, η οποία τους επιτρέπει να αναθέτουν συγκεκριμένα φορτία εργασίας σε κατάλληλες μονάδες επεξεργασίας, αξιοποιώντας τη δυναμική του κάθε υποσυστήματος. Για παράδειγμα, οι GPUs είναι εξειδικευμένες για παράλληλη επεξεργασία, κάτι που τις καθιστά

Πίνακας 2.1: Κατηγοριοποίηση των πεδίων εφαρμογής της ΑΙ.

Τομέας Εφαρμογής	Ενδεικτικές Χρήσεις / Εφαρμογές
Υγεία	Ιατρικά ρομπότ και διαγνωστική απεικόνιση. Διαλογή ασθενών και ταξινόμηση νόσων με υποστήριξη ΑΙ. Ευφυή συστήματα διαχείρισης υγείας.
Δημόσιος τομέας / Διακυβέρνηση	Ψηφιακές δημόσιες υπηρεσίες και ενιαίες πύλες εξυπηρέτησης πολιτών. Προγνωστική αστυνόμευση και πρόβλεψη εγκληματικότητας. Ανίχνευση και καταπολέμηση απάτης.
Εκπαίδευση	Ευφυή συστήματα διδασκαλίας και προσαρμοστικές πλατφόρμες μάθησης. Αυτόματη αξιολόγηση κειμένων και ανίχνευση λογοκλοπής. Αναλυτική της μάθησης και εξατομικευμένη ανατροφοδότηση.
Κλάδος ασφάλειας	Ανάλυση ανθρώπινης συμπεριφοράς, στάσης σώματος και βάδισης. Αναγνώριση προσώπου και ευφυής επιτήρηση. Συστήματα πρόβλεψης απειλών.
Χρηματοπιστωτικός κλάδος	Ανάλυση μεγάλων δεδομένων για χρηματιστηριακές τάσεις. Μοντελοποίηση κινδύνου και πρόβλεψη επενδύσεων.
Τηλεπικοινωνίες	Πρόβλεψη κίνησης δικτύου και ανίχνευση ανωμαλιών. Ευφυής δρομολόγηση σε δίκτυα 5G/6G. Ανάλυση συμπεριφοράς πελατών και βελτιστοποίηση υπηρεσιών.
Μεταποίηση	Πρόβλεψη βλαβών εξοπλισμού και προληπτική συντήρηση. Εντοπισμός ελαττωμάτων και μηχανική όραση. Συστήματα συνεργασίας ανθρώπου-μηχανής.
Ηλεκτρονικό εμπόριο	Αυτοματοποίηση εξυπηρέτησης πελατών. Συστήματα συστάσεων και πρόβλεψη ζήτησης. Ανάλυση πωλήσεων και βελτιστοποίηση τιμολόγησης.
Λιανική	Καταστήματα και αποθήκες χωρίς προσωπικό. Ευφυής διαχείριση εφοδιαστικής αλυσίδας. Στατιστική και ανάλυση ροής πελατών.
Εφοδιαστική Μεταφορές	Παραδόσεις με UAV και συστήματα αυτόματου πιλότου. Αυτόνομη διαλογή και διαχείριση φορτίων. Ευφυείς αλγόριθμοι χρονοπρογραμματισμού logistics.
Οικιακός αυτοματισμός	Ευφυείς φωνητικοί βοηθοί και προγραμματιζόμενος έλεγχος. Παρακολούθηση περιβάλλοντος και συστήματα ειδοποίησης. Πλατφόρμες απομακρυσμένου ελέγχου και λειτουργίας.
Λοιποί κλάδοι	Γεωργία: παρακολούθηση καλλιεργειών σε πραγματικό χρόνο. Προστασία περιβάλλοντος: βελτιστοποίηση ενέργειας, παρακολούθηση της κλιματικής αλλαγής, διαχείριση απορριμμάτων. Αστική διαχείριση: ευφυής δρομολόγηση κυκλοφορίας.

ιδανικές για εκπαίδευση CNNs και επεξεργασία εικόνων υψηλής ανάλυσης. Από την άλλη πλευρά, τα FPGAs μπορούν να προσαρμοστούν ώστε να εκτελούν συγκεκριμένες υπολογιστικές λειτουργίες, επιτρέποντας εξαιρετικά χαμηλούς χρόνους καθυστέρησης και χαμηλό ενεργειακό αποτύπωμα. Αυτός ο συνδυασμός διαφορετικών αρχιτεκτονικών επιτρέπει στα HCS να αναλάβουν απαιτητικά φορτία εργασίας, διατηρώντας υψηλά επίπεδα απόδοσης και αποδοτικότητας, καθώς και σημαντική εξοικονόμηση πόρων [52, 53].

Η ενοποίηση της AI με HCS δεν ενισχύει απλώς την απόδοση, αλλά ανοίγει νέους δρόμους για την ανάπτυξη καινοτόμων εφαρμογών σε ένα ευρύ φάσμα τομέων. Ειδικότερα, οι δυνατότητες αυτές είναι ιδιαίτερα αξιοποιήσιμες στην ιατρική διάγνωση, όπου η ταχύτητα και η ακρίβεια στην επεξεργασία των δεδομένων είναι ζωτικής σημασίας. Για παράδειγμα, τα HCS μπορούν να αναλύσουν εικόνες ακτινογραφίας ή μαγνητικές τομογραφίες σε πραγματικό χρόνο, διευκολύνοντας την έγκαιρη και ακριβή διάγνωση ασθενειών. Επίσης, στην αυτόνομη οδήγηση και στις έξυπνες πόλεις, όπου η επεξεργασία δεδομένων από αισθητήρες και κάμερες πρέπει να γίνεται άμεσα και με ασφάλεια, τα HCS προσφέρουν τις δυνατότητες που απαιτούνται για την άμεση ανάλυση δεδομένων, υποστηρίζοντας τη λήψη αποφάσεων σε πραγματικές συνθήκες [54, 55].

Ο κλάδος αυτός της AI, γνωστός ως μηχανική μάθηση (Machine Learning, ML), παρέχει τα θεμέλια για την ανάπτυξη και προσαρμογή των υπολογιστικών δυνατοτήτων σύμφωνα με τις ανάγκες της εκάστοτε εφαρμογής, διευκολύνοντας την ευέλικτη ανταπόκριση των συστημάτων σε δυναμικά και μεταβαλλόμενα περιβάλλοντα. Η προσαρμοστικότητα των HCS επιτρέπει την εξειδίκευση των διαθέσιμων πόρων σε πραγματικό χρόνο, προσφέροντας τη δυνατότητα για αποδοτική και στοχευμένη υλοποίηση αλγορίθμων AI, ικανών να ανταποκριθούν στις απαιτήσεις των ετερογενών λειτουργικών συνθηκών [56].

Αυτή η ευελιξία και δυνατότητα προσαρμογής καθιστά τα HCS σημαντικό παράγοντα για την ανάπτυξη των τεχνολογιών αιχμής, προωθώντας τη δημιουργία ευφύων, αυτοματοποιημένων και αποδοτικών συστημάτων που μπορούν να ανταποκριθούν στις αυξανόμενες ανάγκες της σύγχρονης κοινωνίας. Με τη συνεχή εξέλιξη και βελτίωση των υπολογιστικών υποδομών, αναμένεται τα HCS να παίξουν έναν κεντρικό ρόλο στην υποστήριξη των εξελισσόμενων εφαρμογών και απαιτήσεων, συμβάλλοντας καθοριστικά στην προαγωγή της υπολογιστικής επιστήμης και της τεχνολογικής καινοτομίας.

2.3 Μηχανική Μάθηση

Η μηχανική μάθηση (Machine Learning, ML) αποτελεί έναν θεμελιώδη τομέα της ΑΙ, ο οποίος επικεντρώνεται στην ανάπτυξη και την εφαρμογή αλγορίθμων που επιτρέπουν στα υπολογιστικά συστήματα να μαθαίνουν από δεδομένα και να βελτιώνουν την απόδοσή τους σε συγκεκριμένα καθήκοντα, χωρίς να απαιτείται άμεσος προγραμματισμός για κάθε ενέργεια. Η κεντρική ιδέα της ML είναι η χρήση δεδομένων για την εκπαίδευση αλγορίθμων που μπορούν να εξάγουν πληροφορίες και να εξάγουν συμπεράσματα με βάση πρότυπα και σχέσεις που ανακαλύπτουν στα δεδομένα [57, 58]. Η αποδοτική υλοποίηση των αλγορίθμων αυτών απαιτεί εξειδικευμένες υπολογιστικές αρχιτεκτονικές, όπως αυτές που προσφέρουν τα HCS.

Η ML διακρίνεται σε τρεις κύριες κατηγορίες: επιβλεπόμενη μάθηση (Supervised Learning), μη επιβλεπόμενη μάθηση (Unsupervised Learning) και ενισχυτική μάθηση (Reinforcement Learning). Στην επιβλεπόμενη μάθηση, οι αλγόριθμοι μαθαίνουν από δεδομένα που περιέχουν τόσο εισόδους όσο και τις αντίστοιχες επιθυμητές εξόδους. Τα σύνολα δεδομένων είναι ετικετοποιημένα, πράγμα που σημαίνει ότι κάθε δείγμα συνοδεύεται από την κατηγορία ή την τιμή στόχο που αντιστοιχεί στην είσοδο. Αυτή η διαδικασία επιτρέπει στους αλγόριθμους να αναγνωρίσουν και να μάθουν τη σχέση μεταξύ των εισόδων και των εξόδων, καθιστώντας τους ικανούς να κάνουν ακριβείς προβλέψεις σε νέα, άγνωστα δεδομένα. Η επιβλεπόμενη μάθηση είναι ιδανική για προβλήματα όπως η κατηγοριοποίηση εικόνων, η πρόβλεψη τιμών και η αναγνώριση προτύπων σε δομημένα δεδομένα, όπως ιατρικά αρχεία [59–62].

Αντίθετα, στη μη επιβλεπόμενη μάθηση, οι αλγόριθμοι εκπαιδεύονται χωρίς προκαθορισμένες ετικέτες ή κατηγορίες, πράγμα που τους επιτρέπει να αναζητούν και να εντοπίζουν κρυφά πρότυπα ή δομές μέσα στα δεδομένα. Οι αλγόριθμοι αυτοί αναλαμβάνουν την ομαδοποίηση ή τον διαχωρισμό των δεδομένων σε κατηγορίες, με βάση τα χαρακτηριστικά τους. Έτσι, η μη επιβλεπόμενη μάθηση είναι ιδιαίτερα χρήσιμη για προβλήματα όπου οι σχέσεις και τα πρότυπα δεν είναι γνωστά εκ των προτέρων, όπως στην ανάλυση πελατειακής συμπεριφοράς, την αναγνώριση ομάδων ενδιαφέροντος, ή την αναζήτηση ασυνήθιστων στοιχείων, όπως σε δεδομένα απάτης [63, 64].

Στην ενισχυτική μάθηση, οι αλγόριθμοι μαθαίνουν να λαμβάνουν αποφάσεις σε περιβάλλοντα που αλλάζουν δυναμικά, χρησιμοποιώντας μια διαδικασία δοκιμής και σφάλματος. Αντί να διαθέτουν καθορισμένες ετικέτες για την είσοδο, οι αλγόριθμοι της ενισχυτικής μάθησης λαμβάνουν ανατροφοδότηση σε μορφή

ενισχύσεων ή ποινών για τις ενέργειες που πραγματοποιούν. Η ανατροφοδότηση αυτή τους επιτρέπει να βελτιώνουν τη στρατηγική τους σταδιακά, ώστε να μεγιστοποιήσουν το συνολικό όφελος ή την απόδοση. Η ενισχυτική μάθηση είναι κατάλληλη για προβλήματα όπου οι αποφάσεις έχουν αλυσιδωτά αποτελέσματα, όπως η αυτόνομη οδήγηση και ο έλεγχος των ρομποτικών συστημάτων [65–67].

Οι κύριες κατηγορίες της ML, δηλαδή η επιβλεπόμενη, η μη επιβλεπόμενη και η ενισχυτική μάθηση, αποτελούν τα θεμέλια του πεδίου, αλλά οι εξελίξεις στον τομέα έχουν οδηγήσει στην ανάδειξη νέων υποκατηγοριών με αυξανόμενη σημασία. Μία από τις πιο αξιοσημείωτες από αυτές είναι η ημι–επιβλεπόμενη μάθηση (Semi–Supervised Learning). Η συγκεκριμένη μέθοδος αποτελεί έναν υβριδικό συνδυασμό των χαρακτηριστικών της επιβλεπόμενης και της μη επιβλεπόμενης μάθησης. Αντί να απαιτεί μεγάλο σύνολο επισημασμένων δεδομένων, που συχνά είναι δύσκολο και χρονοβόρο να συλλεχθούν, η ημι–επιβλεπόμενη μάθηση αξιοποιεί ένα μικρό δείγμα δεδομένων με ετικέτες σε συνδυασμό με έναν μεγάλο όγκο μη επισημασμένων δεδομένων. Αυτή η στρατηγική επιτρέπει την εκπαίδευση μοντέλων που ανταγωνίζονται σε απόδοση τα πλήρως επιβλεπόμενα μοντέλα, καθιστώντας την ιδιαίτερα χρήσιμη σε εφαρμογές όπου η συλλογή επισημασμένων δεδομένων είναι δαπανηρή ή μη πρακτική. Η ημι–επιβλεπόμενη μάθηση βρίσκει εφαρμογές σε τομείς όπως η αναγνώριση ομιλίας, η βιοϊατρική και η ανάλυση φυσικής γλώσσας [68].

Παράλληλα, έχει αναπτυχθεί η αυτο–επιβλεπόμενη μάθηση (Self–Supervised Learning), η οποία βασίζεται σε έναν μηχανισμό προεκπαίδευσης, όπου το μοντέλο μαθαίνει να λύνει ένα ενδιάμεσο έργο που έχει οριστεί τεχνητά. Οι ετικέτες για το έργο αυτό παράγονται αυτόματα, επιτρέποντας τη χρήση μη επισημασμένων δεδομένων για την εκπαίδευση. Η διαδικασία προεκπαίδευσης βοηθά το μοντέλο να εξαγάγει πλούσια χαρακτηριστικά από τα δεδομένα, τα οποία στη συνέχεια χρησιμοποιούνται για την εκπαίδευση του κυρίως έργου. Αυτό το στάδιο προεκπαίδευσης έχει αποδειχθεί ιδιαίτερα αποτελεσματικό σε πολλές εφαρμογές, όπως η επεξεργασία φυσικής γλώσσας, η κυβερνοασφάλεια και οι γεννήτριες περιεχομένου (παραγωγή εικόνων, κειμένου ή μουσικής), όπου η αυτο–επιβλεπόμενη μάθηση έχει επιτύχει εξαιρετικά αποτελέσματα [69–71].

Η ποικιλία των προσεγγίσεων στη ML καθιστά τον κλάδο αυτόν εξαιρετικά ευέλικτο και εφαρμόσιμο σε πολυάριθμους τομείς της βιομηχανίας και της έρευνας. Καθώς η τεχνολογία προχωρά, οι αλγόριθμοι ML συνεχίζουν να βελτιώνονται και να εξελίσσονται, ανταποκρινόμενοι στις συνεχώς αυξανόμενες ανάγκες για ανάλυση

δεδομένων και αυτοματοποιημένη λήψη αποφάσεων. Με τη συνδυασμένη χρήση αυτών των τεχνικών, είναι πλέον δυνατή η ανάπτυξη ευφύων συστημάτων που μπορούν να προσαρμοστούν σε ποικίλες συνθήκες και να επιλύσουν πολύπλοκα προβλήματα, ανοίγοντας νέους δρόμους στην υπολογιστική επιστήμη και τη βιομηχανική παραγωγή.

Η συγκριτική αποτίμηση των βασικών κατηγοριών ML, παρουσιάζεται στον Πίνακα 2.2, με τα κύρια πλεονεκτήματα και τους περιορισμούς της κάθε προσέγγισης. Η επιβλεπόμενη μάθηση προσφέρει υψηλή ακρίβεια και ευρεία εφαρμογή, αλλά εξαρτάται σε μεγάλο βαθμό από την ποιότητα και την ποσότητα των εκπαιδευτικών δεδομένων. Η μη επιβλεπόμενη μάθηση επιτρέπει την ανακάλυψη υποκείμενων δομών χωρίς την ανάγκη επισημείωσης, αν και συνοδεύεται από προκλήσεις στην αξιολόγηση της απόδοσης και στην ερμηνεία των αποτελεσμάτων. Η ενισχυτική μάθηση, κατάλληλη για περιβάλλοντα με διαδοχικές αποφάσεις, προσφέρει υψηλή απόδοση σε ρομποτικά ή αυτόνομα συστήματα, αλλά απαιτεί σημαντικούς υπολογιστικούς πόρους και προσεκτικό σχεδιασμό ανταμοιβών. Οι ημι–επιβλεπόμενες και αυτο–επιβλεπόμενες μέθοδοι προσεγγίζουν την αποδοτικότητα των πλήρως επιβλεπόμενων συστημάτων, με μειωμένες απαιτήσεις σε επισημασμένα δεδομένα, καθιστώντας τες ιδιαίτερες χρήσιμες σε εφαρμογές μεγάλης κλίμακας ή με ελλιπή ετικετοποίηση. Η επιλογή της κατάλληλης κατηγορίας ML εξαρτάται από τη φύση του προβλήματος, τη διαθεσιμότητα δεδομένων και τις απαιτήσεις ακρίβειας, επεξηγησιμότητας και αποδοτικότητας.

2.4 Βαθιά Μάθηση

Η βαθιά μάθηση (Deep Learning, DL) αποτελεί έναν από τους πιο δυναμικά εξελισσόμενους υποκλάδους ML, ο οποίος γνωρίζει ραγδαία ανάπτυξη τα τελευταία χρόνια. Οι αλγόριθμοι DL χαρακτηρίζονται από τη χρήση πολυεπίπεδων αρχιτεκτονικών, όπως τα CNNs, τα οποία επιτρέπουν την εξαγωγή υψηλού επιπέδου χαρακτηριστικών από τα δεδομένα. Οι αλγόριθμοι αυτοί διακρίνονται για την ικανότητά τους να αναλύουν και να κατανοούν πολύπλοκα πρότυπα και σχέσεις μέσα στα δεδομένα, κάτι που είναι αδύνατο για τους παραδοσιακούς αλγόριθμους ML που βασίζονται σε περιορισμένα επίπεδα επεξεργασίας [72–74].

Πίνακας 2.2: Συγκριτική αποτίμηση πλεονεκτημάτων και μειονεκτημάτων των βασικών κατηγοριών ML.

Κατηγορία	Πλεονεκτήματα	Μειονεκτήματα
Επιβλεπόμενη	Υψηλή ακρίβεια σε επισημασμένα δεδομένα. Ιδανική για ταξινόμηση και πρόβλεψη. Ευρεία εφαρμογή σε βιοϊατρική, NLP, χρηματοοικονομικά. Πληθώρα καλά εδραιωμένων αλγορίθμων και τεχνικών.	Απαιτεί μεγάλο όγκο επισημασμένων δεδομένων. Επιρρεπής σε υπερπροσαρμογή. Η απόδοση εξαρτάται από την ποιότητα και την αντιπροσωπευτικότητα των επισημασμένων δεδομένων.
Μη Επιβλεπόμενη	Δεν απαιτεί επισημασμένα δεδομένα. Κατάλληλη για ανακάλυψη προτύπων. Χρήσιμη σε clustering και ανίχνευση ανωμαλιών. Μπορεί να αποκαλύψει άγνωστες δομές και σχέσεις.	Χαμηλή ακρίβεια πρόβλεψης. Δύσκολη αξιολόγηση απόδοσης. Ευαισθησία σε παραμέτρους και αρχικοποίηση.
Ενισχυτική	Κατάλληλη για δυναμικά περιβάλλοντα. Μοντελοποιεί στρατηγικές λήψης απόφασης. Ιδανική για ρομποτική και αυτόνομα οχήματα.	Υψηλό υπολογιστικό κόστος και χρονοβόρα εκπαίδευση. Δυσκολία στην επιλογή ανταμοιβών και επίτευξη σύγκλισης.
Ημι–επιβλεπόμενη	Συνδυάζει επιβλεπόμενη και μη επιβλεπόμενη μάθηση. Μειώνει την ανάγκη για μεγάλο όγκο επισημασμένων δεδομένων.	Εξαρτάται από την ποιότητα των επισημασμένων δειγμάτων. Πολύπλοκη διαδικασία εκπαίδευσης.
Αυτο–επιβλεπόμενη	Δεν απαιτεί ανθρώπινη ετικετοποίηση. Ιδανική για προεκπαίδευση μεγάλων μοντέλων. Κατάλληλη για κείμενο, εικόνα, ήχο.	Υψηλές υπολογιστικές απαιτήσεις. Απαιτεί σχεδιασμό κατάλληλων προκαταρκτικών εργασιών (pretext tasks).

Η επιτυχία της DL οφείλεται στη διαστρωματωμένη δομή των αλγορίθμων της, η οποία επιτρέπει την εκπαίδευση σε μη δομημένα δεδομένα, όπως εικόνες, κείμενα και φωνητικά αρχεία, χωρίς να απαιτείται η χειροκίνητη εξαγωγή χαρακτηριστικών. Αυτή η δυνατότητα της αυτοματοποιημένης εξαγωγής χαρακτηριστικών καθιστά τη DL ιδιαίτερα σημαντική σε εφαρμογές όπου τα δεδομένα είναι μεγάλα και σύνθετα. Συγκεκριμένα, τομείς όπως η CV και η NLP έχουν επωφεληθεί σημαντικά από την ενσωμάτωση των αλγορίθμων DL, οι οποίοι επιτυγχάνουν ακρίβεια και αποδοτικότητα σε πολύ υψηλά επίπεδα [75, 76].

Στην CV [77, 78], οι αλγόριθμοι DL επιτρέπουν την αναγνώριση και την κατηγοριοποίηση αντικειμένων σε εικόνες, την ανίχνευση προσώπων, καθώς

και την ανάλυση βίντεο σε πραγματικό χρόνο. Αυτό έχει πολλές πρακτικές εφαρμογές, όπως η αυτόνομη οδήγηση, όπου τα συστήματα DL βοηθούν τα οχήματα να αναγνωρίζουν πεζούς, οδικές πινακίδες και άλλα οχήματα, ενισχύοντας έτσι την ασφάλεια και την απόδοση του οχήματος. Αντίστοιχα, στην NLP [79, 80], οι αλγόριθμοι DL χρησιμοποιούνται σε συστήματα αυτόματης μετάφρασης, αναγνώρισης συναισθημάτων και συνομιλιακά συστήματα (chatbots), όπου επιτυγχάνεται υψηλό επίπεδο κατανόησης και απόκρισης στη φυσική γλώσσα.

Ένα από τα σημαντικότερα πλεονεκτήματα της DL είναι η δυνατότητα της συνεχούς βελτίωσης των αλγορίθμων μέσω της προσθήκης νέων δεδομένων. Καθώς τα δεδομένα αυξάνονται, οι αλγόριθμοι DL προσαρμόζονται και μαθαίνουν, βελτιώνοντας συνεχώς την ακρίβεια και την αποτελεσματικότητά τους. Αυτή η προσαρμοστικότητα καθιστά τη DL μια αποδοτική λύση για εφαρμογές που απαιτούν ακριβή και διαρκώς βελτιούμενη απόδοση, καθώς οι αλγόριθμοι μπορούν να ανταποκριθούν σε μεταβαλλόμενα δεδομένα και νέες συνθήκες [81].

Η DL συνδυάζεται με την αύξηση του διαθέσιμου όγκου δεδομένων και την πρόοδο της υπολογιστικής ισχύος. Η ύπαρξη δεδομένων μεγάλης κλίμακας παρέχει τη βάση για την εκπαίδευση πολύπλοκων CNNs, ενώ οι σύγχρονες υπολογιστικές συσκευές, όπως οι GPUs και οι εξειδικευμένες μονάδες επιτάχυνσης (Tensor Processing Units, TPUs), επιτρέπουν την αποτελεσματική επεξεργασία τεράστιων όγκων δεδομένων σε εύλογους χρόνους [82, 83].

Η σχέση μεταξύ AI, ML και DL απεικονίζεται στο Σχήμα 2.1. Η AI είναι η επιστήμη και η μηχανική δημιουργία ευφύων συστημάτων, δηλαδή υπολογιστικών οντοτήτων ικανών να επιτελούν καθήκοντα που προϋποθέτουν γνωστικές δεξιότητες όπως ο λογισμός, η μάθηση, η αντίληψη και η επίλυση προβλημάτων. Η ML, ως υποπεδίο της AI, πραγματεύεται την ανάπτυξη αλγορίθμων που επιτρέπουν στους υπολογιστές να μαθαίνουν μέσω εμπειρίας και ανάλυσης δεδομένων, χωρίς την ανάγκη ρητής προγραμματιστικής περιγραφής κάθε ενδεχόμενου. Η ML περιλαμβάνει ένα ευρύ φάσμα τεχνικών, όπως η ταξινόμηση, η παλινδρόμηση και η ομαδοποίηση, οι οποίες εφαρμόζονται σε διάφορα προβλήματα εξαγωγής συμπερασμάτων και αναγνώρισης προτύπων. Η DL είναι ένα υποπεδίο της ML και βασίζεται στην αξιοποίηση CNNs, τα οποία έχουν αποδειχθεί αποτελεσματικά στην επίλυση υπολογιστικών προβλημάτων υψηλής πολυπλοκότητας, όπως η αναγνώριση ομιλίας, η ανάλυση εικόνων και βίντεο, καθώς και σε εφαρμογές που προϋποθέτουν την επεξεργασία μεγάλου όγκου δεδομένων.



Σχήμα 2.1: Η ιεραρχική σχέση μεταξύ AI, ML και DL.

2.5 Βασικά στοιχεία των CNNs

Τα CNNs έχουν επιδείξει εξαιρετικές επιδόσεις σε ευρύ φάσμα εργασιών που εμπίπτουν στο πεδίο της CV, με χαρακτηριστικότερο παράδειγμα την αναγνώριση και την ταξινόμηση εικόνων. Η τυπική αρχιτεκτονική ενός CNN περιλαμβάνει ένα επίπεδο εισόδου, διαδοχικά εναλλασσόμενα επίπεδα συνέλιξης (Convolutional Layer) και συγκέντρωσης (Pooling Layer), ένα ή περισσότερα πλήρως συνδεδεμένα επίπεδα (Fully Connected Layer), συναρτήσεις ενεργοποίησης (Activation Functions), καθώς και ένα επίπεδο εξόδου στο τελικό στάδιο.

Το αρχικό τμήμα του δικτύου απαρτίζεται από έναν αριθμό συνελικτικών και συγκεντρωτικών επιπέδων που στοιβάζονται εναλλάξ, συγκροτώντας έναν εξαγωγέα χαρακτηριστικών. Μέσω αυτής της δομής, καθίσταται δυνατή η εφαρμογή πολυεπίπεδων μετασχηματισμών στα ακατέργαστα δεδομένα εισόδου, τα οποία μετασχηματίζονται σταδιακά σε αφηρημένες και υψηλότερης τάξης αναπαραστάσεις χαρακτηριστικών. Τα πλήρως συνδεδεμένα επίπεδα, σε συνδυασμό με συναρτήσεις ενεργοποίησης, αξιοποιούνται για την υλοποίηση της τελικής ταξινόμησης βάσει των εξαγόμενων χαρακτηριστικών. Για τη βελτιστοποίηση της απόδοσης ενός CNN, ενσωματώνονται επιπρόσθετες τεχνικές, όπως η κανονικοποίηση παρτίδας (Batch Normalization) και η απόσυρση (Dropout), καθώς και διάφορες μη γραμμικές συναρτήσεις απεικόνισης. Η κατάλληλη διαμόρφωση και παραμετροποίηση των επιμέρους δομικών στοιχείων ενός CNN είναι κρίσιμη

προϋπόθεση για την ανάπτυξη νέων αρχιτεκτονικών και, κατ' επέκταση, για την επίτευξη υψηλότερης υπολογιστικής απόδοσης [84].

2.5.1 Συνελικτικό Επίπεδο

Το συνελικτικό επίπεδο (Convolutional Layer) αποτελεί το θεμελιώδες δομικό στοιχείο ενός CNN, εντός του οποίου εκτελείται το μεγαλύτερο μέρος των υπολογιστικών πράξεων που απαιτούνται για την επεξεργασία των εισερχόμενων δεδομένων. Κεντρικό στοιχείο της λειτουργικότητάς του συνιστά ένα σύνολο φίλτρων, τα οποία συχνά αποκαλούνται πυρήνες (kernels), και εφαρμόζονται επί των δεδομένων εισόδου με σκοπό την εξαγωγή εννοιολογικά σχετικών και μορφολογικά διακριτών χαρακτηριστικών. Κάθε πυρήνας διαθέτει συγκεκριμένο ύψος, πλάτος και έναν πίνακα βαρών, τα οποία καθορίζουν τον τύπο του χαρακτηριστικού που έχει σχεδιαστεί να ανιχνεύει εντός του δεδομένου πεδίου εισόδου. Κατά τη φάση προώθησης (forward pass) ενός CNN, τα φίλτρα του συνελικτικού επιπέδου εκτελούν συνελίξεις επί των δεδομένων εισόδου (ύψος $\times$ πλάτος $\times$ κανάλια) κατά μήκος των χωρικών τους διαστάσεων. Η συνελικτική πράξη βασίζεται στον υπολογισμό του εσωτερικού γινομένου μεταξύ των βαρών του φίλτρου και των αντίστοιχων τιμών των στοιχείων της εισόδου στο εκάστοτε πεδίο εφαρμογής του φίλτρου [85].

Το αποτέλεσμα της εν λόγω διαδικασίας είναι η παραγωγή ενός δισδιάστατου χάρτη ενεργοποίησης (activation map) για κάθε φίλτρο, ο οποίος κωδικοποιεί την ένταση της ενεργοποίησης του φίλτρου σε κάθε χωρική θέση. Ο χάρτης αυτός αποτυπώνει την απόκριση του φίλτρου και υποδηλώνει την παρουσία του στοχευόμενου χαρακτηριστικού στην αντίστοιχη περιοχή εισόδου. Σε ένα τυπικό συνελικτικό επίπεδο χρησιμοποιούνται πολλαπλοί πυρήνες, καθένας εκ των οποίων εξειδικεύεται στην ανίχνευση διαφορετικών χαρακτηριστικών. Η ταυτόχρονη εφαρμογή αυτών των φίλτρων στο ίδιο δεδομένο εισόδου παράγει αντίστοιχους χάρτες χαρακτηριστικών, οι οποίοι στοιβάζονται κατά μήκος της διάστασης του βάθους, σχηματίζοντας μια τρισδιάστατη αναπαράσταση εξόδου. Η αναπαράσταση αυτή προωθείται στα επόμενα επίπεδα του δικτύου προς περαιτέρω επεξεργασία. Η χρήση πολλαπλών φίλτρων εντός κάθε συνελικτικού επιπέδου ενισχύει τη δυνατότητα του δικτύου να συλλαμβάνει ποικιλία εννοιολογικών και χωρικών χαρακτηριστικών, ακόμη και εντός της ίδιας χωρικής τοποθεσίας, ενδυναμώνοντας κατ' αυτόν τον τρόπο την αναπαραστατική ικανότητα του μοντέλου [86].

2.5.2 Επίπεδο Συγκέντρωσης

Το επίπεδο συγκέντρωσης (Pooling Layer) [87] αποτελεί δομικό συστατικό ενός CNN και εφαρμόζεται συνήθως μετά από ένα ή περισσότερα συνελικτικά επίπεδα, με σκοπό τη σταδιακή μείωση των χωρικών διαστάσεων (ύψος και πλάτος) των χαρτών χαρακτηριστικών που προκύπτουν από τα προηγούμενα επίπεδα. Η εν λόγω μείωση επιτυγχάνεται μέσω της συσσωμάτωσης των τοπικών ομάδων νευρώνων του προηγούμενου επιπέδου σε έναν μόνο νευρώνα στο επόμενο, αποτυπώνοντας μια συνοπτική αναπαράσταση της πληροφορίας στις αντίστοιχες χωρικές περιοχές της εισόδου.

Η ενσωμάτωση επιπέδων συγκέντρωσης εντός της αρχιτεκτονικής συνεπάγεται πολλαπλά πλεονεκτήματα. Πρωτίστως, επιφέρει σημαντική μείωση του αριθμού των παραμέτρων και του υπολογιστικού κόστους του δικτύου, γεγονός που διευκολύνει τόσο τη φάση της εκπαίδευσης όσο και την παραγωγή συμπερασμάτων, συμβάλλοντας στην αποδοτικότερη αξιοποίηση των διαθέσιμων πόρων. Επιπλέον, η μείωση των διαστάσεων λειτουργεί ως ρυθμιστικός μηχανισμός, περιορίζοντας την πιθανότητα υπερεκπαίδευσης, καθώς επιβάλλει στον αλγόριθμο την εκμάθηση πιο γενικεύσιμων αναπαραστάσεων και όχι λεπτομέρειες ή θόρυβο των δεδομένων της εκπαίδευσης.

Σημαντικό επίσης είναι το γεγονός ότι τα επίπεδα συγκέντρωσης προάγουν την ιδιότητα της χωρικής αμεταβλητότητας, καθιστώντας το δίκτυο πιο ανθεκτικό σε μικρές μετατοπίσεις ή κλιμακώσεις της εισόδου. Μέσω τεχνικών όπως η μέγιστη (Max Pooling) ή η μέση συγκέντρωση (Average Pooling), επιτυγχάνεται η διατήρηση της πληροφορίας με τρόπο που δεν εξαρτάται από την ακριβή θέση των χαρακτηριστικών στο επίπεδο εισόδου. Τέλος, η διαδικασία συγκέντρωσης συμβάλλει στην εξαγωγή περισσότερο γενικευμένων αναπαραστάσεων από τα δεδομένα εισόδου και προάγει τη διαμόρφωση μιας ιεραρχικά δομημένης αναπαράστασης χαρακτηριστικών, κατά την οποία τα χαρακτηριστικά χαμηλού επιπέδου (ανιχνευόμενα στα αρχικά συνελικτικά επίπεδα) συνδυάζονται προοδευτικά, ώστε να αναδειχθούν αφηρημένα χαρακτηριστικά υψηλότερης τάξης στα βαθύτερα επίπεδα του δικτύου.

2.5.3 Συναρτήσεις Ενεργοποίησης

Οι συναρτήσεις ενεργοποίησης (Activation Functions) [88] εφαρμόζονται αμέσως μετά από κάθε συνελικτικό ή πλήρως συνδεδεμένο επίπεδο ενός CNN. Ο βασικός ρόλος τους είναι η εισαγωγή μη γραμμικότητας στο υπολογιστικό μοντέλο, μέσω της μη γραμμικής μετασχηματιστικής επεξεργασίας της εξόδου των νευρώνων. Η προσθήκη αυτής της μη γραμμικής συμπεριφοράς καθιστά δυνατή την αναπαράσταση σύνθετων συναρτησιακών σχέσεων και εννοιολογικών δομών που εμπεριέχονται στα δεδομένα εισόδου — ικανότητα που δεν θα μπορούσε να επιτευχθεί μέσω γραμμικών μετασχηματισμών. Χωρίς την ύπαρξη συναρτήσεων ενεργοποίησης, ακόμη και βαθιά δίκτυα με πολλαπλά επίπεδα θα ισοδυναμούσαν ουσιαστικά με μια γραμμική απεικόνιση, περιορίζοντας δραστικά τη δυνατότητά τους να επιλύουν μη γραμμικά υπολογιστικά προβλήματα.

Η επιλογή της κατάλληλης συνάρτησης ενεργοποίησης συνιστά κρίσιμο σχεδιαστικό παράγοντα για τη διαμόρφωση και τη λειτουργικότητα ενός CNN. Κάθε συνάρτηση διαθέτει διακριτές ιδιότητες, οι οποίες επηρεάζουν καθοριστικά τη συμπεριφορά του μοντέλου ως προς τη σύγκλιση κατά την εκπαιδευτική διαδικασία, τη δυνατότητα γενίκευσης και την υπολογιστική αποδοτικότητα.

Στα κρυφά επίπεδα, κυρίαρχη επιλογή αποτελεί η Rectified Linear Unit (ReLU) [89], λόγω της υπολογιστικής της απλότητας και της ικανότητάς της να διατηρεί μη μηδενικές κλίσεις, διευκολύνοντας την εκπαίδευση βαθιών δικτύων. Η συνάρτηση ReLU ορίζεται μαθηματικά ως:

$$\text{ReLU}(x) = \max(0, x) \quad (2.1)$$

Η λειτουργία της είναι ιδιαίτερα απλή, επιστρέφει την είσοδο x αν αυτή είναι θετική, διαφορετικά επιστρέφει μηδέν. Αυτό έχει ως αποτέλεσμα την αραίωση της ενεργοποίησης, καθώς μόνο οι θετικοί νευρώνες παραμένουν ενεργοί, γεγονός που μειώνει το υπολογιστικό φορτίο. Ωστόσο, η ReLU παρουσιάζει ένα σημαντικό μειονέκτημα για αρνητικές τιμές εισόδου, η έξοδος είναι σταθερά μηδέν, οδηγώντας στην πλήρη απενεργοποίηση κάποιων νευρώνων.

Για την αντιμετώπιση του παραπάνω προβλήματος, έχουν προταθεί παραλλαγές της ReLU, όπως η Leaky ReLU [90] και η Exponential Linear Unit (ELU) [91], οι οποίες επιτρέπουν την μη μηδενική έξοδο και για αρνητικές τιμές εισόδου, με στόχο την

αποφυγή της απενεργοποίησης των νευρώνων. Η Leaky ReLU εισάγει μια μικρή γραμμική κλίση για τις αρνητικές τιμές της εισόδου, διατηρώντας έτσι ενεργούς τους νευρώνες σε όλο το πεδίο τιμών του x . Ο μαθηματικός της ορισμός είναι:

$$\text{Leaky ReLU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{if } x < 0 \end{cases} \quad (2.2)$$

όπου α είναι μια μικρή θετική σταθερά (συνήθως $\alpha = 0.01$). Η ύπαρξη μη μηδενικής παραγώγου για τις αρνητικές τιμές εισόδου συμβάλλει στη βελτίωση της εκπαίδευσης, αποφεύγοντας τη στασιμότητα των βαρών.

Αντίστοιχα, η ELU αποτελεί μια πιο προηγμένη μορφή, η οποία χρησιμοποιεί εκθετική κλίση για τις αρνητικές τιμές εισόδου. Ο μαθηματικός ορισμός της είναι:

$$\text{ELU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha(e^x - 1), & \text{if } x < 0 \end{cases} \quad (2.3)$$

Η βασική της διαφορά με τη Leaky ReLU έγκειται στο ότι η εκθετική μορφή της για τις αρνητικές τιμές εισόδου φέρνει την έξοδο της συνάρτησης πιο κοντά στο μηδέν, γεγονός που ευνοεί την ταχύτερη και πιο σταθερή σύγκλιση κατά την εκπαίδευση. Επιπλέον, λόγω της συνέχειας και της ομαλότητας της παραγώγου της, προάγει πιο ομαλές και σταθερές μεταβολές στα βάρη του δικτύου, περιορίζοντας φαινόμενα αστάθειας κατά την εκπαιδευτική διαδικασία.

Αντιθέτως, στο επίπεδο εξόδου, η επιλογή συνάρτησης ενεργοποίησης εξαρτάται από τη φύση του προβλήματος ταξινόμησης. Σε προβλήματα δυαδικής ταξινόμησης εφαρμόζεται ευρέως η συνάρτηση ενεργοποίησης Sigmoid, η οποία μετασχηματίζει την έξοδο στο διάστημα $(0,1)$, καθιστώντας δυνατή την αποτύπωσή της ως πιθανότητα. Ο μαθηματικός ορισμός της συνάρτησης Sigmoid είναι ο εξής:

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2.4)$$

Σε πολυκατηγορικά προβλήματα, η συνάρτηση ενεργοποίησης Softmax λειτουργεί ως μηχανισμός κανονικοποίησης, ώστε η έξοδος να αναπαριστά έγκυρη κατανομή πιθανότητας (άθροισμα ίσο με 1), διευκολύνοντας την απόδοση κατηγορίας στο προς ταξινόμηση δείγμα. Ο μαθηματικός ορισμός της συνάρτησης Softmax είναι ο εξής:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^N e^{x_j}}, \quad \text{για } i = 1, \dots, N \quad (2.5)$$

όπου x_i είναι το i -οστό στοιχείο του διανύσματος εισόδου, N το πλήθος των κατηγοριών και j ο δείκτης που διατρέχει όλα τα στοιχεία του διανύσματος.

Ο Πίνακας 2.3 παρουσιάζει την συγκριτική αποτίμηση των επικρατέστερων συναρτήσεων ενεργοποίησης στην αρχιτεκτονική των CNNs, αναδεικνύοντας τα πλεονεκτήματα και τους περιορισμούς της κάθε προσέγγισης.

Πίνακας 2.3: Συγκριτική αποτίμηση βασικών συναρτήσεων ενεργοποίησης ως προς τα πλεονεκτήματα και τα μειονεκτήματά τους.

Συνάρτηση	Πλεονεκτήματα	Μειονεκτήματα
ReLU	Υπολογιστικά απλή και ταχεία στη σύγκλιση. Αποτρέπει το πρόβλημα της εξαφάνισης των κλίσεων. Ενισχύει την παραγωγή αραιών αναπαραστάσεων.	Πιθανότητα ανενεργών νευρώνων. Έξοδος μη κεντροθετημένη στο μηδέν. Αγνοεί τις αρνητικές τιμές.
Leaky ReLU	Διατηρεί θετική κλίση και για αρνητικές τιμές. Αντιμετωπίζει το πρόβλημα των ανενεργών νευρώνων.	Απαιτεί ρύθμιση της παραμέτρου α . Η απόδοσή της δεν είναι πάντα ανώτερη της ReLU.
ELU	Κεντροθετημένη στο μηδέν έξοδος. Διατηρεί μη μηδενική κλίση για αρνητικές τιμές. Ταχύτερη και πιο σταθερή σύγκλιση σε σχέση με ReLU.	Υπολογιστικά πιο απαιτητική Απαιτεί ρύθμιση της παραμέτρου α . Πιθανή αριθμητική αστάθεια για πολύ μικρές τιμές.
Sigmoid	Κατάλληλη για δυαδική ταξινόμηση. Ερμηνεύσιμη ως πιθανότητα (έξοδος στο διάστημα (0,1)).	Κορεσμός και εξαφάνιση κλίσεων για ακραίες τιμές. Μη κεντροθετημένη στο μηδέν έξοδος.
Softmax	Παράγει κατανομή πιθανότητας για πολλαπλές κατηγορίες. Ιδανική για πολυκατηγορική ταξινόμηση.	Υπολογιστικά πιο απαιτητική. Ακατάλληλη για πολυετικετοποίηση.

2.5.4 Κανονικοποίηση Παρτίδας

Η κανονικοποίηση παρτίδας (Batch Normalization) [92] αποτελεί μια τεχνική, η οποία στοχεύει στη σταθεροποίηση και την επιτάχυνση της διαδικασίας εκπαίδευσης. Βασίζεται στην ιδέα της κανονικοποίησης των εξόδων κάθε ενδιάμεσου επιπέδου κατά μήκος κάθε χαρακτηριστικού, εντός της κάθε παρτίδας δεδομένων εκπαίδευσης. Συγκεκριμένα, για κάθε χαρακτηριστικό υπολογίζονται η μέση τιμή και η διασπορά σε επίπεδο παρτίδας, και στη συνέχεια κάθε έξοδος κανονικοποιείται, ώστε να αποκτήσει μηδενική μέση τιμή και διακύμανση ίση με ένα. Στη συνέχεια, εφαρμόζεται μια γραμμική μετασχηματιστική λειτουργία μέσω παραμέτρων κλιμάκωσης και μετατόπισης (Scale and Shift), επιτρέποντας στο δίκτυο να διατηρεί τη δυνατότητα εκμάθησης της αρχικής κατανομής, εφόσον αυτό ευνοεί τη μάθηση.

Η εφαρμογή της κανονικοποίησης παρτίδας αντιμετωπίζει αποτελεσματικά το πρόβλημα της εσωτερικής μεταβολής κατανομών (Internal Covariate Shift), δηλαδή της συνεχούς αλλαγής της κατανομής των εισόδων ενός επιπέδου κατά τη διάρκεια της εκπαίδευσης, το οποίο μπορεί να οδηγήσει σε ασταθή ή βραδεία σύγκλιση. Επιπρόσθετα, συμβάλλει στη μείωση του προβλήματος εξαφάνισης και υπερβολικής αύξησης των κλίσεων (Vanishing/Exploding Gradients), επιτρέπει τη χρήση υψηλότερων ρυθμών εκμάθησης (Learning Rates), επιταχύνει τη σύγκλιση, βελτιώνει την ικανότητα γενίκευσης του μοντέλου και διευκολύνει την αρχικοποίηση των βαρών.

2.5.5 Απόσυρση

Η απόσυρση (Dropout) [93] αποτελεί μία τεχνική, η οποία στοχεύει στη μείωση της υπερπροσαρμογής στα δεδομένα εκπαίδευσης. Ενσωματώνεται στην αρχιτεκτονική του μοντέλου με τη μορφή επιπέδου, το οποίο τοποθετείται συνήθως μετά από πλήρως συνδεδεμένα ή συνελκτικά επίπεδα. Η βασική αρχή της μεθόδου έγκειται στην τυχαία απενεργοποίηση ενός υποσυνόλου νευρώνων σε κάθε επανάληψη της φάσης εκπαίδευσης. Συγκεκριμένα, με πιθανότητα $p \in [0, 1]$ (που καθορίζεται ως υπερπαραμέτρος), κάθε νευρώνας είτε διατηρείται ενεργός είτε μηδενίζεται προσωρινά η συνεισφορά του προς τα επόμενα επίπεδα. Η διαδικασία αυτή εφαρμόζεται αποκλειστικά κατά την διάρκεια της εκπαίδευσης, ενώ κατά τη φάση

αξιολόγησης το δίκτυο λειτουργεί με όλους τους νευρώνες ενεργούς, αποτρέποντας έτσι την απώλεια πληροφορίας.

Η συμβολή της τεχνικής της απόσυρσης είναι διττή. Πρώτον, η τυχαία απενεργοποίηση ενός νευρώνα κατά τη διάρκεια μιας επανάληψης εκπαίδευσης σημαίνει ότι ο συγκεκριμένος νευρώνας δεν ενημερώνει τα βάρη του εκείνη τη χρονική στιγμή, επιβάλλοντας έτσι έναν σκόπιμο περιορισμό που αποτρέπει την υπερπροσαρμογή. Δεύτερον, η μέθοδος αναγκάζει τα επόμενα επίπεδα του δικτύου να επεξεργαστούν την πληροφορία υπό συνθήκες ελλειπών εισόδων, γεγονός που ενισχύει την ανθεκτικότητα της αναπαράστασης και ευνοεί την εκμάθηση γενικευμένων χαρακτηριστικών. Αυτή η εισαγωγή στοχαστικού θορύβου στη φάση εκπαίδευσης αποτελεί ουσιαστικό παράγοντα βελτίωσης της γενίκευσης και της τελικής απόδοσης του μοντέλου.

2.5.6 Πλήρως Συνδεδεμένο Επίπεδο

Το πλήρως συνδεδεμένο επίπεδο (Fully Connected Layer) αποτελεί θεμελιώδες στοιχείο των CNNs, εντός του οποίου κάθε νευρώνας συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου, υλοποιώντας έναν πλήρη γραμμικό μετασχηματισμό της μορφής:

$$y = Wx + b. \quad (2.6)$$

όπου, το διάνυσμα εισόδου x πολλαπλασιάζεται με τον πίνακα βαρών W , με διαστάσεις (αριθμός νευρώνων εξόδου) $\times$ (αριθμός νευρώνων εισόδου), και στη συνέχεια προστίθεται το διάνυσμα μετατόπισης b . Το αποτέλεσμα αυτής της γραμμικής πράξης είναι το διάνυσμα εξόδου y , το οποίο περιλαμβάνει τις ενεργοποιήσεις των νευρώνων του πλήρως συνδεδεμένου επιπέδου. Η διαδικασία αυτή επιτρέπει στο επίπεδο να μοντελοποιήσει πολύπλοκες αλληλεπιδράσεις μεταξύ των νευρώνων, παρέχοντας στο δίκτυο τη δυνατότητα να συνδυάσει και να ενοποιήσει τις πληροφορίες που εξάγονται από τα προηγούμενα επίπεδα.

Τα πλήρως συνδεδεμένα επίπεδα τοποθετούνται συνήθως στο τελικό στάδιο της αρχιτεκτονικής, με σκοπό τη χαρτογράφηση των χαρακτηριστικών υψηλού επιπέδου, τα οποία έχουν εξαχθεί από τα προηγούμενα συνελκτικά και συγκεντρωτικά επίπεδα, στις κατηγορίες εξόδου της ταξινόμησης. Παρά την

ενίσχυση της εκφραστικής ικανότητας του δικτύου, τα επίπεδα αυτά εισάγουν σημαντικό αριθμό παραμέτρων, γεγονός που συνεπάγεται αυξημένες υπολογιστικές απαιτήσεις και μεγαλύτερο κίνδυνο υπερεκπαίδευσης, ιδίως σε δίκτυα μεγάλης κλίμακας.

2.6 Επιτάχυνση Υλικού στην AI

Η επιτάχυνση υλικού στην AI συνίσταται στην αξιοποίηση εξειδικευμένων ετερογενών υπολογιστικών αρχιτεκτονικών για την αποδοτικότερη εκτέλεση απαιτητικών αλγορίθμων, σε σύγκριση με τη χρήση της συμβατικής CPU. Επιταχυντές όπως οι GPUs, NPU, FPGAs, ASICs και SOMs διαδραματίζουν καθοριστικό ρόλο στην υποστήριξη και την βελτιστοποίηση εφαρμογών AI, τόσο κατά τη φάση της εκπαίδευσης όσο και κατά την εξαγωγή συμπερασμάτων. Η σχεδίαση αυτών των αρχιτεκτονικών στοχεύει στη βέλτιστη αναλογία της υπολογιστικής ισχύος προς την κατανάλωση ενέργειας, εξασφαλίζοντας παράλληλα μειωμένο κόστος υλοποίησης και επιτάχυνση της χρονικής απόκρισης. Η χρήση τέτοιων επιταχυντών καθιστά εφικτή την υλοποίηση σύνθετων αλγορίθμων AI, οι οποίοι απαιτούν υψηλό βαθμό παραλληλίας ξεπερνώντας τους περιορισμούς των συμβατικών επεξεργαστών [94, 95].

2.6.1 Ετερογενείς Αρχιτεκτονικές Επιτάχυνσης

Η επιτάχυνση της AI υλοποιείται μέσω διαφόρων ετερογενών αρχιτεκτονικών υλικού, οι οποίες διαφοροποιούνται ως προς τα χαρακτηριστικά, τα πλεονεκτήματα και τα μειονεκτήματά τους. Οι κυριότερες ετερογενείς αρχιτεκτονικές περιλαμβάνουν GPUs, NPU, TPU, FPGAs, ASICs και SOMs [96–99]. Ο Πίνακας 2.4 παρουσιάζει μία συγκριτική αξιολόγηση των σημαντικότερων ετερογενών αρχιτεκτονικών επιτάχυνσης στην AI.

Οι GPUs χαρακτηρίζονται από μια μαζικά παράλληλη αρχιτεκτονική με πολλούς υπολογιστικούς πυρήνες και υψηλή διακίνηση δεδομένων. Προσφέρουν υψηλή υπολογιστική ισχύ, εκτεταμένη υποστήριξη από εργαλεία ανάπτυξης, καθώς και ένα ώριμο οικοσύστημα λογισμικού. Ωστόσο, παρουσιάζουν υψηλή κατανάλωση ενέργειας και δεν επιτρέπουν προσαρμογές στην αρχιτεκτονική τους, γεγονός που

περιορίζει την αποδοτικότητά τους σε εφαρμογές που απαιτούν εξειδίκευση σε επίπεδο υλικού, όπως η υλοποίηση CNNs σε φορητές ή ενσωματωμένες συσκευές.

Οι NPUs είναι εξειδικευμένες μονάδες σχεδιασμένες ειδικά για NNs, προσφέροντας αποτελεσματικούς υπολογισμούς σε πολλαπλασιασμούς πινάκων. Διακρίνονται από χαμηλή κατανάλωση ενέργειας, μικρή καθυστέρηση για εξαγωγή συμπερασμάτων και ενεργειακή αποδοτικότητα. Ωστόσο, είναι λιγότερο ευέλικτες σε γενικούς υπολογισμούς και ενδέχεται να απαιτούν περαιτέρω προσαρμογές ή βελτιστοποιήσεις για την υποστήριξη νέων ή μη τυποποιημένων μοντέλων.

Τα TPUs είναι εξειδικευμένοι επεξεργαστές που έχουν σχεδιαστεί για την επιτάχυνση των υπολογισμών σε NNs, με έμφαση στους πολλαπλασιασμούς πινάκων, οι οποίοι αποτελούν βασική λειτουργία στην εκπαίδευση και την εξαγωγή συμπερασμάτων. Η αρχιτεκτονική τους περιλαμβάνει μεγάλο αριθμό υπολογιστικών πυρήνων, προσφέροντας υψηλή απόδοση και χαμηλή καθυστέρηση. Ωστόσο, η εξειδίκευσή τους περιορίζει την ευελιξία σε εφαρμογές γενικού σκοπού, ενώ η ενεργειακή τους αποδοτικότητα είναι βέλτιστη μόνο σε μοντέλα συμβατά με τον αρχικό σχεδιασμό.

Τα FPGAs είναι επαναδιαμορφώσιμα ολοκληρωμένα κυκλώματα που παρέχουν υψηλή ευελιξία και δυνατότητα παράλληλης επεξεργασίας. Πλεονεκτούν στην προγραμματιστική ευελιξία, τη δυνατότητα επαναπρογραμματισμού και τη χαμηλή καθυστέρηση για συγκεκριμένες εργασίες. Αντίθετα, μειονεκτούν στο υψηλό κόστος ανάπτυξης, τη δυσκολία προγραμματισμού και τη συγκριτικά χαμηλότερη απόδοση σε γενικούς φόρτους εργασίας.

Τα ASICs είναι πλήρως εξειδικευμένα ολοκληρωμένα κυκλώματα με σχεδιασμό ειδικά προσαρμοσμένο για συγκεκριμένες εργασίες ή μοντέλα. Προσφέρουν βέλτιστη ενεργειακή αποδοτικότητα, μικρή καθυστέρηση και υψηλή απόδοση ανά Watt. Ωστόσο, χαρακτηρίζονται από έλλειψη ευελιξίας, μεγάλο κόστος σχεδιασμού και κατασκευής, καθώς και χρονοβόρο κύκλο ανάπτυξης.

Τα SOMs είναι ενσωματωμένα συστήματα που περιέχουν CPU, GPU και NPU σε μια ενιαία συσκευή, καθιστώντας τα ιδανικά για εφαρμογές υπολογιστικής στην άκρη του δικτύου (Edge Computing). Προσφέρουν χαμηλή κατανάλωση, υψηλή απόδοση και ευκολία ενσωμάτωσης. Ωστόσο, οι θερμικοί περιορισμοί και το υψηλό κόστος μονάδας είναι μειονεκτήματα που πρέπει να λαμβάνονται υπόψη.

Πίνακας 2.4: Σύγκριση ετερογενών αρχιτεκτονικών υλικού για επιτάχυνση ΑΙ.

Υλικό	Χαρακτηριστικά	Πλεονεκτήματα	Μειονεκτήματα
GPU	Μαζικά παράλληλη αρχιτεκτονική. Πολλοί πυρήνες. Υψηλή διακίνηση δεδομένων.	Υψηλή υπολογιστική ισχύς. Ευρεία υποστήριξη λογισμικού. Ωριμο οικοσύστημα.	Υψηλή κατανάλωση ενέργειας. Περιορισμένη ευελιξία σε επίπεδο υλικού.
NPU	Εξειδικευμένη μονάδα για νευρωνικά δίκτυα. Αποτελεσματικός πολλαπλασιασμός πινάκων.	Χαμηλή κατανάλωση. Χαμηλή καθυστέρηση. Ενεργειακή αποδοτικότητα.	Λιγότερη ευελιξία σε γενικούς υπολογισμούς. Απαιτεί προσαρμογή για νέα μοντέλα.
TPU	Εξειδικευμένος επεξεργαστής για νευρωνικά δίκτυα. Βελτιστοποιημένος πολλαπλασιασμός πινάκων.	Υψηλή υπολογιστική απόδοση. Χαμηλή καθυστέρηση σε εκπαίδευση/συμπεράσματα.	Περιορισμένη ευελιξία σε γενικούς υπολογισμούς. Βέλτιστη ενεργειακή απόδοση μόνο σε συμβατά μοντέλα.
FPGA	Επαναδιαμορφώσιμη αρχιτεκτονική. Παράλληλη επεξεργασία. Υψηλή ευελιξία.	Επαναπρογραμματισμός. Χαμηλή καθυστέρηση. Βέλτιστη ενεργειακή αποδοτικότητα.	Υψηλό κόστος ανάπτυξης. Δυσκολία προγραμματισμού. Χαμηλότερη απόδοση σε γενικές εργασίες.
ASIC	Πλήρως εξειδικευμένη αρχιτεκτονική. Σχεδιασμένη για συγκεκριμένα μοντέλα ή εργασίες.	Καλό ενεργειακό προφίλ. Χαμηλή καθυστέρηση. Υψηλή απόδοση ανά Watt.	Έλλειψη ευελιξίας. Υψηλό κόστος σχεδιασμού και κατασκευής. Χρονοβόρος κύκλος ανάπτυξης.
SoM	Ενσωματωμένο σύστημα με επιταχυντές. Κατάλληλο για edge computing.	Χαμηλή κατανάλωση ενέργειας. Υψηλή απόδοση. Ωριμο οικοσύστημα.	Απαιτεί θερμική διαχείριση. Υψηλό κόστος μονάδας.

2.6.2 Τεχνικές Επιτάχυνσης Υλικού και Βελτιστοποίηση

Η επιτάχυνση υλικού στην ΑΙ αξιοποιεί διάφορες προηγμένες τεχνικές και στρατηγικές βελτιστοποίησης, με στόχο την ενίσχυση της απόδοσης, τη μείωση της κατανάλωσης ενέργειας και την αύξηση της ταχύτητας εκτέλεσης των αλγορίθμων ML και DL. Οι τεχνικές αυτές εφαρμόζονται ευρέως σε αρχιτεκτονικές GPUs, FPGAs, ASICs και SOMs [100].

- Παράλληλη επεξεργασία (Parallel Processing): Οι επιταχυντές υλικού αξιοποιούν παράλληλα πολλαπλές υπολογιστικές μονάδες για τη μείωση του χρόνου επεξεργασίας και την αύξηση της υπολογιστικής απόδοσης.

Η παράλληλη εκτέλεση είναι ιδιαίτερα αποτελεσματική σε εφαρμογές AI, όπως η εκτέλεση μεγάλων συνελίξεων ή πολλαπλασιασμών πινάκων, που αποτελούν τη βάση των αλγορίθμων DL. Η ενίσχυση της παραλληλίας όχι μόνο επιταχύνει την επεξεργασία, αλλά και αυξάνει την ενεργειακή απόδοση, αφού επιτυγχάνεται υψηλότερος ρυθμός υπολογισμών ανά μονάδα ενέργειας (operations per Watt).

- Μείωση αριθμητικής ακρίβειας (Quantization): Η κβαντοποίηση αναφέρεται στη χρήση της αριθμητικής αναπαράστασης χαμηλότερης ακρίβειας (π.χ. FP16, INT8) έναντι της πλήρους ακρίβειας (FP32), με στόχο τη μείωση του υπολογιστικού κόστους, του αποθηκευτικού χώρου και της κατανάλωσης ενέργειας. Αντί να χρησιμοποιούνται 32-bit κινητής υποδιαστολής για όλους τους υπολογισμούς και τα βάρη, τα δεδομένα αναπαρίστανται με λιγότερα bit, επιτρέποντας περισσότερους υπολογισμούς ανά κύκλο και μικρότερο εύρος μεταφοράς στην μνήμη. Η μετάβαση από FP32 σε INT8 ή FP16 μπορεί να επιφέρει σημαντική αύξηση στην ταχύτητα εκτέλεσης και στη μείωση της ενεργειακής κατανάλωσης χωρίς ουσιαστική απώλεια στην απόδοση του μοντέλου. Η κβαντοποίηση δεν περιορίζεται μόνο στις παραμέτρους (βάρη) του δικτύου, αλλά εφαρμόζεται και στις έξοδους που παράγονται από τους νευρώνες, επιτρέποντας την πλήρη αριθμητική βελτιστοποίηση του μοντέλου.
- Αφαίρεση πλεοναζόντων παραμέτρων (Pruning): Πολλά CNNs περιλαμβάνουν πλεονάζουσες ή μη κρίσιμες παραμέτρους, οι οποίες δεν συμβάλλουν ουσιαστικά στην τελική απόδοση του μοντέλου. Η τεχνική του κλαδέματος επιτρέπει την απομάκρυνση τέτοιων στοιχείων, μειώνοντας τον αριθμό των υπολογιστικών πράξεων και το συνολικό μέγεθος του μοντέλου, χωρίς σημαντική απώλεια στην ακρίβεια εξαγωγής συμπερασμάτων.
- Βελτιστοποίηση της ιεραρχίας μνήμης (Memory Hierarchy Optimization): Η μεταφορά δεδομένων από και προς τη βασική μνήμη είναι συχνά πιο αργή και συνοδεύεται από αυξημένο ενεργειακό και χρονικό κόστος σε σύγκριση με την ίδια την υπολογιστική πράξη. Οι επιταχυντές AI αξιοποιούν πολυεπίπεδες ιεραρχίες μνήμης και βέλτιστα πρότυπα πρόσβασης, επιδιώκοντας να ελαχιστοποιήσουν το ενεργειακό κόστος και τις καθυστερήσεις που σχετίζονται με την προσπέλαση των δεδομένων. Η βελτιστοποίηση της ιεραρχίας μνήμης δεν είναι μόνο τεχνική απαίτηση για υψηλές επιδόσεις, αλλά αποτελεί ουσιώδες στοιχείο για την ανάπτυξη αποδοτικών AI συστημάτων σε περιβάλλοντα περιορισμένων πόρων.

2.6.3 Αλληλεπίδραση Επιτάχυνσης Υλικού και Αποδοτικότητας σε HCS

Η ενσωμάτωση των CNNs σε HCS καθορίζεται σε μεγάλο βαθμό από τις απαιτήσεις και τους περιορισμούς που ενέχει η φάση της εξαγωγής συμπερασμάτων. Παρότι τα CNNs έχουν αποδειχθεί ιδιαίτερα αποτελεσματικά σε προβλήματα CV και ταξινόμησης, η υλοποίησή τους σε συστήματα με περιορισμένους πόρους συνοδεύεται από σημαντικές τεχνικές προκλήσεις [101].

Η υψηλή υπολογιστική πολυπλοκότητα που χαρακτηρίζει τα CNNs οφείλεται στην επαναλαμβανόμενη εφαρμογή των συνελκτικών φίλτρων και των μη γραμμικών συναρτήσεων ενεργοποίησης, σε πολλαπλά επίπεδα του δικτύου. Συγκεκριμένα, κάθε συνελκτικό επίπεδο εκτελεί μεγάλο αριθμό πράξεων πολλαπλασιασμού και άθροισης σε πολυδιάστατους πίνακες χαρακτηριστικών, ενώ η ενσωμάτωση μη γραμμικών μετασχηματισμών (όπως η ReLU ή η ELU) αυξάνει περαιτέρω το υπολογιστικό φορτίο. Η επαναλαμβανόμενη εκτέλεση αυτών των υπολογιστικά εντατικών λειτουργιών καθιστά την υλοποίηση των CNNs ιδιαίτερα απαιτητική σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους, όπως τα ενσωματωμένα συστήματα και τις συσκευές άκρου.

Οι απαιτήσεις σε μνήμη που προκύπτουν από την αποθήκευση των εκπαιδευμένων παραμέτρων (π.χ. βάρη και τιμές πόλωσης) και των ενδιάμεσων ενεργοποιήσεων (δηλαδή των εξόδων κάθε επιπέδου κατά τη διαδικασία εξαγωγής συμπερασμάτων) μπορεί να ξεπερνούν τις δυνατότητες των διαθέσιμων υπολογιστικών πόρων. Το πρόβλημα αυτό καθίσταται εντονότερο σε βαθιά δίκτυα, τα οποία περιλαμβάνουν μεγάλο αριθμό επιπέδων και συνεπώς αυξημένο αποτύπωμα παραμέτρων, καθώς και σε αρχιτεκτονικές που έχουν εκπαιδευτεί να επεξεργάζονται χαρακτηριστικά μεγάλης διαστασιμότητας. Σε τέτοια περιβάλλοντα, η περιορισμένη χωρητικότητα μνήμης —ειδικά σε ενσωματωμένες ή φορητές συσκευές— ενδέχεται να αποτελέσει κρίσιμο εμπόδιο για την αποτελεσματική υλοποίηση του δικτύου.

Επιπλέον, η ενεργειακή αποδοτικότητα αναδεικνύεται ως κρίσιμος παράγοντας κατά την ανάπτυξη εφαρμογών DL σε HCS, ιδίως σε περιβάλλοντα κινητών, φορητών ή αυτόνομων συσκευών χαμηλής ισχύος. Η συνεχής λειτουργία πολύπλοκων νευρωνικών αρχιτεκτονικών, όπως τα CNNs, οδηγεί σε σημαντική κατανάλωση ενέργειας, γεγονός που μπορεί να καταστήσει μη βιώσιμη τη χρήση τους σε σενάρια όπου η παρατεταμένη αυτονομία και η ενεργειακή αποδοτικότητα αποτελούν ουσιώδεις σχεδιαστικές απαιτήσεις.

Παράλληλα, η επίτευξη χαμηλής χρονικής καθυστέρησης αποτελεί καθοριστική απαίτηση σε εφαρμογές πραγματικού χρόνου, όπως η αυτόνομη πλοήγηση, η βιοϊατρική παρακολούθηση και τα συστήματα ανίχνευσης κρίσιμων συμβάντων, όπου η άμεση παραγωγή συμπερασμάτων είναι ζωτικής σημασίας για την αξιοπιστία και την ασφάλεια της λειτουργίας. Ωστόσο, ο αυξημένος υπολογιστικός φόρτος των CNNs μπορεί να οδηγήσει σε ανεπιθύμητες καθυστερήσεις κατά τη φάση εξαγωγής συμπερασμάτων, περιορίζοντας την πρακτική χρησιμότητα του συστήματος σε τέτοιες χρονικά κρίσιμες εφαρμογές.

Επιπρόσθετα, τα βαθιά CNNs χαρακτηρίζονται συχνά από ιδιαίτερα μεγάλο αριθμό παραμέτρων, γεγονός που προκαλεί αυξημένες απαιτήσεις αποθήκευσης, μεταφοράς και μνήμης, ιδίως κατά τη φάση εξαγωγής συμπερασμάτων σε συστήματα με περιορισμένους πόρους. Η συρρίκνωση του μεγέθους των μοντέλων —μέσω τεχνικών όπως η συμπίεση, η κβαντοποίηση και η αραίωση— χωρίς να παρατηρείται ουσιώδης υποβάθμιση της ακρίβειας, αποτελεί ενεργό και διαρκώς εξελισσόμενο πεδίο έρευνας στο πλαίσιο της αποδοτικής ανάπτυξης των CNNs.

Η υλοποίηση και ανάπτυξη των CNNs σε συσκευές άκρου, όπως ενσωματωμένα συστήματα και συσκευές IoT, συνεπάγεται ειδικές τεχνολογικές προκλήσεις, οι οποίες απορρέουν από τη φύση των περιορισμένων υπολογιστικών πόρων και τους ενεργειακούς περιορισμούς. Η αποτελεσματική προσαρμογή των αρχιτεκτονικών DL ώστε να παραμένουν λειτουργικές και αποδοτικές υπό αυτούς τους περιορισμούς αποτελεί κρίσιμη απαίτηση για την υλοποίηση καινοτόμων και βιώσιμων εφαρμογών AI σε περιβάλλοντα αιχμής.

2.7 Σύνοψη Κεφαλαίου

Το παρόν κεφάλαιο εισάγει τις βασικές θεωρητικές έννοιες και τεχνολογίες που αποτελούν το υπόβαθρο της διατριβής, εστιάζοντας στον ρόλο των HCS για τις σύγχρονες εφαρμογές υψηλών απαιτήσεων. Αναδεικνύεται η σημασία των HCS ως υποσχόμενων λύσεων για την ικανοποίηση των απαιτήσεων υψηλής απόδοσης, ταχύτητας και ενεργειακής αποδοτικότητας σε απαιτητικές εφαρμογές. Παρουσιάζεται η εννοιολογική διασύνδεση μεταξύ AI, ML και DL, με στόχο την αποσαφήνιση των διακριτών χαρακτηριστικών και των μεταξύ τους διαφορών. Δίνεται έμφαση στα CNNs, αναλύοντας τόσο τη δομή και τα βασικά συστατικά τους όσο και τις προκλήσεις που προκύπτουν κατά την ενσωμάτωσή τους σε HCS. Τέλος,

διερευνώνται οι τεχνολογικές εξελίξεις που επιτρέπουν την αποδοτική υλοποίηση και επιτάχυνση τέτοιων δικτύων μέσω εξειδικευμένων αρχιτεκτονικών και στρατηγικών βελτιστοποίησης, θέτοντας τα θεμέλια για την προτεινόμενη μεθοδολογία σχεδίασης και ανάπτυξης σε HCS, η οποία παρουσιάζεται στο επόμενο κεφάλαιο.

Κεφάλαιο 3

Μεθοδολογία Σχεδίασης και Ανάπτυξης σε HCS

Η αποδοτική αξιοποίηση των CNNs σε περιβάλλοντα HCS αποτελεί σύνθετη και παραμετρική πρόκληση, καθώς απαιτείται ισορροπία ανάμεσα στην ακρίβεια εξαγωγής συμπεράσματος, την υπολογιστική απόδοση και την κατανάλωση ενέργειας. Σε αυτό το κεφάλαιο, παρουσιάζεται η προτεινόμενη ολοκληρωμένη μεθοδολογία που αναπτύξαμε, η οποία οργανώνεται σε τρεις άξονες: τις μετρικές σχεδίασης (Design Metrics), τις μετρικές ποιότητας (Quality Metrics) και τις μετρικές υπολογιστικού αποτυπώματος (Footprint Metrics). Στόχος της εν λόγω μεθοδολογίας είναι η διαμόρφωση ενός ολοκληρωμένου πλαισίου αναφοράς, το οποίο υποστηρίζει τον βέλτιστο σχεδιασμό και την αποδοτική υλοποίηση εφαρμογών CNNs σε HCS, διασφαλίζοντας ταυτόχρονα την προσαρμοστικότητα και τη λειτουργική αποτελεσματικότητα των CNNs, ικανών να ανταποκριθούν σε ένα ευρύ φάσμα εφαρμογών και περιορισμών.

3.1 Μεθοδολογική Προσέγγιση

Η ραγδαία πρόοδος της DL τα τελευταία χρόνια οφείλεται, σε μεγάλο βαθμό, στην εφαρμογή και την εξέλιξη των CNNs. Τα CNNs έχουν αναδειχθεί σε βασικό πυλώνα για την αντιμετώπιση πληθώρας προβλημάτων αναγνώρισης και κατηγοριοποίησης εικόνων, χάρη στην ικανότητά τους να εξάγουν ιεραρχικά και πολυεπίπεδα χαρακτηριστικά απευθείας από τα δεδομένα εισόδου. Η εγγενή ικανότητά τους να μοντελοποιούν χωρικές σχέσεις, μέσω συνελκτικών πράξεων

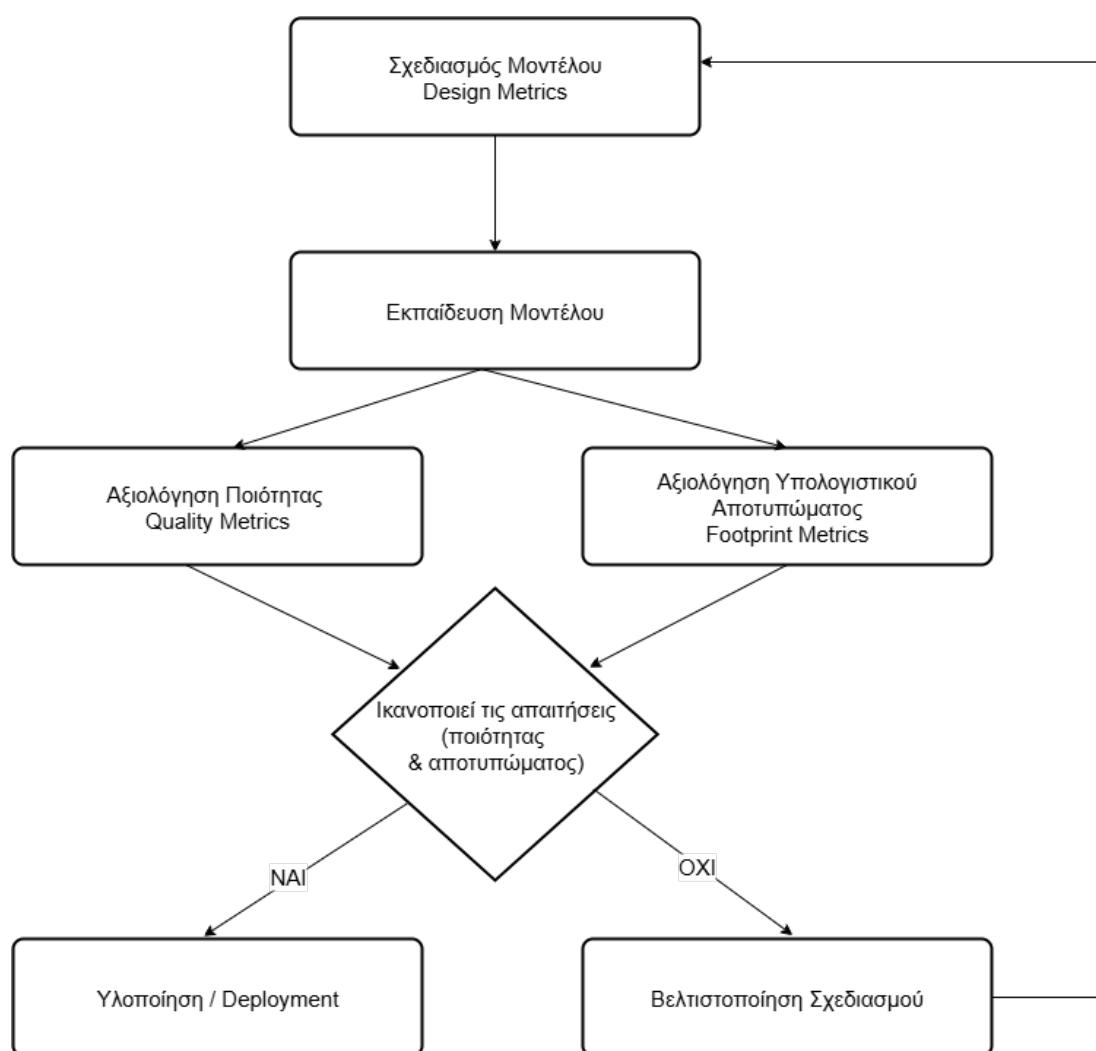
και συγκέντρωσης, επιτρέπει τη διαχείριση μεγάλων όγκων δεδομένων εικόνας με υψηλή αποτελεσματικότητα και ακρίβεια εξαγωγής συμπεράσματος. Ωστόσο, η ανάπτυξη ενός αποδοτικού CNN δεν αποτελεί απλώς ζήτημα μεγιστοποίησης της ακρίβειας σε ένα δεδομένο σύνολο δεδομένων. Αντιθέτως, προϋποθέτει μια σύνθετη και πολυπαραμετρική μεθοδολογία σχεδιασμού και βελτιστοποίησης, η οποία λαμβάνει υπόψη όχι μόνο την ακρίβεια ταξινόμησης αλλά και τις παραμέτρους που σχετίζονται με την υπολογιστική πολυπλοκότητα, την κατανάλωση πόρων, το ενεργειακό αποτύπωμα, καθώς και την εφικτότητα υλοποίησης σε HCS [102, 103].

Η ανάγκη για αποδοτικά CNNs καθίσταται ολοένα και πιο επιτακτική στην εποχή της διάχυτης υπολογιστικής (Pervasive Computing), όπου οι εφαρμογές CV απαιτείται να εκτελούνται σε ένα ευρύ φάσμα ετερογενών συσκευών, από ισχυρούς υπολογιστικούς κόμβους σε κέντρα δεδομένων έως φορητές, ενσωματωμένες και ενεργειακά αυτόνομες συσκευές. Σε αυτό το πολυδιάστατο περιβάλλον, τα HCS διαδραματίζουν καθοριστικό ρόλο, επιτρέποντας τη βέλτιστη αξιοποίηση των υπολογιστικών δυνατοτήτων της εκάστοτε αρχιτεκτονικής, προσαρμοσμένης στις απαιτήσεις της αντίστοιχης εφαρμογής. Ιδιαίτερη έμφαση αποδίδεται στα συστήματα που λειτουργούν υπό περιορισμούς υπολογιστικών και ενεργειακών πόρων, γεγονός που καθιστά κρίσιμη τη σχεδίαση αποδοτικών μοντέλων [104, 105].

Η υλοποίηση των CNNs σε τέτοια HCS απαιτεί την προσαρμογή τόσο της αρχιτεκτονικής δομής όσο και της εκπαιδευτικής διαδικασίας του δικτύου, με απώτερο στόχο την επίτευξη μιας βέλτιστης ισορροπίας μεταξύ της υπολογιστικής απόδοσης και της αποδοτικής διαχείρισης των πόρων. Η ελαχιστοποίηση του συνολικού αποτυπώματος του μοντέλου — σε όρους αριθμού πράξεων κινητής υποδιαστολής ανά δευτερόλεπτο (Floating Point Operations per Second, FLOPs), αριθμού παραμέτρων, απαιτήσεων μνήμης και κατανάλωσης ενέργειας — αναδεικνύεται σε παράμετρο ίσης βαρύτητας με την επίτευξη υψηλής ακρίβειας εξαγωγής συμπεράσματος. Επιπροσθέτως, η αρχιτεκτονική ευελιξία, προκειμένου να προσαρμόζεται δυναμικά στις ιδιαιτερότητες και τις δυνατότητες κάθε υπολογιστικής αρχιτεκτονικής (π.χ. σχεδίαση και υλοποίηση προσαρμοσμένων λογικών πυρήνων σε FPGAs, η βελτιστοποίηση κατανάλωσης ισχύος σε SoCs), αποτελεί κρίσιμο συντελεστή για την επιτυχή ανάπτυξη και εφαρμογή αποδοτικών CNNs σε HCS.

Το Σχήμα 3.1 συνοψίζει τη μεθοδολογική διαδικασία, αποτυπώνοντας την αλληλουχία και τη διασύνδεση των επιμέρους αξόνων αξιολόγησης. Η διαδικασία ξεκινά από τη σχεδίαση της αρχιτεκτονικής, ακολουθείται από την εκπαίδευση

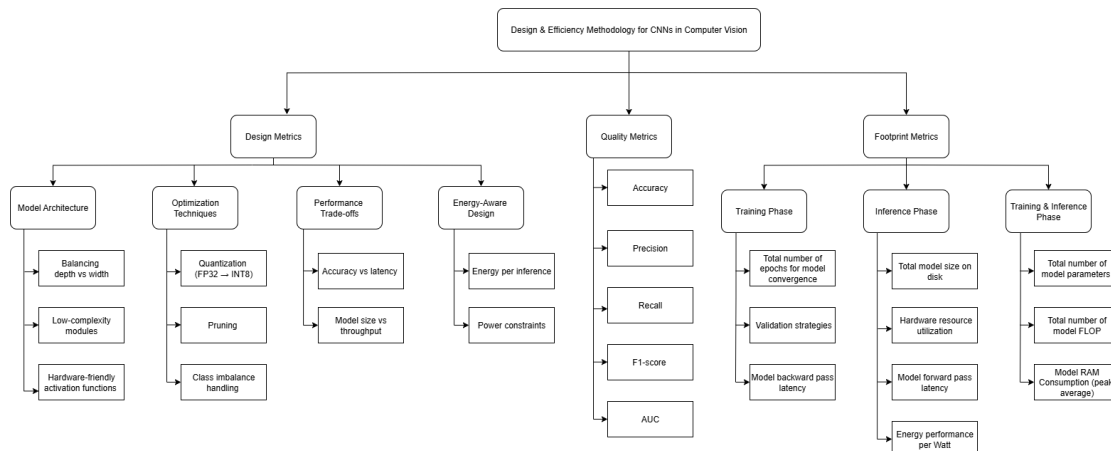
του μοντέλου και κατόπιν πραγματοποιείται διττή αξιολόγηση: αφενός ως προς τις μετρικές ποιότητας και αφετέρου ως προς το υπολογιστικό αποτύπωμα. Στο επόμενο βήμα, πραγματοποιείται ο έλεγχος ως προς την ικανοποίηση όλων των απαιτούμενων κριτηρίων. Εφόσον το μοντέλο πληροί τα κριτήρια που έχουν τεθεί για το εκάστοτε περιβάλλον – στόχο, επιλέγεται για τελική υλοποίηση. Σε αντίθετη περίπτωση, η διαδικασία επιστρέφει στον επανασχεδιασμό του μοντέλου, ακολουθώντας μία επαναληπτική λογική συνεχούς βελτίωσης και προσαρμογής. Η κυκλική αυτή διαδικασία διασφαλίζει ότι η τελική επιλογή της αρχιτεκτονικής ανταποκρίνεται συνολικά στις απαιτήσεις, τόσο ως προς την απόδοση, όσο και ως προς την αποδοτικότητα σε περιβάλλοντα HCS.



Σχήμα 3.1: Διαγραμματική απεικόνιση της πολυκριτηριακής μεθοδολογίας αξιολόγησης καταλληλότητας CNNs σε HCS.

Το Σχήμα 3.2 απεικονίζει τη μεθοδολογική προσέγγιση σχεδίασης και υλοποίησης που αναπτύξαμε. Οι μετρικές διακρίνονται σε τρεις βασικούς άξονες, των οποίων

ο συνδυασμός επιτρέπει την ολοκληρωμένη αποτίμηση της πρακτικής εφικτότητας ενός CNN σε HCS. Η πολυδιάστατη αυτή προσέγγιση είναι καθοριστική για την επιλογή αρχιτεκτονικών λύσεων που επιτυγχάνουν ισορροπία μεταξύ υψηλής υπολογιστικής απόδοσης και περιορισμένου κόστους υλοποίησης, προσαρμοσμένων στις εξειδικευμένες απαιτήσεις του εκάστοτε συστήματος – στόχου.



