ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΔΙΔΑΚΤΟΡΙΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Καινοτόμος χρήση τεχνολογιών IoT και Machine Learning για την Παρακολούθηση και Διαχείριση Έξυπνων Χώρων

Ελένη Γ. Τσαλέρα

ΑΙΓΑΛΕΩ

ΙΑΝΟΥΑΡΙΟΣ 2024

UNIVERSITY OF WEST ATTICA

SCHOOL OF ENGINEERING
DEPARTMENT OF INFORMATICS AND COMPUTER

ENGINEERING DOCTORAL STUDIES

PhD THESIS

Innovative use of IoT and Machine Learning technologies for the Monitoring and Management of Smart Spaces

Eleni G. Tsalera

ATHENS-EGALEO

JANUARY 2024

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Καινοτόμος χρήση των τεχνολογιών IoT και Machine Learning για την Παρακολούθηση και Διαχείριση Έξυπνων Χώρων

Ελένη Γ. Τσαλέρα

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Μαρία Σαμαράκου, Ομότιμη Καθηγήτρια του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, Πανεπιστήμιο Δυτικής Αττικής

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:

Μαρία Σαμαράκου, Ομότιμη Καθηγήτρια, Τμ. ΜΠΥ, ΠαΔΑ
Ιωάννης Βογιατζής, Καθηγητής, Τμ. ΜΠΥ, ΠαΔΑ
Ανδρέας Παπαδάκης, Καθηγητής, Τμ. Εκπ. Ηλ/γων Μηχ. και Εκπ. Ηλ/κών Μηχ., ΑΣΠΑΙΤΕ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Maria Samarakou
Digitally signed by Maria Samarakou
Date: 2024.02.01 13:06:38 +02'00'

Μαρία Σαμαράκου,
Ομότιμη Καθηγήτρια ΠαΔΑ

Ioannis Voyiatzis
Digitally signed by Ioannis Voyiatzis
Date: 2024.02.01 13:27:02 +02'00'

Ιωάννης Βογιατζής,
Καθηγητής ΠαΔΑ

ANDREAS PAPADAKIS
01/02/2024 13:48

Ανδρέας Παπαδάκης,
Καθηγητής ΑΣΠΑΙΤΕ

Kleio Sgouropoulou
Digitally signed by Kleio Sgouropoulou
Date: 2024.02.03 09:43:11 +02'00'

Κλειώ Σγουροπούλου,
Καθηγήτρια ΠαΔΑ

Nikitas Karanikolas
Digitally signed by Nikitas Karanikolas
Date: 2024.02.03 13:23:49 +02'00'

Νικήτας Καρανικόλας,
Καθηγητής ΠαΔΑ

SPYRIDON PANETSOS
04/02/2024 08:43

Σπυρίδων Πανέτσος,
Καθηγητής ΑΣΠΑΙΤΕ

Καθηγήτρια
MARIA NIKOLAIDI
04/02/2024 12:48

Μαρία Νικολαΐδου,
Καθηγήτρια Χαροκόπειο Παν.

Ημερομηνία εξέτασης 22/01/2024

PhD THESIS

Innovative use of IoT and Machine Learning technologies for the Monitoring and Management of Smart Spaces

Eleni G. Tsalera

SUPERVISOR: Maria Samarakou, Professor Emeritus UniWA

THREE-MEMBER ADVISORY COMMITTEE:

Maria Samarakou, Professor Emeritus UniWA
Ioannis Voyiatzis, Professor UniWA
Andreas Papadakis, Professor ASPETE

SEVEN-MEMBER EXAMINATION COMMITTEE

Maria Samarakou
Digitally signed by Maria Samarakou
Date: 2024.02.01 13:07:35 +02'00'

Maria Samarakou,
Professor Emeritus UniWA

Ioannis Voyiatzis
Digitally signed by Ioannis Voyiatzis
Date: 2024.02.01 13:28:14 +02'00'

Ioannis Voyiatzis,
Professor UniWA

ANDREAS PAPADAKIS
01/02/2024 13:50

Andreas Papadakis,
Professor ASPETE

Kleio Sgouropoulou
Digitally signed by Kleio Sgouropoulou
Date: 2024.02.03 00:32:41 +02'00'

Cleo Sgouropoulou,
Professor UniWA

Nikitas Karanikolas
Digitally signed by Nikitas Karanikolas
Date: 2024.02.03 13:25:11 +02'00'

Nikitas Karanikolas,
Professor UniWA

SPYRIDON PANETSOS
04/02/2024 11:26

Spyridon Panetsos,
Professor ASPETE

Καθηγήτρια
MARIA NIKOLAIDI
04/02/2024 12:50

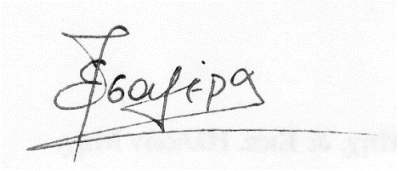
Maria Nikolaidou,
Professor Harokopio Uni.

Examination Date 22/01/2024

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ

Ο/η κάτωθι υπογεγραμμένη Ελένη Τσαλέρα του Γεωργίου, υποψήφια διδάκτορας του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:
«Είμαι συγγραφέας και δικαιούχος των πνευματικών δικαιωμάτων επί της διατριβής και δεν προσβάλω τα πνευματικά δικαιώματα τρίτων. Για τη συγγραφή της διδακτορικής μου διατριβής δεν χρησιμοποίησα ολόκληρο ή μέρος έργου άλλου δημιουργού ή τις ιδέες και αντιλήψεις άλλου δημιουργού χωρίς να γίνεται αναφορά στην πηγή προέλευσης (βιβλίο, άρθρο από εφημερίδα ή περιοδικό, ιστοσελίδα κ.λπ.). Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.
Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Η Δηλούσα



ΠΕΡΙΛΗΨΗ

Οι πρόσφατες εξελίξεις στον τομέα της επεξεργασίας σήματος, των τεχνολογιών των αισθητήρων, των επικοινωνιών καθώς και, της υπολογιστικής υψηλών επιδόσεων (High Performance Computing) καθιστούν δυνατή την υλοποίηση «έξυπνων» χώρων, με την έννοια των φυσικών χώρων οι οποίοι είναι εξοπλισμένοι με τεχνολογία συλλογής, μεταφοράς και επεξεργασίας δεδομένων με στόχο την αύξηση της λειτουργικής αποτελεσματικότητας και την βελτίωση της ποιότητας των διεργασιών που εκτελούνται σε αυτούς. Με την ανάπτυξη εξελιγμένων προσεγγίσεων Μηχανικής Μάθησης (ΜΜ) επωφελείται ένας αυξανόμενος αριθμός εφαρμογών και πραγματοποιείται ένα μέρος του οράματος Internet of Things (IoT). Πλέον, τα συστήματα και οι καταστάσεις μπορούν να παρακολουθούνται και να ελέγχονται αναλόγως των απαιτήσεων.

Η κεντρική ιδέα έγκειται στην καταγραφή δεδομένων πέρα από αυτά που χαρακτηρίζουν τις φυσικές συνθήκες του χώρου. Τα άτομα παράγουν σήματα, τα οποία σχετίζονται με τις δραστηριότητες τους, και ταυτόχρονα αποτελούν δείκτες της ποιότητας της διαδικασίας που εκτελείται. Αυτή η Διδακτορική Διατριβή στοχεύει στην μελέτη των δύο κατεξοχήν σημάτων τα οποία είναι ενδεικτικά της διεργασίας και των συνθηκών που επικρατούν: του Ήχου και της Εικόνας. Η περίπτωση μελέτης είναι οι αίθουσες θεωρητικής διδασκαλίας ή εργαστηριακού πειράματος. Στόχος είναι η αποσαφήνιση των δραστηριοτήτων και του ενδιαφέροντος των εμπλεκόμενων μερών με αποτέλεσμα την βελτίωση και αναβάθμιση της ποιότητας του έργου που συντελείται σε μία έξυπνη αίθουσα. Τα σήματα αυτά είναι πλούσια σε πληροφορίες, χαρακτηρίζονται από διάφορους περιορισμούς, ενώ η επεξεργασία και η κατανόησή τους αποτελεί πρόκληση. Ταυτόχρονα, η συμπαγής κατανόηση των δραστηριοτήτων και της αλληλεπίδρασης των ατόμων μπορεί να υποστηρίξει την αξιολόγηση και συνεπώς την ενίσχυση των επιμέρους διεργασιών.

Στο πρώτο μέρος της διατριβής πραγματοποιείται η ταξινόμηση ηχητικών σημάτων οι οποίοι είναι χαρακτηριστικοί στην εξέλιξη της εκπαιδευτικής διαδικασίας. Αρχικά, αναφέρεται ένα εκτενές σύνολο ηχητικών χαρακτηριστικών (συνολικά 143), υλοποιούνται αλγόριθμοι εξαγωγής των τιμών αυτών, και πραγματοποιείται η ταξινόμηση των ήχων με αλγόριθμους ΜΜ. Δεδομένου του μεγάλου πλήθους των χαρακτηριστικών, της αντίστοιχης υπολογιστικής επιβάρυνσης χρησιμοποιώντας το σύνολο αυτών, αλλά και της πιθανής υποβάθμισης της ακρίβειας ταξινόμησης λόγω υπερπροσαρμογής, προτείνεται μέθοδος ιεράρχησης των χαρακτηριστικών με βάση την περιγραφική τους ικανότητα. Η μέθοδος βασίζεται στην Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis – PCA) και συγκρίνεται με την γνωστή μέθοδο Relief-F. Πραγματοποιούνται πειραματικοί έλεγχοι με πέντε αλγορίθμους ΜΜ χρησιμοποιώντας αυξανόμενο αριθμό χαρακτηριστικών. Τα πειραματικά αποτελέσματα ανέδειξαν την χρησιμότητα της μεθόδου μείωσης της διαστασιολόγησης, επιτυγχάνοντας ακρίβεια ταξινόμησης μεγαλύτερη από 90% χρησιμοποιώντας 25 ηχητικά χαρακτηριστικά.

Στην συνέχεια εφαρμόζονται μηχανισμοί Βαθιάς Μάθησης (ΒΜ) και συγκεκριμένα τα Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ). Προκειμένου να αναβαθμιστεί η ακρίβεια ταξινόμησης, χρησιμοποιούνται καθιερωμένες, ευρέως γνωστές αρχιτεκτονικές. Η χρήση αυτών των ΣΝΔ γίνεται, αφού το ηχητικό σήμα μετατραπεί σε κατάλληλη εικονική αναπαράσταση, μέσω μεταφοράς μάθησης. Σε αυτό το σημείο διερευνάται εκτενώς το σύνολο των τιμών των υπερπαραμέτρων επανεκπαίδευσης των δικτύων σε νέα σύνολα δεδομένων. Το αποτέλεσμα αυτής της διερεύνησης είναι η ρύθμιση των τιμών των υπερπαραμέτρων που οδηγούν σε μεγιστοποίηση της ακρίβειας ταξινόμησης με ταυτόχρονη ελαχιστοποίηση του αντίστοιχου υπολογιστικού χρόνου, επιτυγχάνοντας ακρίβεια ταξινόμησης που ξεπερνά το 90% σε τρεις διαφορετικές βάσεις δεδομένων.

Το δεύτερο μέρος της διατριβής αφορά την μελέτη του σήματος της Εικόνας. Η ανάπτυξη ολοένα και πιο προηγμένων αλγορίθμων και μοντέλων έχει καταστήσει την ανίχνευση και την αναγνώριση αντικειμένων τετριμμένη. Επομένως, η διατριβή αυτή επικεντρώθηκε σε μία πιο εκλεπτυσμένη ανάλυση της εικόνας με στόχο την αναγνώριση της έκφρασης του προσώπου (Facial Emotion Recognition-FER). Με αφορμή αυτή την μελέτη διερευνήθηκαν δύο μηχανισμοί εξαγωγής χαρακτηριστικών της εικόνας: με χειροκίνητο τρόπο, και αυτόματα μέσω Βαθιάς Μάθησης, με βάση τα ΣΝΔ. Και οι δύο μηχανισμοί μελετήθηκαν διεξοδικά και αξιολογήθηκαν σε τρεις βάσεις δεδομένων FER ως προς την απόδοση της ακρίβειας ταξινόμησης και τον αντίστοιχο υπολογιστικό χρόνο. Οι χειροκίνητες μέθοδοι εξετάστηκαν ως προς τις τιμές των εσωτερικών τους παραμέτρων, ενώ η μελέτη των νευρωνικών δικτύων ήταν διττή. Αρχικά, τα χαρακτηριστικά εξήχθησαν χωρίς να επανεκπαιδευτούν τα δίκτυα στα νέα δεδομένα από επίπεδα διαφορετικών βαθών, και στην συνέχεια η εξαγωγή των χαρακτηριστικών έγινε μετά την επανεκπαίδευση των δικτύων στις νέες βάσεις δεδομένων μέσω μεταφοράς μάθησης. Από την έρευνα προέκυψε ότι χωρίς την επανεκπαίδευση των δικτύων η εξαγωγή των χαρακτηριστικών της εικόνας από το βαθύτερο επίπεδο των ΣΝΔ οδηγεί σε υποδεέστερα αποτελέσματα ακρίβειας ταξινόμησης (κατά μέσο όρο 74%) σε σχέση με τα αντίστοιχα των χειροκίνητων μεθόδων (κατά μέσο όρο 86%). Η εξαγωγή χαρακτηριστικών από το 50% ή το 75% του βάθους των ΣΝΔ οδηγεί σε υψηλότερη απόδοση ταξινόμησης (κατά μέσο όρο 90%) για κάθε περίπτωση ποιότητας εικόνων. Η επανεκπαίδευση των ΣΝΔ και η χρήση της μεθόδου της μεταφοράς μάθησης βελτιώνει την ακρίβεια ταξινόμησης στην περίπτωση που είναι διαθέσιμες μεγάλες βάσεις δεδομένων. Επιπρόσθετα, κάθε μέθοδος αξιολογήθηκε ως προς την ανθεκτικότητα της σε δύο συχνά απαντώμενους τύπους θορύβου, τον Gaussian και τον Salt & Pepper, με τα ΣΝΔ να εμφανίζονται πιο ανθεκτικά (η ακρίβειας ταξινόμησης μειώνεται κατά περίπου 10% έναντι της μείωσης της ακρίβειας ταξινόμησης κατά 60% με τις χειροκίνητες μεθόδους). Το αποτέλεσμα ήταν η δημιουργία ενός πλαισίου επιλογής μεθόδου ανάλογα με την ποιότητα των διαθέσιμων εικόνων, τις απαιτήσεις και τις προδιαγραφές της εφαρμογής. Επιλέγοντας την κατάλληλη μέθοδο επιτεύχθηκε ακρίβεια ταξινόμησης που ξεπερνά το 92% για κάθε βάση δεδομένων FER που χρησιμοποιήθηκε.

Η αξία αυτής της διδακτορικής διατριβής επιβεβαιώνεται από την περαιτέρω εφαρμογή των μεθόδων και αλγορίθμων σε ερευνητικές περιπτώσεις ανοιχτών χώρων. Συγκεκριμένα, οι μέθοδοι ταξινόμησης του ήχου με την ιεράρχηση των ηχητικών χαρακτηριστικών εφαρμόστηκε σε έρευνα εστιασμένη στον περιβαλλοντικό θόρυβο σε αστικό τοπίο, επιτυγχάνοντας ακρίβειας ταξινόμησης οκτώ τύπων αστικού θορύβου που φτάνει το 85%. Επιτυγχάνοντας υψηλή απόδοση ταξινόμησης του θορύβου είναι δυνατή η κατανόηση της προέλευσης αυτού και της λήψης αντίστοιχων μέτρων προκειμένου να μειωθεί η ηχορρύπανση. Η ταξινόμηση των εικόνων, σε συνδυασμό με την σημασιολογική κατάτμηση του περιεχομένου τους, εφαρμόστηκε σε έρευνα εστιασμένη στην ανίχνευση πυρκαγιών σε δασικές περιοχές. Με αυτόν τον τρόπο δημιουργείται πλαίσιο ανάλυσης του κινδύνου με βάση τις υποδομές που υπάρχουν στον χώρο αλλά και του τρόπου επέμβασης αναλόγως της πρόσβασης σε αυτόν. Η ανίχνευση αντικειμένων ενδιαφέροντος σε συνδυασμό με τα αποτελέσματα ταξινόμησης που ξεπερνούν το 95% αναδεικνύουν την αξία της προτεινόμενης έρευνας στην έγκαιρη αντιμετώπιση των πυρκαγιών, αλλά και της επεκτασιμότητας του πεδίου εφαρμογών των ανεπτυγμένων αλγορίθμων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μηχανική Μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Βαθιά Μάθηση, μείωση διαστασιολόγησης, μεταφορά μάθησης, Μηχανική Μάθηση, Συνελικτικά Νευρωνικά Δίκτυα, ταξινόμηση σημάτων

ABSTRACT

Recent developments in the field of signal processing, sensor technologies, communications as well as High Performance Computing enable the realization of "smart" spaces, in the sense of physical spaces equipped with technology suitable for the collection, transfer and processing of data with the aim of increasing operational efficiency and improving the quality of the processes performed in them. Applying sophisticated Machine Learning (ML) approaches is benefiting a growing number of applications and a part of the Internet of Things (IoT) vision is being realized. Now, systems and states can be monitored and controlled as required.

The central idea lies in the recording of data beyond those characterizing the physical conditions of the space. Individuals produce signals, that related to their activities and are indicative of the process being carried out, and at the same time, they are indicators of its orderly or not progress. This Doctoral Thesis aims to study the two preeminent signals that are indicative of the process and the prevailing conditions: Sound and Image. The study case is the classrooms of theoretical teaching or laboratory experiments. The aim is to clarify the activities and the interest of the parties involved, resulting in the improvement and upgrading of the quality of the work carried out in a smart room. These signals are rich in information, characterized by various limitations, while processing and understanding them is a challenge. At the same time, a solid understanding of the conditions and consequences of human activity can support the evaluation and therefore the strengthening of processes.

In the first part of the Thesis, the classification of sound signals that are characteristic in the evolution of the educational process. Initially, an extensive set of sound features (143 in total) is reported, algorithms for extracting these values are implemented and the sounds are classified with ML algorithms. Given the large number of audio features, the corresponding computational burden using all of them, but also the possible degradation of the classification accuracy due to overfitting, a method of prioritizing the features based on their descriptive ability is proposed. The method is based on Principal Component Analysis (PCA) and is compared to the well-known Relief-F method. Experimental tests are performed with five ML algorithms using an increasing number of features. The experimental results demonstrated the utility of the dimensionality reduction method by achieving a classification accuracy of more than 90% using 25 audio features.

Then, Deep Learning (DL) mechanisms are employed, specifically Convolutional Neural Networks (CNNs). In order to improve the classification accuracy, well-established, well-known architectures and networks pre-trained on the large ImageNet image dataset are used. The use of these CNNs is done after the audio signal has been converted into a suitable virtual representation through transfer learning. At this point, the set of hyper-parameter values of retraining networks on new datasets is extensively explored. The result of this investigation is the tuning of the hyper-parameter values that lead to the maximization of the classification accuracy while minimizing the corresponding computational time, achieving a classification accuracy exceeding 90% in three different databases.

The research field of sound classification has developed rapidly in recent decades resulting in the creation of sound datasets, which include different, arbitrarily selected sound classes for different case studies. In this regard, two types of systematic associations between sound classes were explored: a) semantic and b) comparative based on sound features. In terms of the first association, audio classes are semantically related taking into account the unifying AudioSet ontology, with audio classes being associated based on the origin (source) of the

audio. Regarding the second correlation, it is based on the calculation of the distance of the values of the audio characteristics. At the same time, sounds originating from realistic environments include classes which are combined in a sequential and/or overlapping manner. In these cases it is necessary to separate (segment) the audio stream in order to achieve the classification of each audio segment. A set of parameters such as the minimum sound duration and sound intervals were defined to achieve the segmentation process of the audio streams.

The second part of the Thesis concerns the study of Image. The development of increasingly advanced algorithms and models has made object detection and recognition trivial. Therefore, this Thesis focused on a more refined analysis of the image with the aim of facial expression recognition (Facial Emotion Recognition). On the occasion of this study, two image feature extraction mechanisms were investigated: manually, with handcrafted methods, and automatically through DL methods, based on CNNs. Both mechanisms were thoroughly studied in terms of their internal parameters and evaluated on FER databases in terms of their classification accuracy performance and the corresponding computational time. Handcrafted methods were examined concerning their internal parameter values, while the study of neural networks was two-fold. First, the features were extracted without retraining the networks on the new data from different depths of their layers, and then the feature extraction was done after retraining the networks on the new databases through transfer learning. The research showed that without retraining the networks, extracting the features from the deepest layer of CNNs leads to inferior classification accuracy results (74% on average) compared to handcrafted methods (86% on average). Extracting image features form 50% or 75% of the depth of the CNNs results in higher classification accuracy (90% on average) for each image quality case. Retraining CNNs and using the transfer learning method improves the classification accuracy when large databases are available. In addition, each method was evaluated for its robustness to two commonly encountered types of noise, Gaussian and Salt & Pepper. CNNs appear to be more robust as the classification accuracy decreases by 10% versus a 60% decrease using handcrafted methods. The result was the creation of a method selection framework according to the quality of available images, application requirements and specifications. By choosing the appropriate method, a classification accuracy exceeding 92% is achieved for each FER database used.