Σχήμα 3.2: Μεθοδολογική προσέγγιση σχεδίασης και ανάπτυξης των CNNs σε HCS.

3.2 Μετρικές Σχεδιασμού

Οι μετρικές σχεδιασμού (Design Metrics) της μεθοδολογίας μας αναφέρονται στις επιλογές που σχετίζονται με την αρχιτεκτονική και τις τεχνικές βελτιστοποίησης ενός CNN, με στόχο την επίτευξη υψηλής απόδοσης υπό περιορισμούς υπολογιστικών πόρων σε HCS. Στη συνέχεια παρουσιάζονται οι βασικοί άξονες σχεδιασμού, οι οποίοι περιλαμβάνουν τη δομή της αρχιτεκτονικής του μοντέλου, τις τεχνικές αποδοτικής βελτιστοποίησης, τους εκάστοτε συμβιβασμούς, καθώς και τα ζητήματα ενεργειακής αποδοτικότητας που ενσωματώνονται στις σχεδιαστικές επιλογές.

3.2.1 Αρχιτεκτονική Μοντέλου

3.2.1.1 Εξισορρόπηση Βάθους και Εύρους

Ένα θεμελιώδες ζήτημα κατά τον σχεδιασμό CNNs αποτελεί η διαμόρφωση μιας βέλτιστης στρατηγικής εξισορρόπησης μεταξύ του βάθους και του εύρους της αρχιτεκτονικής. Το βάθος προσδιορίζεται από τον αριθμό των διαδοχικών

συνελικτικών επιπέδων, τα οποία επιτρέπουν την εξαγωγή χαρακτηριστικών υψηλότερου επιπέδου μέσω διαδοχικών μετασχηματισμών που οδηγούν σε ολοένα και πιο αφηρημένες αναπαραστάσεις των δεδομένων εισόδου. Αντίστοιχα, το εύρος ορίζεται από τον αριθμό των φίλτρων ανά επίπεδο, επηρεάζοντας άμεσα την ποσότητα των χαρακτηριστικών που εξάγονται σε κάθε στάδιο της διαδικασίας επεξεργασίας [106].

Παραδοσιακά, η αύξηση του βάθους ενισχύει τη χωρητικότητα του δικτύου να μοντελοποιεί σύνθετες χωρικές ιεραρχίες, ενώ η αύξηση του εύρους επιτρέπει τη λεπτομερέστερη αποτύπωση ποικιλίας χαρακτηριστικών σε κάθε επίπεδο. Ωστόσο, η μονομερής αύξηση του βάθους πέραν ενός σημείου οδηγεί σε φθίνουσες αποδόσεις, λόγω φαινομένων όπως η υποβάθμιση κλίσεων και η υπερπροσαρμογή, ειδικά όταν δεν συνοδεύεται από αντίστοιχη επαύξηση του εύρους. Η βέλτιστη κατανομή του αριθμού φίλτρων ανά επίπεδο καθίσταται κρίσιμη, καθώς επιδρά άμεσα στην ισορροπία μεταξύ της εκφραστικής ικανότητας του δικτύου και της υπολογιστικής του αποδοτικότητας [107].

Επιπλέον, η ταυτόχρονη αύξηση βάθους και εύρους συνεπάγεται εκθετική αύξηση στον αριθμό παραμέτρων και, συνακόλουθα, στο υπολογιστικό και ενεργειακό κόστος, γεγονός που θέτει σημαντικούς περιορισμούς σε εφαρμογές όπου οι διαθέσιμοι πόροι είναι περιορισμένοι. Η βέλτιστη αναλογία βάθους προς εύρος είναι άμεσα εξαρτώμενη από τη φύση και την πολυπλοκότητα του συνόλου δεδομένων εκπαίδευσης, καθώς και από τους υφιστάμενους περιορισμούς του υπολογιστικού περιβάλλοντος. Η προσαρμογή των αρχιτεκτονικών παραμέτρων σε κάθε συγκεκριμένο περιβάλλον υλοποίησης συνιστά κρίσιμο παράγοντα για την επίτευξη ισορροπίας μεταξύ απόδοσης και αποδοτικότητας [108].

3.2.1.2 Χαμηλής Πολυπλοκότητας Δομοστοιχεία

Η αυξανόμενη ανάγκη για υλοποίηση CNNs σε HCS με περιορισμένους πόρους, όπως ενσωματωμένες συσκευές και υπολογιστικά συστήματα ακμής, έχει οδηγήσει στην ανάπτυξη ειδικών χαμηλής πολυπλοκότητας δομοστοιχείων. Τα δομοστοιχεία αυτά αξιοποιούν ποικίλες αρχιτεκτονικές τεχνικές προκειμένου να μειώσουν την υπολογιστική πολυπλοκότητα, τον αριθμό παραμέτρων, και τη λανθάνουσα κατάσταση, επιτρέποντας τη διατήρηση της απόδοσης εξαγωγής συμπεράσματος σε υψηλά επίπεδα.

Τα βασικά δομοστοιχεία χαμηλής πολυπλοκότητας:

- Συνελίξεις βάθους ανά κανάλι (Depthwise Separable Convolution, DSConv): Το DSConv [109] διαχωρίζεται σε δύο διακριτά στάδια: (i) συνέλιξη που εφαρμόζεται ανεξάρτητα σε κάθε κανάλι εισόδου (depthwise convolution), και (ii) σημειακή συνέλιξη 1×1 (pointwise convolution) για τη συνένωση των αποτελεσμάτων. Η προσέγγιση αυτή μειώνει το πλήθος των απαιτούμενων υπολογισμών, προσφέροντας εξαιρετική αναλογία αποδοτικότητας προς πολυπλοκότητα.
- Ισορροπημένες συνελίξεις διαχωρισμένων σταδίων (Balanced Separable Convolution, BSConv): Το BSConv [110] αποτελεί εξέλιξη των DSConv, στοχεύοντας στη βελτίωση της ποιότητας εκμάθησης μέσω εξισορρόπησης της συμβολής των καναλιών και της αναπαράστασης χαρακτηριστικών. Η βασική διαφοροποίηση του BSConv έγκειται στην εισαγωγή ενός μηχανισμού παραμετρικής αναβάθμισης των καναλιών, ο οποίος αποτρέπει τη μειωμένη αξιοποίηση των κρίσιμων χαρακτηριστικών λόγω υπερβολικής αραίωσης της σημασιολογικής πληροφορίας. Παράλληλα, διατηρείται η χαμηλή υπολογιστική πολυπλοκότητα μέσω διαχωρισμένων συνελίξεων, ενώ η επιπλέον παραμετροποίηση επιτρέπει στο δίκτυο να επιτύχει υψηλότερη εκφραστική δύναμη με περιορισμένο κόστος.
- Ανεστραμμένες συνελικτικές δομές φιαλοειδούς τύπου για φορητά περιβάλλοντα (Mobile Inverted Bottleneck Convolutions, MBConv): Το MBConv [111] αποτελεί εξελιγμένη μορφή των DSConv. Χαρακτηρίζονται από την αρχική επέκταση των καναλιών μέσω συνελίξεων 1×1 , ακολουθούμενη από συνελίξεις ανά κανάλι, και τελική συμπίεση με άλλη συνέλιξη 1×1 . Η προσέγγιση αυτή, σε συνδυασμό με κατάλληλες συνδέσεις παράκαμψης, προσφέρει υψηλή αποδοτικότητα και εκφραστική ισχύ.
- Συγχωνευμένες συνελίξεις ανεστραμμένου φιαλοειδούς τύπου για φορητά περιβάλλοντα (Fused Mobile Inverted Bottleneck Convolutions, Fused-MBConv): Το Fused-MBConv [112] αποτελεί βελτιστοποιημένη παραλλαγή των MBConv. Η κύρια διαφοροποίηση έγκειται στη συγχώνευση των συνελικτικών σταδίων, καθώς η συνελικτική πράξη ανά κανάλι και η σημειακή συνέλιξη 1×1 αντικαθίστανται από μία παραδοσιακή συνελικτική λειτουργία που εφαρμόζεται σε όλα τα κανάλια εισόδου. Με αυτόν τον τρόπο, απλοποιείται η δομή του μοντέλου, μειώνεται η ανάγκη για επιμέρους

βελτιστοποιήσεις στο υλικό, και επιτυγχάνεται χαμηλότερη λανθάνουσα κατάσταση κατά τον συμπερασμό. Παρότι η πολυπλοκότητα αυξάνεται ελαφρώς σε σύγκριση με το MBConv, το Fused—MBConv επιδεικνύει ανώτερη αποδοτικότητα.

- Αντεστραμμένες δομές τύπου Sandglass: Το Sandglass block [113] αποτελεί παραλλαγή των MBConv, με βασική διαφοροποίηση την αντιστροφή της διάταξης των συνελικτικών σταδίων. Σε αντίθεση με την κλασική προσέγγιση του MBConv, όπου η επέκταση καναλιών προηγείται της συνελικτικής πράξης ανά κανάλι, το Sandglass block τοποθετεί την επέκταση των καναλιών στο τέλος του δομοστοιχείου. Συγκεκριμένα, η επεξεργασία ξεκινά με συμπίεση καναλιών μέσω συνελίξεων 1×1 , ακολουθεί η συνελικτική πράξη ανά κανάλι, και ολοκληρώνεται με επέκταση καναλιών. Η διάταξη αυτή διευκολύνει την καλύτερη ροή πληροφορίας μέσω των συνδέσεων παράκαμψης, ιδιαίτερα στα αρχικά επίπεδα του δικτύου, όπου η απώλεια βασικών χαρακτηριστικών μπορεί να επηρεάσει την τελική απόδοση.
- Δομοστοιχεία τύπου Fire: Κάθε Fire module [114] αποτελείται από δύο βασικά στάδια: το στάδιο συμπίεσης, όπου πραγματοποιείται μέσω συνελίξεων 1×1 , και το στάδιο διέγερσης, το οποίο επαναφέρει την πληροφορία με παράλληλες συνελίξεις 1×1 και 3×3 . Η στρατηγική αυτή μειώνει σημαντικά το υπολογιστικό κόστος, καθώς οι συνελίξεις 1×1 έχουν χαμηλότερη πολυπλοκότητα σε σύγκριση με τις τυπικές συνελίξεις μεγαλύτερου πυρήνα. Επιπλέον, η χρήση πυρήνων 3×3 στο στάδιο διέγερσης επιτρέπει στο δίκτυο να διατηρεί χωρική πληροφόρηση υψηλής ανάλυσης.
- Δομοστοιχεία SqueezeNext: Το SqueezeNext [115] αποτελεί βελτιστοποιημένη εκδοχή των Fire modules, σχεδιασμένη με γνώμονα τη μεγιστοποίηση της αποδοτικότητας. Η αρχιτεκτονική τους βασίζεται στη χρήση στενότερων δομών συμπίεσης τύπου bottlenecks, που επιτυγχάνουν περαιτέρω μείωση του αριθμού παραμέτρων, καθώς και στην εκτενή αξιοποίηση συνελίξεων 1×1 , οι οποίες παρουσιάζουν χαμηλή υπολογιστική πολυπλοκότητα. Παράλληλα, γίνεται περιορισμένη χρήση συνελίξεων 3×3 , μόνο όπου είναι απολύτως απαραίτητο, για τη διατήρηση της χωρικής πληροφορίας.
- Μονάδες συμπίεσης και διέγερσης (Squeeze—and—Excitation blocks): Το SE block [116] εισήγαγε έναν αποδοτικό μηχανισμό προσοχής, ο οποίος δρα σε επίπεδο καναλιών και ενισχύει τη σημασιολογική εκφραστικότητα

του δικτύου. Η βασική αρχή των SE blocks συνίσταται στη δυναμική επανεκτίμηση των καναλιών χαρακτηριστικών μέσω δύο διακριτών φάσεων: (i) συμπίεση, όπου η πληροφορία κάθε καναλιού συνοψίζεται με καθολική μέση τιμή συγκέντρωσης (Global Average Pooling), και (ii) διέγερση, όπου υπολογίζεται η σχετική σημαντικότητα κάθε καναλιού μέσω ενός μικρού πλήρως συνδεδεμένου δικτύου. Το αποτέλεσμα είναι ένας προσαρμοστικός συντελεστής για κάθε κανάλι, ο οποίος ενισχύει ή αποδυναμώνει το αντίστοιχο χαρακτηριστικό. Παρά την προσθήκη επιπλέον παραμέτρων και πράξεων, το υπολογιστικό κόστος των SE blocks παραμένει χαμηλό, καθιστώντας τα ιδανικά για χρήση σε αποδοτικά CNNs που στοχεύουν σε βελτιωμένη απόδοση χωρίς σημαντική αύξηση της πολυπλοκότητας.

- Δομοστοιχεία τύπου Inception: Το Inception module [117] εισάγει την έννοια της παραλληλίας συνελκτικών πράξεων με διαφορετικά μεγέθη πυρήνα (1×1 , 3×3 , 5×5) καθώς και πράξεων μέγιστης συγκέντρωσης (Max Pooling) εντός του ίδιου επιπέδου. Η πολυπλοκότητα της πληροφορίας που εξάγεται από τις εικόνες σε διαφορετικές χωρικές κλίμακες απαιτεί δομές που να μπορούν να συλλάβουν χαρακτηριστικά τόσο λεπτομερή όσο και γενικευμένα. Τα Inception modules διευκολύνουν την εξαγωγή πολυκλιμακικών χαρακτηριστικών (multi-scale feature extraction), αυξάνοντας την εκφραστική ικανότητα του δικτύου χωρίς μεγάλη αύξηση της πολυπλοκότητας. Η ενσωμάτωση συνελίξεων 1×1 ως προκαταρκτικό στάδιο πριν από τις συνελίξεις μεγαλύτερης διάστασης λειτουργεί ως μηχανισμός μείωσης διαστάσεων, μειώνοντας τον αριθμό υπολογισμών και παραμέτρων. Νεότερες παραλλαγές, όπως τα Inception-v3 και Inception-v4, εισάγουν επιπλέον τεχνικές, όπως η διάσπαση των συνελίξεων 5×5 σε διαδοχικές συνελίξεις 3×3 , επιτυγχάνοντας περαιτέρω βελτιστοποίηση της αποδοτικότητας.
- Δομοστοιχεία τύπου ESP (Efficient Spatial Pyramid) και EESP (Extremely Efficient Spatial Pyramid): Το ESP module [118] και το EESP unit [119] αποτελούν προηγμένες δομές για αποδοτική εξαγωγή πολυκλιμακικών χαρακτηριστικών με ελαχιστοποίηση της υπολογιστικής πολυπλοκότητας. Τα ESP modules εφαρμόζουν παράλληλες διατεταμένες συνελίξεις (Dilated Convolutions) με διαφορετικούς ρυθμούς διαστολής (Dilation Rates), επεκτείνοντας αποτελεσματικά το πεδίο λήψης (Receptive Field) χωρίς να αυξάνεται ο αριθμός των παραμέτρων ή το υπολογιστικό κόστος.

Τα EESP units αποτελούν εξελιγμένες εκδοχές των ESP, εισάγοντας ομαδοποιημένες συνελίξεις (Grouped Convolutions) και έναν μηχανισμό ιεραρχικής συγχώνευσης χαρακτηριστικών, ο οποίος συνδυάζει τις εξόδους των διαφόρων διαστολών με τρόπο αποδοτικό τόσο σε μνήμη όσο και σε ταχύτητα.

- Δομοστοιχεία τύπου Ghost: Το Ghost module [120] βασίζεται στην παρατήρηση ότι τα περισσότερα χαρακτηριστικά που παράγονται από τις συμβατικές συνελκτικές στρώσεις παρουσιάζουν σημαντικό βαθμό πλεονασμού, καθώς πολλά από αυτά είναι γραμμικά εξαρτώμενα ή επαναλαμβανόμενα. Η βασική ιδέα των Ghost modules είναι ότι ένα μικρό υποσύνολο των απαιτούμενων χαρακτηριστικών μπορεί να παραχθεί μέσω τυπικών συνελκτικών πράξεων, ενώ τα υπόλοιπα — τα αποκαλούμενα "φανταστικά" χαρακτηριστικά — μπορούν να δημιουργηθούν μέσω μετασχηματισμών χαμηλής υπολογιστικής πολυπλοκότητας, όπως σημειακές συνελίξεις 1×1 ή συνελκτικές πράξεις ανά κανάλι με χαμηλή πολυπλοκότητα. Αυτή η προσέγγιση μειώνει δραστικά τον αριθμό των απαιτούμενων υπολογιστικών πράξεων και των παραμέτρων, διατηρώντας παράλληλα την εκφραστική ικανότητα του μοντέλου σε υψηλά επίπεδα.

Η επιλογή κατάλληλου δομοστοιχείου ή συνδυασμού δομοστοιχείων εξαρτάται από το υπολογιστικό περιβάλλον και τις απαιτήσεις της εφαρμογής, ενώ η μείωση της υπολογιστικής πολυπλοκότητας καθιστά εφικτή τη χρήση προηγμένων CNNs σε ενσωματωμένες συσκευές, υπολογιστικά συστήματα ακμής και IoT, διατηρώντας υψηλά επίπεδα ακρίβειας στην εξαγωγή συμπεράσματος. Ο Πίνακας 3.1 συνοψίζει τα βασικότερα δομοστοιχεία που εφαρμόζονται σε HCS με περιορισμένους υπολογιστικούς και ενεργειακούς πόρους.

3.2.1.3 Συναρτήσεις Ενεργοποίησης Φιλικές προς το Υλικό

Οι συναρτήσεις ενεργοποίησης αποτελούν θεμελιώδες συστατικό των CNNs, καθώς διαδραματίζουν καθοριστικό ρόλο στην εισαγωγή μη γραμμικότητας, η οποία είναι απαραίτητη για την ικανότητα των δικτύων να μαθαίνουν και να μοντελοποιούν πολύπλοκες σχέσεις στα δεδομένα. Η εισαγωγή μη γραμμικότητας επιτρέπει στα CNNs να υπερβαίνουν τους περιορισμούς των γραμμικών μοντέλων και να επιλύουν προβλήματα που περιλαμβάνουν περίπλοκες αντιστοιχίσεις μεταξύ εισόδου και εξόδου [121].

Πίνακας 3.1: Συγκριτική ανάλυση χαμηλής πολυπλοκότητας δομοστοιχείων CNNs.

Δομοστοιχείο	Πλεονεκτήματα	Μειονεκτήματα
DSCnv	Σημαντική μείωση υπολογιστικής πολυπλοκότητας.	Δυνητικά μικρή απώλεια ακρίβειας σε σύγκριση με τυπικές συνελίξεις.
BSCnv	Βελτιωμένη αναπαράσταση χαρακτηριστικών. Διατήρηση χαμηλής πολυπλοκότητας.	Ελαφρώς αυξημένο κόστος λόγω επιπλέον παραμετροποίησης.
MBConv	Υψηλή αποδοτικότητα και ευελιξία. Καλή ισορροπία απόδοσης και πολυπλοκότητας.	Απαιτεί παραμετροποίηση για μέγιστη απόδοση.
Fused-MBConv	Απλούστερη υλοποίηση σε υλικό. Χαμηλότερη καθυστέρηση.	Ελαφρώς αυξημένη πολυπλοκότητα σε σχέση με MBConv.
Sandglass block	Βελτιωμένη ροή πληροφορίας σε αρχικά επίπεδα. Αποδοτικότητα για ρηχά δίκτυα.	Η βέλτιστη απόδοση εξαρτάται από το βάθος του δικτύου.
Fire module	Μείωση παραμέτρων και υπολογιστικού κόστους. Απλή δομή.	Περιορισμένη ικανότητα αναπαράστασης.
SqueezeNext	Εξαιρετικά αποδοτικά για ενσωματωμένα συστήματα. Χαμηλό αποτύπωμα μνήμης.	Μειωμένη γενικευσιμότητα σε πολύπλοκες κατανομές δεδομένων.
SE block	Ενίσχυση σημαντικών χαρακτηριστικών. Βελτίωση απόδοσης με μικρό κόστος.	Ελαφρά αύξηση υπολογιστικού φορτίου.
Inception module	Πολυκλιμακική εξαγωγή χαρακτηριστικών. Υψηλή εκφραστική ισχύς.	Αυξημένη πολυπλοκότητα σχεδίασης και παραμετροποίησης.
ESP module / EESP unit	Αποτελεσματική πολυκλιμακική επεξεργασία.	Υπολογιστική επιβάρυνση από τη διαχείριση πολλαπλών κλάδων.
Ghost module	Δραστική μείωση υπολογιστικού κόστους.	Δυνητική απώλεια λεπτομερειών σε δεδομένα που απαιτούν υψηλή ακρίβεια.

Παράλληλα, με την αυξανόμενη πολυπλοκότητα των μοντέλων DL, αναδεικνύεται η ανάγκη για ανάπτυξη τους σε υπολογιστικά περιβάλλοντα περιορισμένων πόρων, όπως ενσωματωμένα συστήματα και συσκευές άκρων. Αυτή η απαίτηση οδηγεί στην αναζήτηση συναρτήσεων ενεργοποίησης που επιτυγχάνουν ισορροπία μεταξύ ακρίβειας εξαγωγής συμπεράσματος και υπολογιστικής αποδοτικότητας, ώστε να μπορούν να ενσωματωθούν αποτελεσματικά σε HCS. Η ανάγκη για επεξεργασία σε πραγματικό χρόνο και ενεργειακή αποδοτικότητα σε τέτοια συστήματα καθιστά την επιλογή της κατάλληλης συνάρτησης ενεργοποίησης κρίσιμο παράγοντα [122].

Οι βασικές συναρτήσεις ενεργοποίησης που χρησιμοποιούν απλές μαθηματικές πράξεις, γεγονός που τις καθιστά ιδιαίτερα κατάλληλες για υλοποίηση σε HCS με περιορισμένους πόρους:

- Rectified Linear Unit (ReLU): Η συνάρτηση ενεργοποίησης ReLU [89] αποτελεί μία από τις πιο διαδεδομένες επιλογές για την εισαγωγή μη γραμμικότητας σε συνελκτικά και πλήρως συνδεδεμένα επίπεδα. Ορίζεται ως:

$$f(x) = \max(0, x) \quad (3.1)$$

όπου x είναι το γραμμικό άθροισμα των σταθμισμένων εισόδων του νευρώνα, δηλαδή $x = \sum_{i=1}^n w_i x_i + b$, όπου w_i τα βάρη, x_i οι είσοδοι και b ο όρος μετατόπισης. Η ReLU αντικαθιστά όλες τις αρνητικές τιμές εισόδου με μηδέν, ενώ διατηρεί αναλλοίωτες τις θετικές. Η απλότητά της ως προς την υπολογιστική απαίτηση, σε συνδυασμό με την ικανότητά της να περιορίζει το φαινόμενο της εξαφάνισης του βαθμωτού σφάλματος, καθιστούν τη ReLU ιδιαίτερα κατάλληλη για μεγάλης κλίμακας CNNs. Επιπλέον, η υλοποίησή της σε υλικό απαιτεί μόνο συγκρίσεις και μηδενισμούς, γεγονός που την καθιστά εξαιρετικά φιλική προς τους επιταχυντές χαμηλής κατανάλωσης.

- Leaky Rectified Linear Unit (Leaky ReLU): Η συνάρτηση Leaky ReLU [90] προτείνεται ως τροποποίηση της κλασικής ReLU, με σκοπό να αντιμετωπιστεί το πρόβλημα των ανενεργών νευρώνων. Ορίζεται ως:

$$f(x) = \begin{cases} x, & x \geq 0 \\ \alpha x, & x < 0 \end{cases} \quad (3.2)$$

όπου α είναι ένας μικρός θετικός παράγοντας, της τάξης του 10^{-2} . Η Leaky ReLU επιτρέπει τη ροή μικρών αρνητικών τιμών, διατηρώντας ενεργά όλα τα νευρωνικά στοιχεία κατά την εκπαίδευση και την εξαγωγή συμπερασμάτων. Η υλοποίησή της στο υλικό παραμένει απλή, ειδικά όταν ο συντελεστής α είναι δύναμη του δύο.

- Hard Hyperbolic Tangent (Hard Tanh): Η συνάρτηση Hard Tanh [123] αποτελεί μία προσεγγιστική, γραμμική εκδοχή της κλασικής Tanh. Ορίζεται ως:

$$f(x) = \max(-1, \min(1, x)) \quad (3.3)$$

Η έξοδος παραμένει σταθερή στις τιμές $(-1, 1)$ για εισόδους εκτός του διαστήματος. Λόγω της απουσίας εκθετικών υπολογισμών, η Hard Tanh προσφέρει σημαντικά πλεονεκτήματα σε υλοποιήσεις.

- Parametric Rectified Linear Unit (PReLU): Η συνάρτηση PReLU [124] αποτελεί γενίκευση της Leaky ReLU, με την διαφορά ότι ο αρνητικός συντελεστής είναι παραμετροποιήσιμος και εκπαιδευσιμος. Ορίζεται ως:

$$f(x) = \begin{cases} x, & x \geq 0 \\ ax, & x < 0 \end{cases} \quad (3.4)$$

όπου a είναι μία παράμετρος που μαθαίνεται κατά την εκπαίδευση μέσω οπισθοδιάδοσης. Η PReLU επιτρέπει στο δίκτυο να προσαρμόσει δυναμικά τη συμπεριφορά των ενεργοποιήσεων χωρίς σημαντική αύξηση της υπολογιστικής πολυπλοκότητας.

- ReLU6: Η συνάρτηση ReLU6 [111] αποτελεί παραλλαγή της κλασικής ReLU, με την προσθήκη ανώτατου ορίου τιμής στην έξοδο, το οποίο έχει καθοριστεί ίσο με 6. Ορίζεται ως:

$$f(x) = \min(\max(0, x), 6) \quad (3.5)$$

Συμπεριφέρεται όπως η ReLU για εισόδους στο διάστημα $[0, 6]$, ενώ περιορίζει κάθε τιμή μεγαλύτερη του 6 στο 6 και κάθε αρνητική τιμή στο μηδέν. Η εισαγωγή του ανώτατου ορίου αποσκοπεί στη μείωση των υψηλών τιμών ενεργοποιήσεων, διευκολύνοντας την αριθμητική αναπαράστασή τους χωρίς σημαντική απώλεια ακρίβειας.

- Hard Swish: Η συνάρτηση Hard Swish [125] αποτελεί προσεγγιστική εκδοχή της Swish, σχεδιασμένη με σκοπό τη μείωση της υπολογιστικής πολυπλοκότητας και τη διευκόλυνση της υλοποίησης σε συστήματα περιορισμένων πόρων. Ορίζεται ως:

$$f(x) = x \cdot \frac{\text{ReLU6}(x + 3)}{6} \quad (3.6)$$

όπου η $\text{ReLU6}(x + 3)$ περιορίζει τη μετατοπισμένη είσοδο στο διάστημα $[0, 6]$, ενώ η διαίρεση με 6 εξασφαλίζει κανονικοποιημένη έξοδο εντός του $[0, x]$.

- **Hard Sigmoid:** Η συνάρτηση Hard Sigmoid [125] αποτελεί γραμμική προσεγγιστική εκδοχή της παραδοσιακής Sigmoid, με στόχο τη μείωση της υπολογιστικής πολυπλοκότητας σε προβλήματα δυαδικής ταξινόμησης. Ορίζεται ως:

$$f(x) = \max(0, \min(1, \frac{x+1}{2})) \quad (3.7)$$

Παρουσιάζει παρόμοια συμπεριφορά με την παραδοσιακή Sigmoid, αποφεύγοντας τους εκθετικούς υπολογισμούς, γεγονός που την καθιστά κατάλληλη για υλοποιήσεις σε συστήματα περιορισμένων πόρων.

Η επιλογή της κατάλληλης συνάρτησης ενεργοποίησης εξαρτάται από την εκάστοτε εφαρμογή—στόχο. Σε HCS με περιορισμούς σε κατανάλωση ενέργειας και υπολογιστικούς πόρους, όπως τα FPGAs ή τα SoMs, προτιμώνται συναρτήσεις με ελάχιστη αριθμητική πολυπλοκότητα. Ιδιαίτερα στο στάδιο της εξαγωγής συμπερασμάτων, οι απλοποιημένες αυτές συναρτήσεις επιτρέπουν την αποδοτική υλοποίηση χωρίς σημαντική απώλεια στην απόδοση του CNN. Ο Πίνακας 3.2 συνοψίζει τις βασικές συναρτήσεις ενεργοποίησης χαμηλής πολυπλοκότητας και φιλικών προς το υλικό.

3.2.2 Τεχνικές Βελτιστοποίησης

3.2.2.1 Κβαντοποίηση

Η κβαντοποίηση (Quantization) [126] αποτελεί μία τεχνική η οποία στοχεύει στη μείωση της αριθμητικής ακρίβειας των παραμέτρων και των ενεργοποιήσεων ενός CNN, από αναπαράσταση κινητής υποδιαστολής 32-bit (FP32) σε αναπαραστάσεις σταθερής υποδιαστολής χαμηλότερης ακρίβειας, όπως 16-bit (FP16), 8-bit (INT8). Όπως φαίνεται και στον Πίνακα 3.3, ο κύριος σκοπός της διαδικασίας αυτής είναι η μείωση των απαιτήσεων αποθήκευσης και των υπολογιστικών πόρων που καταναλώνει ένα δίκτυο, διασφαλίζοντας παράλληλα ότι η ακρίβεια της εξαγωγής συμπερασμάτων επηρεάζεται στο ελάχιστο δυνατό βαθμό.

Υπάρχουν δύο κύριες προσεγγίσεις κβαντοποίησης των CNNs:

- **Κβαντοποίηση μετά την εκπαίδευση (Post-Training Quantization, PTQ):** Κατά τη συγκεκριμένη μέθοδο, ένα ήδη εκπαιδευμένο δίκτυο υψηλής ακρίβειας μετατρέπεται απευθείας σε δίκτυο χαμηλότερης αριθμητικής ακρίβειας, χωρίς

Πίνακας 3.2: Συγκριτική ανάλυση χαμηλής πολυπλοκότητας συναρτήσεων ενεργοποίησης φιλικών προς το υλικό.

Συνάρτηση	Πλεονεκτήματα	Μειονεκτήματα
ReLU	Εξαιρετικά απλή υλοποίηση. Αποτρέπει το φαινόμενο της εξαφάνισης του βαθμωτού σφάλματος.	Κίνδυνος ανενεργών νευρώνων. Μη διαφορίσιμη στο μηδέν.
Leaky ReLU	Διατηρεί ενεργά τα νευρωνικά στοιχεία. Εύκολη υλοποίηση με κατάλληλη επιλογή α .	Εισάγει μικρό επιπλέον κόστος. Η απόδοση εξαρτάται από την τιμή της παραμέτρου α .
Hard Tanh	Ελάχιστη πολυπλοκότητα λόγω μόνο συγκρίσεων.	Μειωμένη ακρίβεια σε μη γραμμικές σχέσεις.
PReLU	Δυναμική προσαρμογή αρνητικής κλίσης. Μειώνει την πιθανότητα ανενεργών νευρώνων.	Αυξάνει την πολυπλοκότητα σε σύγκριση με τη ReLU.
ReLU6	Περιορισμός μεγάλων ενεργοποιήσεων. Βελτιωμένη αριθμητική σταθερότητα.	Δεν είναι διαφορίσιμη στα σημεία αποκοπής $x = 0$ και $x = 6$.
Hard Swish	Ομαλότερη συμπεριφορά από τη ReLU. Χαμηλότερη πολυπλοκότητα από την Swish.	Αυξημένη πολυπλοκότητα έναντι της ReLU.
Hard Sigmoid	Προσεγγίζει καλά την κλασική Sigmoid. Πολύ απλή υλοποίηση.	Περιορισμένη ικανότητα διάκρισης μεταξύ κλάσεων γύρω από το κατώφλι απόφασης.

να απαιτείται περαιτέρω διαδικασία εκπαίδευσης. Η διαδικασία περιλαμβάνει ανάλυση του εύρους των βαρών και ενεργοποιήσεων και τον υπολογισμό των παραμέτρων κβαντοποίησης, όπως τα βήματα κβαντοποίησης και τα όρια εύρους τιμών. Αν και αυτή η μέθοδος είναι ταχεία και εύκολη στην εφαρμογή της, ενδέχεται να προκαλέσει απώλεια ακρίβειας ειδικά όταν χρησιμοποιούνται πολύ χαμηλές αναπαραστάσεις ακρίβειας.

- Κβαντοποίηση κατά τη διάρκεια της εκπαίδευσης (Quantization—Aware Training, QAT): Στην περίπτωση αυτή, η μειωμένη αριθμητική ακρίβεια ενσωματώνεται εξ αρχής κατά τη φάση εκπαίδευσης του μοντέλου. Συγκεκριμένα, οι υπολογισμοί κατά την εμπρόσθια διάδοση (forward

Πίνακας 3.3: Σύγκριση αριθμητικών προτύπων FP32, FP16 και INT8 σε εφαρμογές CNN.

Πρότυπο	Τύπος & Εύρος τιμών	Βασικά χαρακτηριστικά
FP32	Floating Point (32 bits) Εύρος: $\sim \pm 3,4 \times 10^{38}$	Υψηλή αριθμητική ακρίβεια. Αυξημένες απαιτήσεις σε μνήμη και υπολογιστική ισχύ. Τυπικό πρότυπο για εκπαίδευση και αξιολόγηση μοντέλων.
FP16	Half Precision Float (16 bits) Εύρος: $\sim \pm 6,5 \times 10^4$	Μέτρια ακρίβεια. Μειωμένες απαιτήσεις σε μνήμη και ενέργεια. Επιταχύνει την εξαγωγή συμπερασμάτων.
INT8	Integer (8 bits) Εύρος: -128 έως 127	Χαμηλή ακρίβεια. Χαμηλές απαιτήσεις σε μνήμη και υψηλή ταχύτητα υπολογισμών. Ιδανικό για κβαντοποίηση και ανάπτυξη μοντέλων σε συσκευές πραγματικού χρόνου.
INT4	Integer (4 bits) Εύρος: -8 έως 7	Πολύ χαμηλή αριθμητική ακρίβεια. Ελάχιστες απαιτήσεις σε μνήμη και μέγιστη ταχύτητα υπολογισμών. Κατάλληλο για εξαιρετικά περιορισμένο υλικό, αλλά απαιτεί προσεκτική κβαντοποίηση.

pass), δηλαδή την παραγωγή των ενεργοποιήσεων σε κάθε επίπεδο του δικτύου, προσομοιώνονται με χαμηλή ακρίβεια. Αντιθέτως, η οπίσθια διάδοση (backward pass), η οποία περιλαμβάνει τον υπολογισμό των βαθμίδων της συνάρτησης απώλειας και την ενημέρωση των παραμέτρων μέσω οπισθοδιάδοσης (backpropagation), εκτελείται με υψηλότερη αριθμητική ακρίβεια. Η προσέγγιση αυτή επιτρέπει στο δίκτυο να προσαρμοστεί εξ αρχής στις ιδιαιτερότητες της χαμηλής ακρίβειας, επιτυγχάνοντας απόδοση συχνά συγκρίσιμη με εκείνη της πλήρους αριθμητικής ακρίβειας.

Η υιοθέτηση της κβαντοποίησης προσφέρει σημαντικά πλεονεκτήματα σε HCS με περιορισμούς σε υπολογιστικούς πόρους:

- Μείωση μεγέθους μοντέλου: Η αποθήκευση των παραμέτρων σε μορφές χαμηλότερης ακρίβειας περιορίζει σημαντικά τις απαιτήσεις αποθήκευσης,

γεγονός που είναι κρίσιμο για εφαρμογές με περιορισμούς μνήμης, όπως τα ενσωματωμένα συστήματα.

- Ταχύτερη εκτέλεση υπολογισμών: Οι αριθμητικές πράξεις χαμηλής ακρίβειας εκτελούνται με μικρότερη χρονική καθυστέρηση, καθώς οι αλγόριθμοι και οι αντίστοιχες μονάδες υλικού που τις υλοποιούν είναι απλούστερες και συνεπώς αποτελεσματικότερες όσον αφορά το χρόνο εκτέλεσης.
- Χαμηλότερη κατανάλωση ενέργειας: Λόγω της μειωμένης υπολογιστικής πολυπλοκότητας των αριθμητικών πράξεων, τα μοντέλα που χρησιμοποιούν κβαντοποίηση καταναλώνουν λιγότερη ενέργεια, γεγονός ιδιαίτερα σημαντικό για συσκευές που λειτουργούν με ενεργειακούς περιορισμούς.
- Αύξηση της ταχύτητας εξαγωγής συμπερασμάτων: Η χρήση χαμηλότερης αριθμητικής ακρίβειας μειώνει αισθητά τον υπολογιστικό φόρτο κατά την εκτέλεση της διαδικασίας εξαγωγής συμπερασμάτων του μοντέλου, επιτρέποντας έτσι την ταχύτερη παραγωγή προβλέψεων. Αυτό είναι ιδιαίτερα καθοριστικό σε εφαρμογές πραγματικού χρόνου όπου η ελαχιστοποίηση της καθυστέρησης συνιστά βασική απαίτηση.

3.2.2.2 Κλάδεμα

Το κλάδεμα (Pruning) [127] αποτελεί βασική τεχνική βελτιστοποίησης των CNNs, αποσκοπώντας στη μείωση της λειτουργικής και υπολογιστικής πολυπλοκότητας των μοντέλων. Μέσω της αφαίρεσης πλεοναζόντων στοιχείων της αρχιτεκτονικής, το κλάδεμα ενισχύει την αποδοτική υλοποίηση των CNNs, ιδιαίτερα σε συσκευές με περιορισμένους υπολογιστικούς πόρους.

Το κλάδεμα διακρίνεται στις εξής κατηγορίες:

- Μη δομημένο κλάδεμα (Unstructured Pruning): Αφορά την απομάκρυνση μεμονωμένων βαρών ανεξαρτήτως της θέσης τους στο δίκτυο. Η επιλογή των βαρών προς απομάκρυνση βασίζεται συνήθως στο μέτρο της απόλυτης τιμής τους, με τις μικρές τιμές να θεωρούνται λιγότερο σημαντικές. Αν και προσφέρει υψηλά ποσοστά συμπίεσης, το μη δομημένο κλάδεμα μπορεί να οδηγήσει σε αραιές δομές, οι οποίες δεν υποστηρίζονται αποδοτικά από το υλικό.

- Δομημένο κλάδεμα (Structured Pruning): Σε αυτή την περίπτωση, απομακρύνονται ολόκληρες δομές, όπως φίλτρα, κανάλια ή ακόμα και ολόκληρα επίπεδα. Η τεχνική αυτή οδηγεί σε πιο συμπαγή και αποδοτικά μοντέλα, καθώς η νέα δομή του δικτύου μπορεί να εκτελείται πιο αποτελεσματικά από το υλικό.

Η υιοθέτηση τεχνικών κλαδέματος προσφέρει σημαντικά πλεονεκτήματα σε HCS με περιορισμούς σε υπολογιστικούς πόρους:

- Μείωση μεγέθους μοντέλου: Το κλάδεμα οδηγεί σε μείωση του αριθμού των παραμέτρων ενός CNN, καθώς αφαιρούνται βάρη ή ολόκληρες δομές που θεωρούνται μη κρίσιμες για τη συνολική απόδοση του μοντέλου. Η συμπίεση που επιτυγχάνεται μπορεί να μειώσει δραστικά τις απαιτήσεις σε αποθηκευτικό χώρο (π.χ. από δεκάδες MB σε λίγα MB), γεγονός που καθιστά εφικτή την ανάπτυξη του μοντέλου σε περιβάλλοντα με περιορισμένη μνήμη. Επιπλέον, η μικρότερη αποτύπωση του μοντέλου συμβάλλει στη μείωση του χρόνου φόρτωσης και μεταφοράς μεταξύ επιπέδων μνήμης ή ακόμα και μεταξύ συσκευών.
- Μείωση του υπολογιστικού κόστους: Το κλάδεμα οδηγεί σε μείωση των FLOPs, δηλαδή του αριθμού των υπολογιστικών πράξεων κινητής υποδιαστολής που απαιτούνται για την επεξεργασία μιας εισόδου. Ειδικά στις τεχνικές δομημένου κλαδέματος, η απομάκρυνση ολόκληρων φίλτρων ή καναλιών μεταφράζεται σε μικρότερο αριθμό πράξεων ανά επίπεδο, με άμεση μείωση της υπολογιστικής επιβάρυνσης. Αυτή η μείωση είναι κρίσιμη για την επιτάχυνση της εξαγωγής συμπερασμάτων, την αποδοτικότερη χρήση των επιταχυντών και τη βελτιστοποίηση της κατανάλωσης ενέργειας.
- Αύξηση της ταχύτητας εξαγωγής συμπερασμάτων: Η ελαχιστοποίηση του μεγέθους του μοντέλου και του αριθμού των υπολογιστικών πράξεων έχει ως άμεσο αποτέλεσμα την επιτάχυνση της φάσης της εξαγωγής συμπερασμάτων. Αυτό επιτρέπει την ταχύτερη παραγωγή προβλέψεων, κάτι το οποίο είναι καθοριστικό για εφαρμογές πραγματικού χρόνου όπως η ιατρική διάγνωση σε point-of-care συσκευές.

3.2.2.3 Διαχείριση Ανισορροπίας Κλάσεων

Σε πολλές εφαρμογές της CV, τα σύνολα δεδομένων εκπαίδευσης παρουσιάζουν ανισορροπία κλάσεων [128], όπου ορισμένες κλάσεις έχουν σημαντικά λιγότερα δείγματα από άλλες. Αυτή η ανισορροπία οδηγεί σε μοντέλα CNNs που είναι μεροληπτικά προς την πλειοψηφούσα κλάση και αποδίδουν ανεπαρκώς στην αναγνώριση της μειοψηφούσας κλάσης.

Οι προσεγγίσεις για την ανισορροπία κλάσεων διακρίνονται σε τρεις κατηγορίες:

- **Επιπέδου δεδομένων (Data—Level):** Αυτές οι προσεγγίσεις επιχειρούν να εξισορροπήσουν τον αριθμό δειγμάτων μεταξύ των πλειοψηφικών και μειοψηφικών κλάσεων πριν από την εκπαίδευση του μοντέλου. Οι μέθοδοι αυτές δεν επεμβαίνουν στον αλγόριθμο μάθησης, αλλά τροποποιούν το σύνολο εκπαίδευσης, ώστε να μειωθεί η μεροληψία που προκαλείται από την ανομοιόμορφη κατανομή. Μια βασική στρατηγική είναι η υπερδειγματοληψία (oversampling) της μειοψηφικής κλάσης. Τέτοιες τεχνικές είναι η SMOTE [129] και η ADASYN [130], οι οποίες δημιουργούν συνθετικά δείγματα παράγοντας νέα σημεία χαρακτηριστικών ανάμεσα σε υπάρχοντα δείγματα της μειοψηφίας. Αντίθετα, η υποδειγματοληψία (undersampling) μειώνει το πλήθος δειγμάτων της πλειοψηφίας, ώστε να επέλθει ισορροπία. Η ενίσχυση δεδομένων (data augmentation) αποτελεί ακόμη μία τεχνική και περιλαμβάνει μετασχηματισμούς (π.χ. περιστροφές, μετατοπίσεις, αλλαγές φωτεινότητας), οι οποίοι αυξάνουν την ποικιλία των δειγμάτων. Επιπλέον, η χρήση γεννητικών αντιπαραθετικών δικτύων (GANs) [131] επιτρέπει την παραγωγή συνθετικών εικόνων υψηλής ποιότητας για τη μειοψηφική κλάση, προσφέροντας ένα πρόσθετο εργαλείο ενίσχυσης της αναπαράστασης της κλάσης στόχου.
- **Επιπέδου αλγορίθμου (Algorithm—Level):** Αυτές οι προσεγγίσεις επιδιώκουν να ενισχύσουν τη μάθηση της μειοψηφικής κλάσης μέσω της τροποποίησης της συνάρτησης απώλειας, χωρίς να επηρεάζεται το σύνολο δεδομένων. Η πιο διαδεδομένη μέθοδος είναι η χρήση σταθμισμένων συναρτήσεων απώλειας, όπως η σταθμισμένη απώλεια διασταυρούμενης εντροπίας (Weighted Cross—Entropy Loss), όπου τα σφάλματα της μειοψηφικής κλάσης λαμβάνουν υψηλότερη βαρύτητα. Ιδιαίτερης σημασίας είναι η εστιασμένη απώλεια (Focal Loss), η οποία δίνει έμφαση στα δύσκολα προς ταξινόμηση δείγματα,

μειώνοντας τη συνεισφορά των εύκολων περιπτώσεων και ενισχύοντας τη μάθηση του μοντέλου σε περιοχές του χώρου χαρακτηριστικών με μειωμένη εκφραστικότητα. Άλλες τεχνικές περιλαμβάνουν τη μάθηση με κόστος (cost-sensitive learning), όπου ορίζεται πίνακας κόστους για τους διαφορετικούς τύπους σφαλμάτων, καθώς και τη χρήση δυναμικών προσαρμογών στα βάρη ανά δείγμα ή κλάση (sample weights / class weights).

- Υβριδικές (Hybrid): Αυτές οι προσεγγίσεις συνδυάζουν τεχνικές τόσο από το επίπεδο των δεδομένων όσο και από το επίπεδο του αλγορίθμου, επιδιώκοντας να αξιοποιήσουν τα πλεονεκτήματα και των δύο στρατηγικών. Αντί να βασίζονται σε μία μόνο μέθοδο, ενσωματώνουν στοιχεία από διαφορετικές προσεγγίσεις, με στόχο την πιο αποτελεσματική αντιμετώπιση της ανισορροπίας κλάσεων. Μία σημαντική υποκατηγορία αποτελούν οι συναθροιστικές μέθοδοι (ensemble methods), οι οποίες βασίζονται στον συνδυασμό πολλαπλών ταξινομητών ή υπομοντέλων, με στόχο τη μείωση της διακύμανσης, τη βελτίωση της γενικευσιμότητας και την ενίσχυση της σταθερότητας του τελικού μοντέλου.

Στα HCS, όπου οι υπολογιστικοί πόροι είναι περιορισμένοι, η αντιμετώπιση της ανισορροπίας κλάσεων πρέπει να ενσωματώνεται εξ αρχής στον αρχιτεκτονικό σχεδιασμό των CNNs. Η επιλογή τεχνικών διαχείρισης της ανισορροπίας οφείλει να ευθυγραμμίζεται με μετρικές όπως η κατανάλωση μνήμης και το συνολικό ενεργειακό αποτύπωμα. Ως εκ τούτου, η ενσωμάτωση μηχανισμών εξισορρόπησης των κλάσεων αποτελεί αρχιτεκτονική επιλογή, ώστε το τελικό μοντέλο να παραμένει υπολογιστικά χαμηλής πολυπλοκότητας, χωρίς να υποβαθμίζεται η ικανότητά του να αναγνωρίζει τις μειοψηφικές κλάσεις. Ο Πίνακας 3.4 παρουσιάζει συνοπτικά την κατηγοριοποίηση των τεχνικών διαχείρισης ανισορροπίας κλάσεων.

3.2.3 Συμβιβασμοί Επιδόσεων

Η επίδοση ενός CNN δεν αποτιμάται μονοδιάστατα, αντίθετα, συχνά υπάρχει ένας συμβιβασμός μεταξύ διαφορετικών δεικτών απόδοσης, σε HCS, όπου οι υπολογιστικοί πόροι είναι περιορισμένοι. Οι πιο χαρακτηριστικοί είναι ο συμβιβασμός μεταξύ ακρίβειας και καθυστέρησης, καθώς και μεταξύ του μεγέθους μοντέλου και του ρυθμού διεκπεραίωσης. Η εξισορρόπηση αυτών των αντιφάσεων είναι καθοριστική για την αποδοτική υλοποίηση CNNs σε πραγματικά περιβάλλοντα [132, 133].

Πίνακας 3.4: Κατηγοριοποίηση τεχνικών διαχείρισης ανισορροπίας κλάσεων.

Κατηγορία	Πλεονεκτήματα
Data—Level	Ενισχύουν τις μειοψηφικές κλάσεις χωρίς αλλαγές στον αλγόριθμο. Αυξάνουν την ποικιλία και την εκπροσώπηση των δεδομένων. Κατάλληλες όταν υπάρχει αφθονία υπολογιστικών πόρων.
Algorithm—Level	Ενισχύουν τη μάθηση της μειοψηφικής κλάσης μέσω τροποποίησης της συνάρτησης απώλειας. Δεν απαιτούν τροποποίηση των δεδομένων. Είναι πιο αποδοτικές σε HCS περιβάλλοντα.
Hybrid	Ενσωματώνουν πλεονεκτήματα από δεδομένα και αλγόριθμους. Αυξάνουν τη γενικευσιμότητα και μειώνουν την αστάθεια. Κατάλληλες για σύνθετα ή πολύ ανισόρροπα σύνολα δεδομένων.

- Ακρίβεια έναντι καθυστέρησης (Accuracy vs Latency): Δίκτυα με υψηλότερη ικανότητα (π.χ. βαθύτερα ή με περισσότερες παραμέτρους) τείνουν να αποδίδουν καλύτερα σε όρους ακρίβειας ταξινόμησης, όμως συχνά απαιτούν περισσότερους υπολογισμούς και συνεπώς έχουν μεγαλύτερη λανθάνουσα καθυστέρηση στην φάση της εξαγωγής συμπερασμάτων. Αντίστροφα, τα μοντέλα που έχουν σχεδιαστεί για να είναι γρήγορα και να έχουν χαμηλή καθυστέρηση είναι συνήθως απλούστερα, με λιγότερες παραμέτρους, και μπορεί να θυσιάσουν κάποια ακρίβεια προκειμένου να επιτύχουν ταχύτερους χρόνους απόκρισης. Η σχέση μεταξύ ακρίβειας και καθυστέρησης είναι κρίσιμη για πολλές εφαρμογές, ειδικά για εκείνες που απαιτούν επεξεργασία σε πραγματικό χρόνο, όπως η αυτόνομη οδήγηση, η ανίχνευση αντικειμένων σε βίντεο και η ιατρική διάγνωση. Σε αυτές τις περιπτώσεις, μια χαμηλή καθυστέρηση στην επεξεργασία μπορεί να είναι πιο σημαντική από μια οριακή βελτίωση στην ακρίβεια, καθώς η έγκαιρη απόκριση μπορεί να είναι ζωτικής σημασίας.
- Μέγεθος μοντέλου έναντι ρυθμαπόδοσης (Model Size vs Throughput): Ένας εξίσου σημαντικός συμβιβασμός αφορά τη σχέση μεταξύ του μεγέθους του μοντέλου —μετρούμενου σε αριθμό παραμέτρων ή μέγεθος σε bytes— και της ρυθμαπόδοσης, δηλαδή του αριθμού των δειγμάτων που μπορούν να επεξεργαστούν ανά μονάδα χρόνου. Τα μεγαλύτερα μοντέλα τείνουν να απαιτούν περισσότερους υπολογιστικούς και μνημονικούς πόρους για αποθήκευση και επεξεργασία, περιορίζοντας την ταχύτητα επεξεργασίας κατά παρτίδες. Αυτό επηρεάζει αρνητικά τη ρυθμαπόδοση, ειδικά όταν η υλοποίηση πραγματοποιείται σε συστήματα με περιορισμένη επεξεργαστική

ισχύ. Αντίθετα, η μείωση του μεγέθους μέσω χαμηλής πολυπλοκότητας δομοστοιχείων και συμπίεσης μοντέλων μπορεί να βελτιώσει τη ρυθμαπόδοση, αλλά ενδέχεται να περιορίσει την εκφραστική ικανότητα του δικτύου και κατ' επέκταση την τελική ακρίβεια εξαγωγής συμπερασμάτων. Η επιλογή του κατάλληλου μεγέθους μοντέλου συναρτάται άμεσα με το διαθέσιμο υλικό και τον απαιτούμενο όγκο δεδομένων προς επεξεργασία σε εφαρμογές υψηλής ταχύτητας.

3.2.4 Ενεργειακός Σχεδιασμός

Η ανάπτυξη και υλοποίηση CNNs σε ενεργειακά περιορισμένα υπολογιστικά περιβάλλοντα καθιστά επιτακτική την ανάγκη για ενεργειακά αποδοτικό σχεδιασμό. Καθοριστικές παράμετροι στο πλαίσιο αυτό είναι η ενέργεια που απαιτείται ανά εξαγωγή συμπεράσματος και οι περιορισμοί ισχύος του εκάστοτε συστήματος—στόχου [134]. Οι περιορισμοί αυτοί επηρεάζουν άμεσα την επιλογή τόσο της αρχιτεκτονικής του μοντέλου όσο και των τεχνικών βελτιστοποίησης που εφαρμόζονται.

- **Ενέργεια ανά εξαγωγή συμπεράσματος (Energy per Inference):** Αναφέρεται στην ποσότητα ενέργειας που καταναλώνεται κατά την επεξεργασία μιας μεμονωμένης εισόδου από το δίκτυο. Ο δείκτης αυτός εξαρτάται κυρίως από το πλήθος των υπολογιστικών πράξεων, τη συχνότητα προσπέλασης μνήμης και την αποδοτικότητα της υλοποίησης σε επίπεδο υλικού. Δίκτυα με βελτιστοποιημένες προσπελάσεις και μειωμένο υπολογιστικό κόστος τείνουν να παρουσιάζουν αυξημένη ενεργειακή απόδοση. Τεχνικές όπως η κβαντοποίηση και το κλάδεμα μειώνουν σημαντικά την κατανάλωση πόρων, περιορίζοντας τόσο τις αριθμητικές πράξεις όσο και τη μετακίνηση δεδομένων από και προς τη μνήμη. Σε συστημικό επίπεδο, η απόδοση εκφράζεται συχνά με δείκτη που συνδυάζει την καθυστέρηση και την κατανάλωση ισχύος, επιτρέποντας τη συγκριτική αξιολόγηση διαφορετικών υλοποιήσεων (π.χ. CPU, GPU, επιταχυντές).
- **Περιορισμοί ισχύος (Power Constraints):** Οι περιορισμοί κατανάλωσης επηρεάζουν άμεσα τη θερμική συμπεριφορά και τη σταθερότητα του συστήματος. Ανεπαρκής διαχείριση της παραγόμενης θερμότητας μπορεί να επιφέρει μείωση απόδοσης ή και αποτυχία λειτουργίας. Κατά τον σχεδιασμό,

δίκτυα υψηλής πολυπλοκότητας που απαιτούν μεγάλο υπολογιστικό φορτίο ενδέχεται να προκαλέσουν υπερθέρμανση ή ταχεία εξάντληση της διαθέσιμης ενέργειας, ιδίως σε φορητές ή ενσωματωμένες συσκευές. Συνεπώς, η αρχιτεκτονική του μοντέλου οφείλει να ευθυγραμμίζεται με τα διαθέσιμα ενεργειακά όρια, αξιοποιώντας πλήρως τον ενεργειακό προϋπολογισμό χωρίς υπερβάσεις, αλλά και χωρίς να αφήνει ανεκμετάλλευτους υπολογιστικούς πόρους που θα μπορούσαν να συνεισφέρουν σε υψηλότερη ακρίβεια εξαγωγής συμπεράσματος.

3.3 Μετρικές Ποιότητας

Οι μετρικές ποιότητας (Quality Metrics) [135] της μεθοδολογίας μας εστιάζουν στην αποτίμηση της ικανότητας ενός CNN να εκτελεί με ακρίβεια και συνέπεια τη προβλεπτική του λειτουργία, ανεξαρτήτως του υπολογιστικού ή ενεργειακού κόστους που συνεπάγεται η φάση εκπαίδευσης ή η παραγωγική του χρήση. Οι μετρικές αυτές αποτελούν βασικό εργαλείο για την αξιολόγηση της αποτελεσματικότητας ενός μοντέλου από λειτουργική άποψη, με στόχο την αποτύπωση της συσχέτισης μεταξύ της εξαγωγής συμπεράσματος του μοντέλου και της πραγματικής ετικέτας. Οι πλέον διαδεδομένες μετρικές περιλαμβάνουν:

- **Accuracy:** Η συνολική ακρίβεια υπολογίζεται ως ο λόγος των ορθώς ταξινομημένων δειγμάτων προς το σύνολο των δειγμάτων. Αποτελεί μια γενική ένδειξη της επίδοσης του μοντέλου και είναι κατάλληλη κυρίως σε περιπτώσεις όπου οι κατηγορίες είναι ισορροπημένες. Για παράδειγμα, σε ένα σύνολο δεδομένων όπου οι θετικές και αρνητικές κλάσεις εμφανίζονται σε παρόμοια ποσοστά, η ακρίβεια μπορεί να θεωρηθεί αντιπροσωπευτική της συνολικής συμπεριφοράς του ταξινομητή. Ορίζεται ως:

$$\text{Συνολική ακρίβεια} = \frac{\text{Αριθμός σωστών προβλέψεων}}{\text{Συνολικός αριθμός δειγμάτων}} \quad (3.8)$$

- **Precision:** Η ακρίβεια θετικών προβλέψεων μετρά το ποσοστό των αληθινών θετικών επί των συνόλου των προβλέψεων που χαρακτηρίστηκαν ως θετικές από το μοντέλο. Είναι καθοριστική σε περιπτώσεις όπου τα ψευδώς θετικά έχουν σημαντικές επιπτώσεις. Για παράδειγμα, σε ένα σύστημα διάγνωσης ασθενειών, η λανθασμένη διάγνωση υγιών ως ασθενών μπορεί να οδηγήσει σε

άσκοπες ιατρικές παρεμβάσεις, συνεπώς η μεγιστοποίηση αυτής της μετρικής είναι επιθυμητή. Ορίζεται ως:

$$\text{Ακρίβεια θετικών προβλέψεων} = \frac{\text{Αληθώς Θετικά}}{\text{Αληθώς Θετικά} + \text{Ψευδώς Θετικά}} \quad (3.9)$$

- Recall: Η ανάκληση, γνωστή και ως ευαισθησία, μετρά την ικανότητα του μοντέλου να εντοπίζει σωστά όλες τις θετικές περιπτώσεις. Είναι ιδιαίτερα σημαντική σε εφαρμογές όπου οι ψευδώς αρνητικές προβλέψεις είναι ανεπιθύμητες. Ενδεικτικό παράδειγμα αποτελεί η ανίχνευση καρκινικών όγκων, όπου η αποτυχία εντοπισμού ενός θετικού περιστατικού μπορεί να έχει μοιραίες συνέπειες. Ορίζεται ως:

$$\text{Ανάκληση} = \frac{\text{Αληθώς Θετικά}}{\text{Αληθώς Θετικά} + \text{Ψευδώς Αρνητικά}} \quad (3.10)$$

- F1—score: Ο αρμονικός μέσος των precision και recall, ο οποίος προσφέρει εξισορροπημένη εικόνα της επίδοσης του μοντέλου και προτιμάται σε περιπτώσεις ανισορροπίας κλάσεων ή όταν υπάρχει ανάγκη εξισορρόπησης μεταξύ των ψευδώς θετικών και ψευδώς αρνητικών. Παρέχει μία ενιαία μετρική που λαμβάνει υπόψη και τα δύο σφάλματα και χρησιμοποιείται συχνά για τη συνολική εκτίμηση μοντέλων σε περιβάλλοντα χαμηλής πυκνότητας θετικών περιπτώσεων. Ορίζεται ως:

$$\text{F1—score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.11)$$

- AUC: Η περιοχή κάτω από την καμπύλη ROC (Area Under the Curve) καταγράφει την απόδοση του μοντέλου για όλα τα πιθανά κατώφλια απόφασης, ενσωματώνοντας την ευαισθησία και την ειδικότητα. Η AUC θεωρείται ιδιαίτερα χρήσιμη για τη σύγκριση μοντέλων ανεξαρτήτως της τελικής απόφασης κατωφλίου, ιδίως όταν προτεραιοποιείται η διαχωριστική ικανότητα του μοντέλου μεταξύ των κλάσεων.

Οι προαναφερθείσες μετρικές παρέχουν ποσοτική τεκμηρίωση για την αξιολόγηση της συμπεριφοράς ενός ταξινομητή, καθώς αποτυπώνουν με ακρίβεια την ικανότητά του να διακρίνει και να κατηγοριοποιεί τα δεδομένα εισόδου σε σχέση με την πραγματική τους ετικέτα. Αποτελούν το πρώτο επίπεδο ελέγχου της

προβλεπτικής επάρκειας του μοντέλου, πριν εξεταστούν παράμετροι σχετιζόμενες με την υπολογιστική του αποδοτικότητα, την ενεργειακή κατανάλωση, και εν τέλει, τη λειτουργική του καταλληλότητα στο εκάστοτε υπολογιστικό περιβάλλον στόχου. Ο Πίνακας 3.5 παρουσιάζει συνοπτικά τις βασικές μετρικές που χρησιμοποιούνται για την αξιολόγηση των CNNs.

Πίνακας 3.5: Σύνοψη βασικών μετρικών ποιότητας CNN.

Μετρική	Ορισμός	Σημασία και χρήση
Accuracy	Λόγος ορθά ταξινομημένων δειγμάτων προς το σύνολο των δειγμάτων.	Γενική ένδειξη της επίδοσης του μοντέλου. Κατάλληλη όταν οι κλάσεις είναι ισορροπημένες.
Precision	Λόγος των αληθινών θετικών προς το σύνολο των προβλέψεων που χαρακτηρίστηκαν ως θετικές.	Χρήσιμη όταν τα ψευδώς θετικά έχουν υψηλό κόστος (π.χ. ιατρική διάγνωση).
Recall	Λόγος των αληθινών θετικών προς το σύνολο των πραγματικά θετικών περιπτώσεων.	Κρίσιμη όταν τα ψευδώς αρνητικά είναι ανεπιθύμητα (π.χ. εντοπισμός καρκίνου).
F1—score	Αρμονικός μέσος του precision και του recall.	Ισορροπημένη εκτίμηση επίδοσης. Χρήσιμη σε περιβάλλοντα με ανισορροπία κλάσεων.
AUC	Εμβαδόν κάτω από την καμπύλη ROC. Συνδυάζει ευαισθησία και ειδικότητα για όλα τα κατώφλια απόφασης.	Κατάλληλη για συγκριτική αξιολόγηση μοντέλων, ανεξαρτήτως του επιλεγμένου κατωφλίου απόφασης.

3.4 Μετρικές Υπολογιστικού Αποτυπώματος

Οι μετρικές υπολογιστικού αποτυπώματος (Footprint Metrics) [136] της μεθοδολογίας μας αφορούν ποσοτικά μεγέθη που περιγράφουν το υπολογιστικό, ενεργειακό και χρονικό κόστος τόσο στη φάση εκπαίδευσης όσο και στη φάση εξαγωγής συμπερασμάτων ενός CNN. Οι μετρικές αυτές αποκτούν ιδιαίτερη σημασία όταν εξετάζεται η υλοποίηση σε εξειδικευμένες αρχιτεκτονικές υλικού, όπως τα FPGAs και τα SoMs. Σε τέτοιου είδους HCS, οι διαθέσιμοι πόροι είναι περιορισμένοι. Τα FPGAs, για παράδειγμα, προσφέρουν ευελιξία παραλληλισμού,

αλλά περιορίζονται από το εύρος μνήμης και τον ρυθμό επαναδιαμόρφωσης, καθιστώντας κρίσιμο το μέγεθος του μοντέλου και τον αριθμό των πράξεων ανά κύκλο. Τα SoMs, από την άλλη, ενσωματώνουν επιταχυντές, αλλά υπόκεινται σε ενεργειακούς και θερμικούς περιορισμούς, γεγονός που επιβάλλει τη χρήση CNNs χαμηλής πολυπλοκότητας.

Ως εκ τούτου, η επιλογή των CNNs για ανάπτυξη σε τέτοιες συσκευές πρέπει να γίνεται με γνώμονα τη συμβατότητα με τους περιορισμούς υλικού, καθιστώντας τις μετρικές υπολογιστικού αποτυπώματος αναπόσπαστο κριτήριο τεχνικής καταλληλότητας. Η συστηματική αξιολόγηση αυτών των μεγεθών είναι κρίσιμη για την επίτευξη ισορροπίας μεταξύ αποδοτικότητας, ακρίβειας εξαγωγής συμπεράσματος και λειτουργικής βιωσιμότητας στο εκάστοτε υπολογιστικό περιβάλλον στόχο. Οι μετρικές υπολογιστικού αποτυπώματος διακρίνονται σε τρεις υποκατηγορίες, ανάλογα με τη φάση του κύκλου ζωής του CNN στην οποία εφαρμόζονται: (α) μετρικές που αφορούν μόνο τη φάση εκπαίδευσης, (β) μετρικές που αφορούν μόνο τη φάση εξαγωγής συμπερασμάτων, και (γ) μετρικές που αφορούν τη φάση εκπαίδευσης και εξαγωγής συμπερασμάτων.

3.4.1 Φάση Εκπαίδευσης

Οι μετρικές που εμπίπτουν στην παρούσα κατηγορία της μεθοδολογίας μας σχετίζονται αποκλειστικά με τη φάση εκπαίδευσης του μοντέλου και αποτυπώνουν τόσο τον χρονικό όσο και το υπολογιστικό φορτίο που απαιτείται για την επίτευξη της επιθυμητής απόδοσης [137]. Αξιολογούν, δηλαδή, τη διαδικασία μάθησης καθαυτή, ανεξαρτήτως της μεταγενέστερης φάσης χρήσης του μοντέλου. Η ανάλυσή τους είναι κρίσιμη, αφενός για τη βελτιστοποίηση των αλγοριθμικών παραμέτρων (όπως ο ρυθμός μάθησης ή η στρατηγική ενημέρωσης βαρών), και αφετέρου για την εκτίμηση της εφικτότητας και της αποδοτικότητας της εκπαίδευσης σε περιβάλλοντα με χρονικούς, ενεργειακούς ή υπολογιστικούς περιορισμούς.

- Συνολικός αριθμός εποχών για τη σύγκλιση του μοντέλου (Total Number of Epochs for Model Convergence): Ο αριθμός εποχών μέχρι τη σύγκλιση αποτελεί κρίσιμη μετρική για την αποτίμηση της αποδοτικότητας ενός CNN κατά τη φάση εκπαίδευσης, καθώς αποτυπώνει τον αριθμό των πλήρων διελεύσεων του συνόλου εκπαίδευσης που απαιτούνται, ώστε το μοντέλο να φτάσει σε σταθερή κατάσταση ως προς τα σφάλματα της συνάρτησης απώλειας

και τις επιδόσεις. Η μετρική αυτή είναι στενά συνδεδεμένη με τον ρυθμό εκμάθησης, την πολυπλοκότητα του δικτύου, τη φύση των δεδομένων και τον τύπο του βελτιστοποιητή (π.χ., SGD, Adam, AdaGrad, RMSprop). Ένα δίκτυο θεωρείται ότι έχει συγκλίνει όταν η συνάρτηση απώλειας παύει να παρουσιάζει σημαντικές μεταβολές ανά εποχή και οι επιδόσεις στο σύνολο επικύρωσης σταθεροποιούνται ή βελτιώνονται με πολύ μικρό ρυθμό.

Η σημασία αυτής της μετρικής έγκειται στο γεγονός ότι καθορίζει σε μεγάλο βαθμό τη συνολική διάρκεια της εκπαίδευσης και, κατ' επέκταση, το κόστος των υπολογιστικών πόρων και ενέργειας. Σε περιβάλλοντα περιορισμένων πόρων, όπως τα HCS, η μείωση του αριθμού εποχών μέχρι τη σύγκλιση αποτελεί στόχο σχεδιασμού, καθώς μπορεί να περιορίσει σημαντικά το συνολικό ενεργειακό αποτύπωμα της διαδικασίας εκπαίδευσης. Συνεπώς, η επιλογή κατάλληλων αρχιτεκτονικών, τεχνικών κανονικοποίησης, στρατηγικών αρχικοποίησης βαρών, αλλά και προηγμένων τεχνικών όπως ο προγραμματισμός ρυθμού μάθησης (learning rate scheduling) ή η χρήση πρόωρου τερματισμού (early stopping), συμβάλλουν καθοριστικά στη μείωση του αριθμού εποχών χωρίς υποβάθμιση της ποιότητας του τελικού μοντέλου.

Επιπροσθέτως, η συγκεκριμένη μετρική διαδραματίζει καθοριστικό ρόλο στη συγκριτική αξιολόγηση διαφορετικών αρχιτεκτονικών CNNs στο ίδιο υπολογιστικό περιβάλλον, καθώς παρέχει άμεση ένδειξη της αποδοτικότητας μάθησης κάθε μοντέλου. Η δυνατότητα ενός μοντέλου να επιτυγχάνει συγκρίσιμα επίπεδα ακρίβειας με μικρότερο αριθμό εποχών αποτελεί ένδειξη υψηλής προσαρμοστικότητας και ταχείας σύγκλισης, στοιχεία κρίσιμα για εφαρμογές με περιορισμένα χρονικά ή ενεργειακά περιθώρια. Ένα τέτοιο μοντέλο κρίνεται καταλληλότερο σε περιπτώσεις που απαιτούν επαναλαμβανόμενη εκπαίδευση ή περιοδική αναπροσαρμογή, όπως η μεταφορά μάθησης, ή η ενσωμάτωση νέων κατηγοριών. Έτσι, η μείωση του αριθμού εποχών μέχρι τη σύγκλιση δεν ενισχύει μόνο την υπολογιστική αποδοτικότητα, αλλά διευκολύνει και την πρακτική υλοποίηση του μοντέλου σε λειτουργία πραγματικού χρόνου και σε περιορισμένων πόρων υπολογιστικά περιβάλλοντα.

- Στρατηγικές επικύρωσης (Validation Strategies): Κατά τη φάση εκπαίδευσης ενός CNN, ο κατάλληλος διαχωρισμός του συνόλου δεδομένων αποτελεί κρίσιμο βήμα για τη διασφάλιση της γενικευσιμότητας και της αξιοπιστίας των αποτελεσμάτων. Οι στρατηγικές συγκράτησης (Hold-out) και

διασταυρούμενης επικύρωσης (Cross-validation) χρησιμοποιούνται ευρέως για την αποτίμηση της απόδοσης των CNNs, αποτρέποντας το φαινόμενο της υπερπροσαρμογής (overfitting) και παρέχοντας εκτιμήσεις της συμπεριφοράς του μοντέλου σε άγνωστα δεδομένα.

Η στρατηγική συγκράτησης αποτελεί μία από τις πλέον διαδεδομένες τεχνικές διαχωρισμού του συνόλου δεδομένων. Το αρχικό σύνολο διαχωρίζεται σε τρία διακριτά υποσύνολα: το σύνολο εκπαίδευσης (training set), το σύνολο επικύρωσης (validation set) και το σύνολο ελέγχου (test set). Το σύνολο εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου, δηλαδή την προσαρμογή των παραμέτρων του. Το σύνολο επικύρωσης χρησιμοποιείται για τη βελτιστοποίηση των υπερπαραμέτρων του μοντέλου (π.χ., ρυθμός μάθησης, βάθος δικτύου, το μέγεθος παρτίδας, αλγόριθμος βελτιστοποίησης). Το σύνολο ελέγχου χρησιμοποιείται μία μοναδική φορά, μετά την ολοκλήρωση της διαδικασίας εκπαίδευσης και επικύρωσης, για την εκτίμηση της απόδοσης του μοντέλου σε εντελώς νέα, μη ορατά δεδομένα. Μια συνηθισμένη αναλογία είναι 60%–80% για εκπαίδευση, 10%–20% για επικύρωση και 10%–20% για έλεγχο, αναλόγως του μεγέθους και της ποικιλομορφίας των δεδομένων.

Το κύριο πλεονέκτημα της στρατηγικής συγκράτησης είναι η χαμηλή υπολογιστική πολυπλοκότητα και ο μικρός χρόνος εκπαίδευσης, γεγονός που την καθιστά ιδιαίτερα κατάλληλη για μεγάλα σύνολα δεδομένων ή για περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους. Παρόλα αυτά, παρουσιάζει ορισμένους εγγενείς περιορισμούς: το τελικό αποτέλεσμα εξαρτάται από την τυχαία κατανομή των δειγμάτων στα επιμέρους υποσύνολα, γεγονός που μπορεί να οδηγήσει σε υψηλή διακύμανση των μετρικών απόδοσης. Επιπλέον, σε περιπτώσεις ανισορροπίας κλάσεων, ο τυχαίος διαχωρισμός μπορεί να οδηγήσει σε μη αντιπροσωπευτική κατανομή των κατηγοριών στα υποσύνολα, με αποτέλεσμα ορισμένες κλάσεις να υποεκπροσωπούνται στο σύνολο επικύρωσης. Αυτό ενδέχεται να οδηγήσει σε μεροληπτικές εκτιμήσεις της απόδοσης του μοντέλου και να αποκρύψει ενδεχόμενες αδυναμίες στον συμπερασμό των σπανιότερων κατηγοριών.

Η στρατηγική διασταυρούμενη επικύρωσης αποτελεί μία πιο στατιστικά αξιόπιστη προσέγγιση για την αξιολόγηση των μοντέλων. Το αρχικό σύνολο δεδομένων χωρίζεται σε k ισομεγέθη υποσύνολα (folds). Κατά την εκπαίδευση, κάθε ένα από τα k υποσύνολα χρησιμοποιείται διαδοχικά ως σύνολο ελέγχου, ενώ τα υπόλοιπα $k - 1$ ως σύνολο εκπαίδευσης. Η τελική απόδοση του

μοντέλου υπολογίζεται ως ο μέσος όρος των επιδόσεων σε κάθε μία από τις k επαναλήψεις.

Η διασταυρούμενη επικύρωση μειώνει τη διακύμανση στις μετρικές αξιολόγησης που οφείλεται στην τυχαία κατανομή των δειγμάτων και επιτρέπει την πλήρη αξιοποίηση των διαθέσιμων δεδομένων, καθώς κάθε δείγμα συμμετέχει τόσο στη φάση εκπαίδευσης όσο και στη φάση ελέγχου. Η μέθοδος αυτή είναι ιδιαιτέρως χρήσιμη σε περιπτώσεις μικρού ή μεσαίου μεγέθους συνόλων δεδομένων, όπου η απώλεια ακόμη και μικρού αριθμού δειγμάτων μπορεί να επηρεάσει δυσμενώς την εκτιμώμενη απόδοση του μοντέλου.

Ωστόσο, η στρατηγική της διασταυρούμενης επικύρωσης συνεπάγεται σημαντικά αυξημένο υπολογιστικό κόστος, ιδίως όταν ο αριθμός των διαμερίσεων k είναι μεγάλος, καθώς το μοντέλο πρέπει να εκπαιδευτεί k φορές από την αρχή. Η υπολογιστική πολυπλοκότητα αυτή καθιστά τη μέθοδο λιγότερο ελκυστική σε HCS με περιορισμένες δυνατότητες επεξεργασίας ή διαθέσιμης μνήμης. Για τον λόγο αυτό, σε περιβάλλοντα όπως τα FPGA–SoCs ή τα SoMs, χρησιμοποιείται συχνά διασταυρούμενη επικύρωση με μικρότερη τιμή k (low- k fold cross-validation, π.χ. $k = 2$), η οποία επιτυγχάνει καλύτερη ισορροπία ανάμεσα στην ακρίβεια εκτίμησης και στις υπολογιστικές απαιτήσεις.

- Χρονική καθυστέρηση ανάδρομης διάδοσης του μοντέλου (Model Backward Pass Latency): Η χρονική καθυστέρηση του μοντέλου κατά την ανάδρομη διάδοση (backward pass) αναφέρεται στον χρόνο που απαιτείται για την εκτέλεση της διαδικασίας υπολογισμού των κλίσεων (gradients) της συνάρτησης απώλειας και την ενημέρωση των παραμέτρων του δικτύου, δεδομένης μίας παρτίδας δεδομένων κατά την εκπαίδευση. Η ανάδρομη διάδοση είναι η πιο υπολογιστικά απαιτητική φάση της εκπαίδευσης, καθώς περιλαμβάνει διαδικασίες παραγωγίσιμης και μεταδόσεις σφαλμάτων σε όλα τα επίπεδα του δικτύου. Σε συνελκτικές αρχιτεκτονικές με μεγάλο πλήθος παραμέτρων και πολυεπίπεδη δομή, ο φόρτος της ανάδρομης διάδοσης μπορεί να υπερβαίνει τον χρόνο του προσθίου περάσματος (forward pass) κατά 2–3 φορές.

Η μετρική αυτή αποκτά ιδιαίτερη σημασία σε HCS με περιορισμένη μνήμη, περιορισμένο αριθμό υπολογιστικών μονάδων ή χαμηλή συχνότητα λειτουργίας, καθώς επηρεάζει άμεσα τόσο τη συνολική διάρκεια της

εκπαίδευσης όσο και την ενεργειακή της αποδοτικότητα. Η αυξημένη χρονική καθυστέρηση κατά την ανάδρομη διάδοση συνεπάγεται μεγαλύτερη κατανάλωση ενέργειας ανά εποχή και μειωμένη απόδοση σε περιβάλλοντα με χρονικούς ή θερμικούς περιορισμούς. Επιπλέον, σε περιπτώσεις εκπαίδευσης σε συσκευές υπολογισμού στο άκρο (edge training) — όπως έξυπνα κινητά, φορητές συσκευές ή βιοϊατρικοί αισθητήρες — η καθυστέρηση της ανάδρομης διάδοσης καθίσταται κρίσιμος παράγοντας για την πρακτική υλοποιησιμότητα της προσέγγισης.

Η ελαχιστοποίηση αυτής της καθυστέρησης μπορεί να επιτευχθεί μέσω του σχεδιασμού αρχιτεκτονικών με μικρότερο βάθος και πλάτος, καθώς και με την αξιοποίηση δομοστοιχείων χαμηλής υπολογιστικής πολυπλοκότητας. Παράλληλα, η μείωση του συνολικού αριθμού παραμέτρων, η πιο αποδοτική διαχείριση της μνήμης και η χρήση αριθμητικών αναπαραστάσεων χαμηλότερης ακρίβειας συμβάλλουν καθοριστικά στη μείωση της χρονικής καθυστέρησης της ανάδρομης διάδοσης. Ως εκ τούτου, η συγκεκριμένη μετρική λειτουργεί όχι μόνο ως δείκτης υπολογιστικής πολυπλοκότητας, αλλά και ως βασικό κριτήριο επιλογής των αρχιτεκτονικών σε περιβάλλοντα όπου η εκπαίδευση απαιτείται να είναι ταχεία, ενεργειακά αποδοτική και τεχνικά υλοποιήσιμη.

Ο Πίνακας 3.6 παρουσιάζει μια συγκριτική ανάλυση των τριών μετρικών — συνολικός αριθμός εποχών για τη σύγκλιση του μοντέλου, στρατηγικές επικύρωσης και χρονική καθυστέρηση ανάδρομης διάδοσης του μοντέλου — που σχετίζονται με τη φάση εκπαίδευσης των CNNs σε HCS περιορισμένων πόρων.

3.4.2 Φάση Εξαγωγής Συμπερασμάτων

Η παρούσα κατηγορία μετρικών της μεθοδολογίας μας επικεντρώνεται αποκλειστικά στη φάση εξαγωγής συμπερασμάτων, κατά την οποία το μοντέλο εφαρμόζεται για την πρόβλεψη σε νέα, άγνωστα δεδομένα [138]. Οι μετρικές αυτές είναι κρίσιμες για την αποτίμηση της καταλληλότητας ενός μοντέλου σε ενσωματωμένα συστήματα ή κινητές συσκευές, όπου η υπολογιστική και ενεργειακή αποδοτικότητα της εξαγωγής αποτελεί καθοριστικό παράγοντα.

Σε περιπτώσεις εξαγωγής συμπερασμάτων σε πραγματικό χρόνο (real-time inference) [139, 140], ακόμη και μικρές καθυστερήσεις ή αυξημένο αποθηκευτικό

Πίνακας 3.6: Συγκριτική ανάλυση μετρικών εκπαίδευσης για CNNs σε HCS.

Μετρική	Πλεονεκτήματα	Μειονεκτήματα
Total Number of Epochs for Model Convergence	Άμεση ένδειξη της αποδοτικότητας μάθησης και της ταχύτητας σύγκλισης του μοντέλου. Η μείωσή του οδηγεί σε λιγότερο χρόνο εκπαίδευσης, χαμηλότερο κόστος υπολογιστικών πόρων και ενέργειας.	Εξαρτάται από πολλούς παράγοντες (ρυθμό μάθησης, αρχιτεκτονική, σύνολο δεδομένων, βελτιστοποιητή). Υψηλός αριθμός εποχών αυξάνει τον χρόνο εκπαίδευσης και την κατανάλωση ενέργειας.
Validation Strategies	Η Hold-out έχει χαμηλή υπολογιστική πολυπλοκότητα και χρόνο εκπαίδευσης. Η Cross-validation είναι στατιστικά πιο αξιόπιστη στην εκτίμηση της απόδοσης.	Η Hold-out είναι ευάλωτη σε διακύμανση και μεροληψία σε περίπτωση ανισορροπίας κλάσεων. Η Cross-validation έχει αυξημένο υπολογιστικό κόστος, ειδικά για μεγάλο k .
Model Backward Pass Latency	Η μείωσή της οδηγεί σε ταχύτερη εκπαίδευση ανά εποχή. Μπορεί να μειωθεί μέσω αξιοποίησης δομοστοιχείων χαμηλής υπολογιστικής πολυπλοκότητας.	Αποτελεί το πιο απαιτητικό μέρος της εκπαίδευσης. Υψηλή καθυστέρηση αυξάνει την κατανάλωση ενέργειας.

αποτύπωμα μπορούν να επηρεάσουν τη συνολική απόδοση του συστήματος, είτε λόγω περιορισμένων υλικών πόρων είτε λόγω χρονικών περιορισμών. Συνεπώς, η φάση της εξαγωγής συμπερασμάτων δεν πρέπει να θεωρείται δευτερεύουσα, αλλά αποτελεί κρίσιμο στάδιο για τη βιώσιμη ενσωμάτωση του μοντέλου σε HCS.

- **Συνολικό μέγεθος μοντέλου στο δίσκο (Total Model Size on Disk):** Το αποθηκευτικό αποτύπωμα του μοντέλου σε επίπεδο δίσκου αποτελεί καθοριστικό παράγοντα για την ενσωμάτωσή του σε συστήματα με περιορισμένους πόρους αποθήκευσης. Το μέγεθος του μοντέλου στον δίσκο αναφέρεται στον αποθηκευτικό χώρο (συνήθως σε MB ή GB) που απαιτείται για την αποθήκευση όλων των παραμέτρων του εκπαιδευμένου CNN. Αυτές οι παράμετροι περιλαμβάνουν, τα βάρη (weights) και τις μετατοπίσεις (biases). Η τιμή αυτή καθορίζεται κυρίως από τον αριθμό των παραμέτρων του μοντέλου και τον τύπο δεδομένων που χρησιμοποιείται για την αριθμητική αναπαράσταση (π.χ., FP32, FP16, INT8).