The heterogeneity of available algorithms for signal classification is reflected in the computational resources required to train the models and extract the results. From this point of view, the computational resource requirements can be one of the criteria for choosing the classification method. So a set of training time values for different neural network architectures, with different training configurations and for different datasets was generated. Five neural network-based regression models were trained to estimate the training time. Models were evaluated based on correlation coefficient and root mean square error. The result was that the two-layer neural network yielded the highest correlation coefficient with the smallest error, indicating that this model can provide a good approximation of the computational time required for a case of adjacent data.

The available algorithms differ from each other in terms of their architecture and in particular, the number of their layers, the complexity of the connection between them, the number and size of filters. The heterogeneity of the available algorithms is reflected in the computational resources required to train the models and derive the inferences. From this point of view, the computational resource requirements can be the one of the criteria for choosing the classification method. For this purpose, a set of training time values for different CNN architectures, with different training configurations, and for different datasets was constructed. Five neural network-based regression models were trained to estimate the

training time. Models were evaluated based on correlation coefficient and root mean square error.

The value of this PhD Thesis is confirmed by further application of the methods and algorithms in open-space research cases. In particular, sound classification methods with the prioritization of sound features were applied to research focused on environmental noise in an urban landscape, achieving a classification accuracy of eight types of urban noise that reaches 85%. By achieving high noise classification performance it is possible to understand its origin and take appropriate measures to reduce noise pollution. Image classification, combined with semantic segmentation of their content, was applied to research focused on forest fire detection. In this way, a risk analysis framework is created based on the infrastructure that exist in the area, but also the method of intervention depending on the access to it. The detection of objects of interest combined with the classification results that exceed 95% highlight the value of the proposed research in the early response to fires, but also the extensibility of the field of applications of the developed algorithms.

SUBJECT AREA: Machine Learning

KEYWORDS: Convolutional Neural Networks, Deep Learning, dimensionality reduction, Machine Learning, signal classification, transfer learning

Στον Ηλία,

εννοείται.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω καταρχάς την Μαρία Σαμαράκου, Ομότιμη Καθηγήτρια του ΠαΔΑ, η οποία μου άνοιξε την πόρτα του Πανεπιστημίου Δυτικής Αττικής, και που έδωσε την ευκαιρία σε αυτό το όνειρο να γίνει πραγματικότητα. Όλα αυτά τα χρόνια ήταν παρούσα σε κάθε δύσκολη στιγμή, να κρίνει, να καθοδηγεί και να κατευθύνει σύμφωνα με την πολύτιμη εμπειρία της. Σας ευχαριστώ από τα βάθη της καρδιάς μου.

Θα ήθελα, επίσης, να ευχαριστήσω τον Ανδρέα Παπαδάκη, Καθηγητή της ΑΣΠΑΙΤΕ, ο οποίος ήταν ο μόνιμος συνοδοιπόρος, δάσκαλος και ο εμπνευστής αυτής της διατριβής. Οι ατελείωτες ώρες συνεργασίας, η καθοδήγηση, η συνεχής επικοινωνία και η μετάδοση των γνώσεων και της εμπειρίας του θα μου μείνουν αξέχαστα. Δεν θα μπορούσα να ελπίζω σε καλύτερο συνεργάτη και καθοδηγητή.

Ένα μεγάλο ευχαριστώ και στον Ιωάννη Βογιατζή, Καθηγητή του ΠαΔΑ, που με πίστεψε, με ενθάρρυνε και με υποστήριξε σε κάθε δύσκολη στιγμή. Η αμεσότητα της επικοινωνίας, και η εύρεση λύσεων αποτέλεσαν αρωγοί αυτής της διατριβής. Σας ευχαριστώ.

Για το τέλος αφήνω τους δικούς μου ανθρώπους. Τον αδελφικό μου φίλο Μανώλη Κυπραίο ο οποίος ήταν δίπλα μου σε κάθε δύσκολο σημείο αυτής της διαδρομής. Οι γνώσεις του και η βοήθειά του κάθε φορά που τον χρειάστηκα είναι αυτά στα οποία οφείλω το ότι έφτασα ως εδώ. Μανώλη σ’ ευχαριστώ, σε εσένα οφείλω το ότι προχώρησα όταν βρέθηκα σε αδιέξοδο. Τέλος, θέλω να ευχαριστήσω την μητέρα μου Αύρα, και τον σύζυγό μου Ηλία για την υπομονή που έκαναν, για την συμπαράστασή τους, και για όσα στερήθηκαν αυτά τα χρόνια. Χωρίς εσάς δεν θα τα είχα καταφέρει, και το μεγαλύτερο «ευχαριστώ» πηγαίνει σε εσάς.

Ελένη Γ. Τσαλέρα, 2024

ΛΙΣΤΑ ΔΗΜΟΣΙΕΥΣΕΩΝ

1. Andreas Papadakis, **Eleni Tsalera**, and Maria Samarakou, "Survey on sound and video analysis methods for monitoring face-to-face module delivery", *International Journal of Emerging Technologies in Learning (Online), 14.8,* **2019**, pp. 229. https://doi.org/10.3991/ijet.v14i08.9813

2. **Eleni Tsalera**, Andreas Papadakis, and Maria Samarakou, "Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm", *Energy Reports*, vol. 6, suppl. 6, Nov. **2020**, pp. 223-230. https://doi.org/10.1016/j.egyr.2020.08.045

3. **Eleni Tsalera**, Andreas Papadakis, and Maria Samarakou, "Novel principal component analysis-based feature selection mechanism for classroom sound classification", *Computational Intelligence, 37.4,* **2021**, pp. 1827-1843. https://doi.org/10.1111/coin.12468

4. **Eleni Tsalera**, Andreas Papadakis, and Maria Samarakou, "Comparison of pre-trained CNNs for audio classification using transfer learning", *Journal of Sensor and Actuator Networks, 10.4,* **2021**, pp. 72. https://doi.org/10.3390/jsan10040072

5. **Eleni Tsalera**, Andreas Papadakis, Maria Samarakou, and Ioannis Voyiatzis, "Feature extraction with handcrafted methods and convolutional neural networks for facial emotion recognition", *Applied Sciences, 12.17,* **2022**, pp. 8455. https://doi.org/10.3390/app12178455

6. **Eleni Tsalera**, Andreas, Papadakis, Maria Samarakou, and Ioannis Voyiatzis, "CNN-based Segmentation and Classification of Sound Streams under realistic conditions", In *Proceedings of the 26th Pan-Hellenic Conference on Informatics*, **2022**, pp. 373-378. https://doi.org/10.1145/3575879.3576020

7. **Eleni Tsalera**, Andreas Papadakis, Ioannis Voyiatzis, and Maria Samarakou, "CNN-based, contextualized, real-time fire detection in computational resource-constrained environments", *Energy Reports,* vol.9, suppl.9, Sept. **2023**, pp. 247-257. https://doi.org/10.1016/j.egyr.2023.05.260

8. **E. Tsalera**, D. Stratogiannis, A. Papadakis, I. Voyiatzis, and M. Samarakou, "Evaluation and Prediction of Resource Usage for multi-parametric Deep Learning training and inference", In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, **2023**. https://doi.org/10.1145/3635059.3635070

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
High Performance Computing	Υπολογιστική Υψηλών Επιδόσεων
Machine Learning	Μηχανική Μάθηση
Internet of Things	Διαδίκτυο των Πραγμάτων
Case study	Περίπτωση μελέτης
Deep Learning	Βαθιά Μάθηση
Convolutional Neural Networks	Συνελικτικά Νευρωνικά Δίκτυα
Facial emotion recognition	Αναγνώριση συναισθήματος προσώπου
Feature	Χαρακτηριστικό
Classification	Ταξινόμηση
Spectrogram	Φασματογράφημα
Scalogram	Διάγραμμα κλίμακας-χρόνου
Overfitting	Υπερπροσαρμογή
Dimensionality reduction	Μείωση διαστασιολόγησης
Label	Ετικέτα
Support Vector Machines	Μηχανισμοί Υποστήριξης Διανυσμάτων
k-Nearest Neighbors	κ-Πλησιέστεροι Γείτονες
Principal Component Analysis (Network)	(Δίκτυο) Ανάλυση Κύριων Συνιστωσών
Linear Discriminant Analysis	Γραμμική Διακριτική Ανάλυση
Boosted Trees	Ενισχυμένα Δένδρα
Random Forest	Τυχαίο Δάσος
Specificity	Ειδικότητα
Sensitivity	Ευαισθησία
Confusion matrix	Πίνακας σύγχυσης
Wrapper	Περιτύλιγμα
Embedded	Ενσωματωμένος
Kernel	Πυρήνας
Root Mean Square (Error)	Μέση Τετραγωνική Ρίζα (Σφάλματος)
Standard Deviation	Τυπική Απόκλιση
Time Maximum Autocorrelation Coefficients	Συντελεστές Μέγιστης Χρονικής Αυτοσυσχέτισης
Time Autocorrelation Coefficients	Χρονικοί Συντελεστές Αυτοσυσχέτισης
Peak Values	Τιμές Αιχμής
Peak Program Meter	Μετρητής Αιχμής Προγράμματος
Attack time	Χρόνος επίθεσης
Release time	Χρόνος αποδέσμευσης
Time Predictivity Ratio	Λόγος Χρονικής Πρόβλεψης
Zero Crossing Rate	Ρυθμός Διέλευσης από το Μηδέν
Slope	Κλίση
Flux	Ροή
Spectral Decrease	Φασματική Μείωση
Tonal Power Ratio	Λόγος Τονικής Ισχύος
Flatness	Επιπεδότητα
Kurtosis	Κύρτωση
Crest	Κορυφή
Skewness	Ασυμμετρία
Centroid	Κεντροειδές

Brightness	Φωτεινότητα
Spread	Έκταση
Roll-off	Πτώση
Mel Frequency Cepstral Coefficients	Συντελεστές Συχνοτήτων Mel
D (DD) Mel Frequency Cepstral Coefficients	Πρώτη (Δεύτερη) Παράγωγος των Συντελεστών Συχνοτήτων Mel
Discrete Cosine Transformation	Διακριτός Μετασχηματισμός Συνημίτονου
Pitch Spectral Harmonic Product Spectrum	Τονικό Φασματικό Αρμονικό Φάσμα Προϊόντων
Autocorrelation Function	Συνάρτηση Φασματικής Αυτοσυσχέτισης
Pitch Time Autocorrelation Function	Τονική Χρονική Συνάρτηση Αυτοσυσχέτισης
Pitch Time Average Magnitude Difference Function	Τονική Χρονική Συνάρτηση Διαφοράς Μέσου Πλάτους
Supervised	Εποπτευόμενη
Singular Value Decomposition	Διάσπαση Ιδιόμορφων Τιμών
Transfer learning	Μεταφορά μάθησης
Late fusion	Ύστερη συγχώνευση
Rectified Linear Units	Διορθωμένες Γραμμικές Μονάδες
Fire modules	Μονάδες πυρκαγιάς
Network Architecture Search	Αναζήτηση Αρχιτεκτονικής Δικτύου
Convolutional Recurrent Neural Networks	Συνελικτικά Επαναλαμβανόμενα Νευρωνικά Δίκτυα
Sound Event Detection	Ανίχνευση Συμβάντων Ήχου
Pooling layer	Επίπεδο συγκέντρωσης
Hyperparameters	Υπερπαράμετροι
Training (set)	(Σετ) Εκπαίδευσης
Validation (set)	(Σετ) Επαλήθευσης
Test (set)	(Σετ) Ελέγχου
Hop	Μήκος αναπήδησης
Continuous Wavelet Transform	Συνεχής Μετασχηματισμός Περιβάλλουσας
Full length	Πλήρες μήκος
Segment length	Μήκος τμήματος
Overlapping ratio	Λόγος επικάλυψης
Augmentation	Ενίσχυση
Loss function	Συνάρτηση απώλειας
Back-propagation	Οπισθοδιάδοση
Learning rate	Ρυθμός εκμάθησης
Epoch	Εποχή
Validation patience/accuracy	Υπομονή/ακρίβεια επικύρωσης
Stochastic Gradient Descent with Momentum	Στοχαστική Κλίση Καθόδου με Ορμή
Adaptive Moment Estimation	Προσαρμοστική Εκτίμηση Ροπής
Root Mean Square Propagation	Μέση Τετραγωνική Ρίζα Διάδοσης
Adaptive Gradient Algorithm	Αλγόριθμος Προσαρμοστικής Κλίσης
Generalization gap	Χάσμα γενίκευσης
Optimizer	Βελτιστοποιητής
Classification accuracy	Ακρίβεια ταξινόμησης
Training time	Χρόνος εκπαίδευσης
Average value	Μέση τιμή
Segmentation	Κατάτμηση
Discrete Wavelet Transform	Διακριτός Μετασχηματισμός Κυματιδίων
Electroencephalography	Ηλεκτροεγκεφαλογράφημα

Bayesian Information Criterion	Κριτήριο Πληροφορίας Bayesian
Confidence level	Στάθμη εμπιστοσύνης
Virtualization	Εικονικοποίηση
Edge computing	Ακροδικτυακή υπολογιστική
Sum of squares total	Άθροισμα τετραγώνων συνολικά
Sum of squares regression	Άθροισμα τετραγώνων παλινδρόμησης
Sum of squares error	Άθροισμα τετραγώνων σφάλματος
Bilayered neural network	Νευρωνικό δίκτυο δύο επιπέδων
Trilayered neural network	Νευρωνικό δίκτυο τριών επιπέδων
Inference time	Χρόνος εξαγωγής συμπερασμάτων
Affective computing	Συναισθηματική υπολογιστική
Facial action coding system	Σύστημα κωδικοποίησης δράσης του προσώπου
Action units	Μονάδες δράσης
Handcrafted methods	Χειροκίνητες μέθοδοι
Local Binary Patterns	Τοπικά Δυαδικά Μοτίβα
Histogram of Oriented Histograms	Ιστογράμματα Προσανατολισμένων Κλίσεων
Pixel	Εικονοστοιχείο
Pooling	Συγκεντρωτικός
Fully connected	Πλήρως συνδεδεμένα
Feature map	Χάρτης χαρακτηριστικών
Activation map	Χάρτης ενεργοποίησης
Aggregation function	Συνάρτηση συνάθροισης
Auxiliary classifier	Βοηθητικός ταξινομητής
Batch normalization	Κανονικοποίηση δέσμης
Label smoothing regularization	Συστηματοποίηση εξομάλυνσης ετικετών
Peak Signal-to-Noise Ratio	Λόγος Σήματος-προς-Θόρυβο
Pointwise group convolution	Σημειακή ομαδική συνέλιξη
Depth-wise separable convolution	Διαχωρισμός συνέλιξης κατά βάθος
Negative log likelihood	Αρνητική λογαριθμική πιθανότητα
Quantization noise	Θόρυβος κβαντοποίησης
Clinical Neuroscience	Κλινική Νευροεπιστήμης
Behavioural Science Institute	Ινστιτούτο Επιστήμης της Συμπεριφοράς
Big Data	Μεγάλα Δεδομένα
Cloud computing	Υπολογιστικό νέφος
Attention mechanisms	Μηχανισμοί προσοχής
Speech emotion recognition	Αναγνώριση συναισθημάτων ομιλίας

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

ΠαΔΑ	Πανεπιστήμιο Δυτικής Αττικής
UniWA	University of West Attica
ΤΜΠΥ	Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών
ΑΣΠΑΙΤΕ	Ανώτατη Σχολή Παιδαγωγικής και Τεχνολογικής Εκπαίδευσης
ΜΜ	Μηχανική Μάθηση
ΒΜ	Βαθιά Μάθηση
(Σ)ΝΔ	(Συνελικτικά) Νευρωνικά Δίκτυα
ΑΤ	Ακρίβεια Ταξινόμησης
ΧΕ	Χρόνος Εκπαίδευσης
ASR	Automatic Speech Recognition
SID	Speaker Identification
MIR	Music Information Retrieval
(Q)SVM	(Quadratic) Support Vector Machines
kNN	k Nearest Neighbors
PCA(N)	Principal Component Analysis (Network)
LDA	Linear Discriminant Analysis
mRMR	Maximum Relevance Minimum Redundancy
t-SNE	t-distributed Stochastic Neighbor Embedding
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ANN	Artificial Neural Network
RMS(E)	Root Mean Square (Error)
Time (Max) ACF	Time (Maximum) Autocorrelation Coefficients
PPM	Peak Program Meter
ZCR	Zero Crossing Rate
STFT	Short Time Fourier Transform
(D)MFCC	(Derivative) Mel Frequency Cepstral Coefficients
DCT	Discrete Cosine Transformation
Pitch Spectral HPS	Pitch Spectral Harmonic Product Spectrum
ACF	Autocorrelation Function
Pitch Time ACF	Pitch Time Autocorrelation Function

Pitch Time AMDF	Pitch Time Average Magnitude Difference Function
SVD	Singular Value Decomposition
BSS	Blind Source Separation
ReLU	Rectified Linear Units
FLOPs	Float-point(s)
NAS	Network Architecture Search
CRNN	Convolutional Recurrent Neural Networks
SED	Sound Event Detection
CWT	Continuous Wavelet Transform
SGDM	Stochastic Gradient Descent with Momentum
Adam	Adaptive Moment Estimation
RMSProp	Root Mean Square Propagation
AdaGrad	Adaptive Gradient Algorithm
Std	Standard Deviation
Avg	Average
DWT	Discrete Wavelet Transform
EEG	Electroencephalography
IoT	Internet of Things
SST	Sum of squares total
SSR	Sum of squares regression
SSE	Sum of squares error
FER	Facial Emotion Recognition
SIFT	Scale Invariant Feature Transform
FAST	Features from Accelerated Segment Test
SURF	Speed-Up Robust Feature
BRIEF	Binary Robust Independent Elementary Features
ORB	Oriented Fast and Rotated Robust Independent Elementary Features
LBP	Local Binary Patterns
HOG	Histogram of Oriented Gradients
CBD	Compact Binary Descriptor
LTP	Local Ternary Pattern
LPQ	Local Phase Quantization
BoVW	Bag-of-Visual-Words
LFW	Labeled Faces in the Wild

KDEF	Karolinska Directed Emotional Faces
JAFFE	Japanese Female Facial Expression
RaFD	Radboud Faces Database
PSNR	Peak Signal-to-Noise Ratio
CoV	Coefficient of Variance
WASN	Wireless Acoustic Sensor Network
YOLO	You Only Look Once
Mixed-NMS	Mixed Network Management System
Fps	Frames per second
ESE-ShuffleNet	Expanded Squeeze-and Excitation ShuffleNet
NLL	Negative Log Likelihood
SSD	Single Shot Detection
SER	Speech Emotion Recognition

1. Ταξινόμηση του ήχου με μηχανισμούς Μηχανικής Μάθησης

Ο ήχος είναι πανταχού παρόν, τόσο σε δραστηριότητες εξωτερικών χώρων όσο και εσωτερικών, είτε αυτός παράγεται από ανθρώπινες δραστηριότητες είτε από πηγές του περιβάλλοντος. Αποτελεί βασικό σήμα αναγνώρισης του χώρου, των δραστηριοτήτων που λαμβάνουν χώρα και των πηγών που παράγουν τους εκάστοτε ήχους. Ο ήχος είναι ένα σύνθετο, πλούσιο σε χαρακτηριστικά σήμα, και η ταξινόμησή του έχει προσελκύσει έντονο ερευνητικό ενδιαφέρον αποτελώντας ένα ξεχωριστό ερευνητικό πεδίο με αυξανόμενο αριθμό εφαρμογών, όπως η παρακολούθηση αστικών δραστηριοτήτων, η αξιολόγηση του θορύβου, η επίβλεψη βιομηχανικών διεργασιών, καθώς και η αλληλεπίδραση ανθρώπου-υπολογιστή. Η αυτόματη ταξινόμηση (classification) του ήχου μπορεί να βασίζεται σε χαρακτηριστικά (features) τα οποία εξάγονται από το ηχητικό σήμα, χρησιμοποιώντας τεχνικές Μηχανικής Μάθησης (ΜΜ) ή σε τεχνικές Βαθιάς Μάθησης (ΒΜ) με χρήση κυρίως Συνελικτικών Νευρωνικών Δικτύων (ΣΝΔ). Στην παρούσα διατριβή έχουν αναπτυχθεί τόσο αλγόριθμοι ΜΜ όσο και τεχνικές ΒΜ.