Το μέγεθος του μοντέλου επηρεάζει άμεσα τη χρηστικότητα κατά την εξαγωγή συμπερασμάτων, ιδίως στον χρόνο φόρτωσης (όσο μεγαλύτερο είναι

το μοντέλο, τόσο περισσότερος χρόνος απαιτείται για τη φόρτωσή του από τον αποθηκευτικό χώρο στη μνήμη RAM), στις απαιτήσεις σε μνήμη (μεγάλα μοντέλα έχουν υψηλότερες απαιτήσεις RAM), στην ενεργειακή απόδοση και θερμική επιβάρυνση (η μεταφορά μεγάλων αρχείων μοντέλων στη μνήμη και η συνεχής λειτουργία τους επιβαρύνει ενεργειακά το σύστημα, αυξάνοντας την κατανάλωση ενέργειας και τη θερμική εκπομπή).

Η ανάγκη περιορισμού του μεγέθους του μοντέλου στον δίσκο, ιδίως σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους, έχει οδηγήσει στην ανάπτυξη ενός συνόλου τεχνικών βελτιστοποίησης. Αυτές οι τεχνικές εφαρμόζονται είτε μετά την εκπαίδευση είτε κατά τη φάση της εκπαίδευσης, με στόχο τη μείωση του αποθηκευτικού αποτυπώματος χωρίς ουσιώδη απώλεια στην απόδοση. Τέτοιες τεχνικές περιλαμβάνουν, μεταξύ άλλων, την κβαντοποίηση και το κλάδεμα.

- Χρήση πόρων υλικού (Hardware Resource Utilization): Η χρήση των πόρων του υλικού κατά τη φάση εξαγωγής συμπερασμάτων αποκτά ιδιαίτερη σημασία όταν η ανάπτυξη και η εκτέλεση των CNNs πραγματοποιείται σε HCS, όπως τα SoMs και τα FPGA—SoCs. Οι δύο αυτοί τύποι αρχιτεκτονικών ενσωματώνουν διαφορετικά μοντέλα υπολογιστικής επιτάχυνσης και απαιτούν ειδική μέριμνα για τη βέλτιστη αξιοποίηση των διαθέσιμων πόρων υπολογιστικής ισχύος, μνήμης και ενέργειας κατά την ευθεία διάδοση του μοντέλου.

Κατά την εξαγωγή συμπερασμάτων σε SoMs, η GPU υλοποιεί τις πράξεις με υψηλό βαθμό παραλληλισμού, η CPU διαχειρίζεται τη φόρτωση δεδομένων, ο DLA (Deep Learning Accelerator) επιτρέπει την εκτέλεση με χρήση INT8 ή FP16 μειώνοντας τη χρήση της GPU και την κατανάλωση ενέργειας και η RAM χρησιμοποιείται για την αποθήκευση των ενδιάμεσων αποτελεσμάτων. Κατά την εξαγωγή συμπερασμάτων σε FPGA—SoCs, τα LUTs υλοποιούν τις λογικές λειτουργίες (ReLU, MaxPooling), τα DSPs εκτελούν τις πράξεις όπως πολλαπλασιασμούς, η BRAM χρησιμοποιείται για ενδιάμεσες τιμές και προσωρινή αποθήκευση βαρών και ο ARM επεξεργαστής αναλαμβάνει τον συντονισμό και την είσοδο/έξοδο.

Η χρήση των πόρων του υλικού κατά τη φάση εξαγωγής συμπερασμάτων διαφοροποιείται ανάλογα με την υποκείμενη συσκευή. Σε περιβάλλοντα περιορισμένων πόρων η αξιοποίηση του υλικού δεν επηρεάζει μόνο την ταχύτητα εξαγωγής συμπερασμάτων αλλά και την κατανάλωση ενέργειας και τη θερμική συμπεριφορά. Στόχος είναι να επιτευχθεί η μέγιστη απόδοση

ανά διαθέσιμη μονάδα υπολογιστικού ή μνημονικού πόρου, χωρίς συμφόρηση, σπατάλη ή μη ισορροπημένη χρήση.

- Χρονική καθυστέρηση ευθείας διάδοσης του μοντέλου (Model Forward Pass Latency): Η χρονική καθυστέρηση του μοντέλου αναφέρεται στον συνολικό χρόνο που απαιτείται για την ολοκλήρωση ενός κύκλου ευθείας διάδοσης, από τη στιγμή που τα δεδομένα εισόδου δίνονται στο μοντέλο έως τη στιγμή που αυτό παράγει την αντίστοιχη έξοδο. Η καθυστέρηση αυτή επηρεάζεται από πλήθος παραγόντων, όπως η πολυπλοκότητα της αρχιτεκτονικής (π.χ. το πλήθος των επιπέδων και η δομή των συνελκτικών φίλτρων), το μέγεθος και οι διαστάσεις των δεδομένων εισόδου, η υπολογιστική αρχιτεκτονική στην οποία εκτελείται το μοντέλο (π.χ. FPGA, SoM) καθώς και το είδος της αριθμητικής αναπαράστασης που χρησιμοποιείται.

Η καθυστέρηση αυτή επηρεάζει άμεσα τη χρονική απόκριση του συστήματος και τη δυνατότητα λειτουργίας σε πραγματικό χρόνο, γεγονός που καθίσταται κρίσιμο σε εφαρμογές υψηλών απαιτήσεων, όπως η ιατρική διάγνωση, τα αυτόνομα οχήματα και η ανίχνευση επιθέσεων στον τομέα της κυβερνοασφάλειας. Η βελτιστοποίησή της δεν εξαρτάται αποκλειστικά από τα χαρακτηριστικά του ίδιου του μοντέλου, αλλά απαιτεί μια ολιστική προσέγγιση, η οποία συνδυάζει τον κατάλληλο αρχιτεκτονικό σχεδιασμό—με επιλογή δικτύων μικρότερου βάθους και πλάτους και αξιοποίηση δομοστοιχείων χαμηλής υπολογιστικής πολυπλοκότητας— με την εφαρμογή τεχνικών μετα-εκπαιδευτικής βελτιστοποίησης, όπως η κβαντοποίηση και το κλάδεμα, οι οποίες αποσκοπούν στη μείωση της χρονικής καθυστέρησης χωρίς σημαντική απώλεια στην ακρίβεια εξαγωγής συμπεράσματος.

- Ενεργειακή απόδοση ανά Watt (Energy Performance per Watt): Η ενεργειακή απόδοση ανά Watt κατά τη φάση εξαγωγής συμπερασμάτων σε CNNs αποτελεί έναν κρίσιμο δείκτη για την αξιολόγηση της λειτουργικής και υπολογιστικής αποδοτικότητας ενός συστήματος AI. Ειδικά σε HCS, όπως τα SoMs και τα FPGA—SoCs, η αποδοτική χρήση ενέργειας ανά μονάδα απόδοσης είναι καθοριστική για την επίτευξη πραγματικού χρόνου εκτέλεσης, τη θερμική σταθερότητα, και τη μακροπρόθεσμη βιωσιμότητα του συστήματος.

Η ενεργειακή απόδοση ανά Watt εξαρτάται από ένα πλήθος παραμέτρων, με κυριότερες την αρχιτεκτονική του μοντέλου όπου δίκτυα χαμηλής πολυπλοκότητας εμφανίζουν μειωμένες απαιτήσεις σε πράξεις, γεγονός που συνεπάγεται μικρότερη κατανάλωση ενέργειας κατά την ευθεία διάδοση. Η

ενεργειακή βελτιστοποίηση των μοντέλων σχετίζεται στενά με το πλήθος και το είδος των επιπέδων, τη χρήση συνελκτικών φίλτρων και την απουσία πυκνών συνδέσεων υψηλής διάστασης. Επίσης ο τύπος αριθμητικής αναπαράστασης όπου η χρήση αριθμητικής μειωμένης ακρίβειας, όπως FP16 ή INT8, επιτρέπει την εκτέλεση πράξεων με χαμηλότερο ενεργειακό κόστος.

Η ενεργειακή απόδοση ανά Watt δεν αποτελεί απλώς τεχνική παράμετρο, αλλά πολυδιάστατο μέγεθος στρατηγικής σημασίας για τον σχεδιασμό αλγορίθμων AI που προορίζονται για ανάπτυξη σε περιορισμένων πόρων HCS περιβάλλοντα. Η επίτευξη βέλτιστης σχέσης μεταξύ υπολογιστικής απόδοσης και ενεργειακής κατανάλωσης αποτελεί αναγκαία συνθήκη για την ενσωμάτωση της AI σε κρίσιμους τομείς όπως η ιατρική και τα αυτόνομα συστήματα.

Ο Πίνακας 3.7 παρουσιάζει μια συγκριτική ανάλυση των μετρικών — συνολικό μέγεθος του μοντέλου στο δίσκο, χρήση των πόρων του υλικού, χρονική καθυστέρηση ευθείας διάδοσης του μοντέλου και ενεργειακή απόδοση ανά Watt — που σχετίζονται με τη φάση εξαγωγής συμπερασμάτων των CNNs σε HCS περιορισμένων πόρων.

3.4.3 Φάση Εκπαίδευσης και Εξαγωγής Συμπερασμάτων

Η κατηγορία αυτή της μεθοδολογίας μας καλύπτει μετρικές που επηρεάζουν τόσο το υπολογιστικό φορτίο όσο και την κατανάλωση μνήμης σε πραγματικές συνθήκες λειτουργίας του μοντέλου, ανεξαρτήτως φάσης. Οι εν λόγω μετρικές διαδραματίζουν κρίσιμο ρόλο σε περιπτώσεις όπου η βελτιστοποίηση δεν επικεντρώνεται μόνο στην επίδοση εξαγωγής συμπερασμάτων, αλλά και στην αποδοτική αξιοποίηση των υπολογιστικών πόρων. Σε εφαρμογές που εκτελούνται σε ενεργειακά περιορισμένες συσκευές η ανάγκη για μοντέλα με χαμηλό αριθμό πράξεων, περιορισμένο πλήθος παραμέτρων και ελεγχόμενη κατανάλωση μνήμης καθίσταται επιτακτική.

- Συνολικός αριθμός των παραμέτρων του μοντέλου (Total Number of Model Parameters): Ο αριθμός παραμέτρων σε ένα CNN σχετίζεται άμεσα με την ικανότητα εκμάθησης, την εκφραστικότητα του μοντέλου, καθώς και με την υπολογιστική του πολυπλοκότητα, τόσο κατά τη φάση εκπαίδευσης όσο και κατά τη φάση εξαγωγής συμπερασμάτων.

Πίνακας 3.7: Συγκριτική ανάλυση μετρικών εξαγωγής συμπερασμάτων για CNNs σε HCS.

Μετρική	Πλεονεκτήματα	Μειονεκτήματα
Total Model Size on Disk	Βελτιστοποιείται μέσω τεχνικών όπως κβαντοποίηση και κλάδεμα.	Μεγάλα μοντέλα απαιτούν αυξημένο χρόνο φόρτωσης και μνήμη. Επηρεάζει αρνητικά την ενεργειακή και θερμική συμπεριφορά.
Hardware Resource Utilization	Ενισχύει την ενεργειακή και λειτουργική αποδοτικότητα του συστήματος. Επιτρέπει βέλτιστη κατανομή υπολογιστικών και μνημονικών πόρων σε HCS.	Διαφοροποιείται σημαντικά ανάλογα με την συσκευή (SoM, FPGA). Η μη ομοιόμορφη χρήση πόρων προκαλεί συμφόρηση ή σπατάλη.
Model Forward Pass Latency	Μπορεί να μειωθεί με χρήση χαμηλής πολυπλοκότητας και αρχιτεκτονικών και βελτιστοποιήσεων.	Επηρεάζεται από το είδος αριθμητικής αναπαράστασης, το υλικό και τα δεδομένα εισόδου. Η υπερβολική μείωση ενδέχεται να υποβαθμίσει την ακρίβεια.
Energy Performance per Watt	Αξιολογεί τη σχέση μεταξύ απόδοσης και κατανάλωσης ενέργειας.	Εξαρτάται από την αρχιτεκτονική και τον τύπο αριθμητικής αναπαράστασης.

Κατά τη φάση εκπαίδευσης, το σύνολο των παραμέτρων επηρεάζει τη δυναμική εκμάθησης διότι ένα μοντέλο με περισσότερες παραμέτρους έχει αυξημένη ικανότητα προσαρμογής σε περίπλοκα πρότυπα εισόδου, αλλά ταυτόχρονα απαιτεί περισσότερη μνήμη RAM, μεγαλύτερο αριθμό πράξεων και αυξημένο χρόνο εκπαίδευσης. Επιπλέον, η υψηλή παραμετρικότητα μπορεί να οδηγήσει σε φαινόμενα υπερπροσαρμογής, ειδικά σε περιπτώσεις μικρών συνόλων δεδομένων.

Αντίστοιχα, κατά τη φάση εξαγωγής συμπερασμάτων, οι παράμετροι του μοντέλου δεν μεταβάλλονται, αλλά πρέπει να φορτωθούν από τον αποθηκευτικό χώρο στη μνήμη. Το πλήθος τους καθορίζει τόσο το μέγεθος του μοντέλου στον δίσκο, όσο και τη ζήτηση σε μνήμη κατά την εκτέλεση. Επιπλέον, επηρεάζουν άμεσα την καθυστέρηση φόρτωσης και την απόδοση του συστήματος σε περιβάλλοντα με περιορισμένους πόρους.

- Συνολικός αριθμός πράξεων κινητής υποδιαστολής του μοντέλου (Total Number of Model FLOP): Ο αριθμός των FLOP λειτουργεί ως δείκτης της απαίτησης ενός μοντέλου σε υπολογιστικό φόρτο και συσχετίζεται άμεσα με

τις επιδόσεις, τη διάρκεια εκπαίδευσης, τη λανθάνουσα καθυστέρηση και την ενεργειακή κατανάλωση. Πρόκειται για τον συνολικό αριθμό των αριθμητικών πράξεων κινητής υποδιαστολής (προσθέσεις και πολλαπλασιασμοί) που απαιτούνται για την επεξεργασία ενός δείγματος εισόδου σε ένα CNN. Κατά την εκπαίδευση, ένα CNN εκτελεί δύο βασικά στάδια, την ευθεία διάδοση και την οπίσθια διάδοση.

Η συνολική υπολογιστική πολυπλοκότητα ανά εποχή κατά τη φάση εκπαίδευσης ορίζεται από τη σχέση:

$$\text{FLOP}_{\text{training}} = N \cdot B \cdot (F + B) \quad (3.12)$$

όπου N είναι ο αριθμός των δειγμάτων, B είναι ο αριθμός των παρτίδων, F είναι ο αριθμός των FLOP της ευθείας διάδοσης και B είναι ο αριθμός των FLOP της οπίσθιας διάδοσης.

Κατά τη φάση της εξαγωγής συμπερασμάτων, εκτελείται μόνο η ευθεία διάδοση. Ο αριθμός των FLOP ανά συμπερασμό υπολογίζεται ως το άθροισμα της υπολογιστικής πολυπλοκότητας όλων των επιπέδων του δικτύου:

$$\text{FLOP}_{\text{inference}} = \sum_{\ell=1}^L \text{FLOP}(\ell) \quad (3.13)$$

όπου L είναι το συνολικό πλήθος των επιπέδων του μοντέλου και κάθε $\text{FLOP}(\ell)$ εξαρτάται από τον τύπο του επιπέδου (συνέλιξη, πλήρως συνδεδεμένο, κανονικοποίηση, ενεργοποίηση κ.λπ.).

Η πολυπλοκότητα των συνελικτικών επιπέδων ορίζεται από τη σχέση:

$$\text{FLOP}_{\text{conv}} = 2 \cdot H_o \cdot W_o \cdot C_o \cdot K_h \cdot K_w \cdot C_i \quad (3.14)$$

όπου H_o , W_o είναι οι χωρικές διαστάσεις της εξόδου, C_o είναι ο αριθμός των καναλιών εξόδου, C_i είναι ο αριθμός των καναλιών εισόδου, K_h , K_w είναι οι διαστάσεις του πυρήνα συνέλιξης.

Ο αριθμός των FLOP χρησιμοποιείται ως δείκτης σύγκρισης των μοντέλων διαφορετικής αρχιτεκτονικής, ως εργαλείο επιλογής της κατάλληλης συσκευής υλικού και ως εκτιμητής της ενεργειακής κατανάλωσης. Όσο μικρότερος είναι ο αριθμός των FLOP, χωρίς σημαντική απώλεια στην

ακρίβεια συμπερασμού, τόσο πιο αποδοτικό θεωρείται το μοντέλο, ιδίως σε συστήματα με περιορισμένους υπολογιστικούς και ενεργειακούς πόρους.

- Χρήση μνήμης RAM (μέγιστη, μέση) του μοντέλου (Model RAM Consumption (peak, average)): Η μέγιστη και μέση χρήση μνήμης κατά τις φάσεις εκπαίδευσης και εξαγωγής συμπερασμάτων επηρεάζει καθοριστικά τη δυνατότητα ενσωμάτωσης ενός CNN σε υπολογιστικά συστήματα με περιορισμένη μνημονική υποδομή.

Κατά τη φάση εκπαίδευσης ενός CNN, το αποτύπωμα μνήμης είναι σημαντικά αυξημένο, καθώς απαιτείται η διατήρηση των ενεργοποιήσεων για την οπισθοδιάδοση, καθώς και των προσωρινών μεταβλητών που σχετίζονται με τον αλγόριθμο βελτιστοποίησης. Αντίθετα, κατά τη φάση εξαγωγής συμπερασμάτων, η χρήση μνήμης είναι σημαντικά περιορισμένη, καθώς εκτελείται αποκλειστικά η ευθεία διάδοση χωρίς την αποθήκευση των ενδιάμεσων μεταβλητών ή τον υπολογισμό παραγώγων.

Ο Πίνακας 3.8 παρουσιάζει μια συγκριτική ανάλυση των μετρικών – συνολικός αριθμός των παραμέτρων, αριθμός πράξεων κινητής υποδιαστολής, μέγιστη και μέση χρήση μνήμης RAM – που σχετίζονται με τη φάση εκπαίδευσης και εξαγωγής συμπερασμάτων των CNNs σε HCS περιορισμένων πόρων.

Πίνακας 3.8: Συγκριτική ανάλυση μετρικών εκπαίδευσης και εξαγωγής συμπερασμάτων για CNNs σε HCS.

Μετρική	Πλεονεκτήματα	Μειονεκτήματα
Total Number of Model Parameters	Αντικατοπτρίζει την εκφραστικότητα και την ικανότητα γενίκευσης του μοντέλου.	Μεγάλος αριθμός παραμέτρων οδηγεί σε αυξημένες απαιτήσεις RAM και κίνδυνο υπερπροσαρμογής.
Total Number of Model FLOP	Προσφέρει σαφή εκτίμηση της υπολογιστικής απαίτησης ανά δείγμα. Χρησιμοποιείται για σύγκριση μοντέλων και εκτίμηση ενεργειακής κατανάλωσης.	Υψηλές τιμές FLOP συνεπάγονται αυξημένη καθυστέρηση και ενεργειακή επιβάρυνση.
Model RAM Consumption (peak, avg)	Μπορεί να μειωθεί με τεχνικές όπως κβαντοποίηση και κλάδεμα.	Η υψηλή χρήση μνήμης RAM καθιστά αδύνατη την ανάπτυξη σε χαμηλού κόστους συσκευές.

3.5 Συνολική Αξιολόγηση Καταλληλότητας Μοντέλων ως προς τις Μετρικές Αποδοτικότητας

Η συνολική αξιολόγηση της καταλληλότητας ενός μοντέλου σε HCS προϋποθέτει την ταυτόχρονη θεώρηση ενός πλήθους μετρικών απόδοσης που υπερβαίνουν τη συμβατική ακρίβεια συμπερασμού. Η συνδυαστική ενσωμάτωση μετρικών — σχεδίασης, ποιότητας και υπολογιστικού αποτυπώματος — επιτρέπει τη συγκροτημένη εκτίμηση της πραγματικής αποδοτικότητας ενός μοντέλου ως προς την υλοποίηση και τη λειτουργία του σε περιβάλλοντα με περιορισμένους υπολογιστικούς και ενεργειακούς πόρους. Μέσω αυτής της πολυκριτηριακής προσέγγισης είναι δυνατή η ανάδειξη εκείνων των αρχιτεκτονικών που επιτυγχάνουν ισορροπία μεταξύ των επιδόσεων και της πρακτικής βιωσιμότητας σε HCS.

Για παράδειγμα, ένα μοντέλο που επιτυγχάνει υψηλές τιμές ακρίβειας συμπερασμού, ενδέχεται να συνοδεύεται από δυσανάλογο υπολογιστικό και ενεργειακό αποτύπωμα, καθιστώντας το μη βιώσιμο για υλοποίηση σε συσκευές όπως SoMs ή FPGA—SoCs, οι οποίες διαθέτουν περιορισμένο αποθεματικό μνήμης, ισχύος και χρονικών καθυστερήσεων. Αντιστρόφως, μοντέλα που παρουσιάζουν ελαφρύτερη δομή — χαμηλότερο αριθμό παραμέτρων, μειωμένα FLOP — ενδέχεται να υστερούν οριακά σε μετρικές ποιότητας (π.χ. F1—score, AUC), αλλά να αποδεικνύονται καταλληλότερα για ενσωμάτωση σε περιβάλλοντα όπως ενσωματωμένα ή φορητά συστήματα AI.

Η αξιολόγηση της συνολικής καταλληλότητας απαιτεί επομένως τη θεώρηση της πολυδιάστατης αλληλεπίδρασης μεταξύ της ακρίβειας εξαγωγής συμπεράσματος, του κόστους υλοποίησης (σε υπολογιστικό και ενεργειακό επίπεδο), καθώς και της απόκρισης του μοντέλου στο πραγματικό περιβάλλον—στόχο. Η ισορροπία αυτή καθίσταται καθοριστική σε πεδία εφαρμογής όπως η φορητή ιατρική διάγνωση και η αυτόνομη πλοήγηση, στα οποία η ικανότητα του μοντέλου να λειτουργεί υπό πραγματικούς χρονικούς και ενεργειακούς περιορισμούς αποκτά μεγαλύτερη σημασία από τη βέλτιστη θεωρητική του επίδοση. Σε τέτοιες περιπτώσεις, η επιτυχής ενσωμάτωση ενός CNN απαιτεί τη σύνθεση όλων των προαναφερθεισών μετρικών σε ένα ενιαίο πλαίσιο αξιολόγησης, το οποίο αντανακλά την ουσιαστική χρηστικότητα και το λειτουργικό αποτύπωμα της κάθε αρχιτεκτονικής.

3.6 Σύνοψη Κεφαλαίου

Στο παρόν κεφάλαιο παρουσιάστηκε η προτεινόμενη μεθοδολογία για τον αποδοτικό σχεδιασμό και την υλοποίηση των CNNs σε HCS. Η μεθοδολογία οργανώθηκε σε τρεις θεμελιώδεις άξονες: τις μετρικές σχεδίασης, τις μετρικές ποιότητας και τις μετρικές υπολογιστικού αποτυπώματος. Κάθε άξονας αναφέρεται σε συγκεκριμένες παραμέτρους και δείκτες, οι οποίοι συνδυαστικά επιτρέπουν την ολιστική αξιολόγηση και βελτιστοποίηση των CNNs σε απαιτητικά περιβάλλοντα.

Συγκεκριμένα, οι μετρικές σχεδίασης εστιάζουν στα δομικά και λειτουργικά χαρακτηριστικά των μοντέλων, επιτρέποντας τη βελτιστοποίηση της αρχιτεκτονικής με στόχο τη μείωση της πολυπλοκότητας χωρίς απώλεια στην απόδοση. Οι μετρικές ποιότητας αξιολογούν την ακρίβεια και τη γενικευσιμότητα του μοντέλου, διασφαλίζοντας ότι οι προβλέψεις είναι αξιόπιστες και συνεπείς. Τέλος, οι μετρικές υπολογιστικού αποτυπώματος παρέχουν ουσιαστική πληροφόρηση σχετικά με την κατανάλωση πόρων, τον χρόνο εκτέλεσης και την ενεργειακή αποδοτικότητα, στοιχεία ζωτικής σημασίας για την ενσωμάτωση των CNNs σε περιορισμένα ή φορητά συστήματα.

Η συστηματική ενσωμάτωση των παραπάνω μετρικών οδηγεί στη δημιουργία ενός ευέλικτου και επεκτάσιμου πλαισίου αξιολόγησης και σχεδίασης. Μέσω αυτής της προσέγγισης, διασφαλίζεται ότι τα CNNs που υλοποιούνται σε HCS ανταποκρίνονται με αποτελεσματικότητα τόσο στις τεχνικές απαιτήσεις όσο και στους περιορισμούς του εκάστοτε υπολογιστικού περιβάλλοντος, ενώ παράλληλα διατηρούν υψηλή ακρίβεια συμπερασμού, υπολογιστική απόδοση και χαμηλή κατανάλωση ενέργειας. Το επόμενο κεφάλαιο παρουσιάζει τη σχεδίαση, υλοποίηση και βελτιστοποίηση CNN σε FPGA—SoC, με αξιολόγηση των μετρικών που ορίζονται από την προτεινόμενη μεθοδολογία.

Κεφάλαιο 4

Σχεδιασμός και Υλοποίηση Επιτάχυνσης σε FPGA—SoC

Στο προηγούμενο κεφάλαιο παρουσιάστηκε αναλυτικά η προτεινόμενη μεθοδολογία για τον αποδοτικό σχεδιασμό και την υλοποίηση των CNNs σε HCS. Σε αυτό το κεφάλαιο¹, εφαρμόζεται η παραπάνω μεθοδολογία για τον σχεδιασμό και την υλοποίηση επιτάχυνσης CNN σε FPGA—SoC. Αρχικά, περιγράφεται η αρχιτεκτονική που επιλέγεται σύμφωνα με τις κατευθυντήριες γραμμές της μεθοδολογίας, ώστε να επιτευχθεί αποδοτική εκτέλεση του δικτύου. Στη συνέχεια, αναλύεται η διαδικασία βελτιστοποίησης που βασίζεται στην εφαρμογή προηγμένων τεχνικών επιτάχυνσης, πάντα με γνώμονα τις προδιαγραφές και τις μετρικές που καθορίζει η προτεινόμενη μεθοδολογία. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση και την συζήτηση των αποτελεσμάτων αξιολόγησης της προτεινόμενης υλοποίησης, με βάση καθιερωμένες μετρικές απόδοσης, καθώς και με συγκρίσεις έναντι εναλλακτικών υλοποιήσεων σε άλλες υπολογιστικές συσκευές.

4.1 Περίληψη

Το πεδίο της AI έχει ραγδαία ανάπτυξη τα τελευταία χρόνια, επιτρέποντας την επίλυση σύνθετων προβλημάτων και την υλοποίηση καινοτόμων εφαρμογών σε διάφορους τομείς, όπως η γεωργία. Η αναγνώριση των ασθενειών των φύλλων στη γεωργία θεωρείται σημαντική πτυχή για τη διασφάλιση της ποσότητας, της

¹Μέρος του κεφαλαίου έχει δημοσιευθεί στο άρθρο [141]

ποιότητας και της παραγωγής τροφίμων. Κατά κανόνα, τα φύλλα των καλλιεργειών τομάτας είναι ευαίσθητα και επιρρεπή σε διάφορες ασθένειες όπως η μούχλα των φύλλων, η στοχευμένη κηλίδα, η όψιμη λοίμωξη ή η βακτηριακή κηλίδα. Ωστόσο, αυτές οι ασθένειες των φυτών τομάτας είναι δύσκολο να αναγνωριστούν και η έγκαιρη διάγνωση τους είναι ζωτικής σημασίας.

Ταυτόχρονα, η συνεχής ανάπτυξη των προσεγγίσεων των CNNs έχει βοηθήσει σημαντικά στη διάγνωση των ασθενειών των φυτών, παρέχοντας έναν αξιόπιστο μηχανισμό με ακριβή αποτελέσματα. Από την άλλη πλευρά, ο αριθμός των συλλεγόμενων εικόνων με φύλλα ασθενειών είναι συχνά μη ισορροπημένος και η διάγνωση τους με ένα τόσο μη ισορροπημένο σύνολο δεδομένων είναι περίπλοκη. Έτσι, έχουν προταθεί πολυάριθμα μοντέλα για τη διάγνωση ασθενειών τομάτας. Ωστόσο, κανένα από αυτά δεν αντιμετωπίζει το πρόβλημα της ανισορροπίας των κλάσεων και, ως αποτέλεσμα, δεν παράγει ευρήματα με αμερόληπτη ακρίβεια.

Αυτό το κεφάλαιο παρουσιάζει μια αξιόπιστη και αποτελεσματική εφαρμογή της προτεινόμενης μεθοδολογίας, με στόχο την επιτάχυνση στην ετερογενή συσκευή PYNQ-Z1. Οι τεχνικές βελτιστοποίησης—συμπεριλαμβανομένου του ξετυλίγματος βρόχων (Unroll), της διασωλήνωσης (Pipeline), της κατάτμησης πινάκων (Array Partitioning) και της επιπέδωσης βρόχων (Loop Flattening)—ενισχύουν την ταχύτητα υπολογισμού στα στρώματα συνέλιξης, πλήρως συνδεδεμένα και μέγιστης συγκέντρωσης του δικτύου. Η προτεινόμενη προσέγγιση που αναπτύξαμε περιλαμβάνει ένα δίκτυο χαμηλής πολυπλοκότητας με οκτώ επίπεδα, το οποίο ονομάζεται MiniTomatoNet.

Αυτό το δίκτυο χαρακτηρίζεται από τη χαμηλή πολυπλοκότητα του, διαθέτοντας λιγότερες από 23 χιλιάδες παραμέτρους, με το σύνολο των βαρών να καταλαμβάνει 89,51 KB μνήμης σε αναπαράσταση κινητής υποδιαστολής FP32. Επιπλέον, το CNN εκπαιδεύτηκε με τη χρήση της συνάρτησης επανασταθμισμένης εστιακής απώλειας και πέτυχε συνολική ακρίβεια 97,63% και βαθμολογία AUC 98,51%. Ο χρόνος εξαγωγής συμπεράσματος είναι 0,068 δευτερόλεπτα ανά είσοδο εικόνας διαστάσεων 128×128 εικονοστοιχείων, με κατανάλωση ισχύος 2,35 Watt. Τέλος, το μοντέλο είναι χαμηλής κατανάλωσης, αξιόπιστο και υψηλής ακρίβειας εξαγωγής συμπερασμάτων, καθιστώντας το μια πολλά υποσχόμενη λύση για τη διάγνωση ασθενειών της τομάτας.

4.2 Εισαγωγή

Η γεωργία αποτελεί θεμελιώδη οικονομική βάση σε πολλές χώρες, ιδιαίτερα σε περιοχές που στηρίζονται κυρίως σε αγροτικές δραστηριότητες. Ως εκ τούτου, ο αποτελεσματικός έλεγχος και η διαχείριση των ασθενειών είναι ουσιαστικής σημασίας για την πρόληψη των απωλειών της παραγωγής και τη διασφάλιση των βιώσιμων γεωργικών πρακτικών. Αυτό απαιτεί συνεχή παρακολούθηση και ακριβή διάγνωση των ασθενειών των φυτών. Η κύρια πρόκληση στη γεωργία είναι η αναγνώριση των συμπτωμάτων των ασθενειών που επηρεάζουν σημαντικά τις αποδόσεις των καλλιεργειών [142–146].

Οι παραδοσιακές χειροκίνητες και μηχανικές μέθοδοι αποτυγχάνουν να καλύψουν επαρκώς εκτεταμένες γεωργικές εκτάσεις ή να παρέχουν τα απαραίτητα αρχικά δεδομένα για τη λήψη τεκμηριωμένων αποφάσεων. Κατά συνέπεια, ανακύπτει επιτακτική ανάγκη για την ανάπτυξη αξιόπιστων, οικονομικά αποδοτικών και αυτοματοποιημένων συστημάτων, τα οποία θα επιτρέπουν την αποτελεσματική παρακολούθηση της υγείας των καλλιεργειών [147, 148].

Τα αυτοματοποιημένα συστήματα παρακολούθησης ασθενειών παρέχουν ουσιώδεις πληροφορίες που παρέχουν στους αγρότες και στους γεωργικούς φορείς τα απαραίτητα εργαλεία για τη λήψη τεκμηριωμένων αποφάσεων. Αυτά τα συστήματα συμβάλλουν στη διευκόλυνση της πρώιμης ανίχνευσης των ασθενειών, στην πρόληψη της εξάπλωσης των λοιμώξεων και στη μείωση των απωλειών των καλλιεργειών. Επιπλέον, συμβάλλουν στη βελτιστοποίηση των πόρων—όπως η χρήση νερού και λιπασμάτων—παρέχοντας ακριβή δεδομένα για την υγεία και την ανάπτυξη των καλλιεργειών, βελτιώνοντας έτσι τη γεωργική αποδοτικότητα [149, 150].

Η καλλιέργεια τομάτας αποτελεί μια από τις σημαντικότερες γεωργικές δραστηριότητες σε παγκόσμια κλίμακα, συμβάλλοντας ουσιαστικά στη συνολική αγροτική παραγωγή, το εμπόριο και την οικονομία [151, 152]. Οι τομάτες εκτιμώνται ιδιαίτερα για τις αντιοξειδωτικές τους ιδιότητες και τα θρεπτικά τους οφέλη, γεγονός που τις καθιστά βασικό διατροφικό στοιχείο σε πολλές χώρες. Η κατανάλωσή τους έχει συσχετιστεί με μείωση του κινδύνου εμφάνισης καρκίνου, χρόνιων νοσημάτων και φλεγμονωδών καταστάσεων, ενώ προσφέρει ευεργετικές επιδράσεις στην καρδιαγγειακή υγεία, συμβάλλοντας στη μείωση του κινδύνου υπέρτασης, καρδιαγγειακών παθήσεων, παχυσαρκίας και διαβήτη [153].

Ωστόσο, η καλλιέργεια τομάτας αντιμετωπίζει σημαντικές προκλήσεις λόγω της ευαισθησίας της σε ασθένειες, οι οποίες μπορεί να κυμαίνονται από ήπιες έως καταστροφικές. Παράγοντες όπως η αυξημένη υγρασία και η μειωμένη ηλιοφάνεια επιδεινώνουν αυτά τα προβλήματα, οδηγώντας συχνά σε εξάρσεις ασθενειών και παρασίτων. Τα πρώιμα στάδια πολλών ασθενειών της τομάτας είναι δύσκολο να διακριθούν με βάση την οπτική παρατήρηση, γεγονός που περιπλέκει την αναγνώρισή τους και συχνά απαιτεί ειδική ανάλυση από εμπειρογνώμονες και συνεχή παρακολούθηση—διαδικασίες που είναι ταυτόχρονα δαπανηρές και χρονοβόρες [154]. Κατά συνέπεια, η ανάπτυξη αυτοματοποιημένων συστημάτων για την πρώιμη διάγνωση των ασθενειών των φυτών της τομάτας είναι ζωτικής σημασίας. Αυτά τα αυτοματοποιημένα συστήματα επιτρέπουν την έγκαιρη παρέμβαση για τη διαχείριση και πρόληψη της εξάπλωσης των ασθενειών, διασφαλίζοντας την απόδοση των καλλιεργειών και συμβάλλοντας στην επίτευξη της επισιτιστικής ασφάλειας [155].

Η γεωργική βιομηχανία αξιοποιεί την ML και την AI σε διάφορους τομείς [156–160], συμπεριλαμβανομένης της διάγνωσης ασθενειών, του ελέγχου ζιζανίων, της παρακολούθησης του εδάφους, των drones για ανάλυση καλλιεργειών, της διάγνωσης παρασίτων και της πρόβλεψης του καιρού [161–163]. Πρόσφατα, τα CNNs αποτελούν τη νέα τάση στην ML και έχουν λάβει μεγάλη προσοχή και δημοτικότητα [164–167] για την επίλυση πολλών περίπλοκων προβλημάτων [168]. Τα προβλήματα αυτά επικεντρώνονται στην ταξινόμηση ή την αντιστοίχιση ενός αντικειμένου σε μία από τις προκαθορισμένες κατηγορίες. Η ταξινόμηση είναι ουσιώδης σε διάφορους τομείς, από την αναγνώριση εικόνων έως τις κοινωνικές επιστήμες [169–173]. Η επιτυχία τους οφείλεται στην αυξημένη διαθεσιμότητα δεδομένων, στις βελτιώσεις του λογισμικού και του υλικού, καθώς και σε διάφορα αλγοριθμικά επιτεύγματα που επιταχύνουν την εκπαίδευση των μοντέλων [174–180].

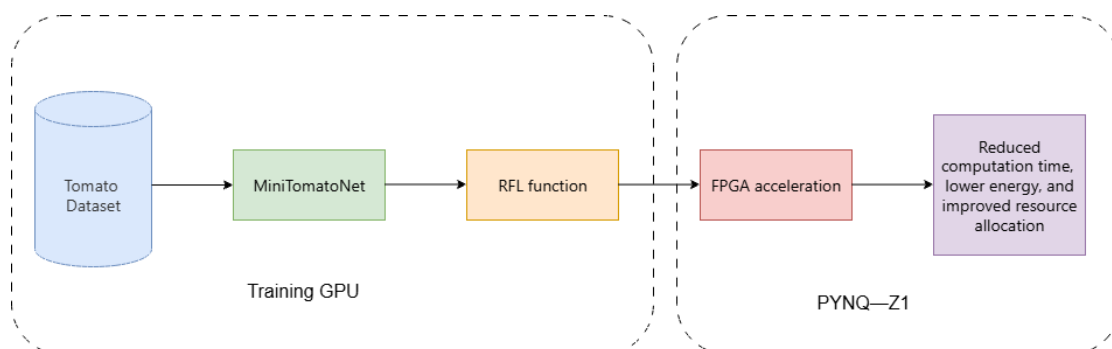
Κατά συνέπεια, με την εφαρμογή των CNNs στη γεωργία, οι αγρότες μπορούν να αναγνωρίσουν και να αντιμετωπίσουν τις ασθένειες των φυτών σε πρωιμότερα στάδια, αυξάνοντας τις αποδόσεις και μειώνοντας τον κίνδυνο απώλειας των καλλιεργειών. Επιπλέον, η εφαρμογή αυτών των τεχνολογιών στο γεωργικό περιβάλλον συμβάλλει στις βιώσιμες γεωργικές πρακτικές μειώνοντας τη χρήση φυτοφαρμάκων και χημικών. Επομένως, η ενσωμάτωση των CNNs στη γεωργία είναι ένα σημαντικό βήμα προς την αύξηση της φυτικής παραγωγής, ελαχιστοποιώντας ταυτόχρονα τις επιπτώσεις στο περιβάλλον [181–185].

Η μελέτη αυτή έχει κίνητρο, όπως φαίνεται στο Σχήμα 4.1, τις προκλήσεις που αντιμετωπίζει η καλλιέργεια τομάτας εξαιτίας της εμφάνισης ασθενειών. Η καινοτομία μας έγκειται στην ανάπτυξη του MiniTomatoNet, ενός μοντέλου CNN χαμηλής πολυπλοκότητας που έχει διαμορφωθεί ειδικά για την αποτελεσματική και ακριβή ταξινόμηση των ασθενειών στα φύλλα της τομάτας. Το προτεινόμενο μοντέλο έρχεται να αντιμετωπίσει τους περιορισμούς των συμβατικών τεχνικών στην γεωργική διάγνωση, οι οποίες συχνά υπολείπονται ως προς την αποτελεσματικότητα στην παρακολούθηση και την έγκαιρη ανίχνευση ασθενειών σε εκτεταμένες καλλιέργειες.

Επιπλέον, αντιμετωπίζουμε το πρόβλημα της ανισορροπίας των δεδομένων, το οποίο προκύπτει από την άνιση κατανομή των δειγμάτων μεταξύ των διαφορετικών κατηγοριών (όπως αποτυπώνεται στην Εικόνα 4.4), βελτιστοποιώντας τη συνάρτηση απώλειας, ώστε να εστιάζει περισσότερο στις υποεκπροσωπούμενες τάξεις, γεγονός που οδηγεί σε σημαντική αύξηση της ακρίβειας του μοντέλου. Αξιοποιώντας την τεχνολογία των FPGAs, στοχεύουμε να παρέχουμε μια ακριβή, αποτελεσματική και επεκτάσιμη λύση. Η επιλογή μας να επικεντρωθούμε στο FPGA ως συσκευή οφείλεται στην ικανότητά του να ενισχύει την υπολογιστική απόδοση των CNNs χωρίς να διακυβεύεται η ακρίβεια. Τα FPGAs προσφέρουν παράλληλες υπολογιστικές δυνατότητες που είναι ιδιαίτερα ωφέλιμες για την επεξεργασία των CNNs, επιτρέποντας την ανάλυση σε πραγματικό χρόνο για να υποστηρίξει έγκαιρες γεωργικές αποφάσεις διαχείρισης. Αυτό δεν μειώνει τη σημασία της ακρίβειας ταξινόμησης, αντίθετα, τη συμπληρώνει διασφαλίζοντας ότι η υψηλή ακρίβεια δεν συνεπάγεται με απαγορευτικούς χρόνους απόκρισης [186–188]. Παράλληλα, επιτρέπουν στοχευμένες βελτιστοποιήσεις υλικού, συμβάλλοντας στη σημαντική μείωση του χρόνου υπολογισμού και της κατανάλωσης ενέργειας. Ειδικότερα, στο πεδίο της διάγνωσης ασθενειών σε φυτά τομάτας, όπου η έγκαιρη και ενεργειακά αποδοτική ανάλυση μπορεί να οδηγήσει σε πρώιμη παρέμβαση και βέλτιστη κατανομή των πόρων, τα πλεονεκτήματα των FPGAs καθίστανται ιδιαίτερα σημαντικά [189, 190].

Αυτή η μελέτη παρουσιάζει τρεις συνεισφορές στην αναγνώριση των ασθενειών της τομάτας, οι οποίες μπορούν να συνοψιστούν ως εξής:

- Προτείνουμε ένα καινοτόμο, χαμηλής πολυπλοκότητας CNN για την ανίχνευση ασθενειών στα φύλλα τομάτας, το οποίο ονομάζεται MiniTomatoNet. Το MiniTomatoNet έχει σχεδιαστεί βάσει της μεθοδολογίας



Σχήμα 4.1: Διάγραμμα ροής εργασίας που απεικονίζει τα βήματα για την επίτευξη των στόχων αναγνώρισης ασθενειών στα φύλλα τομάτας.

που αναπτύξαμε στο Κεφάλαιο 3, και ειδικότερα σύμφωνα με τις μετρικές σχεδιασμού της Ενότητας 3.2. Το δίκτυο στοχεύει στην ταξινόμηση εικόνων φύλλων τομάτας σε δέκα κατηγορίες— εννέα διαφορετικών ασθενειών και μίας υγιούς κατάστασης. Η αρχιτεκτονική του είναι προσαρμοσμένη, ώστε να εξάγει πληροφορίες υψηλού επιπέδου από τις εικόνες εισόδου, οι οποίες είναι απαραίτητες για την ακριβή αναγνώριση και τη διάκριση των ασθενειών.

- Επιλύουμε το πρόβλημα της ανισορροπίας κλάσεων, το οποίο αποτελεί συχνό φαινόμενο σε σύνολα δεδομένων, υιοθετώντας τη συνάρτηση επανασταθμισμένης εστιακής απώλειας (RFL) αντί της παραδοσιακής συνάρτησης απώλειας διασταυρούμενης εντροπίας (CE). Η προσέγγιση αυτή διασφαλίζει αμερόληπτα αποτελέσματα ταξινόμησης, καθώς δίνει μεγαλύτερη βαρύτητα στις υποεκπροσωπούμενες κλάσεις και ελαχιστοποιεί τον αρνητικό αντίκτυπο της ανισορροπίας κλάσεων στην απόδοση του μοντέλου (βλ. Ενότητα 3.2.2.3).
- Επιταχύνουμε το MiniTomatoNet, το οποίο έχει σχεδιαστεί και υλοποιηθεί στην συσκευή Xilinx FPGA PYNQ—Z1. Αυτός ο σχεδιασμός αξιοποιεί τις εγγενείς δυνατότητες της παράλληλης επεξεργασίας της συσκευής, μέσω τεχνικών βελτιστοποίησης: ξετύλιγμα βρόχων, διασωλήνωση, κατάτμησης πινάκων και επιπέδωση βρόχων. Αυτές οι μεθοδολογικές βελτιώσεις ενισχύουν την αποτελεσματικότητα και την απόδοση της επεξεργασίας του MiniTomatoNet στο FPGA, διατηρώντας παράλληλα εξαιρετικά χαμηλή κατανάλωση ενέργειας.

Τα υπόλοιπα μέρη αυτού του κεφαλαίου είναι δομημένα ως εξής: Η ενότητα 4.3 εξετάζει τις υπάρχουσες προσεγγίσεις των CNNs που σχετίζονται με την αναγνώριση

των ασθενειών στα φύλλα τομάτας. Η ενότητα 4.4 παρουσιάζει το σύνολο δεδομένων εικόνων τομάτας, συζητά τις σχετικές έννοιες που χρησιμοποιούνται για την προτεινόμενη αρχιτεκτονική και αναλύει τη συνάρτηση επανασταθμισμένης εστιακής απώλειας. Στη συνέχεια, στην ενότητα 4.5, παρουσιάζουμε την αρχιτεκτονική του συστήματος, των αρθρωτών κυκλωμάτων του MiniTomatoNet και τις τεχνικές βελτιστοποίησης. Η ενότητα 4.6, παρουσιάζει τα πειραματικά αποτελέσματα που υποδεικνύουν την αποτελεσματικότητα της προτεινόμενης μεθοδολογίας καθώς τη συγκρίνουμε με άλλες μεθόδους και συσκευές. Στη συνέχεια, στην ενότητα 4.7 συζητάμε τα αποτελέσματα της έρευνάς μας. Τέλος, στην ενότητα 4.8 ολοκληρώνεται το κεφάλαιο και αναφέρει τις μελλοντικές εργασίες.

4.3 Επισκόπηση Επιστημονικού Πεδίου

Οι ασθένειες των καλλιεργειών τομάτας αποτελούν ένα σημαντικό πρόβλημα που αντιμετωπίζουν οι αγρότες παγκοσμίως, το οποίο οδηγεί σε σημαντικές απώλειες απόδοσης και χαμηλό περιθώριο κέρδους. Για την αντιμετώπιση αυτού του ζητήματος, οι περισσότεροι ερευνητές έχουν διεξάγει εκτεταμένες μελέτες για τον εντοπισμό μιας βέλτιστης λύσης με τη χρήση τεχνικών CNNs για την αναγνώριση ασθενειών σε γεωργικά περιβάλλοντα [191]. Σε αυτή την ενότητα, εξετάζουμε τις πιο πρόσφατες ακαδημαϊκές μελέτες σχετικά με την εφαρμογή των CNNs για την αναγνώριση και ταξινόμηση των ασθενειών που επηρεάζουν τις καλλιέργειες τομάτας.

Οι Elfatimi et al. [192] παρουσιάζουν μια προσέγγιση για την αυτόματη κατηγοριοποίηση ασθενειών φύλλων στις τομάτες χρησιμοποιώντας ένα βαθύ CNN πολλαπλής κλίμακας (Deep Multi-Scale CNN, DMCNN). Αυτή η προσέγγιση περιλαμβάνει 11.000 εικόνες για 10 διακριτές κατηγορίες συνθηκών φύλλων τομάτας και αξιοποιεί μεθόδους επαύξησης δεδομένων για να επεκτείνει τεχνητά το σύνολο δεδομένων. Το προτεινόμενο μοντέλο αναφέρει συνολική ακρίβεια 99,10% μετά από 100 επαναλήψεις. Οι συγγραφείς στο [193] παρουσιάζουν μια μέθοδο διάγνωσης ασθενειών φύλλων φυτών τομάτας χρησιμοποιώντας την αρχιτεκτονική DenseNet σε συνδυασμό με στρατηγικές επαύξησης δεδομένων για ενίσχυση της αναγνώρισης. Το σύνολο δεδομένων που χρησιμοποιήθηκε προέρχεται από το αποθετήριο PlantVillage, το οποίο αποτελείται από 18.160 εικόνες για τον εντοπισμό δέκα κατηγοριών ασθενειών που σχετίζονται με τα φυτά τομάτας. Το μοντέλο DenseNet έδωσε συνολική βαθμολογία ακρίβειας 95,40% μετά από 30 επαναλήψεις.

Στη μελέτη [194] συγκρίθηκαν δύο διαφορετικές τροποποιημένες αρχιτεκτονικές δικτύου (VGG16 και VGG19) για την αναγνώριση ασθενειών των φυτών τομάτας σε δέκα κατηγορίες. Η μελέτη χρησιμοποιεί 18.160 εικόνες φύλλων φυτών τομάτας από το σύνολο δεδομένων PlantVillage, οι οποίες υποβάλλονται σε προεπεξεργασία χρησιμοποιώντας στρατηγικές επαύξησης δεδομένων. Το δίκτυο VGG16 πέτυχε ακρίβεια 92,00%, ενώ το δίκτυο VGG19 πέτυχε ακρίβεια 95,00% μετά από 100 επαναλήψεις. Επίσης, στο [195], εξετάστηκε η χρήση των πλήρως συνελκτικών δικτύων διπλής διαδρομής με εναλλασσόμενη κανονικοποίηση (FC—SNDPN) για την αναγνώριση των ασθενειών των φυτών της τομάτας σε δέκα κατηγορίες. Η εργασία αυτή χρησιμοποιεί 18.160 εικόνες φύλλων φυτών τομάτας από το σύνολο δεδομένων PlantVillage, οι οποίες υποβάλλονται σε προεπεξεργασία χρησιμοποιώντας μεθόδους επαύξησης δεδομένων. Το μοντέλο FC—SNDPN παρουσίασε ακρίβεια 97,59% μετά από 10 επαναλήψεις.

Οι Ahmed et al. [196] παρουσιάζουν μια μεθοδολογία για τον εντοπισμό ασθενειών στα φυτά τομάτας χρησιμοποιώντας το μοντέλο MobileNetV2, που συμπληρώνεται από μεθόδους αύξησης δεδομένων. Η μελέτη χρησιμοποιεί 18.160 εικόνες φύλλων τομάτας από το σύνολο δεδομένων PlantVillage και στοχεύει να αναγνωρίσει δέκα κατηγορίες, συμπεριλαμβανομένων εννέα ασθενειών και μίας υγιούς κατηγορίας. Το μοντέλο MobileNetV2 πέτυχε συνολική ακρίβεια 99,30% μετά από 1.000 επαναλήψεις. Ομοίως, στη μελέτη [197] έχουν σχεδιάσει μια προσαρμοσμένη αρχιτεκτονική που αποτελείται από έξι στρώματα για να αναγνωρίζει δέκα παθήσεις φυτών τομάτας — εννέα παθολογικές και μία υγιούς κατάσταση. Το σύνολο δεδομένων που χρησιμοποίησαν ήταν 18.160 εικόνες από φύλλα τομάτας από το σύνολο δεδομένων PlantVillage. Το δίκτυό τους πέτυχε ακρίβεια 98,44% και η βαθμολογία F1 είναι 88,45% μετά από 100 επαναλήψεις.

Οι συγγραφείς στο [198] παρουσιάζουν την εκπαίδευση και την αξιολόγηση τεσσάρων πρόσφατων μοντέλων CNNs (MobileNetV3, NasNetMobile, MobileNetV2 και Xception) για την ταξινόμηση ασθενειών στα φύλλα τομάτας σε δέκα κατηγορίες σε συσκευές χαμηλής κατανάλωσης. Το σύνολο δεδομένων που χρησιμοποίησαν ήταν 18.215 δεδομένα εικόνας φύλλων τομάτας από το PlantVillage και με την τεχνική της αύξησης δεδομένων, ολόκληρο το σύνολο δεδομένων αυξήθηκε στις 109.290 εικόνες. Στη συνέχεια, το επαυξημένο σύνολο δεδομένων χωρίστηκε σε 70% για εκπαίδευση και 30% για σκοπούς αξιολόγησης. Μετά από 10 επαναλήψεις στο επαυξημένο σύνολο δεδομένων δοκιμής, το MobileNetV3 πέτυχε ακρίβεια 98,00%, το NasNetMobile πέτυχε ακρίβεια 84,00%, το MobileNetV2 πέτυχε ακρίβεια 75,00%

και το Xception πέτυχε ακρίβεια 100,00%. Στο [199] συνέκριναν την αρχιτεκτονική DenseNet121 με και χωρίς την μέθοδο της αύξησης δεδομένων για την αναγνώριση ασθενειών στα φυτά τομάτας σε δέκα κατηγορίες. Τα δεδομένα εικόνας των φύλλων τομάτας συλλέγονται από το σύνολο δεδομένων PlantVillage και αποτελείται από 16.012 εικόνες. Το μοντέλο DenseNet121 απέδωσε ακρίβεια 97,11% με την χρήση της αύξησης δεδομένων και 94,34% χωρίς την αύξηση δεδομένων μετά από 100 επαναλήψεις.

Ταυτόχρονα, στο [200] χρησιμοποίησαν δύο αρχιτεκτονικές την Attention—embedded Residual CNN και την Residual CNN στο σύνολο δεδομένων εικόνας PlantVillage για να ταξινομήσουν τις ασθένειες των φυτών τομάτας σε τέσσερις κατηγορίες (τρεις ασθένειες και μία υγιή). Στο σύνολο δεδομένων εικόνων τομάτας από το PlantVillage, το μοντέλο Residual CNN πέτυχε ακρίβεια 95,00% μετά από 140 επαναλήψεις, ενώ το μοντέλο Attention—embedded Residual CNN πέτυχε ακρίβεια 98,00% μετά από 140 επαναλήψεις. Στο [201] σχεδιάστηκε η αρχιτεκτονική ToLeD που αποτελείται από τρία συνελκτικά επίπεδα και τρία επίπεδα συγκέντρωσης ακολουθούμενα από δύο πλήρως συνδεδεμένα επίπεδα για την αναγνώριση δέκα ασθενειών των φυτών τομάτας (εννέα ασθένειες και μία υγιής). Το σύνολο δεδομένων που χρησιμοποίησαν ήταν 17.500 εικόνες από φύλλα τομάτας από το PlantVillage. Το δίκτυο ToLeD πέτυχε ακρίβεια 91,20% μετά από 1.000 επαναλήψεις.

Επίσης, στο [202], αξιολογήθηκε το ResNet χρησιμοποιώντας μια ποικιλία τεχνικών αύξησης δεδομένων, όπως τυχαία υποδειγματοληψία, τυχαία υπερδειγματοληψία, SMOTE και ένα GAN στο σύνολο δεδομένων εικόνας PlantVillage για να ταξινομήσουν τις ασθένειες των φυτών τομάτας σε δύο κατηγορίες (μία ασθένεια και μία υγιή), χρησιμοποιώντας ένα σύνολο δεδομένων που περιλαμβάνει 1.960 εικόνες. Η εκπαίδευση του GAN ήταν 200.000 επαναλήψεις και το μοντέλο ResNet πέτυχε ακρίβεια 85,56% μετά από 3.000 επαναλήψεις. Η εκτεταμένη εκπαίδευση του GAN με συνολικά 200.000 επαναλήψεις υπογραμμίζει την πολυπλοκότητα της δημιουργίας ρεαλιστικών δειγμάτων δεδομένων σε συνθήκες με μεγάλες ανισορροπίες κλάσεων.

Οι συγγραφείς στο [203] ανέπτυξαν δύο διαφορετικές αρχιτεκτονικές δικτύων DL (S—CNN και F—CNN) με σκοπό την ανίχνευση ασθενειών φυτών τομάτας σε δέκα κατηγορίες. Τα δεδομένα εικόνων φύλλων τομάτας προήλθαν από το σύνολο δεδομένων PlantVillage, ενώ οι προτεινόμενες αρχιτεκτονικές αξιολογήθηκαν σε εικόνες από το διαδίκτυο. Το μοντέλο F—CNN πέτυχε ακρίβεια 42,30%, ενώ το μοντέλο S—CNN σημείωσε ακρίβεια 98,60%. Επιπλέον, στο [204], οι

συγγραφείς πρότειναν την αρχιτεκτονική τριών καναλιών στο CNN (TCCNN) για την αναγνώριση ασθενειών φυτών τομάτας σε οχτώ κατηγορίες. Τα δεδομένα εικόνων φύλλων τομάτας προέρχονταν από το σύνολο δεδομένων PlantVillage, που περιλάμβανε 15.817 εικόνες. Το μοντέλο TCCNN πέτυχε ακρίβεια 91,15% μετά από 50 επαναλήψεις.

Ταυτόχρονα, στη μελέτη [205], συγκρίθηκαν δύο διαφορετικές αρχιτεκτονικές δικτύων (Faster R—CNN με ResNet101 και Mask R—CNN με ResNet101) για την ανίχνευση ασθενειών φυτών τομάτας σε έντεκα κατηγορίες. Συνολικά 286 εικόνες φύλλων τομάτας συλλέχθηκαν από το διαδίκτυο. Το δίκτυο Faster R—CNN με ResNet101 πέτυχε συνολική ακρίβεια 88,53% μετά από 70.000 επαναλήψεις, ενώ το δίκτυο Mask R—CNN με ResNet101 πέτυχε ακρίβεια 99,64% μετά από 500 επαναλήψεις. Ο χρόνος εκπαίδευσης για το μοντέλο Faster R—CNN με ResNet101 ήταν 12,52 ώρες, ενώ για την αρχιτεκτονική Mask R—CNN με ResNet101 ανήλθε σε 23,82 ώρες. Επιπλέον, στο [206], εξετάστηκαν δύο διαφορετικές αρχιτεκτονικές δικτύων (Faster R—CNN και AlexNet) για την αναγνώριση ασθενειών φυτών τομάτας σε τέσσερις κατηγορίες. Τα δεδομένα αποτελούνταν από 4.923 εικόνες φύλλων τομάτας, που συλλέχθηκαν υπό ελεγχόμενες συνθήκες. Το μοντέλο Faster R—CNN πέτυχε συνολική ακρίβεια 80,00% μετά από 50 επαναλήψεις, ενώ η αρχιτεκτονική AlexNet σημείωσε ακρίβεια 95,75% μετά από 65 επαναλήψεις.

Επίσης, στο [207], οι Sardogan et al. εφάρμοσαν ένα CNN σε συνδυασμό με τον αλγόριθμο Learning Vector Quantization, επιτυγχάνοντας συνολική ακρίβεια 86,00% στην αναγνώριση πέντε κατηγοριών ασθενειών τομάτας, χρησιμοποιώντας 500 εικόνες από το σύνολο δεδομένων PlantVillage μετά από 300 επαναλήψεις. Στο [208], συγκρίθηκαν οι αρχιτεκτονικές GoogLeNet, AlexNet και ResNet με τη χρήση των βελτιστοποιητών SGD και Adam, σε 5.550 εικόνες από το σύνολο PlantVillage. Η αρχιτεκτονική ResNet με τον βελτιστοποιητή SGD πέτυχε μέγιστη ακρίβεια 97,28% μετά από 4.992 επαναλήψεις.

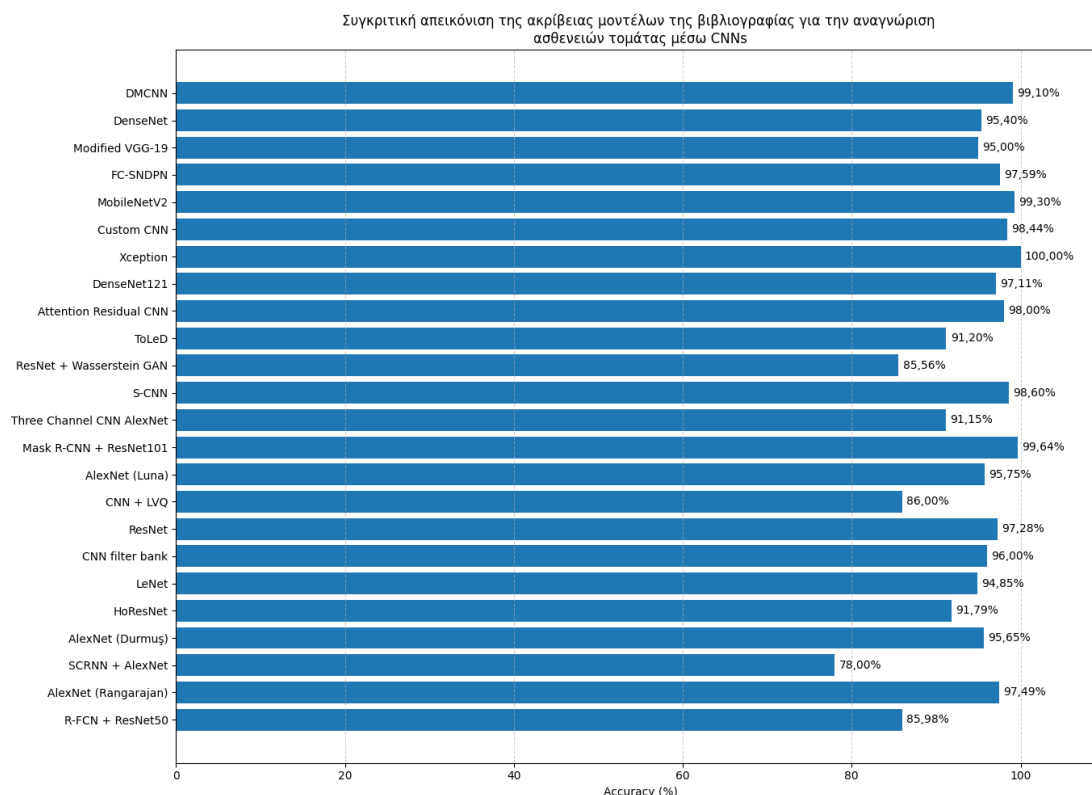
Οι συγγραφείς στο [209], εξέτασαν μια αρχιτεκτονική φίλτρων CNN filter bank στο σύνολο δεδομένων Tomato Diseases and Pest, επιτυγχάνοντας ακρίβεια 96,00% στην ανίχνευση ασθενειών σε δέκα κατηγορίες, μετά από 20.000 επαναλήψεις. Οι συγγραφείς του [210] διερεύνησαν την αρχιτεκτονική LeNet για την ανίχνευση ασθενειών σε δέκα κατηγορίες, χρησιμοποιώντας 18.160 εικόνες από το σύνολο PlantVillage, επιτυγχάνοντας ακρίβεια 94,85% μέσα σε 30 επαναλήψεις. Επιπλέον, στο [211], προτάθηκε η αρχιτεκτονική High—Order Residual CNN (HOREsNet) για την ανίχνευση ασθενειών σε έξι κατηγορίες, χρησιμοποιώντας 10.478 εικόνες από

το σύνολο PlantVillage, επιτυγχάνοντας συνολική ακρίβεια 91,79% μετά από 100 επαναλήψεις.

Οι Yamamoto et al. [212] πρότειναν το Super-Resolution CNN (SRCNN), βασισμένο στην αρχιτεκτονική AlexNet, χρησιμοποιώντας 18.149 εικόνες από το σύνολο δεδομένων PlantVillage για την ανίχνευση των ασθενειών τομάτας σε εννέα κατηγορίες, επιτυγχάνοντας ακρίβεια 78,00% μετά από 50 επαναλήψεις. Οι συγγραφείς στο [213] εξέτασαν τα προεκπαιδευμένα δίκτυα AlexNet και VGG16 για την ανίχνευση ασθενειών σε έξι κατηγορίες με χρήση 13.262 εικόνων από το PlantVillage, όπου το AlexNet πέτυχε ακρίβεια 97,49% μετά από 10 επαναλήψεις. Επιπλέον, στο [214], εξετάστηκε η αρχιτεκτονική Region-Based Fully Convolutional Network (R-FCN) σε συνδυασμό με το μοντέλο ResNet50, το οποίο δοκιμάστηκε σε 5.000 εικόνες από αγροτικές εκτάσεις, επιτυγχάνοντας ακρίβεια 85,98% μετά από 200.000 επαναλήψεις στην ανίχνευση ασθενειών σε εννέα κατηγορίες.

Στον Πίνακα 4.1 συνοψίζονται οι διάφορες προσεγγίσεις των CNNs που έχουν εφαρμοστεί για τη διάγνωση ασθενειών τομάτας, ενώ στο Σχήμα 4.2 παρουσιάζεται η αναλυτική συγκριτική απεικόνιση της συνολικής ακρίβειας εξαγωγής συμπερασμάτων που επιτυγχάνεται από τις διαφορετικές αρχιτεκτονικές. Η ανασκόπηση της βιβλιογραφίας αποδεικνύει τη δυνατότητα των CNNs να αναγνωρίζουν με ακρίβεια τις ασθένειες της τομάτας. Ωστόσο, τα υφιστάμενα μοντέλα CNNs, αν και ισχυρά, απαιτούν βελτιστοποίηση ως προς την ενεργειακή και υπολογιστική τους αποδοτικότητα, κάτι που είναι κρίσιμο για αγροτικές εφαρμογές με περιορισμένους πόρους. Επιπλέον, ένα σημαντικό ερευνητικό κενό προκύπτει από την ανεπαρκή αντιμετώπιση του προβλήματος της ανισορροπίας των κατηγοριών στα δεδομένα εκπαίδευσης. Η έλλειψη αντιμετώπισης αυτού του ζητήματος οδηγεί σε διαγνωστική μεροληψία, που ευνοεί τις πιο συχνά εμφανιζόμενες ασθένειες, παραμελώντας τις πιο σπάνιες αλλά εξίσου κρίσιμες παθήσεις.

Στην παρούσα μελέτη, προτείνουμε το MiniTomatoNet για την κάλυψη αυτών των κενών. Το MiniTomatoNet διακρίνεται για την ικανότητά του να αναγνωρίζει με ακρίβεια ένα ευρύ φάσμα κατηγοριών ασθενειών τομάτας, χωρίς να απαιτεί παρατεταμένες περιόδους εκπαίδευσης ή την εισαγωγή περίπλοκων στοιχείων στην αρχιτεκτονική του. Η αρχιτεκτονική του μοντέλου έχει βελτιστοποιηθεί για FPGAs, επιτρέποντας ταχύτερη και πιο ενεργειακά αποδοτική αναγνώριση ασθενειών, ιδανική για αγροτικά περιβάλλοντα με περιορισμένους πόρους. Επιπλέον, η ενσωμάτωση της συνάρτησης επανασταθμισμένης εστιακής απώλειας στο



Σχήμα 4.2: Συγκριτική απεικόνιση της συνολικής ακρίβειας που επιτεύχθηκε από διάφορα αρχιτεκτονικά σχήματα CNNs, όπως έχουν παρουσιαστεί στη διεθνή βιβλιογραφία, για την αναγνώριση ασθενειών τομάτας.

MiniTomatoNet προσφέρει μια νέα προσέγγιση για τη διαχείριση της ανισορροπίας των κατηγοριών, εξασφαλίζοντας ότι ακόμη και οι υποεκπροσωπούμενες κατηγορίες αναγνωρίζονται με υψηλή ακρίβεια. Με αυτόν τον τρόπο, το μοντέλο διατηρεί την αμεροληψία του, αποτρέποντας την υπερβολική έμφαση σε συχνότερες ασθένειες εις βάρος των σπανιότερων αλλά εξίσου σημαντικών καταστάσεων. Αυτή η ενίσχυση στη φάση εκπαίδευσης εξασφαλίζει ότι το MiniTomatoNet παραμένει ισορροπημένο στο πλήρες φάσμα παθολογιών των φυτών τομάτας.

4.4 Υλικά και Μέθοδοι

4.4.1 Σύνολο Δεδομένων Εικόνας

Τα δεδομένα εικόνας σχετικά με τις ασθένειες των καλλιεργειών τομάτας που χρησιμοποιήθηκαν στα πειράματα προέρχονται από το σύνολο δεδομένων PlantVillage [216]. Το PlantVillage περιλαμβάνει περισσότερες από 50.000 εικόνες

Πίνακας 4.1: Σύνοψη της ερευνητικής βιβλιογραφίας σε προσεγγίσεις CNNs για την αναγνώριση των ασθενειών τομάτας.

Paper	Year	Best model	Dataset	Classes	Epochs	Best accuracy
[192]	2023	DMCNN	Tomato PlantVillage Subset 11.000 Images	10	100	99,10%
[193]	2023	DenseNet	Tomato PlantVillage Subset 18.160 Images	10	30	95,40%
[194]	2023	Modified VGG—19	Tomato PlantVillage Subset 18.160 Images	10	100	95,00%
[195]	2023	FC—SNDPN	Tomato PlantVillage Subset 18.160 Images	10	10	97,59%
[196]	2022	MobileNetV2	Tomato PlantVillage Subset 18.160 Images	10	1.000	99,30%
[197]	2022	Custom CNN	Tomato PlantVillage Subset 18.160 Images	10	100	98,44%
[198]	2021	Xception	Tomato PlantVillage Subset 18.215 Images	10	10	100,00%
[199]	2021	DenseNet121	Tomato PlantVillage Subset 16.012 Images	10	100	97,11%
[200]	2020	Attention embedded Residual CNN model	Tomato PlantVillage Subset	4	140	98,00%
[201]	2020	ToLeD	Tomato PlantVillage Subset 17.500 Images	10	1.000	91,20%
[202]	2020	ResNet with Wasserstein GAN	Tomato PlantVillage Subset 1.960 Images	2	3.000	85,56%
[203]	2020	S—CNN	Tomato PlantVillage Subset 15.817 Images	10	—	98,60%
[204]	2019	Three Channel CNN AlexNet	Tomato PlantVillage Subset 15.817 Images	8	50	91,15%
[205]	2019	Mask R—CNN with ResNet101	Customized 286 Images	11	500	99,64%
[206]	2018	AlexNet	Customized 4.923 Images	4	65	95,75%
[207]	2018	CNN with Learning Vector Quantization algorithm	Tomato PlantVillage Subset 500 Images	5	300	86,00%
[208]	2018	ResNet	Tomato PlantVillage Subset 5.550 Images	9	4.992	97,28%
[209]	2018	CNN filter bank	Customized 8.927 Images	10	20.000	96,00%
[210]	2018	LeNet	Tomato PlantVillage Subset 18.160 Images	10	30	94,85%
[211]	2018	HoResNet	Tomato PlantVillage Subset 10.478 Images	6	100	91,79%
[215]	2018	AlexNet	Tomato PlantVillage Subset	10	—	95,65%
[212]	2017	SCRNN with AlexNet	Tomato PlantVillage Subset 18.149 Images	9	50	78,00%
[213]	2018	AlexNet	Tomato PlantVillage Subset 13.262 Images	6	10	97,49%
[214]	2017	R—FCN with ResNet50	Customized 5.000 Images	9	200.000	85,98%

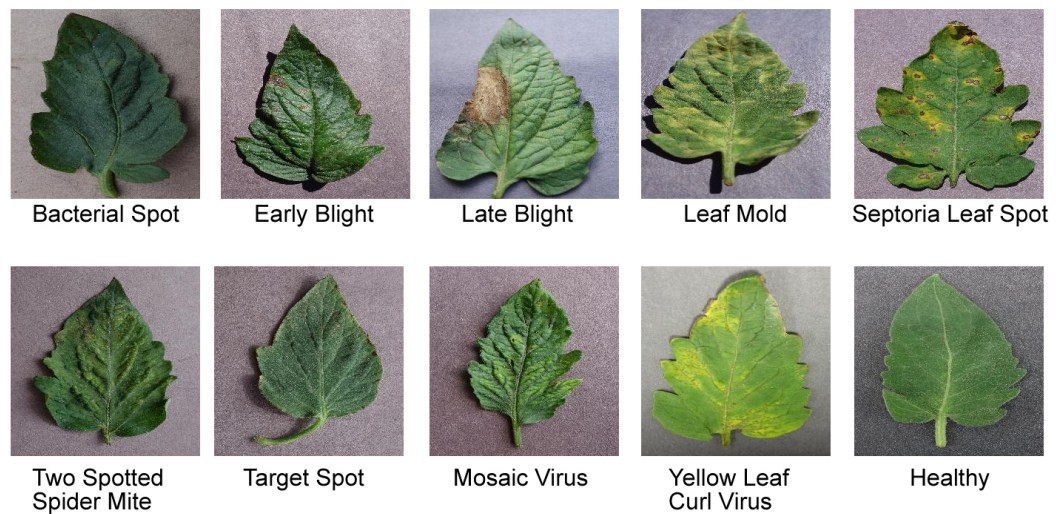
που αντιπροσωπεύουν δεκατέσσερα διαφορετικά είδη γεωργικών καλλιεργειών, όπως ντομάτες, μήλα, σταφύλια, πατάτες, καλαμπόκι και σόγια. Κάθε εικόνα του συνόλου δεδομένων έχει μέγεθος 256×256 εικονοστοιχείων. Συγκεκριμένα, το σύνολο δεδομένων για τις ντομάτες περιέχει συνολικά 18.160 εικόνες, από τις οποίες οι 14.525 (80%) ανήκουν στο σύνολο εκπαίδευσης και οι 3.635 (20%) στο σύνολο ελέγχου. Το σύνολο εκπαίδευσης χωρίζεται περαιτέρω σε σύνολο εκπαίδευσης (60%) και σύνολο επικύρωσης (20%).

Το σύνολο δεδομένων για τις ασθένειες των φύλλων τομάτας αποτελείται από δέκα διαφορετικές κατηγορίες, στις οποίες περιλαμβάνονται: 952 εικόνες με μούχλα φύλλων (leaf mold), 1.771 εικόνες με κηλίδες φύλλων σηπτορίας (septoria leaf spot), 1.909 με όψιμη προσβολή (late blight), 2.127 με βακτηριακές κηλίδες (bacterial spot), 1.000 με πρόωμη προσβολή (early blight), 5.357 με τον ιό κιτρινίσματος των φύλλων (yellow leaf curl virus), 373 με τον ιό του μωσαϊκού (mosaic virus), 1.404 με στόχο κηλίδας (target spot), 1.676 με ακάρεα δύο σημείων (two spotted spider mite) και 1.591 εικόνες υγιών φύλλων (healthy).

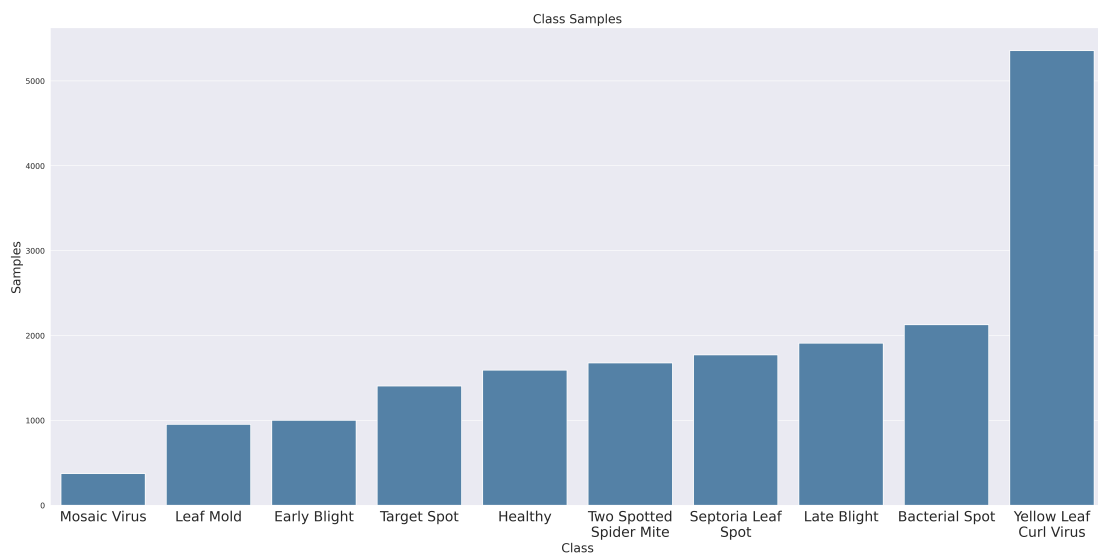
Στον Πίνακα 4.2 παρουσιάζεται η κατανομή των εννέα διαφορετικών κατηγοριών ασθενειών που επηρεάζουν τα φύλλα της τομάτας, καθώς και η κατηγορία των υγιών φύλλων. Στην Εικόνα 4.3 απεικονίζει ένα δείγμα εικόνας από κάθε κατηγορία, ενώ στην Εικόνα 4.4 παρουσιάζεται η συχνότητα των κατηγοριών στο σύνολο δεδομένων τομάτας. Όπως παρατηρείται στην Εικόνα 4.4, τα δεδομένα είναι μη ισορροπημένα, με αποτέλεσμα οι κατηγορίες να εμφανίζουν διαφορετικά μοτίβα κατανομής. Αυτό το φαινόμενο μπορεί να οδηγήσει σε μεροληψία κατά την εκπαίδευση του μοντέλου, δίνοντας προτεραιότητα στις συχνότερες κατηγορίες εις βάρος των σπανιότερων, οι οποίες ενδέχεται να είναι εξίσου κρίσιμες.

4.4.2 Αρχιτεκτονική Δικτύου MiniTomatoNet

Ο αρχικός στόχος στην ανάπτυξη του MiniTomatoNet είναι να διασφαλιστεί ότι το μοντέλο επιτυγχάνει υψηλό ποσοστό ακρίβειας στον εντοπισμό της νόσου στα φύλλα της τομάτας, διατηρώντας ταυτόχρονα μια βελτιστοποιημένη αρχιτεκτονική, έτσι ώστε ακόμη και FPGAs με περιορισμένους πόρους να μπορούν να ανταποκρίνονται επαρκώς στις υπολογιστικές απαιτήσεις και τις ανάγκες αποθήκευσης του μοντέλου [217–219]. Προκειμένου να ελαχιστοποιηθούν οι υπολογιστικές απαιτήσεις και οι παράμετροι του μοντέλου, το MiniTomatoNet



Σχήμα 4.3: Ενδεικτικά δείγματα διαφορετικών φύλλων τομάτας. Ένα φύλλο για κάθε ασθένεια και ένα υγιές φύλλο τομάτας.



Σχήμα 4.4: Συχνότητα κάθε κλάσης στο σύνολο δεδομένων τομάτας. Η σημαντική διαφοροποίηση στο πλήθος δειγμάτων μεταξύ των κλάσεων αναδεικνύει την ύπαρξη ανισορροπίας στο σύνολο δεδομένων.

υιοθέτησε τη συνέλιξη βάθους ανά κανάλι αντί για τις παραδοσιακές συνελικτικές μεθόδους.

Η διαδικασία συνέλιξης βάθους ανά κανάλι [220] χωρίζει τη συνέλιξη σε δύο διακριτά στάδια: κατά βάθος και σημειακή συνέλιξη. Στη συνέλιξη σε βάθος, κάθε ξεχωριστό κανάλι εισόδου υφίσταται συνέλιξη χρησιμοποιώντας έναν πυρήνα ενός καναλιού και τα αποτελέσματα στη συνέχεια συγχωνεύονται για να σχηματίσουν μια

Πίνακας 4.2: Αναλυτική κατανομή του αριθμού εικόνων που χρησιμοποιήθηκαν για εκπαίδευση και έλεγχο κατά τη διαδικασία διαχωρισμού βάσει της μεθόδου συγκράτησης.

Class	Number of images	Train/Validation set	Test set
Bacterial spot	2.127	1.702	425
Yellow leaf curl virus	5.357	4.287	1.070
Late blight	1.909	1.529	380
Two spotted spider mite	1.676	1.340	336
Septoria leaf Spot	1.771	1.415	356
Target spot	1.404	1.124	280
Healthy	1.591	1.271	320
Leaf mold	952	760	192
Early blight	1.000	800	200
Mosaic virus	373	297	76
Total	18.160	14.525	3.635

ενδιάμεση έξοδο. Στη συνέχεια, η σημειακή συνέλιξη περιλαμβάνει την εφαρμογή πολλαπλών πυρήνων 1×1 σε αυτήν την ενδιάμεση έξοδο. Όταν αντιπαρατίθεται με την τυπική συνέλιξη, έχοντας τις ίδιες διαστάσεις εισόδου και εξόδου, η διαχωριστική συνέλιξη προσφέρει μείωση της υπολογιστικής πολυπλοκότητας και των παραμέτρων σχεδόν 9 φορές λιγότερο [109]. Έτσι, αυτό έχει ως αποτέλεσμα την επιτάχυνση του χρόνου εκπαίδευσης και συμπερασμάτων του MiniTomatoNet.

Το MiniTomatoNet σχεδιάστηκε ακολουθώντας την μεθοδολογία που παρουσιάστηκε στο Κεφάλαιο 3, και ειδικότερα βάσει των μετρικών σχεδιασμού (βλ. Ενότητα 3.2). Όπως απεικονίζεται στο Σχήμα 4.5, η αρχιτεκτονική περιλαμβάνει συνολικά οχτώ επίπεδα, τρία διαχωριστικά συνελκτικά επίπεδα, τρία επίπεδα συγκέντρωσης, και δύο πλήρως συνδεδεμένα επίπεδα. Η επιλογή τριών διαχωριστικών συνελκτικών στρωμάτων αντανακλά τον στόχο της μεθοδολογίας μας για επίτευξη επαρκούς βάθους, ικανής εξαγωγής αφηρημένων χαρακτηριστικών, χωρίς υπερβολική αύξηση της υπολογιστικής πολυπλοκότητας (βλ. Ενότητα 3.2.1.1). Η χρήση διαχωριστικών συνελίξεων (βλ. Ενότητα 3.2.1.2) μειώνει σημαντικά τον αριθμό παραμέτρων και το υπολογιστικό φορτίο, ακολουθώντας τις βέλτιστες πρακτικές σχεδιασμού που προτείνει η μεθοδολογία μας (βλ. Ενότητα 3.4.3).