1.1 Εισαγωγή

Η ταξινόμηση του ήχου με μηχανισμούς ΜΜ περιλαμβάνει πολλαπλά βήματα. Ανάλογα με την περίπτωση μελέτης, προσδιορίζεται ένα σύνολο βασικών τύπων (κλάσεων) ήχου και συγκεντρώνεται ένα σετ ήχων με ετικέτες (labels) τις κλάσεις αυτές. Οι αλγόριθμοι ΜΜ για την ταξινόμηση του ήχου μπορούν να υποστηριχθούν από το μεγάλο πλήθος των χαρακτηριστικών αυτού. Το ηχητικό σήμα συνήθως δεν χρησιμοποιείται στην ακατέργαστη μορφή του λόγω της πολυπλοκότητάς του και του αυξανόμενου αριθμού δεδομένων δειγματοληψίας. Αντιθέτως, χρησιμοποιούνται τα εξαγόμενα χαρακτηριστικά γνωρίσματα του ήχου ή/και άλλες αναπαραστάσεις αυτού (όπως π.χ. το φασματογράφημα (spectrogram) και το διάγραμμα κλίμακας-χρόνου (scalogram)), για την εκπαίδευση και τον έλεγχο των μοντέλων ταξινόμησης.

Η επιλογή των χαρακτηριστικών είναι ένα κρίσιμο σημείο της διαδικασίας, καθώς η ακρίβεια της ταξινόμησης (classification accuracy) συνδέεται άμεσα με τον αριθμό και την περιγραφική *ικανότητα* αυτών. Από την άλλη πλευρά, ένας περιττά μεγάλος αριθμός χαρακτηριστικών περιπλέκει την ανάπτυξη και την εκτέλεση των μοντέλων ταξινόμησης και, σε ορισμένες περιπτώσεις, π.χ. λόγω υπερπροσαρμογής (overfitting), υποβαθμίζει την απόδοση τέτοιων μοντέλων. Η ταχεία ανάπτυξη των τεχνικών ΜΜ αυξάνει το ενδιαφέρον για τις μεθοδολογίες επιλογής των χαρακτηριστικών και την μείωση της διαστασιολόγησης (dimensionality reduction). Η συσχέτιση του τύπου και του αριθμού των επιλεγμένων χαρακτηριστικών με την αναμενομένη ακρίβεια των μοντέλων ταξινόμησης, υποστηριζόμενη από μεθοδολογίες εκτίμησης των βαρών των χαρακτηριστικών, μπορεί να προωθήσει την τεκμηριωμένη επιλογή χαρακτηριστικών.

Η μελέτη της περίπτωσης των εσωτερικών χώρων και συγκεκριμένα μίας αίθουσας διδασκαλίας ή ενός εργαστηρίου, κατά την διάρκεια της δια ζώσης εκπαίδευσης, μπορεί να αποτελέσει πρόκληση λόγω της πολλαπλής και δυναμικής αλλαγής των καταστάσεων και των τύπων ήχου που μπορεί να συνδέονται με την συμμετοχή και την προσοχή των σπουδαστών, τα χαρακτηριστικά της παράδοσης και τα στυλ διδασκαλίας (π.χ. διάλεξη, συζήτηση, πείραμα κ.λπ.).

1.2 Στόχοι

Οι στόχοι της έρευνας είναι οι εξής:

- Ο προσδιορισμός των βασικότερων μεγεθών που αποτελούν τα χαρακτηριστικά του ηχητικού σήματος με βάση τις χρονικές, φασματικές και αντιληπτές ιδιότητες αυτού
- Η ανάπτυξη αλγορίθμων για την εξαγωγή των τιμών των χαρακτηριστικών αυτών
- Η ανάπτυξη μεθόδου ιεράρχησης των ηχητικών χαρακτηριστικών με βάση την περιγραφική τους ικανότητα

- Η ανάπτυξη αλγορίθμων ΜΜ για την ταξινόμηση του ήχου χρησιμοποιώντας την προτεινόμενη κατάταξη ηχητικών χαρακτηριστικών
- Η αξιολόγηση της προτεινόμενης μεθόδου με βάση τα αποτελέσματα της ακρίβειας ταξινόμησης και σύγκριση αυτών με τα αντίστοιχα αποτελέσματα της μεθόδου Relief-F

1.3 Συναφής έρευνα

Ο ήχος είναι ένα σήμα από το οποίο μπορούν να εξαχθούν πολλαπλά χαρακτηριστικά [1]. Τα χαρακτηριστικά αυτά υποστηρίζουν την αναγνώριση και την ταξινόμηση του ήχου στο οικιακό περιβάλλον [2], στον χώρο εργασίας [3], και σε αστικά περιβάλλοντα [4]. Οι εξωτερικοί ήχοι, με την μορφή της ηχορρύπανσης, υποβάλλονται σε επεξεργασία και αναλύονται στα [5], [6], καθώς και οι ήχοι του περιβάλλοντος [7]. Για την ταξινόμηση μπορούν να χρησιμοποιηθούν μπορούν να χρησιμοποιηθούν περίπλοκα μοντέλα ΜΜ, συμπεριλαμβανομένης της πλατφόρμας ταξινόμησης νευρωνικού δικτύου δύο επιπέδων που σχεδιάστηκε στο [8] για τον περιβαλλοντικό θόρυβο. Η αναγνώριση του τοπίου εσωτερικών και εξωτερικών χώρων που πραγματοποιείται στα [9] και [10] τα οποία επικεντρώνονται στην ανάκτηση 10 κοινών αστικών ήχων συνδυάζοντας χρονικά και φασματικά χαρακτηριστικά αλλά και περιγραφικά στατιστικά μεγέθη. Οι τοποθεσίες και οι δραστηριότητες του περιβάλλοντος διακρίνονται στο [11]. Περαιτέρω εφαρμογές περιλαμβάνουν την αυτόματη αναγνώριση ομιλίας (Automatic Speech Recognition – ASR) [12], την αναγνώριση της ταυτότητας του ομιλητή (Speaker Identification – SID) [13], και την ανάκτηση πληροφοριών μουσικής (Music Information Retrieval – MIR) [14]. Παράλληλα, κόμβοι περιορισμένων δυνατοτήτων μπορούν να χρησιμοποιηθούν σε συνδυασμό με αλγορίθμους ΜΜ, όπως στο [15] όπου χρησιμοποιήθηκαν κόμβοι Raspberry-pi και οι Μηχανισμοί Υποστήριξης Διανυσμάτων (Support Vector Machines – SVM) και κ-Πλησιέστερων Γειτόνων (k-Nearest Neighbors – kNN).

Η αναγνώριση ήχων σε αίθουσα διδασκαλίας έχει προσελκύσει το ερευνητικό ενδιαφέρον για την εξαγωγή υψηλού επιπέδου εννοιολογικών πληροφοριών. Ο εντοπισμός των ερωτήσεων του καθηγητή κατά τη διάρκεια της παράδοσης περιγράφεται στο [16]. Στο [17] οι συγγραφείς χρησιμοποιούν την ανίχνευση ομιλίας και την καταγραφή των ομιλητών για την ανάλυση δραστηριοτήτων και των συνθηκών στην τάξη. Σε αυτή την εργασία επιλέγεται και αξιολογείται ένας σημαντικός αριθμός χαρακτηριστικών χαμηλού επιπέδου πριν την τροφοδότηση με αυτά σε αλγορίθμους ταξινόμησης. Στο [18] χρησιμοποιούνται νευρωνικά δίκτυα συνελικτικού τύπου για την έξυπνη παρατήρηση της τάξης, με βάση δεδομένα ήχου, μετά την μετατροπή τους σε εικόνες φασματογράμματος Mel. Οι κλάσεις δραστηριότητας προς αναγνώριση σχετίζονται με την ροή εργασιών στην τάξη. Στο [19] οι συγγραφείς χρησιμοποιούν ταξινομητές ΒΜ για τον αυτόματο σχολιασμό δραστηριοτήτων, αξιολογώντας μία συλλογή καταγραφών από αίθουσες διδασκαλίας ως «μονοφωνικά», «πολυφωνικά», «χωρίς φωνή» ή «άλλα». Ο σχεδιασμός αυτοματοποιημένων εργαλείων για την αναγνώριση δραστηριοτήτων με βάση την ΜΜ εμφανίζεται ως ο κοινός παρονομαστής των ερευνών σε αυτόν τον τομέα.

Όσον αφορά τις ηχητικές παραμέτρους, το ηχητικό σήμα μπορεί να περιγραφεί χρησιμοποιώντας πολλαπλά χρονικά, φασματικά και αντιληπτά χαρακτηριστικά. Τα χαρακτηριστικά αυτά περιλαμβάνουν στατιστικές ιδιότητες, ηχόχρωμα, χαρακτηριστικά που σχετίζονται με την ένταση (περιβάλλουσα, στάθμη και ένταση), τονικά και, χρονικά χαρακτηριστικά (ρυθμικές και χρονικές ιδιότητες). Ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται στην ταξινόμηση μπορεί να επηρεάσει θετικά την ακρίβεια ταξινόμησης αλλά, θέτει επίσης προκλήσεις όσον αφορά τους υπολογιστικούς πόρους και την πολυπλοκότητα του συστήματος. Οι μεθοδολογίες επιλογής χαρακτηριστικών μπορούν να προσφέρουν μία συμβιβαστική λύση.

Ο αλγόριθμος Relief-F χρησιμοποιείται για την επιλογή χαρακτηριστικών με βάση την συνάφεια και τον πλεονασμό κάθε χαρακτηριστικού (ανεξάρτητα από τον αλγόριθμο μοντελοποίησης των δεδομένων). Ο αλγόριθμος υπολογίζει τα βάρη των στοιχείων πρόβλεψης, δίνοντας μειονεκτική θέση στα στοιχεία που δίνουν διαφορετικές τιμές σε γείτονες της ίδιας κλάσης και πλεονεκτική

θέση σε αυτά τα στοιχεία που δίνουν διαφορετικές τιμές σε γείτονες διαφορετικών κλάσεων. Η απόδοση του αλγορίθμου έχει αναλυθεί στα [20], [21]. Οι ταξινομητές μονής μεταβλητής χρησιμοποιούνται για την κατάταξη και την επιλογή χαρακτηριστικών λαμβάνοντας υπόψη την πιθανή σύνδεση με τον ταξινομητή [22]. Ο αλγόριθμος βελτιστοποίησης Forest έχει χρησιμοποιηθεί για την επιλογή χαρακτηριστικών σε συνεργασία με την προεπεξεργασία δεδομένων η οποία βασίζεται στην τεχνική ελάχιστου πλεονασμού και μέγιστης συνάφειας (maximum Relevance Minimum Redundancy - mRMR) για την αφαίρεση των λιγότερο σημαντικών χαρακτηριστικών από το σύνολο των χαρακτηριστικών [23]. Η τ-διανεμημένη στοχαστική ενσωμάτωση γειτόνων (t-distributed Stochastic Neighbor Embedding – t-SNE) [24], απεικονίζει τα δεδομένα υψηλής διάστασης σε χώρο χαμηλότερων διαστάσεων (συνήθως 2D ή 3D) για σκοπούς οπτικοποίησης. Στην συγκεκριμένη δημοσίευση υλοποιείται ο αλγόριθμος Relief-F για σύγκριση με την προτεινόμενη μέθοδο επιλογής χαρακτηριστικών.

Η επιλογή της μεθοδολογίας ταξινόμησης και η ρύθμιση των υπερπαραμέτρων μπορεί να δημιουργήσει προκλήσεις ανά τομέα εφαρμογής και τύπο προβλήματος. Οι παραγωγικοί αλγόριθμοι που ορίζονται σε πλαίσιο Bayesian περιλαμβάνουν τα Gaussian Mixture Models (GMM) και τα Hidden Markov Models (HMM). Οι διακριτικοί αλγόριθμοι περιλαμβάνουν τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks – ANN) και τους μηχανισμούς SVM. Υβριδικές περιπτώσεις έχουν διερευνηθεί στο [25], ενώ ταξινομητές διαφορετικών κατηγοριών (SVM και GMM) στο [26]. Οι τεχνικές εκτίμησης της ακρίβειας μπορούν να υποστηρίξουν την επιλογή του αλγορίθμου ταξινόμησης. Η επίδραση των υπερπαραμέτρων στην απόδοση του αλγορίθμου, π.χ. η ασυμπτωτική συμπεριφορά των SVM με Gaussian πυρήνα (kernel), διερευνάται στο [27].

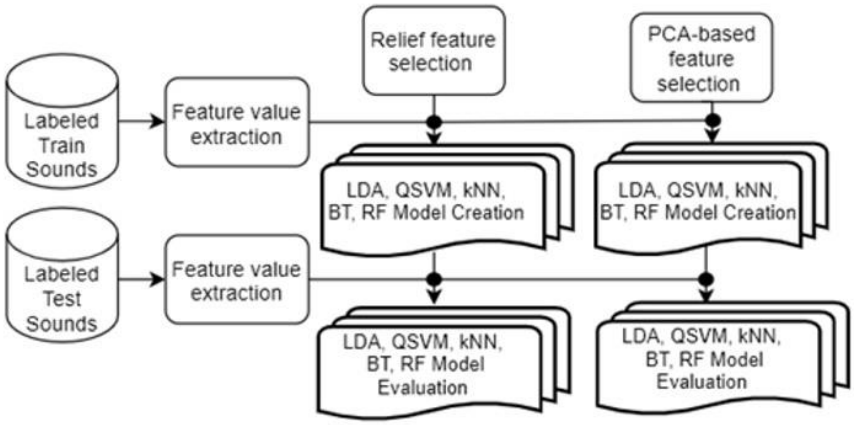
1.4 Ροή εργασιών - Μεθοδολογία

Αρχικά προσδιορίστηκε ένα σύνολο 143 χαρακτηριστικών τα οποία περιγράφουν το ηχητικό σήμα με βάση τις χρονικές, φασματικές και αντιληπτές ιδιότητές του. Στην συνέχεια, προτάθηκε μέθοδος ιεράρχησης αυτών των ηχητικών χαρακτηριστικών, η οποία βασίζεται στην Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis – PCA). Καθορίστηκαν έξι κλάσεις ήχων ενδιαφέροντος για την περίπτωση της μελέτης της εκπαίδευσης εκ του σύνεγγυς. Οι κλάσεις αυτές περιλαμβάνουν την ομιλία ενός (είτε του καθηγητή, είτε μαθητή), την συνομιλία (δύο ή περισσότεροι ομιλούντες), τον ήχο από πόρτα/καρέκλα/θρανίο, των ήχο βιβλίων και χαρτιών, τον θόρυβο (πολλοί ταυτόχρονοι ήχοι) και την ησυχία. Ένα σύνολο αρχείων ηχογραφήθηκε υπό ρεαλιστικές και ελεγχόμενες συνθήκες και εμπλουτίστηκε από υπάρχοντα σύνολα δεδομένων που συμπεριλαμβάνουν τις έξι κατηγορίες ενδιαφέροντος. Τα αρχεία ήχου κατατμήθηκαν σε πλαίσια (της ίδιας κλάσης) διάρκειας ενός δευτερολέπτου και έχουν επισημανθεί ως μία από τις κλάσεις ήχου που προαναφέρθηκαν. Πραγματοποιήθηκε η εξαγωγή των τιμών των 143 ηχητικών χαρακτηριστικών από κάθε πλαίσιο χρησιμοποιώντας τεχνικές επεξεργασίας ήχου.

Η προτεινόμενη μεθοδολογία κατάταξης χαρακτηριστικών με βάση τα βάρη της PCA, καθώς και ο υπάρχον μηχανισμός επιλογής χαρακτηριστικών Relief-F εφαρμόστηκαν για την δημιουργία δύο κατατάξεων των ηχητικών χαρακτηριστικών. Αυτές οι κατατάξεις χρησιμοποιήθηκαν για την δημιουργία μοντέλων ταξινόμησης με πέντε αλγορίθμους ΜΜ. Συγκεκριμένα υλοποιήθηκαν, α) μοντέλο Γραμμικής Διακριτικής Ανάλυσης (Linear Discriminant Analysis – LDA), β) μοντέλο Τετραγωνικού Μηχανισμού Διανυσματικής Υποστήριξης (Quadratic SVM – QSVM), γ) μοντέλο kNN, δ) μοντέλο των Ενισχυμένων Δένδρων (Boosted Trees) και δ) το μοντέλο Τυχαίου Δάσους (Random Forest). Δημιουργήθηκαν 143 μοντέλα ανά αλγόριθμο με αυξανόμενο αριθμό χαρακτηριστικών, με βάση και τις δύο κατατάξεις χαρακτηριστικών, δημιουργώντας συνολικά 1430 μοντέλα.

Οι ταξινομητές αξιολογούνται ως προς την αποδιδόμενη ακρίβεια ταξινόμησης. Οι Πίνακες Ειδικότητας (Specificity), Ευαισθησίας (Sensitivity) και Σύγχυσης (Confusion Matrix)

κατασκευάζονται για τους δύο αλγορίθμους με τις καλύτερες επιδόσεις (LDA και QSVM). Με βάση την ακρίβεια ταξινόμησης των μοντέλων, αξιολογείται η κατάταξη των χαρακτηριστικών και η επίδραση του αριθμού των χαρακτηριστικών που συμμετέχουν στην ταξινόμηση. Στο Σχήμα 1.1 παρουσιάζεται η ροή εργασιών της πειραματικής διαδικασίας.

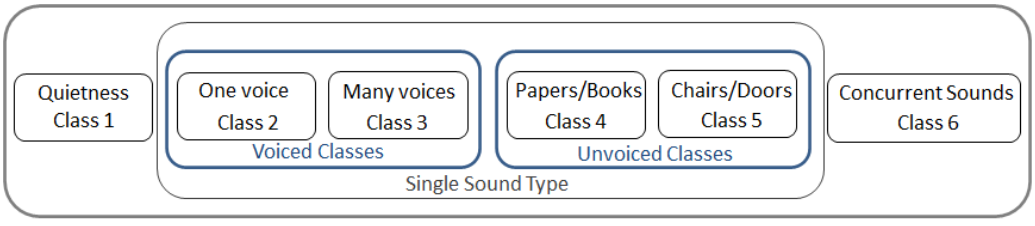


Σχήμα 1.1: Ροή εργασιών για την ταξινόμηση του ήχου με χρήση πέντε αλγορίθμων ΜΜ και δύο κατατάξεων των ηχητικών παραμέτρων.

1.4.1 Κλάσεις ταξινόμησης και σετ ήχου

Οι τύποι των ήχων που ορίζονται είναι αυτοί που μπορούν να υποστηρίξουν την αναγνώριση των δραστηριοτήτων οι οποίες συμβαίνουν στον χώρο του μαθήματος. Κάθε τύπος ήχου είναι ενδεικτικός για διαφορετική φάση της εκπαιδευτικής διαδικασίας. Αυτοί περιλαμβάνουν φωνητικούς ήχους, όπως παραδείγματος χάριν την διάλεξη του καθηγητή ή τον διάλογο με τους σπουδαστές, καθώς και μη φωνητικούς ήχους οι οποίοι σχετίζονται με κινήσεις, όπως το άνοιγμα και το κλείσιμο των θυρών, η μετακίνηση καρεκλών καθώς και ήχοι που σχετίζονται με δραστηριότητες των σπουδαστών όπως η χρήση βιβλίων και εγγράφων. Τέλος, περιλαμβάνεται η περίπτωση που συνυπάρχουν οι προηγούμενοι τύποι ήχου όπως συμβαίνει κατά τη διάρκεια ενός διαλείμματος. Συνολικά ορίστηκαν έξι κλάσεις ήχου όπως απεικονίζονται στο Σχήμα 1.2, και είναι οι εξής:

α. Ησυχία (κανένας ήχος ή ήχοι χαμηλής έντασης)
β. Ένας ομιλητής (καθηγητής ή σπουδαστής)
γ. Πολλοί ομιλητές (δύο ή και περισσότερες φωνές ταυτόχρονα)
δ. Χαρτιά/βιβλία
ε. Θύρες/καρέκλες (άνοιγμα/κλείσιμο θυρών και μετακίνηση καρεκλών)
στ. Θόρυβος (ήχοι που ακούγονται ταυτόχρονα χωρίς να υπάρχει επικρατέστερος)



Σχήμα 1.2: Κλάσεις ήχου.