Το MiniTomatoNet δέχεται ως είσοδο εικόνες διαστάσεων 128×128 εικονοστοιχείων. Το πρώτο συνελκτικό επίπεδο περιλαμβάνει 18 φίλτρα διαστάσεων 9×9 , ακολουθούμενο από ένα επίπεδο συγκέντρωσης με παράθυρο 4×4 . Το δεύτερο συνελκτικό επίπεδο περιέχει 36 φίλτρα διαστάσεων 5×5 και ακολουθείται από ένα επίπεδο συγκέντρωσης, παραθύρου 3×3 . Το τρίτο συνελκτικό στάδιο είναι

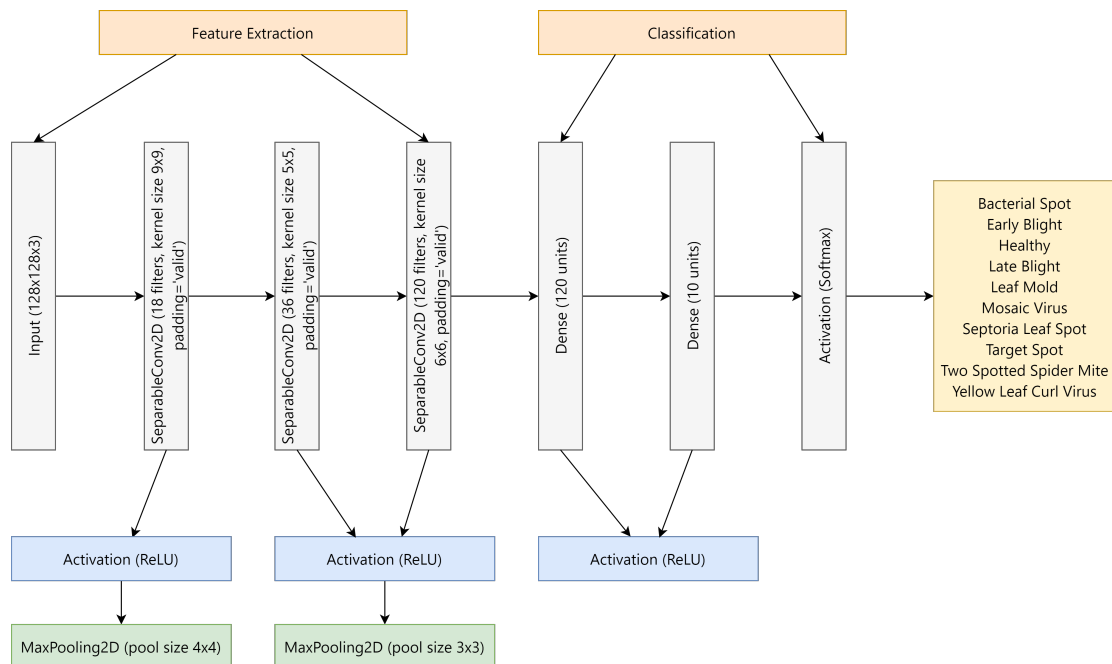
όμοια δομημένο, με στόχο την περαιτέρω συμπίεση των χωρικών διαστάσεων της εικόνας και την εξαγωγή πιο αφηρημένων χαρακτηριστικών. Στη συνέχεια, η αρχιτεκτονική περιλαμβάνει δύο πλήρως συνδεδεμένα στρώματα. Το προτελευταίο επίπεδο αποτελείται από 120 νευρώνες, ενώ το τελικό επίπεδο έχει 10 νευρώνες με συνάρτηση ενεργοποίησης softmax, η οποία είναι κατάλληλη για την ταξινόμηση πολλαπλών κλάσεων.

Η στρατηγική συνέλιξης σε βάθος μεγιστοποιεί την αποδοτικότητα μέσω της μείωσης των απαιτήσεων επεξεργασίας, διασφαλίζοντας ότι το μοντέλο MiniTomatoNet διατηρεί χαμηλό αποτύπωμα, γεγονός που το καθιστά ιδανικό για υλοποίηση σε περιβάλλοντα με περιορισμένους πόρους ή σε εφαρμογές που απαιτούν ταχεία και αποδοτική λειτουργία. Μια ολοκληρωμένη ανάλυση των προδιαγραφών του MiniTomatoNet παρουσιάζεται στον Πίνακα 4.3. Στον Πίνακα 4.3, το βήμα μετατόπισης (stride) του πυρήνα κατά τη συνελκτική πράξη επηρεάζει τις χωρικές διαστάσεις της εξόδου κάθε επιπέδου. Οι παράμετροι μετατόπισης (biases) προστίθενται σε κάθε φίλτρο εξόδου στα συνελκτικά επίπεδα και σε κάθε νευρώνα στα πλήρως συνδεδεμένα επίπεδα, μετατοπίζοντας τη γραμμική έξοδο πριν από την εφαρμογή της συνάρτησης ενεργοποίησης. Τα εκπαιδευσιμα βάρη (weights) καθορίζονται από το μέγεθος του πυρήνα (kernel size), τον αριθμό των καναλιών εισόδου και το πλήθος των φίλτρων εξόδου. Το συνολικό πλήθος παραμέτρων ανά επίπεδο (Param #) προκύπτει από το άθροισμα των βαρών και των μετατοπίσεων του αντίστοιχου επιπέδου.

Για παράδειγμα, στο επίπεδο Conv1 του MiniTomatoNet, στο στάδιο της κατά βάθος συνέλιξης εφαρμόζονται 3 φίλτρα 9×9 , ένα για κάθε κανάλι, οδηγώντας σε $3 \times 9 \times 9 = 243$ βάρη (weights). Στη συνέχεια, στο στάδιο της σημειακής συνέλιξης εφαρμόζονται 18 φίλτρα 1×1 σε όλα τα κανάλια εισόδου, δηλαδή $3 \times 18 = 54$ πρόσθετα βάρη. Συνολικά, το επίπεδο περιλαμβάνει $243 + 54 = 297$ εκπαιδευσιμα βάρη και 18 παραμέτρους μετατόπισης (biases), μία για κάθε φίλτρο εξόδου, με αποτέλεσμα $297 + 18 = 315$ συνολικές παραμέτρους (Param #), όπως παρουσιάζεται στον Πίνακα 4.3.

4.4.3 Συνάρτηση Επανασταθμισμένης Εστιακής Απώλειας

Το πρόβλημα της ανισορροπίας κλάσεων αποτελεί μία σημαντική πρόκληση στις εργασίες ταξινόμησης [221, 222], καθώς συχνά οδηγεί σε μη βέλτιστη απόδοση των μοντέλων, ειδικά για τις λιγότερο εκπροσωπούμενες κατηγορίες [223]. Το φαινόμενο



Σχήμα 4.5: Δομική απεικόνιση του MiniTomatoNet για την αναγνώριση ασθενειών της τομάτας.

Layer (type)	Output Size	Kernel Size	Stride	Weights	Biases	Param #
Input Layer	$128 \times 128 \times 3$	—	—	—	—	—
Conv1 (SeparableConv2D)	$120 \times 120 \times 18$	9×9	1	297	18	315
Pool1 (MaxPooling2D)	$30 \times 30 \times 18$	4×4	4	—	—	—
Conv2 (SeparableConv2D)	$26 \times 26 \times 36$	5×5	1	1098	36	1134
Pool2 (MaxPooling2D)	$8 \times 8 \times 36$	3×3	3	—	—	—
Conv3 (SeparableConv2D)	$3 \times 3 \times 120$	6×6	1	5616	120	5736
Pool3 (MaxPooling2D)	$1 \times 1 \times 120$	3×3	3	—	—	—
FC1 (Dense)	120	—	—	14,400	120	14,520
FC2 (Dense)	10	—	—	1200	10	1210
Total	—	—	—	22,611	304	22,915

Πίνακας 4.3: Οι προδιαγραφές της αρχιτεκτονικής MiniTomatoNet για την αναγνώριση ασθενειών της τομάτας.

αυτό είναι γνωστό ότι προκαλεί δυσκολίες στην επίτευξη υψηλής ακρίβειας, δεδομένου ότι τα μοντέλα τείνουν να ευνοούν τις πολυπληθέστερες κλάσεις [224].

Μία ευρέως χρησιμοποιούμενη στρατηγική για την αντιμετώπιση της ανισορροπίας κλάσεων είναι η επαύξηση δεδομένων. Η τεχνική αυτή περιλαμβάνει την επέκταση του συνόλου δεδομένων εκπαίδευσης μέσω διαφόρων μετασχηματισμών και τροποποιήσεων των αρχικών δειγμάτων. Παρότι η μέθοδος αυτή συμβάλλει στη βελτίωση της ανθεκτικότητας των μοντέλων, είναι απαραίτητο να τονιστεί ότι αυξάνει επίσης το υπολογιστικό φορτίο, επεκτείνοντας τον απαιτούμενο χρόνο εκπαίδευσης [225].

Η αναγκαιότητα εφαρμογής κατάλληλων στρατηγικών γίνεται ακόμη πιο επιτακτική σε περιπτώσεις έντονης ανισορροπίας, όπως αυτή που παρατηρείται στο σύνολο δεδομένων της παρούσας έρευνας. Ειδικότερα, στην περίπτωση της αναγνώρισης ασθενειών στα φύλλα τομάτας, η κλάση που σχετίζεται με τον ιό της κηλίδωσης των φύλλων περιέχει 5.357 δείγματα, ενώ αντίθετα η κλάση που αφορά τον ιό του μωσαϊκού περιέχει μόλις 373 δείγματα, όπως φαίνεται στον Πίνακα 4.2. Αυτές οι έντονες ανισορροπίες επισημαίνουν την ανάγκη υιοθέτησης ειδικά σχεδιασμένων τεχνικών, ώστε να διασφαλιστεί η ισότιμη προγνωστική απόδοση των μοντέλων για όλες τις κατηγορίες.

Η συνάρτηση εστιακής απώλειας (FL) [226] έχει αναδειχθεί ως μία ιδιαίτερα αποτελεσματική τεχνική για την αντιμετώπιση του προβλήματος της ταξικής ανισορροπίας, κυρίως στο πεδίο της ταξινόμησης αντικειμένων. Η βασική ιδέα πίσω από τη συνάρτηση εστιακής απώλειας είναι διπλή: αφενός μειώνει τη συνεισφορά των εύκολα προβλέψιμων δειγμάτων στη συνάρτηση απώλειας, και αφετέρου δίνει μεγαλύτερη έμφαση σε εκείνα που παρουσιάζουν δυσκολία στην ταξινόμηση.

Η μαθηματική διατύπωση της εστιακής απώλειας περιλαμβάνει δύο βασικούς όρους: την προβλεπόμενη πιθανότητα, που αντιστοιχεί στην πραγματική κλάση του δείγματος, και τον συντελεστή εξισορρόπησης, ο οποίος αντανακλά στη συχνότητα εμφάνισης της κάθε κλάσης. Επιπλέον, χρησιμοποιείται η παράμετρος εστίασης, για να ρυθμίσει περαιτέρω την έμφαση στα πιο δύσκολα δείγματα.

Συγκεκριμένα:

$$FL(p_t) = - \sum_{i=1}^N \alpha_t (1 - p_t)^\gamma \log(p_t) \quad (4.1)$$

όπου N είναι το πλήθος των δειγμάτων του συνόλου δεδομένων, i είναι ο δείκτης δείγματος, $i = 1, 2, \dots, N$, t είναι η σωστή κλάση του δείγματος i , p_t αντιπροσωπεύει την προβλεπόμενη από το μοντέλο πιθανότητα της σωστής κλάσης t για ένα δεδομένο δείγμα i . Η παράμετρος α_t χρησιμεύει ως συντελεστής εξισορρόπησης, παρέχοντας διαφορετικά βάρη στις κλάσεις t , με βάση τη συχνότητα ή τη σημασία τους, ενώ γ καθορίζει την ένταση της εστίασης στα δύσκολα προς ταξινόμηση δείγματα.

Η προτεινόμενη συνάρτηση επανασταθμισμένης εστιακής απώλειας, αναπτύχθηκε με στόχο την αποτελεσματικότερη αντιμετώπιση του προβλήματος της

ανισορροπίας των κλάσεων. Η βασική ιδέα πίσω από τη συνάρτηση επανασταθμισμένης εστιακής απώλειας είναι η χρήση ενός προσαρμοσμένου συντελεστή εξισορρόπησης α , ο οποίος είναι αντιστρόφως ανάλογος της συχνότητας εμφάνισης κάθε κλάσης στο σύνολο εκπαίδευσης. Συγκεκριμένα, σε ένα σύνολο δεδομένων που περιέχει N κλάσεις, έστω n_i ο αριθμός των δειγμάτων της κλάσης i , ο συνολικός αριθμός των δειγμάτων N_{total} υπολογίζεται από τη σχέση:

$$N_{total} = \sum_{i=1}^N n_i \quad (4.2)$$

Ο συντελεστής εξισορρόπησης α υπολογίζεται ως εξής:

$$\alpha_t = \frac{N_{total} - n_i}{N_{total}} \quad (4.3)$$

Η τελική συνάρτηση επανασταθμισμένης εστιακής απώλειας διατυπώνεται ως ακολούθως:

$$\begin{aligned} RFL(p_t) &= - \sum_{i=1}^N \alpha_i (1 - p_{t,i})^\gamma \log(p_{t,i}) \\ &= - \sum_{i=1}^N \frac{N_{total} - n_i}{N_{total}} (1 - p_{t,i})^\gamma \log(p_{t,i}) \end{aligned} \quad (4.4)$$

Η ανάπτυξη της συνάρτησης επανασταθμισμένης εστιακής απώλεια σχεδιάστηκε με γνώμονα τις αρχές της μεθοδολογίας μας (βλ. Ενότητα 3.2.2.3), όπου δίνεται ιδιαίτερη έμφαση στη διαμόρφωση κατάλληλων μηχανισμών εξισορρόπησης με βάση τα χαρακτηριστικά του συνόλου δεδομένων και τους υπολογιστικούς περιορισμούς των HCS. Η συγκεκριμένη προσέγγιση επιτρέπει τη δικαιότερη κατανομή των βαρών, παρέχοντας μεγαλύτερη βαρύτητα στις λιγότερο εκπροσωπούμενες κλάσεις, και συμβάλλει στην αποφυγή μεροληψιών του μοντέλου, εξασφαλίζοντας υψηλότερη συνολική απόδοση.

4.5 Επισκόπηση Αρχιτεκτονικής Συστήματος

4.5.1 Μεθοδολογία Σχεδιασμού και Αρχιτεκτονική Υλοποίησης

Οι συσκευές που βασίζονται στις GPUs έχουν καθιερωθεί ως η παραδοσιακή λύση για την εκτέλεση εργασιών DL, ιδιαίτερα στον τομέα των CNNs. Οι GPUs διαθέτουν υψηλή υπολογιστική ισχύ, ιδανική για τις πολύπλοκες μαθηματικές πράξεις που απαιτούνται από τα CNNs, καθιστώντας τις κατάλληλες για εφαρμογές που απαιτούν μεγάλη επεξεργαστική ισχύ και ανάλυση δεδομένων σε πραγματικό χρόνο. Η αρχιτεκτονική παράλληλης επεξεργασίας που χαρακτηρίζει τις GPUs, σε συνδυασμό με εξειδικευμένα πλαίσια ανάπτυξης, όπως το CUDA και το TensorFlow, επιτρέπουν την ταχεία εκτέλεση πολύπλοκων υπολογιστικών διαδικασιών, επιτυγχάνοντας σημαντική μείωση του απαιτούμενου χρόνου εκπαίδευσης των CNNs [227].

Παρόλα αυτά, η ανάγκη για ακόμα υψηλότερη ενεργειακή απόδοση και ταχύτητα έχει οδηγήσει την ερευνητική κοινότητα στη διερεύνηση εναλλακτικών τεχνολογιών, όπως είναι τα FPGAs. Τα FPGA αποτελούν μια πολλά υποσχόμενη λύση λόγω της δυνατότητάς τους να επιταχύνουν συγκεκριμένες υπολογιστικές διεργασίες μέσω εξειδικευμένων, υλικού επιπέδου υλοποιήσεων. Ο στόχος, λοιπόν, σχεδιασμού και ανάπτυξης ενός μοντέλου CNN για χρήση σε FPGA είναι διττός: αφενός η επίτευξη υψηλών ταχυτήτων επεξεργασίας, που είναι απαραίτητες για εφαρμογές πραγματικού χρόνου, και αφετέρου η επίτευξη υψηλής ενεργειακής απόδοσης, που είναι κρίσιμη σε περιβάλλοντα με περιορισμένη διαθεσιμότητα ισχύος [228, 229].

Η ανάγκη για υψηλή ταχύτητα επεξεργασίας είναι ιδιαίτερα κρίσιμη σε εφαρμογές πραγματικού ή σχεδόν πραγματικού χρόνου, όπως είναι η αυτόνομη οδήγηση, η ιατρική διάγνωση μέσω εικόνας, και η ανίχνευση ασθενειών στη γεωργία. Ταυτόχρονα, η ενεργειακή απόδοση αποτελεί ολοένα και πιο σημαντική απαίτηση, ειδικά όταν η διαθέσιμη ισχύς είναι περιορισμένη, επεκτείνοντας τη διάρκεια λειτουργίας των συσκευών και μειώνοντας την ανάγκη για συχνές επαναφορτίσεις ή αντικατάσταση μπαταριών.

Στην παρούσα έρευνα, η επιλογή της πλατφόρμας FPGA PYNQ—Z1 βασίστηκε στη δυνατότητά της να προσφέρει σημαντική επιτάχυνση υλικού στις υπολογιστικές εργασίες. Η χρήση αυτής της πλατφόρμας επιτρέπει την υλοποίηση και εκτέλεση

των CNNs με υψηλή ενεργειακή απόδοση και ταχύτητα, καλύπτοντας τις απαιτήσεις για άμεση λήψη αποφάσεων, όπως απαιτείται σε εφαρμογές γεωργίας ακριβείας.

4.5.2 Πλακέτα PYNQ—Z1

Η πλακέτα PYNQ—Z1 [230] αποτελεί μια ευέλικτη πλατφόρμα, βασισμένη στο SoC ZYNQ XC7Z020. Διαθέτει έναν επεξεργαστή διπλού πυρήνα ARM Cortex—A9, ο οποίος συνδυάζεται με τις ισχυρές δυνατότητες του FPGA της σειράς 7 της Xilinx, παρέχοντας έτσι υψηλές επιδόσεις και σημαντική επιτάχυνση υλικού σε υπολογιστικά απαιτητικές εφαρμογές.

Το SoC ZYNQ XC7Z020 διαθέτει έναν διπύρρηνο επεξεργαστή ARM Cortex—A9, ο οποίος παρέχει υψηλή υπολογιστική ισχύ και μπορεί να αξιοποιηθεί για την εκτέλεση λογισμικού και την επικοινωνία με το υλικό του FPGA. Ο συγκεκριμένος επεξεργαστής λειτουργεί παράλληλα με τις δυνατότητες του FPGA της πλατφόρμας, το οποίο περιλαμβάνει 13.300 λογικές τομές (Logic Slices), 53.200 πίνακες αναζήτησης (Look—Up Tables, LUTs) και 220 τομές επεξεργασίας ψηφιακού σήματος (DSP Slices). Αυτή η υλική υποδομή προσφέρει μεγάλη ευελιξία στον σχεδιασμό ψηφιακών κυκλωμάτων και στην υλοποίηση σύνθετων αλγορίθμων.

Η πλακέτα PYNQ—Z1 παρέχει επίσης εκτεταμένες δυνατότητες συνδεσιμότητας. Ενσωματώνει θύρες επικοινωνίας όπως Gigabit Ethernet, HDMI για οπτική απεικόνιση, USB—UART, καθώς και δύο θύρες USB, εκ των οποίων η μία λειτουργεί και ως host. Επιπλέον, προσφέρει ηχητική σύνδεση μέσω ενσωματωμένου κωδικοποιητή ήχου και θύρες εισόδου—εξόδου, ενώ περιλαμβάνει και δύο υποδοχές 40 ακίδων συμβατές με τα αντίστοιχα επεκτατικά συστήματα του Raspberry Pi, αυξάνοντας περαιτέρω τις δυνατότητες επέκτασης και ευελιξίας του συστήματος.

Μία επιπλέον σημαντική δυνατότητα της πλατφόρμας PYNQ—Z1 είναι η υποστήριξη του πλαισίου ανάπτυξης PYNQ (Python Productivity for ZYNQ). Το πλαίσιο αυτό επιτρέπει στους προγραμματιστές να αξιοποιήσουν τις δυνατότητες της επιτάχυνσης υλικού μέσω FPGA, χρησιμοποιώντας τη γλώσσα προγραμματισμού Python. Με τον τρόπο αυτό επιτυγχάνεται μεγαλύτερη παραγωγικότητα, ευκολία στη χρήση και σημαντική μείωση του χρόνου ανάπτυξης εφαρμογών που απαιτούν εντατική υπολογιστική ισχύ.

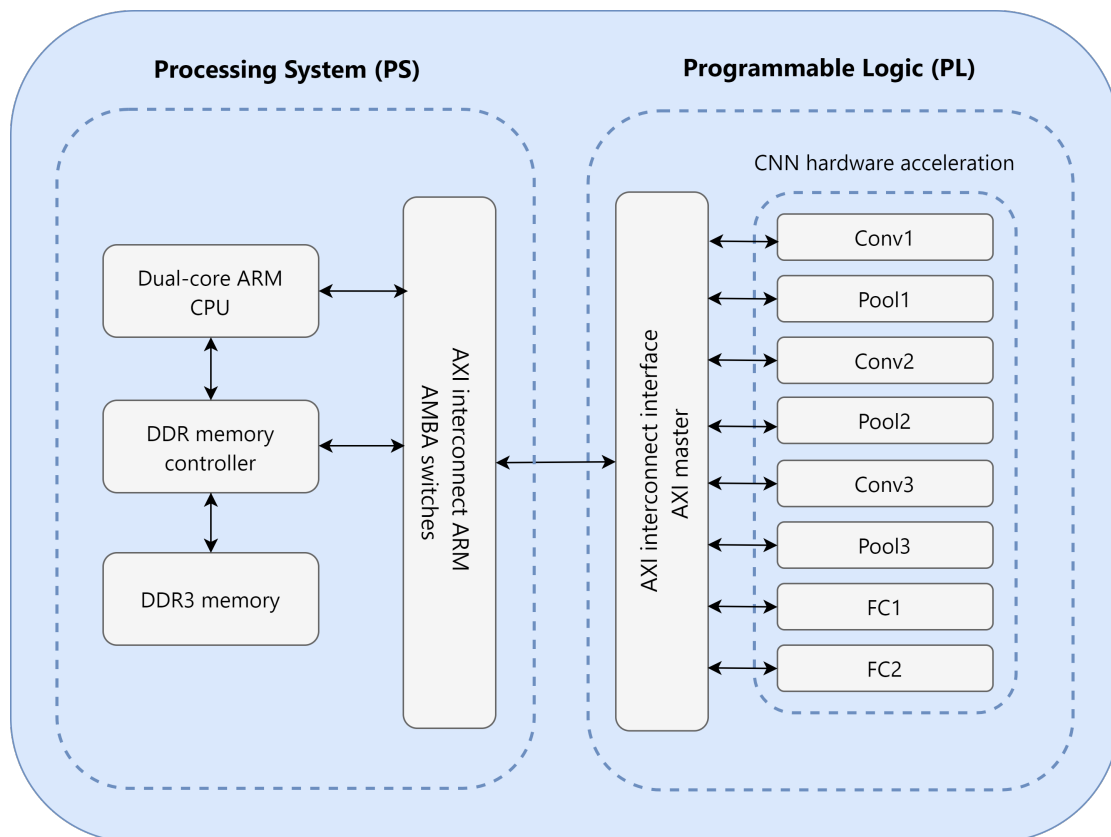
4.5.3 Αρχιτεκτονική Υλικού

Η αρχιτεκτονική της πλακέτας PYNQ-Z1 [230], όπως παρουσιάζεται στο Σχήμα 4.6, διακρίνεται σε δύο κύρια τμήματα: το σύστημα επεξεργασίας (Processing System, PS) και το τμήμα λογικού προγραμματισμού (Programmable Logic, PL). Το PS περιλαμβάνει έναν διπύρρηνο επεξεργαστή ARM Cortex-A9, ο οποίος διαχειρίζεται τη μνήμη DDR3 μέσω του DDR Memory Controller, αναλαμβάνοντας τον έλεγχο της εισαγωγής και της προεπεξεργασίας των δεδομένων εισόδου, όπως η φόρτωση εικόνων και η αποθήκευση των προεκπαιδευμένων παραμέτρων (βάρη, μεροληψίες). Τα δεδομένα αυτά αποθηκεύονται στη μνήμη DDR3, που λειτουργεί ως κοινός αποθηκευτικός χώρος τόσο για το PS, όσο και για το PL. Η επικοινωνία μεταξύ των δύο υποσυστημάτων επιτυγχάνεται μέσω του διαύλου AXI Interconnect, ο οποίος διασφαλίζει την αμφίδρομη και ταχύτατη μεταφορά δεδομένων και εντολών, με τη χρήση του AXI DMA Controller να επιτρέπει την μεταφορά μεγάλων όγκων δεδομένων χωρίς την άμεση εμπλοκή της CPU, μειώνοντας έτσι τη χρονική καθυστέρηση και τον υπολογιστικό φόρτο του επεξεργαστή.

Το δεύτερο τμήμα, PL, αποτελεί το FPGA μέρος του συστήματος ZYNQ SoC, το οποίο περιλαμβάνει ένα AXI Interconnect Interface (AXI Master) για την επικοινωνία με το PS και τα IP blocks που αντιστοιχούν στα επιμέρους στρώματα του CNN (Conv1, Pool1, Conv2, Pool2, Conv3, Pool3, FC1, FC2). Η παραλληλία που προσφέρει το FPGA επιτρέπει την αποδοτική εκτέλεση των υπολογιστικά απαιτητικών στρωμάτων του δικτύου, με αποτέλεσμα την σημαντική μείωση του χρόνου επεξεργασίας και της κατανάλωσης ενέργειας. Οι εικόνες και τα βάρη φορτώνονται από την κάρτα SD στη DDR3 μνήμη μέσω της CPU και στη συνέχεια, μέσω του AXI DMA Controller, αποστέλλονται στα αντίστοιχα IP blocks του PL για επεξεργασία. Μετά την ολοκλήρωση των υπολογισμών στο PL, τα παραγόμενα αποτελέσματα αποθηκεύονται στη μνήμη DDR3 μέσω του AXI DMA Controller, ώστε να είναι άμεσα προσβάσιμα από την CPU για μετεπεξεργασία ή παρουσίαση.

Ένα ιδιαίτερο χαρακτηριστικό της πλακέτας PYNQ-Z1 είναι η υποστήριξη του πλαισίου ανάπτυξης PYNQ, το οποίο επιτρέπει τη χρήση της γλώσσας προγραμματισμού Python για τον άμεσο έλεγχο και την παραμετροποίηση της PL. Η δυνατότητα αυτή καθιστά την πλατφόρμα ιδιαίτερα ισχυρή και προσιτή για την υλοποίηση και αξιολόγηση εφαρμογών AI, διευκολύνοντας την ενιαία διασύνδεση μεταξύ λογισμικού και επιτάχυνσης υλικού. Στο PS της πλατφόρμας PYNQ-Z1 εκτελείται κώδικας σε Python (μέσω του PYNQ framework), ο οποίος αναλαμβάνει

τη διαχείριση της επικοινωνίας με το PL. Ο κώδικας αυτός φορτώνει το bitstream του FPGA, τα δεδομένα εισόδου (ανά batch) και τα βάρη στη μνήμη, δρομολογεί τις αντίστοιχες εντολές προς τους επιταχυντές υλικού για την επεξεργασία τους και παραλαμβάνει τα αποτελέσματα για οπτικοποίηση.



Σχήμα 4.6: Αρχιτεκτονική επισκόπηση υλικού της πλατφόρμας PYNQ—Z1 που χρησιμοποιείται για την αναγνώριση ασθενειών της τομάτας.

4.5.4 Σχεδίαση Αρθρωτών Κυκλωμάτων του MiniTomatoNet

Ο μηχανισμός επιτάχυνσης για την αύξηση της υπολογιστικής απόδοσης στα συνελικτικά, πλήρως συνδεδεμένα και μέγιστης συγκέντρωσης επίπεδα του MiniTomatoNet υλοποιήθηκε μέσω του Vivado High-Level Synthesis (HLS), έκδοση 2018.1, με χρήση της γλώσσας προγραμματισμού C++ [231]. Η προσέγγιση αυτή απαλλάσσει τους σχεδιαστές από την ανάγκη εμπλοκής στις λεπτομέρειες της κυκλωματικής αρχιτεκτονικής, επιτρέποντάς τους να εργαστούν σε αλγοριθμικό επίπεδο. Η τεχνική HLS επιλέχθηκε αντί της παραδοσιακής συγγραφής σε VHDL, καθώς προσφέρει σημαντική μείωση στον χρόνο ανάπτυξης και επιτρέπει την ενσωμάτωση βελτιστοποιήσεων, διευκολύνοντας την αποδοτική υλοποίηση πολύπλοκων υπολογιστικών δομών, όπως αυτές των CNNs, σε υλικό.

Το κύκλωμα του MiniTomatoNet ακολουθεί μια αρθρωτή σχεδιαστική προσέγγιση, όπου κάθε ξεχωριστή μονάδα αντιστοιχεί σε ένα επίπεδο του MiniTomatoNet. Δύο βασικές αρχές υποστηρίζουν αυτή τη σχεδιαστική επιλογή:

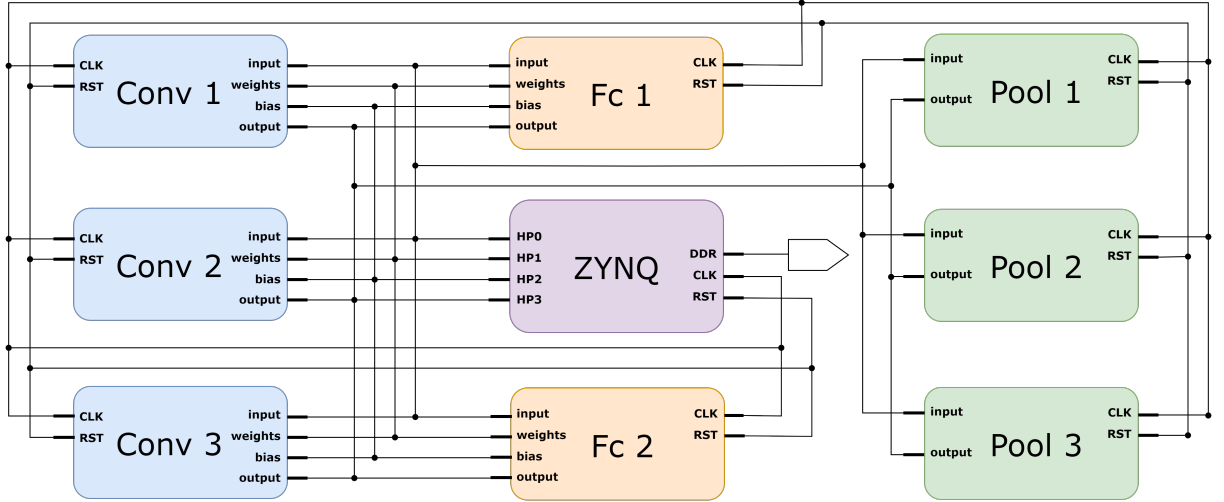
- Βελτιστοποίηση αποθήκευσης: Καθώς κάθε επίπεδο του MiniTomatoNet παράγει ενδιάμεσα δεδομένα χαρακτηριστικών, η αρθρωτή σχεδίαση του κυκλώματος επιτρέπει τη μεταφορά αυτών των δεδομένων στην εξωτερική μνήμη DDR3 DRAM για προσωρινή αποθήκευση. Αυτή η προσέγγιση μειώνει την κατανάλωση ενσωματωμένων πόρων μνήμης στο PL, αποδεσμεύοντας πολύτιμο χώρο για άλλες υπολογιστικές μονάδες και βελτιώνοντας τη συνολική δυνατότητα παραλληλοποίησης του συστήματος.
- Επεκτασιμότητα και αναβαθμισιμότητα: Η αρθρωτή σχεδίαση διευκολύνει την εξέλιξη και επέκταση του δικτύου, επιτρέποντας την εύκολη προσθήκη ή τροποποίηση επιπέδων του MiniTomatoNet χωρίς να επηρεάζεται η συνολική αρχιτεκτονική του κυκλώματος.

Οι αλγόριθμοι 4.5.1 — 4.5.4 παρουσιάζουν τον HLS ψευδοκώδικα που αντιστοιχεί σε κάθε επίπεδο του δικτύου MiniTomatoNet. Στη συνέχεια, οι ξεχωριστοί κυκλωματικοί κώδικες που αντιστοιχούν σε κάθε επίπεδο ενσωματώνονται μέσα σε IP cores.

Οι IP cores, μέσα στο περιβάλλον ανάπτυξης του Vivado, διασυνδέονται μεταξύ τους για τη διαμόρφωση της συνολικής κυκλωματικής αρχιτεκτονικής του MiniTomatoNet, όπως φαίνεται στο Σχήμα 4.7. Το σύστημα ZYNQ PS, το οποίο είναι εξοπλισμένο με επεξεργαστή ARM, επικοινωνεί με κάθε IP core που είναι κρίσιμο για τη λειτουργία του MiniTomatoNet. Αυτό επιτυγχάνεται μέσω των διαύλων AXI HP, διασφαλίζοντας τη βέλτιστη ροή δεδομένων μεταξύ των PS και PL. Με αυτόν τον τρόπο, η επικοινωνία μεταξύ των ετερογενών μονάδων επεξεργασίας γίνεται αποτελεσματικά, εξασφαλίζοντας χαμηλό λανθάνοντα χρόνο και υψηλή απόδοση στην επεξεργασία του CNN.

4.5.5 Τεχνικές Βελτιστοποίησης HLS του MiniTomatoNet σε Ετερογενή FPGA—SoC

Η εφαρμογή των τεχνικών βελτιστοποίησης παίζει καθοριστικό ρόλο στην αξιοποίηση του ενσωματωμένου παραλληλισμού των FPGAs, διασφαλίζοντας



Σχήμα 4.7: Μπλοκ διάγραμμα υλικού με τους IP πυρήνες της αρχιτεκτονικής MiniTomatoNet.

Αλγόριθμος 4.5.1 Depthwise convolution layer — Επίπεδο συνέλιξης κατά βάθος.

- 1: **DepthwiseConv**(*input_dds*, *weights_dds*, *output_dds*)
- 2: ▷ Define data structures for input and kernel weights
- 3: *input*[*KernelSize*][*KernelSize*]
- 4: *weights*[*KernelSize*][*KernelSize*]
- 5: **for** *in_depth* = 0 **to** *InDepth* - 1 **do**
- 6: Load weights from *weights_dds* for depth *in_depth* and size *KernelSize*
- 7: **for** *out_height* = 0 **to** *OutHeight* - 1 **do**
- 8: Load input slice from *input_dds* corresponding to height *out_height*
- 9: **for** *out_width* = 0 **to** *OutWidth* - 1 **do**
- 10: Update a column of *input* by *out_width*
- 11: $conv_result \leftarrow$ Calculate convolution of *input* and *weights* of size $KernelSize \times KernelSize$
- 12: Save *conv_result* to *output_dds* for depth *in_depth*, height *out_height*, and width *out_width*
- 13: **end for**
- 14: **end for**
- 15: **end for**

την αποτελεσματική ροή δεδομένων και μεγιστοποιώντας την απόδοση των διαθέσιμων πόρων. Στα πλαίσια της σύνθεσης του HLS για επιταχυντές CNN βασισμένους σε FPGA, η ανάγκη για ταχεία υπολογιστική επεξεργασία είναι υψίστης σημασίας. Τα CNNs, λόγω της πολύπλοκης αλληλεπίδρασης γραμμικών και μη γραμμικών λειτουργιών μεταξύ συνελκτικών, πλήρως συνδεδεμένων και επιπέδων συγκέντρωσης, απαιτούν τεράστιες υπολογιστικές δυνατότητες. Για την κάλυψη αυτής της ανάγκης και τη βελτιστοποίηση της ταχύτητας εκτέλεσης των

Αλγόριθμος 4.5.2 Pointwise convolution layer — Επίπεδο σημειακής συνέλιξης.

```

1: PointwiseConv(input_ddr, weights_ddr, bias_ddr, output_ddr)
2: ▷ Compute the pointwise convolution
3: for out_depth = 0 to OutDepth−1 do
4:   for out_height = 0 to OutHeight−1 do
5:     for out_width = 0 to OutWidth−1 do
6:       for in_depth = 0 to InDepth−1 do
7:         output_ddr[out_depth][out_height][out_width] +=
8:           weights_ddr[in_depth][out_depth] ×
9:           input_ddr[in_depth][out_height][out_width]
10:        end for
11:      end for
12:    end for
13: ▷ Apply bias and ReLU activation
14: for out_depth = 0 to OutDepth−1 do
15:   for out_height = 0 to OutHeight−1 do
16:     for out_width = 0 to OutWidth−1 do
17:       val_rl = output_ddr[out_depth][out_height][out_width] +
18:       bias_ddr[out_depth]
19:       if val_rl < 0 then
20:         output_ddr[out_depth][out_height][out_width] = 0
21:       else
22:         output_ddr[out_depth][out_height][out_width] = val_rl
23:       end if
24:     end for
25:   end for

```

υπολογισμών, εφαρμόζονται τέσσερις βασικές τεχνικές HLS: ξετύλιγμα βρόχων, διασωλήνωση, κατατμήση πινάκων και επιπέδωση βρόχων, οι οποίες αξιοποιούνται στο μοντέλο MiniTomatoNet.

4.5.5.1 Ξετύλιγμα Βρόχων

Η τεχνική του ξετυλίγματος βρόχων αναπαράγει το σώμα του βρόχου πολλές φορές, μειώνοντας τον αριθμό των επαναλήψεων και εκθέτοντας ευκαιρίες για ταυτόχρονη εκτέλεση των αρχικά σειριακών λειτουργιών. Έτσι, με την εφαρμογή της ανάπτυξης βρόχων σε εμφωλευμένους βρόχους εντός των στρωμάτων συνελκτικής συνέλιξης, συγκέντρωσης και πλήρως συνδεδεμένων επιπέδων, επιτυγχάνεται σημαντική αύξηση του παραλληλισμού. Αυτή η αύξηση της παράλληλης επεξεργασίας

Αλγόριθμος 4.5.3 Max-pool layer – Επίπεδο μέγιστης συγκέντρωσης.

```

1: MaxPool(input_ddr, output_ddr)
2: input[InD][InH][InW]
3: output[OutD][OutH][OutW]
4: for out_depth = 0 to OutDepth - 1 do
5:   for out_height = 0 to OutHeight - 1 do
6:     for out_width = 0 to OutWidth - 1 do
7:       for in_depth = 0 to InDepth - 1 do
8:         Load input values from input_ddr corresponding to current
           in_depth, out_height, and out_width
9:         output[out_depth][out_height][out_width] =  $-\infty$  ▷ Initialize to a
           very small value
10:        for k_height = 0 to KernelSize - 1 do
11:          for k_width = 0 to KernelSize - 1 do
12:            in_temp = input[in_depth][in_height +
              k_height][in_width + k_width]
13:            output[out_depth][out_height][out_width] =
              max(output[out_depth][out_height][out_width], in_temp)
14:          end for
15:        end for
16:        Save output[out_depth][out_height][out_width] to output_ddr
17:      end for
18:    end for
19:  end for
20: end for

```

Αλγόριθμος 4.5.4 Fully connected layer – Πλήρως συνδεδεμένο επίπεδο.

```

1: FullyConnected(input_ddr, weights_ddr, output_ddr)
2: input[IN]
3: output[OUT]
4: weights[IN][OUT]
5: for input_i = 0 to InputNum - 1 do
6:   for in_i = 0 to IN - 1 do
7:     Load input[in_i] value from input_ddr[input_i][in_i]
8:     for output_i = 0 to OutputNum - 1 do
9:       Initialize sum = 0
10:      for out_i = 0 to OUT - 1 do
11:        Load weight value from weights_ddr[in_i][out_i]
12:        sum + = input[in_i] × weight
13:      end for
14:      Save sum to output_ddr[input_i][output_i]
15:    end for
16:  end for
17: end for

```

συνεπάγεται μεγαλύτερη ροή των δεδομένων και συνεπώς βελτιωμένη απόδοση του συστήματος. Ωστόσο, παρά τα οφέλη στην ταχύτητα, αυτή η τεχνική συνοδεύεται από αύξηση της χρήσης πόρων, γεγονός που πρέπει να λαμβάνεται υπόψη κατά τη σχεδίαση των FPGA επιταχυντών.

4.5.5.2 Διασωλήνωση

Η τεχνική διασωλήνωσης διαχωρίζει τη διαδικασία σε διακριτά, σειριακά βήματα, τα οποία εκτελούνται επικαλυπτόμενα. Με τον τρόπο αυτό, μειώνεται το διάστημα εκκίνησης μεταξύ των επαναλήψεων βρόχων ή των κλήσεων συναρτήσεων, επιτρέποντας την ταυτόχρονη εκκίνηση πολλαπλών επαναλήψεων χωρίς την ανάγκη πλήρους ολοκλήρωσης της προηγούμενης.

Στα CNNs, όπου η έξοδος ενός επιπέδου αποτελεί την είσοδο του επόμενου, η τεχνική της διασωλήνωσης διασφαλίζει ότι, μόλις ένα επίπεδο παράγει επαρκή δεδομένα εξόδου, το επόμενο επίπεδο μπορεί να ξεκινήσει την επεξεργασία χωρίς καθυστέρηση. Αυτό εξασφαλίζει σταθερή και αποδοτική ροή των δεδομένων σε όλο το δίκτυο. Στο FPGA, ο παραλληλισμός που παρέχεται από τη διασωλήνωση βελτιώνει ακόμα περισσότερο τη συνολική υπολογιστική αποδοτικότητα.

4.5.5.3 Κατάτμηση Πινάκων

Η προσπέλαση δεδομένων από πίνακες υπόκειται συχνά σε περιορισμούς του εύρους ζώνης της μνήμης στο υλικό, ιδίως όταν απαιτείται ταυτόχρονη πρόσβαση σε πολλαπλά στοιχεία. Η τεχνική της κατάτμησης πινάκων επιχειρεί να αντιμετωπίσει αυτή την πρόκληση μέσω της διαίρεσης των μεγάλων πινάκων σε μικρότερα, ανεξάρτητα τμήματα. Με αυτόν τον τρόπο, καθίσταται δυνατή η παράλληλη προσπέλαση των διαφορετικών τμημάτων, γεγονός που αυξάνει το συνολικό εύρος ζώνης της μνήμης και μειώνει τη συμφόρηση στην πρόσβαση.

Η κατάτμηση πινάκων είναι ιδιαίτερης σημασίας κατά την υλοποίηση των συνελικτικών και πλήρως συνδεδεμένων επιπέδων, στα οποία τα μοτίβα αποθήκευσης και προσπέλασης των βαρών και των μεροληψιών (biases) επηρεάζουν άμεσα την υπολογιστική απόδοση. Δεδομένου ότι τα εν λόγω επίπεδα συνιστούν τον βασικό υπολογιστικό πυρήνα ενός CNN, η δυνατότητα ταυτόχρονης χρήσης πολλαπλών στοιχείων βάρους και μεροληψίας επιταχύνει σημαντικά τις πράξεις των πινάκων και ενισχύει τη συνολική αποδοτικότητα του μοντέλου.

4.5.5.4 Επιπέδωση Βρόχων

Οι εμφωλευμένοι βρόχοι δημιουργούν επιπλέον κύκλους ρολογιού κάθε φορά που υπάρχει μετάβαση μεταξύ του εσωτερικού και του εξωτερικού βρόχου. Για την αντιμετώπιση αυτού του ζητήματος, εφαρμόζεται η τεχνική της επιπέδωσης βρόχων. Με τη μετατροπή αυτών των εμφωλευμένων βρόχων σε έναν ενιαίο βρόχο, εξαλείφεται η επιβάρυνση που προκύπτει από την εναλλαγή μεταξύ των βρόχων, οδηγώντας σε σημαντική εξοικονόμηση του χρόνου εκτέλεσης και τη βελτίωση της συνολικής απόδοσης του συστήματος.

Η αποτελεσματική εφαρμογή των παραπάνω τεχνικών HLS στο MiniTomatoNet επιτρέπει τη μέγιστη αξιοποίηση των πόρων του FPGA και τη δραστική μείωση του συνολικού χρόνου επεξεργασίας, καθιστώντας το μοντέλο ιδιαίτερα αποδοτικό για εφαρμογές DL σε HCS.

4.6 Πειραματικά Αποτελέσματα

4.6.1 Φάση Εκπαίδευσης και Επικύρωσης

4.6.1.1 Παράμετροι Εκπαίδευσης

Ο πειραματικός σχεδιασμός που χρησιμοποιήθηκε για την εκπαίδευση και την αξιολόγηση του MiniTomatoNet περιλάμβανε έναν επεξεργαστή Intel® Core™ i7–11700F με συχνότητα λειτουργίας 2.5 GHz, 16 GB μνήμης RAM και κάρτα γραφικών NVIDIA GeForce RTX 3050 GPU. Η αρχιτεκτονική του MiniTomatoNet εκπαιδεύτηκε και δοκιμάστηκε σε περιβάλλον προγραμματισμού Python, αξιοποιώντας τις δυνατότητες των σύγχρονων υπολογιστικών μονάδων για την αποτελεσματική εκπαίδευση και υλοποίηση του μοντέλου.

Η εκπαίδευση του MiniTomatoNet δεν υλοποιήθηκε στο PYNQ-Z1, καθώς το PYNQ-Z1 – αν και ιδανικό για την επιτάχυνση της φάσης εξαγωγής συμπερασμάτων – παρουσιάζει σημαντικούς περιορισμούς στη φάση της εκπαίδευσης, λόγω απουσίας κατάλληλων υποδομών για την αποδοτική υλοποίηση των απαιτητικών πράξεων της εκπαίδευσης (όπως η οπισθοδιάδοση και η ενημέρωση βαρών).

Για την φάση εκπαίδευσης του μοντέλου υιοθετήθηκε η στρατηγική βελτιστοποίησης μέσω του αλγορίθμου Adam [232], ο οποίος είναι γνωστός για την ικανότητα του

να ενημερώνει δυναμικά τα βάρη του δικτύου, προσφέροντας ταχεία σύγκλιση και σταθερότητα σε όλη τη διάρκεια της διαδικασίας. Αυτή η τεχνική επιτρέπει την αποτελεσματική προσαρμογή των παραμέτρων του CNN, μειώνοντας σταδιακά τη διαφορά μεταξύ των πραγματικών ετικετών των δεδομένων και των προβλέψεων του μοντέλου.

Ο αριθμός των εποχών καθορίστηκε σε 30 (βλ. Ενότητα 3.4.1), προκειμένου να επιτραπεί στο μοντέλο να συγκλίνει σε σταθερό σημείο με ελάχιστη υπερπροσαρμογή. Ο ρυθμός μάθησης ορίστηκε στο 0,001, όπως προτείνεται εξ ορισμού στον αλγόριθμο Adam, με στόχο την επίτευξη βέλτιστης ισορροπίας μεταξύ ταχύτητας σύγκλισης και σταθερότητας της εκπαίδευσης. Το μέγεθος παρτίδας των 16 δειγμάτων διασφαλίζει την αποδοτική αξιοποίηση των διαθέσιμων πόρων, ενώ ταυτόχρονα διατηρεί τη συνοχή στη διαδικασία μάθησης του δικτύου.

Κατά την εφαρμογή της εστιακής συνάρτησης απώλειας, η τιμή της παραμέτρου α ορίστηκε σε 0,5 και της γ σε 2, που αποτελούν τις τιμές – προεπιλογές της συγκεκριμένης συνάρτησης [226]. Η παράμετρος α καθορίζει το βάρος που αποδίδεται στα θετικά και αρνητικά δείγματα κατά τον υπολογισμό της απώλειας και με την επιλογή 0,5 να επιτυγχάνεται η ισότιμη μεταχείριση και των δύο ομάδων. Η παράμετρος γ , από την άλλη, ελέγχει τον βαθμό εστίασης της απώλειας στα πιο δύσκολα ή λανθασμένα ταξινομημένα δείγματα, ενισχύοντας τον ρόλο τους στην εκπαίδευση του δικτύου και περιορίζοντας τη συνεισφορά των εύκολα ταξινομημένων περιπτώσεων.

Η επιλογή των υπερπαραμέτρων για την εκπαίδευση του MiniTomatoNet έγινε με στόχο τη βέλτιστη απόδοση του μοντέλου, λαμβάνοντας υπόψη τη σταθερότητα και την αποδοτικότητα του αλγορίθμου μάθησης. Συγκεκριμένα, οι βασικές παράμετροι εκπαίδευσης παρουσιάζονται στον Πίνακα 4.4.

Πίνακας 4.4: Καθορισμός των υπερπαραμέτρων που χρησιμοποιήθηκαν για την εκπαίδευση του MiniTomatoNet.

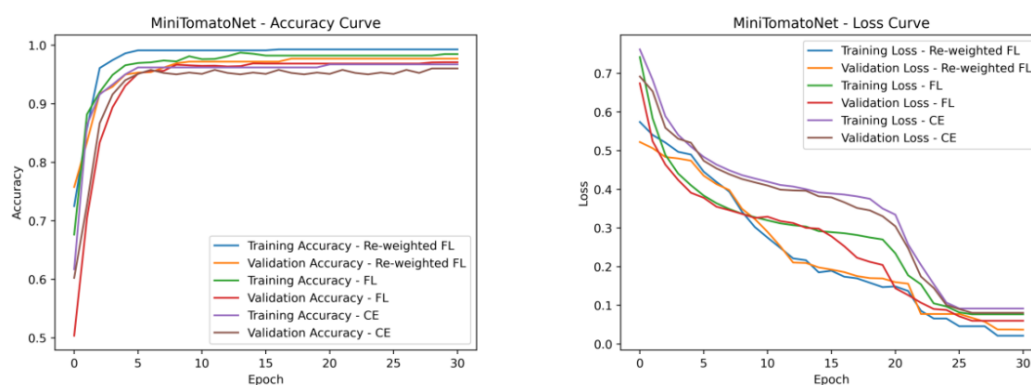
Hyperparameter	Value
Optimizer	Adam
Batch size	16
Loss function	CE, FL, RFL
Epochs	30
Learning rate	0,001

4.6.1.2 Καμπύλες Ακρίβειας/Απώλειας Εκπαίδευσης και Επικύρωσης

Οι καμπύλες ακρίβειας/απώλειας κατά τη διάρκεια της φάση εκπαίδευσης και επικύρωσης, όπως προέκυψαν με την εφαρμογή των συναρτήσεων της διασταυρούμενης εντροπίας, εστιακής απώλειας και επανασταθμισμένης εστιακής απώλειας, παρουσιάζονται στο Σχήμα 4.8.

Όπως παρατηρείται στις καμπύλες ακρίβειας, όλες οι παραλλαγές της συνάρτησης απώλειας οδηγούν σε ταχεία αύξηση τόσο στο σύνολο εκπαίδευσης, όσο και στο σύνολο επικύρωσης, γεγονός που καταδεικνύει ότι το MiniTomatoNet μπορεί να εκπαιδευτεί αποτελεσματικά στα δεδομένα. Ιδιαίτερα, η επανασταθμισμένη εστιακή απώλεια παρουσιάζει ανώτερη τελική ακρίβεια τόσο στην εκπαίδευση, όσο και στην επικύρωση, υποδηλώνοντας ότι η αναπροσαρμογή των βαρών ως προς τις κλάσεις βοηθά στην καλύτερη γενίκευση του μοντέλου, ειδικά σε δεδομένα με ανισορροπία κλάσεων. Η εστιακή απώλεια επίσης υπερτερεί της διασταυρούμενης εντροπίας, ενώ η διασταυρούμενη εντροπία παρουσιάζει τη χαμηλότερη απόδοση στο τέλος της διαδικασίας εκπαίδευσης.

Όλες οι καμπύλες απώλειας στις παραλλαγές της συνάρτησης απώλειας, ακολουθούν φθίνουσα πορεία κατά τη διάρκεια της εκπαίδευσης, αποδεικνύοντας ότι το μοντέλο βελτιώνει προοδευτικά την απόδοσή του. Στο τέλος της διαδικασίας εκπαίδευσης, οι καμπύλες συγκλίνουν σε χαμηλές τιμές, με την επανασταθμισμένη εστιακή απώλεια και την εστιακή απώλεια να εμφανίζουν τις μικρότερες τελικές τιμές απώλειας σε σχέση με την διασταυρούμενη εντροπία, επιβεβαιώνοντας τη βελτιωμένη προσαρμοστικότητα σε ανισόρροπες κατανομές.



Σχήμα 4.8: Συγκριτική απεικόνιση των καμπυλών ακρίβειας/απώλειας εκπαίδευσης και επικύρωσης, με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.

4.6.2 Φάση Ελέγχου

4.6.2.1 Αξιολόγηση με Μήτρες Σύγχυσης

Η αξιολόγηση του προτεινόμενου CNN χαμηλής πολυπλοκότητας, πραγματοποιήθηκε με χρήση των μητρών σύγχυσης, όπως προέκυψαν με την εφαρμογή των συναρτήσεων της διασταυρούμενης εντροπίας, εστιακής απώλειας και επανασταθμισμένης εστιακής απώλειας, όπως παρουσιάζονται στο Σχήμα 4.9.

Στη φάση ελέγχου αξιολογήθηκαν συνολικά 3.635 δείγματα, με στόχο την κατηγοριοποίησή τους σε δέκα διακριτές κατηγορίες ασθενειών φύλλων τομάτας. Το MiniTomatoNet, σε συνδυασμό με την επανασταθμισμένη εστιακή απώλεια, κατέγραψε μόλις 86 λανθασμένες ταξινομήσεις, γεγονός που καταδεικνύει τον ιδιαίτερα υψηλό βαθμό ακρίβειας στην αναγνώριση των δέκα κατηγοριών. Αντίστοιχα, με την εφαρμογή της εστιακής συνάρτησης απώλειας, το προτεινόμενο συνελικτικό δίκτυο εμφάνισε 108 εσφαλμένες ταξινομήσεις, ενώ με τη χρήση της συνάρτησης διασταυρούμενης εντροπίας καταγράφηκαν συνολικά 125 λανθασμένες ταξινομήσεις.

Τα ευρήματα υποδηλώνουν ότι το MiniTomatoNet, ιδιαίτερα όταν συνδυάζεται με την επανασταθμισμένη εστιακή απώλεια, παρουσιάζει αυξημένη ικανότητα γενίκευσης και υψηλή ακρίβεια ταξινόμησης σε άγνωστα δεδομένα. Το μοντέλο επιδεικνύει τη δυνατότητα να διακρίνει με αξιοπιστία τις επιμέρους κατηγορίες ασθενειών των φυτών, γεγονός που αποτελεί κρίσιμο παράγοντα για την αποτελεσματική διάγνωση και διαχείριση σε πραγματικές γεωργικές συνθήκες.

4.6.2.2 Σύγκριση Μετρικών Ποιότητας

Οι ποιοτικές μετρικές σύμφωνα με την μεθοδολογία που παρουσιάστηκε στο Κεφάλαιο 3, και ειδικότερα βάσει των μετρικών ποιότητας (βλ. Ενότητα 3.3), με εφαρμογή των συναρτήσεων της διασταυρούμενης εντροπίας, της εστιακής απώλειας και της επανασταθμισμένης εστιακής απώλειας, παρουσιάζονται στον Πίνακα 4.5 και στο Σχήμα 4.10.

Από τα δεδομένα του Πίνακα 4.5, παρατηρούμε μια σταθερή βελτίωση του μέτρου της συνολικής ακρίβειας καθώς μεταβαίνουμε από τη χρήση της συνάρτησης απώλειας διασταυρούμενης εντροπίας στην επανασταθμισμένη εστιακή απώλεια.

		Confusion Matrix										
Actual	Bacterial Spot	419	1	2	0	0	0	1	2	0	0	0
	Early Blight	3	189	0	2	0	1	3	2	0	0	0
	Healthy	0	0	319	0	0	0	0	0	1	0	0
	Late Blight	0	21	0	351	4	0	1	2	1	0	0
	Leaf Mold	0	0	0	0	188	0	2	1	1	0	0
	Mosaic Virus	0	0	0	0	0	75	0	1	0	0	0
	Septoria Leaf Spot	2	3	0	0	0	0	350	1	0	0	0
	Target Spot	1	1	7	0	0	0	4	258	9	0	0
	Two Spotted Spider Mite	0	1	0	0	0	2	0	17	315	1	0
	Yellow Leaf Curl Virus	16	1	0	0	2	2	0	1	2	0	1046
	Predicted	Bacterial Spot	Early Blight	Healthy	Late Blight	Leaf Mold	Mosaic Virus	Septoria Leaf Spot	Target Spot	Two Spotted Spider Mite	Yellow Leaf Curl Virus	

		Confusion Matrix										
Actual	Bacterial Spot	404	0	0	0	0	1	0	15	4	1	0
	Early Blight	3	174	0	8	2	2	5	5	1	0	0
	Healthy	0	0	320	0	0	0	0	0	0	0	0
	Late Blight	0	7	0	367	3	0	1	1	1	0	0
	Leaf Mold	0	0	0	0	190	0	0	2	0	0	0
	Mosaic Virus	0	0	0	0	0	76	0	0	0	0	0
	Septoria Leaf Spot	0	2	0	0	0	2	1	350	1	0	0
	Target Spot	1	0	5	0	0	0	1	270	3	0	0
	Two Spotted Spider Mite	0	0	0	0	1	0	0	10	323	2	0
	Yellow Leaf Curl Virus	11	0	0	0	2	1	1	0	2	0	1053
	Predicted	Bacterial Spot	Early Blight	Healthy	Late Blight	Leaf Mold	Mosaic Virus	Septoria Leaf Spot	Target Spot	Two Spotted Spider Mite	Yellow Leaf Curl Virus	

(α) Διασταυρούμενη εντροπία (CE)

(β) Εστιακή απώλεια (FL)

		Confusion Matrix										
Actual	Bacterial Spot	420	0	0	0	0	0	3	2	0	0	0
	Early Blight	1	189	0	4	1	0	3	2	0	0	0
	Healthy	0	0	320	0	0	0	0	0	0	0	0
	Late Blight	0	7	1	372	0	0	0	0	0	0	0
	Leaf Mold	0	0	0	0	191	0	1	0	0	0	0
	Mosaic Virus	0	0	0	0	0	75	1	0	0	0	0
	Septoria Leaf Spot	1	1	0	1	1	0	351	1	0	0	0
	Target Spot	1	0	15	0	0	0	6	254	4	0	0
	Two Spotted Spider Mite	0	1	1	1	1	0	2	9	321	0	0
	Yellow Leaf Curl Virus	10	0	0	0	2	0	0	0	2	0	1056
	Predicted	Bacterial Spot	Early Blight	Healthy	Late Blight	Leaf Mold	Mosaic Virus	Septoria Leaf Spot	Target Spot	Two Spotted Spider Mite	Yellow Leaf Curl Virus	

(γ) Επανασταθμισμένη εστιακή απώλεια (RFL)

Σχήμα 4.9: Μήτρες σύγχυσης με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.

Συγκεκριμένα, η συνολική ακρίβεια αυξάνεται από 96,56% με την διασταυρούμενη εντροπία σε 97,63% με την επανασταθμισμένη εστιακή απώλεια. Αυτή η σταδιακή βελτίωση υποδηλώνει ότι με την εφαρμογή της επανασταθμισμένης εστιακής απώλειας, το δίκτυο διαχειρίζεται καλύτερα την ανισορροπία στο σύνολο δεδομένων, οδηγώντας σε μεγαλύτερο αριθμό ορθών προβλέψεων.

Η μετρική της ακρίβειας των θετικών προβλέψεων, αξιολογεί την ικανότητα του μοντέλου να αναγνωρίζει τις θετικές περιπτώσεις μεταξύ όλων των περιπτώσεων που χαρακτηρίστηκαν ως θετικές, παρουσιάζει επίσης αυξητική πορεία. Ξεκινά από 95,26% με την διασταυρούμενη εντροπία και φτάνει 97,19% με την επανασταθμισμένη εστιακή απώλεια. Αντίστοιχα, η ανάκληση, δηλαδή η ικανότητα του μοντέλου να ανιχνεύει όλες τις θετικές περιπτώσεις, εμφανίζει ανοδική τάση, αυξανόμενη από 96,37% με την διασταυρούμενη εντροπία σε 97,29% με την επανασταθμισμένη εστιακή απώλεια.

Η βαθμολογία F1, αξιολογεί την ισορροπία μεταξύ της ακρίβειας των θετικών προβλέψεων και της ανάκλησης, παρουσιάζει θετική πορεία, μεταβαίνοντας από 95,77% με την διασταυρούμενη εντροπία σε 97,22% με την επανασταθμισμένη εστιακή απώλεια. Η αύξηση της βαθμολογίας F1 είναι ιδιαίτερα σημαντική, καθώς επιβεβαιώνει τη συνολική βελτίωση της απόδοσης του μοντέλου τόσο στην ακρίβεια των θετικών προβλέψεων, όσο και στην ανάκληση, χωρίς να θυσιάζεται το ένα για την ενίσχυση του άλλου.

Επιπλέον, η βαθμολογία AUC δείχνει βελτίωση μεταξύ των συναρτήσεων απώλειας, καταλήγοντας σε τιμή 98,51% με την εφαρμογή της επανασταθμισμένης εστιακής απώλειας. Η βαθμολογία AUC αποτελεί μια συνολική μέτρηση απόδοσης για όλα τα κατώφλια ταξινόμησης και η αύξηση αυτής της τιμής υποδηλώνει την ανώτερη ικανότητα του μοντέλου να διακρίνει μεταξύ ασθενών και υγιών φύλλων σε διαφορετικά επίπεδα κατωφλίων.

Οι Πίνακες 4.6–4.8 παρουσιάζουν συγκριτικά τις μετρικές ποιότητας ανα κατηγορία — ακρίβειας των θετικών προβλέψεων, ανάκλησης και βαθμολογίας F1 — μετά την εφαρμογή της διασταυρούμενης εντροπίας, εστιακής απώλειας και επανασταθμισμένης εστιακής απώλειας. Τα αποτελέσματα αυτά μεταξύ των διαφορετικών συναρτήσεων απωλειών αναδεικνύουν τα πλεονεκτήματα της χρήσης της επανασταθμισμένης εστιακής απώλειας στην αντιμετώπιση της ανισορροπίας των κλάσεων. Αυτό έχει ιδιαίτερη σημασία για τις εφαρμογές σε πραγματικά περιβάλλοντα, όπου το κόστος των ψευδώς αρνητικών μπορεί να μεταφραστεί σε σημαντικές οικονομικές απώλειες λόγω της μειωμένης παραγωγικότητας των καλλιεργειών.

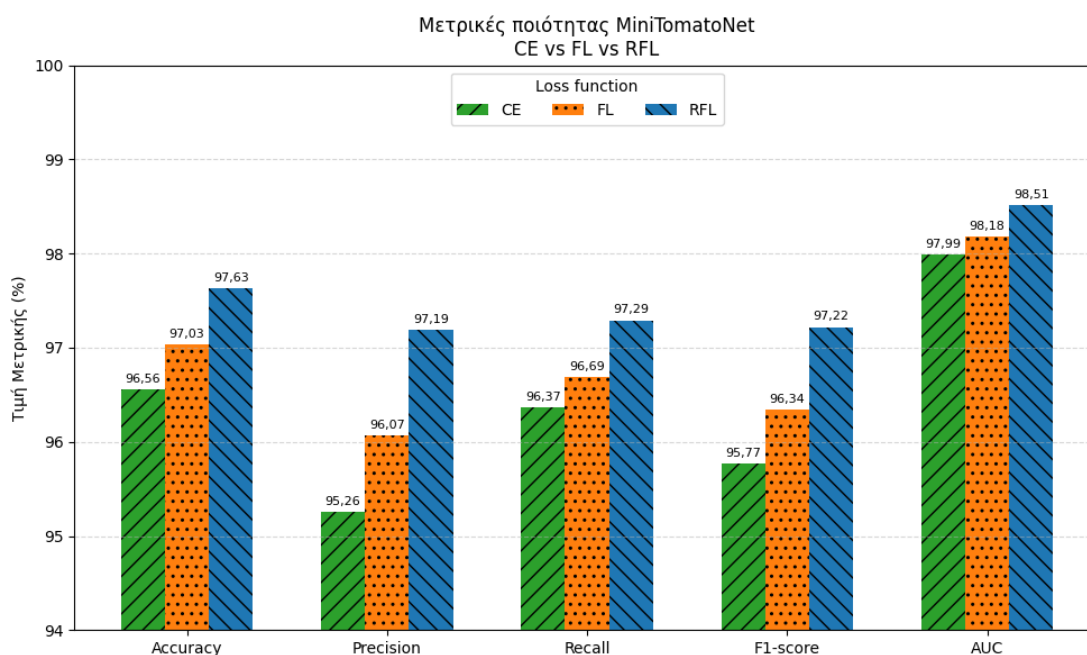
Η προσέγγιση της επανασταθμισμένης εστιακής απώλειας επιτυγχάνει βέλτιστη ισορροπία, οδηγώντας σε ένα πιο αξιόπιστο και ισχυρό μοντέλο χαμηλής πολυπλοκότητας για την αναγνώριση ασθενειών στις τομάτες. Η υιοθέτηση αυτής της μεθόδου είναι ιδιαίτερα κρίσιμη για την έγκαιρη διάγνωση και την αποτελεσματική διαχείριση των ασθενειών στη γεωργία, διασφαλίζοντας τη μέγιστη απόδοση και βιωσιμότητα των καλλιεργειών.

4.6.3 Επιλογή Μεθόδου Συνελικτικής Επεξεργασίας

Η επιλογή της κατάλληλης μεθόδου συνέλιξης αποτελεί κρίσιμο παράγοντα για την ανάπτυξη CNNs σε FPGA, δεδομένου ότι το μέγεθος του μοντέλου επηρεάζει

Πίνακας 4.5: Μέσοι όροι των μετρικών ποιότητας του MiniTomatoNet στα δεδομένα ελέγχου, με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.

Average quality metrics	CE (%)	FL (%)	RFL (%)
Accuracy	96,56	97,03	97,63
Precision	95,26	96,07	97,19
Recall	96,37	96,69	97,29
F1—score	95,77	96,34	97,22
AUC	97,99	98,18	98,51



Σχήμα 4.10: Συγκριτική απεικόνιση των μέσων όρων των μετρικών ποιότητας, με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.

άμεσα τη χρήση των διαθέσιμων υπολογιστικών πόρων και την αποδοτικότητα της υλοποίησης, σύμφωνα με την μεθοδολογία (βλ. Ενότητα 3.4.2). Ο Πίνακας 4.9 και το Σχήμα 4.11 συνοψίζει την αξιολόγηση της εφαρμογής της παραδοσιακής συνέλιξης και της συνέλιξης βάθους ανά κανάλι στο μοντέλο MiniTomatoNet κατά την υλοποίησή του σε FPGA.

Η χρήση της παραδοσιακής συνέλιξης έχει ως αποτέλεσμα μεγαλύτερο μέγεθος μοντέλου, το οποίο ανέρχεται στα 1.198,08 KB, σε σύγκριση με τα 89,51 KB για τη συνέλιξη βάθους ανά κανάλι. Αυτή η μείωση του μεγέθους του μοντέλου είναι ιδιαίτερα σημαντική για τα FPGAs, τα οποία διαθέτουν περιορισμένη ενσωματωμένη μνήμη. Συνεπώς, η ελαχιστοποίηση του μεγέθους του μοντέλου μέσω της συνέλιξης

Πίνακας 4.6: Ακρίβεια των θετικών προβλέψεων ανά κατηγορία για το MiniTomatoNet με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.

Category	CE (%)	FL (%)	RFL (%)
Bacterial Spot	95,01	96,42	97,00
Early Blight	87,10	95,08	95,45
Healthy	97,26	98,46	94,96
Late Blight	99,43	97,87	98,41
Leaf Mold	96,91	94,53	97,45
Mosaic Virus	93,75	95,00	100,00
Septoria Leaf Spot	96,95	93,83	95,64
Target Spot	90,53	92,15	94,78
Two Spotted Spider Mite	95,74	97,58	98,17
Yellow Leaf Curl Virus	99,90	99,81	100,00

Πίνακας 4.7: Ανάκληση ανά κατηγορία για το MiniTomatoNet, με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.

Category	CE (%)	FL (%)	RFL (%)
Bacterial Spot	98,59	95,06	98,82
Early Blight	94,50	87,00	94,50
Healthy	99,69	100,00	100,00
Late Blight	92,37	96,58	97,89
Leaf Mold	97,92	98,96	99,48
Mosaic Virus	98,68	100,00	98,68
Septoria Leaf Spot	98,31	98,31	98,60
Target Spot	92,14	96,43	90,71
Two Spotted Spider Mite	93,75	96,13	95,54
Yellow Leaf Curl Virus	97,76	98,41	98,69

Πίνακας 4.8: Βαθμολογία F1 ανά κατηγορία για το MiniTomatoNet με εφαρμογή των συναρτήσεων της (α) διασταυρούμενης εντροπίας, (β) εστιακής απώλειας και (γ) επανασταθμισμένης εστιακής απώλειας.

Category	CE (%)	FL (%)	RFL (%)
Bacterial Spot	96,77	95,73	97,90
Early Blight	90,65	90,86	94,97
Healthy	98,46	99,22	97,41
Late Blight	95,77	97,22	98,15
Leaf Mold	97,41	96,69	98,45
Mosaic Virus	96,15	97,44	99,34
Septoria Leaf Spot	97,63	96,02	97,10
Target Spot	91,33	94,24	92,70
Two Spotted Spider Mite	94,74	96,85	96,83
Yellow Leaf Curl Virus	98,82	99,11	99,34

βάθους ανά κανάλι επιτρέπει την καλύτερη αξιοποίηση των υπολογιστικών πόρων του FPGA, μειώνοντας τις απαιτήσεις σε μνήμη και αυξάνοντας την αποδοτικότητα της υπολογιστικής διαδικασίας.

Επιπλέον, ο αριθμός των παραμέτρων του μοντέλου, ο οποίος ανέρχεται σε 307.198 K για την παραδοσιακή συνέλιξη και μόλις 22.915 K για τη τη συνέλιξη βάθους ανά κανάλι, έχει άμεσες επιπτώσεις στην υλοποίηση στο FPGA, σύμφωνα με την μεθοδολογία (βλ. Ενότητα 3.4.3). Ένας μικρότερος αριθμός παραμέτρων σημαίνει ότι το μοντέλο απαιτεί μικρότερο εύρος ζώνης για τη φόρτωση των βαρών στο FPGA και μπορεί να εκτελεί υπολογισμούς πιο αποτελεσματικά, μειώνοντας την ανάγκη πρόσβασης στη μνήμη.

Επιπροσθέτως, το συνολικό πλήθος πράξεων κινητής υποδιαστολής που απαιτούνται για την εκτέλεση όλων των συνελικτικών επιπέδων (ConvFLOPs) αποτελεί άμεσο δείκτη της υπολογιστικής πολυπλοκότητας, σύμφωνα με την μεθοδολογία (βλ. Ενότητα 3.4.3). Η παραδοσιακή συνέλιξη απαιτεί 125,70 M FLOP, ενώ η συνέλιξη βάθους ανά κανάλι απαιτεί μόλις 16,954 M FLOP. Η διαφορά αυτή καταδεικνύει ότι η συνέλιξη βάθους ανά κανάλι είναι πιο αποδοτική υπολογιστικά, γεγονός που μεταφράζεται σε μειωμένη κατανάλωση ισχύος.

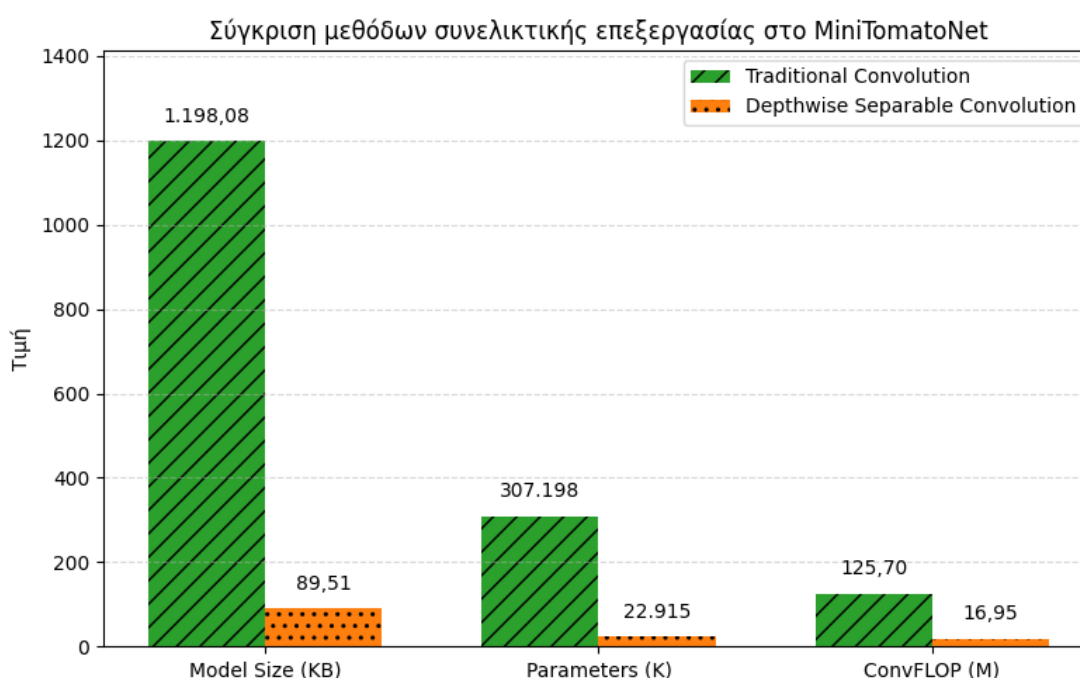
Η εφαρμογή της συνέλιξης βάθους ανά κανάλι στην υλοποίηση του MiniTomatoNet σε FPGA προσφέρει πολλαπλά οφέλη. Συγκεκριμένα:

- Μείωση του μεγέθους του μοντέλου: Ελαχιστοποιεί τις απαιτήσεις αποθήκευσης, διευκολύνοντας την πλήρη αξιοποίηση των διαθέσιμων πόρων ενσωματωμένης μνήμης και καθιστώντας δυνατή την εκτέλεση του δικτύου χωρίς την ανάγκη εξωτερικής μνήμης.
- Μείωση του αριθμού παραμέτρων: Κατακερματίζοντας τη συνέλιξη σε πράξεις ανά κανάλι, μειώνεται ο συνολικός αριθμός παραμέτρων. Αυτό περιορίζει την ανάγκη μεταφοράς δεδομένων από και προς τη μνήμη, γεγονός που βελτιώνει τη ροή και επιταχύνει την εκτέλεση.
- Περιορισμός της υπολογιστικής πολυπλοκότητας: Η αρχιτεκτονική της συνέλιξης βάθους ανά κανάλι μειώνει τον αριθμό των απαιτούμενων FLOP, γεγονός που μεταφράζεται σε ταχύτερους χρόνους επεξεργασίας και μειωμένη ενεργειακή απαίτηση.

Αυτή η βελτιστοποίηση καθιστά τη συγκεκριμένη αρχιτεκτονική πιο κατάλληλη για εφαρμογές με περιορισμένους υπολογιστικούς πόρους, επιτυγχάνοντας υψηλή αποδοτικότητα με μειωμένες απαιτήσεις σε υλικό και ενέργεια.

Πίνακας 4.9: Σύγκριση διαφορετικών μεθόδων συνελικτικής επεξεργασίας στο μοντέλο MiniTomatoNet.

	Traditional Convolution	Depthwise Separable Convolution	Optimization (%)
Model Size (KB)	1.198,08	89,51	92,5%
Parameters (K)	307.198	22.915	92,5%
ConvFLOP (M)	125,70	16,954	86,5%



Σχήμα 4.11: Συγκριτική απεικόνιση των διαφορετικών μεθόδων συνελικτικής επεξεργασίας στο μοντέλο MiniTomatoNet.

4.6.4 Υπολογιστικό Κόστος των Τύπων Δεδομένων

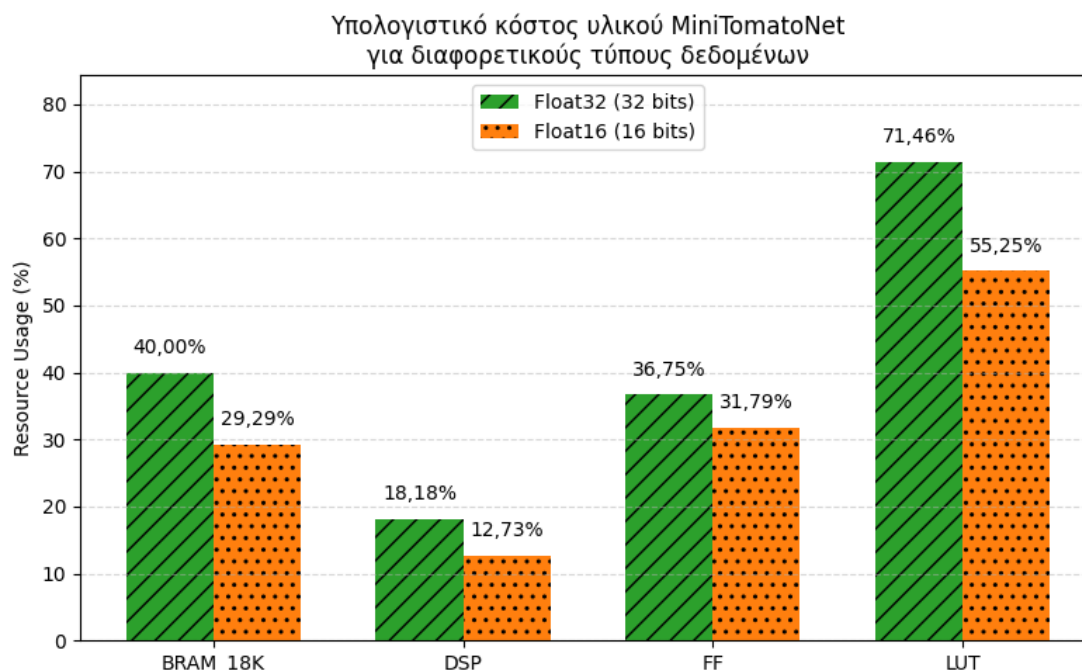
Η επιλογή των τύπων δεδομένων για τις παραμέτρους του μοντέλου και την αναπαράσταση των χαρακτηριστικών επηρεάζει σημαντικά την υπολογιστική απόδοση του MiniTomatoNet. Για να εκτιμηθεί η επίδραση των διαφορετικών τύπων δεδομένων στο κυκλωματικό κόστος του MiniTomatoNet, πραγματοποιήθηκε πειραματική ανάλυση με έμφαση την σύγκριση μεταξύ της αναπαράστασης της κινητής υποδιαστολής 32-bit (FP32) μονής ακρίβειας και 16-bit (FP16) μισής

ακρίβειας, σύμφωνα με την μεθοδολογία (βλ. Ενότητα 3.2.2.1). Η υλοποίηση FP16 πραγματοποιήθηκε με τη χρήση της βιβλιοθήκης `hls_half.h` της Xilinx.

Η αρχιτεκτονική του MiniTomatoNet, όπως και κάθε CNN, χαρακτηρίζεται από μεγάλο πλήθος πράξεων πολλαπλασιασμού και πρόσθεσης, και συνεπώς ωφελείται από την υλοποίηση αυτών των πράξεων σε μειωμένη αριθμητική ακρίβεια. Οι αριθμητικές μονάδες μισής ακρίβειας, όπως πολλαπλασιαστές και αθροιστές, απαιτούν λιγότερους υπολογιστικούς πόρους του FPGA σε σχέση με τις αντίστοιχες μονάδες μονής ακρίβειας. Η ανάλυση της χρήσης πόρων του MiniTomatoNet για τους διαφορετικούς τύπους δεδομένων, παρουσιάζεται στον Πίνακα 4.10 και στο Σχήμα 4.12.

Πίνακας 4.10: Υπολογιστικό κόστος υλικού του MiniTomatoNet για διαφορετικούς τύπους δεδομένων (FP32 και FP16).

Resource	Float (32 bits)	Half (16 bits)	Available	Optimization (%)
BRAM_18K	112 (40%)	82 (29,29%)	280	26,79%
DSP	40 (18,18%)	28 (12,73%)	220	30%
FF	39.102 (36,75%)	33.824 (31,79%)	106.400	13,49%
LUT	38.012 (71,46%)	29.396 (55,25%)	53.200	22,65%



Σχήμα 4.12: Συγκριτική απεικόνιση του υπολογιστικού κόστους υλικού των διαφορετικών τύπων δεδομένων (FP32 και FP16) στο μοντέλο MiniTomatoNet.

Από τα δεδομένα του Πίνακα 4.10, παρατηρούμε ότι η υλοποίηση του MiniTomatoNet με τη μορφή της αναπαράστασης κινητής υποδιαστολής μονής ακρίβειας

καταναλώνει σημαντικά περισσότερους πόρους σε όλες τις κατηγορίες, σε σύγκριση με την υλοποίηση της μισής ακρίβειας. Ειδικότερα:

- Χρήση μπλοκ μνήμης RAM (BRAM_18K): Η υλοποίηση με τη μορφή κινητής υποδιαστολής μονής ακρίβειας απαιτεί 112 μπλοκ, αντιπροσωπεύοντας το 40% των διαθέσιμων 280 μπλοκ, ενώ η αντίστοιχη με μισής ακρίβειας χρησιμοποιεί μόνο 82 μπλοκ (29,29%), επιτυγχάνοντας ποσοστό βελτιστοποίησης 26,79%.
- Κατανάλωση μονάδων επεξεργασίας ψηφιακού σήματος (DSPs): Στην περίπτωση της μονής ακρίβειας, χρησιμοποιούνται 40 slices (18,18%) των διαθέσιμων πόρων, ενώ της μισής ακρίβειας καταναλώνει μόλις 28 slices (12,73%), επιτυγχάνοντας μείωση κατά 30%.
- Χρήση στοιχείων μνήμης καταχωρητών (FFs): Η υλοποίηση με τη μορφή μονής ακρίβειας καταναλώνει 39.102 flip—flops (36,75%), ενώ εκείνη με μισή ακρίβεια περιορίζεται σε 33.824 flip—flops (31,79%), με αντίστοιχη βελτιστοποίηση της τάξης του 13,49%.
- Κατανάλωση πινάκων αναζήτησης (LUTs): Η μεγαλύτερη εξοικονόμηση πόρων καταγράφεται στους πίνακες αναζήτησης, με την μονή ακρίβεια να καταναλώνει 71,46% των διαθέσιμων LUTs, ενώ η υλοποίηση με μισή ακρίβεια απαιτεί μόλις 55,25%, οδηγώντας σε βελτιστοποίηση 22,65%.

Η εφαρμογή του MiniTomatoNet με τη μορφή της κινητής υποδιαστολής μισής ακρίβειας επιτυγχάνει σημαντική εξοικονόμηση πόρων υλικού, μειώνοντας τόσο το υπολογιστικό κόστος, όσο και την κατανάλωση ισχύος. Αυτή η βελτιστοποίηση καθιστά το δίκτυο πιο αποδοτικό για υλοποιήσεις σε FPGA, όπου οι διαθέσιμοι πόροι είναι περιορισμένοι. Συνεπώς, η χρήση της κινητής υποδιαστολής μισής ακρίβειας είναι ιδιαίτερα επωφελής για εφαρμογές που απαιτούν υψηλή αποδοτικότητα σε όρους ενέργειας και επιδόσεων, καθιστώντας δυνατή την ανάπτυξη χαμηλής πολυπλοκότητας και αποδοτικών επιταχυντών CNN σε HCS.