Το σετ ήχου που χρησιμοποιήθηκε για την εκπαίδευση και τον έλεγχο των μοντέλων ταξινόμησης αποτελείται από α) αρχεία ήχου τα οποία έχουν εγγραφεί υπό ρεαλιστικές και ελεγχόμενες συνθήκες (με περιορισμένο θόρυβο) κατά την διάρκεια δια ζώσης εκπαιδευτικής διαδικασίας στο Ανώτατο Εκπαιδευτικό Ίδρυμα της ΑΣΠΑΙΤΕ στην διάρκεια εργαστηριακών μαθημάτων και β) αρχεία ήχου από υπάρχουσες βάσεις δεδομένων και διαδικτυακές πηγές. Στην πρώτη κατηγορία

η ηχογράφηση ήταν μονοφωνική και χρησιμοποιήθηκε φορητή συσκευή υψηλής ακρίβειας εγγραφής (Philips DVT4010). Η συχνότητα δειγματοληψίας είναι τα 44.1 kHz και η διάρκεια της ηχογράφησης περίπου δύο ώρες. Στην δεύτερη κατηγορία έχουν συμπεριληφθεί αρχεία από το Google Audio και από το σύνολο ήχων TUT Acoustic Scenes [28]. Από το πλήρες σύνολο των δύο αυτών βάσεων δεδομένων συμπεριλήφθηκε το υποσύνολο των ήχων εσωτερικών χώρων που σχετίζονται με την συγκεκριμένη περίπτωση μελέτης, όπως παραδείγματος χάριν του σπιτιού, του γραφείου και της βιβλιοθήκης.

Το σετ ήχου περιλαμβάνει όλες τις κλάσεις σε ελαφρώς άνισα μέρη (ησυχία = 17%, ένας ομιλητής = 17%, πολλοί ομιλητές = 18.6%, χαρτιά/βιβλία = 17%, θύρες/καρέκλες = 16.4%, θόρυβος = 14.7%). Το σετ του ήχου έχει επιμεριστεί σε πλαίσια (της ίδιας κλάσης) διάρκειας ενός δευτερολέπτου τα οποία έχουν ετικέτα μία από τις έξι κλάσεις ενδιαφέροντος. Κάθε πλαίσιο έχει υποστεί επεξεργασία για να εξαχθούν οι 143 τιμές των χαρακτηριστικών του ήχου που τροφοδοτούν του ταξινομητές.

1.4.2 Σύνολο χαρακτηριστικών ήχου

Τα ηχητικά σήματα χαρακτηρίζονται από ένα πλούσιο σύνολο χαρακτηριστικών χαμηλού-επιπέδου τα οποία ανήκουν σε τρεις κατηγορίες: χρονικά, φασματικά και αντιληπτά [29]. Τα **χρονικά** χαρακτηριστικά περιγράφουν την χρονικής εξέλιξη του σήματος και υπολογίζονται από την χρονική περιβάλλουσα της κυματομορφής του σήματος. Στην παρούσα έρευνα περιλαμβάνονται:

1. Η Μέση Τετραγωνική Ρίζα (Time Root Mean Square – Time RMS) ανά χρονικό πλαίσιο.
2. Η Τυπική Απόκλιση (Standard Deviation)
3. Οι Συντελεστές Μέγιστης Χρονικής Αυτοσυσχέτισης (Time Maximum Autocorrelation Coefficients – Time Max ACF), οι οποίοι υποδεικνύουν την ομοιότητα ενός σήματος με αντίγραφό του το οποίο είναι χρονικά μετατοπισμένο για την αναγνώριση της περιοδικότητας. Όσο λιγότερο περιοδικό (επομένως και λιγότερο τονικό) είναι ένα σήμα τόσο μικρότερη είναι και η τιμή τέτοιων μεγίστων [30].
4. Οι Χρονικοί Συντελεστές Αυτοσυσχέτισης (Time ACF Coefficients), οι οποίοι περιγράφουν την φασματική κατανομή του σήματος στο πεδίο του χρόνου [31].
5. Οι Τιμές Αιχμής (Peak Values) σε απόλυτη τιμή της περιβάλλουσας για κάθε χρονικό πλαίσιο, οι οποίες κυμαίνονται στο εύρος τιμών [0, 1].
6. Ο Μετρητής Αιχμής Προγράμματος (Peak Program Meter – PPM), ο οποίος λειτουργεί με διαφορετικό χρόνο ολοκλήρωσης για τον χρόνο «επίθεσης» (attack time) και για τον χρόνο «αποδέσμευσης» (release time). Οι τιμές επίσης κυμαίνονται στο εύρος [0, 1].
7. Ο Λόγος Χρονικής Πρόβλεψης (Time Predictivity Ratio), ο οποίος μετράει το σφάλμα μεταξύ του αρχικού δείγματος και του δείγματος που έχει προβλεφθεί με βάση την προηγούμενη τιμή του δείγματος. Όσο περισσότερο θορυβώδες είναι το ηχητικό σήμα τόσο μεγαλύτερο είναι το σφάλμα, δεν υπάρχουν στατιστικές σχέσεις μεταξύ των δειγμάτων και επομένως δεν μπορεί να γίνει πρόβλεψη.
8. Ο Ρυθμός Διέλευσης από το Μηδέν (Zero Crossing Rate - ZCR), ο οποίος υποδεικνύει τον αριθμό των διαδοχικών δειγμάτων που έχουν διαφορετικό πρόσημο.

Τα **φασματικά** χαρακτηριστικά τα οποίο περιγράφουν το φασματικό *σχήμα* του ηχητικού σήματος είναι στενά συνδεδεμένα με το ηχόχρωμα ή αλλιώς την ποιότητα ή υφή του ήχου. Οι τιμές των χαρακτηριστικών υπολογίζονται χρησιμοποιώντας τον σύντομο μετασχηματισμό Fourier (Short-Time Fourier Transform – STFT) του σήματος. Τα φασματικά χαρακτηριστικά που χρησιμοποιήθηκαν είναι τα εξής:

1. Η φασματική Κλίση (Slope), η οποία αντιπροσωπεύει το φασματικό πλάτος σε σχέση με την συχνότητα (dB/octave), και υπολογίζεται με γραμμική παλινδρόμηση (linear regression) του φασματικού πλάτους.

2. Η φασματική Ροή (Flux), η οποία μετράει το μέγεθος της αλλαγής του φασματικού σχήματος μεταξύ διαδοχικών πλαισίων STFT, και ορίζεται ως το τετράγωνο της μέσης τιμής της διαφοράς μεταξύ κανονικοποιημένων πλατών φάσματος.
3. Η Φασματική Μείωση (Spectral Decrease), η οποία μετράει την κλίση της μείωσης της φασματικής περιβάλλουσας ως προς την συχνότητα.
4. Ο Λόγος Τονικής Ισχύος (Tonal Power Ratio), είναι ο λόγος της τονικής ισχύος προς την συνολική ισχύ. Χαμηλές τιμές του Λόγου Τονικής Ισχύος υποδεικνύουν ένα σήμα με θόρυβο ή ένα σιωπηλό σήμα.
5. Η φασματική Επιπεδότητα (Flatness), είναι ο λόγος του γεωμετρικού μέσου προς τον αριθμητικό μέσο.
6. Η φασματική Κύρτωση (Kurtosis), η οποία μετρά την επιπεδότητα (ή την «αιχμηρότητα» αντίστοιχα) της κατανομής σε σύγκριση με την κατανομή Gauss.
7. Η φασματική Κορυφή (Crest), η οποία είναι μέτρο της τονικότητας και συγκρίνει το μέγιστο φασματικό πλάτος με το άθροισμα του φασματικού πλάτους. Χαμηλές τιμές της Κορυφής υποδεικνύουν επίπεδο φασματικό πλάτος ενώ αντίθετα, υψηλές τιμές ημιτονοειδές ηχητικό σήμα.
8. Η φασματική Ασυμμετρία (Skewness), η οποία υπολογίζεται για να χαρακτηρίσει την ασυμμετρία της κατανομής γύρω από τη μέση τιμή της. Η τιμή μηδέν χαρακτηρίζει την συμμετρική κατανομή, αρνητικές τιμές τις δεξιο-σταθμισμένες, και θετικές τις αριστερο-σταθμισμένες κατανομές [32].
9. Το φασματικό Κεντροειδές (Centroid) ή αλλιώς την Φωτεινότητα (Brightness), η οποία αποτελεί το κέντρο βάρους της φασματικής ενέργειας.
10. Η φασματική Έκταση (Spread), η οποία περιγράφει την φασματική συγκέντρωση της φασματικής ισχύος γύρω από το Κεντροειδές, δηλαδή αποτελεί την τυπική απόκλιση του φάσματος γύρω από την μέση τιμή του.
11. Η φασματική Πτώση (Roll-off), η οποία ορίζεται ως η συχνότητα συγκεκριμένου ποσοστού της συνολικής ενέργειας του σήματος.

Τα **αντιληπτά** χαρακτηριστικά βασίζονται σε μοντέλα προσομοίωσης της διαδικασίας ακοής του ανθρώπου. Τα αντιληπτά χαρακτηριστικά που χρησιμοποιήθηκαν στην παρούσα εργασία είναι τα εξής:

1. Οι συχνότητες Mel. Το ανθρώπινο σύστημα ακοής δεν ερμηνεύει όλες τις ζώνες συχνοτήτων εξίσου. Οι συχνότητες Mel αντιστοιχίζουν την πραγματική συχνότητα με τον αντιληπτό τόνο. Η μετατροπή από Hz σε κλίμακα Mel πραγματοποιείται με βάση την σχέση:

$$f_{MEL} = 2595log[(f/700) + 1] \quad (1.1)$$

όπου f είναι η πραγματική συχνότητα σε Hz και f_{MEL} η προκύπτουσα σε κλίμακα Mel. Στην ανάλυση Fourier ο όρος cepstrum είναι το αποτέλεσμα του υπολογισμού του αντίστροφου μετασχηματισμού Fourier του λογαρίθμου του εκτιμώμενου φασματικού σήματος. Για τον υπολογισμό του Συντελεστών Συχνοτήτων Mel Cepstral (Mel Frequency Cepstral Coefficients – MFCC) καθώς και της πρώτης και δεύτερης παραγώγου τους (DMEL, DDMEL) [33] το ηχητικό σήμα διαιρείται σε πλαίσια (frames). Οι συχνότητες μετατρέπονται στην κλίμακα Mel μέσω της σχέσης (1.1) και το φάσμα αναπαρίσταται σε άξονα συχνότητας Mel. Ο άξονας Mel χωρίζεται σε τριγωνικά φίλτρα ζώνης. Το άθροισμα των φασματικών πλατών κάθε μπάντας υπολογίζεται και οι cepsrtal συντελεστές εξάγονται εφαρμόζοντας Διακριτό Μετασχηματισμό Συνημιτόνου (Discrete Cosine Transform – DCT) στον λογάριθμο των σταθμισμένων κατά Mel φίλτρο ζωνών. Στην παρούσα εργασία υπολογίστηκαν 40 συντελεστές Mel για κάθε ηχητικό πλαίσιο οι οποίοι καλύπτουν το εύρος των συχνοτήτων [133, 6854] Hz, με τις 13 πρώτες μπάντες να χωρίζονται γραμμικά και οι υπόλοιπες 27 λογαριθμικά. Επίσης, υπολογίστηκαν οι παράγωγοι $1^{ης}$ και $2^{ης}$ τάξης των συντελεστών αυτών με αποτέλεσμα συνολικά 120 τιμές.

2. Το Τονικό Φασματικό Αρμονικό Γινόμενο Φάσματος (Pitch Spectral Harmonic Product Spectrum – Pitch Spectral HPS), το οποίο ανιχνεύει τους τόνους μετρώντας τον μέγιστο συγχρονισμό των αρμονικών για κάθε φασματικό πλαίσιο.
3. Η Συνάρτηση της Φασματικής Αυτοσυσχέτισης (Autocorrelation Function – ACF), η οποία υπολογίζει το μέγιστο της συνάρτησης της φασματικής αυτοσυσχέτισης.
4. Η Τονική Χρονική Συνάρτηση Αυτοσυσχέτισης (Pitch Time ACF), η οποία υπολογίζει την υστέρηση της συνάρτησης αυτοσυσχέτισης.
5. Η Τονική Χρονική Συνάρτηση Διαφοράς Μέσου Πλάτους (Pitch Time Average Magnitude Difference Function – Pitch Time AMDF), η οποία υπολογίζει την υστέρηση της συνάρτησης της μέσης διαφοράς πλάτους. Η τελευταία υπερτερεί λόγω της μικρότερης πολυπλοκότητας, ενώ η Τονική Χρονική Συνάρτηση Αυτοσυσχέτισης είναι πιο κατάλληλη για την περίπτωση θορυβωδών ήχων [34].

Ο Πίνακας 1.1 περιλαμβάνει τα χαρακτηριστικά του ήχου διαχωρισμένα στις τρεις βασικές κατηγορίες. Στην παρένθεση χρησιμοποιείται ο αγγλικός όρος των χαρακτηριστικών με σύντομο τρόπο όπου είναι εφικτό. Αυτός ο όρος θα χρησιμοποιηθεί στην συνέχεια του κειμένου.

Πίνακας 1.1: Τα χαρακτηριστικά του ήχου ανά κατηγορία

Χρονικά		**Φασματικά**		**Αντιληπτά**	
1.	Μέση Τετραγωνική Ρίζα (Time RMS)	**9.**	Κλίση (Slope)	**20.**	Τονικό Φασματικό Αρμονικό Γινόμενο Φάσματος (Pitch Spectral HPS)
2.	Τυπική Απόκλιση (Time Std)	**10.**	Ροή (Flux)	**21.**	Συνάρτηση της Φασματικής Αυτοσυσχέτισης (Pitch Spectral ACF)
3.	Συντελεστές Μέγιστης Χρονικής Αυτοσυσχέτισης (Time Max ACF)	**11.**	Φασματική Μείωση (Spectral Decrease)	**22.**	Τονική Χρονική Συνάρτηση Αυτοσυσχέτισης (Pitch Time ACF)
4.	Συντελεστές Αυτοσυσχέτισης (Time ACF)	**12.**	Λόγος Τονικής Ισχύος (Tonal Power Ratio)	**23.**	Τονική Χρονική Συνάρτηση Διαφοράς Μέσου Πλάτους (Pitch Time AMDF)
5.	Τιμές Αιχμής (Peak Values)	**13.**	Επιπεδότητα (Flatness)	**24-63.**	Συντελεστές συχνότητας Mel (MFCC)
6.	Μετρητής Αιχμής Προγράμματος (PPM)	**14.**	Κύρτωση (Kurtosis)	**64-103.**	Πρώτες παράγωγοι των συντελεστών συχνότητας Mel (DMEL)
7.	Λόγος Χρονικής Πρόβλεψης (Time Predictivity Ratio)	**15.**	Κορυφή (Crest)	**103-143.**	Δεύτερες παράγωγοι των συντελεστών συχνότητας Mel (DDMEL)
8.	Ρυθμός Διέλευσης από το Μηδέν (Time ZCR)	**16.**	Ασυμμετρία (Skewness)		

	17.	Κεντροειδές (Centroid)	
	18.	Έκταση (Spread)	
	19.	Πτώση (Roll-off)	

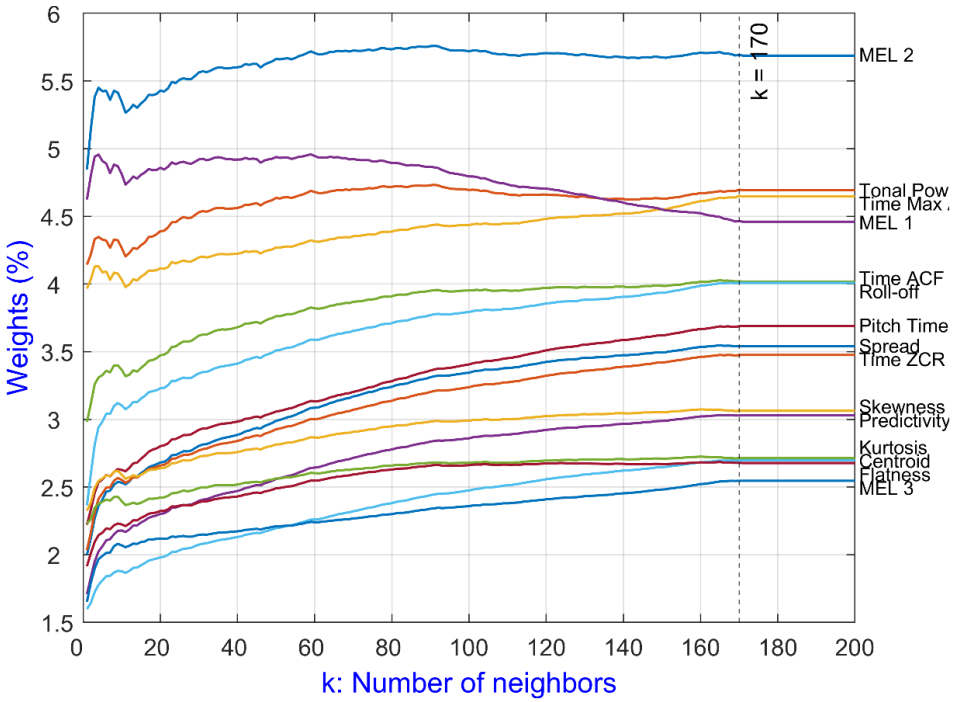
1.5 Μείωση της διαστασιολόγησης με επιλογή χαρακτηριστικών

Ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται στην ταξινόμηση μπορεί να επηρεάσει θετικά την ακρίβεια ταξινόμησης αλλά ταυτόχρονα ο μεγάλος αριθμός αυτών μπορεί να επιβαρύνει την πολυπλοκότητα, και κατ' επέκταση τους απαιτούμενους υπολογιστικούς πόρους. Οι τεχνικές μείωσης της διαστασιολόγησης μπορούν να εναρμονίσουν τον αριθμό αυτών με μεγιστοποίηση της ακρίβειας ταξινόμησης. Αυτές οι τεχνικές διακρίνονται σε δύο μεγάλες κατηγορίες: της εποπτευόμενης (supervised) μεθόδου και της χωρίς επίβλεψη (unsupervised) μεθόδου. Στην περίπτωση που η επιλογή των χαρακτηριστικών αγνοεί την μεταβλητή εξόδου η τεχνική ανήκει στις χωρίς επίβλεψη μεθόδους. Οι τεχνικές εποπτευόμενης μεθόδου διακρίνονται σε τέσσερις κύριες κατηγορίες, α) περιτυλίγματος (wrapper), στην οποία χρησιμοποιούνται μοντέλα μάθησης για την αξιολόγηση επιλεγμένων υποσυνόλων χαρακτηριστικών, β) φίλτρου (filter) στην οποία αξιολογούνται τα χαρακτηριστικά με στατιστικά μοντέλα, γ) ενσωματωμένα (embedded) στην οποία χρησιμοποιούνται μοντέλα μάθησης και οι εσωτερικοί παράγοντες ως κριτήρια αξιολόγησης και δ) υβριδικές μέθοδοι [35].

1.5.1 Επιλογή χαρακτηριστικών με την μέθοδο Relief-F

Ο αλγόριθμος Relief-F [36] εκτιμά την ποιότητα των αρχικών χαρακτηριστικών (IF) σύμφωνα με την διακριτική τους ικανότητα συγκρινόμενα με κοντινές περιπτώσεις. Ο αλγόριθμος εντοπίζει για κάθε περίπτωση τους δύο πλησιέστερους γείτονες της ίδιας και διαφορετικής κλάσης και ενημερώνει τις εκτιμήσεις ποιότητας για όλα τα χαρακτηριστικά. Συγκεκριμένα, ο αλγόριθμος αναζητά για κάθε περίπτωση τους k πλησιέστερους γείτονες της ίδιας κλάσης και τους k πλησιέστερους γείτονες από κάθε μία από τις διαφορετικές κλάσεις και υπολογίζει την μέση εκτίμηση ποιότητας για κάθε χαρακτηριστικό [21].

Μελέτες σχετικά με την συμβολή της παραμέτρου k (αριθμού των πλησιέστερων γειτόνων) στις εκτιμήσεις της μεθόδου Relief-F έχουν δείξει ότι η ποιότητα της εκτίμησης ακολουθεί διαφορετική τάση ανάλογα με το αν τα χαρακτηριστικά είναι εξαρτημένα ή όχι. Εξετάσαμε τα βάρη των 15 πρώτων χαρακτηριστικών μεταβάλλοντας την τιμή της παραμέτρου k από 1 έως 200. Τα αποτελέσματα απεικονίζονται στο Σχήμα 1.3. Παρατηρούμε ότι από την τιμή $k = 170$ και μετά η κατάταξη και τα βάρη των χαρακτηριστικών είναι σταθερά.