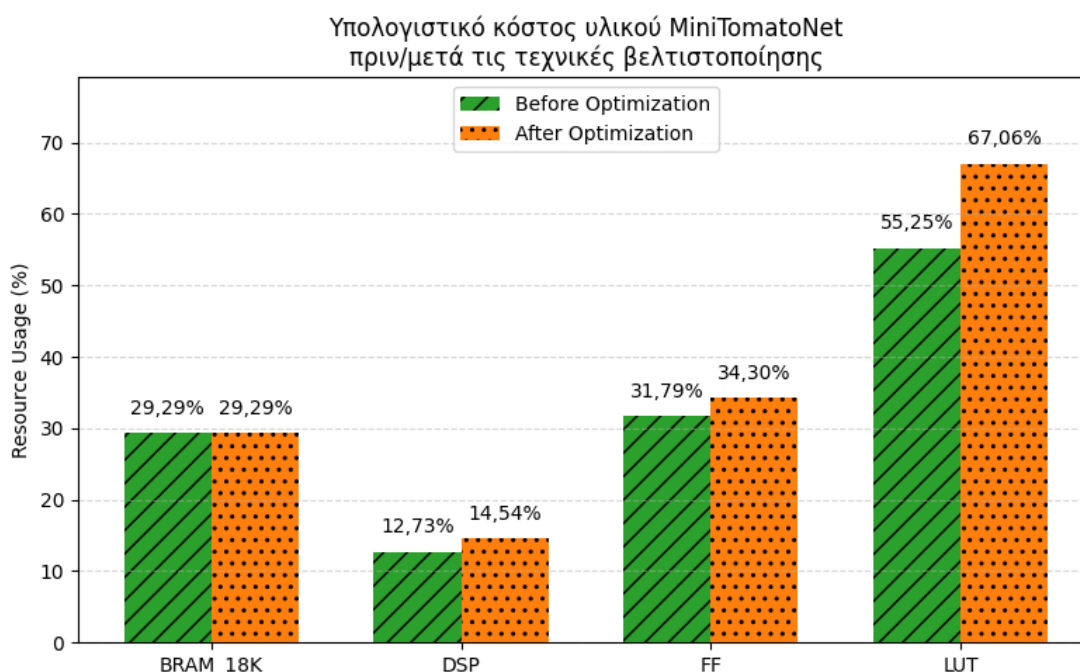
4.6.5 Υπολογιστικό Κόστος των Τεχνικών Βελτιστοποίησης

Στην προσπάθεια μεγιστοποίησης της αξιοποίησης των πόρων του FPGA και της επιτάχυνσης της διαδικασίας ανίχνευσης ασθενειών τομάτας, η αρχιτεκτονική MiniTomatoNet έχει υποβληθεί σε μια σειρά τεχνικών βελτιστοποίησης. Οι τεχνικές

αυτές περιλαμβάνουν το ξετύλιγμα βρόχων, τη διασωλήνωση, την κατάτμηση πινάκων και την επιπέδωση βρόχων, οι οποίες εφαρμόστηκαν για τη βελτίωση της απόδοσης του κυκλώματος. Η υλοποίηση των τεχνικών βελτιστοποίησης πραγματοποιήθηκε με τις οδηγίες pragma του περιβάλλοντος HLS. Ειδικότερα, εφαρμόστηκαν οι εντολές #pragma HLS UNROLL, #pragma HLS PIPELINE, #pragma HLS ARRAY_PARTITION και #pragma HLS LOOP_FLATTEN.

Πίνακας 4.11: Υπολογιστικό κόστος και καθυστέρηση του MiniTomatoNet των τεχνικών βελτιστοποίησης (ξετύλιγμα βρόχων, διασωλήνωση, κατάτμηση πινάκων και επιπέδωση βρόχων).

Resource	Before	After	Available
BRAM_18K	82 (29,29%)	82 (29,29%)	280
DSP	28 (12,73%)	32 (14,54%)	220
FF	33.824 (31,79%)	36.504 (34,30%)	106.400
LUT	29.396 (55,25%)	35.676 (67,06%)	53.200
Latency (M cycles)	8,42	1,84	—



Σχήμα 4.13: Συγκριτική απεικόνιση του υπολογιστικού κόστους υλικού των τεχνικών βελτιστοποίησης (ξετύλιγμα βρόχων, διασωλήνωση, κατάτμηση πινάκων και επιπέδωση βρόχων) στο μοντέλο MiniTomatoNet.

Ο Πίνακας 4.11 και το Σχήμα 4.13 παρέχει μια συγκριτική ανάλυση που παρουσιάζει την κατανομή των πόρων και τις μετρικές χρονικής καθυστέρησης πριν και μετά την εφαρμογή αυτών των στρατηγικών βελτιστοποίησης. Από τον Πίνακα 4.11, παρατηρούμε ότι η χρήση της BRAM παρέμεινε σταθερή στις 82 μονάδες, που

αντιστοιχούν στο 29,29% της συνολικής διαθέσιμης μνήμης BRAM του FPGA. Αυτό υποδηλώνει ότι οι τεχνικές βελτιστοποίησης δεν επηρέασαν τις απαιτήσεις σε μνήμη block RAM του MiniTomatoNet. Αντιθέτως, η κατανάλωση DSP αυξήθηκε ελαφρώς από 28 σε 32 μετά τη βελτιστοποίηση, σημειώνοντας αύξηση από 12,73% σε 14,54% των διαθέσιμων DSP πόρων. Επιπλέον, η χρήση των FFs αυξήθηκε από 33.824 σε 36.504, καταλαμβάνοντας πλέον το 34,30% των συνολικών διαθέσιμων FFs, σε σχέση με το 31,79% πριν την εφαρμογή των βελτιστοποιήσεων. Η μεγαλύτερη αύξηση αφορά την κατανάλωση LUTs, όπου η χρήση τους αυξήθηκε από 29.396 (55,25%) σε 35.676 (67,06%) του συνόλου των διαθέσιμων πόρων.

Το πιο σημαντικό αποτέλεσμα, ωστόσο, είναι η δραστική μείωση της χρονικής καθυστέρησης. Συγκεκριμένα, η καθυστέρηση μειώθηκε από 8,42 M κύκλους σε μόλις 1,84 M κύκλους, σημειώνοντας μείωση σχεδόν 78%. Αυτή η μείωση επιβεβαιώνει την αποτελεσματικότητα των τεχνικών βελτιστοποίησης που εφαρμόστηκαν. Η μείωση της χρονικής καθυστέρησης δεν σημαίνει μόνο ταχύτερους χρόνους επεξεργασίας αλλά και σημαντική εξοικονόμηση ενέργειας, καθώς το κύκλωμα λειτουργεί για μικρότερες χρονικές περιόδους, σύμφωνα με την μεθοδολογία (βλ. Ενότητα 3.4.2).

Αυτή η βελτίωση καθιστά το MiniTomatoNet εξαιρετικά κατάλληλο για εφαρμογές όπου η ταχύτητα λήψης αποφάσεων είναι κρίσιμη. Μέσω της μείωσης της χρονικής καθυστέρησης και της εξοικονόμησης πόρων, το μοντέλο μπορεί να λειτουργεί αποδοτικότερα σε αρχιτεκτονικές FPGA, όπου οι υπολογιστικοί και ενεργειακοί περιορισμοί είναι κρίσιμοι παράγοντες. Έτσι, οι βελτιστοποιήσεις που εφαρμόστηκαν στο MiniTomatoNet δεν επηρέασαν την απαιτούμενη χωρητικότητα μνήμης, αλλά πέτυχαν σημαντική αύξηση στην υπολογιστική αποδοτικότητα και τη μείωση της χρονικής καθυστέρησης.

4.6.6 Συγκριτική Ανάλυση σε Διαφορετικές Υπολογιστικές Συσκευές

Ο Πίνακας 4.12 παρουσιάζει τη σύγκριση των μετρικών απόδοσης της υλοποίησης του MiniTomatoNet σε τρεις διαφορετικές υπολογιστικές συσκευές: το ZYNQ XC7Z020, το Jetson AGX Xavier και την GeForce RTX 3050 GPU. Αυτή η ανάλυση εστιάζει στις διαφορές μεταξύ της συχνότητας ρολογιού, της ταχύτητας εξαγωγής συμπερασμάτων και της κατανάλωσης ενέργειας.

Στην υλοποίηση του MiniTomatoNet στο FPGA ZYNQ XC7Z020 η συχνότητα του ρολογιού είναι 100 MHz, με χρόνο εξαγωγής συμπερασμάτων 0,068 s (βλ. Ενότητα 3.2.3). Παρά την σχετικά χαμηλή συχνότητα ρολογιού σε σύγκριση με τις άλλες συσκευές, το FPGA παραμένει εξαιρετικά αποδοτικό σε όρους ισχύος, καταναλώνοντας μόλις 2,35 W (βλ. Ενότητα 3.2.4). Αυτό το χαρακτηριστικό το καθιστά ιδανικό για εφαρμογές όπου η ενεργειακή απόδοση είναι πρωταρχικός στόχος, όπως τα ενσωματωμένα συστήματα.

Στην περίπτωση του Jetson AGX Xavier, η συχνότητα ρολογιού αυξάνεται σημαντικά στα 1.377 MHz, επιτρέποντας την ταχύτερη επεξεργασία με χρόνο εξαγωγής συμπερασμάτων 0,024 s. Ωστόσο, αυτή η αύξηση στη συχνότητα ρολογιού συνεπάγεται και μεγαλύτερη κατανάλωση ενέργειας, η οποία ανέρχεται σε 8,58 W. Αυτό το σύστημα αποτελεί μια ενδιαφέρουσα λύση που συνδυάζει υψηλότερη υπολογιστική ισχύ από το FPGA, καθιστώντας το κατάλληλο για εφαρμογές όπου η ταχύτητα επεξεργασίας είναι πιο κρίσιμη, ενώ η κατανάλωση ενέργειας παραμένει σημαντική παράμετρος.

Η GeForce RTX 3050 GPU εμφανίζει την υψηλότερη συχνότητα ρολογιού στα 1.552 MHz, επιτυγχάνοντας τη μικρότερη καθυστέρηση με χρόνο εξαγωγής συμπερασμάτων 0,016 s. Ωστόσο, αυτό το πλεονέκτημα στην ταχύτητα έχει ως αντίτιμο μια εξαιρετικά αυξημένη κατανάλωση ισχύος, η οποία φτάνει τα 130 W. Παρόλο που η RTX 3050 προσφέρει την ταχύτερη επεξεργασία, το υψηλό ενεργειακό κόστος την καθιστά λιγότερο ιδανική για συστήματα με περιορισμούς κατανάλωσης ενέργειας.

Από τη συγκριτική ανάλυση, προκύπτουν τα εξής συμπεράσματα:

- Το FPGA (ZYNQ XC7Z020) παρέχει την καλύτερη ενεργειακή αποδοτικότητα, διατηρώντας παράλληλα ικανοποιητική υπολογιστική απόδοση. Είναι ιδανικό για εφαρμογές με περιορισμένα ενεργειακά αποθέματα.
- Το Jetson AGX Xavier αποτελεί μια ισορροπημένη επιλογή μεταξύ ισχύος και κατανάλωσης, καθιστώντας το κατάλληλο για εφαρμογές που απαιτούν υψηλή απόδοση, αλλά ελεγχόμενη ενεργειακή κατανάλωση.
- Η GeForce RTX 3050 GPU προσφέρει την υψηλότερη υπολογιστική ισχύ και την ταχύτερη επεξεργασία, όμως η κατανάλωση ενέργειας είναι πολλαπλάσια σε σχέση με τις άλλες επιλογές, γεγονός που περιορίζει τη χρήση της σε συστήματα όπου η ενεργειακή απόδοση δεν είναι προτεραιότητα.

Αυτά τα αποτελέσματα καταδεικνύουν τη σημασία της επιλογής της κατάλληλης υπολογιστικής συσκευής ανάλογα με τις απαιτήσεις μιας εφαρμογής. Για εφαρμογές χαμηλής κατανάλωσης και πραγματικού χρόνου, το FPGA είναι η βέλτιστη λύση. Αντίθετα, αν η ταχύτητα αποτελεί τον καθοριστικό παράγοντα, τότε τα ισχυρότερα υπολογιστικά συστήματα, όπως οι GPUs, μπορούν να αξιοποιηθούν με το αντίστοιχο ενεργειακό κόστος.

Πίνακας 4.12: Επίδραση της υλοποίησης MiniTomatoNet σε διαφορετικές υπολογιστικές συσκευές.

Device	Clock (MHz)	Inference speed (s)	Power (W)
ZYNQ XC7Z020	100	0,068	2,35
Jetson AGX Xavier	1.377	0,024	8,58
GeForce RTX 3050	1.552	0,016	130

4.7 Συζήτηση Αποτελεσμάτων

Στην επιδίωξη της προόδου στην αγροτεχνολογία, και ειδικότερα στη διάγνωση ασθενειών των φυτών, εξακολουθεί να υφίσταται ένα σημαντικό ερευνητικό κενό. Το κενό αυτό εντοπίζεται στην ανάγκη για ταχύτατες και ακριβείς μεθόδους αναγνώρισης, οι οποίες ταυτόχρονα πρέπει να είναι οικονομικά αποδοτικές και πρακτικά εφαρμόσιμες σε πραγματικές συνθήκες καλλιέργειας. Το σημερινό τοπίο έρευνας προσφέρει λύσεις που είτε συμβιβάζουν την ακρίβεια για χάρη της απλότητας είτε απαιτούν εξελιγμένο και δαπανηρό υλικό. Αυτή η διάσταση αποτελεί μια σημαντική ερευνητική πρόκληση, η οποία μπορεί να αντιμετωπιστεί μέσω προσέγγισης που αξιοποιεί τη συνεργασία μεταξύ CNN μοντέλων και FPGA συστημάτων.

Η αρχιτεκτονική MiniTomatoNet σχεδιάστηκε για την ανίχνευση ασθενειών στα φύλλα τομάτας, επιτυγχάνοντας έναν ισορροπημένο συνδυασμό υπολογιστικής αποδοτικότητας και ακρίβειας εξαγωγής συμπερασμάτων. Παρόλο που η ανάπτυξη ενός αποτελεσματικού CNN αποτελεί σημαντικό βήμα, δεν αποτελεί από μόνη της ολοκληρωμένη λύση για τις προκλήσεις της ανίχνευσης ασθενειών. Η υλοποίηση ενός τέτοιου δικτύου σε πραγματικές συνθήκες απαιτεί υπολογιστικές συσκευές που να είναι ισχυρές, ενεργειακά αποδοτικές και οικονομικά βιώσιμες, διατηρώντας παράλληλα την ικανότητα λειτουργίας σε ποικίλες περιβαλλοντικές συνθήκες.

Τα FPGA προσφέρουν ένα σημαντικό πλεονέκτημα χάρη στην επαναπρογραμματιζόμενη φύση τους, επιτρέποντας εξειδικευμένο συν—σχεδιασμό υλικού—λογισμικού, το οποίο αξιοποιεί πλήρως τον ενσωματωμένο παραλληλισμό των CNNs. Αυτό σημαίνει ότι η υλοποίηση του MiniTomatoNet σε FPGA επιτρέπει την ακριβή προσαρμογή του υλικού στις ανάγκες του δικτύου, οδηγώντας σε αποδοτικότερη χρήση των πόρων και ταχύτερους χρόνους εξαγωγής συμπερασμάτων σε σχέση με γενικού σκοπού επεξεργαστές. Επιπλέον, η ενεργειακή αποδοτικότητα των FPGAs είναι ιδιαίτερα σημαντική για γεωργικές εφαρμογές, όπου οι διαθέσιμες πηγές ισχύος είναι περιορισμένες.

Το MiniTomatoNet, υλοποιημένο σε FPGA, μπορεί να λειτουργήσει σε πραγματικό χρόνο, παρέχοντας άμεση ανίχνευση ασθενειών και βοηθώντας στην πρόληψη εξάπλωσης ασθενειών στις καλλιέργειες. Ωστόσο, η προσέγγιση αυτή δεν είναι χωρίς προκλήσεις. Παρόλο που τα FPGAs διατηρούν μια καλή ισορροπία μεταξύ απόδοσης και κατανάλωσης ενέργειας, υπολείπονται των GPU σε υπολογιστική ισχύ. Συνεπώς, η συνεχής βελτιστοποίηση των στρατηγικών υλοποίησης CNNs σε FPGAs αποτελεί ερευνητική πρόκληση, με στόχο την περαιτέρω μείωση της απόκλισης των μετρικών απόδοσης.

Στην συγκριτική ανάλυση που παρουσιάζεται στον Πίνακα 4.9, διερευνάται η επίδραση της χρήσης της παραδοσιακής συνέλιξης έναντι της συνέλιξης βάθους ανά κανάλι στο MiniTomatoNet σε FPGA. Η σημαντική μείωση του μεγέθους του μοντέλου, του αριθμού των παραμέτρων και των απαιτούμενων υπολογιστικών πόρων καταδεικνύει την αξία της βελτιστοποίησης στο FPGA. Η μετάβαση από την παραδοσιακή συνέλιξη με υψηλές απαιτήσεις πόρων (1.198,08 KB σε μέγεθος και 125,70 M ConvFLOP) σε συνέλιξη βάθους ανά κανάλι (89,51 KB και 16,954 M ConvFLOP) υπογραμμίζει τη σημαντική αύξηση της ενεργειακής αποδοτικότητας. Αυτή η βελτιστοποίηση δεν είναι απλώς μια μείωση των αριθμητικών τιμών αλλά ενσωματώνει μια κρίσιμη προσαρμογή σε περιβάλλοντα με περιορισμένους υπολογιστικούς και αποθηκευτικούς πόρους.

Επιπλέον, όταν εξετάζουμε τις παραλλαγές αριθμητικής ακριβείας που επισημαίνονται στον Πίνακα 4.10, η οικονομία πόρων της μορφής αναπαράστασης μισής ακρίβειας 16—bit είναι εμφανής, ιδιαίτερα όσον αφορά την κατανάλωση των τμημάτων LUT και DSP. Παρά τη μικρή αύξηση της χρήσης DSP μετά τη βελτιστοποίηση, η σημαντική μείωση της χρονικής καθυστέρησης κατά περίπου 78% αποδεικνύει την ενίσχυση της υπολογιστικής απόδοσης και της ταχείας απόκρισης του συστήματος. Η συνεχής χρήση της BRAM σε συνδυασμό με τη

μειωμένη χρονική καθυστέρηση υποδηλώνει ότι η χωρητικότητα της μνήμης διατηρείται ενώ βελτιώνεται η απόδοση του συστήματος — μια κρίσιμη ισορροπία στη σχεδίαση FPGA.

Τέλος, από τα αποτελέσματα στον Πίνακα 4.12, παρατηρούμε ότι το FPGA ZYNQ XC7Z020 με συχνότητα 100 MHz επιτυγχάνει χρόνο εξαγωγής συμπερασμάτων 0,068 s και χαμηλή κατανάλωση ισχύος 2,35 W. Αυτό επιβεβαιώνει ότι το FPGA είναι ιδανικό για ενεργειακά αποδοτικές εφαρμογές, ειδικά σε αυτόνομα αγροτικά συστήματα με περιορισμένη πρόσβαση σε ενέργεια. Αντίθετα, η GPU NVIDIA GeForce RTX 3050, με πολύ υψηλότερη συχνότητα ρολογιού (1.552 MHz), πετυχαίνει χρόνο εξαγωγής συμπερασμάτων 0,016 s, αλλά με σημαντικά αυξημένη κατανάλωση ισχύος στα 130 W.

Αντίστοιχα, το Jetson AGX Xavier προσφέρει μια ενδιάμεση λύση, συνδυάζοντας υψηλή υπολογιστική ισχύ με χαμηλότερη κατανάλωση ισχύος από τη GPU, γεγονός που το καθιστά κατάλληλο για εφαρμογές όπου η ταχύτητα επεξεργασίας είναι πρωταρχικός στόχος, αλλά η ενεργειακή αποδοτικότητα εξακολουθεί να είναι σημαντική παράμετρος. Αυτό τοποθετεί το Xavier ως μια ισορροπημένη επιλογή για εφαρμογές που απαιτούν ταχύτατη υπολογιστική απόδοση χωρίς υπερβολική κατανάλωση ενέργειας, γεφυρώνοντας το χάσμα μεταξύ της υψηλής απόδοσης συστημάτων που βασίζονται σε FPGA και της υψηλής ταχύτητας επεξεργασίας των GPU με μεγαλύτερη ένταση ισχύος.

Η ανάλυση αυτής της έρευνας καταδεικνύει ότι η προσεκτική επιλογή τεχνικών βελτιστοποίησης και η χρήση χαμηλής πολυπλοκότητας δομοστοιχείων CNNs μπορούν να αποφέρουν σημαντικά οφέλη στην απόδοση του FPGA. Οι βελτιστοποιημένες κυκλωματικές υλοποιήσεις του MiniTomatoNet συνδυάζουν την αποδοτική αξιοποίηση των διαθέσιμων πόρων με την ικανότητα εκτέλεσης πολύπλοκων υπολογισμών. Αυτό ανοίγει το δρόμο για την ανάπτυξη πιο εξελιγμένων νευρωνικών μοντέλων που μπορούν να εκτελούνται σε FPGA, χωρίς συμβιβασμούς σε ταχύτητα ή ενεργειακή αποδοτικότητα.

4.8 Συμπεράσματα και Μελλοντικές Επεκτάσεις

Η παρούσα έρευνα σχεδιάστηκε με βάση την πολυκριτηριακή μεθοδολογία που αναπτύξαμε, η οποία συνδυάζει μετρικές σχεδίασης, ποιότητας και υπολογιστικού αποτυπώματος. Η μεθοδολογική αυτή προσέγγιση επέτρεψε τη διερεύνηση των

αρχιτεκτονικών επιλογών, των τεχνικών βελτιστοποίησης και των παραμέτρων αποδοτικότητας του συστήματος, τόσο σε επίπεδο CNN όσο και σε επίπεδο υλοποίησης σε FPGA–SoC. Η αξιολόγηση πραγματοποιήθηκε με έμφαση στην ισορροπία μεταξύ συνολικής ακρίβειας, ταχύτητας εξαγωγής συμπερασμάτων και απαιτήσεων σε πόρους, διασφαλίζοντας την εφαρμοσιμότητα του προτεινόμενου συστήματος σε πραγματικές συνθήκες αγροτικής παραγωγής.

Σχεδιάσαμε και υλοποιήσαμε ένα χαμηλής πολυπλοκότητας CNN, το MiniTomatoNet, για τη διάγνωση ασθενειών φύλλων τομάτας σε μια ετερογενή πλατφόρμα FPGA–SoC. Το MiniTomatoNet χαρακτηρίζεται από μια βελτιστοποιημένη αρχιτεκτονική σχεδίαση, με λιγότερες από 23 K παραμέτρους, συμπεριλαμβανομένων όλων των βαρών του δικτύου. Ένα κρίσιμο στοιχείο της προτεινόμενης προσέγγισης αποτελεί η αντιμετώπιση του προβλήματος ανισορροπίας δεδομένων, το οποίο αποτελεί μόνιμη πρόκληση στη διάγνωση ασθενειών των φυτών. Η χρήση βελτιστοποιημένων τεχνικών επιτάχυνσης κυκλωμάτων διασφάλισε ότι η υλοποίηση σε PYNQ–Z1 FPGA–SoC πέτυχε ταχύτερους χρόνους εξαγωγής συμπερασμάτων, ελάχιστη κατανάλωση ισχύος και υψηλή ακρίβεια στη διάγνωση ασθενειών της τομάτας.

Το MiniTomatoNet παρουσίασε υψηλές επιδόσεις, επιτυγχάνοντας συνολική ακρίβεια εξαγωγής συμπερασμάτων 97,63% και βαθμολογία AUC 98,51% στη διάγνωση των ασθενειών της τομάτας. Αυτά τα αποτελέσματα επιβεβαιώνουν τη δυνατότητα ανάπτυξης αποδοτικών και εξειδικευμένων αλγορίθμων DL που να ανταποκρίνονται στις απαιτήσεις πραγματικού χρόνου και στις περιορισμένες υπολογιστικές δυνατότητες των FPGA.

Στο πεδίο της μελλοντικής έρευνας, η επικέντρωση θα πρέπει να εστιάσει στα εξής:

- Εξέταση της δυνατότητας γενίκευσης του MiniTomatoNet μέσω της εφαρμογής του σε ετερογενείς γεωργικές καλλιέργειες πέραν της τομάτας, με στόχο την αξιολόγηση της επεκτασιμότητας και της διαλειτουργικότητας του μοντέλου σε ποικίλα αγρονομικά περιβάλλοντα.
- Βελτιστοποίηση της ενεργειακής αποδοτικότητας μέσω πιο προηγμένων τεχνικών διαχείρισης ισχύος στο FPGA, διασφαλίζοντας ακόμη χαμηλότερη κατανάλωση ενέργειας για εφαρμογές σε απομακρυσμένα και αυτόνομα αγροτικά συστήματα.

- Διερεύνηση της ενσωμάτωσης του MiniTomatoNet με συστήματα IoT, με σκοπό τη δημιουργία έξυπνων αγροτικών συστημάτων που να συνδυάζουν ανάλυση εικόνας σε πραγματικό χρόνο, συλλογή δεδομένων από αισθητήρες και απομακρυσμένη επεξεργασία μέσω συσκευών υπολογισμού στο άκρο.
- Περαιτέρω βελτιώσεις στην ανίχνευση και την ταξινόμηση φυτικών ασθενειών μέσω της αξιοποίησης προηγμένων τεχνικών DL, όπως τα μοντέλα βασισμένα σε μετασχηματιστές και οι μηχανισμοί προσοχής, με στόχο τη μεγιστοποίηση της ακρίβειας εξαγωγής συμπερασμάτων σε ποικίλα περιβάλλοντα καλλιέργειας.

Η επέκταση της έρευνας σε ευρύτερες αγροτικές εφαρμογές θα επιτρέψει τη δημιουργία ενός ολοκληρωμένου συστήματος διαχείρισης ασθενειών στον γεωργικό τομέα. Ένα τέτοιο σύστημα, συνδυάζοντας ευφυείς αλγορίθμους, χαμηλή κατανάλωση ενέργειας και ταχεία εξαγωγή συμπερασμάτων, μπορεί να συμβάλει σημαντικά στην αύξηση της αποδοτικότητας της αγροτικής παραγωγής και στη μείωση των απωλειών λόγω ασθενειών.

Κεφάλαιο 5

Σχεδιασμός και Υλοποίηση Επιτάχυνσης σε SoM

Στο προηγούμενο κεφάλαιο εφαρμόστηκε η προτεινόμενη μεθοδολογία για τον σχεδιασμό και την υλοποίηση επιτάχυνσης CNN σε FPGA–SoC. Στο παρόν κεφάλαιο¹ εφαρμόζεται η ίδια μεθοδολογία στον σχεδιασμό και την υλοποίηση επιτάχυνσης CNN σε σύστημα τύπου SoM. Αρχικά, παρουσιάζεται η αρχιτεκτονική του μοντέλου που επιλέγεται σύμφωνα με τις κατευθυντήριες γραμμές της μεθοδολογίας, με στόχο τη μείωση της πολυπλοκότητας. Στη συνέχεια, αναλύεται η διαδικασία βελτιστοποίησης με βάση τις μετρικές που ορίζει η μεθοδολογία. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση και τη συζήτηση των αποτελεσμάτων της αξιολόγησης της προτεινόμενης υλοποίησης, χρησιμοποιώντας καθιερωμένες μετρικές ποιότητας, καθώς και με συγκρίσεις έναντι εναλλακτικών υλοποιήσεων αρχιτεκτονικών CNN.

5.1 Περίληψη

Η παγκόσμια εξάπλωση των επιδημικών πνευμονικών νοσημάτων, συμπεριλαμβανομένου της νόσου του κορωνοϊού (COVID–19), υπογραμμίζει την ανάγκη για αποτελεσματικές διαγνωστικές μεθόδους. Στο πλαίσιο αυτής της ανάγκης, αναπτύξαμε και αξιολογήσαμε ένα υπολογιστικά αποδοτικό και χαμηλής πολυπλοκότητας CNN για την ταχεία και ακριβή αναγνώριση των πνευμονικών

¹Μέρος του κεφαλαίου έχει δημοσιευθεί στο άρθρο [233].

παθήσεων μέσω 29.131 ακτινογραφιών θώρακα, οι οποίες περιλαμβάνουν επτά διαφορετικές κατηγορίες ασθενειών.

Για τη διασφάλιση της αξιοπιστίας των αποτελεσμάτων, εφαρμόστηκε η μέθοδος της διασταυρούμενης επικύρωσης πέντε πτυχών. Το προτεινόμενο μοντέλο έχει βελτιστοποιηθεί για ετερογενείς ενσωματωμένες συσκευές και παρουσίασε ανώτερη διαγνωστική απόδοση. Συγκεκριμένα, επιτεύχθηκε συνολική ακρίβεια εξαγωγής συμπερασμάτων 98,56%, ξεπερνώντας την απόδοση καθιερωμένων αρχιτεκτονικών όπως τα ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT-B/16, σε όλες τις ποιοτικές μετρικές, συμπεριλαμβανομένων της ακρίβειας των θετικών προβλέψεων, ανάκλησης, βαθμολογία F1 και βαθμολογία AUC.

Ένα από τα πλέον αξιοσημείωτα χαρακτηριστικά του προτεινόμενου μοντέλου είναι η μειωμένη υπολογιστική απαίτηση σε σύγκριση με τις άλλες αρχιτεκτονικές. Η μέση διάρκεια εκπαίδευσης ανά πτυχή είναι μόλις 55 λεπτά, γεγονός που καταδεικνύει ότι η διαδικασία παραμένει εφικτή και σε συστήματα με περιορισμένη υπολογιστική ισχύ. Η παρούσα μελέτη συμβάλλει στη βελτίωση της ιατρικής ανάλυσης εικόνων μέσω της ανάπτυξης ενός αποδοτικού και χαμηλής πολυπλοκότητας δικτύου με περιορισμένο αριθμό παραμέτρων, υπογραμμίζοντας τη δυνατότητά του να αναβαθμίσει τις διαγνωστικές διαδικασίες στο σημείο περίθαλψης. Τα αποτελέσματα της έρευνας ενισχύουν τη σημασία της χρήσης των προηγμένων τεχνικών σε πραγματικές ιατρικές εφαρμογές, προσφέροντας μια εναλλακτική λύση υψηλής ακρίβειας εξαγωγής συμπερασμάτων και χαμηλού υπολογιστικού κόστους για τη διάγνωση των πνευμονικών νοσημάτων.

5.2 Εισαγωγή

Οι επιδημικές πνευμονικές νόσοι, όπως η αδιαφάνεια των πνευμόνων, η ιογενής πνευμονία, η ίνωση, η βακτηριακή πνευμονία, ο COVID-19 και η φυματίωση, έχουν αναδειχθεί ως σημαντικές παγκόσμιες προκλήσεις για τη δημόσια υγεία, προκαλώντας υψηλή νοσηρότητα και θνησιμότητα σε πολλές χώρες [234, 235]. Τα άτομα που έχουν προσβληθεί από αυτές τις ασθένειες παρουσιάζουν ποικίλα συμπτώματα, τα οποία κυμαίνονται από ήπιες έως μέτριες εκδηλώσεις, όπως πυρετός, βήχας και δύσπνοια. Ωστόσο, σε ορισμένες περιπτώσεις, η μόλυνση μπορεί να εξελιχθεί σε σοβαρές πνευμονικές επιπλοκές, οι οποίες ενδέχεται να οδηγήσουν σε θνησιμότητα [236, 237].

Η πανδημία του COVID—19 υπήρξε ιδιαίτερα ανησυχητική, καθώς ένα μεγάλο ποσοστό των κρουσμάτων παρουσίασε έντονη συμφόρηση στον θώρακα και σημαντική μείωση των επιπέδων οξυγόνου, οδηγώντας σε σοβαρές καρδιαγγειακές επιπλοκές [238–240]. Αυτή η εξέλιξη καθιστά απολύτως αναγκαία την έγκαιρη διάγνωση και αντιμετώπιση της νόσου, προκειμένου να περιοριστεί η επιβάρυνση στο αναπνευστικό σύστημα και να αποτραπούν οι περαιτέρω επιπλοκές [241–243].

Αντίστοιχα, η πνευμονία είναι μία σοβαρή πνευμονική λοίμωξη, η οποία χαρακτηρίζεται από φλεγμονή στα μικροσκοπικά κυψελιδικά κυστίδια εντός των πνευμόνων. Η ασθένεια μπορεί να προκληθεί από διάφορους παθογόνους παράγοντες, συμπεριλαμβανομένων ιογενών λοιμώξεων (όπως η γρίπη, ο COVID—19 ή ο αναπνευστικός συγκυτιακός ιός), ιογενείς λοιμώξεις κοινού κρυολογήματος και βακτηριακές λοιμώξεις [244–246]. Η έγκαιρη διάγνωση και η εφαρμογή της κατάλληλης θεραπείας είναι απολύτως κρίσιμες για την αντιμετώπιση αυτών των επιδημικών πνευμονικών νόσων, συμβάλλοντας στη μείωση της θνησιμότητας και της συνολικής επιβάρυνσης στη δημόσια υγεία [247, 248].

Η εξέλιξη της ιατρικής τεχνολογίας, σε συνδυασμό με τις πρόσφατες προόδους στην ML και την ανάλυση των ιατρικών εικόνων, επιτρέπουν την ανάπτυξη αποτελεσματικών μεθόδων για την έγκαιρη διάγνωση και πρόγνωση [249–251]. Τα συστήματα που βασίζονται στην DL και ειδικότερα στα CNNs έχουν επιδείξει εξαιρετική απόδοση στη διάγνωση των αναπνευστικών νόσων, επιτρέποντας την ταχεία και αξιόπιστη ανάλυση των ακτινογραφιών θώρακα. Μέσω της χρήσης αυτών των τεχνικών, είναι εφικτή η ακριβής διαφοροποίηση μεταξύ των διαφορετικών κατηγοριών των πνευμονικών λοιμώξεων, βελτιώνοντας έτσι τη διαδικασία της διάγνωσης και θεραπείας [252–255]. Η δυνατότητα της ταχείας αναγνώρισης των παθήσεων του αναπνευστικού συστήματος επιτρέπει την άμεση παρέμβαση των επαγγελματιών υγείας, οι οποίοι διαθέτουν την εξειδικευμένη γνώση για τη χορήγηση στοχευμένων θεραπειών βάσει της συγκεκριμένης αιτιολογίας της λοίμωξης [256–258].

Ανάλογα με τον παθογόνο παράγοντα που προκαλεί τη νόσο, η θεραπευτική αγωγή [259–261] μπορεί να περιλαμβάνει:

- Αντιβιοτικά, για την καταπολέμηση των βακτηριακών λοιμώξεων, όπως η βακτηριακή πνευμονία.
- Αντιϊικά φάρμακα, για την αντιμετώπιση των ιογενών λοιμώξεων, όπως ο COVID—19 ή η ιογενής πνευμονία.

- Συμπτωματική αγωγή και υποστηρικτικές παρεμβάσεις, για τη διαχείριση της λοίμωξης και την αποφυγή επιπλοκών.

Ο πρωταρχικός στόχος της θεραπείας είναι η εξάλειψη της λοίμωξης, η ανακούφιση των συμπτωμάτων και η πρόληψη των επιπλοκών που μπορεί να προκύψουν από την εξέλιξη της νόσου [262]. Οι επιπλοκές [263–265] αυτές μπορεί να είναι ήπιες έως πολύ σοβαρές, οι οποίες κυμαινόμενες από:

- Αναπνευστική ανεπάρκεια, λόγω της προοδευτικής επιδείνωσης της πνευμονικής λειτουργίας.
- Πνευμονία, η οποία μπορεί να οδηγήσει σε συσσώρευση υγρού στις κυψελίδες και μειωμένη ανταλλαγή αερίων.
- Μακροχρόνιες βλάβες των πνευμόνων, που μπορεί να προκαλέσουν χρόνιες αναπνευστικές παθήσεις ή προδιάθεση για μελλοντικές λοιμώξεις.

Η άμεση διάγνωση των πνευμονικών νόσων και η χορήγηση της κατάλληλης θεραπείας μπορούν να μειώσουν σημαντικά τον κίνδυνο αυτών των επιπλοκών. Παράλληλα, η αποτελεσματική διαχείριση των παθήσεων αυτών συμβάλλει στη δραστηκή μείωση της θνησιμότητας που σχετίζεται με σοβαρές λοιμώξεις του αναπνευστικού συστήματος [266, 267].

Η πρόοδος στις διαγνωστικές τεχνικές, οι καινοτομίες στην DL και στην ανάλυση των ιατρικών εικόνων προσφέρουν νέες δυνατότητες για την ανίχνευση και την παρακολούθηση των πνευμονικών παθήσεων. Η ενσωμάτωση ευφών διαγνωστικών συστημάτων στις κλινικές διαδικασίες επιτρέπει την αξιόπιστη και ταχεία αξιολόγηση της υγείας των πνευμόνων, προάγοντας τη βελτίωση των θεραπευτικών εκβάσεων [268, 269].

Η ακτινογραφία θώρακα αποτελεί μια από τις πλέον ευρέως χρησιμοποιούμενες διαγνωστικές μεθόδους για την αξιολόγηση της κατάστασης των πνευμόνων και την ανίχνευση των παθολογικών ευρημάτων. Πρόκειται για ένα απολύτως απαραίτητο εργαλείο στη σύγχρονη ιατρική, καθώς παρέχει πολύτιμες διαγνωστικές πληροφορίες σχετικά με την ύπαρξη και την έκταση των διαφόρων πνευμονικών παθήσεων και ανωμαλιών, επιτρέποντας στους κλινικούς ιατρούς να λαμβάνουν τεκμηριωμένες αποφάσεις για τη διαχείριση της υγείας των ασθενών [270–272].

Ωστόσο, η ακριβής διαφοροποίηση μεταξύ των διαφόρων επιδημικών πνευμονικών νόσων αποκλειστικά μέσω της οπτικής παρατήρησης των ακτινογραφικών εικόνων μπορεί να αποδειχθεί ιδιαίτερα απαιτητική. Η διάγνωση αυτών των παθήσεων μέσω ακτινογραφιών απαιτεί εξειδικευμένη γνώση και βαθιά κατανόηση των ακτινολογικών χαρακτηριστικών και προτύπων της κάθε νόσου [273–276].

Ένα από τα μεγαλύτερα διαγνωστικά εμπόδια είναι η επικάλυψη των ακτινολογικών χαρακτηριστικών μεταξύ των διαφορετικών πνευμονικών νόσων [277–279]. Συγκεκριμένα:

- Οπτικά χαρακτηριστικά της ιογενούς πνευμονίας μπορεί να μοιάζουν με εκείνα της βακτηριακής πνευμονίας, καθιστώντας δύσκολη τη διαφοροποίηση.
- Η αδιαφάνεια των πνευμόνων που παρατηρείται στον COVID–19 μπορεί να είναι παρόμοια με εκείνη που συναντάται σε περιπτώσεις ινώδους πνευμονίας ή φυματίωσης.
- Η ύπαρξη πολλών και διαφορετικών παραγόντων που επηρεάζουν τη διάγνωση, όπως οι παραλλαγές στη μορφολογία των βλαβών που προκαλούν οι πνευμονικές παθήσεις, οδηγεί σε ακόμη μεγαλύτερη διαγνωστική δυσκολία και αυξάνει τον βαθμό πολυπλοκότητας της διαδικασίας.

Η επίδραση αυτής της ακτινολογικής επικάλυψης μπορεί να οδηγήσει σε διαγνωστικά σφάλματα και χρονικές καθυστερήσεις στην έναρξη της κατάλληλης θεραπείας. Η λανθασμένη ή η καθυστερημένη διάγνωση μπορεί να επιφέρει σοβαρές επιπτώσεις στην υγεία των ασθενών, επιδεινώνοντας την πρόγνωση και αυξάνοντας τον κίνδυνο των επιπλοκών [280, 281]. Για την αντιμετώπιση αυτών των προκλήσεων, η ενσωμάτωση των τεχνικών της DL και των CNNs στην ανάλυση των ακτινογραφιών του θώρακα αποτελεί καινοτόμα προσέγγιση για τη βελτίωση της διαγνωστικής ακρίβειας. Τα συστήματα [282–286] αυτά είναι ικανά να:

- Αναγνωρίζουν τις λεπτές διαφοροποιήσεις στα ακτινολογικά πρότυπα που δεν είναι εύκολα αντιληπτές από τον ανθρώπινο οφθαλμό.
- Μειώνουν τα ποσοστά των ψευδών θετικών και των ψευδών αρνητικών διαγνώσεων.
- Επιταχύνουν τη διαγνωστική διαδικασία, επιτρέποντας την άμεση κλινική παρέμβαση.

Η αξιοποίηση αυτών των τεχνολογιών σε ένα κλινικό περιβάλλον μπορεί να οδηγήσει σε αξιοσημείωτη βελτίωση των διαγνωστικών διαδικασιών, επιτρέποντας την ταχύτερη και την ακριβέστερη διάγνωση των πνευμονικών παθήσεων.

Οι τεχνολογίες της ML [287, 288], της DL [289–291], των CNNs [292] και της AI [293, 294] έχουν επιφέρει σημαντικές εξελίξεις στην ανάλυση των ιατρικών εικόνων, συμπεριλαμβανομένης της διάγνωσης των σύνθετων ιατρικών απεικονίσεων, όπως οι ακτινογραφίες θώρακα. Αυτές οι τεχνολογίες και οι αλγόριθμοι έχουν αποδείξει ιδιαίτερα υψηλές δυνατότητες στην εξαγωγή των κρίσιμων προτύπων και των χαρακτηριστικών από ιατρικές εικόνες [295–297]. Τα πρότυπα αυτά δεν είναι πάντα άμεσα αντιληπτά από τον ανθρώπινο παρατηρητή, γεγονός που καθιστά τις τεχνικές ML ιδιαίτερα χρήσιμες για την ακριβή και την αντικειμενική διάγνωση των ασθενειών [298–301]. Με τη χρήση αυτών των μεθόδων, οι επαγγελματίες υγείας έχουν στη διάθεσή τους ένα ισχυρό διαγνωστικό εργαλείο, το οποίο μπορεί να μειώσει τον κίνδυνο διαγνωστικών σφαλμάτων και να ενισχύσει την αξιοπιστία των ιατρικών αποφάσεων.

Επιπλέον, οι τεχνολογίες αυτές συμβάλλουν στη βελτιστοποίηση της διαγνωστικής διαδικασίας μέσω:

- Επισήμανσης των κρίσιμων περιοχών ενδιαφέροντος, καθοδηγώντας τους κλινικούς ιατρούς προς τις πλέον ύποπτες ζώνες της απεικόνισης.
- Παροχής ποσοτικών μετρήσεων, επιτρέποντας μια πιο αντικειμενική και αριθμητικά τεκμηριωμένη ανάλυση των ιατρικών δεδομένων.
- Πρόβλεψης των πιθανών διαγνώσεων, στηριζόμενη σε πρότυπα που έχουν αντληθεί από μεγάλα σύνολα δεδομένων, λειτουργώντας ως υποστηρικτικό εργαλείο λήψης των κλινικών αποφάσεων.

Η ενσωμάτωση αυτών των τεχνολογιών στην ανάλυση των ιατρικών εικόνων μπορεί να επιταχύνει τη διαδικασία της διάγνωσης και να επιτρέψει στους επαγγελματίες υγείας την ταχύτερη και ακριβέστερη αξιολόγηση των ασθενών. Επιπλέον, η χρήση αυτοματοποιημένων αλγορίθμων μειώνει τον κίνδυνο των υποκειμενικών αποκλίσεων, οι οποίες είναι πιθανές σε περιπτώσεις όπου η διάγνωση βασίζεται αποκλειστικά στην ανθρώπινη κρίση [302–304]. Είναι σημαντικό να τονιστεί ότι αυτές οι τεχνολογίες δεν έχουν σκοπό να αντικαταστήσουν τους επαγγελματίες υγείας, αλλά να ενισχύσουν την εμπειρογνωμοσύνη τους. Ο συνδυασμός της

ανθρώπινης κλινικής εμπειρίας με τους προηγμένους αλγορίθμους ML μπορεί να οδηγήσει σε ακριβέστερες διαγνώσεις, βελτιωμένη περίθαλψη των ασθενών και μεγαλύτερη αποτελεσματικότητα στη διαχείριση των επιδημικών πνευμονικών νοσημάτων [305–308].

Στη σύγχρονη ιατρική πρακτική, υπάρχει ιδιαίτερη έμφαση στην ιατρική στο σημείο περίθαλψης, η οποία περιλαμβάνει την παροχή ιατρικής φροντίδας και τη διενέργεια διαγνωστικών εξετάσεων απευθείας στο σημείο όπου βρίσκεται ο ασθενής ή κοντά σε αυτόν [309, 310]. Η χρήση φορητών ακτινογραφικών μηχανημάτων ή συσκευών χειρός επιτρέπει την άμεση απόκτηση ιατρικών εικόνων, εξαλείφοντας την ανάγκη μεταφοράς των ασθενών σε κεντρικές ακτινολογικές μονάδες [311–313]. Αυτή η προσέγγιση:

- Εξοικονομεί πολύτιμο χρόνο στη διαγνωστική διαδικασία.
- Μειώνει τη φυσική καταπόνηση των ασθενών, ειδικά για όσους είναι κρίσιμα πάσχοντες ή έχουν περιορισμένη κινητικότητα.
- Επιτρέπει την ταχύτερη έναρξη των θεραπευτικών παρεμβάσεων, καθώς μειώνει τη χρονική καθυστέρηση μεταξύ της λήψης και της ερμηνείας των ιατρικών εικόνων.

Η προσέγγιση της ιατρικής στο σημείο περίθαλψης στοχεύει στη βελτίωση της αμεσότητας των ιατρικών υπηρεσιών και στη διευκόλυνση της πρόσβασης στις απεικονιστικές διαγνωστικές μεθόδους [314–316]. Αυτό είναι ιδιαίτερα σημαντικό σε καταστάσεις έκτακτης ανάγκης, όπου η άμεση διάγνωση είναι ζωτικής σημασίας για την ταχεία λήψη κλινικών αποφάσεων.

Επιπλέον, η υιοθέτηση αυτής της προσέγγισης έχει τη δυνατότητα να βελτιώσει τη συνολική απόδοση των συστημάτων της υγειονομικής περίθαλψης, συμβάλλοντας στην αποδοτικότερη κατανομή των ιατρικών πόρων και στη μείωση του χρόνου αναμονής των ασθενών. Ιδιαίτερα ωφέλιμη είναι η εφαρμογή της σε περιοχές με περιορισμένους ιατρικούς πόρους, σε απομακρυσμένες ή υποεξυπηρετούμενες περιοχές [317–319]. Σε τέτοια περιβάλλοντα, η έγκαιρη πρόσβαση σε διαγνωστικές υπηρεσίες μέσω φορητών ιατρικών τεχνολογιών μπορεί να βελτιώσει σημαντικά τα θεραπευτικά αποτελέσματα και να μειώσει τη συνολική θνησιμότητα [320, 321].

Η ανάπτυξη και η ενσωμάτωση των έξυπνων διαγνωστικών συστημάτων βασισμένων στην ΑΙ σε συνδυασμό με τις φορητές απεικονιστικές τεχνολογίες μπορεί να

ενισχύσει ακόμη περισσότερο την συνολική ακρίβεια, την ταχύτητα και την προσβασιμότητα της ιατρικής στο σημείο περίθαλψης. Καθώς τα CNNs συνεχίζουν να εξελίσσονται, η ικανότητά τους να παρέχουν αυτοματοποιημένη διάγνωση υψηλής ακρίβειας εξαγωγής συμπερασμάτων θα μπορούσε να καταστήσει το σημείο περίθαλψης ακόμη πιο αποτελεσματικό και προσιτό παγκοσμίως [322–324].

Η παρούσα έρευνα συνιστά τρεις κύριες συνεισφορές στον τομέα της ιατρικής απεικόνισης και της διάγνωσης των ασθενειών, οι οποίες μπορούν να συνοψιστούν ως εξής:

- Δημιουργία ενός ενιαίου συνόλου δεδομένων υψηλής ποικιλομορφίας: Συγχωνεύσαμε πολλαπλά υπάρχοντα σύνολα δεδομένων, δημιουργώντας μια πιο αντιπροσωπευτική και ολοκληρωμένη συλλογή ακτινογραφιών θώρακα. Το αποτέλεσμα είναι η διαμόρφωση ενός προβλήματος ταξινόμησης επτά κατηγοριών, γεγονός που ενισχύει τη γενικευσιμότητα και την ανθεκτικότητα της προτεινόμενης προσέγγισης. Αναπτύξαμε ένα χαμηλής πολυπλοκότητας CNN, το οποίο έχει σχεδιαστεί βάσει της μεθοδολογίας που παρουσιάστηκε στο Κεφάλαιο 3, και ειδικότερα σύμφωνα με τις μετρικές σχεδιασμού της Ενότητας 3.2.1.1, κατάλληλο για υλοποίηση σε ετερογενή ενσωματωμένα συστήματα. Η συμπαγής αρχιτεκτονική του μοντέλου το καθιστά ιδανικό για χρήση σε συστήματα περιορισμένων πόρων, όπως φορητές διαγνωστικές συσκευές και εφαρμογές στο σημείο περίθαλψης, χωρίς να θυσιάζεται η διαγνωστική ακρίβεια της εξαγωγής συμπερασμάτων.
- Αντιμετώπιση της ανισορροπίας των κλάσεων: Αντιμετωπίζουμε το ευρέως διαδεδομένο πρόβλημα της ανισορροπίας των κλάσεων στις ιατρικές απεικονίσεις (βλ. Ενότητα 3.2.2.3), υιοθετώντας τη συνάρτηση εστιακής απώλειας κατά τη φάση εκπαίδευσης του μοντέλου. Η στρατηγική αυτή δίνει έμφαση στα δύσκολα δείγματα και μειώνει την προκατάληψη των υπερεκπροσωπούμενων κλάσεων. Το αποτέλεσμα είναι η βελτίωση της ικανότητας του μοντέλου στην ακριβή εξαγωγή συμπερασμάτων σπάνιων ή υποεκπροσωπούμενων παθήσεων.
- Συγκριτική αξιολόγηση με σύγχρονες αρχιτεκτονικές αιχμής: Συγκρίναμε την απόδοση του προτεινόμενου μοντέλου με γνωστές αρχιτεκτονικές μεταφοράς μάθησης, όπως ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT–B/16. Για την αξιολόγηση, εφαρμόσαμε την τεχνική της πενταπλής διασταυρούμενης επικύρωσης (βλ. Ενότητα 3.4.1), προκειμένου να

διασφαλίσουμε τη σταθερότητα και την αξιοπιστία των αποτελεσμάτων. Τα πειραματικά αποτελέσματα ανέδειξαν ότι το προτεινόμενο μοντέλο υπερτερεί σημαντικά στις ποιοτικές μετρικές, υπογραμμίζοντας τη χρηστικότητα και την αξιοπιστία του ως υποστηρικτικό εργαλείο για τους επαγγελματίες υγείας.

Το υπόλοιπο του κεφαλαίου οργανώνεται ως εξής: Η ενότητα 5.3 παρουσιάζει την σχετική βιβλιογραφία. Η ενότητα 5.4 αναφέρεται στα υλικά και τα δεδομένα που χρησιμοποιήθηκαν. Η ενότητα 5.5 περιγράφει τη μεθοδολογία σχεδιασμού και υλοποίησης του συστήματος. Στη συνέχεια, στην ενότητα 5.6, παρουσιάζουμε τα πειραματικά αποτελέσματα και τις αξιολογήσεις που υποδεικνύουν την αποτελεσματικότητα της προτεινόμενης μεθοδολογίας. Στην ενότητα 5.7 αναλύονται και σχολιάζονται τα ευρήματα της έρευνας. Τέλος, η ενότητα 5.8 περιλαμβάνει τα συμπεράσματα και προτείνει κατευθύνσεις για μελλοντική έρευνα.

5.3 Επισκόπηση Επιστημονικού Πεδίου

Τα τελευταία χρόνια, η πρόγνωση και η διάγνωση των πνευμονικών νόσων έχουν αποτελέσει αντικείμενο εκτεταμένης ερευνητικής δραστηριότητας και τεχνολογικής εξέλιξης από ιατρούς, ερευνητές και επιστήμονες παγκοσμίως [325–330]. Η ραγδαία πρόοδος στον τομέα της ιατρικής απεικόνισης, σε συνδυασμό με τις εξελίξεις στην ML, την AI και τα CNNs, έχει οδηγήσει σε σημαντική βελτίωση της συνολικής ακρίβειας εξαγωγής συμπερασμάτων, της ταχύτητας και της αποδοτικότητας στην κατηγοριοποίηση των πνευμονικών νοσημάτων [331–338].

Οι σύγχρονες τεχνολογίες επιτρέπουν:

- Πρώιμη ανίχνευση των πνευμονικών παθήσεων, προτού αυτές εξελιχθούν σε πιο σοβαρά στάδια.
- Βελτιωμένη συνολική ακρίβεια διάγνωσης, μειώνοντας τον κίνδυνο των διαγνωστικών λαθών μέσω των αυτοματοποιημένων αλγορίθμων.
- Ταχύτερη εξαγωγή συμπερασμάτων, υποστηρίζοντας τη λήψη έγκαιρων και τεκμηριωμένων ιατρικών αποφάσεων.

Παρά την πρόοδο που έχει σημειωθεί, είναι σημαντικό να επισημανθεί ότι η έρευνα στον τομέα αυτόν βρίσκεται σε διαρκή εξέλιξη [339–341]. Οι προσπάθειες για την

ανάπτυξη ακόμη πιο εξελιγμένων και πρακτικά εφαρμόσιμων διαγνωστικών λύσεων συνεχίζονται [342, 343]. Οι ερευνητικές κοινότητες επικεντρώνονται στη:

- Διεύρυνση και ενοποίηση δεδομένων για τη βελτίωση της γενίκευσης των μοντέλων.
- Ανάπτυξη εξειδικευμένων αλγορίθμων χαμηλού κόστους για εφαρμογές σε περιβάλλοντα με περιορισμένους πόρους.
- Ενσωμάτωση συστημάτων ΑΙ σε διαγνωστικές ροές εργασίας, διασφαλίζοντας τη διαλειτουργικότητα και την αποδοτικότητα.

Συνεπώς, η συνδυασμένη χρήση τεχνολογιών ΑΙ και ιατρικής απεικόνισης αποτελεί ένα δυναμικά εξελισσόμενο πεδίο, με προοπτική την αναβάθμιση της διαγνωστικής ακρίβειας εξαγωγής συμπερασμάτων και την ποιοτική βελτίωση των παρεχόμενων υπηρεσιών υγείας.

Στη μελέτη των [344], οι συγγραφείς πρότειναν τη χρήση ενός ανταγωνιστικού δικτύου υπό όρους (cGAN), σε συνδυασμό με ένα βελτιστοποιημένο προεκπαιδευμένο CNN τύπου ResNet50. Η εν λόγω αρχιτεκτονική σχεδιάστηκε για την αναγνώριση έξι διακριτών κατηγοριών πνευμονικών νόσων, μέσω ακτινογραφιών θώρακα. Το σύνολο δεδομένων που χρησιμοποιήθηκε στις πειραματικές διαδικασίες περιλάμβανε 1.229 εικόνες πνευμονικών νοσημάτων. Το προτεινόμενο μοντέλο εκπαιδεύτηκε για 30 εποχές χρησιμοποιώντας τη συνάρτηση απώλειας κατηγορικής διασταυρούμενης εντροπίας. Ως αποτέλεσμα, το μοντέλο πέτυχε ακρίβεια εξαγωγής συμπερασμάτων 93,67%.

Επιπλέον, στη μελέτη των [345], προτάθηκε ένα εξατομικευμένο CNN με την ονομασία DWTMBConvNet. Η αρχιτεκτονική του μοντέλου στόχευε στην ανίχνευση πέντε κατηγοριών που σχετίζονται με αναπνευστικές παθήσεις. Το σύνολο δεδομένων περιλάμβανε 11.819 εικόνες ακτινογραφιών θώρακα. Το προτεινόμενο μοντέλο εκπαιδεύτηκε για 500 εποχές με τη χρήση της συνάρτησης απώλειας κατηγορικής διασταυρούμενης εντροπίας. Η τελική συνολική ακρίβεια του DWTMBConvNet που επιτεύχθηκε ήταν 95,50%.

Ταυτόχρονα, στο [346], παρουσιάστηκε το μοντέλο FC–DenseNet103 για την ταξινόμηση πέντε διακριτών κατηγοριών που αφορούν παθήσεις του αναπνευστικού συστήματος. Το σύνολο δεδομένων που χρησιμοποιήθηκε στις πειραματικές διαδικασίες περιλάμβανε 502 εικόνες ακτινογραφιών θώρακα.

Το μοντέλο εκπαιδεύτηκε για 100 εποχές χρησιμοποιώντας τη συνάρτηση απώλειας κατηγορικής διασταυρούμενης εντροπίας. Ως αποτέλεσμα, το μοντέλο FC–DenseNet103 πέτυχε συνολική ακρίβεια 88,90%. Επιπλέον, οι συγγραφείς πρότειναν τη χρήση του μοντέλου ResNet18 για την αναγνώριση τεσσάρων κατηγοριών πνευμονικών λοιμώξεων. Το αντίστοιχο σύνολο δεδομένων περιλάμβανε 482 εικόνες. Το μοντέλο ResNet18 εκπαιδεύτηκε για 100 εποχές με χρήση της ίδιας συνάρτησης απώλειας. Η ακρίβεια ταξινόμησης που επιτεύχθηκε από το μοντέλο ResNet18 ανήλθε σε 91,90%.

Οι συγγραφείς στη μελέτη [347], παρουσιάζουν ένα προσαρμοσμένο CNN με την ονομασία CoroDet, το οποίο σχεδιάστηκε ειδικά για την ταξινόμηση τεσσάρων διακριτών κατηγοριών πνευμονολογικών διαταραχών. Το σύνολο δεδομένων που χρησιμοποιήθηκε στις πειραματικές διαδικασίες περιλάμβανε 2.100 εικόνες προερχόμενες από ακτινογραφίες θώρακα. Το μοντέλο CoroDet εκπαιδεύτηκε για 50 εποχές με στόχο τη βελτιστοποίηση της ακρίβειας ταξινόμησης. Ως αποτέλεσμα, επιτεύχθηκε συνολική ακρίβεια εξαγωγής συμπερασμάτων 91,20%.

Αντίστοιχα, στη μελέτη των [348], παρουσιάστηκε το CoroNet, ένα βαθύ CNN βασισμένο στην αρχιτεκτονική του μοντέλου Xception. Το CoroNet σχεδιάστηκε για την ταξινόμηση τεσσάρων κατηγοριών με παθολογίες των πνευμόνων. Η αρχιτεκτονική του μοντέλου περιλάμβανε συνολικά 33 εκατομμύρια παραμέτρους. Για τις ανάγκες της πειραματικής αξιολόγησης, οι συγγραφείς χρησιμοποίησαν ένα σύνολο δεδομένων που περιλάμβανε 1.251 εικόνες. Το CoroNet εκπαιδεύτηκε για 80 εποχές, επιτυγχάνοντας τελικά συνολική ακρίβεια 89,60%.

Στη μελέτη των [349], παρουσιάστηκε ένα προσαρμοσμένο CNN σχεδιασμένο για την ταξινόμηση τεσσάρων κατηγοριών που αφορούν ασθένειες του αναπνευστικού. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιλάμβανε 6.414 εικόνες, οι οποίες ελήφθησαν από ακτινογραφίες θώρακα. Το προτεινόμενο CNN εκπαιδεύτηκε για 70 εποχές. Η πειραματική αξιολόγηση ανέδειξε διαγνωστική επίδοση, με συνολική ακρίβεια 97,94%. Παράλληλα, στη μελέτη των [350], προτάθηκε ένα τροποποιημένο CNN βασισμένο στην αρχιτεκτονική VGG19, με σκοπό την ταξινόμηση τεσσάρων κατηγοριών με πνευμονολογικές διαταραχές. Το σύνολο δεδομένων περιλάμβανε 33.676 εικόνες και το μοντέλο εκπαιδεύτηκε για 500 εποχές με την χρήση της συνάρτησης απώλειας κατηγορικής διασταυρούμενης εντροπίας. Η τελική ακρίβεια που επιτεύχθηκε στο τροποποιημένο VGG19 ήταν 98,05%.

Οι συγγραφείς στη μελέτη [351], πρότειναν ένα CNN βασισμένο στο προεκπαιδευμένο μοντέλο AlexNet. Το μοντέλο σχεδιάστηκε για την ταξινόμηση τεσσάρων κατηγοριών που αφορά πνευμονικά νοσήματα. Για τις ανάγκες της πειραματικής αξιολόγησης, οι συγγραφείς χρησιμοποίησαν ένα σύνολο δεδομένων που περιλάμβανε 11.568 εικόνες ακτινογραφιών θώρακα. Η εκπαίδευση του μοντέλου πραγματοποιήθηκε για 20 εποχές, με τη χρήση της συνάρτησης απώλειας κατηγορικής διασταυρούμενης εντροπίας. Το προτεινόμενο CNN πέτυχε ακρίβεια εξαγωγής συμπερασμάτων 93,42%.

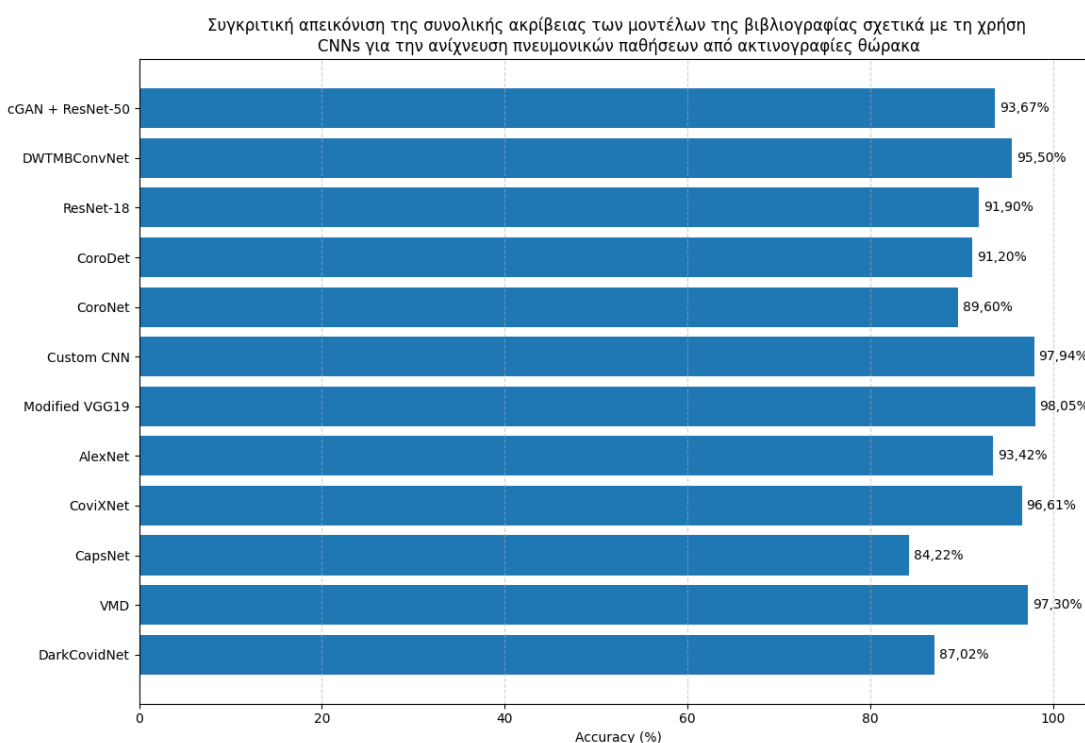
Παράλληλα, στη μελέτη των [352] παρουσιάστηκε το CoviXNet, ένα μοντέλο CNN χαμηλής πολυπλοκότητας, σχεδιασμένο για την ταξινόμηση τριών κατηγοριών παθήσεων του αναπνευστικού συστήματος. Το σύνολο δεδομένων που χρησιμοποιήθηκε περιλάμβανε 6.207 εικόνες ακτινογραφιών θώρακα. Το CoviXNet πέτυχε συνολική ακρίβεια 96,61%. Επιπροσθέτως, στη μελέτη των [353], προτάθηκε ένα CNN τύπου CapsNet για την ταξινόμηση τριών κατηγοριών που αφορούν νοσήματα του αναπνευστικού συστήματος. Το αρχικό σύνολο δεδομένων περιλάμβανε 2.331 εικόνες, οι οποίες αυξήθηκαν με την εφαρμογή των τεχνικών της επαύξησης δεδομένων για να επεκτείνει τεχνητά το σύνολο δεδομένων σε 3.150 εικόνες. Το μοντέλο εκπαιδεύτηκε για 50 εποχές, επιτυγχάνοντας συνολική ακρίβεια εξαγωγής συμπερασμάτων 84,22%.

Στη μελέτη των [354], αξιολογήθηκαν τέσσερα προεκπαιδευμένα CNN, συγκεκριμένα τα VGG19, InceptionV3, MobileNetV2, και DenseNet. Επιπροσθέτως, οι ερευνητές ανέπτυξαν τέσσερα υβριδικά μοντέλα, συνδυάζοντας διαφορετικές αρχιτεκτονικές CNNs: VID (VGG19, Inception, DenseNet), VMI (VGG19, MobileNet, Inception), VMD (VGG19, MobileNet, DenseNet), και IMD (Inception, MobileNet, DenseNet). Για την πειραματική αξιολόγηση των μοντέλων, χρησιμοποιήθηκε ένα σύνολο δεδομένων που περιλάμβανε 4.500 εικόνες για την ταξινόμηση τριών κατηγοριών πνευμονικών νοσημάτων. Όλα τα μοντέλα εκπαιδεύτηκαν με τις ίδιες υπερπαραμέτρους για 10 εποχές, με την χρήση της συνάρτησης απώλειας κατηγορικής διασταυρούμενης εντροπίας. Τα αποτελέσματα των πειραμάτων ανέδειξαν ότι το υβριδικό μοντέλο VMD παρουσίασε την καλύτερη απόδοση μεταξύ όλων των μοντέλων που εξετάστηκαν, επιτυγχάνοντας συνολική ακρίβεια 97,30%.

Αντιστοίχως, στη μελέτη των [355], παρουσιάστηκε το DarkCovidNet, ένα προσαρμοσμένο CNN για την ταξινόμηση τριών κατηγοριών που αφορούν ασθένειες του αναπνευστικού. Το σύνολο δεδομένων που χρησιμοποιήθηκε

περιλάμβανε 1.127 εικόνες ακτινογραφιών θώρακα. Το μοντέλο εκπαιδεύτηκε για 100 εποχές και πέτυχε συνολική ακρίβεια εξαγωγής συμπερασμάτων 87,02%.

Στον Πίνακα 5.1 συνοψίζονται οι διάφορες προσεγγίσεις των CNNs για την αναγνώριση των πνευμονικών παθήσεων μέσω ακτινογραφιών θώρακα, ενώ στο Σχήμα 5.1 παρουσιάζεται η αναλυτική συγκριτική απεικόνιση της συνολικής ακρίβειας εξαγωγής συμπερασμάτων που επιτυγχάνεται από τις διαφορετικές αρχιτεκτονικές. Παρατηρείται ότι, τα μοντέλα Modified VGG19, Custom CNN και το υβριδικό μοντέλο VMD καταγράφουν τις υψηλότερες τιμές ακρίβειας εξαγωγής συμπερασμάτων, με ποσοστά 98,05%, 97,94% και 97,30% αντίστοιχα. Αντίθετα, τα μοντέλα CapsNet και DarkCovidNet παρουσιάζουν σημαντικά χαμηλότερες επιδόσεις, με ακρίβειες 84,22% και 87,02% αντίστοιχα. Η διαφοροποίηση στις επιδόσεις των μοντέλων καταδεικνύει την επίδραση τόσο της επιλογής αρχιτεκτονικής όσο και της ποιότητας και της ποικιλομορφίας των δεδομένων εκπαίδευσης στη συνολική διαγνωστική ικανότητα των αρχιτεκτονικών στον τομέα της ιατρικής απεικόνισης.



Σχήμα 5.1: Συγκριτική απεικόνιση της συνολικής ακρίβειας που επιτεύχθηκε από διάφορα αρχιτεκτονικά σχήματα CNNs, όπως έχουν παρουσιαστεί στη διεθνή βιβλιογραφία, για την αναγνώριση των πνευμονικών παθήσεων.

Οι προηγούμενες μελέτες αξιοποίησαν τεχνικές DL για την αναγνώριση των πνευμονικών νοσημάτων, χρησιμοποιώντας ακτινογραφίες θώρακα, οι οποίες

Πίνακας 5.1: Σύνοψη της σύγχρονης βιβλιογραφίας με τη χρήση CNNs για την ανίχνευση πνευμονικών παθήσεων μέσω ακτινογραφιών θώρακα.

Paper	Classes	Best model	Image Count	Accuracy (%)
[344]	6	cGAN with a ResNet50	40 COVID—Mild, 42 COVID—Medium, 36 COVID—Severe, 348 normal, 500 pneumonia, and 263 tuberculosis	93,67
[345]	5	DWTMBConvNet	3593 normal, 1837 tuberculosis, 2098 Covid—19, 2786 bacterial pneumonia, and 1505 viral pneumonia	95,50
[346]	4	ResNet—18	180 COVID—19, 191 normal, 54 bacterial pneumonia, and 20,000 lung cancer	91,90
[347]	4	CoroDet	500 COVID—19, 800 normal, 400 bacterial pneumonia, and 400 viral pneumonia	91,20
[348]	4	CoroNet	284 COVID—19, 310 normal, 330 bacterial pneumonia, and 327 viral pneumonia	89,60
[349]	4	Custom CNN	558 COVID—19, 1583 normal, 2780 bacterial pneumonia, and 1493 viral pneumonia	97,94
[350]	4	modified VGG19	4320 COVID—19, 3500 normal, 5856 bacterial pneumonia, and 400 viral pneumonia	98,05
[351]	4	pre—trained AlexNet	371 COVID—19, 2882 normal, 4078 bacterial pneumonia, and 4237 viral pneumonia	93,42
[352]	3	CoviXNet	1281 COVID—19, 3270 normal, and 1656 pneumonia	96,61
[353]	3	CapsNet	1050 COVID—19, 1050 no—findings, and 1050 pneumonia	84,22
[354]	3	VMD	1500 COVID—19, 1500 normal, and 1500 pneumonia	97,30
[355]	3	DarkCovidNet	127 COVID—19, 500 no—findings, and 500 pneumonia	87,02

ωστόσο δεν διαχειρίζονται αποτελεσματικά τις περιορισμένες δυνατότητες της κατηγορικής διασταυρούμενης εντροπίας ως συνάρτηση απώλειας. Αξιοσημείωτο είναι το γεγονός ότι πολλές από τις υπάρχουσες μελέτες στον τομέα της αναγνώρισης των πνευμονικών παθήσεων περιορίζονται από το μικρό εύρος συνόλου δεδομένων που χρησιμοποιούν, επικεντρώνοντας τη διάγνωση σε τρεις ή τέσσερις κατηγορίες παθήσεων. Η εξάρτηση από τέτοιου τύπου σύνολα δεδομένων γεννά ερωτήματα ως προς τη δυνατότητα γενίκευσης των εν λόγω μοντέλων σε πιο σύνθετα, ετερογενή και πολυδιάστατα κλινικά περιβάλλοντα. Συγκεκριμένα, η ακρίβεια εξαγωγής συμπερασμάτων και η αξιοπιστία των μοντέλων αυτών παραμένουν αμφίβολες όταν εφαρμόζονται σε μεγαλύτερα και πολυπλοκότερα σύνολα δεδομένων, καθώς η επίδοσή τους ενδέχεται να μειωθεί αισθητά εκτός του πεδίου εκπαίδευσής τους.

Ένα πρόσθετο και κρίσιμο πρόβλημα κατά την εκπαίδευση CNNs σε μικρού μεγέθους σύνολα δεδομένων αποτελεί το φαινόμενο της υπερπροσαρμογής, φαινόμενο που αποτελεί ένδειξη περιορισμένης δυνατότητας μεταφοράς γνώσης. Στην περίπτωση αυτή, το μοντέλο μαθαίνει υπερβολικά τις ιδιαιτερότητες των δεδομένων εκπαίδευσης και αδυνατεί να αποδώσει ικανοποιητικά σε άγνωστα ή νέα δεδομένα, γεγονός που επηρεάζει αρνητικά τη γενικευσιμότητά του.

Λαμβάνοντας υπόψη όλους τους παραπάνω περιορισμούς, η παρούσα μελέτη αποσκοπεί στην υπέρβαση των εν λόγω προκλήσεων μέσω της ανάπτυξης μιας αρχιτεκτονικής CNN χαμηλής πολυπλοκότητας, κατάλληλης για ετερογενή ενσωματωμένα συστήματα. Η εν λόγω προσέγγιση καθιστά το προτεινόμενο μοντέλο ιδανικό για χρήση σε ιατρικά περιβάλλοντα περιορισμένων πόρων, και ιδιαίτερα στο σημείο περίθαλψης, όπου απαιτείται ταχύτητα και ακρίβεια στη διάγνωση.

Για την ενίσχυση της ποικιλομορφίας και της αντιπροσωπευτικότητας του συνόλου δεδομένων εκπαίδευσης, πραγματοποιήθηκε συγχώνευση πολλαπλών συνόλων δεδομένων από ακτινογραφίες θώρακα, διαμορφώνοντας ένα ενιαίο και εμπλουτισμένο σύνολο που καλύπτει επτά κατηγορίες πνευμονικών παθήσεων. Η προσέγγιση αυτή ενισχύει τη γενικευσιμότητα του μοντέλου, καθιστώντας το ικανό να ανταποκριθεί σε ένα ευρύτερο φάσμα κλινικών περιπτώσεων.

Επιπλέον, δόθηκε ιδιαίτερη έμφαση στην αντιμετώπιση του φαινομένου της ανισορροπίας των κλάσεων, το οποίο αποτελεί σύνηθες πρόβλημα στην ανάλυση ιατρικών εικόνων, καθώς ορισμένες κατηγορίες παθήσεων εμφανίζονται με

μικρότερη συχνότητα. Για την αντιμετώπιση αυτής της πρόκλησης, υιοθετήθηκε η συνάρτηση της εστιακής απώλειας κατά τη διάρκεια της εκπαίδευσης, με στόχο τον μετριασμό της προκατάληψης προς τις υπερεκπροσωπούμενες κλάσεις και τη βελτίωση της ακρίβειας στην εξαγωγή συμπερασμάτων για τις σπάνιες παθήσεις.

Η μεθοδολογία αξιολόγησης βασίστηκε στη χρήση της πενταπλής διασταυρούμενης επικύρωσης, ενισχύοντας τη στατιστική αξιοπιστία των αποτελεσμάτων και τη γενικευσιμότητα του μοντέλου σε διαφορετικά υποσύνολα του συνόλου δεδομένων, σύμφωνα με την μεθοδολογία (βλ. Ενότητα 3.2.2.3). Τέλος, η υλοποίηση του προτεινόμενου συστήματος πραγματοποιήθηκε λαμβάνοντας υπόψη τους υπολογιστικούς περιορισμούς της ενσωματωμένης πλατφόρμας NVIDIA Jetson Xavier, εξασφαλίζοντας την αποδοτικότητα και τη δυνατότητα της άμεσης ανάπτυξης του συστήματος σε πραγματικά ιατρικά περιβάλλοντα.

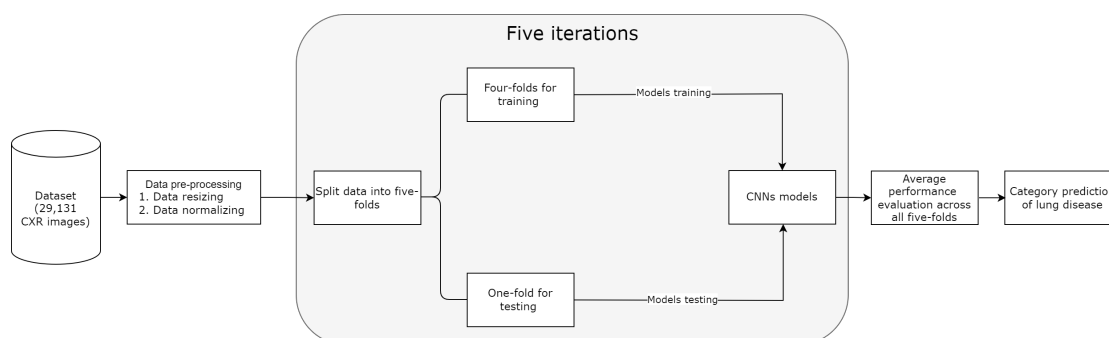
5.4 Υλικά και Δεδομένα

Στο Σχήμα 5.2 παρουσιάζεται η ολοκληρωμένη διαδικασία για την ταξινόμηση των πνευμονικών παθήσεων, καθώς και της κατηγορίας υγιούς ατόμου, μέσω ακτινογραφιών θώρακα. Αρχικά, συλλέγεται ένα ποικιλόμορφο σύνολο ακτινογραφιών από ελεύθερα προσβάσιμες δημόσιες πηγές. Οι εικόνες αυτές συγχωνεύονται ώστε να σχηματίσουν ένα ενιαίο και εκτεταμένο σύνολο δεδομένων, το οποίο καλύπτει επτά διακριτές κατηγορίες που αντιστοιχούν σε συγκεκριμένες πνευμονικές παθήσεις.

Για τη διασφάλιση της αξιοπιστίας του σχεδιασμένου CNN ταξινόμησης, σύμφωνα με την μεθοδολογία (βλ. Ενότητα 3.2.2.3), εφαρμόζεται η διαδικασία της πενταπλής διασταυρούμενης επικύρωσης. Το σύνολο δεδομένων διαχωρίζεται σε πέντε ίσα υποσύνολα, εκ των οποίων τα τέσσερα χρησιμοποιούνται για την εκπαίδευση και το ένα για τον έλεγχο του μοντέλου. Η διαδικασία επαναλαμβάνεται πέντε φορές, έτσι ώστε κάθε υποσύνολο να χρησιμοποιείται μία φορά ως σύνολο ελέγχου. Στη συνέχεια, δημιουργούνται τα προεκπαιδευμένα μοντέλα ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT-B/16, καθώς και η προτεινόμενη αρχιτεκτονική χαμηλής πολυπλοκότητας CNN.

Ακολουθεί η φάση της αξιολόγησης των μοντέλων, κατά την οποία υπολογίζονται η ακρίβεια εξαγωγής συμπερασμάτων και η διαγνωστική αποτελεσματικότητα κάθε μοντέλου, σύμφωνα με την μεθοδολογία που παρουσιάστηκε στο Κεφάλαιο 3, και

ειδικότερα βάσει των μετρικών ποιότητας (βλ. Ενότητα 3.3). Τα αποτελέσματα υπολογίζονται ως ο μέσος όρος των επιδόσεων σε όλα τα υποσύνολα, λαμβάνοντας υπόψη πολλαπλές μετρικές αποδοτικότητας. Η εφαρμογή της πενταπλής διασταυρούμενης επικύρωσης διασφαλίζει ότι η αξιολόγηση των μοντέλων λαμβάνει υπόψη τη μεταβλητότητα του συνόλου δεδομένων, παρέχοντας μια πιο αξιόπιστη και γενικεύσιμη εκτίμηση της ικανότητάς τους στην ταξινόμηση των πνευμονικών παθήσεων.



Σχήμα 5.2: Δομή του συστήματος για την αναγνώριση των πνευμονικών παθήσεων με πενταπλή διασταυρούμενη επικύρωση από το ενοποιημένο σύνολο δεδομένων ακτινογραφιών θώρακα.

5.4.1 Σύνολο Δεδομένων Εικόνων

Η παρούσα μελέτη επικεντρώθηκε στην πολυκατηγορική ταξινόμηση των πνευμονικών παθήσεων, αξιοποιώντας συνδυαστικά τέσσερα διαφορετικά ιατρικά σύνολα δεδομένων με ακτινογραφίες θώρακα. Τα δεδομένα αντλήθηκαν από ενοποιημένες ανοικτές βάσεις εικόνων της διεθνούς βιβλιογραφίας [356]. Συνολικά συγκεντρώθηκαν 29.131 εικόνες ακτινογραφιών, οι οποίες περιλαμβάνουν τόσο φυσιολογικά περιστατικά όσο και περιπτώσεις έξι διαφορετικών πνευμονικών νοσημάτων.

Το πρώτο ιατρικό σύνολο δεδομένων που χρησιμοποιήθηκε προέρχεται από τις μελέτες [357, 358]. Το συγκεκριμένο σύνολο περιλαμβάνει 3.616 περιπτώσεις θετικών σε COVID—19 ακτινογραφιών, 6.012 εικόνες με ενδείξεις πνευμονικής θόλωσης, 10.192 φυσιολογικές εικόνες και 1.345 περιπτώσεις ιογενούς πνευμονίας. Στην παρούσα μελέτη, αξιοποιήθηκαν όλες οι κατηγορίες του εν λόγω συνόλου. Το δεύτερο ιατρικό σύνολο δεδομένων προέρχεται από τη μελέτη [359] και περιλαμβάνει 2.780 εικόνες βακτηριακής πνευμονίας και 1.493 εικόνες ιογενούς πνευμονίας. Ωστόσο, στην παρούσα ανάλυση χρησιμοποιήθηκαν μόνο οι εικόνες

που αντιστοιχούν στη βακτηριακή πνευμονία. Το τρίτο ιατρικό σύνολο δεδομένων προέρχεται από τη μελέτη [360] και περιλαμβάνει 3.500 εικόνες φυματίωσης και ισάριθμες φυσιολογικές ακτινογραφίες θώρακα. Για τη συγκεκριμένη μελέτη, χρησιμοποιήθηκαν μόνο οι εικόνες που απεικονίζουν περιπτώσεις φυματίωσης. Το τέταρτο ιατρικό σύνολο δεδομένων προέρχεται από τη μελέτη [361] και από αυτό το σύνολο αξιοποιήθηκαν μόνο οι 1.686 εικόνες που αφορούν περιπτώσεις της πνευμονικής ίνωσης.

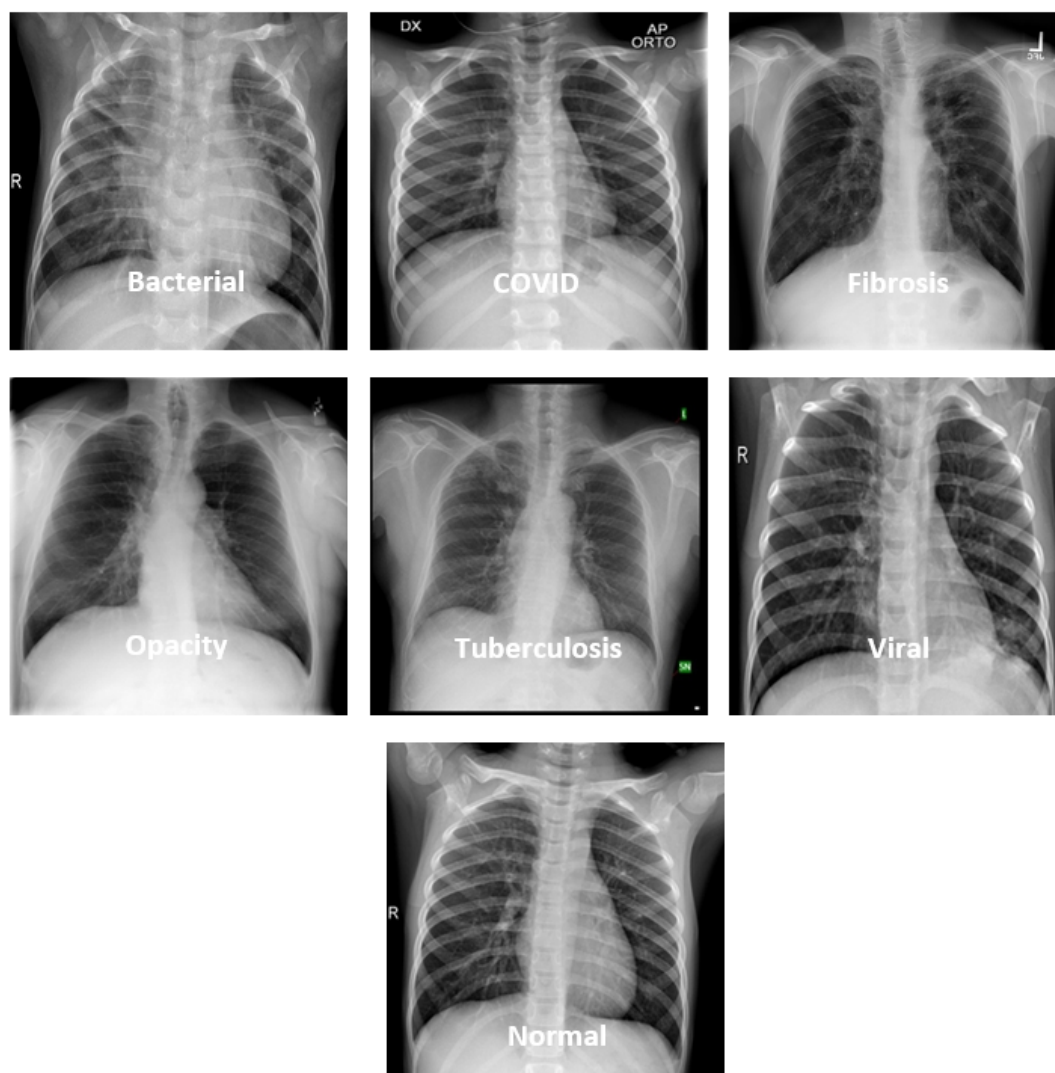
Στο Σχήμα 5.3 παρουσιάζονται ενδεικτικές εικόνες ακτινογραφιών θώρακα για κάθε μία από τις κλάσεις του ενοποιημένου συνόλου δεδομένων. Ο Πίνακας 5.2 συνοψίζει την κατανομή των περιπτώσεων ανά κλάση, προσδιορίζοντας το πλήθος των ακτινογραφιών που αντιστοιχούν σε φυσιολογικά περιστατικά και σε έξι διαφορετικούς τύπους πνευμονικής νόσου. Επιπλέον, στο Σχήμα 5.4 αποδίδεται γραφικά η κατανομή των δειγμάτων ανά κλάση, αναδεικνύοντας την ανισορροπία του πλήθους εικόνων μεταξύ των κλάσεων, η οποία απαιτεί ειδική μεθοδολογική αντιμετώπιση.

Πίνακας 5.2: Κατανομή του συνολικού αριθμού των ακτινογραφιών θώρακα ανά κλάση στο τελικό ενοποιημένο σύνολο δεδομένων.

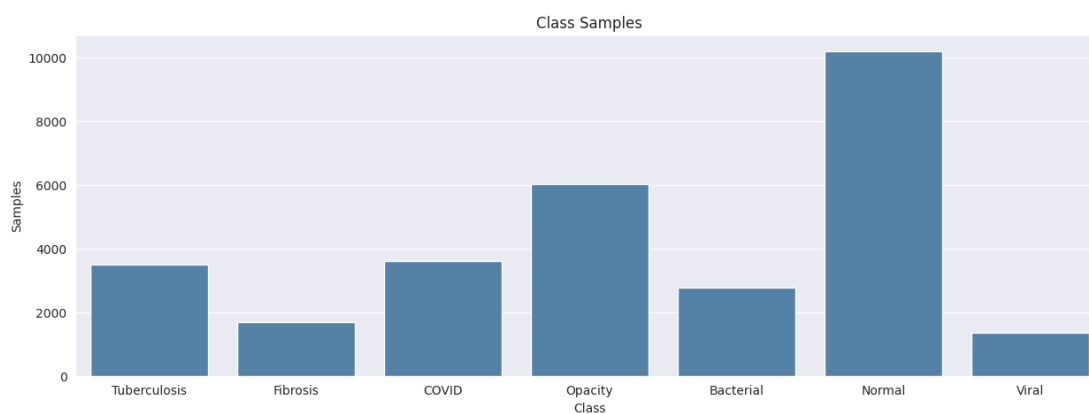
Class	Total Number of Images
Bacterial	2.780
COVID	3.616
Fibrosis	1.686
Normal	10.192
Opacity	6.012
Tuberculosis	3.500
Viral	1.345
Total	29.131

5.4.2 Προεπεξεργασία Δεδομένων

Στα προβλήματα ταξινόμησης εικόνων με χρήση CNNs, αποτελεί ευρέως διαδεδομένη πρακτική η εκπαίδευση των μοντέλων σε εικόνες σταθερών διαστάσεων 224×224 εικονοστοιχείων [362]. Η συγκεκριμένη ανάλυση έχει καθιερωθεί ως πρότυπο στον τομέα της CV, καθώς προσφέρει ικανοποιητική ισορροπία μεταξύ υπολογιστικής αποδοτικότητας και διατήρησης κρίσιμων μορφολογικών χαρακτηριστικών, απαραίτητων για την ακριβή κατηγοριοποίηση. Ως εκ τούτου, όλες οι εικόνες του ενοποιημένου συνόλου δεδομένων αναπροσαρμόστηκαν στις



Σχήμα 5.3: Ενδεικτικά δείγματα ακτινογραφιών θώρακα από το ενοποιημένο σύνολο δεδομένων. Κάθε δείγμα αντιστοιχεί σε μία από τις επτά κλάσεις.



Σχήμα 5.4: Κατανομή του πλήθους εικόνων ακτινογραφιών θώρακα ανά κλάση στο ενοποιημένο σύνολο δεδομένων. Η σημαντική διαφοροποίηση στο πλήθος δειγμάτων μεταξύ των κλάσεων αναδεικνύει την ύπαρξη ανισορροπίας στο σύνολο δεδομένων.

διαστάσεις 224×224 εικονοστοιχείων, προκειμένου να εξασφαλιστεί η συμβατότητα με τις απαιτήσεις των επιλεγμένων συνελκτικών αρχιτεκτονικών.

Πέραν της αναπροσαρμογής του μεγέθους, οι εικόνες αναπαρίστανται ως πίνακες ακέραιων αριθμών, όπου κάθε στοιχείο αντιστοιχεί στην τιμή φωτεινότητας ενός εικονοστοιχείου. Για εικόνες 8-bit σε κλίμακα του γκρι, οι τιμές αυτές κυμαίνονται από 0 (μαύρο) έως 255 (λευκό). Η συγκεκριμένη αναπαράσταση αποτελεί το πρότυπο μορφοποίησης για ψηφιακές εικόνες, καθώς παρέχει επαρκή ανάλυση για την απεικόνιση λεπτομερειών. Η χρήση ακέραιων τιμών επιτρέπει την αποδοτική αποθήκευση και την ταχεία αριθμητική επεξεργασία από τους αλγόριθμους DL και ιδίως από τα CNNs. Επιπλέον, η συγκεκριμένη μορφή δεδομένων είναι άμεσα συμβατή με τα πρότυπα των βιβλιοθηκών της CV, οι οποίες αναμένουν ως είσοδο έναν δισδιάστατο ή τρισδιάστατο πίνακα, με κάθε στοιχείο να αντιπροσωπεύει την ένταση φωτεινότητας ή τη χρωματική πληροφορία ενός εικονοστοιχείου ανά κανάλι [363, 364].

Με σκοπό τη μείωση της υπολογιστικής πολυπλοκότητας και τη βελτίωση της αριθμητικής σταθερότητας κατά την εκπαίδευση, εφαρμόστηκε κανονικοποίηση των τιμών φωτεινότητας στο συνεχές διάστημα $[0, 1]$. Η κανονικοποίηση επιτυγχάνεται μέσω διαίρεσης της κάθε τιμής με το μέγιστο επιτρεπτό όριο (255). Η τεχνική αυτή ενισχύει τη σταθερότητα της διαδικασίας βελτιστοποίησης, επιταχύνει τη σύγκλιση του αλγορίθμου εκπαίδευσης και ενισχύει τη συμβατότητα μεταξύ των επιμέρους στρωμάτων του CNN [365].

5.4.3 Διαχωρισμός του Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιείται στην παρούσα εργασία αποτελείται συνολικά από 29.131 ακτινογραφίες θώρακα. Η συλλογή αυτή υποβάλλεται σε διαδικασία πενταπλής διασταυρούμενης επικύρωσης [366], με στόχο τη διασφάλιση της αξιοπιστίας και της γενικευσιμότητας του προτεινόμενου CNN κατηγοριοποίησης. Το πλήρες σύνολο διαχωρίζεται σε πέντε ισομεγέθη υποσύνολα (πτύχες). Σε κάθε επανάληψη της διαδικασίας, οι τέσσερις πτυχές χρησιμοποιούνται για την εκπαίδευση του μοντέλου και η πέμπτη για την αξιολόγηση της απόδοσής του. Η διαδικασία επαναλαμβάνεται πέντε φορές, έτσι ώστε κάθε πτυχή να χρησιμοποιηθεί ακριβώς μία φορά ως σύνολο ελέγχου.

Η μέθοδος της πενταπλής διασταυρούμενης επικύρωσης [367] παρουσιάζει σημαντικά πλεονεκτήματα. Πρώτον, αξιοποιείται στο μέγιστο δυνατό βαθμό τόσο το σύνολο εκπαίδευσης όσο και το σύνολο ελέγχου, επιτρέποντας στο μοντέλο να αξιολογηθεί πάνω στο πλήρες σύνολο δεδομένων, γεγονός που ενισχύει τη γενικευσιμότητα των ευρημάτων. Δεύτερον, παρέχει πιο αξιόπιστη εκτίμηση της απόδοσης του μοντέλου σε άγνωστα δεδομένα, περιορίζοντας τη μεταβλητότητα που μπορεί να προκύψει από έναν μεμονωμένο τυχαίο διαχωρισμό. Η συγκεκριμένη διαδικασία οδηγεί στην κατασκευή ενός ανθεκτικού μοντέλου, μειώνοντας τον κίνδυνο υπερπροσαρμογής και ενισχύοντας την αξιοπιστία του σε εναλλασσόμενα υποσύνολα δεδομένων.

Κατά τη φάση της εκπαίδευσης, το CNN βελτιστοποιεί τα βάρη του μέσω της ελαχιστοποίησης της απόκλισης μεταξύ των προβλεπόμενων και των πραγματικών εξόδων συμπερασμάτων. Η διαδικασία αυτή περιλαμβάνει επαναληπτικές προσαρμογές των παραμέτρων του δικτύου, με στόχο τη βελτίωση της ικανότητάς του να ταξινομεί με ακρίβεια τις ακτινογραφίες θώρακα. Το σύνολο εκπαίδευσης διαδραματίζει καθοριστικό ρόλο σε αυτό το στάδιο, καθώς προσφέρει ένα εκτενές και ποικιλόμορφο σύνολο δειγμάτων, το οποίο αξιοποιείται από το δίκτυο για την εκμάθηση και τη βελτίωση των εσωτερικών αναπαραστάσεών του.

Αντίστοιχα, το σύνολο ελέγχου λειτουργεί ως ουδέτερο μέτρο αξιολόγησης της απόδοσης του εκπαιδευμένου μοντέλου. Ο Πίνακας 5.3 παρουσιάζει την κατανομή του αριθμού των ακτινογραφιών θώρακα ανά κατηγορία πνευμονικής πάθησης σε κάθε πτυχή της διαδικασίας. Το εκάστοτε σύνολο ελέγχου περιλαμβάνει δείγματα που δεν έχουν χρησιμοποιηθεί κατά την εκπαίδευση, επιτρέποντας την αποτίμηση της απόδοσης του μοντέλου σε εντελώς άγνωστα δεδομένα, προσομοιώνοντας έτσι πραγματικές συνθήκες κλινικής εφαρμογής.

Πίνακας 5.3: Αναλυτική κατανομή του αριθμού ακτινογραφιών θώρακα που χρησιμοποιήθηκαν για εκπαίδευση (Tr) και έλεγχο (Ts) σε κάθε μία από τις πέντε πτυχές της διαδικασίας πενταπλής διασταυρούμενης επικύρωσης.

Fold	Normal (Tr/Ts)	Opacity (Tr/Ts)	COVID (Tr/Ts)	Tuberculosis (Tr/Ts)	Bacterial (Tr/Ts)	Fibrosis (Tr/Ts)	Viral (Tr/Ts)
1	8.190/2.002	4.804/1.208	2.904/712	2.817/683	2.203/577	1.313/373	1.073/272
2	8.100/2.092	4.842/1.170	2.905/711	2.804/696	2.236/544	1.346/340	1.072/273
3	8.191/2.001	4.786/1.226	2.851/765	2.827/673	2.219/561	1.342/344	1.089/256
4	8.106/2.086	4.832/1.180	2.866/750	2.785/715	2.248/532	1.371/315	1.097/248
5	8.181/2.011	4.784/1.228	2.938/678	2.767/733	2.214/566	1.372/314	1.049/296

5.5 Μεθοδολογία Σχεδιασμού και Υλοποίησης του Συστήματος

5.5.1 Συνάρτηση Εστιακής Απώλειας

Στα προβλήματα αναγνώρισης ιατρικών εικόνων, η διαδικασία της ταξινόμησης διατυπώνεται συνήθως ως ένα πολυκατηγορικό πρόβλημα, στο οποίο στόχος είναι η ακριβής και αποδοτική αντιστοίχιση ενός δείγματος σε μία από τις προδιαγεγραμμένες κατηγορίες. Τα CNNs αποτελούν τη βασική αρχιτεκτονική επιλογή για την αντιμετώπιση τέτοιων προβλημάτων, καθώς προσφέρουν υψηλή ικανότητα εξαγωγής και αξιοποίησης των μορφολογικών χαρακτηριστικών από εικόνες.

Η πλέον διαδεδομένη συνάρτηση απώλειας που εφαρμόζεται σε περιπτώσεις πολυκατηγορικής ταξινόμησης είναι η διασταυρούμενη εντροπία [368]. Η συνάρτηση αυτή επιτρέπει τη μέτρηση της απόκλισης μεταξύ της πραγματικής κατανομής ετικετών και της προβλεπόμενης κατανομής πιθανοτήτων που εξάγεται από το μοντέλο. Κατά συνέπεια, αποτελεί βασικό μηχανισμό βελτιστοποίησης στη διαδικασία εκπαίδευσης μοντέλων επιβλεπόμενης μάθησης.

Η συνάρτηση απώλειας διασταυρούμενης εντροπίας (Cross-entropy) για πολυκατηγορική ταξινόμηση ορίζεται ως εξής:

$$\text{Cross-entropy}(p_c) = - \sum_{c=1}^N y_c \log(p_c) \quad (5.1)$$

όπου N είναι ο συνολικός αριθμός κατηγοριών, y_c είναι η μεταβλητή ένδειξης (με τιμή 1 όταν το δείγμα ανήκει στην κατηγορία c , και 0 διαφορετικά), και p_c η προβλεπόμενη πιθανότητα το δείγμα να ταξινομηθεί στην κατηγορία c .

Ωστόσο, η κλασική συνάρτηση απώλειας διασταυρούμενης εντροπίας δεν θεωρείται βέλτιστη σε περιπτώσεις όπου παρατηρείται έντονη ανισορροπία στην κατανομή των κατηγοριών, όπως συμβαίνει συχνά σε ιατρικά σύνολα δεδομένων, στα οποία οι σπάνιες παθολογικές καταστάσεις υποεκπροσωπούνται. Η άνιση κατανομή των δειγμάτων μπορεί να οδηγήσει σε μοντέλα που εμφανίζουν υψηλή ακρίβεια ως προς τις κυρίαρχες τάξεις, αλλά αποτυγχάνουν να ανιχνεύσουν επαρκώς τις σπάνιες

κατηγορίες, γεγονός που περιορίζει σημαντικά τη χρηστικότητα τους σε πραγματικά κλινικά περιβάλλοντα.

Για την αντιμετώπιση αυτής της πρόκλησης, οι Lin et al. [226] πρότειναν τη συνάρτηση εστιακής απώλειας ως εναλλακτική της διασταυρούμενης εντροπίας. Η συνάρτηση εστιακής απώλειας αποτελεί μια τροποποιημένη μορφή της απώλειας διασταυρούμενης εντροπίας, η οποία ενσωματώνει συντελεστές βαρύτητας και έναν μηχανισμό ενίσχυσης, με σκοπό να διαφοροποιήσει τη συμβολή των εύκολα και δύσκολα ταξινομήσιμων δειγμάτων. Ειδικότερα, η συνάρτηση εστιακής απώλειας μειώνει την επίδραση των παραδειγμάτων που ταξινομούνται με υψηλή βεβαιότητα και αποδίδει μεγαλύτερη βαρύτητα στα παραδείγματα που παρουσιάζουν υψηλότερη αβεβαιότητα, ενισχύοντας έτσι την ικανότητα του μοντέλου να μάθει από δύσκολες ή υποεκπροσωπούμενες περιπτώσεις.

Ο μαθηματικός ορισμός της συνάρτησης εστιακής απώλειας (Focal-loss) είναι ο ακόλουθος:

$$\text{Focal-loss}(p_c) = -\alpha_c(1 - p_c)^\gamma \log(p_c) \quad (5.2)$$

όπου p_c είναι η προβλεπόμενη πιθανότητα, α_c ο συντελεστής βαρύτητας για την κατηγορία c και γ είναι μία υπερπαραμέτρος που ελέγχει τον βαθμό ενίσχυσης της εστίασης.

Παρ' όλα αυτά, η αρχική μορφή της συνάρτησης εστιακής απώλειας έχει σχεδιαστεί κατά κύριο λόγο για την αντιμετώπιση προβλημάτων δυαδικής ταξινόμησης, όπου το σύστημα καλείται να διακρίνει μεταξύ δύο κατηγοριών. Η απευθείας εφαρμογή της σε πολυκατηγορικά προβλήματα, όπως είναι η ταξινόμηση πνευμονικών παθήσεων με βάση δεδομένα απεικόνισης, δεν επαρκεί για την αποτελεσματική αντιμετώπιση της εγγενούς πολυπλοκότητας και της κατηγορικής ανισορροπίας που τα χαρακτηρίζει. Συνεπώς, απαιτείται η επέκταση της εστιακής συνάρτησης απώλειας σε μορφή κατάλληλη για πολυκατηγορική ταξινόμηση, ώστε να επιτρέπει τη διαφοροποιημένη στάθμιση των επιμέρους κλάσεων και να ενισχύει τη συμβολή των σπάνιων ή δύσκολα προβλέψιμων περιπτώσεων στη διαδικασία μάθησης.

Η γενικευμένη έκφραση της συνάρτησης εστιακής απώλειας για πολλαπλές κατηγορίες διατυπώνεται ως εξής:

$$\text{Focal-loss}_{\text{multiple}}(FL) = - \sum_{c=1}^N \alpha_c (1 - p_c)^\gamma y_c \log(p_c) \quad (5.3)$$

$$y_c = \begin{cases} 1, & \text{αν } c \text{ είναι η πραγματική κατηγορία} \\ 0, & \text{διαφορετικά} \end{cases} \quad (5.4)$$

όπου α_c είναι ο παράγοντας βαρύτητας ανά κατηγορία, p_c η προβλεπόμενη πιθανότητα για την κατηγορία c , γ ο παράγοντας ενίσχυσης της εστίασης, και N ο συνολικός αριθμός των κατηγοριών (τύποι των πνευμονικών παθήσεων).

Στην παρούσα εργασία, υιοθετείται η προσαρμοσμένη εκδοχή της συνάρτησης εστιακής απώλειας, η οποία παραμετροποιήθηκε κατάλληλα προκειμένου να λαμβάνει υπόψη τις ιδιαιτερότητες των ιατρικών δεδομένων, την ετερογένεια των κατηγοριών και την έντονη ανισορροπία που τις χαρακτηρίζει, με γνώμονα τις αρχές της μεθοδολογίας μας (βλ. Ενότητα 3.2.2.3). Η ενσωμάτωση της προσαρμοσμένης συνάρτησης στη διαδικασία εκπαίδευσης συμβάλλει στη βελτίωση της ικανότητας του μοντέλου να ανιχνεύει ορθά τόσο τις θετικές όσο και τις αρνητικές περιπτώσεις. Ως αποτέλεσμα, επιτυγχάνεται μια εξισορροπημένη απόδοση μεταξύ των επιμέρους κατηγοριών, μειώνοντας τη μεροληψία υπέρ των υπερεκπροσωπούμενων κλάσεων και ενισχύοντας τη διαγνωστική αξιοπιστία του συστήματος.

5.5.2 Σχεδιασμός CNN Χαμηλής Πολυπλοκότητας

Το προτεινόμενο CNN που απεικονίζεται στο Σχήμα 5.5, σχεδιάστηκε ακολουθώντας την μεθοδολογία που στοχεύει στην επίτευξη ισορροπίας μεταξύ της εκφραστικής ισχύος του μοντέλου και της υπολογιστικής αποδοτικότητας, όπως αναλύεται στο Κεφάλαιο 3. Βασικό κριτήριο αποτέλεσε η υιοθέτηση της στρατηγικής εξισορρόπησης μεταξύ βάθους και εύρους (βλ. Ενότητα 3.2.1.1), έτσι ώστε το δίκτυο να μπορεί να αποτυπώνει κρίσιμες διαβαθμίσεις και πολυπλοκότητες των δεδομένων, αξιοποιώντας επαρκώς τα δομικά του χαρακτηριστικά για την εξαγωγή αφηρημένων και ποικιλόμορφων αναπαραστάσεων, χωρίς να οδηγείται σε υπερβολικό αριθμό παραμέτρων ή σε αυξημένο υπολογιστικό και ενεργειακό κόστος που θα καθιστούσε την υλοποίηση μη αποδοτική ή απαγορευτική για περιβάλλοντα με περιορισμένους πόρους.

Το προτεινόμενο χαμηλής πολυπλοκότητας μοντέλο ξεκινά με ένα στρώμα εισόδου, στο οποίο καθορίζεται το σχήμα των εικόνων εισόδου ως $(224 \times 224 \times 3)$, δηλαδή εικόνες με διαστάσεις 224 εικονοστοιχείων ανά πλευρά και τρία κανάλια χρώματος. Το πρώτο συνελκτικό μπλοκ περιλαμβάνει 64 φίλτρα διαστάσεων 5×5 , ακολουθούμενο από ένα στρώμα κανονικοποίησης κατά παρτίδες για τη σταθεροποίηση της κατανομής των ενεργοποιήσεων και το στρώμα συγκέντρωσης με μέγεθος παραθύρου 2×2 για μείωση των χωρικών διαστάσεων. Το δεύτερο συνελκτικό μπλοκ αποτελείται από 32 φίλτρα διαστάσεων 3×3 και επαναλαμβάνεται η ίδια ακολουθία, στρώμα κανονικοποίησης κατά παρτίδες και στρώμα συγκέντρωσης. Το τρίτο συνελκτικό μπλοκ περιλαμβάνει 16 φίλτρα διαστάσεων 3×3 και παρόμοια διαμόρφωση. Μετά την εξαγωγή χαρακτηριστικών μέσω των συνελκτικών στρωμάτων, προστίθεται το στρώμα ισοπέδωσης, το οποίο μετατρέπει την έξοδο των προηγούμενων στρωμάτων σε μονοδιάστατο διάνυσμα, προετοιμάζοντας τα δεδομένα για την επεξεργασία από τα πλήρως συνδεδεμένα στρώματα.

Το πρώτο πλήρως συνδεδεμένο στρώμα περιλαμβάνει 256 νευρώνες, ακολουθούμενο από κανονικοποίηση κατά παρτίδες και στρώμα απόσυρσης με ποσοστό 0.3 για την αποτροπή της υπερπροσαρμογής. Το δεύτερο πλήρως συνδεδεμένο στρώμα περιλαμβάνει 128 νευρώνες και επαναλαμβάνει την ίδια ακολουθία. Η έξοδος του δικτύου παράγεται από ένα τελικό στρώμα εξόδου με επτά νευρώνες, οι οποίες αντιστοιχούν στις επτά κατηγορίες των πνευμονικών παθήσεων. Η συνάρτηση ενεργοποίησης softmax εφαρμόζεται στο στρώμα εξόδου, μετατρέποντας τις ενεργοποιήσεις σε πιθανότητες και αποτυπώνοντας την πιθανότητα κάθε κατηγορίας ως έκβαση του ταξινομητή.

Τα συνελκτικά στρώματα με την κανονικοποίηση κατά παρτίδες επιτελούν τη βασική εξαγωγή χαρακτηριστικών, εφαρμόζοντας φίλτρα σε πολλαπλά επίπεδα αφαίρεσης και ανιχνεύοντας πρότυπα όπως άκρα, υφές και σχήματα. Η υποδειγματοληψία μέσω του στρώματος συγκέντρωσης συμβάλλει στη μείωση της υπολογιστικής πολυπλοκότητας και των χωρικών διαστάσεων, διατηρώντας παράλληλα τα σημαντικότερα χαρακτηριστικά. Τα πλήρως συνδεδεμένα στρώματα που ακολουθούν, εκτελούν τη διαδικασία κατηγοριοποίησης, αντιστοιχίζοντας τα εξαγόμενα χαρακτηριστικά σε συγκεκριμένες κατηγορίες.

Ο Πίνακας 5.4 συνοψίζει τα επιμέρους στρώματα και τις συνολικές παραμέτρους του προτεινόμενου μοντέλου χαμηλής πολυπλοκότητας. Ο συνολικός αριθμός παραμέτρων (FP32) του μοντέλου ανέρχεται σε 3.275.255. Αυτές οι παράμετροι

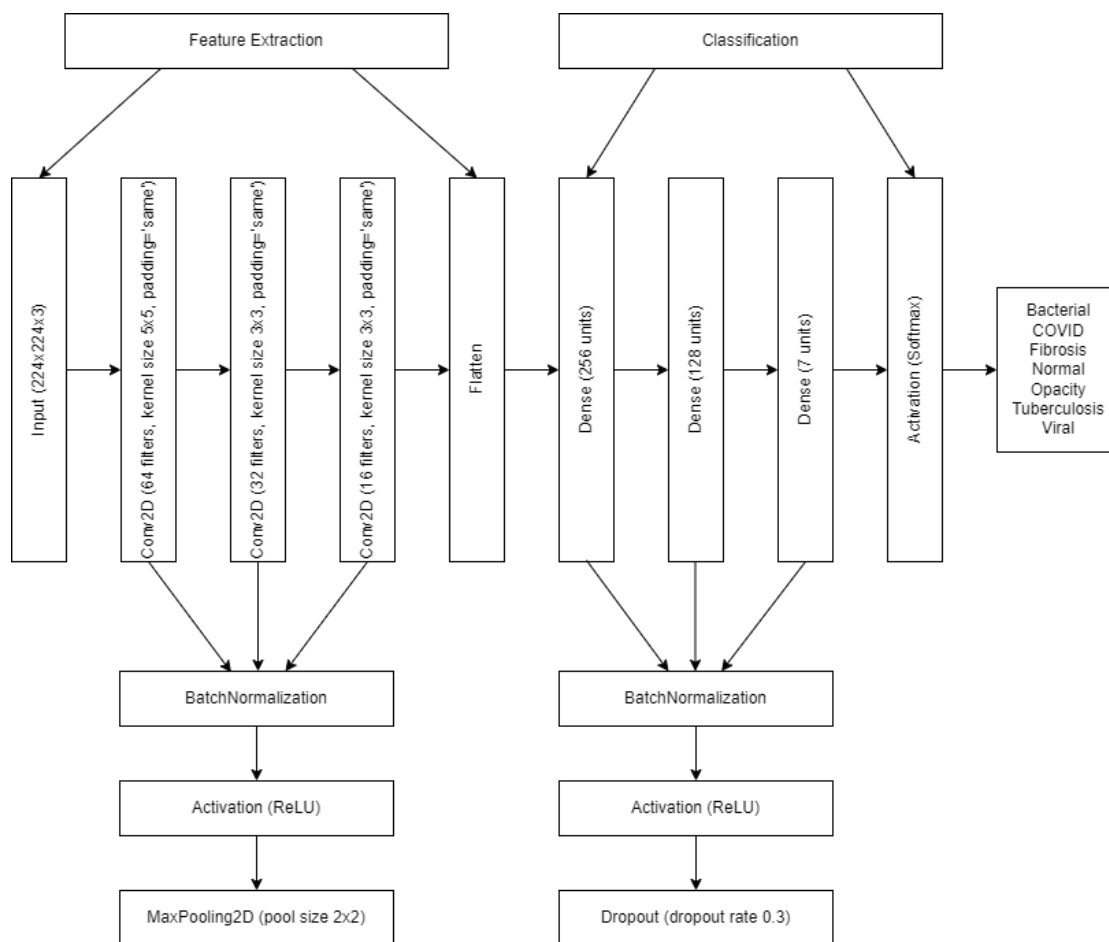
περιλαμβάνουν όλα τα βάρη και τις τιμές μετατόπισης που σχετίζονται με τα συνελικτικά φίλτρα, τα πλήρως συνδεδεμένα επίπεδα, καθώς και τα στρώματα κανονικοποίησης. Από το σύνολο αυτό, οι 3.274.263 χαρακτηρίζονται ως εκπαιδεύσιμες παράμετροι, δηλαδή μεταβλητές που τροποποιούνται δυναμικά κατά τη διάρκεια της διαδικασίας εκπαίδευσης μέσω της αντίστροφης διάδοσης σφάλματος και του επιλεγμένου αλγορίθμου βελτιστοποίησης. Σε αυτές περιλαμβάνονται κυρίως τα βάρη των συνελικτικών και πλήρως συνδεδεμένων στρωμάτων, καθώς και οι παράμετροι κλίμακας και μετατόπισης που χρησιμοποιούνται στα στρώματα κανονικοποίησης κατά παρτίδες.

Οι υπόλοιπες 992 αποτελούν τις μη εκπαιδεύσιμες παραμέτρους. Στο προτεινόμενο μοντέλο, αυτές οι παράμετροι προέρχονται από τα πέντε στρώματα κανονικοποίησης κατά παρτίδες, τα οποία, εκτός από τις εκπαιδεύσιμες παραμέτρους (κλίμακα και μετατόπιση), διατηρούν και εσωτερικές στατιστικές μεταβλητές — τη μέση τιμή και τη διασπορά — οι οποίες χαρακτηρίζονται ως μη εκπαιδεύσιμες. Οι συγκεκριμένες μη εκπαιδεύσιμες παράμετροι χρησιμοποιούνται κυρίως στο στάδιο της πρόβλεψης, συμβάλλοντας στη βελτίωση της γενικευτικής ικανότητας του μοντέλου, χωρίς να επιβαρύνουν τον υπολογιστικό φόρτο κατά τη διαδικασία εκπαίδευσης.

Η καινοτομία του παρόντος μοντέλου έγκειται στη δομική του σχεδίαση, η οποία έχει βελτιστοποιηθεί με γνώμονα την ανάπτυξη σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους, όπως οι συσκευές NVIDIA Jetson Xavier. Ο σχεδιασμός του μοντέλου επιτυγχάνει την ισορροπία μεταξύ της ακρίβειας και της υπολογιστικής αποδοτικότητας, γεγονός κρίσιμο για εφαρμογές σε φορητά διαγνωστικά συστήματα και απομακρυσμένες υπηρεσίες υγειονομικής φροντίδας που απαιτούν υψηλή απόδοση χωρίς να υπερφορτώνουν τη διαθέσιμη υποδομή. Η βελτιστοποίηση της αρχιτεκτονικής αντικατοπτρίζεται στην προσεκτική επιλογή του αριθμού των συνελικτικών και πλήρως συνδεδεμένων στρωμάτων, καθώς και στην ενσωμάτωση τεχνικών όπως η κανονικοποίηση κατά παρτίδες και η απόσυρση, που ενισχύουν τη γενικευτική ικανότητα του μοντέλου.

5.5.3 Σχεδιασμός Προεκπαιδευμένων CNNs

Στην παρούσα μελέτη αξιολογήθηκαν έξι προεκπαιδευμένα CNNs: ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT-B/16. Κάθε προεκπαιδευμένο μοντέλο αξιοποιήθηκε με τα ήδη εκπαιδευμένα βάρη του, τα



Σχήμα 5.5: Δομική απεικόνιση του CNN χαμηλής πολυπλοκότητας που αναπτύχθηκε σύμφωνα με τη μεθοδολογία, για την κατηγοριοποίηση πνευμονικών παθήσεων από ακτινογραφίες θώρακα.

οποία έχουν προκύψει μέσω εκπαίδευσης στο ευρέως χρησιμοποιούμενο σύνολο δεδομένων ImageNet [369]. Το εν λόγω σύνολο περιλαμβάνει 1.000 κατηγορίες, και προσφέρει τη δυνατότητα στα δίκτυα να μάθουν αναπαραστάσεις υψηλού επιπέδου που σχετίζονται με γενικά οπτικά χαρακτηριστικά (όπως ακμές, υφές και σύνθετα σχήματα).

Για τη μετάβαση από την αρχική ταξινόμηση γενικών αντικειμένων σε ένα νέο πρόβλημα ταξινόμησης πνευμονικών παθήσεων μέσω ακτινογραφιών θώρακα, πραγματοποιείται τροποποίηση στο τελικό τμήμα των προεκπαιδευμένων δικτύων. Συγκεκριμένα, αφαιρούνται τα τελικά πλήρως συνδεδεμένα επίπεδα, τα οποία έχουν εκπαιδευτεί για έξοδο σε 1.000 κατηγορίες του ImageNet, καθώς η παραγόμενη πληροφορία δεν είναι σχετική με τις υπό μελέτη κατηγορίες παθήσεων.

Η αποκοπή των αρχικών στρωμάτων εξόδου επιτρέπει την αντικατάστασή τους με ένα νέο προσαρμοσμένο ταξινομητή, κατάλληλο για το νέο πρόβλημα

Πίνακας 5.4: Οι προδιαγραφές της αρχιτεκτονικής του CNN χαμηλής πολυπλοκότητας για την αναγνώριση πνευμονικών παθήσεων από ακτινογραφίες θώρακα.

Layer (type)	Output Shape	Param #	Trainable
InputLayer	(None, 224, 224, 3)	0	
Conv2D_1	(None, 224, 224, 64)	4.864	True
BatchNormalization_1	(None, 224, 224, 64)	256	True
Activation_1	(None, 224, 224, 64)	0	
MaxPooling2D_1	(None, 112, 112, 64)	0	
Conv2D_2	(None, 112, 112, 32)	18.464	True
BatchNormalization_2	(None, 112, 112, 32)	128	True
Activation_2	(None, 112, 112, 32)	0	
MaxPooling2D_2	(None, 56, 56, 32)	0	
Conv2D_3	(None, 56, 56, 16)	4.624	True
BatchNormalization_3	(None, 56, 56, 16)	64	True
Activation_3	(None, 56, 56, 16)	0	
MaxPooling2D_3	(None, 28, 28, 16)	0	
Flatten_1	(None, 12544)	0	
Dense_1	(None, 256)	3.211.520	True
BatchNormalization_4	(None, 256)	1.024	True
Activation_4	(None, 256)	0	
Dropout_1	(None, 256)	0	
Dense_2	(None, 128)	32.896	True
BatchNormalization_5	(None, 128)	512	True
Activation_5	(None, 128)	0	
Dropout_2	(None, 128)	0	
Dense_3	(None, 7)	903	True
Activation_6	(None, 7)	0	
Total params: 3.275.255			
Trainable params: 3.274.263			
Non-trainable params: 992			

κατηγοριοποίησης. Έτσι, διατηρείται η προεκπαιδευμένη ικανότητα του μοντέλου στην εξαγωγή γενικών οπτικών χαρακτηριστικών, ενώ παράλληλα παρέχεται η δυνατότητα εκμάθησης εξειδικευμένων αναπαραστάσεων που σχετίζονται με τις κατηγορίες των πνευμονικών παθήσεων.

Για τη βέλτιστη προσαρμογή των μοντέλων στο πρόβλημα κατηγοριοποίησης των πνευμονικών παθήσεων, τα προεκπαιδευμένα βάρη διατηρήθηκαν σταθερά, και προστέθηκαν τρία προσαρμοσμένα πλήρως συνδεδεμένα επίπεδα. Το πρώτο επίπεδο περιλαμβάνει 256 νευρώνες, ενώ ακολουθεί το στρώμα κανονικοποίησης κατά παρτίδες και απόσυρσης με ποσοστό 0,3, με στόχο την αποτροπή της υπερπροσαρμογής. Το επόμενο επίπεδο αποτελείται από 128 νευρώνες και συνοδεύεται από τα ίδια στρώματα κανονικοποίησης κατά παρτίδες και απόσυρσης. Η έξοδος του κάθε μοντέλου προκύπτει από ένα τελικό επίπεδο με 7 νευρώνες, όσες

και οι κατηγορίες ταξινόμησης, και ενεργοποιείται μέσω της συνάρτησης softmax, η οποία μετατρέπει τις τιμές εξόδου σε κατανομή πιθανοτήτων, υποδηλώνοντας τη σχετική πιθανότητα κάθε εισόδου να ανήκει σε συγκεκριμένη κατηγορία.

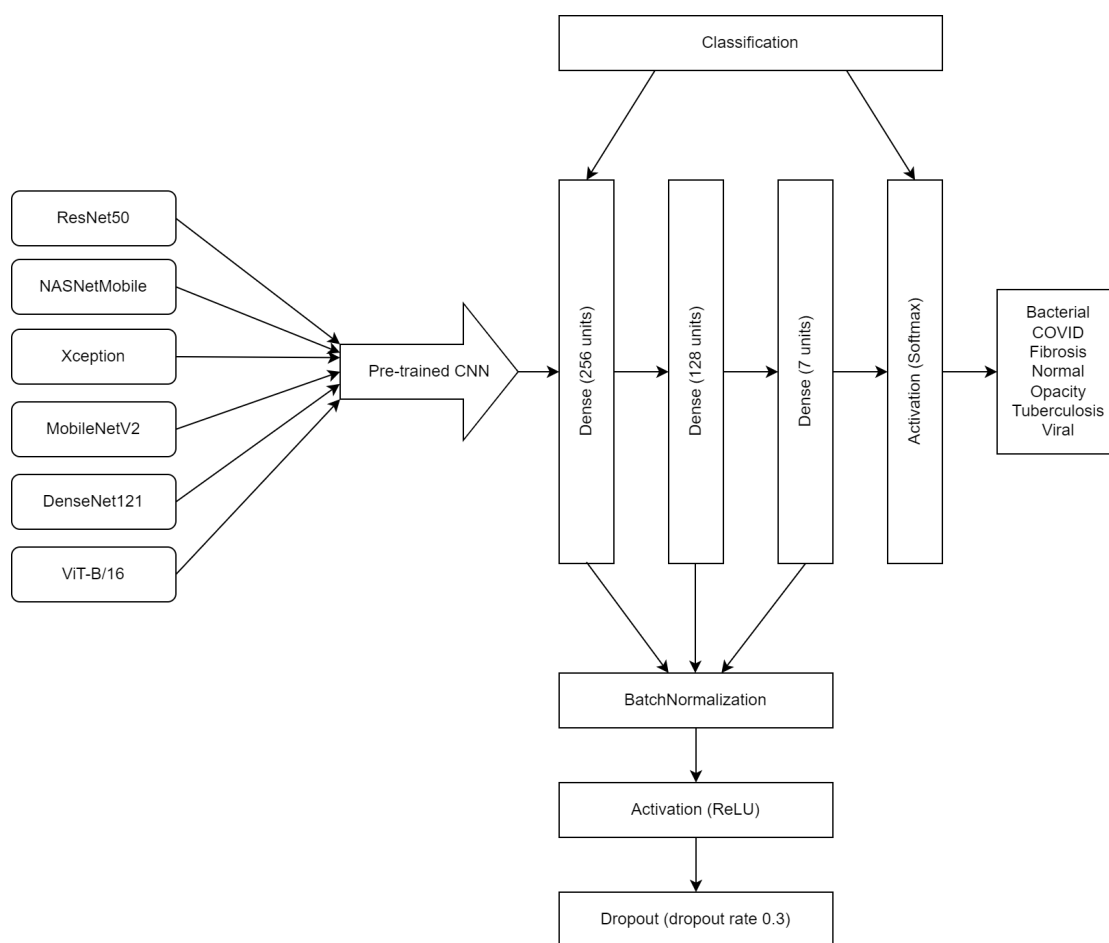
Ο αρχιτεκτονικός σχεδιασμός των έξι προεκπαιδευμένων μοντέλων απεικονίζεται στο Σχήμα 5.6. Κατά τη διαδικασία παραμετροποίησης, βασική προτεραιότητα αποτέλεσε η διατήρηση υψηλών επιδόσεων, τόσο για το προτεινόμενο CNN χαμηλής πολυπλοκότητας όσο και για τα προεκπαιδευμένα. Ο Πίνακας 5.5 παρουσιάζει αναλυτικά τα τεχνικά χαρακτηριστικά των μοντέλων που χρησιμοποιήθηκαν και αξιολογήθηκαν στην παρούσα μελέτη. Συγκεκριμένα, παρατίθεται για κάθε αρχιτεκτονική ο συνολικός αριθμός παραμέτρων, οι εκπαιδεύσιμες παράμετροι, οι μη εκπαιδεύσιμες παράμετροι, η απαιτούμενη μνήμη καθώς και οι υπολογιστικές απαιτήσεις σε πράξεις κινητής υποδιαστολής.

Το προτεινόμενο CNN χαμηλής πολυπλοκότητας διακρίνεται για τη μειωμένη παραμετρική και υπολογιστική του επιβάρυνση, γεγονός που το καθιστά ιδιαίτερα κατάλληλο για ανάπτυξη και υλοποίηση σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους. Συγκεκριμένα, ο συνολικός αριθμός παραμέτρων ανέρχεται σε 3.275.255 (βλ. Ενότητα 3.4.3), γεγονός που συνεπάγεται εξαιρετικά χαμηλές απαιτήσεις αποθήκευσης και επεξεργασίας σε σχέση με τα αντίστοιχα προεκπαιδευμένα μοντέλα. Οι απαιτήσεις σε υπολογιστικές πράξεις κινητής υποδιαστολής είναι επίσης περιορισμένες, με συνολικό κόστος 3,745 GFLOP ανά εξαγωγή συμπεράσματος (βλ. Ενότητα 3.4.3), ενώ η απαιτούμενη μνήμη δεν υπερβαίνει τα 12,5 MB (βλ. Ενότητα 3.4.2).

Αντίθετα, τα προεκπαιδευμένα μοντέλα (ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT-B/16), παρά τη βελτιωμένη τους απόδοση στην ταξινόμηση γενικού σκοπού, εμφανίζουν σημαντικά αυξημένη υπολογιστική και μνημονική απαίτηση. Ενδεικτικά, το μοντέλο ResNet50 περιλαμβάνει 49.313.415 παραμέτρους — αριθμό σχεδόν 15 φορές μεγαλύτερο από εκείνον του προτεινόμενου μοντέλου — με μνήμη που φτάνει τα 98,13 MB και υπολογιστικό κόστος 8,043 GFLOP ανά εξαγωγή συμπεράσματος. Παρόμοια εικόνα παρουσιάζουν και τα μοντέλα Xception και NASNetMobile, με τις απαιτήσεις σε μνήμη να υπερβαίνουν τα 50 MB και το υπολογιστικό κόστος να κυμαίνεται άνω των 6 GFLOP ανά δείγμα.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η περίπτωση του μοντέλου ViT-B/16, το οποίο ανήκει στην οικογένεια των οπτικών μετασχηματιστών (Vision Transformers) και αξιοποιείται συχνά σε προβλήματα ταξινόμησης εικόνων υψηλής ακρίβειας.

Παρότι παρουσιάζει εξαιρετικά υψηλό συνολικό αριθμό παραμέτρων 90.032.647, μόλις το 4,7% αυτών είναι εκπαιδύσιμο κατά την διαδικασία εκπαίδευσης, καθώς η πλειονότητα αφορά παγωμένα βάρη που έχουν προκύψει από την προεκπαίδευση στο σύνολο δεδομένων ImageNet. Το σχετικά περιορισμένο αποτύπωμα μνήμης 16,15 MB εξηγείται από το γεγονός ότι μεγάλο μέρος της υποδομής του μοντέλου παραμένει στατικό κατά την εκπαίδευση. Παρ' όλα αυτά, οι υπολογιστικές απαιτήσεις είναι αυξημένες, αγγίζοντας τα 5,853 GFLOP ανά εξαγωγή συμπεράσματος, γεγονός που καθιστά την υλοποίησή του πιο απαιτητική σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους.



Σχήμα 5.6: Δομική απεικόνιση της αρχιτεκτονικής των προεκπαιδευμένων CNNs για την ταξινόμηση των πνευμονικών παθήσεων από ακτινογραφίες θώρακα.

Πίνακας 5.5: Συγκριτική παρουσίαση προδιαγραφών των CNNs για την ταξινόμηση πνευμονικών παθήσεων από ακτινογραφίες θώρακα.

Model	Total Params	Trainable Params	Non-Trainable Params	Memory (MB)	Giga flop
Proposed lightweight CNN (CE)	3.275.255	3.274.263	992	12,49	3,745
Proposed lightweight CNN (FE)	3.275.255	3.274.263	992	12,49	3,745
ResNet50 (CE)	49.313.415	25.724.935	23.588.480	98,13	8,043
ResNet50 (FE)	49.313.415	25.724.935	23.588.480	98,13	8,043
NASNetMobile (CE)	17.551.771	13.281.287	4.270.484	50,66	6,604
NASNetMobile (FE)	17.551.771	13.281.287	4.270.484	50,66	6,604
Xception (CE)	46.587.183	25.724.935	20.862.248	98,13	8,493
Xception (FE)	46.587.183	25.724.935	20.862.248	98,13	8,493
MobileNetV2 (CE)	18.349.895	16.091.143	2.258.752	61,38	5,679
MobileNetV2 (FE)	18.349.895	16.091.143	2.258.752	61,38	5,679
DenseNet121 (CE)	19.918.151	12.879.879	7.038.272	49,13	4,689
DenseNet121 (FE)	19.918.151	12.879.879	7.038.272	49,13	4,689
ViT-B/16 (CE)	90.032.647	4.233.223	85.799.424	16,15	5,853
ViT-B/16 (FE)	90.032.647	4.233.223	85.799.424	16,15	5,853

5.6 Πειραματικά Αποτελέσματα

5.6.1 Ετερογενές Σύστημα Επιτάχυνσης NVIDIA Jetson AGX Xavier

Στο σύγχρονο ερευνητικό πεδίο της AI, παρατηρείται έντονη στροφή προς την αξιοποίηση παραλληλισμένων υπολογιστικών αρχιτεκτονικών. Μια ιδιαίτερως ελκυστική προσέγγιση περιλαμβάνει τη χρήση GPUs, που διαθέτουν πολλαπλούς πυρήνες επεξεργασίας και υποστηρίζουν παράλληλη εκτέλεση εντολών μέσω εξειδικευμένων λογισμικών. Οι GPUs έχουν καθιερωθεί ως θεμελιώδης υπολογιστική υποδομή για την εκτέλεση υπολογιστικά απαιτητικών εφαρμογών, όπως είναι η DL και η ανάλυση ιατρικών δεδομένων, λόγω της ικανότητάς τους να εκτελούν μαζικά παραλληλοποιημένες πράξεις σε υψηλούς ρυθμούς.

Ενδεικτικό παράδειγμα αυτής της τεχνολογικής κατεύθυνσης αποτελεί το ετερογενές σύστημα επιτάχυνσης NVIDIA Jetson AGX Xavier [370], το οποίο αποτελεί μια ολοκληρωμένη ενσωματωμένη λύση υψηλής απόδοσης για εφαρμογές AI και DL σε περιβάλλοντα με περιορισμένους πόρους. Η πλατφόρμα συνδυάζει έναν υψηλών επιδόσεων επεξεργαστή GPU με έναν προσαρμοσμένο επεξεργαστή CPU και εξειδικευμένους επιταχυντές για CNNs [371].

Πιο συγκεκριμένα, το Jetson AGX Xavier ενσωματώνει έναν οκταπύρηνο επεξεργαστή ARMv8.2 64-bit με αρχιτεκτονική Carmel, προσφέροντας ισχυρές δυνατότητες γενικής υπολογιστικής επεξεργασίας. Η μονάδα GPU βασίζεται στην

αρχιτεκτονική NVIDIA Volta και περιλαμβάνει 512 πυρήνες CUDA και 64 πυρήνες Tensor, παρέχοντας σημαντική επιτάχυνση για εργασίες DL.

Επιπλέον, διαθέτει δύο ανεξάρτητες μονάδες NVDLA (NVIDIA Deep Learning Accelerator), οι οποίες υποστηρίζουν την αποδοτική εκτέλεση συνελκτικών και πλήρως συνδεδεμένων στρωμάτων για CNNs με υψηλή ενεργειακή απόδοση. Η μνήμη της πλατφόρμας ανέρχεται στα 16 GB τύπου LPDDR4x, επιτρέποντας την εκτέλεση μεγάλων μοντέλων με υψηλές απαιτήσεις εύρους ζώνης. Ο αποθηκευτικός χώρος φτάνει τα 32 GB eMMC, εξυπηρετώντας την αποθήκευση του λειτουργικού συστήματος, των εφαρμογών και των δεδομένων των μοντέλων.

Η πλατφόρμα υποστηρίζει ποικίλες δυνατότητες συνδεσιμότητας, όπως Gigabit Ethernet, θύρες USB 3.1, PCIe Gen3 x16, καθώς και εξόδους HDMI και DisplayPort. Οι δυνατότητες αυτές εξασφαλίζουν υψηλό βαθμό ευελιξίας και επεκτασιμότητας, καθιστώντας την πλατφόρμα κατάλληλη για προηγμένες αρχιτεκτονικές ενσωματωμένων συστημάτων και εφαρμογών AI αιχμής. Επιπλέον, είναι πλήρως συμβατή με το οικοσύστημα παράλληλης υπολογιστικής CUDA της NVIDIA, καθώς και με γνωστά πλαίσια ανάπτυξης DL όπως τα TensorFlow, PyTorch και Caffe.

Το ιδιαίτερο χαρακτηριστικό της πλατφόρμας έγκειται στην σχέση της υπολογιστικής απόδοσης και της ενεργειακής κατανάλωσης, η οποία κυμαίνεται μεταξύ 10 και 30 Watt, ανάλογα με τη φύση και την ένταση του υπολογιστικού φορτίου. Η δυνατότητα λειτουργίας υπό συνθήκες χαμηλής ενεργειακής κατανάλωσης, σε συνδυασμό με την υψηλή απόδοση επεξεργασίας, καθιστά την πλατφόρμα κατάλληλη για εφαρμογές οι οποίες υλοποιούνται σε φορητά διαγνωστικά συστήματα ή απομακρυσμένα ιατρικά περιβάλλοντα, όπου οι διαθέσιμοι ενεργειακοί και υπολογιστικοί πόροι είναι περιορισμένοι [372].

Στην παρούσα μελέτη, πραγματοποιήθηκε πειραματική αξιολόγηση της πλατφόρμας NVIDIA Jetson AGX Xavier, αξιοποιώντας τις προχωρημένες δυνατότητες επεξεργασίας της GPU για την υλοποίηση μοντέλων DL σε ιατρικά δεδομένα εικόνας. Η ενσωματωμένη μονάδα Volta GPU με Tensor Cores προσφέρει την απαραίτητη υπολογιστική ισχύ για την εκτέλεση χαμηλού λανθάνοντος χρόνου και υψηλής απόδοσης, κριτήρια τα οποία είναι κρίσιμα για εφαρμογές σε περιβάλλοντα σημείου φροντίδας. Οι συγκριτικές αξιολογήσεις περιλάμβαναν μετρήσεις ταχύτητας συμπερασμάτων, ρυθμό καρτέ ανά δευτερόλεπτο (FPS), και υπολογιστική αποδοτικότητα.

5.6.2 Μετρικές Ποιότητας

Η αξιολόγηση της απόδοσης ενός μοντέλου είναι κρίσιμη για τη συνολική εκτίμηση της λειτουργικότητας και της αξιοπιστίας του, ιδίως σε εφαρμογές με υψηλές απαιτήσεις ακρίβειας εξαγωγής συμπερασμάτων, όπως η ιατρική διάγνωση. Οι μετρικές ποιότητας, σύμφωνα με την μεθοδολογία (βλ. Ενότητα 3.3), παρέχουν ποσοτικά κριτήρια για την αποτίμηση της ικανότητας του μοντέλου να πραγματοποιεί ορθή εξαγωγή συμπερασμάτων σε προβλήματα ταξινόμησης.

Η αποτίμηση της απόδοσης του προτεινόμενου CNN χαμηλής πολυπλοκότητας, καθώς και των έξι προεκπαιδευμένων μοντέλων, πραγματοποιήθηκε με βάση ένα σύνολο καθιερωμένων μετρικών αξιολόγησης. Οι μετρικές που χρησιμοποιήθηκαν περιλαμβάνουν την συνολική ακρίβεια, την ακρίβεια των θετικών προβλέψεων, την ανάκληση, την βαθμολογία F1 και την βαθμολογία AUC [373]. Κάθε μία από τις εν λόγω μετρικές αποτυπώνει διαφορετική πτυχή της συμπεριφοράς του μοντέλου και η συνδυαστική τους χρήση καθίσταται απαραίτητη, ιδίως σε περιβάλλοντα με κατηγορική ανισορροπία ή ιδιαίτερη ευαισθησία στις ψευδείς προβλέψεις.

Η συνολική ακρίβεια ορίζεται ως ο λόγος του πλήθους των ορθά ταξινομημένων δειγμάτων προς το συνολικό αριθμό των προβλέψεων που παράγονται από το μοντέλο. Αποτελεί μία συνολική μετρική αξιολόγησης της απόδοσης ενός ταξινομητή, καθώς συνυπολογίζει τόσο τις αληθώς θετικές όσο και τις αληθώς αρνητικές περιπτώσεις. Η ακρίβεια προσφέρει μια συνολική εκτίμηση της ορθότητας των προβλέψεων του μοντέλου σε όλο το φάσμα των κατηγοριών, καθιστώντας την ιδιαίτερα χρήσιμη όταν οι κατηγορίες είναι ισομερώς κατανομημένες.

Η ακρίβεια των θετικών προβλέψεων αποτιμά το ποσοστό των προβλεπόμενων θετικών περιπτώσεων που αντιστοιχούν πραγματικά σε αληθείς θετικές περιπτώσεις. Αξιολογεί, επομένως, την αξιοπιστία των θετικών προβλέψεων του μοντέλου και την ικανότητα του να αποφεύγει τα ψευδώς θετικά σφάλματα. Η ανάκληση ποσοτικοποιεί την ικανότητα του μοντέλου να εντοπίζει σωστά τα θετικά δείγματα του συνόλου δεδομένων. Συγκεκριμένα, εκφράζει το ποσοστό των πραγματικά θετικών περιπτώσεων που αναγνωρίζονται επιτυχώς από τον ταξινομητή ως πραγματικά θετικές.

Η βαθμολογία F1 αποτελεί έναν ενοποιημένο δείκτη απόδοσης που συνδυάζει την ακρίβεια των θετικών προβλέψεων και την ανάκληση, υπολογίζοντας τον μέσο των δύο. Είναι ιδιαίτερα χρήσιμη σε περιπτώσεις ανισορροπίας των δεδομένων

κατηγοριών εισόδου, καθώς προσφέρει εξισορροπημένη εκτίμηση της ικανότητας του μοντέλου τόσο ως προς την ορθή αναγνώριση των θετικών περιπτώσεων, όσο και ως προς την αποφυγή ψευδώς θετικών προβλέψεων.

Η βαθμολογία AUC αποτελεί έναν συνολικό δείκτη απόδοσης που αποτυπώνει την ικανότητα του μοντέλου να διακρίνει μεταξύ των διαφόρων κατηγοριών, ανεξαρτήτως της ρύθμιση του κατωφλίου ταξινόμησης που επιλέγεται. Υπολογίζεται ως το εμβαδόν κάτω από την καμπύλη ROC, η οποία απεικονίζει τη σχέση μεταξύ του ποσοστού των αληθώς θετικών και των ψευδώς θετικών προβλέψεων για όλα τα δυνατά κατώφλια. Η AUC προσφέρει μια αντικειμενική αξιολόγηση της ταξινομητικής ικανότητας του μοντέλου, γεγονός που την καθιστά εξαιρετικά χρήσιμη σε περιπτώσεις κατηγορικής ανισορροπίας ή σε προβλήματα όπου η επιλογή του βέλτιστου κατωφλίου δεν είναι εκ των προτέρων γνωστή. Μια υψηλή τιμή AUC υποδηλώνει ότι το μοντέλο μπορεί να διακρίνει αποτελεσματικά μεταξύ των θετικών και των αρνητικών περιπτώσεων, εξασφαλίζοντας αξιόπιστη και ισορροπημένη απόδοση.

5.6.3 Φάση Εκπαίδευσης και Επικύρωσης

Η διαδικασία εκπαίδευσης του κάθε CNN πραγματοποιείται για 30 εποχές ανά πτυχή της πενταπλής διασταυρούμενης επικύρωσης, διατηρώντας σταθερό ρυθμό εκμάθησης ίσο με 0,0001 (βλ. Ενότητα 3.4.1). Η βελτιστοποίηση των παραμέτρων του μοντέλου πραγματοποιείται μέσω του αλγορίθμου Adam [374], ο οποίος αποτελεί μια προσαρμοστική εκδοχή της στοχαστικής καθόδου μέγιστης κλίσης (Stochastic Gradient Descent, SGD) [375]. Ο αλγόριθμος Adam επιταχύνει τη διαδικασία εκπαίδευσης, προσαρμόζοντας δυναμικά τους ρυθμούς μάθησης για κάθε παράμετρο, εξασφαλίζοντας ταχύτερη και σταθερότερη σύγκλιση, ακόμη και σε προβλήματα με θορυβώδη δεδομένα ή περιορισμένο αριθμό δειγμάτων.

Το μέγεθος παρτίδας ορίστηκε σε 16 δείγματα, προκειμένου να επιτευχθεί μια βέλτιστη ισορροπία μεταξύ υπολογιστικής επιβάρυνσης και αποτελεσματικότητας της διαδικασίας μάθησης. Η επιλογή αυτού του μικρού μεγέθους παρτίδας επιτρέπει την αποδοτική αξιοποίηση των διαθέσιμων πόρων της πλατφόρμας NVIDIA Jetson AGX Xavier, η οποία, παρότι εξειδικευμένη για εφαρμογές AI, διαθέτει περιορισμένους υπολογιστικούς πόρους σε σχέση με τα συμβατικά υπολογιστικά περιβάλλοντα γενικής χρήσης.

Για την εστιακή συνάρτηση απώλειας, οι υπερπαραμέτροι α και γ επιλέχθηκαν ως 0,5 και 2, αντίστοιχα, που αποτελούν τις προεπιλεγμένες τιμές της συγκεκριμένης συνάρτησης [226]. Η παράμετρος α ρυθμίζει τη σχετική βαρύτητα των θετικών και αρνητικών δειγμάτων κατά τον υπολογισμό της απώλειας, με την τιμή 0,5 να υποδηλώνει ισόρροπη στάθμιση. Αντίστοιχα, η παράμετρος γ λειτουργεί ως ενισχυτικός συντελεστής εστίασης για τη στόχευση των δυσκολότερα προς ταξινόμηση δειγμάτων, ενισχύοντας τη συμβολή τους στη διαδικασία εκμάθησης και περιορίζοντας την επιρροή των δειγμάτων που ταξινομούνται με υψηλή βεβαιότητα [226].

Ο Πίνακας 5.6 συνοψίζει τις τιμές των υπερπαραμέτρων που εφαρμόστηκαν κατά την εκπαίδευση των μοντέλων.

Πίνακας 5.6: Καθορισμός των υπερπαραμέτρων που χρησιμοποιήθηκαν για την εκπαίδευση των CNNs.

Hyperparameter	Value
Optimizer	Adam
Batch size	16
Loss function	CE, FL
Activation function	ReLu, Softmax
Epochs	30
Learning rate	0,0001
α	0,5
γ	2

5.6.3.1 Χρόνος Εκπαίδευσης και Ελέγχου

Η αποδοτικότητα ενός μοντέλου ΑΙ δεν εξαρτάται μόνο από τη διαγνωστική του ακρίβεια, αλλά και από τον απαιτούμενο χρόνο για την εκπαίδευση και τον έλεγχο του, ιδίως όταν η υλοποίηση πραγματοποιείται σε συστήματα με περιορισμένους υπολογιστικούς πόρους. Στον Πίνακα 5.7 παρουσιάζονται οι μέσοι συνολικοί χρόνοι εκπαίδευσης για 30 εποχές, καθώς και οι χρόνοι ελέγχου ανά πτυχή, για κάθε ένα από τα συγκρινόμενα μοντέλα.

Από την ανάλυση των αποτελεσμάτων προκύπτει ότι το προτεινόμενο CNN χαμηλής πολυπλοκότητας παρουσιάζει σημαντικά μειωμένους χρόνους τόσο ως προς την εκπαίδευση όσο και ως προς τη διαδικασία ελέγχου, σε σχέση με τα υπόλοιπα εξεταζόμενα προεκπαιδευμένα μοντέλα. Ειδικότερα, η εκπαίδευση του μοντέλου ολοκληρώνεται κατά μέσο όρο σε λιγότερο από 56 λεπτά σε περιβάλλον NVIDIA

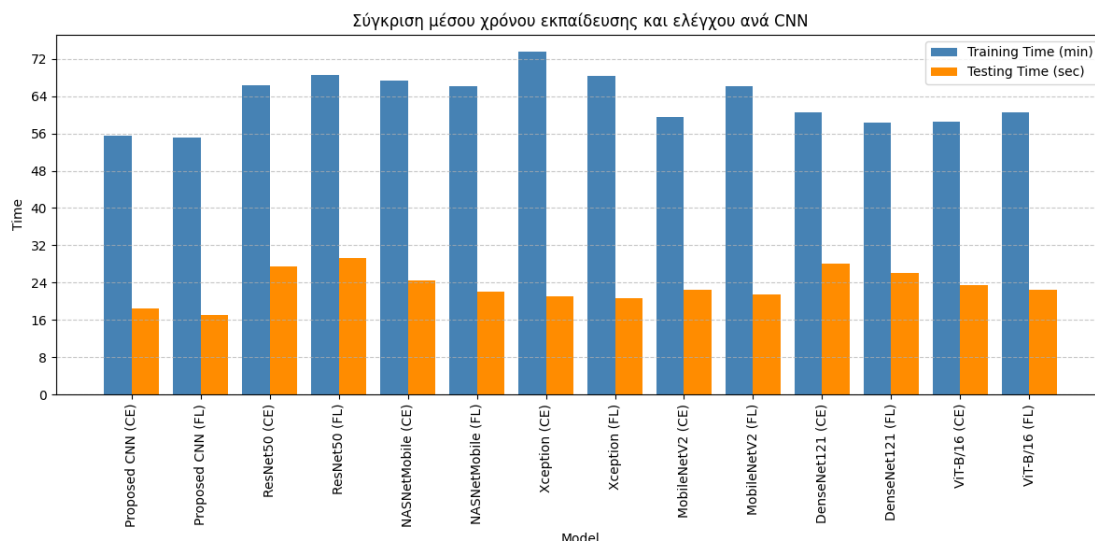
Jetson AGX Xavier, ενώ ο αντίστοιχος χρόνος ελέγχου δεν υπερβαίνει τα 18 δευτερόλεπτα ανά πτυχή. Η επίδοση αυτή αναδεικνύει τη σημαντική υπολογιστική αποδοτικότητα του μοντέλου, γεγονός που το καθιστά ιδιαίτερα κατάλληλο για εφαρμογές DL σε ενσωματωμένες πλατφόρμες, όπως η NVIDIA Jetson AGX Xavier, όπου η χαμηλή καθυστέρηση και ο χρόνος απόκρισης είναι κρίσιμα χαρακτηριστικά.

Αντιθέτως, τα προεκπαιδευμένα CNNs, όπως το ResNet50, το DenseNet121 και το Xception, παρουσιάζουν αυξημένο υπολογιστικό κόστος, με μέσους χρόνους εκπαίδευσης που υπερβαίνουν τα 60 λεπτά και με διαδικασία ελέγχου που διαρκούν έως και 29 δευτερόλεπτα ανά πτυχή. Η διαφορά αυτή αντανακλά τόσο στη δομική πολυπλοκότητα των εν λόγω αρχιτεκτονικών όσο και τις υψηλότερες απαιτήσεις σε υπολογιστικούς πόρους.

Πίνακας 5.7: Αναλυτική παρουσίαση της μέσης διάρκειας εκπαίδευσης (σε λεπτά) και χρόνου ελέγχου (σε δευτερόλεπτα) για καθένα από τα εξεταζόμενα CNNs.

Model	Average training time for 30 epoch (min)	Average total testing time (sec)
Proposed lightweight CNN (CE)	55,56	18,42
Proposed lightweight CNN (FL)	55,15	17,14
ResNet50 (CE)	66,34	27,48
ResNet50 (FL)	68,57	29,36
NASNetMobile (CE)	67,27	24,38
NASNetMobile (FL)	66,08	22,12
Xception (CE)	73,52	21,08
Xception (FL)	68,34	20,59
MobileNetV2 (CE)	59,48	22,53
MobileNetV2 (FL)	66,24	21,48
DenseNet121 (CE)	60,54	28,04
DenseNet121 (FL)	58,37	26,14
ViT-B/16 (CE)	58,54	23,43
ViT-B/16 (FL)	60,47	22,52

Στο Σχήμα 5.7 παρουσιάζεται συγκριτικά η μέση διάρκεια εκπαίδευσης (σε λεπτά) και διαδικασίας ελέγχου (σε δευτερόλεπτα) για κάθε ένα από τα μοντέλα που αξιολογήθηκαν. Η γραφική απεικόνιση καταδεικνύει τη σημαντικά μειωμένη υπολογιστική επιβάρυνση του προτεινόμενου μοντέλου χαμηλής πολυπλοκότητας, σε σύγκριση με τα προεκπαιδευμένα μοντέλα. Αυτό επιβεβαιώνει την καταλληλότητα του προτεινόμενου μοντέλου για ανάπτυξη και εφαρμογή σε συστήματα με περιορισμένους υπολογιστικούς πόρους.



Σχήμα 5.7: Συγκριτική γραφική απεικόνιση της μέσης συνολικής διάρκειας εκπαίδευσης (σε λεπτά) και του χρόνου ελέγχου (σε δευτερόλεπτα) για κάθε ένα από τα εξεταζόμενα CNNs.

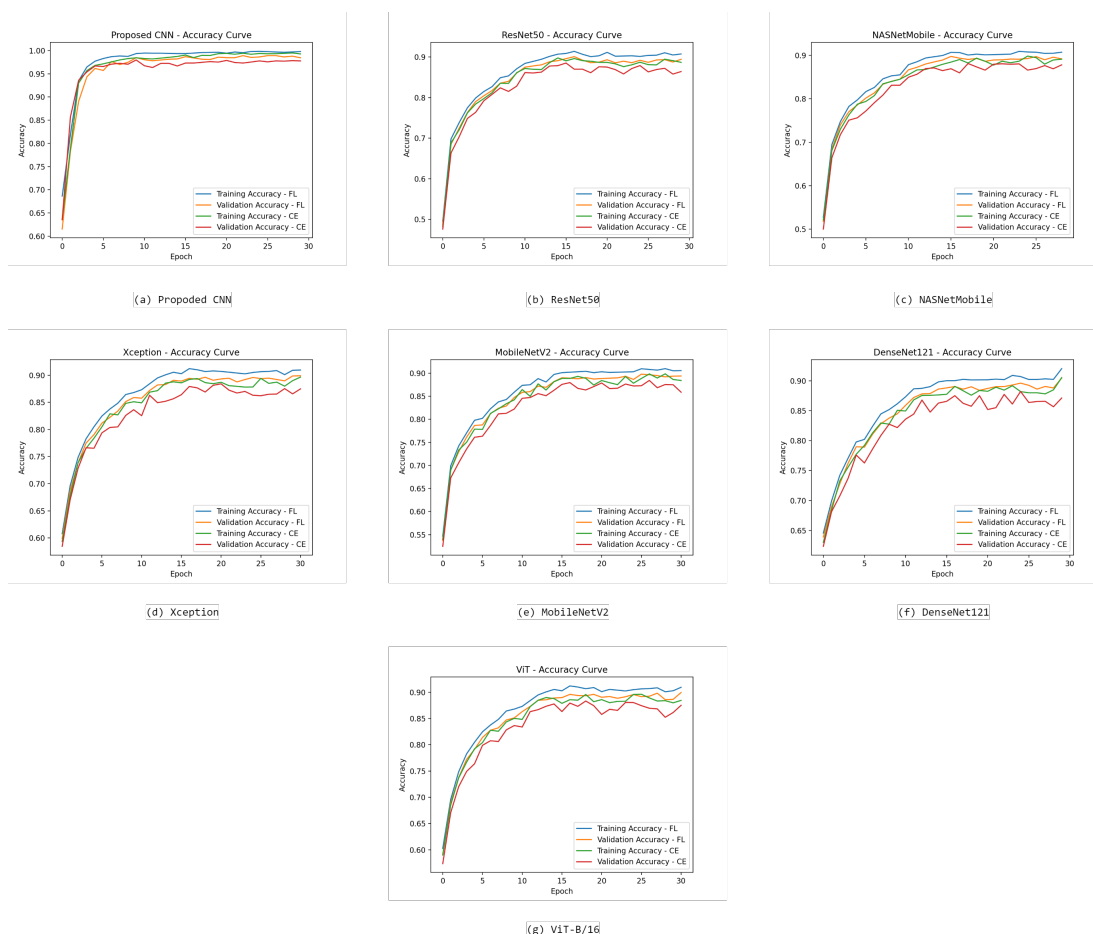
5.6.3.2 Καμπύλες Μέσης Ακρίβειας Εκπαίδευσης και Επικύρωσης

Η μέση καμπύλη ακρίβειας κατά τη φάση της εκπαίδευσης και της επικύρωσης, όπως προέκυψε από τη διαδικασία της πενταπλής διασταυρούμενης επικύρωσης των εξεταζόμενων μοντέλων με την εφαρμογή της συνάρτησης διασταυρούμενης εντροπίας και της εστιακής απώλειας, απεικονίζεται στο Σχήμα 5.8.

Παρατηρείται ότι σε όλα τα μοντέλα η σύγκλιση επιτυγχάνεται ταχύτερα όταν εφαρμόζεται η εστιακή απώλεια, σε αντίθεση με τη διασταυρούμενη εντροπία που οδηγεί σε βραδύτερη βελτίωση της ακρίβειας. Το αποτέλεσμα αυτό αποδίδεται στο γεγονός ότι η εστιακή απώλεια είναι καταλληλότερη για προβλήματα με έντονη κατηγορική ανισορροπία, χαρακτηριστικό που συναντάται συχνά σε ιατρικά σύνολα δεδομένων. Συγκεκριμένα, η εστιακή απώλεια ενισχύει τη συμβολή των υποεκπροσωπούμενων δειγμάτων, δηλαδή εκείνων που το μοντέλο τείνει να ταξινομεί λανθασμένα. Η ιδιότητα αυτή οδηγεί σε ταχύτερη μάθηση, με αποτέλεσμα την ταχύτερη σύγκλιση του μοντέλου.

Μεταξύ των μοντέλων που αξιολογήθηκαν, το προτεινόμενο CNN χαμηλής πολυπλοκότητας με τη χρήση της εστιακής απώλειας πέτυχε μέση ακρίβεια επικύρωσης της τάξης του 98,56%, υπερβαίνοντας τα υπόλοιπα προεκπαιδευμένα μοντέλα. Αντίστοιχα, τα ποσοστά επικύρωσης για τα μοντέλα ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT-B/16 ήταν 86,25%, 85,43%, 90,11%, 88,19%, 88,69% και 88,69%, αντίστοιχα.

Όταν χρησιμοποιήθηκε η συνάρτηση απώλειας διασταυρούμενης εντροπίας, το προτεινόμενο μοντέλο εξακολούθησε να παρουσιάζει υψηλή απόδοση, με ακρίβεια επικύρωσης 96,57%. Τα αντίστοιχα ποσοστά για τα υπόλοιπα μοντέλα ήταν 85,05% (ResNet50), 84,98% (NASNetMobile), 88,79% (Xception), 87,01% (MobileNetV2), 88,54% (DenseNet121) και 86,65% (ViT-B/16).



Σχήμα 5.8: Συγκριτική απεικόνιση της μέσης καμπύλης ακρίβειας εκπαίδευσης και επικύρωσης για τα εξεταζόμενα CNNs, μετά τη διαδικασία της πενταπλής διασταυρούμενης επικύρωσης, με εφαρμογή της συνάρτησης διασταυρούμενης εντροπίας και της εστιακής απώλειας.

5.6.3.3 Καμπύλες Μέσης Απώλειας Εκπαίδευσης και Επικύρωσης

Στο Σχήμα 5.9 παρουσιάζονται οι μέσες καμπύλες απώλειας εκπαίδευσης και επικύρωσης όπως προέκυψε από τη διαδικασία της πενταπλής διασταυρούμενης επικύρωσης των εξεταζόμενων μοντέλων, με την εφαρμογή της συνάρτησης διασταυρούμενης εντροπίας και της εστιακής συνάρτησης απώλειας.

Το προτεινόμενο CNN χαμηλής πολυπλοκότητας σε συνδυασμό με την εστιακή συνάρτηση απώλειας κατέγραψε ελάχιστη τιμή απώλειας επικύρωσης, ίση με 0,2262, υποδηλώνοντας την αυξημένη ικανότητά του να περιορίζει τα σφάλματα κατά τη διαδικασία μάθησης. Αντίθετα, τα υπόλοιπα προεκπαιδευμένα μοντέλα απέδωσαν υψηλότερες τιμές, με τις αντίστοιχες απώλειες να ανέρχονται σε 0,3026 (ResNet50), 0,3207 (NASNetMobile), 0,2868 (Xception), 0,3264 (MobileNetV2), 0,2908 (DenseNet121) και 0,3114 (ViT–B/16).

Όταν χρησιμοποιήθηκε η συνάρτηση απώλειας διασταυρούμενης εντροπίας, το προτεινόμενο μοντέλο εμφάνισε επίσης αξιοσημείωτη επίδοση με τιμή απώλειας επικύρωσης 0,2457. Οι τιμές απώλειας των υπολοίπων μοντέλων κυμάνθηκαν υψηλότερα, 0,3471 (ResNet50), 0,3238 (NASNetMobile), 0,3282 (Xception), 0,3646 (MobileNetV2), 0,3402 (DenseNet121) και 0,3785 (ViT–B/16).

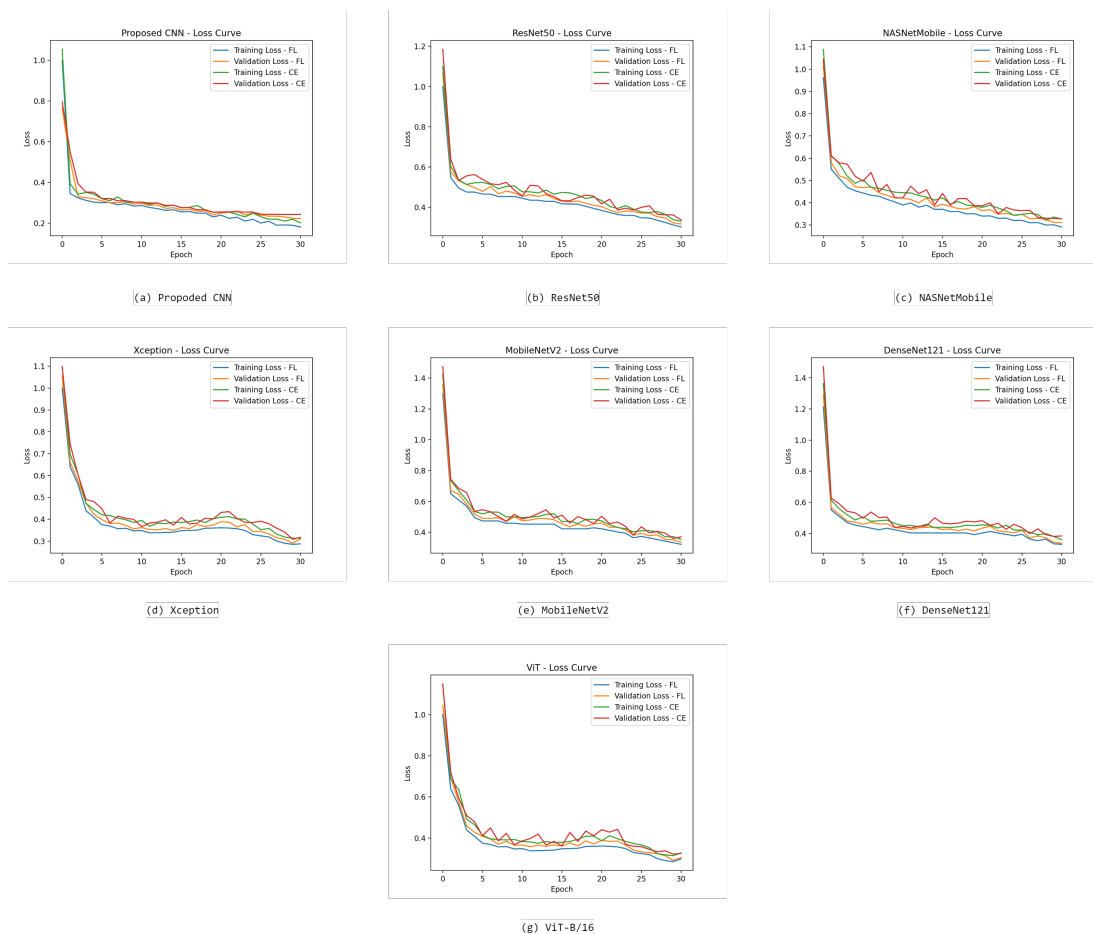
Τα αποτελέσματα αυτά υποδηλώνουν ότι το προτεινόμενο μοντέλο χαμηλής πολυπλοκότητας, ιδίως όταν συνδυάζεται με την εστιακή συνάρτηση απώλειας, επιτυγχάνει υψηλότερο βαθμό βελτιστοποίησης και γενίκευσης σε σύγκριση με τα πιο σύνθετα προεκπαιδευμένα μοντέλα. Διακρίνεται για την ανώτερη απόδοσή του τόσο ως προς την ακρίβεια επικύρωσης όσο και ως προς τη μείωση της απώλειας, καθιστώντας το ιδιαίτερα κατάλληλο για περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους.

5.6.4 Φάση Ελέγχου

5.6.4.1 Αξιολόγηση με Μήτρες Σύγχυσης

Η αξιολόγηση του προτεινόμενου CNN χαμηλής πολυπλοκότητας, καθώς και των αντίστοιχων προεκπαιδευμένων μοντέλων, πραγματοποιήθηκε με χρήση των μητρών σύγχυσης μετά την ολοκλήρωση της διαδικασίας πενταπλής διασταυρούμενης επικύρωσης. Οι αντίστοιχες απεικονίσεις παρουσιάζονται στα Σχήματα 5.10–5.11.

Στη φάση ελέγχου ελήφθησαν υπόψη συνολικά 5.825 δείγματα, με στόχο την κατηγοριοποίησή τους σε επτά διακριτές κατηγορίες παθολογικών και φυσιολογικών καταστάσεων. Το προτεινόμενο συνελκτικό μοντέλο σε συνδυασμό με την εστιακή συνάρτηση απώλειας εμφάνισε επίδοση, με μόλις 84 λανθασμένες ταξινομήσεις, γεγονός που υποδηλώνει τον υψηλό βαθμό ακρίβειας στην αναγνώριση των επτά κατηγοριών.



Σχήμα 5.9: Συγκριτική απεικόνιση της μέσης καμπύλης απώλειας εκπαίδευσης και επικύρωσης για τα εξεταζόμενα CNNs, μετά τη διαδικασία της πενταπλής διασταυρούμενης επικύρωσης, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.

Αντιθέτως, τα προεκπαιδευμένα μοντέλα παρουσίασαν αυξημένο πλήθος εσφαλμένων ταξινομήσεων. Συγκεκριμένα το δίκτυο NASNetMobile παρουσίασε 847 λανθασμένες προβλέψεις, το ResNet50 κατέγραψε 797 περιπτώσεις εσφαλμένης ταξινόμησης, το DenseNet121 εμφάνισε 598 λανθασμένες προβλέψεις, το MobileNetV2 σημείωσε 684 σφάλματα, το μοντέλο ViT-B/16 ταξινόμησε εσφαλμένα 655 δείγματα και το Xception παρουσίασε 573 λανθασμένες ταξινομήσεις.

Με την εφαρμογή της συνάρτησης διασταυρούμενης εντροπίας, το προτεινόμενο συνελκτικό δίκτυο κατέγραψε 201 εσφαλμένες ταξινομήσεις, διατηρώντας υψηλό επίπεδο απόδοσης. Αντίθετα, τα προεκπαιδευμένα μοντέλα παρουσίασαν μεγαλύτερο αριθμό σφαλμάτων: ResNet50: 869, NASNetMobile: 874, Xception: 650, MobileNetV2: 755, DenseNet121: 665 και ViT-B/16: 774.