Σχήμα 1.3: Γραφική παράσταση των βαρών των 15 πρώτων χαρακτηριστικών σε συνάρτηση με τον αριθμό των γειτόνων (k)

Εφαρμόζουμε τον αλγόριθμο Relief-F στα δεδομένα εκπαίδευσης με $k = 170$ και η κατάταξη που προκύπτει (τα 50 πρώτα χαρακτηριστικά) παρουσιάζονται στον Πίνακα 1.2:

Πίνακας 1.2: Κατάταξη των χαρακτηριστικών με βάση την μέθοδο Relief-F (Τα 50 πρώτα)

	Χαρακτηριστικό	**Βάρος (%)**		**Χαρακτηριστικό**	**Βάρος (%)**		**Χαρακτηριστικό**	**Βάρος (%)**
1	MEL 2	5.69	**18**	MEL 17	2.17	**35**	MEL 9	0.89
2	Tonal Power Ratio	4.70	**19**	Pitch Spectral ACF	2.06	**36**	MEL 10	0.82
3	Time Max ACF	4.65	**20**	Pitch Spectral HPS	1.82	**37**	MEL 28	0.71
4	MEL 1	4.46	**21**	MEL8	1.74	**38**	MEL 6	0.55
5	Time ACF Coefficient	4.02	**22**	Time Std	1.49	**39**	MEL 5	0.54
6	Roll-off	4.01	**23**	Time RMS	1.45	**40**	MEL 15	0.53
7	Pitch Time ACF	3.69	**24**	Flux	1.45	**41**	MEL 27	0.48
8	Spread	3.54	**25**	Decrease	1.44	**42**	MEL 23	0.44
9	Time ZCR	3.47	**26**	MEL 14	1.41	**43**	MEL 18	0.44
10	Skewness	3.06	**27**	MEL 13	1.37	**44**	MEL 20	0.36
11	Predictivity	3.03	**28**	MEL 12	1.22	**45**	MEL 11	0.33
12	Kurtosis	2.71	**29**	Peak Program Meter	1.17	**46**	MEL 26	0.32

13	Centroid	2.70	**30**	Peak Values	1.15	**47**	MEL 31	0.32
14	Flatness	2.68	**31**	Slope	1.04	**48**	MEL 32	0.32
15	MEL 3	2.55	**32**	MEL 4	0.99	**49**	MEL 33	0.32
16	Crest	2.43	**33**	MEL 24	0.98	**50**	M3L 21	0.31
17	MEL 7	2.20	**34**	Pitch Time AMDF	0.91			

Τα συσσωρευμένα βάρη για την κατάταξη Relief-F για τα 25 και 50 πρώτα χαρακτηριστικά είναι αντίστοιχα: $\sum_1^{25} w_i = 73.21\%$ και $\sum_1^{50} w_i = 91.13\%$.

1.5.2 Μέθοδος επιλογής των χαρακτηριστικών βασισμένη στην PCA

Στην παρούσα εργασία προτείνεται μία μεθοδολογία επιλογής χαρακτηριστικών βασισμένη στην PCA. Η μέθοδος PCA είναι μία από τις πιο γνωστές και ευρέως διαδεδομένες μεθόδους στατιστικής για την μείωση διαστασιολόγησης, η οποία αντιστοιχίζει τα δεδομένα μεγάλων διαστάσεων σε δεδομένα μικρότερων διαστάσεων ενώ ταυτόχρονα η διακύμανση των δεδομένων στον χώρο των μικρότερων διαστάσεων γίνεται μέγιστη. Οι Κύριες Συνιστώσες είναι οι νέες μεταβλητές οι οποίες κατασκευάζονται σαν γραμμικοί συνδυασμοί των αρχικών μεταβλητών. Αυτές οι νέες μεταβλητές δημιουργούνται έτσι ώστε να είναι μεταξύ τους ορθογώνιες (μη συσχετισμένες) και οργανώνονται έτσι ώστε το μεγαλύτερο μέρος της πληροφορίας (η μεγαλύτερη διακύμανση) να περιέχεται στις πρώτες από αυτές. Αγνοώντας τις συνιστώσες των τελευταίων θέσεων αλλά παράλληλα, διατηρώντας την μεγαλύτερη ποσότητα πληροφορίας, επιτυγχάνεται η μείωση των διαστάσεων με μικρό κόστος στην ακρίβειας της ταξινόμησης αλλά με σημαντική εξοικονόμηση στους απαιτούμενους υπολογιστικούς πόρους και την πολυπλοκότητα.

Η προτεινόμενη μέθοδος εφαρμόζεται στο σύνολο των 143 χαρακτηριστικών του ήχου που αναφέρθηκαν στην Παράγραφο 1.3.2, με στόχο την φθίνουσα κατάταξη αυτών ως προς την περιγραφική τους ικανότητα. Εν συνεχεία μελετώνται τα αποτελέσματα της ταξινόμησης των ήχων κάνοντας χρήση περιορισμένου αριθμού χαρακτηριστικών, επιλέγοντας αυτά με την μεγαλύτερη περιγραφική ικανότητα (αυτά τα οποία βρίσκονται στις πρώτες θέσεις της κατάταξης).

Η PCA, η οποία βασίζεται στον αλγόριθμο Διάσπασης Ιδιόμορφων Τιμών (Singular Value Decomposition – SVD), εφαρμόζεται σε ένα σύνολο χαρακτηριστικών και παρέχει ένα σύνολο νέων ανεξάρτητων Κύριων Συνιστωσών οι οποίες εξαρτώνται γραμμικά από τα αρχικά χαρακτηριστικά (Initial Features – IF). Οι προκύπτουσες κύριες συνιστώσες (Principal Components – PC) δημιουργούνται κατά φθίνουσα σειρά της διακύμανσης (ή αλλιώς της διασποράς και κατ' επέκταση της περιεχόμενης πληροφορίας) των συνιστωσών. Ο υπολογιζόμενος πίνακας συντελεστών (Coefficient Matrix – CM) περιγράφει την εξάρτηση των κύριων συνιστωσών από τα αρχικά χαρακτηριστικά. Οι προκύπτουσες κύριες συνιστώσες PC και τα αρχικά χαρακτηριστικά IF είναι διανύσματα $N \times 1$, ενώ ο πίνακας συντελεστών CM είναι $N \times N$.

$$PC = CM * IF \tag{1.2}$$

Κάθε κύρια συνιστώσα εξηγεί/περιέχει ένα ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής και η συνολική διακύμανση είναι το άθροισμα των επιμέρους συνιστωσών. Έστω V ($N \times 1$ διάνυσμα) είναι το *διάνυσμα ποσοστού της διακύμανσης*, το οποίο περιλαμβάνει τους συντελεστές της διακύμανσης της εν λόγω κύριας συνιστώσας. Το μέρος της διακύμανσης που περιέχεται σε κάθε κύρια συνιστώσα είναι ο λόγος της διακύμανσης αυτής της κύριας

συνιστώσας προς την συνολική διακύμανση.

Εκτιμούμε την «αξία» ενός χαρακτηριστικού μέσω του υπολογισμού του ποσοστού διακύμανσης που εξηγείται από αυτό το χαρακτηριστικό. Αυτό επιτυγχάνεται μέσω του πολλαπλασιασμού (προβολής) της συνεισφοράς του σε κάθε κύρια συνιστώσα (όπως αναφέρεται στην αντίστοιχη στήλη του πίνακα CM) με το διάνυσμα ποσοστού της διακύμανσης. Για το αρχικό χαρακτηριστικό j θεωρούμε την j στήλη του πίνακα CM. Καθώς ο πολλαπλασιασμός $1 \times N * N \times 1$ έχει σαν αποτέλεσμα ένα μονόμετρο μέγεθος αυτό αντιστοιχεί στην εκτιμώμενη αξία του αρχικού χαρακτηριστικού j.

$$V' * (j\ col\ of\ CM) \tag{1.3}$$

Με βάση αυτούς του υπολογισμούς έχουμε την ακόλουθη κατάταξη για τα πρώτα 50 αρχικά χαρακτηριστικά και τα βάρη τους (Πίνακας 1.3). Τα βάρη συγκεντρώνονται στα αρχικά χαρακτηριστικά καθώς παρατηρούμε ότι $\sum_1^{25} w_i = 94\%$ και $\sum_1^{50} w_i = 99\%$.

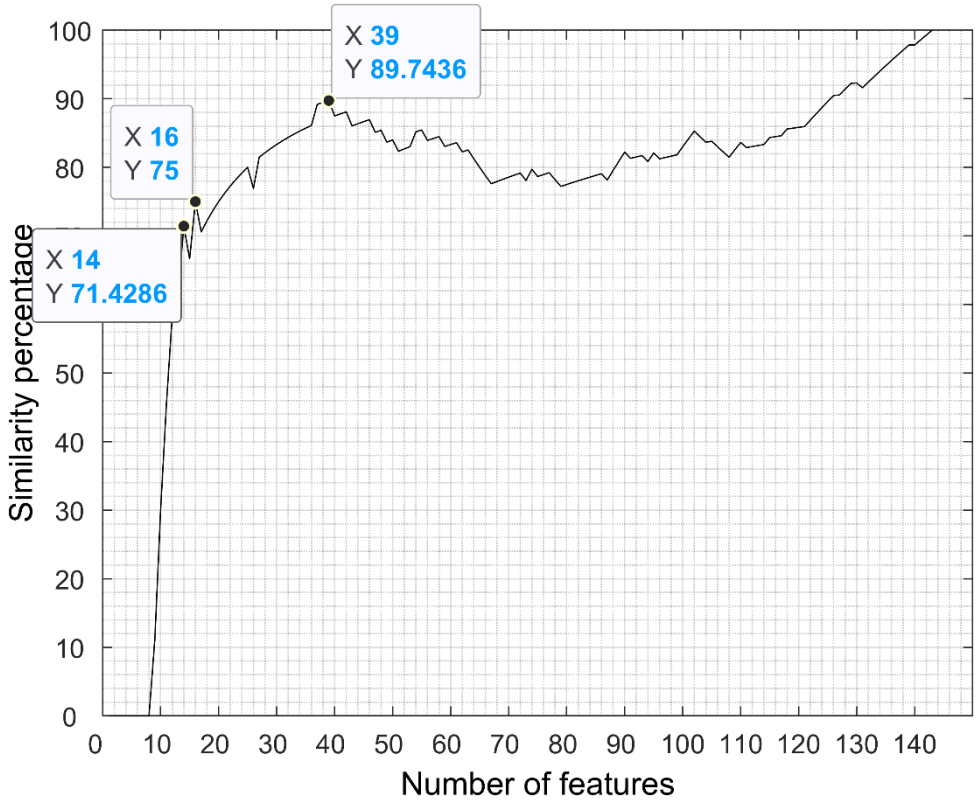
Πίνακας 1.3: Κατάταξη των χαρακτηριστικών με βάση την PCA (Τα 50 πρώτα)

	Χαρακτηριστικό	**Βάρος (%)**		**Χαρακτηριστικό**	**Βάρος (%)**		**Χαρακτηριστικό**	**Βάρος (%)**
1	Flatness	36.90	**18**	MEL 2	0.85	**35**	MEL 1	0.21
2	Time Std	9.42	**19**	Pitch spectral ACF	0.77	**36**	MEL 26	0.20
3	Time RMS	7.58	**20**	Slope	0.72	**37**	MEL 14	0.19
4	Kurtosis	5.10	**21**	Time max ACF	0.67	**38**	MEL 30	0.17
5	Predictivity	4.56	**22**	Flux	0.61	**39**	MEL 21	0.15
6	Skewness	3.70	**23**	Pitch time AMDF	0.57	**40**	MEL 35	0.15
7	Crest	3.28	**24**	MEL 4	0.54	**41**	MEL 8	0.14
8	Peak Program Meter	3.04	**25**	MEL 13	0.49	**42**	MEL 17	0.13
9	Spread	2.87	**26**	Peak Values	0.45	**43**	MEL 32	0.13
10	Time ZCR	2.51	**27**	Tonal power ratio	0.40	**44**	MEL 12	0.12
11	Time ACF coeffients	2.10	**28**	MEL 7	0.35	**45**	MEL 11	0.11
12	Roll-off	1.86	**29**	MEL 6	0.34	**46**	MEL 36	0.11
13	Pitch time ACF	1.36	**30**	MEL 5	0.31	**47**	DMEL 30	0.10
14	Centroid	1.33	**31**	MEL 24	0.29	**48**	MEL 25	0.09
15	Pitch spectral HPS	1.21	**32**	MEL 10	0.27	**49**	MEL 29	0.09
16	MEL 3	1.04	**33**	MEL 19	0.25	**50**	DMEL 28	0.08
17	Decrease	0.92	**34**	MEL 28	0.24			

Κατατάσσοντας λοιπόν τα χαρακτηριστικά σύμφωνα με την προτεινόμενη μέθοδο επιτυγχάνουμε να λάβουμε το 99% της πληροφορίας που αποδίδουν τα 143 συνολικά χαρακτηριστικά λαμβάνοντας υπόψη περίπου το 1/3 αυτών.

1.5.3 Σύγκριση των κατατάξεων

Οι κατατάξεις που προκύπτουν από τις δύο μεθόδους έχουν κοινά χαρακτηριστικά αλλά με διαφορετική σειρά. Ο συντελεστής συσχέτισης Spearman μεταξύ των δύο κατατάξεων υπολογίζεται στο 35%. Η τομή των συνόλων των χαρακτηριστικών απεικονίζεται στο Σχήμα 1.4:



Σχήμα 1.4: Ομοιότητα της κατάταξης με την μέθοδο που βασίζεται στην PCA με την μέθοδο Relief-F

Τα πρώτα 16 χαρακτηριστικά έχουν ποσοστό ομοιότητας 75%, ενώ το υψηλότερο ποσοστό ομοιότητας φτάνει το 89.7% με τα πρώτα 39 χαρακτηριστικά.

Σύμφωνα με τα συσσωρευμένα βάρη είναι εμφανές ότι η κατάταξη των χαρακτηριστικών με την μέθοδο βασισμένη στην PCA υπερτερεί αφού τα 25 και 50 πρώτα χαρακτηριστικά περιέχουν μεγαλύτερο ποσοστό της πληροφορίας σε σχέση με την κατάταξη της μεθόδου Relief-F.

1.6 Αλγόριθμοι Μηχανικής Μάθησης

Η ταξινόμηση πραγματοποιείται με αλγορίθμους ΜΜ εποπτευόμενου τύπου (supervised learning). Σε αυτόν τον τύπο παρέχονται στην είσοδο των αλγορίθμων ΜΜ δεδομένα με γνωστή την ετικέτα τους (κλάση) προκειμένου να χρησιμοποιηθούν ως δεδομένα εκπαίδευσης (training data). Στην προκειμένη έρευνα τα δεδομένα εισόδου είναι τα χαρακτηριστικά του ήχου που προαναφέρθηκαν και οι ετικέτες είναι οι αντίστοιχες κλάσεις. Επιλέχθηκαν πέντε αλγόριθμοι ταξινόμησης και συγκεκριμένα (i) LDA, (ii) QSVM, (iii) kNN, (iv) Boosted Trees και, (v) Random Forest. Οι συγκεκριμένοι ταξινομητές επιλέχθηκαν λόγω της υψηλής ακρίβειας ταξινόμησης. Για την εκτίμηση της αναμενόμενης ακρίβειας ταξινόμησης το σύνολο των δεδομένων διαιρείται σε k υποσύνολα (folds). Τα $k-1$ από αυτά χρησιμοποιούνται για την εκπαίδευση και ένα υποσύνολο για τον έλεγχο. Η διαδικασία αυτή εκτελείται επαναλαμβανόμενα

έτσι ώστε κάθε παρατήρηση (ηχητικό αρχείο) να χρησιμοποιείται μία φορά για έλεγχο και $k-1$ φορές για εκπαίδευση. Η μέση προβλεπόμενη ακρίβεια και η διακύμανση αυτών για διασταυρωμένη επικύρωση (cross-validation) $k = 5, 10$ και 20 φορές απεικονίζεται στον Πίνακα 1.4. Η αναμενόμενη ακρίβεια (μέσος όρος) και η διακύμανση μειώνονται με μεγαλύτερο αριθμό υποσυνόλων, ενώ ταυτόχρονα αυξάνεται ο υπολογιστικός φόρτος. Λαμβάνοντας υπόψη τα αποτελέσματα του Πίνακα 1.4 αλλά και προηγούμενες μελέτες [37] σχετικά με τον τυπικά χρησιμοποιούμενο αριθμό υποσυνόλων, επιλέχθηκε για την πειραματική μελέτη $k = 10$.

Πίνακας 1.4: Εκτιμώμενη ακρίβεια των αλγορίθμων ταξινόμησης (μέση τιμή και διακύμανση)

	k = 5		**k = 10**		**k = 20**	
	Μέση Τιμή	Διακύμανση	Μέση Τιμή	Διακύμανση	Μέση Τιμή	Διακύμανση
LDA	98.87	0.052	98.03	0.042	97.84	0.035
QSVM	95.94	0.127	96.13	0.036	96.38	0.009
kNN	85.06	0.391	86.32	0.289	85.55	0.177
Boosted Trees	91.48	0.244	91.56	0.168	91.99	0.086
Random Forest	86.48	0.533	87.43	0.384	87.76	0.127

Ο ταξινομητής **LDA** εκτιμά την πιθανότητα το δεδομένο ελέγχου (test file) να ανήκει σε μία κλάση χρησιμοποιώντας το θεώρημα Bayes. Η κλάση εξόδου είναι αυτή που έχει την μεγαλύτερη πιθανότητα. Στον ταξινομητή **QSVM** οι τιμές των δεδομένων μετασχηματίζονται σε έναν χώρο μεγαλύτερων διαστάσεων χρησιμοποιώντας την τετραγωνική συνάρτηση kernel για να επιτευχθεί γραμμικός διαχωρισμός. Η εργασία ταξινόμησης πολλαπλών κλάσεων αντιμετωπίζεται με την διαίρεση του προβλήματος σε ένα σύνολο δυαδικών υπο-προβλημάτων. Η τεχνική κωδικοποίησης «ένας-εναντίον-ενός» (one-against-one) μειώνει το πρόβλημα ταξινόμησης πολλαπλών κλάσεων σε πολλαπλά δυαδικά προβλήματα ταξινόμησης. Με αυτόν τον τρόπο ελέγχονται όλα τα πιθανά ζεύγη κλάσεων οδηγώντας σε $n(n-1)/2$ ταξινομητές, όπου n είναι ο αριθμός των κλάσεων (στην προκειμένη περίπτωση $n = 6$). Στον ταξινομητή **kNN** επιλέχθηκε η μετρική απόστασης του συνημίτονου καθώς παρέχει καλύτερα αποτελέσματα από αυτά σε σχέση με τις μετρικές της Ευκλείδειας και της απόστασης Chebychev. Ο αριθμός των γειτόνων έχει οριστεί σε $n = 10$ μετά από σύγκριση των αποτελεσμάτων για $n = 5, 10, 15$ και 20. Για $n = 10$ το κατώφλι της ακρίβειας του 85% επιτεύχθηκε ταχύτερα (δηλαδή με μικρότερο αριθμό χαρακτηριστικών), ενώ παράλληλα επιτεύχθηκε και το μεγαλύτερο ποσοστό ακρίβειας.

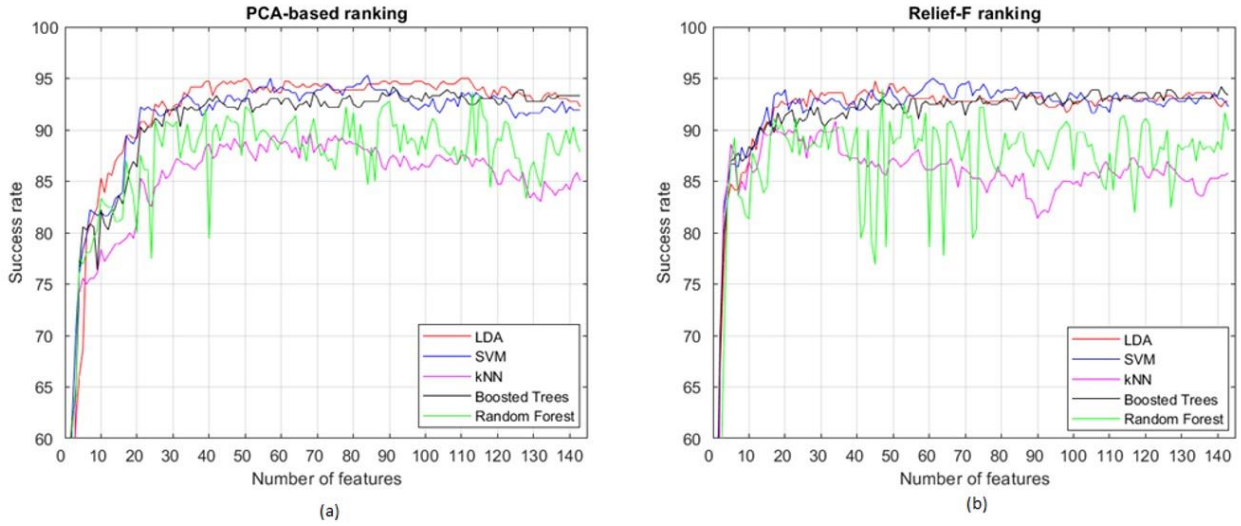
Από τους **Ensemble** ταξινομητές δοκιμάστηκαν οι μέθοδοι Bagging και Boosting. Αυτές οι δύο μέθοδοι διαφέρουν ουσιαστικά ως προς τον χρόνο εκπαίδευσης κάθε «αδύναμου» (weak) εκπαιδευόμενου σε σχέση με τους άλλους. Στην περίπτωση της μεθόδου **Boosted**, οι αδύναμοι εκπαιδευόμενοι εκπαιδεύονται διαδοχικά δίνοντας κάθε φορά μεγαλύτερη βαρύτητα στις προηγούμενες λανθασμένες ταξινομήσεις (adaptive boosting) [38]. Στην παρούσα μελέτη η μέθοδος bagging χρησιμοποιεί τον αλγόριθμο **Random Forest** [39]. Το σύνολο των αρχείων εκπαίδευσης (training data) έχει χρησιμοποιηθεί για την εκπαίδευση κάθε αδύναμου εκπαιδευόμενου (δέντρο) του δάσους. Επιλέχθηκαν οι 100 κύκλοι εκπαίδευσης σύμφωνα με την αξιολόγηση της απόδοσης και του αντίστοιχου υπολογιστικού χρόνου στην σύγκριση μεταξύ των 100, 300 και 500 κύκλων. Η μέθοδος adaptive boosting εκτελείται επανειλημμένα (για 100 κύκλους) και σε κάθε επανάληψη ένας νέος εκπαιδευόμενος, ο οποίος εστιάζει στις λανθασμένες ταξινομήσεις, ενισχύει το μοντέλο ensemble.

Κάθε μοντέλο εκπαιδεύεται 143 φορές χρησιμοποιώντας κάθε φορά αυξανόμενο αριθμό χαρακτηριστικών (από 1 έως 143) σύμφωνα με την κατάταξη των χαρακτηριστικών βασισμένη στην PCA και την κατάταξη Relief-F. Συνολικά, ο αριθμός των μοντέλων που εκπαιδεύονται ισούται με:

$$(\#\ ταξινομητών) * (\#\ χαρακτηριστικών) * (\#\ κατατάξεων) = 1430 \text{ μοντέλα} \quad (1.4)$$

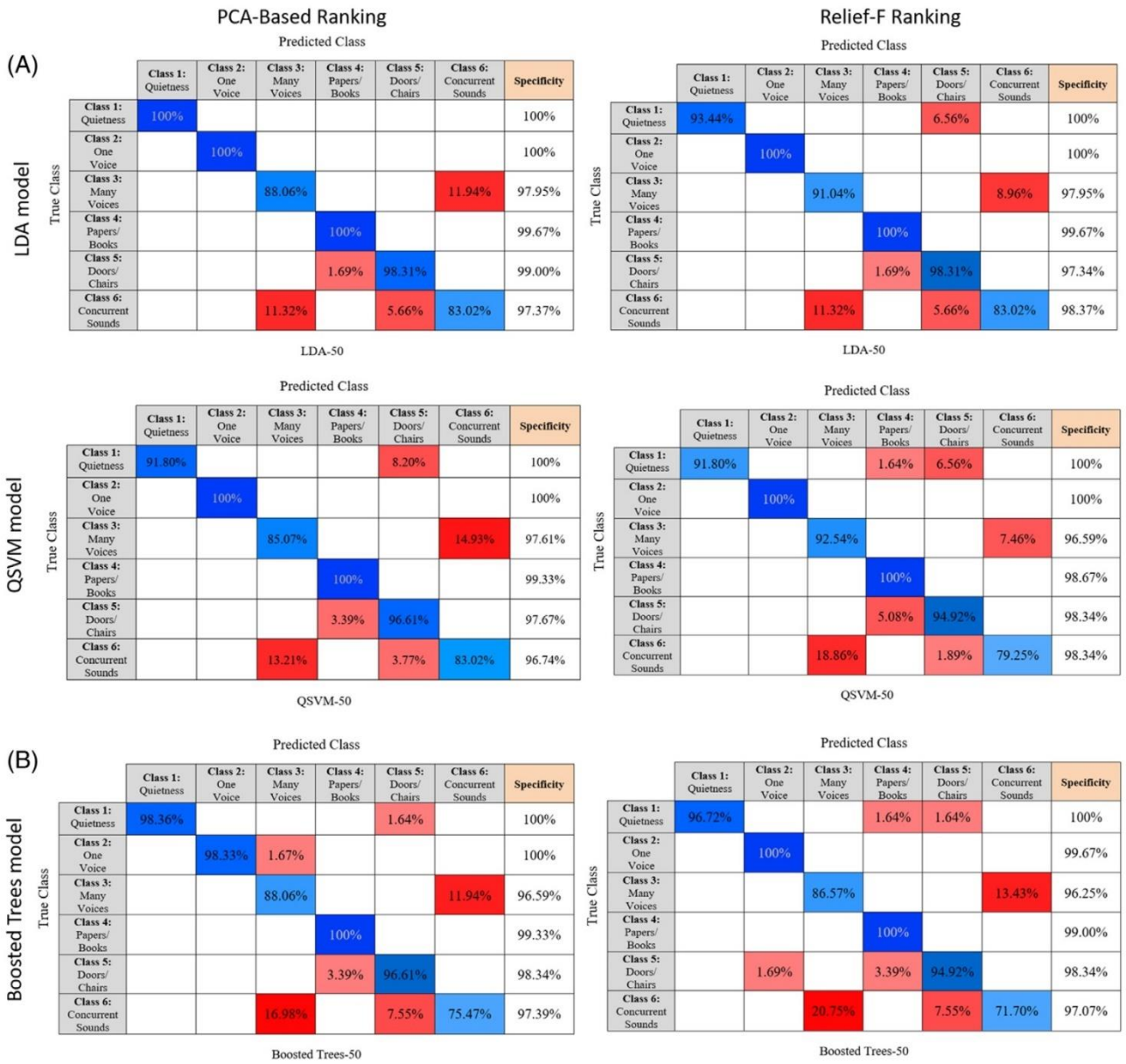
1.6.1 Ακρίβεια ταξινόμησης μοντέλων

Στο Σχήμα 1.5 παρουσιάζεται η ακρίβεια ταξινόμησης (%) των μοντέλων και για τις δύο κατατάξεις με αυξανόμενο αριθμό των χρησιμοποιούμενων χαρακτηριστικών. Καθώς η ακρίβεια αυξάνεται ραγδαία με την αύξηση των χαρακτηριστικών, το σχήμα περιλαμβάνει τιμές από το 60% και μετά ώστε να διευκολύνεται η σύγκριση. Η ακρίβεια ταξινόμησης των τριών από τους πέντε αλγορίθμους (LDA, QSVM και Boosted Trees) ξεπερνά το 90% τόσο με την κατάταξη βασισμένη στην PCA όσο και με την Relief-F. Ορισμένα μοντέλα του αλγορίθμου Random Forest φτάνουν και ξεπερνούν το όριο του 90% με συγκεκριμένο αριθμό χαρακτηριστικών, αλλά η απόδοσή τους παρουσιάζει διακυμάνσεις. Τα μοντέλα του αλγορίθμου ταξινόμησης kNN επιτυγχάνει εν γένει ακρίβεια μικρότερη του 90%, εκτός από ένα μοντέλο που σημειώνει 90.8% χρησιμοποιώντας 34 χαρακτηριστικά με την κατάταξη Relief-F.



Σχήμα 1.5: Η ακρίβεια ταξινόμησης των μοντέλων ταξινόμησης LDA, QSVM, kNN, Boosted Trees και Random Forest με (a) την κατάταξη της μεθόδου που βασίζεται στην PCA και (b) την κατάταξη με την μέθοδο Relief-F.

Εξετάζουμε τα αποτελέσματα που προέκυψαν με την κατάταξη χαρακτηριστικών βασισμένη στην PCA και με την μέθοδο Relief-F όσον αφορά α) την ακρίβεια ταξινόμησης, β) την κλίση της ακρίβειας που σχετίζεται με το πλήθος των απαιτούμενων χαρακτηριστικών ώστε να επιτευχθεί υψηλή ακρίβεια ταξινόμησης και γ) την σταθερότητα των αποτελεσμάτων σε σχέση με το πλήθος των χρησιμοποιούμενων χαρακτηριστικών. Όσον αφορά (α) την ακρίβεια της ταξινόμησης, τα μοντέλα LDA, QSVM και Boosted Trees επιτυγχάνουν και διατηρούν ακρίβεια μεγαλύτερη του 90% υποδεικνύοντας υψηλή ακρίβεια και σταθερότητα (93-95%). Η απόδοση αυτή επαληθεύεται επίσης μέσω των Πινάκων Σύγχυσης (Σχήμα 1.6).



Σχήμα 1.6: Οι Πίνακες Σύγχυσης για τα μοντέλα ταξινόμησης LDA, QSVM και Boosted Trees χρησιμοποιώντας 50 χαρακτηριστικά και με τις δύο κατατάξεις ηχητικών χαρακτηριστικών.

Προκειμένου να υποστηριχθεί η αξιολόγηση των κατατάξεων των χαρακτηριστικών παρουσιάζεται ένα σύνολο μετρικών αξιολόγησης για τους τρεις ταξινομητές με τις υψηλότερες αποδόσεις (LDA, QSVM και Boosted Trees) για 50 χαρακτηριστικά (για αυτόν τον αριθμό χαρακτηριστικών επιτυγχάνεται η καλύτερη απόδοση και για τις δύο κατατάξεις). Οι Αληθώς Θετικές (ΑΘ), οι Αληθώς Αρνητικές (ΑΑ), οι Ψευδώς Θετικές (ΨΘ) και οι Ψευδώς Αρνητικές (ΨΑ) προβλέψεις υπολογίζονται ανά κλάση ήχου για να προσδιοριστούν η Ευαισθησία (Sensitivity) και η Ειδικότητα (Specificity) με βάση τις σχέσεις (5) και (6) και για τις δύο κατατάξεις.

$$Ευαισθησία = \frac{AΘ}{AΘ+ΨA} \qquad (1.5) \qquad\qquad Ειδικότητα = \frac{AA}{AA+ΨΘ} \qquad (1.6)$$

Τα διαγώνια στοιχεία στο Σχήμα 1.6 αντιπροσωπεύουν την Ευαισθησία κάθε κλάσης, ενώ η Ειδικότητα για κάθε κλάση παρουσιάζεται στην τελευταία στήλη.

Συγκρίνεται η αποτελεσματικότητα των κατατάξεων μέσω του λόγου της ακρίβειας προς τον αριθμό των χρησιμοποιούμενων χαρακτηριστικών, λαμβάνοντας υπόψη 25, 50 και 75

χαρακτηριστικά. Τα αποτελέσματα αυτής της σύγκρισης είναι αξιοσημείωτα παραπλήσια για τις δύο κατατάξεις και παρουσιάζονται στον Πίνακα 1.5.

Πίνακας 1.5: Ο λόγος της ακρίβειας ταξινόμησης προς τον αριθμό των χρησιμοποιούμενων χαρακτηριστικών για όλα τα μοντέλα ταξινόμησης και τις κατατάξεις χαρακτηριστικών

	$i = 25$		$i = 50$		$i = 75$	
	PCA-based	Relief-F	PCA-based	Relief-F	PCA-based	Relief-F
LDA	3.70	3.72	1.90	1.89	1.25	1.24
QSVM	3.68	3.67	1.86	1.87	1.25	1.26
kNN	3.38	3.57	1.78	1.74	1.19	1.13
Boosted Trees	3.64	3.63	1.86	1.84	1.23	1.23
Random Forest	3.68	3.50	1.83	1.77	1.15	1.21

Όσον αφορά (β) την κλίση της ανόδου της ακρίβειας ταξινόμησης, τα μοντέλα που χρησιμοποιούν την κατάταξη Relief-F προσεγγίζουν ορισμένα επίπεδα ακρίβειας ταχύτερα (δηλαδή η ακρίβεια ακολουθεί μια πιο απότομη αύξηση) από αυτά που χρησιμοποιούν την κατάταξη βασισμένη στην PCA. Με την μέθοδο Relief-F ξεπερνούν το 85% της ακρίβειας ταξινόμησης με λιγότερα από 10 χαρακτηριστικά, ενώ με την μέθοδο βασισμένη στην PCA χρησιμοποιώντας περισσότερα από 15 χαρακτηριστικά (με εξαίρεση το μοντέλο LDA που φτάνει στο 85% με 12 χαρακτηριστικά).

Σχετικά με (γ) την σταθερότητα της ακρίβειας ταξινόμησης σε σχέση με τον αριθμό των χαρακτηριστικών, υπολογίστηκαν η διακύμανση και η τυπική απόκλιση της ακρίβειας ταξινόμησης για το κάθε μοντέλο. Για να αντιμετωπιστούν οι (τυχαίες) αλλαγές λόγων της αρχικής ανόδου θεωρήθηκαν από $i = 25$ έως 143 χαρακτηριστικά. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 1.6 και επαληθεύεται ότι η ακρίβεια ταξινόμησης είναι πιο σταθερή για τα μοντέλα LDA και QSVM με χαρακτηριστικά που έχουν επιλεγεί με την κατάταξη Relief-F, ενώ για τα μοντέλα Boosted Trees, kNN και Random Forest με βάση την κατάταξη PCA.

Πίνακας 1.6: Η τυπική απόκλιση για κάθε μοντέλο ταξινόμησης με $i \geq 25$

$i = 50$	**PCA-based ranking**	**Relief-F ranking**
LDA	0.74	0.50
QSVM	0.93	0.72
kNN	1.58	1.64
Boosted Trees	0.66	0.75
Random Forest	2.19	3.46

1.6.2 Αξιολόγηση των κατατάξεων

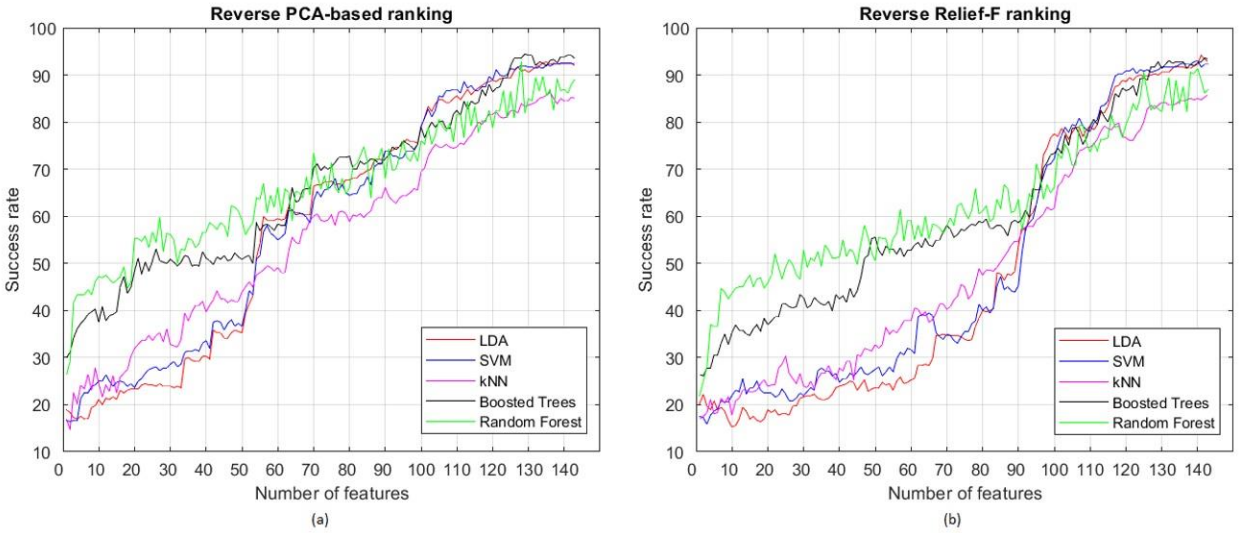
Το μοντέλο LDA επιτυγχάνει τέλεια πρόβλεψη με ποσοστό 100% για την κλάση 1 (ησυχία) με την κατάταξη βασισμένη στην PCA, ενώ με την κατάταξη της μεθόδου Relief-F αναγνωρίζει λανθασμένα κάποια δείγματα ως κλάση 5 (ήχοι από πόρτες/καρέκλες κλπ.) επιτυγχάνοντας ακρίβεια ταξινόμησης 93.44%. Οι κλάσεις 2 (μία φωνή) και 4 (ήχοι από βιβλία και χαρτιά) αναγνωρίζονται τέλεια και με τις δύο κατατάξεις. Η κλάση 3 (πολλές φωνές) συγχέεται με την κλάση 6 (ταυτόχρονοι ήχοι) και το αντίστροφο και με τις δύο κατατάξεις. Ωστόσο, η κλάση 3 αναγνωρίζεται με ελαφρώς μεγαλύτερη ακρίβεια στην περίπτωση της κατάταξης Relief-F. Οι

κλάσεις 5 και 6 αναγνωρίζονται με την ακριβώς ίδια ακρίβεια ταξινόμησης και από τις δύο κατατάξεις και τις ίδιες λανθασμένες προβλέψεις. Οι δύο κατατάξεις διαφέρουν μόνο ως προς την Ειδικότητα (δηλαδή την ικανότητα αναγνώρισης αληθώς ψευδών περιπτώσεων), με την κατάταξη PCA να είναι πιο αποτελεσματική για την αναγνώριση της κλάσης 5 και την Relief-F στην κλάση 6.

Το μοντέλο QSVM ταξινομεί με τον ίδιο ποσοστό την κλάση 1 και για τις δύο περιπτώσεις κατατάξεων: η μεν κατάταξη βασισμένη στην PCA αναγνωρίζει λανθασμένα ορισμένα αρχεία ως κλάση 5, ενώ η Relief-F αναγνωρίζει λανθασμένα και ως κλάση 4. Οι κλάσεις 2 και 4 αναγνωρίζονται απόλυτα και με τις δύο κατατάξεις. Η κατάταξη Relief-F επιτυγχάνει μεγαλύτερη ακρίβεια για την κλάση 3, ενώ PCA για τις κλάσεις 5 και 6.

Το μοντέλο Boosted Trees ταξινομεί τις κλάσεις 1, 2 και 5 με υψηλή ακρίβεια, ενώ η ακρίβεια ταξινόμησης για την κλάση 4 είναι 100% και με τις δύο κατατάξεις. Όπως και στην περίπτωση των μοντέλων LDA και QSVM, η κλάση 3 συγχέεται με την κλάση 6 και αντίστροφα και με τις δύο κατατάξεις. Η κλάση 6 ταξινομείται λανθασμένα με τα χαμηλότερα ποσοστά από όλες τις κλάσεις και με τις δύο κατατάξεις. Με αυτό το μοντέλο η PCA εμφανίζει αυξημένη ακρίβεια σε σχέση με την Relief-F, ενώ η συνολική ακρίβεια του μοντέλου Boosted Trees είναι χαμηλότερη από εκείνη των LDA και QSVM.

Για να επικυρωθεί η αποτελεσματικότητα της προτεινόμενης μεθόδου κατάταξης χαρακτηριστικών συγκρίθηκαν τα αποτελέσματα της ακρίβειας ταξινόμησης με διαφορετικές κατατάξεις χαρακτηριστικών, τυχαίες και επιλεγμένες. Σε αυτές τις περιπτώσεις τα μοντέλα διατηρούν την σειρά τους όσον αφορά την ακρίβεια ταξινόμησης (LDA, QSVM, Boosted Trees ακολουθούμενα από Random Forest και kNN), αλλά η κλίση ανόδου είναι λιγότερο απότομη σε σύγκριση με τις κατατάξεις βασισμένης στην PCA και Relief-F. Το Σχήμα 1.7 απεικονίζει την ακρίβεια ταξινόμησης η οποία επιτυγχάνεται με την αντίστροφη σειρά από αυτή των κατατάξεων PCA και Relief-F, οι οποίες θεωρήθηκαν τα σενάρια αναφοράς. Όπως φαίνεται, ο αριθμός των χρησιμοποιούμενων χαρακτηριστικών επηρεάζει θετικά την ακρίβεια ταξινόμησης αλλά οι μέγιστες αποδόσεις επιτυγχάνονται με τα περισσότερα χαρακτηριστικά. Το ποσοστό ακρίβειας φτάνει το 92.5% χρησιμοποιώντας σχεδόν όλες τις παραμέτρους, δηλαδή η ακρίβεια συγκλίνει όταν έχουν χρησιμοποιηθεί όλα τα χαρακτηριστικά.