Τα ευρήματα υποδηλώνουν ότι το προτεινόμενο συνελκτικό μοντέλο, ιδιαίτερα όταν

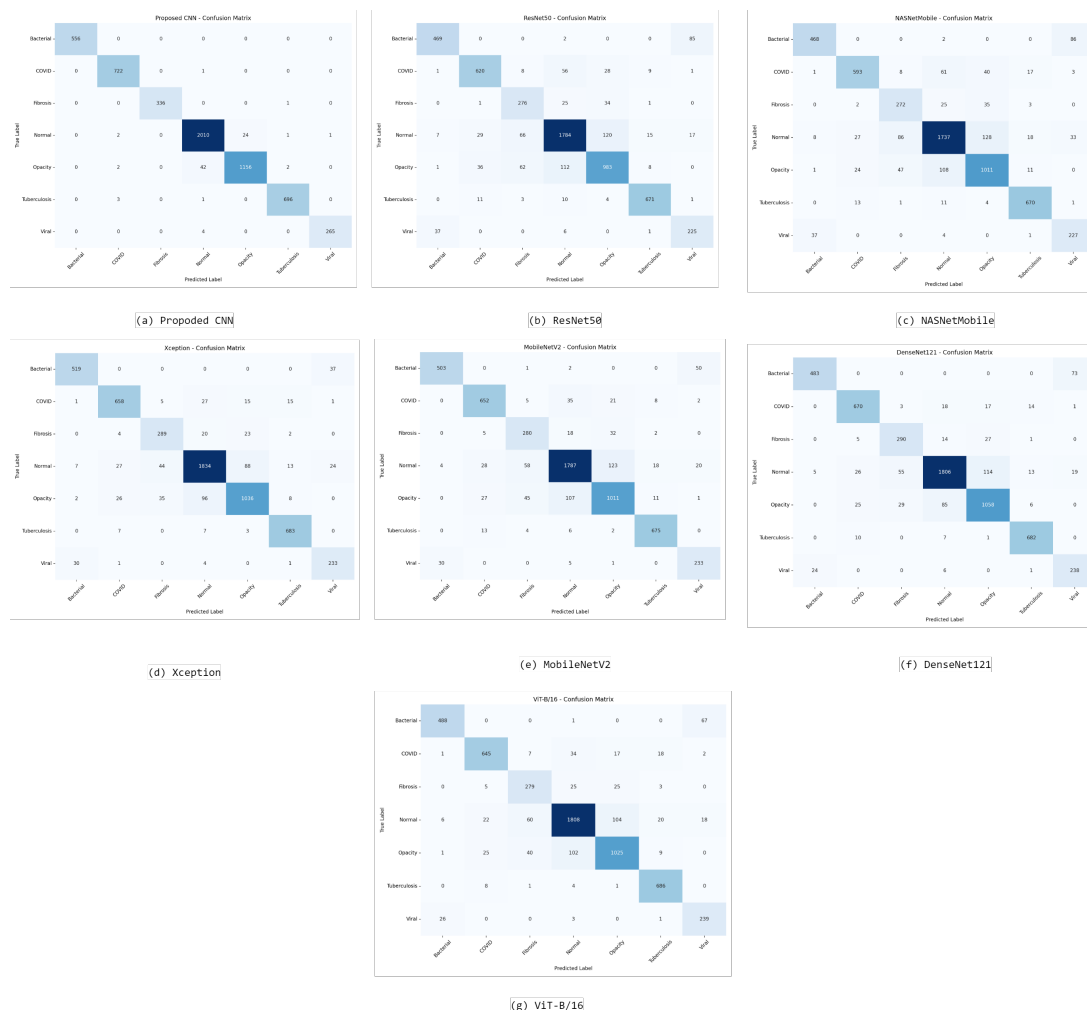
συνδυάζεται με την εστιακή συνάρτηση απώλειας, εμφανίζει σημαντικά αυξημένη ικανότητα γενίκευσης και ακρίβεια ταξινόμησης σε άγνωστα δεδομένα, καθιστώντας το ιδιαίτερα κατάλληλο για εφαρμογές ιατρικής διάγνωσης με πολλαπλές κλάσεις και κατηγορική ανισορροπία. Στο Σχήμα 5.12 απεικονίζεται συγκριτικά ο συνολικός αριθμός των λανθασμένων ταξινομήσεων για κάθε CNN που αξιολογήθηκε, με εφαρμογή των συναρτήσεων διασταυρούμενης εντροπίας και εστιακής απώλειας, κατά τη διαδικασία της πενταπλής διασταυρούμενης επικύρωσης.



Σχήμα 5.10: Μήτρες σύγχυσης για κάθε εξεταζόμενο CNN μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης διασταυρούμενης εντροπίας.

5.6.4.2 Διακριτική Ικανότητα μέσω Καμπυλών ROC

Οι καμπύλες ROC για τα εξεταζόμενα CNNs, όπως προέκυψαν μετά από την πενταπλή διασταυρούμενη επικύρωση για τη διάκριση μεταξύ επτά διαφορετικών

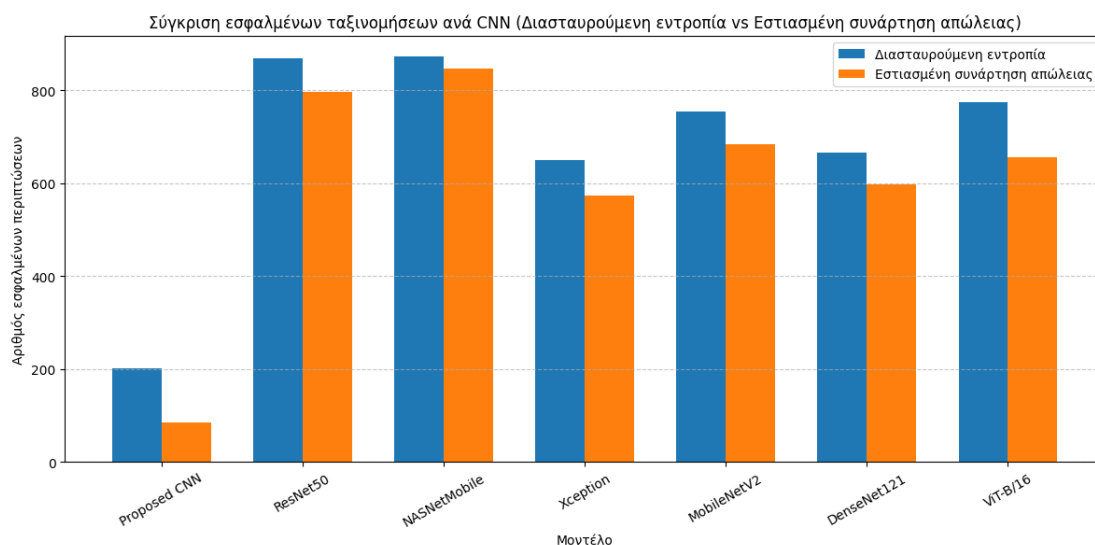


Σχήμα 5.11: Μήτρες σύγχυσης για κάθε εξεταζόμενο CNN μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της εστιακής συνάρτησης απώλειας.

κατηγοριών, παρουσιάζονται στα Σχήματα 5.13–5.14. Οι καμπύλες ROC απεικονίζουν τη συσχέτιση μεταξύ του ποσοστού των αληθώς θετικών και των ψευδώς θετικών προβλέψεων ανά κατηγορία, επιτρέποντας την εκτίμηση της διακριτικής ικανότητας του κάθε μοντέλου.

Στο Σχήμα 5.15 παρουσιάζονται οι τιμές AUC για κάθε εξεταζόμενο συνελκτικό μοντέλο, όπως προέκυψαν μετά την εφαρμογή της πενταπλής διασταυρούμενης επικύρωσης, τόσο με τη συνάρτηση απώλειας διασταυρούμενης εντροπίας όσο και με την εστιακή συνάρτηση απώλειας.

Το προτεινόμενο CNN χαμηλής πολυπλοκότητας σε συνδυασμό με την εστιακή απώλεια παρουσίασε διακριτική ικανότητα, καταγράφοντας AUC ίσο με 99,30%, τιμή που υποδηλώνει την ικανότητά του να διαχωρίζει αποτελεσματικά τις επτά κατηγορίες παθολογιών. Αντίστοιχα, τα προεκπαιδευμένα μοντέλα ResNet50,



Σχήμα 5.12: Συγκριτική απεικόνιση του αριθμού λανθασμένων ταξινομήσεων για καθένα από τα επτά CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.

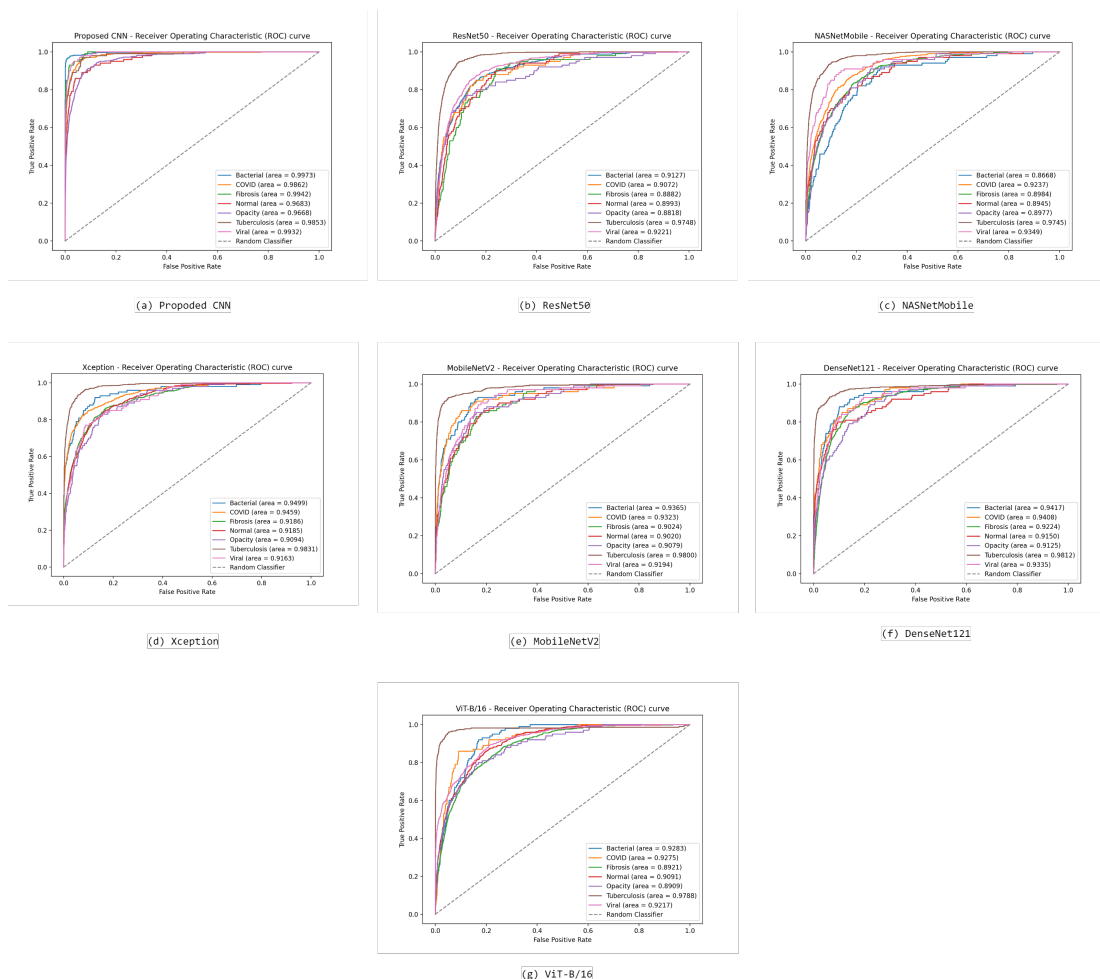
NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT–B/16 απέδωσαν χαμηλότερες τιμές AUC, ήτοι 91,68%, 91,30%, 94,15%, 93,13%, 93,96% και 93,32%, αντίστοιχα.

Αξιοσημείωτη είναι και η επίδοση του προτεινόμενου μοντέλου με την εφαρμογή της διασταυρούμενης εντροπίας, με AUC ίσο με 98,44%, τιμή που εξακολουθεί να υπερτερεί έναντι των υπολοίπων προεκπαιδευμένων μοντέλων. Οι αντίστοιχες τιμές για τα προεκπαιδευμένα μοντέλα ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT–B/16 ανέρχονται σε 91,23%, 91,29%, 93,45%, 92,57%, 93,53% και 92,11%, αντίστοιχα.

5.6.4.3 Σύγκριση Μετρικών Ποιότητας

Στον Πίνακα 5.8 και στο Σχήμα 5.16 παρουσιάζονται τα αποτελέσματα των μετρικών ποιότητας για κάθε μοντέλο, όπως προέκυψαν μετά τη διαδικασία της πενταπλής διασταυρούμενης επικύρωσης. Τα αποτελέσματα αυτά αποτυπώνουν την επίδοση των μοντέλων για την αναγνώριση των επτά κατηγοριών πνευμονοπαθειών.

Το προτεινόμενο CNN χαμηλής πολυπλοκότητας, σύμφωνα με την απώλεια της διασταυρούμενης εντροπίας, σημειώνει μέση συνολική ακρίβεια 96,57%, αποδεικνύοντας την ικανότητά του να προβλέπει με συνέπεια τις σωστές κατηγορίες

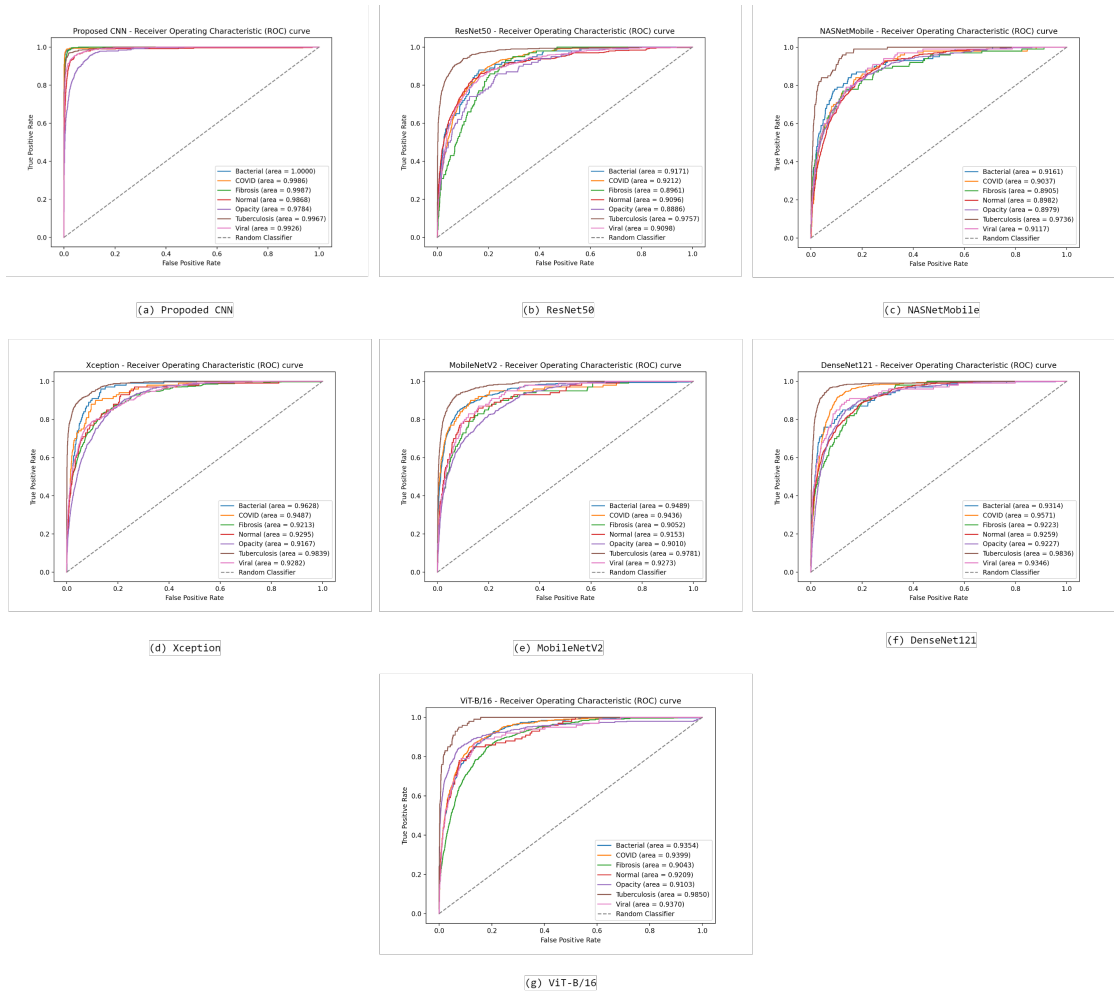


Σχήμα 5.13: Καμπύλες ROC για τα επτά εξεταζόμενα CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης απώλειας διασταυρούμενης εντροπίας.

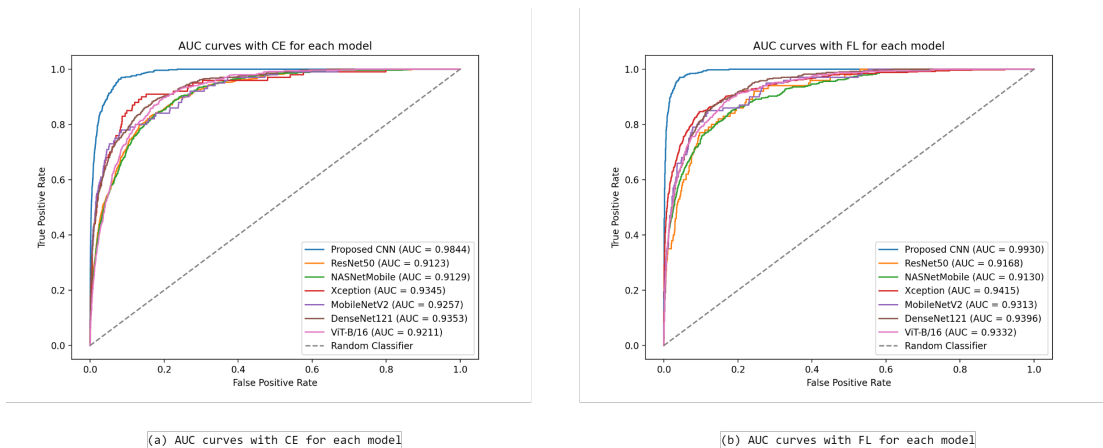
στο σύνολο των επτά κλάσεων. Παράλληλα, καταγράφει μέση βαθμολογία F1 ίση με 97,23%, υποδηλώνοντας ισορροπημένη απόδοση ως προς την ακρίβεια των θετικών προβλέψεων και την ανάκληση, στοιχείο καθοριστικό για την αποτελεσματική ταξινόμηση.

Ωστόσο, με την εφαρμογή της εστιακής απώλειας, η απόδοση του μοντέλου βελτιώνεται περαιτέρω, φθάνοντας σε μέση συνολική ακρίβεια 98,56% και μέση βαθμολογία F1 ίση με 99,01%. Η εν λόγω βελτίωση αναδεικνύει την αποτελεσματικότητα της εστιακής συνάρτησης απώλειας στη βελτιστοποίηση της προβλεπτικής ικανότητας του μοντέλου, ιδίως σε περιπτώσεις δύσκολων ή υποεκπροσωπούμενων δειγμάτων στο σύνολο των δεδομένων.

Επιπροσθέτως, η συνάρτηση της εστιακής απώλειας έναντι της διασταυρούμενης εντροπίας σε όλα τα εξεταζόμενα μοντέλα καταδεικνύει τη χρησιμότητά της



Σχήμα 5.14: Καμπύλες ROC για τα επτά εξεταζόμενα CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της εστιακής συνάρτησης απώλειας.

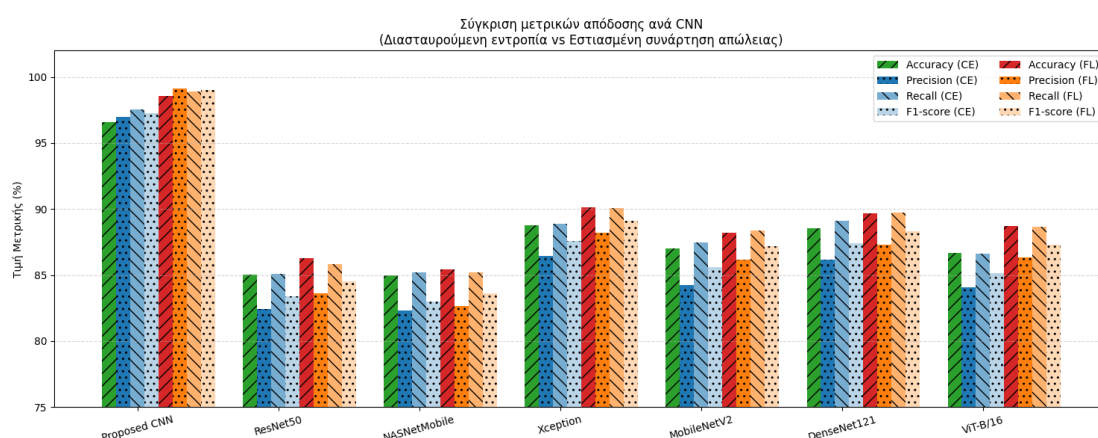


Σχήμα 5.15: Συγκριτική απεικόνιση των τιμών AUC για κάθε μοντέλο, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.

στην αντιμετώπιση προβλημάτων κατηγορικής ανισορροπίας, δίνοντας έμφαση σε δυσδιάκριτες περιπτώσεις. Η εν λόγω ιδιότητα καθίσταται ιδιαίτερος κρίσιμη στο πεδίο της ιατρικής απεικόνισης, όπου η παρουσία κατηγοριών με περιορισμένο αριθμό δειγμάτων επιβάλλει την ανάγκη για αξιόπιστη ανίχνευση όλων των παθολογικών κλάσεων, συμπεριλαμβανομένων και των λιγότερο εκπροσωπημένων.

Πίνακας 5.8: Μέσοι όροι των μετρικών ποιότητας, για τα εξεταζόμενα CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.

Model	Average accuracy (%)	Average precision (%)	Average recall (%)	Average F1-score (%)
Proposed lightweight CNN (CE)	96,57	96,97	97,53	97,23
Proposed lightweight CNN (FL)	98,56	99,10	98,91	99,01
ResNet50 (CE)	85,05	82,44	85,11	83,40
ResNet50 (FL)	86,25	83,64	85,82	84,51
NASNetMobile (CE)	84,98	82,34	85,22	83,01
NASNetMobile (FL)	85,43	82,67	85,20	83,63
Xception (CE)	88,79	86,47	88,89	87,58
Xception (FL)	90,11	88,21	90,08	89,10
MobileNetV2 (CE)	87,01	84,25	87,44	85,60
MobileNetV2 (FL)	88,19	86,18	88,37	87,18
DenseNet121(CE)	88,54	86,18	89,08	87,41
DenseNet121 (FL)	89,69	87,29	89,74	88,34
ViT-B/16 (CE)	86,65	84,05	86,61	85,13
ViT-B/16 (FL)	88,69	86,31	88,66	87,32



Σχήμα 5.16: Συγκριτική απεικόνιση των μέσων όρων των μετρικών ποιότητας, για τα εξεταζόμενα CNNs, μετά την πενταπλή διασταυρούμενη επικύρωση, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.

Οι Πίνακες 5.9–5.14 παρουσιάζουν συγκριτικά τους μέσους όρους των μετρικών ποιότητας – ακρίβειας των θετικών προβλέψεων, ανάκλησης και βαθμολογίας F1 – για τα εξεταζόμενα CNNs, μετά την εφαρμογή της πενταπλής διασταυρούμενης επικύρωσης και με χρήση των συναρτήσεων διασταυρούμενης εντροπίας και εστιακής απώλειας. Το προτεινόμενο CNN εμφανίζει σταθερά υψηλές επιδόσεις σε όλες τις κατηγορίες, ιδιαιτέρως υπό τη χρήση της συνάρτησης εστιακής απώλειας, η οποία συμβάλει στην καλύτερη διαχείριση των περιπτώσεων αβεβαιότητας, προσφέροντας υψηλότερες τιμές και, κατ' επέκταση, βελτιωμένη απόδοση ταξινόμησης σε όλες τις κατηγορίες.

Πίνακας 5.9: Μέσοι όροι ακρίβειας των θετικών προβλέψεων για τα εξεταζόμενα CNNs, με συνάρτηση απώλειας διασταυρούμενης εντροπίας και πενταπλή διασταυρούμενη επικύρωση.

Category	Proposed lightweight CNN	ResNet50	NASNetMobile	Xception	MobileNetV2	DenseNet121	ViT–B/16
Bacterial	99,94	92,94	94,24	92,12	92,25	93,55	91,94
COVID	95,54	87,78	90,13	89,82	89,40	91,39	88,66
Fibrosis	98,66	62,53	68,22	73,90	68,43	71,37	68,48
Normal	97,13	88,46	89,51	91,58	91,37	92,20	89,99
Opacity	92,82	83,45	82,11	87,34	84,04	85,22	84,50
Tuberculosis	99,56	93,78	93,38	94,40	94,10	95,71	93,44
Viral	95,05	68,12	58,75	76,16	70,18	73,85	71,34

Πίνακας 5.10: Μέσοι όροι ακρίβειας των θετικών προβλέψεων για τα εξεταζόμενα CNNs, με χρήση της εστιακής συνάρτησης απώλειας και πενταπλή διασταυρούμενη επικύρωση.

Category	Proposed lightweight CNN	ResNet50	NASNetMobile	Xception	MobileNetV2	DenseNet121	ViT–B/16
Bacterial	100,00	91,03	90,84	92,82	93,64	94,31	93,46
COVID	99,01	88,64	89,70	90,75	89,52	90,66	91,24
Fibrosis	100,00	68,69	67,95	79,21	73,29	78,68	74,10
Normal	97,66	89,28	89,03	92,17	91,03	93,25	91,27
Opacity	97,98	84,01	82,94	88,82	84,95	86,92	87,44
Tuberculosis	99,42	95,12	93,11	94,54	94,48	95,06	93,13
Viral	99,63	68,67	65,16	79,19	76,38	72,16	73,56

Πίνακας 5.11: Μέσοι όροι ανάκλησης για τα εξεταζόμενα CNNs, με συνάρτηση απώλειας διασταυρούμενης εντροπίας και πενταπλή διασταυρούμενη επικύρωση.

Category	Proposed lightweight CNN	ResNet50	NASNetMobile	Xception	MobileNetV2	DenseNet121	ViT–B/16
Bacterial	99,46	83,21	73,83	90,79	88,09	88,99	86,46
COVID	97,86	83,02	86,02	90,58	87,87	89,30	87,02
Fibrosis	98,93	80,97	82,31	85,79	83,11	86,86	80,97
Normal	95,17	85,84	84,16	88,01	84,66	86,93	86,98
Opacity	95,28	80,55	84,35	85,10	85,84	86,42	82,12
Tuberculosis	97,11	95,81	95,81	97,40	96,82	96,82	96,68
Viral	98,90	86,40	90,07	84,56	85,66	88,24	86,03

Πίνακας 5.12: Μέσοι όροι ανάκλησης για τα εξεταζόμενα CNNs, με χρήση της εστιακής συνάρτησης απώλειας και πενταπλή διασταυρούμενη επικύρωση.

Category	Proposed lightweight CNN	ResNet50	NASNetMobile	Xception	MobileNetV2	DenseNet121	ViT-B/16
Bacterial	100,00	84,30	84,12	93,32	90,43	86,82	87,73
COVID	99,86	85,73	82,03	91,01	90,16	92,72	89,16
Fibrosis	99,73	81,77	80,70	85,79	83,14	86,06	82,84
Normal	98,62	87,52	85,25	89,99	87,67	88,60	88,70
Opacity	96,19	81,79	84,11	86,18	84,11	88,00	85,26
Tuberculosis	99,42	95,81	95,66	97,54	96,39	97,40	97,98
Viral	98,53	83,82	84,56	86,76	86,76	88,60	88,97

Πίνακας 5.13: Μέσοι όροι βαθμολογίας F1 για τα εξεταζόμενα CNNs, με συνάρτηση απώλειας διασταυρούμενης εντροπίας και πενταπλή διασταυρούμενη επικύρωση.

Category	Proposed lightweight CNN	ResNet50	NASNetMobile	Xception	MobileNetV2	DenseNet121	ViT-B/16
Bacterial	99,73	87,81	82,79	91,45	90,12	91,21	89,12
COVID	96,69	85,34	88,03	90,20	88,63	90,33	87,83
Fibrosis	98,80	70,56	74,61	79,40	75,06	78,36	74,20
Normal	96,14	87,13	86,75	89,76	87,89	89,49	88,46
Opacity	94,04	81,97	83,22	86,21	84,93	85,82	83,29
Tuberculosis	98,32	94,78	94,58	95,87	95,44	96,26	95,03
Viral	96,94	76,18	71,12	80,14	77,15	80,40	78,00

Πίνακας 5.14: Μέσοι όροι βαθμολογίας F1 για τα εξεταζόμενα CNNs, με χρήση της εστιακής συνάρτησης απώλειας και πενταπλή διασταυρούμενη επικύρωση.

Category	Proposed lightweight CNN	ResNet50	NASNetMobile	Xception	MobileNetV2	DenseNet121	ViT-B/16
Bacterial	100,00	87,54	87,35	93,07	92,01	90,41	90,50
COVID	99,43	87,16	85,69	90,88	89,84	91,68	90,19
Fibrosis	99,87	74,66	73,77	82,37	77,89	82,20	78,23
Normal	98,13	88,39	87,10	91,06	89,32	90,87	89,97
Opacity	97,08	82,89	83,52	87,48	84,53	87,45	86,34
Tuberculosis	99,42	95,46	94,37	96,02	95,42	96,22	95,49
Viral	99,08	75,50	73,60	82,81	81,24	79,54	80,53

5.6.4.4 Ανάλυση Χρόνου Εκτέλεσης Εξαγωγής Συμπερασμάτων

Δεδομένης της υψίστης σημασίας του χρόνου εκτέλεσης εξαγωγής συμπερασμάτων σε εφαρμογές πραγματικών συνθηκών, στις οποίες η άμεση απόκριση αποτελεί ουσιώδες κριτήριο λειτουργικότητας, οι μετρήσεις που παρατίθενται στον Πίνακα 5.15 αποτελούν βασικό δείκτη για την αξιολόγηση της υπολογιστικής αποδοτικότητας των εξεταζόμενων μοντέλων (βλ. Ενότητα 3.4.2). Ειδικότερα, καταδεικνύεται η ικανότητά τους να ανταποκρίνονται αποτελεσματικά σε περιβάλλοντα όπου απαιτείται ταχεία ανάλυση και λήψη αποφάσεων.

Η επίδοση του προτεινόμενου CNN χαμηλής πολυπλοκότητας παρουσιάζει σημαντικά μικρότερους χρόνους εκτέλεσης σε σύγκριση με τα προεκπαιδευμένα μοντέλα. Συγκεκριμένα, κατά την εκτέλεση στη CPU, καταγράφεται μέσος χρόνος πρόβλεψης μόλις 0,0322 ms, γεγονός που υποδηλώνει την ικανότητα του

μοντέλου να επεξεργάζεται την είσοδο και να αποδίδει έξοδο με εξαιρετικά υψηλή ταχύτητα. Αντιστοίχως, όταν εκτελείται σε περιβάλλον GPU, η απόδοσή του ενισχύεται περαιτέρω, με χρόνο πρόβλεψης 0,0165 ms. Επιπροσθέτως, οι τιμές καρτέ ανά δευτερόλεπτο που καταγράφει το προτεινόμενο μοντέλο, σε αμφότερες τις αρχιτεκτονικές υποδομές, αποτυπώνουν την ταχεία διαχείριση και επεξεργασία δεδομένων. Συγκριτικά, τα μοντέλα MobileNetV2, DenseNet121, NASNetMobile, ViT-B/16, ResNet50 και Xception εμφανίζουν υψηλότερους χρόνους εξαγωγής συμπερασμάτων, ιδίως όταν εκτελούνται σε CPU, γεγονός που περιορίζει τη χρηστικότητα τους σε περιβάλλοντα με περιορισμένη υπολογιστική ισχύ.

Ο Πίνακας 5.15 και τα Σχήματα 5.17–5.18 καταδεικνύουν ότι το προτεινόμενο μοντέλο παρουσιάζει αξιοσημείωτη αποδοτικότητα σε συστήματα με περιορισμένες δυνατότητες παραλληλοποίησης, όπως οι CPUs. Η αποδοτικότητα αυτή αποδίδεται στην αρχιτεκτονική σχεδίαση του μοντέλου, η οποία χαρακτηρίζεται από περιορισμένο αριθμό επιπέδων και παραμέτρων. Το γεγονός αυτό συνεπάγεται χαμηλή υπολογιστική πολυπλοκότητα, διατηρώντας παράλληλα υψηλά επίπεδα ακρίβειας στην κατηγοριοποίηση.

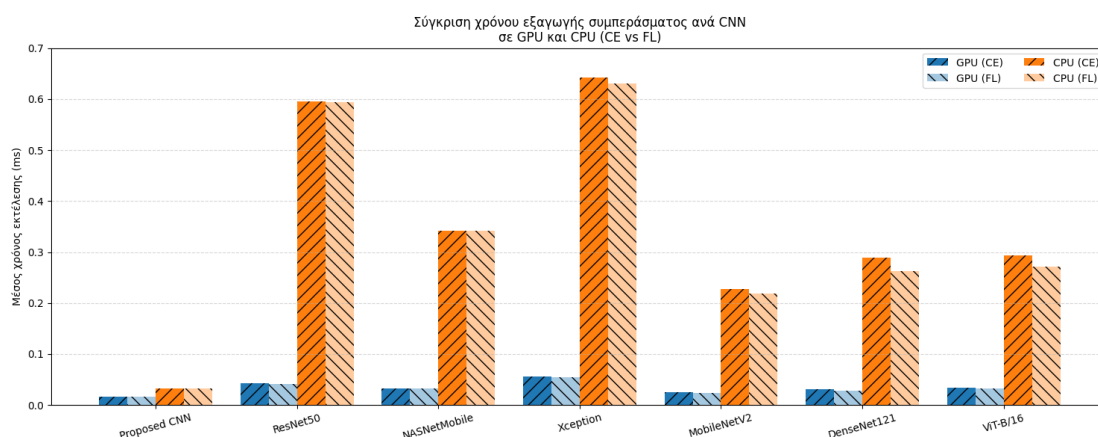
Αντιθέτως, τα συγκριτικά προεκπαιδευμένα μοντέλα, αν και αποδοτικά σε GPU λόγω της εγγενούς παραλληλότητας που διαθέτουν, δεν διατηρούν αντίστοιχη απόδοση σε CPU, εξαιτίας της αυξημένης υπολογιστικής επιβάρυνσης και της έλλειψης επαρκούς παραλληλισμού. Τα αποτελέσματα αυτά υπογραμμίζουν την αποτελεσματικότητα του σχεδιασμού του προτεινόμενου μοντέλου, το οποίο συνδυάζει απλοποιημένη αρχιτεκτονική με υψηλή λειτουργική απόδοση τόσο σε GPU όσο και σε CPU περιβάλλοντα, ενισχύοντας τη δυναμική του εφαρμογή σε συστήματα πραγματικού χρόνου με περιορισμένους υπολογιστικούς πόρους.

5.7 Συζήτηση Ευρημάτων

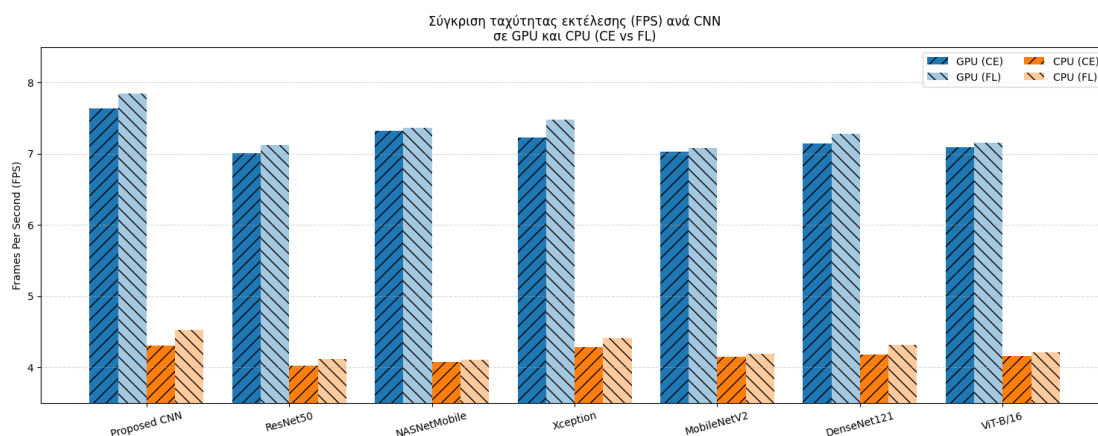
Στη σύγχρονη ιατρική πρακτική, παρατηρείται αυξανόμενο ενδιαφέρον για την προσέγγιση στο σημείο φροντίδας. Η ραγδαία εξέλιξη των φορητών ακτινογραφικών μηχανημάτων και των απεικονιστικών συσκευών χειρός έχει επιφέρει σημαντική μεταβολή σε αυτό το πεδίο. Οι εν λόγω τεχνολογίες αποδεικνύονται ιδιαίτερες επωφελείς για την επείγουσα διαγνωστική προσπάθεια, ιδίως σε περιπτώσεις ασθενών με σοβαρή κλινική εικόνα ή περιορισμένη κινητικότητα, για τους

Πίνακας 5.15: Σύγκριση του μέσου χρόνου εκτέλεσης εξαγωγής συμπερασμάτων και των καρτέ ανά δευτερόλεπτο για κάθε CNN, σε GPU και CPU, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.

Model	Average inference time on GPU (in MS)	FPS on GPU	Average inference time on CPU (in MS)	FPS on CPU
Proposed lightweight CNN (CE)	0,0169	7,64	0,0328	4,31
Proposed lightweight CNN (FL)	0,0165	7,85	0,0322	4,53
ResNet50 (CE)	0,0424	7,01	0,5949	4,03
ResNet50 (FL)	0,0414	7,12	0,5935	4,12
NASNetMobile (CE)	0,0327	7,32	0,3424	4,08
NASNetMobile (FL)	0,0321	7,37	0,3421	4,11
Xception (CE)	0,0564	7,23	0,6424	4,29
Xception (FL)	0,0542	7,48	0,6302	4,41
MobileNetV2 (CE)	0,0249	7,03	0,2278	4,15
MobileNetV2 (FL)	0,0236	7,08	0,2184	4,19
DenseNet121 (CE)	0,0306	7,15	0,2890	4,18
DenseNet121 (FL)	0,0282	7,28	0,2624	4,32
ViT-B/16 (CE)	0,0342	7,09	0,2932	4,16
ViT-B/16 (FL)	0,0326	7,16	0,2708	4,21



Σχήμα 5.17: Σύγκριση του μέσου χρόνου εκτέλεσης εξαγωγής συμπερασμάτων για κάθε CNN, τόσο σε GPU όσο και σε CPU, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.



Σχήμα 5.18: Σύγκριση των καρτέ ανά δευτερόλεπτο για τα εξεταζόμενα CNN, κατά την εκτέλεση εξαγωγής συμπεράσματος σε GPU και CPU, τόσο σε GPU όσο και σε CPU, με εφαρμογή της συνάρτησης (α) διασταυρούμενης εντροπίας και (β) εστιακής απώλειας.

οποίους η μετακίνηση μεταξύ νοσοκομειακών τμημάτων είναι συχνά δυσχερής και επιβαρυντική.

Η αμεσότητα που προσφέρουν τα απεικονιστικά μέσα στο σημείο φροντίδας ενισχύει την αποδοτικότητα των παρεχόμενων υπηρεσιών υγείας, διασφαλίζοντας την ταχεία διάγνωση και την άμεση ιατρική παρέμβαση σε κρίσιμες καταστάσεις. Η δυνατότητα ανάλυσης απεικονιστικών δεδομένων πλησίον του ασθενούς, χωρίς την ανάγκη διακομιδής, συμβάλλει στην ελαχιστοποίηση του χρόνου λήψης αποφάσεων και στη βελτίωση της έκβασης για τον ασθενή. Ως εκ τούτου, η απεικόνιση στο σημείο φροντίδας αναδεικνύεται ως βασικός πυλώνας για την προώθηση μιας υγειονομικής περίθαλψης που είναι άμεσα προσβάσιμη, αποτελεσματική και επικεντρωμένη στις εξατομικευμένες ανάγκες του ασθενούς, ενισχύοντας τη συνολική ποιότητα και βιωσιμότητα των υπηρεσιών υγείας.

Ένα σημαντικό κενό στη σχετική βιβλιογραφία είναι η απουσία αποδοτικών και υψηλής ακρίβειας μεθόδων για την κατηγοριοποίηση των ιατρικών εικόνων σε ενσωματωμένες συσκευές σημείου φροντίδας. Οι υπάρχουσες προσεγγίσεις βασίζονται κυρίως σε προεκπαιδευμένα μοντέλα που δεν είναι βελτιστοποιημένα για συγκεκριμένα ιατρικά καθήκοντα, ούτε είναι κατάλληλα για υλοποίηση σε συστήματα περιορισμένων πόρων. Επιπλέον, η κατηγορική ανισορροπία περιπλέκει περαιτέρω την ανάπτυξη αξιόπιστων μοντέλων. Για την κάλυψη αυτού του κενού, προτείνεται ένα CNN με βελτιστοποιημένη αρχιτεκτονική και χαμηλή υπολογιστική πολυπλοκότητα, σχεδιασμένο για το ενσωματωμένο σύστημα NVIDIA Jetson AGX Xavier. Το δίκτυο αυτό συνδυάζεται με την εστιακή συνάρτηση απώλειας, ώστε να αντιμετωπίζει αποτελεσματικά το πρόβλημα της κατηγορικής ανισορροπίας, στοιχείο που αποτελεί και τη βασική συνεισφορά της παρούσας εργασίας.

Το κίνητρο της παρούσας ερευνητικής προσπάθειας εδράζεται στην επιτακτική ανάγκη για την ανάπτυξη προηγμένων, αξιόπιστων και υπολογιστικά αποδοτικών τεχνικών κατηγοριοποίησης ιατρικών εικόνων, με έμφαση στην συνολική ακρίβεια, την ταχύτητα επεξεργασίας και τη δυνατότητα ενσωμάτωσης σε εφαρμογές στο σημείο φροντίδας. Η σχεδίαση και υλοποίηση υψηλής ακρίβειας, αλλά ταυτόχρονα υπολογιστικά αποδοτικών μοντέλων AI, δύναται να ενισχύσει το έργο των επαγγελματιών υγείας, διευκολύνοντας την άμεση και τεκμηριωμένη διάγνωση παθολογικών καταστάσεων σε συνθήκες κλινικής καθημερινότητας.

Παράλληλα, η παρούσα εργασία αποσκοπεί στη συμβολή της στη διαρκώς εξελισσόμενη ερευνητική περιοχή της αποδοτικής DL, προτείνοντας αρχιτεκτονικές

και μεθοδολογίες που επιτυγχάνουν ισορροπία μεταξύ συνολικής ακρίβειας εξαγωγής συμπεράσματος και χαμηλής υπολογιστικής επιβάρυνσης. Οι προτεινόμενες προσεγγίσεις καθίστανται ιδιαίτερα κατάλληλες για ενσωμάτωση σε συστήματα με περιορισμένους υπολογιστικούς πόρους και για ανάπτυξη εφαρμογών σε περιβάλλοντα πραγματικού χρόνου, όπως οι ενσωματωμένες συσκευές χαμηλής ισχύος και κινητές ιατρικές συσκευές.

Τα ευρήματα επιβεβαιώνουν την επιτάχυνση που επιτυγχάνεται μέσω της χρήσης του ετερογενούς συστήματος NVIDIA Jetson AGX Xavier σε εφαρμογές ΑΙ στον ιατρικό τομέα, αξιοποιώντας τον συνδυασμό ενός προσαρμοσμένου CPU και της ισχυρής Volta GPU. Οι μετρικές ποιότητας που εφαρμόστηκαν – συνολική ακρίβεια, ακρίβεια των θετικών προβλέψεων, ανάκληση, βαθμολογία F1 και βαθμολογία AUC – παρέχουν μια συνολική αποτίμηση της επίδοσης των μοντέλων. Η ανάλυση των αποτελεσμάτων ανέδειξε ότι το προτεινόμενο CNN, ιδίως όταν εκπαιδεύεται με εστιακή συνάρτηση απώλειας, υπερέρχει απο τα προεκπαιδευμένα μοντέλα ως προς την ικανότητα ταξινόμησης, επιτυγχάνοντας υψηλή ακρίβεια και αποτελεσματικότητα στην αντιμετώπιση της κατηγορικής ανισορροπίας.

Κατά τη φάση της εκπαίδευσης και επικύρωσης, το προτεινόμενο μοντέλο εμφάνισε σημαντικά χαμηλότερους χρόνους εκπαίδευσης και ελέγχου. Η καλύτερη αποδοτικότητα επιτυγχάνεται με τη χρήση της εστιακής απώλειας, η οποία επιταχύνει τη σύγκλιση εστιάζοντας σε δύσκολα δείγματα. Το προτεινόμενο CNN σημειώνει συνολική ακρίβεια 98,56% με την εφαρμογή της εστιακής απώλειας, επιβεβαιώνοντας την αποτελεσματικότητά του έναντι των άλλων μοντέλων.

Επιπλέον, η αξιολόγηση του χρόνου εκτέλεσης εξαγωγής συμπερασμάτων και των καρτέ ανά δευτερόλεπτο σε GPU και CPU έδειξε σημαντικά μικρότερους χρόνους για το προτεινόμενο μοντέλο, ενισχύοντας τη χρησιμότητά του σε εφαρμογές πραγματικού χρόνου με ταχείες απαιτήσεις απόκρισης. Η αποδοτικότητα αυτή, σε συνδυασμό με την υψηλή συνολική ακρίβεια και την ενεργειακή αποδοτικότητα, καθιστούν το μοντέλο κατάλληλο για ευρεία εφαρμογή σε ιατρικά συστήματα.

Σε σύγκριση με προηγούμενες μελέτες (Πίνακας 5.1), η προτεινόμενη αρχιτεκτονική διακρίνεται για την χαμηλή πολυπλοκότητα και την υπολογιστική της αποδοτικότητα. Αντιθέτως, προσεγγίσεις όπως το CoroNet [348] και το DWTMBConvNet [345], χαρακτηρίζονται από υψηλό επίπεδο αρχιτεκτονικής πολυπλοκότητας και αυξημένες απαιτήσεις σε πόρους, περιορίζοντας την καταλληλότητά τους για ενσωμάτωση σε συστήματα περιορισμένης υπολογιστικής

ισχύος. Παράλληλα, η επίδοσή μας ανταγωνίζεται επάξια τα καλύτερα αποτελέσματα της βιβλιογραφίας (όπως τα μοντέλα των Karthik et al. [349] και Ibrahim et al. [350]), διατηρώντας ταυτόχρονα χαμηλό υπολογιστικό αποτύπωμα, γεγονός που την καθιστά ιδανική για ενσωμάτωση σε εφαρμογές στο σημείο φροντίδας.

Τέλος, η συνολική ακρίβεια των μοντέλων πρόβλεψης, ιδίως στον τομέα της ιατρικής διάγνωσης, δύναται να επηρεαστεί σημαντικά από εξωγενείς παράγοντες. Έρευνες υπογραμμίζουν τη σημασία ενσωμάτωσης περιβαλλοντικών και κοινωνικοοικονομικών δεδομένων, τα οποία μπορούν να ενισχύσουν τις επιδόσεις των μοντέλων. Στο πεδίο της διάγνωσης των πνευμονοπαθειών, παράγοντες όπως η γεωγραφική θέση, οι κοινωνικοοικονομικές συνθήκες και τα δημογραφικά χαρακτηριστικά των ασθενών δύνανται να προσφέρουν πολύτιμες πληροφορίες για την εξατομίκευση της διάγνωσης και της θεραπευτικής προσέγγισης. Η ένταξη τέτοιων πολυδιάστατων δεδομένων μπορεί να οδηγήσει σε περισσότερο προσωποποιημένες και αποδοτικές υπηρεσίες υγείας, προσαρμοσμένες στις πραγματικές ανάγκες του κάθε ασθενούς.

5.8 Συμπεράσματα και Μελλοντικές Επεκτάσεις

Η παρούσα μελέτη στηρίχθηκε στην πολυδιάστατη μεθοδολογική προσέγγιση που αναπτύξαμε, η οποία ενσωματώνει μετρικές σχεδίασης, ποιότητας και υπολογιστικού αποτυπώματος. Μέσω αυτής της μεθοδολογίας η αξιολόγηση επικεντρώθηκε στην επίτευξη ισορροπίας ανάμεσα στην συνολική ακρίβεια και τη χρήση υπολογιστικών πόρων, ώστε να διασφαλιστεί η πρακτική εφαρμογή του προτεινόμενου συστήματος σε περιβάλλοντα σημείου φροντίδας.

Η έρευνα επικεντρώνεται στην αποτελεσματική αναγνώριση των πνευμονικών παθήσεων και των φυσιολογικών περιπτώσεων μέσω της αξιοποίησης του προτεινόμενου CNN χαμηλής πολυπλοκότητας και έξι προεκπαιδευμένων αρχιτεκτονικών: ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT-B/16. Για την κάλυψη των απαιτήσεων του πειραματικού πλαισίου, κατασκευάστηκε ένα ευρύ σύνολο δεδομένων, το οποίο περιλαμβάνει 29.131 ακτινογραφίες θώρακα προερχόμενες από τη συγχώνευση τεσσάρων δημόσια διαθέσιμων βάσεων δεδομένων με επτά διακριτές κατηγορίες.

Το προτεινόμενο CNN σχεδιάστηκε ώστε να επιτυγχάνει ισορροπία μεταξύ της συνολικής ακρίβειας στην κατηγοριοποίηση των ιατρικών εικόνων και της υπολογιστικής αποδοτικότητας — δύο κρίσιμων παραμέτρων σε περιβάλλοντα υγειονομικής περίθαλψης με περιορισμένους πόρους. Για την υλοποίησή του, επιλέχθηκε το ετερογενές σύστημα NVIDIA Jetson Xavier, δεδομένης της καταλληλότητάς της για την εκτέλεση εφαρμογών DL, γεγονός που ευθυγραμμίζεται με τον στόχο ενσωμάτωσης του προτεινόμενου μοντέλου σε πραγματικές ιατρικές συνθήκες, ακόμη και σε εγκαταστάσεις περιορισμένων υπολογιστικών δυνατοτήτων.

Η βελτιστοποίηση της αρχιτεκτονικής επικεντρώνεται στη μείωση της υπολογιστικής πολυπλοκότητας, αξιοποιώντας τις δυνατότητες επιτάχυνσης που προσφέρει η GPU του NVIDIA Jetson Xavier. Μέσω αυτής της προσέγγισης, επιτυγχάνεται η διατήρηση της υψηλής απόδοσης και ακρίβειας εξαγωγής συμπερασμάτων ακόμη και σε περιβάλλοντα με περιορισμούς υπολογιστικών και ενεργειακών πόρων. Η συγκεκριμένη βελτιστοποίηση καθιστά την αρχιτεκτονική κατάλληλη για ενσωμάτωση σε εφαρμογές απεικόνισης στο σημείο φροντίδας, όπου απαιτείται άμεση επεξεργασία ιατρικών δεδομένων σε πραγματικές κλινικές συνθήκες.

Από την ανάλυση των μετρικών ποιότητας, το προτεινόμενο μοντέλο χαμηλής πολυπλοκότητας ξεπέρασε τις επιδόσεις των προεκπαιδευμένων δικτύων, επιτυγχάνοντας μέση συνολική ακρίβεια 98,56% και τιμές ανάκλησης, ακρίβειας των θετικών προβλέψεων, βαθμολογίας F1 και βαθμολογίας AUC ίσες με 98,91%, 99,10%, 99,01% και 99,30%, αντίστοιχα, για ένα σύστημα επτακατηγορικής ταξινόμησης. Αντιθέτως, τα προεκπαιδευμένα μοντέλα εμφάνισαν σημαντικά χαμηλότερες επιδόσεις, με τις τιμές συνολικής ακρίβειας να κυμαίνονται από 85,43% έως 90,11%.

Περαιτέρω, το προτεινόμενο μοντέλο ολοκλήρωσε την εκπαίδευσή του σε λιγότερο από 60 λεπτά, ενώ ο αριθμός παραμέτρων του είναι κατά 5,35 φορές μικρότερος συγκριτικά με τα προεκπαιδευμένα δίκτυα, καθιστώντας το ιδιαίτερος κατάλληλο για ενσωματωμένες συσκευές. Επιπλέον, ο χρόνος εξαγωγής συμπεράσματος του προτεινόμενου μοντέλου ήταν 17,14 s, ενώ οι αντίστοιχοι χρόνοι για τα ResNet50, NASNetMobile, Xception, MobileNetV2, DenseNet121 και ViT-B/16 ήταν 29,36 s, 22,12 s, 20,59 s, 21,48 s, 26,14 s και 22,52 s, αντίστοιχα. Η βαθμολογία AUC του προτεινόμενου δικτύου ήταν 99,30%, ξεπερνώντας τις αντίστοιχες τιμές των συγκριτικών μοντέλων, οι οποίες κυμάνθηκαν από 91,23% έως 93,53%.

Η μελλοντική κατεύθυνση της παρούσας εργασίας περιλαμβάνει τη διεύρυνση του πλαισίου της κατηγοριοποίησης των πνευμονικών παθήσεων, μέσω της ενσωμάτωσης περισσότερων κατηγοριών παθολογιών και της επέκτασης του συνόλου δεδομένων με μεταβλητές όπως δημογραφικά χαρακτηριστικά ασθενών, περιβαλλοντικά δεδομένα και κοινωνικοοικονομικούς δείκτες. Η διεύρυνση αυτή αναμένεται να ενισχύσει την ικανότητα του μοντέλου να προσαρμόζεται σε ποικίλα ιατρικά περιβάλλοντα και να αποδίδει με υψηλότερη ακρίβεια στην αυξημένη πολυπλοκότητα και την ετερογένεια των δεδομένων, προσεγγίζοντας τη λογική της προσωποποιημένης ιατρικής διάγνωσης.

Κεφάλαιο 6

Συμπεράσματα

Η παρούσα διατριβή διερεύνησε σε βάθος τον απαιτητικό και διαρκώς εξελισσόμενο τομέα της επιτάχυνσης και της υλοποίησης εφαρμογών σε HCS. Κεντρικός άξονας της διατριβής αποτέλεσε ο σχεδιασμός, η βελτιστοποίηση και η υλοποίηση των CNNs σε διάφορες ετερογενείς πλατφόρμες HCS, αξιοποιώντας τη μεθοδολογία που αναπτύξαμε. Η ολοκληρωμένη αυτή μεθοδολογία συνδυάζει πολυδιάστατες μετρικές απόδοσης και προηγμένες στρατηγικές βελτιστοποίησης, εξασφαλίζοντας τη βέλτιστη αξιοποίηση των διαθέσιμων υπολογιστικών πόρων. Μέσω της εφαρμογής της μεθοδολογίας που αναπτύχθηκε, αποδείχθηκε η δυνατότητα επιτάχυνσης εφαρμογών CNNs σε τυπικά HCS, όπως τα FPGA-SoC και τα SoMs. Η διατριβή αυτή προσφέρει μια εμπεριστατωμένη ανάλυση των μηχανισμών που καθορίζουν την αποτελεσματική εκτέλεση αλγορίθμων AI σε περιβάλλοντα με περιορισμένες υπολογιστικές και ενεργειακές δυνατότητες.

Καθώς ο τομέας των HCS και της DL συνεχίζει να εξελίσσεται, η χρήση της μεθοδολογίας μας απέδειξε ότι τα ευρήματα της παρούσας διατριβής ενισχύουν την ερευνητική συζήτηση σχετικά με την υλοποίηση ευέλικτων και αποδοτικών συστημάτων εξαγωγής συμπερασμάτων AI σε πραγματικό χρόνο. Μέσω της μεθοδολογίας, η εις βάθος ανάλυση των παραμέτρων βελτιστοποίησης των CNNs σε HCS που πραγματοποιήθηκε στην παρούσα διατριβή συνιστά ουσιαστική πρόοδο προς τη γεφύρωση του χάσματος μεταξύ της ακρίβειας συμπερασμού, της χρονικής απόδοσης και της ενεργειακής κατανάλωσης. Τα αποτελέσματα της έρευνας ανοίγουν νέες προοπτικές για τη σχεδίαση και υλοποίηση αποτελεσματικών και επεκτάσιμων συστημάτων AI σε σύγχρονες ετερογενείς υπολογιστικές υποδομές.

6.1 Κύριες Συνεισφορές της Διατριβής

Οι κύριες συνεισφορές της παρούσας διατριβής συνοψίζονται ως εξής:

- Ανάπτυξη μιας ολοκληρωμένης μεθοδολογίας για τον σχεδιασμό και την υλοποίηση εφαρμογών σε HCS, βασισμένη σε τρεις κύριους πυλώνες: μετρικές σχεδιασμού, μετρικές ποιότητας και μετρικές υπολογιστικού αποτυπώματος. Η μεθοδολογία προσαρμόζεται στις διαφορετικές αρχιτεκτονικές των HCS, υποστηρίζοντας την αποδοτική επιτάχυνση των εφαρμογών.
- Διαμόρφωση και εφαρμογή πολυδιάστατου πλαισίου μετρικών απόδοσης, καλύπτοντας σχεδίαση, ποιότητα και υπολογιστικό αποτύπωμα σε CNNs για HCS. Το πλαίσιο αυτό επιτρέπει την ολοκληρωμένη αξιολόγηση και βελτιστοποίηση των αρχιτεκτονικών, λαμβάνοντας υπόψη τόσο τις ανάγκες της εφαρμογής όσο και τους περιορισμούς του συστήματος.
- Σχεδίαση και υλοποίηση του MiniTomatoNet, ενός χαμηλής πολυπλοκότητας CNN με εξαιρετικά αποτελέσματα και ελάχιστη κατανάλωση πόρων στο FPGA–SoC. Το MiniTomatoNet ταξινομεί αξιόπιστα δέκα κατηγορίες ασθενειών φύλλων τομάτας, με σημαντικά μικρότερο αριθμό παραμέτρων και κατανάλωση ενέργειας σε σχέση με εναλλακτικές υλοποιήσεις σε άλλες συσκευές.
- Ανάπτυξη βελτιστοποιημένου CNN για ταξινόμηση πνευμονικών παθήσεων σε NVIDIA Jetson Xavier, με επιδόσεις που υπερβαίνουν καθιερωμένα δίκτυα. Το προτεινόμενο μοντέλο διακρίνεται για τον μικρό αριθμό παραμέτρων και την ταχύτητα επεξεργασίας, καθιστώντας το κατάλληλο για υλοποίηση σε συστήματα με περιορισμένη υπολογιστική ισχύ.

6.2 Περίληψη Ευρημάτων

Στην παρούσα διατριβή, αξιοποιώντας τη μεθοδολογία, μελετήθηκαν και υλοποιήθηκαν δύο προσεγγίσεις εφαρμογής CNNs σε HCS. Ιδιαίτερη έμφαση δόθηκε στην αποδοτική επεξεργασία εικόνων σε συσκευές με περιορισμένους πόρους υλικού. Η πρώτη μελέτη αφορούσε τον γεωργικό τομέα, με την ανάπτυξη του MiniTomatoNet, ενός χαμηλής πολυπλοκότητας CNN οχτώ επιπέδων για

την αναγνώριση ασθενειών φύλλων τομάτας σε FPGA–SoC (PYNQ–Z1). Η δεύτερη μελέτη εστίασε στον ιατρικό τομέα, παρουσιάζοντας ένα CNN χαμηλής πολυπλοκότητας για την ταξινόμηση πνευμονικών νοσημάτων από ακτινογραφίες θώρακος σε NVIDIA Jetson Xavier.

Και στις δύο περιπτώσεις, έγινε η υιοθέτηση κατάλληλων μετρικών, οι οποίες λαμβάνουν υπόψη τόσο στα τεχνικά χαρακτηριστικά των μοντέλων όσο και τις ιδιαίτερες απαιτήσεις του υπολογιστικού περιβάλλοντος. Η βέλτιστη επιλογή και εφαρμογή αυτών των μετρικών αποτελεί καθοριστικό παράγοντα για τη συνολική απόδοση κάθε συστήματος, καθώς επιτρέπει τη συγκριτική αποτίμηση εναλλακτικών σχεδιαστικών προσεγγίσεων και στρατηγικών βελτιστοποίησης. Οι μετρικές αποδοτικότητας διαρθρώνονται σε τρεις θεμελιώδεις κατηγορίες, οι οποίες ανταποκρίνονται στις επιμέρους διαστάσεις της σχεδίασης, της λειτουργικής ποιότητας και του υπολογιστικού αποτυπώματος των CNNs σε HCS:

- **Μετρικές σχεδίασης:** Αυτές σχετίζονται με τις επιλογές που αφορούν την αρχιτεκτονική και τις τεχνικές βελτιστοποίησης ενός CNN, με σκοπό τη μεγιστοποίηση της απόδοσης υπό τους περιορισμούς που επιβάλλονται από τα HCS. Σε αυτή την κατηγορία περιλαμβάνονται παράγοντες όπως η δομή του μοντέλου, οι μέθοδοι βελτιστοποίησης της απόδοσης, οι εκάστοτε συμβιβασμοί, καθώς και τα ζητήματα ενεργειακής απόδοσης που λαμβάνονται υπόψη κατά τον σχεδιασμό.
- **Μετρικές ποιότητας:** Επικεντρώνονται στην αξιολόγηση της ικανότητας ενός CNN να εκτελεί προβλεπτικές εργασίες με ακρίβεια και αξιοπιστία, ανεξάρτητα από το υπολογιστικό ή ενεργειακό κόστος που μπορεί να προκύπτει είτε στη φάση της εκπαίδευσης είτε κατά την παραγωγική λειτουργία. Οι μετρικές αυτές είναι καθοριστικές για τη λειτουργική αποτίμηση της αποτελεσματικότητας ενός μοντέλου, καθώς αποτυπώνουν τη σχέση μεταξύ των προβλέψεων του συστήματος και των πραγματικών αποτελεσμάτων.
- **Μετρικές υπολογιστικού αποτυπώματος:** Αφορούν ποσοτικά χαρακτηριστικά που αποδίδουν το υπολογιστικό, ενεργειακό και χρονικό κόστος τόσο κατά την εκπαίδευση όσο και κατά την εξαγωγή συμπερασμάτων ενός CNN. Οι συγκεκριμένες μετρικές διακρίνονται περαιτέρω, ανάλογα με το στάδιο του κύκλου ζωής του μοντέλου στο οποίο εφαρμόζονται: (α) μετρικές που αφορούν αποκλειστικά τη φάση της εκπαίδευσης, (β) μετρικές που σχετίζονται μόνο με

το στάδιο της εξαγωγής συμπερασμάτων και (γ) μετρικές που καλύπτουν και τις δύο φάσεις, παρέχοντας έτσι μία συνολική αποτίμηση της αποδοτικότητας του CNN στο εκάστοτε υπολογιστικό περιβάλλον.

Το MiniTomatoNet κατόρθωσε να ταξινομήει εικόνες φύλλων τομάτας σε δέκα κατηγορίες (εννέα διακριτές ασθένειες και μια κατηγορία υγιούς φύλλου) με εξαιρετική ακρίβεια. Συγκεκριμένα, διαθέτει μόλις 23 χιλιάδες παραμέτρους (μνήμη 89,5 KB) και πέτυχε συνολική ακρίβεια 97,63% στην αναγνώριση των ασθενειών, βαθμολογία AUC 98,51%, χρόνο εξαγωγής αποτελέσματος 0,068 s ανά εικόνα και κατανάλωση ισχύος 2,35 W στο FPGA. Αυτή η εξαιρετικά συμπαγής αρχιτεκτονική επιτεύχθηκε μέσω χαμηλής πολυπλοκότητας σχεδιαστικών επιλογών, όπως η χρήση συνέλιξης βάθους ανά κανάλι (βλ. Ενότητα 3.2.1.2) αντί των συμβατικών, μειώνοντας το πλήθος παραμέτρων και υπολογισμών (βλ. Ενότητα 3.4.3) περίπου κατά 9 φορές σε σύγκριση με ένα ισοδύναμο συμβατικό δίκτυο. Επιπλέον, εφαρμόστηκε η προσαρμοσμένη συνάρτηση επανασταθμισμένης εστιακής απώλειας (βλ. Ενότητα 3.2.2.3) για την αντιμετώπιση της ανισορροπίας των κλάσεων στο σύνολο δεδομένων, δίνοντας μεγαλύτερη έμφαση στις σπανιότερες κατηγορίες ασθενειών. Αυτή η τεχνική συνέβαλε σε αμερόληπτη βελτίωση της ακρίβειας του μοντέλου, εξασφαλίζοντας ότι ακόμα και οι υποεκπροσωπούμενες ασθένειες αναγνωρίζονται αξιόπιστα.

Αντίστοιχα, στη δεύτερη μελέτη, το προτεινόμενο CNN για την ταξινόμηση πνευμονοπαθειών από ακτινογραφίες θώρακος πέτυχε συγκρίσιμα υψηλές επιδόσεις. Επεξεργάστηκε ένα μεγάλο σύνολο 29.131 ακτινογραφιών από διαφορετικές πηγές, ταξινομώντας σε επτά κατηγορίες πνευμονικών παθήσεων. Το μοντέλο αυτό σχεδιάστηκε με γνώμονα την υλοποίηση σε συστήματα περιορισμένων πόρων (βλ. Ενότητα 3.2.1.1) και βελτιστοποιήθηκε για το ενσωματωμένο ετερογενές περιβάλλον του Jetson Xavier. Επιπλέον, εφαρμόστηκε η συνάρτηση εστιακής απώλειας (βλ. Ενότητα 3.2.2.3) για την αντιμετώπιση της ανισορροπίας των κλάσεων στο σύνολο δεδομένων, δίνοντας μεγαλύτερη έμφαση στις υποεκπροσωπούμενες κατηγορίες ασθενειών. Επιτεύχθηκε συνολική ακρίβεια 98,56%, ξεπερνώντας καθιερωμένα DL μοντέλα. Η καινοτομία του προτεινόμενου μοντέλου επιτεύχθηκε μέσω της προσεκτικής αρχιτεκτονικής διαμόρφωση με γνώμονα την ισορροπία μεταξύ του βάθους του δικτύου και της υπολογιστικής αποδοτικότητας. Ο συνολικός αριθμός παραμέτρων (3,3 εκατομμύρια) είναι 5 φορές μικρότερος από των προεκπαιδευμένων CNNs, γεγονός που καθιστά το μοντέλο ιδιαίτερα ευέλικτο για υλοποίηση σε ενσωματωμένες συσκευές (βλ. Ενότητα 3.4.3). Επιπλέον, χάρη στην απλότητα του

δικτύου, ο χρόνος εξαγωγής αποτελέσματος είναι 0,0165 ms και το μοντέλο μπορεί να ταξινομεί εικόνες με ρυθμό περίπου 4,53 FPS.

Τα ανωτέρω αποτελέσματα καταδεικνύουν ότι ο προσεκτικός σχεδιασμός και η προσαρμογή των CNNs στις ιδιαιτερότητες της εκάστοτε ετερογενούς συσκευής μπορούν να αποφέρουν εξαιρετικά επίπεδα ακρίβειας και αποδοτικότητας, ακόμη και σε περιβάλλοντα με περιορισμένους υπολογιστικούς ή ενεργειακούς πόρους. Η αξιοποίηση εξειδικευμένων τεχνικών, όπως οι συνέλιξεις βάθους ανά κανάλι και οι προηγμένες συναρτήσεις απώλειας για την αντιμετώπιση της ανισορροπίας των κλάσεων, διασφαλίζει όχι μόνο τη βέλτιστη εκμετάλλευση των διαθέσιμων πόρων, αλλά και την αξιόπιστη απόδοση του συστήματος ακόμη και σε απαιτητικά ή ανομοιογενή σύνολα δεδομένων.

Τα πειραματικά ευρήματα αποκαλύπτουν ότι η σύγκριση μεταξύ FPGA-SoC και SoM/GPU αρχιτεκτονικών αναδεικνύει διαφορετικά πλεονεκτήματα και μειονεκτήματα, ανάλογα με τις επιμέρους λειτουργικές απαιτήσεις. Η PYNQ-Z1, ως πλατφόρμα FPGA-SoC, επιδεικνύει εξαιρετική επιτάχυνση κατά την φάση της εξαγωγής συμπερασμάτων, ενεργειακή απόδοση και προσαρμοστικότητα, καθιστώντας την ελκυστική για εφαρμογές με περιορισμούς ισχύος και για περιπτώσεις όπου απαιτούνται χαμηλού επιπέδου βελτιστοποιήσεις υλικού. Ωστόσο, η πολυπλοκότητα ανάπτυξης και οι περιορισμένοι πόροι μπορεί να αποτελέσουν πρόκληση για ορισμένες εφαρμογές. Από την άλλη πλευρά, τα συστήματα SoM όπως το NVIDIA Jetson προσφέρουν υψηλή απόδοση και ευκολία ανάπτυξης χάρη στο ώριμο οικοσύστημα λογισμικού και την ισχυρή υποστήριξη της GPU, καθιστώντας τα ιδανικά για περιπτώσεις που απαιτούν υψηλούς ρυθμούς FPS και όπου η ευκολία ανάπτυξης είναι πρωταρχικής σημασίας.

6.3 Προτάσεις για Μελλοντική Έρευνα

Η παρούσα διατριβή θέτει τα θεμέλια για μια σειρά από κατευθύνσεις μελλοντικής έρευνας, οι οποίες αποσκοπούν στη βελτιστοποίηση της αποδοτικότητας και της ευελιξίας εφαρμογών AI σε HCS:

- Αυτοματοποιημένη προσέγγιση εξερεύνησης χώρου σχεδίασης: Η βασική πρόκληση που αντιμετωπίζει ο σχεδιαστής είναι η ύπαρξη ενός τεράστιου χώρου δυνατών παραμέτρων — όπως το πλήθος και το είδος των

επιπέδων σε ένα CNN, οι υπερπαραμέτροι εκπαίδευσης, αλλά και οι αρχιτεκτονικές αποφάσεις σε επίπεδο υλικού. Η αυτοματοποίηση θα αναζητά αποδοτικά τον χώρο των πιθανών σχεδιαστικών επιλογών, ελαχιστοποιώντας το χρόνο και το υπολογιστικό κόστος της αναζήτησης. Μέσω της αξιολόγησης πολυδιάστατων μετρικών – όπως η ακρίβεια, η ταχύτητα, η κατανάλωση ενέργειας, το αποτύπωμα μνήμης και το κόστος υλοποίησης – η αυτοματοποίηση θα καθιστά δυνατή την αυτόματη ανεύρεση της ισορροπίας μεταξύ αντικρουόμενων στόχων ανάλογα με τις απαιτήσεις της εφαρμογής. Επιπλέον, η αυτοματοποιημένη προσέγγιση επιτρέπει την ευελιξία σχεδίασης για εξειδικευμένες συσκευές επιτάχυνσης, όπως FPGA, SoC, SoM, προσαρμόζοντας το μοντέλο στις ιδιαιτερότητες του υλικού και βελτιστοποιώντας την απόδοσή του στο εκάστοτε περιβάλλον. Η ενσωμάτωση τεχνικών αυτοματοποιημένης εξερεύνησης χώρου σχεδίασης στον κύκλο ζωής ανάπτυξης συστημάτων DL και HCS αποτελεί καθοριστικό παράγοντα για την ανάπτυξη λύσεων υψηλής απόδοσης σε απαιτητικές εφαρμογές πραγματικών συνθηκών.

- Διερεύνηση αρχιτεκτονικών μικτής ακρίβειας: Σε αντίθεση με τις παραδοσιακές προσεγγίσεις, όπου όλα τα αριθμητικά δεδομένα και οι παράμετροι των μοντέλων κωδικοποιούνται με μία ενιαία ακρίβεια (συνήθως FP32), οι αρχιτεκτονικές μικτής ακρίβειας υιοθετούν τη στρατηγική χρήσης διαφορετικών επιπέδων αριθμητικής ακρίβειας (π.χ. FP16, INT8 ή ακόμη και INT4) σε επιμέρους τμήματα του CNN ή σε διαδοχικά στάδια της υπολογιστικής διαδικασίας. Η προσέγγιση αυτή επιτρέπει τη δραστηκή μείωση της κατανάλωσης μνήμης, της απαιτούμενης υπολογιστικής ισχύος και της ενεργειακής δαπάνης, χωρίς να θυσιάζεται η τελική ακρίβεια του μοντέλου, καθώς κρίσιμα τμήματα ή συγκεκριμένες παράμετροι μπορούν να διατηρούν υψηλότερη ακρίβεια όταν αυτό απαιτείται. Η υλοποίηση μοντέλων μικτής ακρίβειας είναι ιδιαίτερα κατάλληλη για τα FPGA, ενσωματωμένες συσκευές και εξειδικευμένους επιταχυντές, όπου ο περιορισμός πόρων και η ανάγκη για πραγματικού χρόνου επεξεργασία καθιστούν ζωτικής σημασίας τη βέλτιστη διαχείριση του πλήθους των bits που χρησιμοποιούνται για την αριθμητική αναπαράσταση. Η μελλοντική έρευνα στη συγκεκριμένη κατεύθυνση αναμένεται να συμβάλει στην ανάπτυξη συστημάτων που επιτυγχάνουν την ιδανική ισορροπία ανάμεσα στην ακρίβεια των αποτελεσμάτων και στην αποδοτικότητα ως προς τους υπολογιστικούς και ενεργειακούς πόρους, διευρύνοντας ακόμη περισσότερο το φάσμα εφαρμογών των έξυπνων HCS.

- Διεύρυνση ενεργειακά αποδοτικών προσεγγίσεων εκπαίδευσης και εξαγωγής αποτελεσμάτων: Η υλοποίηση αλγορίθμων CNNs σε HCS αναδεικνύει τη σημασία της μελλοντικής έρευνας στην ανάπτυξη ενεργειακά αποδοτικών λύσεων, οι οποίες αφορούν όχι μόνο τη φάση της εξαγωγής αποτελεσμάτων, αλλά και τη φάση της εκπαίδευσης των μοντέλων. Ειδικότερα, η ανάγκη για συστήματα που να ανταποκρίνονται σε περιορισμούς ενεργειακής κατανάλωσης και διαθεσιμότητας υπολογιστικών πόρων καθιστά απαραίτητη την υιοθέτηση προηγμένων τεχνικών αξιολόγησης και βελτιστοποίησης, όπου η μέτρηση της ενεργειακής και υπολογιστικής απόδοσης πραγματοποιείται απευθείας στο υλικό—στόχο κατά τη διάρκεια τόσο της ανάπτυξης όσο και της αξιολόγησης του συστήματος. Με αυτόν τον τρόπο, επιτυγχάνεται η βέλτιστη παραμετροποίηση και προσαρμογή των μοντέλων, εξασφαλίζοντας ότι οι επιδόσεις τους ανταποκρίνονται στις πραγματικές συνθήκες λειτουργίας και τους περιορισμούς του συστήματος. Τέτοιες προσεγγίσεις αποκτούν ιδιαίτερη αξία σε σύγχρονες εφαρμογές αιχμής, όπως τα edge και IoT συστήματα, όπου η αυτονομία, η χαμηλή κατανάλωση ενέργειας και η δυνατότητα άμεσης προσαρμογής σε μεταβαλλόμενα περιβάλλοντα αποτελούν κρίσιμους παράγοντες για την επιτυχή και αξιόπιστη εφαρμογή λύσεων AI σε HCS.
- Διεύρυνση σε πολυτροπικές τεχνικές DL: Οι πολυτροπικές μέθοδοι ξεπερνούν τους περιορισμούς των μονοτροπικών προσεγγίσεων, καθώς ενσωματώνουν πληροφορίες διαφορετικής φύσης, όπως εικόνες, ήχο, κείμενο, χωρικά και χρονικά δεδομένα, βιομετρικούς δείκτες ή ακόμη και κοινωνικο—οικονομικά δεδομένα, δημιουργώντας πλουσιότερες και πιο ανθεκτικές αναπαραστάσεις της πραγματικότητας. Η δυνατότητα συγχώνευσης πολλαπλών τύπων πληροφορίας επιτρέπει την ανάπτυξη συστημάτων με ανώτερη ακρίβεια, αυξημένη γενικευσιμότητα και βελτιωμένη ανθεκτικότητα σε αβεβαιότητα ή μερικώς ελλιπή δεδομένα, γεγονός που τα καθιστά εξαιρετικά πολύτιμα για εφαρμογές που κυμαίνονται από την ανάλυση πολυμεσικών δεδομένων και την έξυπνη ρομποτική, έως την ανίχνευση περιστατικών σε έξυπνες πόλεις, τη βιομηχανική διάγνωση, την πολυτροπική ανάλυση ανθρώπινης συμπεριφοράς, και την εξατομικευμένη υγεία. Η ενσωμάτωση πολυτροπικών προσεγγίσεων σε HCS, όπως συνδυαστικά περιβάλλοντα CPU, GPU, FPGA και εξειδικευμένων επιταχυντών, δημιουργεί νέες προκλήσεις αλλά και ευκαιρίες για δυναμική κατανομή υπολογιστικού φορτίου, βέλτιστη διαχείριση ενεργειακών και υπολογιστικών πόρων, καθώς και διαλειτουργικότητα σε πραγματικό χρόνο. Προϋποθέτει την ανάπτυξη εξελιγμένων τεχνικών συγχώνευσης δεδομένων,

προηγμένων αλγορίθμων και διαλειτουργικών μοντέλων που μπορούν να διαχειριστούν αποτελεσματικά την ποικιλομορφία, την ετερογένεια και τη δυναμική των εισερχόμενων δεδομένων. Επομένως, η διεύρυνση της έρευνας σε πολυτροπικές προσεγγίσεις αποτελεί στρατηγικό άξονα για την ανάπτυξη ευφύων, ανθεκτικών και πρακτικά εφαρμόσιμων συστημάτων που δύνανται να λειτουργήσουν αποτελεσματικά σε πληθώρα επιστημονικών πεδίων.

Βιβλιογραφία

- [1] Rahul Mishra and Hari Gupta. Transforming large-size to lightweight deep neural networks for IoT applications. *ACM Computing Surveys*, 55(11):1–35, 2023.
- [2] Giosué Cataldo Marinó, Alessandro Petrini, Dario Malchiodi, and Marco Frasca. Deep neural networks compression: A comparative survey and choice recommendations. *Neurocomputing*, 520:152–170, 2023.
- [3] Qing Xu, Zhenghua Chen, Keyu Wu, Chao Wang, Min Wu, and Xiaoli Li. KDnet-RUL: A knowledge distillation framework to compress deep neural networks for machine remaining useful life prediction. *IEEE Transactions on Industrial Electronics*, 69(2):2022–2032, 2021.
- [4] Paul D Rosero-Montalvo, Pınar Tözün, and Wilmar Hernandez. Optimized CNN architectures benchmarking in hardware-constrained edge devices in IoT environments. *IEEE Internet of Things Journal*, 2024.
- [5] Xiaotian Guo, Andy D Pimentel, and Todor Stefanov. Automated exploration and implementation of distributed cnn inference at the edge. *IEEE Internet of Things Journal*, 10(7):5843–5858, 2023.
- [6] Muhammad Zawish, Nouman Ashraf, Rafay Iqbal Ansari, and Steven Davy. Energy-aware AI-driven framework for edge-computing-based IoT applications. *IEEE Internet of Things Journal*, 10(6):5013–5023, 2022.
- [7] Hamza Djigal, Jia Xu, Linfeng Liu, and Yan Zhang. Machine and deep learning for resource allocation in multi-access edge computing: A survey. *IEEE Communications Surveys & Tutorials*, 24(4):2449–2494, 2022.
- [8] Zhuoqing Chang, Shubo Liu, Xingxing Xiong, Zhaohui Cai, and Guoqing Tu. A survey of recent advances in edge-computing-powered artificial intelligence of things. *IEEE Internet of Things Journal*, 8(18):13849–13875, 2021.

- [9] JunKyu Lee, Lev Mukhanov, Amir Sabbagh Molahosseini, Umar Minhas, Yang Hua, Jesus Martinez del Rincon, Kiril Dichev, Cheol-Ho Hong, and Hans Vandierendonck. Resource-efficient convolutional networks: A survey on model-, arithmetic-, and implementation-level techniques. *ACM Computing Surveys*, 55(13s):1–36, 2023.
- [10] Ahmed Mujtaba, Wai-Kong Lee, and Seong Oun Hwang. Low latency implementations of CNN for resource-constrained IoT devices. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(12):5124–5128, 2022.
- [11] Md Maruf Hossain Shuvo, Syed Kamrul Islam, Jianlin Cheng, and Bashir I Morshed. Efficient acceleration of deep learning inference on resource-constrained edge devices: A review. *Proceedings of the IEEE*, 111(1): 42–91, 2022.
- [12] Tianming Zhao, Yucheng Xie, Yan Wang, Jerry Cheng, Xiaonan Guo, Bin Hu, and Yingying Chen. A survey of deep learning on mobile devices: Applications, optimizations, challenges, and research opportunities. *Proceedings of the IEEE*, 110(3):334–354, 2022.
- [13] Jaekwang Cha and Shiho Kim. CNN hardware accelerator architecture design for energy-efficient AI. In *Artificial Intelligence and Hardware Accelerators*, pages 319–357. Springer, 2023.
- [14] Mahboob Elahi, Samuel Olaiya Afolaranmi, Jose Luis Martinez Lastra, and Jose Antonio Perez Garcia. A comprehensive literature review of the applications of AI techniques through the lifecycle of industrial equipment. *Discover Artificial Intelligence*, 3(1):43, 2023.
- [15] Ali Ghoroghi, Yacine Rezgui, Ioan Petri, and Thomas Beach. Advances in application of machine learning to life cycle assessment: a literature review. *The International Journal of Life Cycle Assessment*, 27(3):433–456, 2022.
- [16] Lei Wang, Zhengchao Liu, Ang Liu, and Fei Tao. Artificial intelligence in product lifecycle management. *The International Journal of Advanced Manufacturing Technology*, 114:771–796, 2021.
- [17] Cristina Silvano, Daniele Ielmini, Fabrizio Ferrandi, Leandro Fiorin, Serena Curzel, Luca Benini, Francesco Conti, Angelo Garofalo, Cristian Zambelli, Enrico Calore, et al. A survey on deep learning hardware accelerators for heterogeneous hpc platforms. *ACM Computing Surveys*, 2023.