Σχήμα 1.7: Η ακρίβεια ταξινόμησης με όλα τα μοντέλα ΜΜ με (a) την αντίστροφη κατάταξη βασισμένη στην PCA και (b) την αντίστροφη Relief-F κατάταξη.

1.7 Συμπεράσματα

Οι στόχοι που τέθηκαν στην αρχή του παρόντος κεφαλαίου ικανοποιήθηκαν στο σύνολό τους.

Προσδιορίστηκε ένα ευρύ σύνολο χαρακτηριστικών που περιγράφουν το ηχητικό σήμα και αναπτύχθηκαν αλγόριθμοι εξαγωγής των τιμών αυτών. Οι αλγόριθμοι εφαρμόστηκαν σε ένα σύνολο ηχητικών δεδομένων το οποίο συλλέχθηκε υπό ρεαλιστικές, ελεγχόμενες συνθήκες κατά την διάρκεια εργαστηριακού μαθήματος, και εμπλουτίστηκε με συναφή ηχητικά αρχεία διαδικτυακά διαθέσιμων βιβλιοθηκών. Τα ηχητικά αρχεία ομογενοποιήθηκαν ως προς τον αριθμό καναλιών και την συχνότητα δειγματοληψίας και κατατμήθηκαν σε πλαίσια ενός δευτερολέπτου. Η περίπτωση που μελετήθηκε ήταν η ταξινόμηση ήχων σε μία αίθουσα διδασκαλίας, κατά την διάρκεια σύγχρονης εκπαίδευσης εκ του σύνεγγυς, με βάση έξι κατηγορίες ήχων: κανένας ήχος, ένας ομιλητής, ταυτόχρονες φωνές, χρήση βιβλίων και εγγράφων, κινήσεις στον χώρο και θόρυβος. Κάθε ηχητικό πλαίσιο ετικετοποιήθηκε χειροκίνητα με μία από τις έξι κλάσεις ήχου.

Προτάθηκε μία νέα μέθοδος κατάταξης χαρακτηριστικών που βασίζεται στην PCA και συγκρίθηκαν τα αποτελεσματα με εκείνα της μεθόδου Relief-F. Η μέθοδος εφαρμόστηκε σε 143 χρονικά, φασματικά και αντιληπτά χαρακτηριστικά ήχου. Τα χαρακτηριστικά αυτά χρησιμοποιήθηκαν από πέντε μοντέλα ταξινόμησης και συγκεκριμένα τα Linear Discriminant Analysis, Quadratic Support Vector Machine, k Nearest Neighbors, Boosted Trees και Random Forest. Αυτά τα μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν με όλα τα πιθανά πλήθη χρησιμοποιούμενων χαρακτηριστικών βασισμένα στις δύο μεθόδους κατάταξης αυτών. Ένα σύνολο 1430 μοντέλων εκπαιδεύτηκε και αξιολογήθηκε.

Όπως φαίνεται από τα Σχήματα 1.5 και 1.6 ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται στην ταξινόμηση των κλάσεων έχει σημαντικό αντίκτυπο στην ακρίβεια ταξινόμησης, με αυτή να αυξάνεται ταχέως με αυξανόμενο αριθμό αυτών ανεξαρτήτως της σειράς με την οποία αυτά λαμβάνονται υπόψη. Με την ιεράρχηση όμως των ηχητικών χαρακτηριστικών (Σχήμα 1.5), επιτυγχάνεται η μέγιστη απόδοση ακρίβειας ταξινόμησης με σκορ 95% χρησιμοποιώντας σχεδόν το 1/3 του συνόλου αυτών (50 ηχητικά χαρακτηριστικά), ενώ με την αντίστροφη σειρά χρήσης των χαρακτηριστικών απαιτείται το σύνολο αυτών για την ίδια απόδοση (Σχήμα 1.6).

Τα αποτελέσματα της προτεινόμενης μεθόδου κατάταξης των χαρακτηριστικών είναι συγκρινόμενα με τα αντίστοιχα της μεθόδου κατάταξης Relief-F. Οι διαφορές μεταξύ των δύο μεθόδων έγκεινται (α) στο ότι η μέθοδος Relief-F παρουσιάζει μεγαλύτερη κλίση ανόδου ξεπερνώντας το 85% της ακρίβειας ταξινόμησης με λιγότερα από 10 χαρακτηριστικά, ενώ η προτεινόμενη μέθοδος με περισσότερο από 15 χαρακτηριστικά (με εξαίρεση το μοντέλο LDA το οποίο φτάνει στο 85% με 12 χαρακτηριστικά), και (β) στο ότι η προτεινόμενη μέθοδος επιτυγχάνει καλύτερη σταθερότητα με μέση τυπική απόκλιση της ακρίβειας ταξινόμησης περίπου 1.2 ενώ η Relief-F παρουσιάζει μέση τυπική απόκλιση περίπου 1.4.

Τα πειραματικά αποτελέσματα των ελέγχων που πραγματοποιήθηκαν απέδειξαν την αξία της χρήσης μεθόδων ιεράρχησης των χαρακτηριστικών και αντίστοιχα της μείωσης της διαστασιολόγησης του προβλήματος ταξινόμησης. Χρησιμοποιώντας υποσύνολο των διαθέσιμων χαρακτηριστικών επιτεύχθηκε η μέγιστη ακρίβεια ταξινόμησης, με ταυτόχρονο μετριασμό της πολυπλοκότητας των ανεπτυγμένων συστημάτων. Το σύνολο αυτής της έρευνας, καθώς και τα αποτελέσματα των πειραματικών ελέγχων περιέχονται στην δημοσίευση [40].

2. Ταξινόμηση του ήχου με μηχανισμούς Βαθιάς Μάθησης

2.1 Εισαγωγή

Η ΒΜ αποτελεί υποσύνολο της ΜΜ. Οι αλγόριθμοι ΒΜ εμπνευσμένοι από την δομή του ανθρώπινου εγκεφάλου έχουν μία πολυεπίπεδη δομή νευρώνων που συνιστούν τα Νευρωνικά Δίκτυα (ΝΔ) τα οποία εκπαιδεύονται μέσω παραδειγμάτων ώστε να παίρνουν αποφάσεις. Τα ΝΔ και συγκεκριμένα αυτά που πραγματοποιούν την πράξη της συνέλιξης (ΣΝΔ) αρχικά επικεντρώθηκαν σε εργασίες ταξινόμησης εικόνων, ανίχνευσης αντικειμένων και λειτουργίες αναγνώρισης. Οι εικόνες χρησιμοποιούνται ως είσοδοι και συγκεκριμένα το ImageNet [41] μία βάση δεδομένων 14 εκατομμυρίων εικόνων χρησιμοποιείται για την εκπαίδευση και την αξιολόγηση ορισμένων από τα πιο γνωστά ΣΝΔ. Το πεδίο εφαρμογής έχει επεκταθεί στον ήχο μετά την μετατροπή των χαρακτηριστικών του ακατέργαστου ηχητικού σήματος σε εικονικές αναπαραστάσεις όπως το φασματογράφημα (spectrogram) και το διάγραμμα κλίμακας-χρόνου (scalogram). Μειονέκτημα των ΝΔ είναι η ανάγκη για εκτεταμένους υπολογιστικούς πόρους καθώς απαιτείται μεγάλος όγκος δεδομένων ειδικά κατά την διάρκεια της εκπαίδευσης των δικτύων. Εξαιτίας αυτού αναπτύχθηκε η μεθοδολογία της μεταφοράς μάθησης (transfer learning) με την οποία τα ΣΝΔ επανεκπαιδεύονται ώστε να μπορούν να λειτουργήσουν σε άλλα σύνολα δεδομένων. Το βασικό όφελος της μεταφοράς μάθησης είναι η εξοικονόμηση πόρων καθώς, ως ένα σημείο, το δίκτυο επαναχρησιμοποιείται.

Τα ΣΝΔ εξελίσσονται διαρκώς ως προς το μέγεθος (αριθμός επιπέδων) και την δομή (τύπος και σύνδεση των επιπέδων) τους. Στην συνέχεια γίνεται αναφορά σε ένα σύνολο γνωστών δικτύων ταξινομημένων σε χρονολογική σειρά, από τα οποία επιλέγεται ένα υποσύνολο για την αξιολόγηση στην ταξινόμηση του ήχου. Η πλειοψηφία, όμως, των δικτύων χρησιμοποιείται σε πειραματική έρευνα σε περαιτέρω εφαρμογές στην συνέχεια της διατριβής.

2.2 Στόχοι

Οι στόχοι της έρευνας συνοψίζονται στους εξής:

- Ο προσδιορισμός των ΣΝΔ που θα χρησιμοποιηθούν για ταξινόμηση του ήχου
- Η μετατροπή του ηχητικού σήματος σε κατάλληλη εικονική αναπαράσταση προκειμένου να μπορεί να αποτελέσει είσοδο στα δίκτυα
- Προσδιορισμός μεγάλων βάσεων, δεδομένης της απαίτησης των ΣΝΔ για μεγάλο όγκο αρχείων εκπαίδευσης
- Διερεύνηση των εσωτερικών υπερπαραμέτρων επανεκπαίδευσης και καθορισμός του βέλτιστου συνδυασμού των τιμών αυτών προκειμένου να επιτευχθεί η μεγιστοποίηση της ακρίβειας ταξινόμησης με ταυτόχρονη ελαχιστοποίηση του υπολογιστικού χρόνου
- Διερεύνηση περαιτέρω βελτιώσεων με χρήση σεναρίων συγχώνευσης (fusion) των μεθόδων

2.3 Συναφής έρευνα

Η ταξινόμηση του ήχου μπορεί να αφορά σε εξωτερικούς ήχους, σε περιβαλλοντικά και αστικά πλαίσια [42], [43], ή σε ήχους εσωτερικού χώρου, οι οποίοι αφορούν πιο εξειδικευμένα περιβάλλοντα, όπως επαγγελματικά, οικιακά και εκπαιδευτικά [44], [45]. Όσον αφορά τους ήχους που παράγονται από τον άνθρωπο, οι εφαρμογές περιλαμβάνουν την αυτόματη αναγνώριση ομιλίας και της ταυτότητας του ομιλητή [46] καθώς και την ανάκτηση MIR [47]. Έχει πραγματοποιηθεί έρευνα για την μείωση του θορύβου χρησιμοποιώντας τεχνικές διαχωρισμού τυφλών πηγών (Blind Source Separation – BSS) [48].

Η αναγνώριση του ήχου έχει ενισχυθεί μέσω της διαθεσιμότητας των συνόλων δεδομένων ήχου. Τέτοια σύνολα αποτελούνται από αποσπάσματα ήχου με ετικέτα, συνήθως ως διακριτά στοιχεία.

Η δημιουργία συνόλου δεδομένων χαρακτηρίζεται από τις προκλήσεις του χαρακτηρισμού του ήχου (ετικετοποίηση), εξαιτίας της πιθανής συνύπαρξης δύο ή και περισσότερων ταυτόχρονων τύπων ήχου. Ακόμη και στην περίπτωση των υπαρχόντων συνόλων δεδομένων ήχου, περισσότεροι του ενός τύποι ήχου μπορεί να συνυπάρχουν σε ένα θεωρητικά ομοιογενές απόσπασμα. Τα UrbanSound8K [49], ESC-10 [50], Air Compressor [51] και TUT [28] είναι δημόσια διαθέσιμα σύνολα δεδομένων, τα οποία χρησιμοποιούνται συχνά για την αξιολόγηση των αλγορίθμων ταξινόμησης. Για κάθε ένα από αυτά τα σύνολα, έχουν οριστεί το ηχητικό πλαίσιο και οι αντίστοιχες κλασεις (ήχοι εσωτερικού ή εξωτερικού χώρου, συγκεκριμένες δραστηριότητες ή κατάσταση μηχανής). Έχει γίνει μία συστηματική προσπάθεια, με το όνομα οντολογία AudioSet, ώστε να υπάρχει μία ιεραρχική συλλογή τύπων ήχου η οποία καλύπτει ένα ευρύ φάσμα καθημερινών ήχων (συνολικά 632 κλάσεις). Τέτοιες προσπάθειες ανοίγουν το αποθεματικό βίντεο και ήχων του YouTube-100M και YouTube-8M [52].

Η έρευνα βασίζεται σε τεχνικές ΜΜ και ΒΜ. Στις τεχνικές ΜΜ εξάγονται τα χαρακτηριστικά του ήχου, τα οποία βασίζονται στις ψυχοακουστικές ιδιότητες των ήχων, όπως π.χ. η ένταση, η χροιά και οι συντελεστές Mel καθώς και οι παράγωγοί τους, τροφοδοτούν τους αλγόριθμους [53]. Στην περίπτωση μηχανισμών ΜΜ χρησιμοποιούνται μεθοδολογίες αξιολόγησης και ιεράρχησης των εξαγώμενων χαρακτηριστικών [54], [40]. Πρόσφατα, χρησιμοποιήθηκαν τεχνικές ΒΜ [55]. Ο παράγοντας διαφοροποίησης μεταξύ της τυπικής χρήσης κλασικών μηχανισμών ΜΜ και τεχνικών ΒΜ, είναι ότι στην πρώτη περίπτωση ένα σύνολο τιμών ηχητικών χαρακτηριστικών τροφοδοτεί τους αλγόριθμους, ενώ στην δεύτερη περίπτωση τα ΣΝΔ είναι υπεύθυνα για τον σχηματισμό της αναπαράστασης του ήχου με πιο αδιαφανή τρόπο.

2.3.1 Συνελικτικά Νευρωνικά Δίκτυα

Το AlexNet ήταν το ΣΝΔ που πέτυχε την καλύτερη απόδοση στο ImageNet Challenge το 2012 [59], ένα έτος ορόσημο μετά το δίκτυο του LeCun [60] για την αναγνώριση χειρόγραφων ψηφίων. Το AlexNet όχι μόνο έχει περισσότερα επίπεδα αλλά χρησιμοποιεί και τις Διορθωμένες Γραμμικές Μονάδες (Rectified Linear Units-ReLUs) αντί για την σιγμοειδή ή την υπερβολική εφαπτομένη ως συναρτήσεις ενεργοποίησης, επιτυγχάνοντας ταχύτερη εκπαίδευση. Το μέγεθος της εικόνας εισόδου είναι 227×227×3, έχει βάθος οκτώ στρωμάτων (πέντε συνελικτικά και τρία πλήρως συνδεδεμένα) και 60 εκατομμύρια παραμέτρους [61].

Το 2014 η οικογένεια των νευρωνικών δικτύων VGG ήταν η εξέλιξη του AlexNet, στα οποία χρησιμοποιήθηκε η λειτουργία ενεργοποίησης ReLU αλλά τα πεδία λήψης αντικαταστάθηκαν από ομάδες μικρότερων (3×3 αντί για 11×11 και 5×5 στο AlexNet) με σταθερό βήμα 1 που δημιουργεί συνελικτικά μπλοκ και οδηγεί σε βελτιωμένη απόδοση [62]. Τα VGG-16 και VGG-19 χρησιμοποιούνται συνήθως και έχουν και τα δύο τρία πλήρως συνδεδεμένα και 13, 16 αντίστοιχα συνελικτικά επίπεδα. Έχουν 138 και 144 εκατομμύρια αντίστοιχα παραμέτρους και οι είσοδοί τους δέχονται εικόνες 224×224×3.

Την ίδια χρονιά, την καλύτερη απόδοση στο ImageNet Challenge πέτυχε το GoogleNet (Inception_v1). Πρόκειται για ένα δίκτυο βάθους 22 επιπέδων (συμπεριλαμβανομένων εννέα μονάδων έναρξης) με επτά εκατομμύρια παραμέτρους. Η καινοτομία αυτού του δικτύου ήταν η παράλληλη εφαρμογή φίλτρων διαφόρων μεγεθών, ενώ οι έξοδοι συνενώθηκαν σε μία ενιαία έξοδο (μονάδα έναρξης). Η μονάδα Inception χρησιμοποιεί επίσης ένα 1×1 συνελικτικό επίπεδο, το οποίο οδηγεί σε μείωση του υπολογιστικού κόστους. Τέλος, η μέση συγκέντρωση πριν από το επίπεδο ταξινόμησης οδηγεί σε σημαντική μείωση του αριθμού των παραμέτρων [63]. Λαμβάνει εικόνες εισόδου 224×224×3. Το 2015, η αρχιτεκτονική Inception τροποποιήθηκε με παραγοντοποίηση των συνελίξεων που οδήγησαν στο Inception_v2 με βάθος 42 επιπέδων και η χρήση βοηθητικών ταξινομητών με κανονικοποίηση παρτίδας οδήγησε στο Inception_v3 [64]. Το Inception_v3 λαμβάνει είσοδο εικόνες 299×299×3, έχει βάθος 48 στρωμάτων και 23.9 εκατομμύρια παραμέτρους. Παράλληλα, τα δίκτυα residual, ResNet18 έχει βάθος 18 επιπέδων

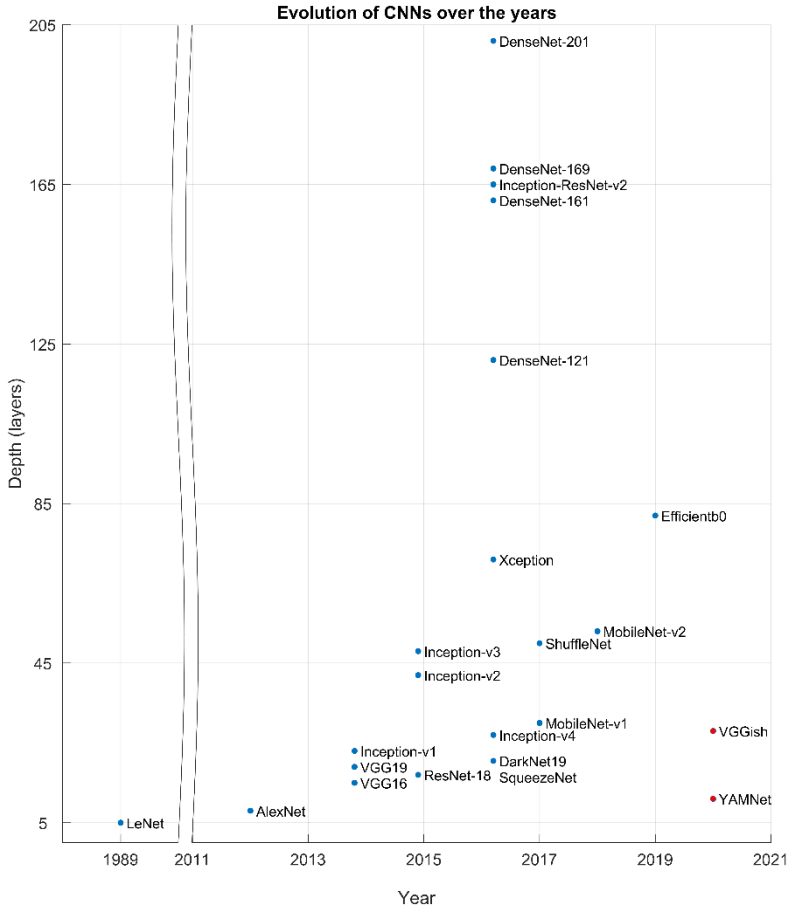
και 11.7 εκατομμύρια παραμέτρους. Τα ResNets αποφεύγουν το αυξανόμενο σφάλμα εκπαίδευσης καθώς το πλήθος των επιπέδων αυξάνεται παρακάμπτοντας τα μη γραμμικά επίπεδα και αντιστοιχώντας τις ταυτότητες με τα residual μπλοκ [65]. Τα ResNets λαμβάνουν ως είσοδο εικόνες 224×224×3.