- [18] Nicolás Hernández, Francisco Almeida, and Vicente Blanco. Optimizing convolutional neural networks for IoT devices: performance and energy efficiency of quantization techniques. *The Journal of Supercomputing*, 80(9): 12686–12705, 2024.
- [19] Fumio Hamanaka, Takashi Odan, Kenji Kise, and Thiem Van Chu. An exploration of state-of-the-art automation frameworks for FPGA-based DNN acceleration. *IEEE Access*, 11:5701–5713, 2023.
- [20] Lingwei Xu, Jingjing Wang, Xingwang Li, Fen Cai, Ye Tao, and T Aaron Gulliver. Performance analysis and prediction for mobile internet-of-things (IoT) networks: a CNN approach. *IEEE Internet of Things Journal*, 8(17): 13355–13366, 2021.
- [21] Biagio Peccerillo, Mirco Mannino, Andrea Mondelli, and Sandro Bartolini. A survey on hardware accelerators: Taxonomy, trends, challenges, and perspectives. *Journal of Systems Architecture*, 129:102561, 2022.
- [22] Ghattas Akkad, Ali Mansour, and Elie Inaty. Embedded deep learning accelerators: A survey on recent advances. *IEEE Transactions on Artificial Intelligence*, 5(5):1954–1972, 2023.
- [23] Meriam Dhouibi, Ahmed Karim Ben Salem, Afef Saidi, and Slim Ben Saoud. Accelerating deep neural networks implementation: A survey. *IET Computers & Digital Techniques*, 15(2):79–96, 2021.
- [24] Hanan Hussain, PS Tamizharasan, and CS Rahul. Design possibilities and challenges of DNN models: a review on the perspective of end devices. *Artificial Intelligence Review*, pages 1–59, 2022.
- [25] Sparsh Mittal and Shraiyysh Vaishay. A survey of techniques for optimizing deep learning on GPUs. *Journal of Systems Architecture*, 99:101635, 2019.
- [26] Sparsh Mittal. A survey of techniques for architecting and managing asymmetric multicore processors. *ACM Computing Surveys (CSUR)*, 48(3):1–38, 2016.
- [27] Ahmed Ghazi Blaiech, Khaled Ben Khalifa, Carlos Valderrama, Marcelo AC Fernandes, and Mohamed Hedi Bedoui. A survey and taxonomy of FPGA-based deep learning accelerators. *Journal of Systems Architecture*, 98:331–345, 2019.

- [28] Kaiyuan Guo, Shulin Zeng, Jincheng Yu, Yu Wang, and Huazhong Yang. [DL] A survey of FPGA-based neural network inference accelerators. *ACM Transactions on Reconfigurable Technology and Systems (TRETs)*, 12(1):1–26, 2019.
- [29] Cristobal Ortega, Lluc Alvarez, Alper Buyuktosunoglu, Ramon Bertran, Todd Rosedahl, Pradip Bose, and Miquel Moreto. Adaptive Power Shifting for Power-Constrained Heterogeneous Systems. *IEEE Transactions on Computers*, 72(3):627–640, 2022.
- [30] Mihaela Malița, George Vlăduț Popescu, and Gheorghe M Ștefan. Heterogeneous computing system for deep learning. *Deep learning: Concepts and architectures*, pages 287–319, 2020.
- [31] Hongzhi Xu, Binlian Zhang, Chen Pan, and Keqin Li. Energy-efficient scheduling for parallel applications with reliability and time constraints on heterogeneous distributed systems. *Journal of Systems Architecture*, 152:103173, 2024.
- [32] Peter Greenhalgh. Big. little processing with arm cortex-a15 & cortex-a7. *ARM White paper*, 17, 2011.
- [33] Tal Ben-Nun and Torsten Hoefler. Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. *ACM Computing Surveys (CSUR)*, 52(4):1–43, 2019.
- [34] Mostafa W Numan, Braden J Phillips, Gavin S Puddy, and Katrina Falkner. Towards automatic high-level code deployment on reconfigurable platforms: A survey of high-level synthesis tools and toolchains. *IEEE Access*, 8: 174692–174722, 2020.
- [35] Anakhi Hazarika, Soumyajit Poddar, and Hafizur Rahaman. Survey on memory management techniques in heterogeneous computing systems. *IET Computers & Digital Techniques*, 14(2):47–60, 2020.
- [36] Dina G Mahmoud, Vincent Lenders, and Mirjana Stojilović. Electrical-level attacks on CPUs, FPGAs, and GPUs: Survey and implications in the heterogeneous era. *ACM Computing Surveys (CSUR)*, 55(3):1–40, 2022.
- [37] Sparsh Mittal. A survey on modeling and improving reliability of DNN algorithms and accelerators. *Journal of Systems Architecture*, 104:101689, 2020.

- [38] Sparsh Mittal. A survey of FPGA-based accelerators for convolutional neural networks. *Neural computing and applications*, 32(4):1109–1139, 2020.
- [39] Qunsong Zeng, Yuqing Du, Kaibin Huang, and Kin K Leung. Energy-efficient resource management for federated edge learning with CPU-GPU heterogeneous computing. *IEEE Transactions on Wireless Communications*, 20(12):7947–7962, 2021.
- [40] Jinchao Chen, Yu He, Ying Zhang, Pengcheng Han, and Chenglie Du. Energy-aware scheduling for dependent tasks in heterogeneous multiprocessor systems. *Journal of Systems Architecture*, 129:102598, 2022.
- [41] Jie Feng, Wenjing Zhang, Qingqi Pei, Jinsong Wu, and Xiaodong Lin. Heterogeneous computation and resource allocation for wireless powered federated edge learning systems. *IEEE Transactions on Communications*, 70(5):3220–3233, 2022.
- [42] Anish Krishnakumar, Umit Ogras, Radu Marculescu, Mike Kishinevsky, and Trevor Mudge. Domain-specific architectures: Research problems and promising approaches. *ACM Transactions on Embedded Computing Systems*, 22(2):1–26, 2023.
- [43] Ramesh Krishnamoorthy, Kalimuthu Krishnan, Bharatiraja Chokkalingam, Sanjeevikumar Padmanaban, Zbigniew Leonowicz, Jens Bo Holm-Nielsen, and Massimo Mitolo. Systematic approach for state-of-the-art architectures and system-on-chip selection for heterogeneous IoT applications. *IEEE Access*, 9:25594–25622, 2021.
- [44] Luis Gomes and Zita Vale. Energy Predictions for System on a Chip. In *16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2021)*, volume 1401, page 243. Springer Nature, 2021.
- [45] Xiaochen Peng, Shanshi Huang, Hongwu Jiang, Anni Lu, and Shimeng Yu. DNN+ NeuroSim V2. 0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(11):2306–2319, 2020.
- [46] Yassin Kortli, Souhir Gabsi, Lew FC Lew Yan Voon, Maher Jridi, Mehrez Merzougui, and Mohamed Atri. Deep embedded hybrid CNN-LSTM network

- for lane detection on NVIDIA Jetson Xavier NX. *Knowledge-based systems*, 240: 107941, 2022.
- [47] Agus Kurniawan. Iot projects with nvidia jetson nano. *IoT Projects with NVIDIA Jetson Nano*, 2021.
- [48] Marcello M Mariani, Isa Machado, Vittoria Magrelli, and Yogesh K Dwivedi. Artificial intelligence in innovation research: A systematic review, conceptual framework, and future research directions. *Technovation*, 122:102623, 2023.
- [49] Ida Merete Enholm, Emmanouil Papagiannidis, Patrick Mikalef, and John Krogstie. Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, 24(5):1709–1734, 2022.
- [50] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53:5455–5516, 2020.
- [51] Nurullah Yüksel, Hüseyin Rıza Börklü, Hüseyin Kürşad Sezer, and Olcay Ersel Canyurt. Review of artificial intelligence applications in engineering design perspective. *Engineering Applications of Artificial Intelligence*, 118:105697, 2023.
- [52] Mohammad Hasan Ahmadilivani, Mahdi Taheri, Jaan Raik, Masoud Daneshtalab, and Maksim Jenihhin. A systematic literature review on hardware reliability assessment methods for deep neural networks. *ACM Computing Surveys*, 56(6):1–39, 2024.
- [53] Pudi Dhilleswararao, Srinivas Boppu, M Sabarimalai Manikandan, and Linga Reddy Cenkeramaddi. Efficient hardware architectures for accelerating deep neural networks: Survey. *IEEE access*, 10:131788–131828, 2022.
- [54] Jia-Hao Syu, Jerry Chun-Wei Lin, Gautam Srivastava, and Keping Yu. A comprehensive survey on artificial intelligence empowered edge computing on consumer electronics. *IEEE Transactions on Consumer Electronics*, 69(4): 1023–1034, 2023.
- [55] Qian Zhang, Jie Lu, and Yaochu Jin. Artificial intelligence in recommender systems. *Complex & Intelligent Systems*, 7(1):439–457, 2021.
- [56] Asad Hayat, Yasir Noman Khalid, Muhammad Siraj Rathore, and Muhammad Nadeem Nadir. A machine learning-based resource-efficient task

- scheduler for heterogeneous computer systems. *The Journal of Supercomputing*, 79(14):15700–15728, 2023.
- [57] Zhiheng Zhong, Minxian Xu, Maria Alejandra Rodriguez, Chengzhong Xu, and Rajkumar Buyya. Machine learning-based orchestration of containers: A taxonomy and future directions. *ACM Computing Surveys (CSUR)*, 54(10s):1–35, 2022.
- [58] Raffaele Pugliese, Stefano Regondi, and Riccardo Marini. Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Science and Management*, 4:19–29, 2021.
- [59] Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(3):1327–1347, 2022.
- [60] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch. A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access*, 9:82146–82168, 2021.
- [61] Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincui Huang, Xin Xu, Bin Dai, and Qiguang Miao. Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5064–5078, 2022.
- [62] Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, 231:120495, 2023.
- [63] Poojan Oza, Vishwanath A Sindagi, Vibashan Vishnukumar Sharmini, and Vishal M Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [64] Ji Liu, Jizhou Huang, Yang Zhou, Xuhong Li, Shilei Ji, Haoyi Xiong, and Dejing Dou. From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, 64(4):885–917, 2022.
- [65] Ngan Le, Vidhiwar Singh Rathour, Kashu Yamazaki, Khoa Luu, and Marios Savvides. Deep reinforcement learning in computer vision: a comprehensive survey. *Artificial Intelligence Review*, pages 1–87, 2022.
- [66] Stephen Adams, Tyler Cody, and Peter A Beling. A survey of inverse reinforcement learning. *Artificial Intelligence Review*, 55(6):4307–4346, 2022.

- [67] M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.
- [68] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE transactions on knowledge and data engineering*, 35(9):8934–8954, 2022.
- [69] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [70] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Jundong Li, and Zi Huang. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):335–355, 2023.
- [71] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.
- [72] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.
- [73] Tala Talaei Khoei, Hadjar Ould Slimane, and Naima Kaabouch. Deep learning: systematic review, models, challenges, and research directions. *Neural Computing and Applications*, 35(31):23103–23124, 2023.
- [74] Shams Forruque Ahmed, Md Sakib Bin Alam, Maruf Hassan, Mahtabin Rodela Rozbu, Taoseef Ishtiaq, Nazifa Rafa, M Mofijur, ABM Shawkat Ali, and Amir H Gandomi. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, 56(11):13521–13617, 2023.
- [75] Junyi Chai, Hao Zeng, Anming Li, and Eric WT Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021.
- [76] Wahab Khan, Ali Daud, Khairullah Khan, Shakoor Muhammad, and Rafiul Haq. Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. *Natural Language Processing Journal*, 4:100026, 2023.

- [77] VG Dhanya, A Subeesh, NL Kushwaha, Dinesh Kumar Vishwakarma, T Nagesh Kumar, G Ritika, and AN Singh. Deep learning based computer vision approaches for smart agricultural applications. *Artificial Intelligence in Agriculture*, 6:211–229, 2022.
- [78] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022.
- [79] Samuel Sousa and Roman Kern. How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing. *Artificial Intelligence Review*, 56(2):1427–1492, 2023.
- [80] Ivano Lauriola, Alberto Lavelli, and Fabio Aiolli. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, 470:443–456, 2022.
- [81] Rasoul Ameri, Chung-Chian Hsu, and Shahab S Band. A systematic review of deep learning approaches for surface defect detection in industrial applications. *Engineering Applications of Artificial Intelligence*, 130:107717, 2024.
- [82] Seyedehfaezeh Hosseininoorbin, Siamak Layeghy, Mohanad Sarhan, Raja Jurdak, and Marius Portmann. Exploring edge TPU for network intrusion detection in IoT. *Journal of Parallel and Distributed Computing*, 179:104712, 2023.
- [83] Bertalan Kovács, Anders D Henriksen, Jonathan Dyssel Stets, and Lazaros Nalpantidis. Object detection on TPU accelerated embedded devices. In *Computer Vision Systems: 13th International Conference, ICVS 2021, Virtual Event, September 22-24, 2021, Proceedings 13*, pages 82–92. Springer, 2021.
- [84] Maria Vakalopoulou, Stergios Christodoulidis, Ninon Burgos, Olivier Colliot, and Vincent Lepetit. Deep learning: basics and convolutional neural networks (CNNs). *Machine learning for brain disorders*, pages 77–115, 2023.
- [85] Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99, 2024.
- [86] Shuang Cong and Yang Zhou. A review of convolutional neural network architectures and their optimizations. *Artificial Intelligence Review*, 56(3): 1905–1969, 2023.

- [87] Lei Zhao and Zhonglin Zhang. A improved pooling method for convolutional neural networks. *Scientific Reports*, 14(1):1589, 2024.
- [88] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 503:92–108, 2022.
- [89] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.
- [90] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.
- [91] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.
- [92] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [93] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [94] Manar Abu Talib, Sohaib Majzoub, Qassim Nasir, and Dina Jamal. A systematic literature review on hardware implementation of artificial intelligence algorithms. *The Journal of Supercomputing*, 77:1897–1938, 2021.
- [95] Hyunbin Park and Shiho Kim. Hardware accelerator systems for artificial intelligence and machine learning. In *Advances in Computers*, volume 122, pages 51–95. Elsevier, 2021.
- [96] Gianmarco Dinelli, Gabriele Meoni, Emilio Rapuano, and Luca Fanucci. Advantages and limitations of fully on-chip CNN FPGA-based hardware accelerator. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2020.
- [97] Hyeonseok Hong, Dahun Choi, Namjoon Kim, Haein Lee, Beomjin Kang, Huibeom Kang, and Hyun Kim. Survey of convolutional neural network

- accelerators on field-programmable gate array platforms: architectures and optimization techniques. *Journal of Real-Time Image Processing*, 21(3):64, 2024.
- [98] Yu Wang, Yibing Liao, Jiamei Yang, Hui Wang, Yuxuan Zhao, Chengyu Zhang, Bende Xiao, Fei Xu, Yifan Gao, Mingzhu Xu, et al. An FPGA-based online reconfigurable CNN edge computing device for object detection. *Microelectronics Journal*, 137:105805, 2023.
- [99] José María Rodríguez Corral, Javier Civit-Masot, Francisco Luna-Perejón, Ignacio Díaz-Cano, Arturo Morgado-Estévez, and Manuel Domínguez-Morales. Energy efficiency in edge TPU vs. embedded GPU for computer-aided medical imaging segmentation and classification. *Engineering Applications of Artificial Intelligence*, 127:107298, 2024.
- [100] SM Mojahidul Ahsan, Tamzidul Hoque, Md Sakib Hasan, Mrityika Chowdhury, and Anurag Dhungel. Hardware Accelerators for Artificial Intelligence. In *AI-Enabled Electronic Circuit and System Design: From Ideation to Utilization*, pages 497–535. Springer, 2025.
- [101] Ashutosh Mishra, Jaekwang Cha, Hyunbin Park, and Shiho Kim. *Artificial Intelligence and Hardware Accelerators*. Springer, 2023.
- [102] Seifeddine Messaoud, Soulef Bouaafia, Amna Maraoui, Ahmed Chiheb Ammari, Lazhar Khriji, and Mohsen Machhout. Deep convolutional neural networks-based Hardware–Software on-chip system for computer vision application. *Computers & Electrical Engineering*, 98:107671, 2022.
- [103] Kh Shahriya Zaman, Mamun Bin Ibne Reaz, Sawal Hamid Md Ali, Ahmad Ashrif A Bakar, and Muhammad Enamul Hoque Chowdhury. Custom hardware architectures for deep learning on portable devices: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6068–6088, 2021.
- [104] Jian-Xun Mi, Jie Feng, and Ke-Yang Huang. Designing efficient convolutional neural network structure: A survey. *Neurocomputing*, 489:139–156, 2022.
- [105] Ehazaz Mustafa, Junaid Shuja, Faisal Rehman, Ahsan Riaz, Mohammed Maray, Muhammad Bilal, and Muhammad Khurram Khan. Deep Neural Networks meet computation offloading in mobile edge networks: Applications, taxonomy, and open issues. *Journal of Network and Computer Applications*, page 103886, 2024.

- [106] Shijun Zhang, Zuowei Shen, and Haizhao Yang. Neural network architecture beyond width and depth. *Advances in Neural Information Processing Systems*, 35: 5669–5681, 2022.
- [107] Feihao Chen and Jin Yeu Tsou. Assessing the effects of convolutional neural network architectural factors on model performance for remote sensing image classification: An in-depth investigation. *International Journal of Applied Earth Observation and Geoinformation*, 112:102865, 2022.
- [108] Chengmin Lin, Pengfei Yang, Quan Wang, Zeyu Qiu, Wenkai Lv, and Zhenyi Wang. Efficient and accurate compound scaling for convolutional neural networks. *Neural Networks*, 167:787–797, 2023.
- [109] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [110] Daniel Haase and Manuel Amthor. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14600–14609, 2020.
- [111] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [112] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.
- [113] Daquan Zhou, Qibin Hou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Rethinking bottleneck structure for efficient mobile network design. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III* 16, pages 680–697. Springer, 2020.
- [114] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. *arXiv preprint arXiv:1602.07360*, 2016.

- [115] Amir Gholami, Kiseok Kwon, Bichen Wu, Zizheng Tai, Xiangyu Yue, Peter Jin, Sicheng Zhao, and Kurt Keutzer. Squeezenext: Hardware-aware neural network design. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1638–1647, 2018.
- [116] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [117] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [118] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 552–568, 2018.
- [119] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9190–9200, 2019.
- [120] Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chunjing Xu, Enhua Wu, and Qi Tian. GhostNets on heterogeneous devices via cheap operations. *International Journal of Computer Vision*, 130(4):1050–1069, 2022.
- [121] Andrea Apicella, Francesco Donnarumma, Francesco Isgrò, and Roberto Prevete. A survey on modern trainable activation functions. *Neural Networks*, 138:14–32, 2021.
- [122] Renato J Cintra, Stefan Duffner, Christophe Garcia, and André Leite. Low-complexity approximate convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 29(12):5981–5992, 2018.
- [123] Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the selection of initialization and activation function for deep neural networks. *arXiv preprint arXiv:1805.08266*, 2018.

- [124] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [125] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.
- [126] Babak Rokh, Ali Azarpeyvand, and Alireza Khanteymoori. A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Transactions on Intelligent Systems and Technology*, 14(6):1–50, 2023.
- [127] Zonglei Lyu, Tong Yu, Fuxi Pan, Yilin Zhang, Jia Luo, Dan Zhang, Yiren Chen, Bo Zhang, and Guangyao Li. A survey of model compression strategies for object detection. *Multimedia tools and applications*, 83(16):48165–48236, 2024.
- [128] Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. The class imbalance problem in deep learning. *Machine Learning*, 113(7):4845–4901, 2024.
- [129] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [130] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.
- [131] Divya Saxena and Jiannong Cao. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.
- [132] André Tse, Lino Oliveira, and João Vinagre. Measuring Latency-Accuracy Trade-Offs in Convolutional Neural Networks. In *EPIA Conference on Artificial Intelligence*, pages 323–334. Springer, 2023.

- [133] Soumyalatha Naveen, Manjunath R Kounte, and Mohammed Riyaz Ahmed. Low latency deep learning inference model for distributed intelligent IoT edge clusters. *IEEE Access*, 9:160607–160621, 2021.
- [134] Manuel F Dolz, Sergio Barrachina, Héctor Martínez, Adrián Castelló, Antonio Maciá, Germán Fabregat, and Andrés E Tomás. Performance–energy trade-offs of deep learning convolution algorithms on ARM processors. *The Journal of Supercomputing*, 79(9):9819–9836, 2023.
- [135] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference*, pages 15–25. Springer, 2023.
- [136] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pages 291–326. Chapman and Hall/CRC, 2022.
- [137] Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [138] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7436–7456, 2021.
- [139] Sumeet Saurav, Prashant Gidde, Ravi Saini, and Sanjay Singh. Dual integrated convolutional neural network for real-time facial expression recognition in the wild. *The Visual Computer*, 38(3):1083–1096, 2022.
- [140] Alexander Menshchikov, Dmitrii Shadrin, Viktor Prutyantov, Daniil Lopatkin, Sergey Sosnin, Evgeny Tsykunov, Evgeny Iakovlev, and Andrey Somov. Real-time detection of hogweed: UAV platform empowered by deep learning. *IEEE Transactions on Computers*, 70(8):1175–1188, 2021.
- [141] Theodora Sanida and Minas Dasygenis. MiniTomatoNet: A lightweight CNN for tomato leaf disease recognition on heterogeneous FPGA-SoC. *The Journal of Supercomputing*, 80(15):21837–21866, 2024.

- [142] Ishana Attri, Lalit Kumar Awasthi, and Teek Parval Sharma. Machine learning in agriculture: a review of crop management applications. *Multimedia Tools and Applications*, 83(5):12875–12915, 2024.
- [143] Md Ismail Hossen, Mohammad Awrangjeb, Shirui Pan, and Abdullah Al Mamun. Transfer learning in agriculture: a review. *Artificial Intelligence Review*, 58(4): 97, 2025.
- [144] Abhishek Upadhyay, Narendra Singh Chandel, Krishna Pratap Singh, Subir Kumar Chakraborty, Balaji M Nandede, Mohit Kumar, A Subeesh, Konga Upendar, Ali Salem, and Ahmed Elbeltagi. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. *Artificial Intelligence Review*, 58(3):1–64, 2025.
- [145] André Abade, Paulo Afonso Ferreira, and Flavio de Barros Vidal. Plant diseases recognition on images using convolutional neural networks: A systematic review. *Computers and Electronics in Agriculture*, 185:106125, 2021.
- [146] Asif Iqbal Khan, SMK Quadri, Saba Banday, and Junaid Latief Shah. Deep diagnosis: A real-time apple leaf disease detection system based on deep learning. *computers and Electronics in Agriculture*, 198:107093, 2022.
- [147] Lin Chen, Zhonghao Chen, Yubing Zhang, Yunfei Liu, Ahmed I Osman, Mohamed Farghali, Jianmin Hua, Ahmed Al-Fatesh, Ikko Ihara, David W Rooney, et al. Artificial intelligence-based solutions for climate change: a review. *Environmental Chemistry Letters*, 21(5):2525–2557, 2023.
- [148] Deming Zhai, Ruifeng Shi, Junjun Jiang, and Xianming Liu. Rectified Meta-learning from Noisy Labels for Robust Image-based Plant Disease Classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s):1–17, 2022.
- [149] Dekera Kenneth Kwaghtyo and Christopher Ifeanyi Eke. Smart farming prediction models for precision agriculture: a comprehensive survey. *Artificial Intelligence Review*, 56(6):5729–5772, 2023.
- [150] Mamillapally Nagaraju and Priyanka Chawla. Systematic review of deep learning techniques in plant disease detection. *International journal of system assurance engineering and management*, 11(3):547–560, 2020.

- [151] Theodora Sanida, Argyrios Sideris, Maria V Sanida, and Minas Dasygenis. Tomato leaf disease identification via two-stage transfer learning approach. *Smart Agricultural Technology*, 5:100275, 2023.
- [152] Maria Vasiliki Sanida, Theodora Sanida, Argyrios Sideris, and Minas Dasygenis. An Efficient Hybrid CNN Classification Model for Tomato Crop Disease. *Technologies*, 11(1):10, 2023.
- [153] Pedro Henrique Gorni, Gustavo Rios de Lima, Lucia Maria de Oliveira Pereira, Kamille Daleck Spera, Allan de Marcos Lapaz, and Ana Cláudia Pacheco. Increasing plant performance, fruit production and nutritional value of tomato through foliar applied rutin. *Scientia Horticulturae*, 294:110755, 2022.
- [154] M Chithambarathanu and MK Jeyakumar. Survey on crop pest detection using deep learning and machine learning approaches. *Multimedia Tools and Applications*, 82(27):42277–42310, 2023.
- [155] Manjunatha Badiger and Jose Alex Mathew. Tomato plant leaf disease segmentation and multiclass disease detection using hybrid optimization enabled deep learning. *Journal of Biotechnology*, 374:101–113, 2023.
- [156] Kodjo Abel Odah, Sèton Calmette Ariane Houetohossou, Vinasetan Ratheil Houndji, and Romain Lucas Glèlè Kakaï. Machine learning techniques for tomato yield prediction: A comprehensive Analysis. *Smart Agricultural Technology*, page 101067, 2025.
- [157] Abbas Dehghani, Ali Kavari, Mahdi Kalbasi, and Keyvan RahimiZadeh. A new approach for design of an efficient FPGA-based reconfigurable convolver for image processing. *The Journal of Supercomputing*, pages 1–19, 2022.
- [158] Rami Ibrahim and M Omair Shafiq. Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. *ACM Computing Surveys*, 55(10): 1–37, 2023.
- [159] Hui-Wen Liu and Chung-An Shen. The design of efficient data flow and low-complexity architecture for a highly configurable cnn accelerator. *Circuits, Systems, and Signal Processing*, 42(8):4759–4783, 2023.
- [160] Iqbal H Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2(6):1–20, 2021.

- [161] Diana Susan Joseph, Pranav M Pawar, and Rahul Pramanik. Intelligent plant disease diagnosis using convolutional neural network: a review. *Multimedia Tools and Applications*, 82(14):21415–21481, 2023.
- [162] Christos Charisis and Dimitrios Argyropoulos. Deep learning-based instance segmentation architectures in agriculture: A review of the scopes and challenges. *Smart Agricultural Technology*, 8:100448, 2024.
- [163] Sumitra Samal and Vijayant Verma. Tomato plant leaf diseases detection and classification using an advanced self-developed CNN model. *Multimedia Tools and Applications*, 83(36):83837–83866, 2024.
- [164] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36, 2018.
- [165] Shuqiang Jiang, Weiqing Min, Yongqiang Lyu, and Linhu Liu. Few-shot food recognition via multi-view representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(3):1–20, 2020.
- [166] Zhineng Chen, Shanshan Ai, and Caiyan Jia. Structure-aware deep learning for product image classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1s):1–20, 2019.
- [167] Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. Model complexity of deep learning: A survey. *Knowledge and Information Systems*, 63:2585–2619, 2021.
- [168] Jou-An Chen, Wei Niu, Bin Ren, Yanzhi Wang, and Xipeng Shen. Survey: Exploiting data redundancy for optimization of deep learning. *ACM Computing Surveys*, 55(10):1–38, 2023.
- [169] Zhi-Hui Zhan, Jian-Yu Li, and Jun Zhang. Evolutionary deep learning: A survey. *Neurocomputing*, 483:42–58, 2022.
- [170] Yu Miao, Haiwei Dong, Jihad Mohamad Al Jaam, and Abdulmotaleb El Saddik. A deep learning system for recognizing facial expression in real-time. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2):1–20, 2019.

- [171] Shui-Hua Wang and Yu-Dong Zhang. DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2s):1–19, 2020.
- [172] Mehedi Masud, Mohammed F Alhamid, and Yin Zhang. A convolutional neural network model using weighted loss function to detect diabetic retinopathy. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s):1–16, 2022.
- [173] Wenjuan Gong, Yue Zhang, Wei Wang, Peng Cheng, and Jordi Gonzalez. Meta-MMFNet: Meta-learning-based multi-model fusion network for micro-expression recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(2):1–20, 2023.
- [174] Zahra Valadanjoz, Hassan Daryanavard, and Abbas Harifi. High-speed YOLOv4-tiny hardware accelerator for self-driving automotive. *The Journal of Supercomputing*, pages 1–26, 2023.
- [175] Kai Zeng, Qian Ma, Jia Wen Wu, Zhe Chen, Tao Shen, and Chenggang Yan. FPGA-based accelerator for object detection: A comprehensive survey. *The Journal of Supercomputing*, 78(12):14096–14136, 2022.
- [176] Sadiq M Sait, Aiman El-Maleh, Mohammad Altakrouri, and Ahmad Shawahna. Optimization of FPGA-based CNN accelerators using metaheuristics. *The Journal of Supercomputing*, 79(4):4493–4533, 2023.
- [177] Adib Hosseiny and Hadi Jahanirad. Hardware acceleration of YOLOv7-tiny using high-level synthesis tools. *Journal of Real-Time Image Processing*, 20(4): 75, 2023.
- [178] Maedeh Nobari and Hadi Jahanirad. FPGA-based implementation of deep neural network using stochastic computing. *Applied Soft Computing*, 137:110166, 2023.
- [179] David Castells-Rufas, Vinh Ngo, Juan Borrego-Carazo, Marc Codina, Carles Sanchez, Debora Gil, and Jordi Carrabina. A Survey of FPGA-Based Vision Systems for Autonomous Cars. *IEEE Access*, 10:132525–132563, 2022.
- [180] Sparsh Mittal. A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. *Journal of Systems Architecture*, 97:428–442, 2019.

- [181] Mohammed Brahimi, Kamel Boukhalfa, and Abdelouahab Moussaoui. Deep learning for tomato diseases: classification and symptoms visualization. *Applied Artificial Intelligence*, 31(4):299–315, 2017.
- [182] Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018.
- [183] Aanis Ahmad, Dharmendra Saraswat, and Aly El Gamal. A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools. *Smart Agricultural Technology*, 3:100083, 2023.
- [184] Yuzhen Lu and Sierra Young. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 178: 105760, 2020.
- [185] Poornima Singh Thakur, Pritee Khanna, Tanuja Sheorey, and Aparajita Ojha. Trends in vision-based machine learning techniques for plant disease identification: A systematic review. *Expert Systems with Applications*, 208: 118117, 2022.
- [186] Donald G Bailey. *Design for embedded image processing on FPGAs*. John Wiley & Sons, 2023.
- [187] Kyubaik Choi and Gerald E Sobelman. An efficient CNN accelerator for low-cost edge systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 21(4): 1–20, 2022.
- [188] Tsung-Han Tsai, Nai-Chieh Tung, and Chun-Yu Chen. An FPGA-Based Reconfigurable Convolutional Neural Network Accelerator for Tiny YOLO-V3. *Circuits, Systems, and Signal Processing*, pages 1–22, 2025.
- [189] Fang Liu, Heyuan Li, Wei Hu, and Yanxiang He. Review of neural network model acceleration techniques based on FPGA platforms. *Neurocomputing*, page 128511, 2024.
- [190] Mingqiang Huang, Ao Shen, Kai Li, Haoxiang Peng, Boyu Li, Yupeng Su, and Hao Yu. Edgellm: A highly efficient cpu-fpga heterogeneous edge accelerator for large language models. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2025.

- [191] Rajasekaran Thangaraj, S Anandamurugan, P Pandiyan, and Vishnu Kumar Kaliappan. Artificial intelligence in tomato leaf disease detection: a comprehensive review and discussion. *Journal of Plant Diseases and Protection*, 129(3):469–488, 2022.
- [192] Elhoucine Elfatimi, Recep Eryiğit, and Lahcen Elfatimi. Deep multi-scale convolutional neural networks for automated classification of multi-class leaf diseases in tomatoes. *Neural Computing and Applications*, pages 1–20, 2023.
- [193] Intan Nurma Yulita, Naufal Ariful Amri, and Akik Hidayat. Mobile Application for Tomato Plant Leaf Disease Detection Using a Dense Convolutional Network Architecture. *Computation*, 11(2):20, 2023.
- [194] Showmick Guha Paul, Al Amin Biswas, Arpa Saha, Md Sabab Zulfiker, Nadia Afrin Ritu, Ifrat Zahan, Mushfiquir Rahman, and Mohammad Ashraful Islam. A real-time application-based convolutional neural network approach for tomato leaf disease classification. *Array*, 19:100313, 2023.
- [195] Xibei Huang, Aibin Chen, Guoxiong Zhou, Xin Zhang, Jianwu Wang, Ning Peng, Na Yan, and Canhui Jiang. Tomato leaf disease detection system based on FC-SNDPN. *Multimedia Tools and Applications*, 82(2):2121–2144, 2023.
- [196] Sabbir Ahmed, Md Bakhtiar Hasan, Tasnim Ahmed, Md Redwan Karim Sony, and Md Hasanul Kabir. Less is more: lighter and faster deep neural architecture for tomato leaf disease classification. *IEEE Access*, 10:68868–68884, 2022.
- [197] N Aishwarya, NG Praveena, S Priyanka, and J Pramod. Smart farming for detection and identification of tomato plant diseases using light weight deep neural network. *Multimedia Tools and Applications*, pages 1–12, 2022.
- [198] Victor Gonzalez-Huitron, José A León-Borges, AE Rodriguez-Mata, Leonel Ernesto Amabilis-Sosa, Blenda Ramírez-Pereda, and Hector Rodriguez. Disease detection in tomato leaves via CNN with lightweight architectures implemented in Raspberry Pi 4. *Computers and Electronics in Agriculture*, 181:105951, 2021.
- [199] Amreen Abbas, Sweta Jain, Mahesh Gour, and Swetha Vankudothu. Tomato plant disease detection using transfer learning with C-GAN synthetic images. *Computers and Electronics in Agriculture*, 187:106279, 2021.

- [200] R Karthik, M Hariharan, Sundar Anand, Priyanka Mathikshara, Annie Johnson, and R Menaka. Attention embedded residual CNN for disease detection in tomato leaves. *Applied Soft Computing*, 86:105933, 2020.
- [201] Mohit Agarwal, Abhishek Singh, Siddhartha Arjaria, Amit Sinha, and Suneet Gupta. ToLeD: Tomato leaf disease detection using convolution neural network. *Procedia Computer Science*, 167:293–301, 2020.
- [202] Nasik Muhammad Nafi and William H Hsu. Addressing class imbalance in image-based plant disease detection: Deep generative vs. sampling-based approaches. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 243–248. IEEE, 2020.
- [203] Parul Sharma, Yash Paul Singh Berwal, and Wiqas Ghai. Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. *Information Processing in Agriculture*, 7(4):566–574, 2020.
- [204] Shanwen Zhang, Wenzhun Huang, and Chuanlei Zhang. Three-channel convolutional neural networks for vegetable leaf disease recognition. *Cognitive Systems Research*, 53:31–41, 2019.
- [205] Qimei Wang, Feng Qi, Minghe Sun, Jianhua Qu, and Jie Xue. Identification of tomato disease types and detection of infected areas based on deep convolutional neural networks and object detection techniques. *Computational intelligence and neuroscience*, 2019:15, 2019.
- [206] Robert G. de Luna, Elmer P. Dadios, and Argel A. Bandala. Automated Image Capturing System for Deep Learning-based Tomato Plant Leaf Disease Detection and Recognition. In *TENCON 2018 - 2018 IEEE Region 10 Conference*, pages 1414–1419, Jeju, Korea (South), 2018. IEEE.
- [207] Melike Sardogan, Adem Tuncer, and Yunus Ozen. Plant leaf disease detection and classification based on CNN with LVQ algorithm. In *2018 3rd international conference on computer science and engineering (UBMK)*, pages 382–385, Sarajevo, Bosnia and Herzegovina, 2018. IEEE.
- [208] Keke Zhang, Qiufeng Wu, Anwang Liu, and Xiangyan Meng. Can deep learning identify tomato leaf disease? *Advances in multimedia*, 2018:10, 2018.

- [209] Alvaro F Fuentes, Sook Yoon, Jaesu Lee, and Dong Sun Park. High-performance deep neural network-based tomato plant diseases and pests diagnosis system with refinement filter bank. *Frontiers in plant science*, 9:1162, 2018.
- [210] Prajwala Tm, Alla Pranathi, Kandiraju SaiAshritha, Nagaratna B. Chittaragi, and Shashidhar G. Koolagudi. Tomato Leaf Disease Detection Using Convolutional Neural Networks. In *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pages 1–5, Noida, India, 2018. IEEE.
- [211] Weihui Zeng, Miao Li, Jian Zhang, Lei Chen, Sisi Fang, and Jingxian Wang. High-order residual convolutional neural network for robust crop disease recognition. In *Proceedings of the 2nd international conference on computer science and application engineering*, pages 1–5, 2018.
- [212] Kyosuke Yamamoto, Takashi Togami, and Norio Yamaguchi. Super-resolution of plant disease images for the acceleration of image-based phenotyping and vigor diagnosis in agriculture. *Sensors*, 17(11):2557, 2017.
- [213] Aravind Krishnaswamy Rangarajan, Raja Purushothaman, and Anirudh Ramesh. Tomato crop disease classification using pre-trained deep learning algorithm. *Procedia computer science*, 133:1040–1047, 2018.
- [214] Alvaro Fuentes, Sook Yoon, Sang Cheol Kim, and Dong Sun Park. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, 17(9):2022, 2017.
- [215] Halil Durmuş, Ece Olcay Güneş, and Mürvet Kırıcı. Disease detection on the leaves of the tomato plants by using deep learning. In *2017 6th International Conference on Agro-Geoinformatics*, pages 1–5, Fairfax, VA, USA, 2017. IEEE.
- [216] David Hughes, Marcel Salathé, et al. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015.
- [217] Giorgos Armeniakos, Georgios Zervakis, Dimitrios Soudris, and Jörg Henkel. Hardware approximate techniques for deep neural network accelerators: A survey. *ACM Computing Surveys*, 55(4):1–36, 2022.
- [218] Xiao Wu, Yufei Ma, Meiqi Wang, and Zhongfeng Wang. A flexible and efficient FPGA accelerator for various large-scale and lightweight CNNs. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 69(3):1185–1198, 2021.

- [219] Wenjin Huang, Huangtao Wu, Qingkun Chen, Conghui Luo, Shihao Zeng, Tianrui Li, and Yihua Huang. FPGA-based high-throughput CNN hardware accelerator with high computing resource utilization ratio. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):4069–4083, 2021.
- [220] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
- [221] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020.
- [222] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.
- [223] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.
- [224] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.
- [225] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.
- [226] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [227] Weizhi Xu, Yintai Sun, Shengyu Fan, Hui Yu, and Xin Fu. Accelerating convolutional neural network by exploiting sparsity on GPUs. *ACM Transactions on Architecture and Code Optimization*, 20(3):1–26, 2023.
- [228] Anouar Nechi, Lukas Groth, Saleh Mulhem, Farhad Merchant, Rainer Buchty, and Mladen Berekovic. FPGA-based Deep Learning Inference Accelerators: Where Are We Standing? *ACM Transactions on Reconfigurable Technology and Systems*, 16(4):1–32, 2023.

- [229] Tamador Mohaidat and Kasem Khalil. A Survey on Neural Network Hardware Accelerators. *IEEE Transactions on Artificial Intelligence*, 2024.
- [230] Xilinx. PYNQ: Python productivity for ZYNQ, 2018. URL <http://www.pynq.io/>. Last Accessed 20 May 2025.
- [231] Xilinx. Vivado HLS Optimization Methodology Guide, 2018. URL <https://docs.xilinx.com/v/u/en-US/ug1270-vivado-hls-opt-methodology-guide>. Last Accessed 25 May 2025.
- [232] Theodora Sanida, Dimitris Tsiktisiris, Argyrios Sideris, and Minas Dasygenis. A Heterogeneous Lightweight Network for Plant Disease Classification. In *2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, pages 1–4. IEEE, 2021.
- [233] Theodora Sanida and Minas Dasygenis. A novel lightweight CNN for chest X-ray-based lung disease identification on heterogeneous embedded system. *Applied Intelligence*, 54(6):4756–4780, 2024.
- [234] Angela J Frank, Ibiayi Dagogo-Jack, Ioana A Dobre, Sarah Tait, Lana Schumacher, Florian J Fintelmann, Leah M Fingerman, Florence K Keane, and Sydney B Montesi. Management of Lung Cancer in the Patient with Interstitial Lung Disease. *The Oncologist*, 28(1):12–22, 2023.
- [235] Shagun Sharma and Kalpna Guleria. A systematic literature review on deep learning approaches for pneumonia detection using chest X-ray images. *Multimedia Tools and Applications*, 83(8):24101–24151, 2024.
- [236] Peter M George, Paolo Spagnolo, Michael Kreuter, Goksel Altinisik, Martina Bonifazi, Fernando J Martinez, Philip L Molyneaux, Elisabetta A Renzoni, Luca Richeldi, Sara Tomassetti, et al. Progressive fibrosing interstitial lung disease: clinical uncertainties, consensus recommendations, and research priorities. *The lancet Respiratory medicine*, 8(9):925–934, 2020.
- [237] Marlies Wijsenbeek, Atsushi Suzuki, and Toby M Maher. Interstitial lung diseases. *The Lancet*, 400(10354):769–786, 2022.
- [238] Qusay Shihab Hamad, Hussein Samma, and Shahrel Azmin Suandi. Feature selection of pre-trained shallow CNN using the QLESCA optimizer: COVID-19 detection as a case study. *Applied Intelligence*, pages 1–23, 2023.

- [239] Mainak Chakraborty, Sunita Vikrant Dhavale, and Jitendra Ingole. Corona-Nidaan: lightweight deep convolutional neural network for chest X-Ray based COVID-19 infection detection. *Applied Intelligence*, 51(5): 3026–3043, 2021.
- [240] Raman Sharma, Madhulata Agarwal, Mayank Gupta, Somyata Somendra, and Shailendra K Saxena. Clinical characteristics and differential clinical diagnosis of novel coronavirus disease 2019 (COVID-19). *Coronavirus Disease 2019 (COVID-19) Epidemiology, Pathogenesis, Diagnosis, and Therapeutics*, pages 55–70, 2020.
- [241] Christian Holm Hansen, Ida Rask Moustsen-Helms, Morten Rasmussen, Bolette Søborg, Henrik Ullum, and Palle Valentiner-Branth. Short-term effectiveness of the XBB. 1.5 updated COVID-19 vaccine against hospitalisation in Denmark: a national cohort study. *The Lancet Infectious Diseases*, 24(2):e73–e74, 2024.
- [242] Md Alamin Talukder, Md Abu Layek, Mohsin Kazi, Md Ashraf Uddin, and Sunil Aryal. Empowering COVID-19 detection: Optimizing performance through fine-tuned efficientnet deep learning architecture. *Computers in Biology and Medicine*, 168:107789, 2024.
- [243] Jayesh Beladiya, Anup Kumar, Yogesh Vasava, Krupanshu Parmar, Dipanshi Patel, Sandip Patel, Sandip Dholakia, Devang Sheth, Sai HS Boddu, and Chirag Patel. Safety and efficacy of COVID-19 vaccines: A systematic review and meta-analysis of controlled and randomized clinical trials. *Reviews in medical virology*, 34(1):e2507, 2024.
- [244] Ebenezer Jangam, Aaron Antonio Dias Barreto, and Chandra Sekhara Rao Annavarapu. Automatic detection of COVID-19 from chest CT scan and chest X-Rays images using deep learning, transfer learning and stacking. *Applied Intelligence*, pages 1–17, 2022.
- [245] Sabina A Guler and Tamera J Corte. Interstitial lung disease in 2020: a history of progress. *Clinics in chest medicine*, 42(2):229–239, 2021.
- [246] Shivraj Popat Jadhav, Himmat Singh, Salman Hussain, Ritu Gilhotra, Anurag Mishra, Parteek Prasher, Anand Krishnan, and Gaurav Gupta. Introduction to lung diseases. *Targeting cellular Signalling pathways in lung diseases*, pages 1–25, 2021.

- [247] DR Sarvamangala and Raghavendra V Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence*, 15(1):1–22, 2022.
- [248] X Zhao, B Liu, Y Yu, X Wang, Y Du, J Gu, and XJCR Wu. The characteristics and clinical value of chest CT images of novel coronavirus pneumonia. *Clinical radiology*, 75(5):335–340, 2020.
- [249] Subrato Bharati, Prajoy Podder, and M Rubaiyat Hossain Mondal. Hybrid deep learning for detecting lung diseases from X-ray images. *Informatics in Medicine Unlocked*, 20:100391, 2020.
- [250] Simon LF Walsh, Stephen M Humphries, Athol U Wells, and Kevin K Brown. Imaging research in fibrotic lung disease; applying deep learning to unsolved problems. *The Lancet Respiratory Medicine*, 8(11):1144–1153, 2020.
- [251] Yiwen Xu, Tariq M Khan, Yang Song, and Erik Meijering. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artificial Intelligence Review*, 58(3):1–78, 2025.
- [252] GV Eswara Rao, B Rajitha, Parvathaneni Naga Srinivasu, Muhammad Fazal Ijaz, and Marcin Woźniak. Hybrid framework for respiratory lung diseases detection based on classical CNN and quantum classifiers from chest X-rays. *Biomedical Signal Processing and Control*, 88:105567, 2024.
- [253] Mohan Bhandari, Tej Bahadur Shahi, Birat Siku, and Arjun Neupane. Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI. *Computers in Biology and Medicine*, 150:106156, 2022.
- [254] KC Santosh, Siva Allu, Sivaramakrishnan Rajaraman, and Sameer Antani. Advances in deep learning for tuberculosis screening using chest X-rays: the last 5 years review. *Journal of Medical Systems*, 46(11):82, 2022.
- [255] Ahmed Iqbal, Muhammad Usman, and Zohair Ahmed. Tuberculosis chest X-ray detection using CNN-based hybrid segmentation and classification approach. *Biomedical Signal Processing and Control*, 84:104667, 2023.
- [256] Lauren K Troy, Christopher Grainge, Tamera J Corte, Jonathan P Williamson, Michael P Vallely, Wendy A Cooper, Annabelle Mahar, Jeffrey L Myers, Simon Lai, Ellie Mulyadi, et al. Diagnostic accuracy of transbronchial lung cryobiopsy

- for interstitial lung disease diagnosis (COLDICE): a prospective, comparative study. *The lancet Respiratory medicine*, 8(2):171–181, 2020.
- [257] Ole Hilberg, Anna-Maria Hoffmann-Vold, Vanessa Smith, Demosthenes Bouros, Maritta Kilpeläinen, Julien Guiot, Antonio Morais, Susana Clemente, Zoe Daniil, Despina Papakosta, et al. Epidemiology of interstitial lung diseases and their progressive-fibrosing behaviour in six European countries. *ERJ open research*, 8(1), 2022.
- [258] James P Allinson, Shoaib Afzal, Yunus Çolak, Debbie Jarvis, Helena Backman, Maarten van den Berge, H Marike Boezen, Marie-Kathrin Breyer, Robab Breyer-Kohansal, Guy Brusselle, et al. Changes in lung function in European adults born between 1884 and 1996 and implications for the diagnosis of lung disease: a cross-sectional analysis of ten population-based studies. *The Lancet Respiratory Medicine*, 10(1):83–94, 2022.
- [259] Siqin He, Jiajia Gui, Kun Xiong, Meiwan Chen, Huile Gao, and Yao Fu. A roadmap to pulmonary delivery strategies for the treatment of infectious lung diseases. *Journal of nanobiotechnology*, 20(1):101, 2022.
- [260] Mairi MacLeod, Alberto Papi, Marco Contoli, Bianca Beghé, Bartolome R Celli, Jadwiga A Wedzicha, and Leonardo M Fabbri. Chronic obstructive pulmonary disease exacerbation fundamentals: diagnosis, treatment, prevention and disease impact. *Respirology*, 26(6):532–551, 2021.
- [261] Natascha Sommer, Hossein A Ghofrani, Oleg Pak, Sebastien Bonnet, Steve Provencher, Olivier Sitbon, Stephan Rosenkranz, Marius M Hoeper, and David G Kiely. Current and future treatments of pulmonary arterial hypertension. *British journal of pharmacology*, 178(1):6–30, 2021.
- [262] Naru Zhang, Lili Wang, Xiaoqian Deng, Ruiying Liang, Meng Su, Chen He, Lanfang Hu, Yudan Su, Jing Ren, Fei Yu, et al. Recent advances in the detection of respiratory virus infection in humans. *Journal of medical virology*, 92(4):408–417, 2020.
- [263] Yubo Shao, Hong Zhang, Qi Shi, Yongjun Wang, and Qianqian Liang. Clinical prediction models of rheumatoid arthritis and its complications: focus on cardiovascular disease and interstitial lung disease. *Arthritis Research & Therapy*, 25(1):159, 2023.

- [264] Zaid Zoumot, Maria-Fernanda Bonilla, Ali S Wahla, Irfan Shafiq, Mateen Uzbeck, Rania M El-Lababidi, Fadi Hamed, Mohamed Abuzakouk, and Mahmoud ElKaissi. Pulmonary cavitation: an under-recognized late complication of severe COVID-19 lung disease. *BMC Pulmonary Medicine*, 21: 1–8, 2021.
- [265] Ankit Bharat, Tiago N Machuca, Melissa Querrey, Chitaru Kurihara, Rafael Garza-Castillon, Samuel Kim, Adwaiy Manerikar, Andres Pelaez, Mauricio Pipkin, Abbas Shahmohammadi, et al. Early outcomes after lung transplantation for severe COVID-19: a series of the first consecutive cases from four countries. *The lancet respiratory medicine*, 9(5):487–497, 2021.
- [266] Mikko Nuutinen and Riikka-Leena Leskelä. Systematic review of the performance evaluation of clinicians with or without the aid of machine learning clinical decision support system. *Health and Technology*, pages 1–14, 2023.
- [267] Paolo Spagnolo, Christopher J Ryerson, Rachel Putman, Justin Oldham, Margaret Salisbury, Nicola Sverzellati, Claudia Valenzuela, Sabina Guler, Steve Jones, Marlies Wijsenbeek, et al. Early diagnosis of fibrotic interstitial lung disease: challenges and opportunities. *The Lancet Respiratory Medicine*, 9(9): 1065–1076, 2021.
- [268] Nussair Adel Hroub, Ali Nader Alsannaa, Maad Alowaifeer, Motaz Alfarraj, and Emmanuel Okafor. Explainable deep learning diagnostic system for prediction of lung disease from medical images. *Computers in Biology and Medicine*, 170: 108012, 2024.
- [269] Abdennabi Morchid, Marouane Marhoun, Rachid El Alami, and Bensalem Boukili. Intelligent detection for sustainable agriculture: a review of IoT-based embedded systems, cloud platforms, DL, and ML for plant disease detection. *Multimedia Tools and Applications*, 83(28):70961–71000, 2024.
- [270] Sreeparna Das, Ishan Ayus, and Deepak Gupta. A comprehensive review of COVID-19 detection with machine learning and deep learning techniques. *Health and Technology*, pages 1–14, 2023.
- [271] Hadeer A Helaly, Mahmoud Badawy, and Amira Y Haikal. A review of deep learning approaches in clinical and healthcare systems based on medical image analysis. *Multimedia Tools and Applications*, pages 1–42, 2023.

- [272] Balaji Annamalai, Prabakeran Saravanan, and Indumathi Varadharajan. ABOA-CNN: auction-based optimization algorithm with convolutional neural network for pulmonary disease prediction. *Neural Computing and Applications*, 35(10):7463–7474, 2023.
- [273] RKPMTKR Rajagopal, R Karthick, P Meenalochini, and T Kalaichelvi. Deep Convolutional Spiking Neural Network optimized with Arithmetic optimization algorithm for lung disease detection using chest X-ray images. *Biomedical Signal Processing and Control*, 79:104197, 2023.
- [274] Ishak Pacal and Serhat Kılıcarslan. Deep learning-based approaches for robust classification of cervical cancer. *Neural Computing and Applications*, 35(25):18813–18828, 2023.
- [275] Ajay Sharma and Pramod Kumar Mishra. DRI-UNet: dense residual-inception UNet for nuclei identification in microscopy cell images. *Neural Computing and Applications*, pages 1–34, 2023.
- [276] Cafer Budak, Vasfiye Mençik, and Osman Varışlı. Online diagnosis of COVID-19 from chest radiography images by using deep learning algorithms. *Neural Computing and Applications*, pages 1–18, 2023.
- [277] Md Kawsher Mahbub, Milon Biswas, Loveleen Gaur, Fayadh Alenezi, and KC Santosh. Deep features to detect pulmonary abnormalities in chest X-rays due to infectious diseaseX: Covid-19, pneumonia, and tuberculosis. *Information Sciences*, 592:389–401, 2022.
- [278] Tamara Vojinovic, Ilaria Cavazzana, Paolo Ceruti, Micaela Fredi, Denise Modina, Marialma Berlendis, and Franco Franceschini. Predictive features and clinical presentation of interstitial lung disease in inflammatory myositis. *Clinical Reviews in Allergy & Immunology*, 60:87–94, 2021.
- [279] Michael H Cho, Brian D Hobbs, and Edwin K Silverman. Genetics of chronic obstructive pulmonary disease: understanding the pathobiology and heterogeneity of a complex disorder. *The Lancet Respiratory Medicine*, 10(5):485–496, 2022.
- [280] Amira Ehtioui and Yassine Ben Ayed. Automated detection of COVID-19 based on transfer learning. *Multimedia Tools and Applications*, pages 1–21, 2023.

- [281] Vinayakumar Ravi, Vasundhara Acharya, and Mamoun Alazab. A multichannel EfficientNet deep learning-based stacking ensemble approach for lung disease detection using chest X-ray images. *Cluster Computing*, 26(2):1181–1203, 2023.
- [282] Saad I Nafisah and Ghulam Muhammad. Tuberculosis detection in chest radiograph using convolutional neural network architecture and explainable artificial intelligence. *Neural Computing and Applications*, 36(1):111–131, 2024.
- [283] Hassaan Malik, Tayyaba Anees, Muizzud Din, and Ahmad Naeem. CDC_Net: Multi-classification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung Cancer, and tuberculosis using chest X-rays. *Multimedia Tools and Applications*, 82(9):13855–13880, 2023.
- [284] Hassaan Malik and Tayyaba Anees. BDCNet: Multi-classification convolutional neural network model for classification of COVID-19, pneumonia, and lung cancer from chest radiographs. *Multimedia Systems*, 28(3):815–829, 2022.
- [285] Mukesh Mann, Rakesh P Badoni, Harsh Soni, Mohammed Al-Shehri, Aman Chandra Kaushik, and Dong-Qing Wei. Utilization of deep convolutional neural networks for accurate chest X-ray diagnosis and disease detection. *Interdisciplinary Sciences: Computational Life Sciences*, 15(3):374–392, 2023.
- [286] Himadri Mukherjee, Subhankar Ghosh, Ankita Dhar, Sk Md Obaidullah, KC Santosh, and Kaushik Roy. Deep neural network to detect COVID-19: one architecture for both CT Scans and Chest X-rays. *Applied Intelligence*, 51: 2777–2789, 2021.
- [287] Joan Jonathan and Alcardo Alex Barakabitze. ML technologies for diagnosing and treatment of tuberculosis: a survey. *Health and Technology*, 13(1):17–33, 2023.
- [288] Zaid Abdi Alkareem Alyasseri, Mohammed Azmi Al-Betar, Iyad Abu Doush, Mohammed A Awadallah, Ammar Kamal Abasi, Sharif Naser Makhadmeh, Osama Ahmad Alomari, Karrar Hameed Abdulkareem, Afzan Adam, Robertas Damasevicius, et al. Review on COVID-19 diagnosis models based on machine learning and deep learning approaches. *Expert systems*, 39(3):e12759, 2022.
- [289] Tejalal Choudhary, Shubham Gujar, Anurag Goswami, Vipul Mishra, and Tapas Badal. Deep learning-based important weights-only transfer learning approach for COVID-19 CT-scan classification. *Applied Intelligence*, 53(6):7201–7215, 2023.

- [290] Abobaker Mohammed Qasem Farhan and Shangming Yang. Automatic lung disease classification from the chest X-ray images using hybrid deep learning algorithm. *Multimedia Tools and Applications*, pages 1–27, 2023.
- [291] Ding-Xuan Zhou. Theory of deep convolutional neural networks: Downsampling. *Neural Networks*, 124:319–327, 2020.
- [292] Gerosh Shibu George, Pratyush Raj Mishra, Panav Sinha, and Manas Ranjan Prusty. COVID-19 detection on chest X-ray images using Homomorphic Transformation and VGG inspired deep convolutional neural network. *Biocybernetics and Biomedical Engineering*, 43(1):1–16, 2023.
- [293] Scott Askin, Denis Burkhalter, Gilda Calado, and Samar El Dakrouni. Artificial Intelligence Applied to clinical trials: opportunities and challenges. *Health and Technology*, 13(2):203–213, 2023.
- [294] Arun Sharma, Sheeba Rani, and Dinesh Gupta. Artificial intelligence-based classification of chest X-ray images into COVID-19 and other infectious diseases. *International journal of biomedical imaging*, 2020:1–10, 2020.
- [295] Seyed Hesamoddin Hosseini, Reza Monsefi, and Shabnam Shadroo. Deep learning applications for lung cancer diagnosis: a systematic review. *Multimedia Tools and Applications*, 83(5):14305–14335, 2024.
- [296] Scott J Adams, Emily Stone, David R Baldwin, Rozemarijn Vliegenthart, Pyng Lee, and Florian J Fintelmann. Lung cancer screening. *The Lancet*, 401(10374): 390–408, 2023.
- [297] Theodoros Karampitsakos, Paolo Spagnolo, Nesrin Mogulkoc, Wim A Wuyts, Sara Tomassetti, Elisabeth Bendstrup, Maria Molina-Molina, Effrosyni D Manali, Ömer Selim Unat, Francesco Bonella, et al. Lung cancer in patients with idiopathic pulmonary fibrosis: A retrospective multicentre study in Europe. *Respirology*, 28(1):56–65, 2023.
- [298] Jawad Rasheed, Alaa Ali Hameed, Chawki Djeddi, Akhtar Jamil, and Fadi Al-Turjman. A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. *Interdisciplinary Sciences: Computational Life Sciences*, 13:103–117, 2021.

- [299] Fatih Demir, Abdulkadir Sengur, and Varun Bajaj. Convolutional neural networks based efficient approach for classification of lung diseases. *Health information science and systems*, 8:1–8, 2020.
- [300] Tawsifur Rahman, Muhammad EH Chowdhury, Amith Khandakar, Zaid Bin Mahbub, Md Sakib Abrar Hossain, Abraham Alhatou, Eynas Abdalla, Sreekumar Muthiyil, Khandaker Farzana Islam, Saad Bin Abul Kashem, et al. BIO-CXRNET: a robust multimodal stacking machine learning technique for mortality risk prediction of COVID-19 patients using chest X-ray images and clinical data. *Neural Computing and Applications*, pages 1–23, 2023.
- [301] Farhan Ullah, Jihoon Moon, Hamad Naeem, and Sohail Jabbar. Explainable artificial intelligence approach in combating real-time surveillance of COVID19 pandemic from CT scan and X-ray images using ensemble model. *The Journal of Supercomputing*, 78(17):19246–19271, 2022.
- [302] Zhicheng Jiao, Ji Whae Choi, Kasey Halsey, Thi My Linh Tran, Ben Hsieh, Dongcui Wang, Feyisope Eweje, Robin Wang, Ken Chang, Jing Wu, et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. *The Lancet Digital Health*, 3(5):e286–e294, 2021.
- [303] Sergio Varela-Santos and Patricia Melin. A new modular neural network approach with fuzzy response integration for lung disease classification based on multiple objective feature optimization in chest X-ray images. *Expert Systems with Applications*, 168:114361, 2021.
- [304] Bomi Kim, Stephan Romeijn, Mark van Buchem, Mohammad Hosein Rezazade Mehrizi, and Willem Grootjans. A holistic approach to implementing artificial intelligence in radiology. *Insights Into Imaging*, 15(1):22, 2024.
- [305] Alireza Tahamtan and Abdollah Ardebili. Real-time RT-PCR in COVID-19 detection: issues affecting the results. *Expert review of molecular diagnostics*, 20(5):453–454, 2020.
- [306] Xin Li, Lei Zhang, Jingsi Yang, and Fei Teng. Role of artificial intelligence in medical image analysis: A review of current trends and future directions. *Journal of Medical and Biological Engineering*, 44(2):231–243, 2024.
- [307] Elias P Medeiros, Marcos R Machado, Emannuel Diego G de Freitas, Daniel S da Silva, and Renato William R de Souza. Applications of machine learning

- algorithms to support COVID-19 diagnosis using X-rays data information. *Expert Systems With Applications*, 238:122029, 2024.
- [308] Yousif Ahmed Al-Naser. The impact of artificial intelligence on radiography as a profession: A narrative review. *Journal of Medical Imaging and Radiation Sciences*, 54(1):162–166, 2023.
- [309] Shariq Suleman, Sudheesh K Shukla, Nitesh Malhotra, Shikandar D Bukkitgar, Nagaraj P Shetti, Roberto Pilloton, Jagriti Narang, Yen Nee Tan, and Tejraj M Aminabhavi. Point of care detection of COVID-19: Advancement in biosensing and diagnostic methods. *Chemical Engineering Journal*, 414:128759, 2021.
- [310] MD Abdur Rahman, M Shamim Hossain, Nabil A Alrajeh, and BB Gupta. A multimodal, multimedia point-of-care deep learning framework for COVID-19 diagnosis. *ACM Transactions on Multimedia Computing Communications and Applications*, 17(1s):1–24, 2021.
- [311] Golla Madhu, Sandeep Kautish, Yogita Gupta, G Nagachandrika, Soly Mathew Biju, and Manoj Kumar. XCovNet: An optimized xception convolutional neural network for classification of COVID-19 from point-of-care lung ultrasound images. *Multimedia Tools and Applications*, 83(11):33653–33674, 2024.
- [312] Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, et al. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE transactions on medical imaging*, 39(8):2676–2687, 2020.
- [313] Navchetan Awasthi, Aveen Dayal, Linga Reddy Cenkeramaddi, and Phaneendra K Yalavarthy. Mini-COVIDNet: efficient lightweight deep neural network for ultrasound based point-of-care detection of COVID-19. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 68(6): 2023–2037, 2021.
- [314] Suchanat Boonkaew, Ilhoon Jang, Eka Noviana, Weena Siangproh, Orawon Chailapakul, and Charles S Henry. Electrochemical paper-based analytical device for multiplexed, point-of-care detection of cardiovascular disease biomarkers. *Sensors and Actuators B: Chemical*, 330:129336, 2021.
- [315] Toru Kameda, Yoshihiro Mizuma, Hayato Taniguchi, Masato Fujita, and Nobuyuki Taniguchi. Point-of-care lung ultrasound for the assessment of

- pneumonia: a narrative review in the COVID-19 era. *Journal of Medical Ultrasonics*, 48:31–43, 2021.
- [316] Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, 35(22):16157–16173, 2023.
- [317] Danilo Buonsenso, Antonio Chiaretti, Antonietta Curatola, Rosa Morello, Martina Giacalone, and Niccolò Parri. Pediatrician performed point-of-care ultrasound for the detection of ingested foreign bodies: case series and review of the literature. *Journal of Ultrasound*, 24:107–114, 2021.
- [318] F Dornaika and V Truong Hoang. A novel graph-based multi-view spectral clustering: application to X-ray image analysis for COVID-19 recognition. *Neural Computing and Applications*, pages 1–11, 2023.
- [319] Amirhosein Jahanshir, Shida Mohajer Moghari, Ayat Ahmadi, Pejman Z Moghadam, and Maryam Bahreini. Value of point-of-care ultrasonography compared with computed tomography scan in detecting potential life-threatening conditions in blunt chest trauma patients. *The Ultrasound Journal*, 12:1–10, 2020.
- [320] Asmaa H Rabie, Alaa M Mohamed, MA Abo-Elsoud, and Ahmed I Saleh. A new Covid-19 diagnosis strategy using a modified KNN classifier. *Neural Computing and Applications*, pages 1–25, 2023.
- [321] Mingjing Wang, Xiaoping Li, Long Chen, and Huiling Chen. Medical machine learning based on multiobjective evolutionary algorithm using learning decomposition. *Expert Systems with Applications*, 216:119450, 2023.
- [322] Kalaimani Markandan, Yong Wei Tiong, Revathy Sankaran, Sakthinathan Subramanian, Uma Devi Markandan, Vishal Chaudhary, Arshid Numan, Mohammad Khalid, and Rashmi Walvekar. Emergence of infectious diseases and role of advanced nanomaterials in point-of-care diagnostics: a review. *Biotechnology and Genetic Engineering Reviews*, 40(4):3438–3526, 2024.
- [323] Ronil J Rath, Jack O Herrington, Muhammad Adeel, Firat Güder, Fariba Dehghani, and Syamak Farajikhah. Ammonia detection: A pathway towards potential point-of-care diagnostics. *Biosensors and Bioelectronics*, 251:116100, 2024.

- [324] Qiya Gao and Shuang Li. Intelligent point of care testing for medicine diagnosis. *Interdisciplinary Medicine*, 2(1):e20230031, 2024.
- [325] Rakesh Kumar, Pooja Kumbharkar, Sandeep Vanam, and Sanjeev Sharma. Medical images classification using deep learning: a survey. *Multimedia Tools and Applications*, pages 1–46, 2023.
- [326] John R Hurst, MeiLan K Han, Barinder Singh, Sakshi Sharma, Gagandeep Kaur, Enrico de Nigris, Ulf Holmgren, and Mohd Kashif Siddiqui. Prognostic risk factors for moderate-to-severe exacerbations in patients with chronic obstructive pulmonary disease: a systematic literature review. *Respiratory research*, 23(1):213, 2022.
- [327] Theodoros Karampitsakos, Brenda M Juan-Guardela, Argyris Tzouvelekis, and Jose D Herazo-Maya. Precision medicine advances in idiopathic pulmonary fibrosis. *EBioMedicine*, 95, 2023.
- [328] Meghavi Rana and Megha Bhushan. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimedia Tools and Applications*, 82(17):26731–26769, 2023.
- [329] Muhammad Ajmal, Muhammad Attique Khan, Tallha Akram, Abdullah Alqahtani, Majed Alhaisoni, Ammar Armghan, Sara A Althubiti, and Fayadh Alenezi. BF2SkNet: Best deep learning features fusion-assisted framework for multiclass skin lesion classification. *Neural Computing and Applications*, pages 1–17, 2022.
- [330] CR Asswin, Dharshan Kumar KS, Avinash Dora, Vinayakumar Ravi, V Sowmya, EA Gopalakrishnan, KP Soman, et al. Transfer learning approach for pediatric pneumonia diagnosis using channel attention deep CNN architectures. *Engineering Applications of Artificial Intelligence*, 123:106416, 2023.
- [331] Alyaa Hamel Sfayyih, Nasri Sulaiman, and Ahmad H Sabry. A review on lung disease recognition by acoustic signal analysis with deep learning networks. *Journal of big Data*, 10(1):101, 2023.
- [332] Pedro Celard, EL Iglesias, JM Sorribes-Fdez, Rubén Romero, A Seara Vieira, and L Borrajo. A survey on deep learning applied to medical images: from simple artificial neural networks to generative models. *Neural Computing and Applications*, 35(3):2291–2323, 2023.

- [333] Zengchen Yu, Ke Wang, Zhibo Wan, Shuxuan Xie, and Zhihan Lv. Popular deep learning algorithms for disease prediction: a review. *Cluster Computing*, 26(2): 1231–1251, 2023.
- [334] Sher Lyn Tan, Ganeshsree Selvachandran, Raveendran Paramesran, and Weiping Ding. Lung Cancer Detection systems Applied to Medical images: a state-of-the-art survey. *Archives of Computational Methods in Engineering*, 32(1):343–380, 2025.
- [335] Yasin Kaya and Ercan Gürsoy. A MobileNet-based CNN model with a novel fine-tuning mechanism for COVID-19 infection detection. *Soft Computing*, 27(9):5521–5535, 2023.
- [336] J Arun Prakash, CR Asswin, Vinayakumar Ravi, V Sowmya, and KP Soman. Pediatric pneumonia diagnosis using stacked ensemble learning on multi-model deep CNN architectures. *Multimedia Tools and Applications*, 82(14):21311–21351, 2023.
- [337] Anupam Das. Adaptive UNet-based lung segmentation and ensemble learning with CNN-based deep features for automated COVID-19 diagnosis. *Multimedia Tools and Applications*, 81(4):5407–5441, 2022.
- [338] Mehdi Gheisari, Fereshteh Ebrahimzadeh, Mohamadtaghi Rahimi, Mahdieh Moazzamigodarzi, Yang Liu, Pijush Kanti Dutta Pramanik, Mohammad Ali Heravi, Abolfazl Mehbodniya, Mustafa Ghaderzadeh, Mohammad Reza Feylizadeh, et al. Deep learning: Applications, architectures, models, tools, and frameworks: A comprehensive survey. *CAAI Transactions on Intelligence Technology*, 2023.
- [339] Rammah Yousef, Gaurav Gupta, Nabhan Yousef, and Manju Khari. A holistic overview of deep learning approach in medical imaging. *Multimedia Systems*, 28(3):881–914, 2022.
- [340] Apeksha Koul, Rajesh K Bawa, and Yogesh Kumar. Artificial intelligence techniques to predict the airway disorders illness: a systematic review. *Archives of Computational Methods in Engineering*, 30(2):831–864, 2023.
- [341] Kavita Thakur, Manjot Kaur, and Yogesh Kumar. A comprehensive analysis of deep learning-based approaches for prediction and prognosis of infectious diseases. *Archives of Computational Methods in Engineering*, 30(7):4477–4497, 2023.

- [342] Yong Peng, Enbin Liu, Shanbi Peng, Qikun Chen, Dangjian Li, and Dianpeng Lian. Using artificial intelligence technology to fight COVID-19: a review. *Artificial intelligence review*, pages 1–37, 2022.
- [343] Bernd Carsten Stahl, Josephina Antoniou, Nitika Bhalla, Laurence Brooks, Philip Jansen, Blerta Lindqvist, Alexey Kirichenko, Samuel Marchal, Rowena Rodrigues, Nicole Santiago, et al. A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*, 56(11):12799–12831, 2023.
- [344] Tirth Mehta and Ninad Mehendale. Classification of X-ray images into COVID-19, pneumonia, and TB using cGAN and fine-tuned deep transfer learning models. *Research on Biomedical Engineering*, 37:803–813, 2021.
- [345] Rajendra D Bhosale and DM Yadav. Customized convolutional neural network for pulmonary multi-disease classification using chest x-ray images. *Multimedia Tools and Applications*, pages 1–35, 2023.
- [346] Yujin Oh, Sangjoon Park, and Jong Chul Ye. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE transactions on medical imaging*, 39(8):2688–2700, 2020.
- [347] Emtiaz Hussain, Mahmudul Hasan, Md Anisur Rahman, Ickjai Lee, Tasmi Tamanna, and Mohammad Zavid Parvez. CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images. *Chaos, Solitons & Fractals*, 142:110495, 2021.
- [348] Asif Iqbal Khan, Junaid Latief Shah, and Mohammad Mudasir Bhat. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Computer methods and programs in biomedicine*, 196:105581, 2020.
- [349] R Karthik, R Menaka, and M Hariharan. Learning distinctive filters for COVID-19 detection from chest X-ray using shuffled residual CNN. *Applied Soft Computing*, 99:106744, 2021.
- [350] Dina M Ibrahim, Nada M Elshennawy, and Amany M Sarhan. Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Computers in biology and medicine*, 132:104348, 2021.

- [351] Abdullahi Umar Ibrahim, Mehmet Ozsoz, Sertan Serte, Fadi Al-Turjman, and Polycarp Shizawaliyi Yakoi. Pneumonia classification using deep learning from chest X-ray images during COVID-19. *Cognitive Computation*, pages 1–13, 2021.
- [352] Gaurav Srivastava, Anindita Chauhan, Mahesh Jangid, and Sandeep Chaurasia. CoviXNet: A novel and efficient deep learning model for detection of COVID-19 using chest X-Ray images. *Biomedical Signal Processing and Control*, 78:103848, 2022.
- [353] Suat Toraman, Talha Burak Alakus, and Ibrahim Turkoglu. Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks. *Chaos, Solitons & Fractals*, 140: 110122, 2020.
- [354] Harsh Gupta, Naman Bansal, Swati Garg, Hritesh Mallik, Anju Prabha, and Jyoti Yadav. A hybrid convolutional neural network model to detect COVID-19 and pneumonia using chest X-ray images. *International Journal of Imaging Systems and Technology*, 33(1):39–52, 2023.
- [355] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in biology and medicine*, 121:103792, 2020.
- [356] Junaid Shuja, Eisa Alanazi, Waleed Alasmay, and Abdulaziz Alashaikh. COVID-19 open source data sets: a comprehensive survey. *Applied Intelligence*, 51:1296–1325, 2021.
- [357] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can AI help in screening viral and COVID-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020.
- [358] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in biology and medicine*, 132:104319, 2021.
- [359] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Labeled optical coherence tomography (OCT) and chest X-ray images for classification (2018).

- Mendeley Data*, v2 <https://doi.org/10.17632/rscbjbr9sj> <https://nihcc.app.box.com/v/ChestXray-NIHCC>, 2018.
- [360] Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020.
- [361] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [362] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.
- [363] Wilhelm Burger and Mark J Burge. *Digital image processing: An algorithmic introduction*. Springer Nature, 2022.
- [364] R Archana and PS Eliahim Jeevaraj. Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, 57(1):11, 2024.
- [365] Christian Garbin, Xingquan Zhu, and Oge Marques. Dropout vs. batch normalization: an empirical study of their impact to deep learning. *Multimedia Tools and Applications*, 79:12777–12815, 2020.
- [366] Andrew Holbrook, Thomas Lumley, and Daniel Gillen. Estimating prediction error for complex samples. *Canadian Journal of Statistics*, 48(2):204–221, 2020.
- [367] Jerzy Wieczorek, Cole Guerin, and Thomas McMahon. K-fold cross-validation for complex sample surveys. *Stat*, 11(1):e454, 2022.
- [368] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26, 2020.

- [369] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.
- [370] Nvidia Developer. Jetson AGX Xavier Developer Kit. URL <https://developer.nvidia.com/embedded/jetson-agx-xavier-developer-kit>. Last accessed 20 May 2025.
- [371] EunJin Jeong, Jangryul Kim, and Soonhoi Ha. Tensorrt-based framework and optimization methodology for deep learning inference on jetson boards. *ACM Transactions on Embedded Computing Systems (TECS)*, 21(5):1–26, 2022.
- [372] KG Satheeshkumar, V Arunachalam, and S Deepika. Hand-held GPU accelerated device for multiclass classification of X-ray images using CNN model. *Microprocessors and Microsystems*, 106:105046, 2024.
- [373] Ghulam Murtaza, Liyana Shuib, Ainuddin Wahid Abdul Wahab, Ghulam Mujtaba, Ghulam Mujtaba, Henry Friday Nweke, Mohammed Ali Al-garadi, Fariha Zulfiqar, Ghulam Raza, and Nor Aniza Azmi. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. *Artificial Intelligence Review*, 53:1655–1720, 2020.
- [374] Utkarsh Mahadeo Khair and R Dhanalakshmi. High-dimensional microarray dataset classification using an improved adam optimizer (iAdam). *Journal of Ambient Intelligence and Humanized Computing*, 11(11):5187–5204, 2020.
- [375] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: tricks of the trade: second edition*, pages 421–436. Springer, 2012.

Παράρτημα Α΄

Δημοσιεύσεις

Κατά τη διάρκεια της διδακτορικής διατριβής δημοσιεύθηκαν οι ακόλουθες εργασίες σε επιστημονικά περιοδικά και πρακτικά διεθνών συνεδρίων με κρίση.

Α΄.1 Περιοδικά

- [Π1] Sanida, T., & Dasygenis, M. (2024). A novel lightweight CNN for chest X-ray-based lung disease identification on heterogeneous embedded system. *Applied Intelligence*, 54(6), 4756-4780. **(Scimago Q2)**
- [Π2] Sanida, T., & Dasygenis, M. (2024). MiniTomatoNet: A lightweight CNN for tomato leaf disease recognition on heterogeneous FPGA-SoC. *The Journal of Supercomputing*, 80(15), 21837-21866. **(Scimago Q2)**
- [Π3] Sideris, A., Sanida, T., & Dasygenis, M. (2024). Hardware acceleration design of the SHA-3 for high throughput and low area on FPGA. *Journal of Cryptographic Engineering*, 14(2), 193-205. **(Scimago Q2)**
- [Π4] Sanida, T., Sanida, M. V., Sideris, A., & Dasygenis, M. (2024). Optimizing Lung Condition Categorization through a Deep Learning Approach to Chest X-ray Image Analysis. *BioMedInformatics*, 4(3), 2002-2021. **(Scimago Q2)**
- [Π5] Sanida, T., Sanida, M. V., Sideris, A., & Dasygenis, M. (2024). Enhancing Pulmonary Diagnosis in Chest X-rays through Generative AI Techniques. *J*, 7(3), 302-318.

- [Π6] Sanida, M. V., Sanida, T., Sideris, A., & Dasygenis, M. (2024). An advanced deep learning framework for multi-class diagnosis from chest X-ray images. *J*, 7(1), 48-71.
- [Π7] Sanida, T., Sideris, A., Sanida, M. V., & Dasygenis, M. (2023). Tomato leaf disease identification via two-stage transfer learning approach. *Smart Agric. Technol.* 5, 100275 (2023). **(Scimago Q1)**
- [Π8] Sideris, A., Sanida, T., & Dasygenis, M. (2023). A novel hardware architecture for enhancing the keccak hash function in FPGA devices. *Information*, 14(9), 475. **(Scimago Q2)**
- [Π9] Sanida, T., Tabakis, I. M., Sanida, M. V., Sideris, A., & Dasygenis, M. (2023). A robust hybrid deep convolutional neural network for COVID-19 disease identification from chest X-ray images. *Information*, 14(6), 310. **(Scimago Q2)**
- [Π10] Sanida, M. V., Sanida, T., Sideris, A., & Dasygenis, M. (2023). An efficient hybrid CNN classification model for tomato crop disease. *Technologies*, 11(1), 10. **(Scimago Q1)**
- [Π11] Sanida, T., Tsiktisiris, D., Sideris, A., & Dasygenis, M. (2022). A heterogeneous implementation for plant disease identification using deep learning. *Multimedia Tools and Applications*, 81(11), 15041-15059. **(Scimago Q1)**
- [Π12] Sanida, T., Sideris, A., Tsiktisiris, D., & Dasygenis, M. (2022). Lightweight neural network for COVID-19 detection from chest X-ray images implemented on an embedded system. *Technologies*, 10(2), 37.
- [Π13] Sideris, A., Sanida, T., Tsiktisiris, D., & Dasygenis, M. (2022). Acceleration of Image Processing with SHA-3 (Keccak) Algorithm using FPGA. *J. Eng. Res. Sci.* 1, 20-28.
- [Π14] Sideris, A., Sanida, T., & Dasygenis, M. (2020). High throughput implementation of the keccak hash function using the NIOS-II processor. *Technologies*, 8(1), 15.

A'.2 Συνέδρια

- [Σ1] Sanida, T., Sideris, A., Sanida, M. V., Dossis, M., & Dasygenis, M. (2024, September). Accelerating CNNs for Pneumonia Disease Diagnosis via

- Heterogeneous FPGA Systems. In 2024 9th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 159-162). IEEE.
- [Σ2] Sanida, M. V., Sanida, T., Sideris, A., Dossis, M., & Dasygenis, M. (2024, September). Fake News Detection Approach Using Hybrid Deep Learning Framework. In 2024 9th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 81-84). IEEE.
- [Σ3] Chatzisavvas, A., Sanida, T., Dossis, M., & Dasygenis, M. (2024, September). Improving Efficiency in Agricultural UGVs Through Enhanced Pathfinding Techniques. In 2024 9th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 179-182). IEEE.
- [Σ4] Sanida, T., Sideris, A., Sanida, M. V., Dossis, M., & Dasygenis, M. (2023, November). An Efficiency CNN Solution for Olive Disease Management Through FPGA. In 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
- [Σ5] Sanida, T., Sanida, M. V., Sideris, A., Dossis, M., & Dasygenis, M. (2023, November). Efficient categorization of pneumonia diagnosis using low-power embedded devices. In 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
- [Σ6] Sideris, A., Sanida, T., Sanida, M. V., Dossis, M., & Dasygenis, M. (2023, November). Accelerate processing of image with the keccak-512 algorithm on cryptoprocessor. In 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
- [Σ7] Sanida, T., Sideris, A., Sanida, M. V., Dossis, M., & Dasygenis, M. (2023, November). Acceleration of GANs for Potato Crop Disease Identification via FPGA. In 2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.

- [Σ8] Sanida, T., Sanida, M. V., Sideris, A., & Dasygenis, M. (2023, June). A lightweight CNN model for tomato crop diseases on heterogeneous embedded system. In 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST) (pp. 1-4). IEEE.
- [Σ9] Sanida, T., Sideris, A., Chatzisavvas, A., Dossis, M., & Dasygenis, M. (2022, September). Radiography images with transfer learning on embedded system. In 2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
- [Σ10] Sideris, A., Sanida, T., Chatzisavvas, A., Dossis, M., & Dasygenis, M. (2022, September). High throughput of image processing with keccak algorithm using microprocessor on FPGA. In 2022 7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM) (pp. 1-4). IEEE.
- [Σ11] Sideris, A., Sanida, T., Tsiktisiris, D., & Dasygenis, M. (2022). Image hashing based on SHA-3 implemented on FPGA. In Recent Advances in Manufacturing Modelling and Optimization: Select Proceedings of RAM 2021 (pp. 521-530). Singapore: Springer Nature Singapore.
- [Σ12] Tsiktisiris, D., Sanida, T., Sideris, A., & Dasygenis, M. (2022). Accelerated defective product inspection on the edge using deep learning. In Recent Advances in Manufacturing Modelling and Optimization: Select Proceedings of RAM 2021 (pp. 185-191). Singapore: Springer Nature Singapore.
- [Σ13] Sanida, T., Tsiktisiris, D., Sideris, A., & Dasygenis, M. (2021, July). A heterogeneous lightweight network for plant disease classification. In 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST) (pp. 1-4). IEEE.
- [Σ14] Sideris, A., Sanida, T., & Dasygenis, M. (2020, December). High throughput pipelined implementation of the SHA-3 cryptoprocessor. In 2020 32nd International Conference on Microelectronics (ICM) (pp. 1-4). IEEE.
- [Σ15] Sanida, T., Sideris, A., & Dasygenis, M. (2020, September). Accelerating the AES algorithm using OpenCL. In 2020 9th International conference on modern circuits and systems technologies (MOCAST) (pp. 1-4). IEEE.

- [Σ16] Sanida, T., Sideris, A., & Dasygenis, M. (2020, September). A heterogeneous implementation of the Sobel edge detection filter using OpenCL. In 2020 9th International conference on modern circuits and systems technologies (MOCAST) (pp. 1-4). IEEE.
- [Σ17] Sideris, A., Sanida, T., & Dasygenis, M. (2019, November). Hardware acceleration of the AES algorithm using NIOS-II processor. In 2019 Panhellenic Conference on Electronics & Telecommunications (PACET) (pp. 1-5). IEEE.
- [Σ18] Sideris, A., Sanida, T., & Dasygenis, M. (2019, May). Hardware acceleration of SHA-256 algorithm using NIOS-II processor. In 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST) (pp. 1-4). IEEE.