Το 2016 σχεδιάστηκαν τα SqueezeNet, Xception, Inception_v4, DenseNet και DarkNet19. Το SqueezeNet έχει βάθος 18 επιπέδων (δύο συνελικτικά και 16 μονάδες πυρκαγιάς (fire modules)) και έχει 1.24 εκατομμύρια παραμέτρους. Επιτυγχάνει περίπου την ίδια ακρίβεια ταξινόμησης με το AlexNet, αλλά είναι περίπου 50 φορές μικρότερο ως προς τον αριθμό των παραμέτρων και 44 φορές ως προς το μέγεθος που καταλαμβάνει στον δίσκο. Αυτή η μείωση επιτεύχθηκε με την μονάδα πυρκαγιάς η οποία α) συμπιέζει τη διάσταση του χάρτη των χαρακτηριστικών (δηλαδή τον αριθμό των καναλιών) αντικαθιστώντας τα περισσότερα από τα 3×3 φίλτρα με 1×1 και β) συνενώνει τους χάρτες των χαρακτηριστικών η οποία είναι μία τεχνική παρόμοια με την inception [66]. Λαμβάνει ως είσοδο εικόνες με ανάλυση 227×227×3. Το Xception, το ΣΝΔ Extreme Inception, βασίζεται στο Inception_v3 και αντικατέστησε τις μονάδες έναρξης με διαχωρίσιμα σε βάθος επίπεδα συνέλιξης, παρουσιάζοντας βελτίωση στην ακρίβεια [67]. Το Xception λαμβάνει για είσοδο εικόνες RGB 299×299×3, έχει βάθος 71 επιπέδων και 22.9 εκατομμύρια παραμέτρους. Το Inception_v4 βάθους 27 επιπέδων και το υβριδικό μοντέλο Inception-ResNet-v2 βάθους 164 επιπέδων δοκιμάστηκαν και έδειξαν παρόμοιο υπολογιστικό κόστος και την ίδια απόδοση αναγνώρισης [68]. Το Inception-ResNet-v2 λαμβάνει εικόνες μεγέθους 299×299×3 και έχει 55.9 εκατομμύρια παραμέτρους. Το DenseNet [69] συνδέει όλα τα επίπεδα (με ταιριαστά μεγέθη) μεταξύ τους με αποτέλεσμα μία πυκνή διάταξη συνδεσιμότητας. Έτσι αντιμετωπίζεται η πρόκληση της εξασθένισης της πληροφορίας καθώς περνά μέσα από τα επίπεδα μεγάλων δικτύων. Όλες οι εκδόσεις του DenseNet έχουν περισσότερα από 100 επίπεδα (DenseNet-121, DenseNet-169, DenseNet-201 και DenseNet-161), ενώ ο αριθμός των παραμέτρων κυμαίνεται από 8 έως 28.7 εκατομμύρια με εικόνες εισόδου 224×224×3. Το DarkNet19 είναι παρόμοιο με τα μοντέλα VGG όσον αφορά το μέγεθος του φίλτρου. Έχει βάθος 19 επιπέδων, λαμβάνει εικόνες μεγέθους 256×256×3 και έχει 41.6 εκατομμύρια παραμέτρους [70].

Το 2017, για να περιορίσει την ανάγκη για πόρους, η Google εισήγαγε το MobileNet για κινητές συσκευές και ενσωματωμένες εφαρμογές [71]. Το δίκτυο περιορίζει τον αριθμό των παραμέτρων (όγκος δίσκου) και την πολυπλοκότητα των λειτουργιών (ισχύς και καθυστέρηση). Το MobileNet χρησιμοποιεί διαχωρίσιμες σε βάθος συνέλιξη και μια σημειακή συνέλιξη, με 7 φορές μείωση στον αριθμό των παραμέτρων και μόνο 1% μικρότερη ακρίβεια σε σύγκριση με μοντέλα πλήρους συνέλιξης. Το MobileNet-v1 παίρνει μέγεθος εικόνας εισόδου 224×224×3, έχει βάθος 28 στρωμάτων και έχει 4.2 εκατομμύρια παραμέτρους. Το ShuffleNet έχει μέγεθος εικόνας εισόδου 224×224×3, έχει βάθος 50 στρωμάτων και έχει 1.4 εκατομμύρια παραμέτρους. Αυτή η αρχιτεκτονική χρησιμοποιεί ανακάτεμα καναλιών και σημειακή συνέλιξη, με αποτέλεσμα υψηλότερη ακρίβεια και μικρότερο υπολογιστικό κόστος σε μικρά δίκτυα [72].

Το 2018, το MobileNet-v2 εμφανίστηκε ως η εξέλιξη του προηγούμενου μοντέλου, βελτιώνοντάς το σημαντικά, μειώνοντας την απαιτούμενη μνήμη αλλά διατηρώντας τα επίπεδα ακρίβειας. Αυτό επιτεύχθηκε με μια ανεστραμμένη υπολειμματική δομή και συνδέσεις συντόμευσης μεταξύ γραμμικών σημείων συμφόρησης [73]. Το MobileNet-v2 λαμβάνει εικόνες 224×224×3, έχει βάθος 53 στρωμάτων και έχει 3.5 εκατομμύρια παραμέτρους. Στο ShuffleNet-v2, ο σχεδιασμός της αρχιτεκτονικής βασίζεται στην ταχύτητα αντί για τις λειτουργίες float-point (FLPOs) [74]. Το 2019, η τρίτη γενιά MobileNets, το MobileNet-v3, αξιοποίησε τις συνεργατικές τεχνικές Αναζήτησης Αρχιτεκτονικής Δικτύου (Network Architecture Search - NAS) και τον αλγόριθμο NetAdapt και δημιούργησε δύο νέα μοντέλα (Μεγάλο και Μικρό) ανάλογα με τους πόρους των χρηστών [75]. Τα EfficientNets είναι μια ομάδα μοντέλων όπου η επέκταση των δικτύων γίνεται με τέτοιο τρόπο ώστε όλες οι διαστάσεις να κλιμακώνονται σε σταθερή αναλογία. Το Efficientb0

έχει βάθος 82 στρωμάτων, η είσοδος είναι εικόνα ανάλυσης 224×224×3 και έχει 5.3 εκατομμύρια παραμέτρους [76]. Η χρονολογική εξέλιξη των ΣΝΔ απεικονίζεται στο Σχήμα 2.1.



Σχήμα 2.1: Η εξέλιξη των ΣΝΔ με την πάροδο του χρόνου. Οι μπλε κουκίδες αναφέρονται στα ΣΝΔ Εικόνας και οι κόκκινες κουκίδες στα ΣΝΔ Ήχου.

Το 2020, η απόδοση των ΣΝΔ στην ταξινόμηση εικόνων ενέπνευσε την ιδέα δημιουργίας νέων ή προσαρμογής υπαρχουσών αρχιτεκτονικών για την ταξινόμηση σημάτων ήχου, τα οποία στην συνέχεια θα αποκαλούνται ως ΣΝΔ «Ήχου» (σε αντίθεση με την αρχιτεκτονική των ΣΝΔ που επικεντρώνεται στην εικόνα, και στην συνέχεια θα αναφέρονται ως ΣΝΔ «Εικόνας»). Το VGGish και το YAMNet είναι δύο τέτοια χαρακτηριστικά ΣΝΔ, με το πρώτο να είναι ένα δίκτυο βάθους 24 επιπέδων, βασισμένο στην αρχιτεκτονική VGG, και το δεύτερο (YAMNet) είναι ένα δίκτυο βάθους 28 επιπέδων που χρησιμοποιεί την αρχιτεκτονική MobileNet-v1. Η εκπαίδευση και ο έλεγχος των ΣΝΔ Ήχου χρησιμοποιούν σύνολα δεδομένων ήχου, μετά από κατάλληλη μετατροπή του ηχητικού σήματος σε εικόνα ή σύνολο εικόνων. Συνδυασμός ΣΝΔ και επαναλαμβανόμενων νευρωνικών δικτύων (Recurrent Neural Networks - CRNN) προτάθηκε στο [77] για την ανίχνευση συμβάντων ήχου (Sound Event Detection - SED). Η απόδοση της ταξινόμησης σχετίζεται με την ποσότητα των δεδομένων και την ανισορροπία μεταξύ των δεδομένων των κλάσεων. Αυτό το πρόβλημα αντιμετωπίζεται με ρυθμίσεις ευαίσθητες στο κόστος [78] ή χρησιμοποιώντας την μέθοδο μεταφοράς μάθησης.

2.3.2 Μεταφορά μάθησης

Η μεταφορά μάθησης εφαρμόζει τη γνώση που αποκτήθηκε από έναν τομέα προέλευσης (πηγή) σε έναν άλλο τομέα προορισμού (στόχος) για να αξιοποιήσει το εκπαιδευμένο μοντέλο στον

τομέα προέλευσης ή/και να αντισταθμίσει τα περιορισμένα ή ανεπαρκή δεδομένα στον τομέα προορισμού (στόχος). Το βασικό όφελος αυτής της τεχνικής είναι η εξοικονόμηση πόρων καθώς τα μοντέλα ως έναν βαθμό επαναχρησιμοποιούνται. Οι τεχνικές μεταφοράς μάθησης έχουν χρησιμοποιηθεί σε προβλήματα ταξινόμησης, παλινδρόμησης και ομαδοποίησης [79] σε διαφορετικούς τομείς, και συνήθως περιλαμβάνουν εικόνες, όπως στο [80], με χαρακτηριστικά που προέρχονται από εικόνες του ImageNet τα οποία εξάγονται για χρήση στο σύνολο δεδομένων PASCAL VOC. Για να χρησιμοποιηθεί η μεταφορά μάθησης στην ταξινόμηση ήχου, οι εικονικές αναπαραστάσεις ήχου χρησιμοποιούνται ως κοινός παρονομαστής μεταξύ των εκπαιδευμένων δικτύων και των ακατέργαστων δεδομένων (ήχων). Συγκεκριμένα, παρεμβάλλεται το βήμα προ-επεξεργασίας κατά το οποίο τα ηχητικά αποσπάσματα, μετά από κατάλληλη τμηματοποίηση, μετατρέπονται σε εικόνες, ως φασματογράμματα ή/και διαγράμματα κλίμακας-χρόνου.

Τα ΣΝΔ τυπικά αποτελούνται από (α) το συνελικτικό και το επίπεδο συγκέντρωσης (pooling layer), που είναι υπεύθυνα για την εξαγωγή των χαρακτηριστικών, και (β) τα επίπεδα ταξινόμησης τα οποία συνδέονται με το πρώτο μέρος. Θεωρώντας, λοιπόν, μία τέτοια λειτουργική διάκριση, μια ανοιχτή ερώτηση σχετίζεται με το ποια μέρη των εκπαιδευμένων μοντέλων θα επαναχρησιμοποιηθούν ως έχουν, και ποια θα εκπαιδευτούν εκ νέου σύμφωνα με το σύνολο δεδομένων προορισμού [81]. Υπάρχει ένα ευρύ φάσμα επιλογών, συμπεριλαμβανομένων των εξής:

i. Επανεκπαίδευση του συνολικού δικτύου, διατηρώντας την αρχιτεκτονική και την τοπολογία, για την προσαρμογή των υπαρχόντων ή τον υπολογισμό νέων βαρών και για τα δύο μέρη (του συνελικτικού και του ταξινομητή) των ΣΝΔ.

ii. Εκπαίδευση μόνο ένα μέρους του τμήματος συνέλιξης καθώς και του τμήματος ταξινομητή. Αυτό μπορεί να περιλαμβάνει πιθανές αλλαγές της αρχιτεκτονικής, με συμπερίληψη, αφαίρεση ή αναδιαμόρφωση επιπέδων.

iii. Επανεκπαίδευση του ταξινομητή του αρχικού ΣΝΔ, χωρίς προσαρμογές στο συνελικτικό και στο επίπεδο συγκέντρωσης.

Η επανεκπαίδευση τμημάτων του δικτύου (εκτός από τον ταξινομητή), όπως στις επιλογές (i) και (ii), επιτρέπει ευελιξία και μπορεί να υπόσχεται μεγαλύτερη ακρίβεια ταξινόμησης, ενώ η τρίτη επιλογή διατηρεί τις περισσότερες σταθμίσεις δικτύου και εξοικονομεί υπολογιστικούς πόρους. Ανάλογα με το υπό εξέταση πρόβλημα και την ομοιότητα μεταξύ των προβλημάτων προέλευσης και προορισμού, η τρίτη επιλογή δεν υστερεί από την άποψη της ακρίβειας ταξινόμησης, με το πρόσθετο πλεονέκτημα ότι τυποποιεί το χρησιμοποιούμενο δίκτυο (εκτός από το τμήμα ταξινομητή). Ταυτόχρονα, υπάρχει το μειονέκτημα ότι η είσοδος (δηλαδή οι διαστάσεις της εικόνας) πρέπει να είναι αυτή που χρησιμοποιείται από το αρχικά εκπαιδευμένο δίκτυο.

2.4 Μεθοδολογία

Αρχικά επιλέγονται τα ΣΝΔ τα οποία θα επανεκπαιδευτούν χρησιμοποιώντας την τεχνική της μεταφοράς μάθησης. Δεδομένου ότι η εκπαίδευση των δικτύων απαιτεί μεγάλο όγκο δεδομένων, καθορίζονται τρία δημόσια διαθέσιμα σύνολα δεδομένων ήχου. Στην συνέχεια είναι απαραίτητη η μετατροπή του ηχητικού σήματος σε εικόνα ώστε να εξαχθούν τα κατάλληλα χαρακτηριστικά τα οποία θα τροφοδοτήσουν τα ΣΝΔ. Ακολούθως, πραγματοποιείται διερεύνηση των εσωτερικών παραμέτρων επανεκπαίδευσης των δικτύων, εκτελώντας διαδοχικά συνδυασμούς των τιμών αυτών με κριτήρια την μεγιστοποίηση της ακρίβειας ταξινόμησης με ταυτόχρονη την ελαχιστοποίηση του αντίστοιχου υπολογιστικού χρόνου.

2.4.1 Επιλογή ΣΝΔ και συνόλων δεδομένων

Επιλέχθηκαν τρία ΣΝΔ Εικόνας ως αντιπροσωπευτικά δείγματα της εξέλιξης της αρχιτεκτονικής των ΣΝΔ. Συγκεκριμένα επιλέχθηκαν τα: GoogleNet, SqueezeNet και ShuffleNet, με τον αριθμό των επιπέδων να κυμαίνεται από 18 έως 50. Παράλληλα, χρησιμοποιήθηκαν δύο ΣΝΔ Ήχου, τα VGGish και YAMNet (Πίνακας 2.1).

Πίνακας 2.1: Επιλεγμένα ΣΝΔ για την ταξινόμηση του ήχου με μηχανισμούς ΒΜ

ΣΝΔ	Τύπος	Εκπαίδευση σε	Αριθμός επιπέδων	Παράμετροι (εκατομμύρια)
GoogleNet	Image	ImageNet	22	7
SqueezeNet	Image	ImageNet	18	1.24
ShuffleNet	Image	ImageNet	50	1.4
VGGish	Sound	YouTube	24	72.1
YAMNet	Sound	YouTube	28	3.7

Αναφορικά με τα σενάρια μεταφοράς μάθησης, επιλέχθηκε το (iii) σενάριο διατηρώντας τον προκαθορισμένο αριθμό στρωμάτων και τις τιμές των βαρών. Το τελευταίο επίπεδο του ταξινομητή αντικαταστάθηκε με νέο επίπεδο το οποίο εκπαιδεύτηκε σύμφωνα με τις κλάσεις των νέων συνόλων δεδομένων.

Τα σύνολα των ηχητικών δεδομένων που επιλέχθηκαν για την πειραματική διαδικασία είναι τρία δημόσια διαθέσιμα σύνολα και συγκεκριμένα τα: UrbanSound8K, ESC-10 και Air Compressor. Το σύνολο UrbanSound8K αποτελείται από 8732 αρχεία τύπου wav με 10 κλάσεις τους οποίους συναντάμε σε εξωτερικούς χώρους. Οι κλάσεις και των ποσοστό των αντίστοιχων ηχητικών αρχείων (ανά κλάση) στο σύνολο της βάσης δεδομένων καθώς και η ελάχιστη, η μέση και η μέγιστη διάρκεια (σε δευτερόλεπτα) των αρχείων παρουσιάζονται στον Πίνακα 2.2. Οι κλάσεις έχουν την πρωτότυπη αγγλική ονομασία.

Το σύνολο ESC-10 (υποσύνολο του ESC-50) περιέχει εγγραφές εξωτερικών χώρων και έχει 10 κλάσεις. Κάθε κλάση αποτελείται από 40 αρχεία ήχου ogg διάρκειας πέντε δευτερολέπτων. Το σύνολο Air Compressor περιέχει οκτώ κλάσεις οι οποίες περιγράφουν την κατάσταση μίας μηχανής, με μία «υγιή» κατάσταση και επτά ελαττωματικές καταστάσεις, η κάθε μία εκ των οποίων περιγράφει συγκεκριμένη δυσλειτουργία. Κάθε κλάση αποτελείται από 255 αρχεία ήχου wav διάρκειας τριών δευτερολέπτων. Οι κλάσεις κάθε συνόλου με τις υπόλοιπες πληροφορίες παρουσιάζονται στους Πίνακες 2.3 και 2.4.

Πίνακας 2.2: Το σύνολο ηχητικών δεδομένων UrbanSound8K

Κλάση	Αριθμός αρχείων	Ποσοστό (%)	Μέση διάρκεια (s)	Ελάχιστη διάρκεια (s)	Μέγιστη διάρκεια (s)
AC	1000	11.45	3.99	2.39	4.00
Car horn	429	4.91	2.51	0.06	4.00
Children playing	1000	11.45	3.96	1.05	4.04
Dog bark	1000	11.45	3.15	0.12	4.00
Drilling	1000	11.45	3.55	0.41	4.01
Engine idling	1000	11.45	3.93	0.77	4.00

Gunshot	374	4.28	1.62	0.17	4.00
Jackhammer	1000	11.45	3.59	0.39	4.00
Siren	929	10.64	3.90	0.26	4.00
Street music	1000	11.45	4.00	4.00	4.00

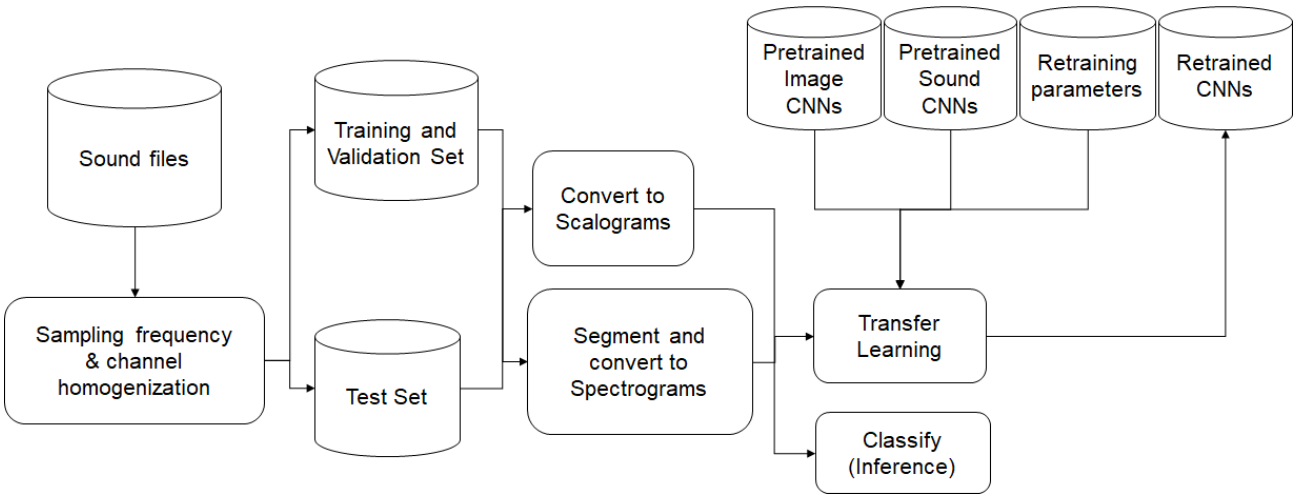
Πίνακας 2.3: Τα σύνολα δεδομένων ήχου ESC-10 και Air Compressor

ESC-10		**Air Compressor**	
Κλάση	**Αριθμός αρχείων**	**Κλάση**	**Αριθμός αρχείων**
Dog bark	40	Bearing	225
Rain	40	Flywheel	225
Sea waves	40	Healthy	225
Baby cry	40	LIV	225
Clock tick	40	LOV	225
Person sneeze	40	NRV	225
Helicopter	40	Piston	225
Chainsaw	40	Riderbelt	225
Rooster	40		
Fire crackling	40		

Πίνακας 2.4: Οι κλάσεις, ο αριθμός των αρχείων και ο τύπος των αρχείων κάθε ηχητικού συνόλου

Σύνολο ήχων	**Κλάσεις**	**Αριθμός αρχείων**	**Τύπος αρχείων**
UrbanSound8K	10	8732	wav
ESC-10	10	400	ogg
Air Compressor	8	1800	wav

Η ροή εργασιών φαίνεται στο Σχήμα 2.2. Συγκεκριμένα: (1) πραγματοποιήθηκε η προεπεξεργασία των ηχητικών αρχείων και η μετατροπή τους σε εικόνες, (2) επανεκπαιδεύτηκαν τα ΣΝΔ με διάφορους συνδυασμούς των τιμών των υπερπαραμέτρων (hyperparameters) τους χρησιμοποιώντας τα προαναφερθέντα σύνολα ήχου και (3) αξιολογήθηκαν οι αντίστοιχοι συνδυασμοί υπερπαραμέτρων και ΣΝΔ.



Σχήμα 2.2: Ροή εργασιών για την ταξινόμηση του ήχου με μηχανισμούς ΒΜ.

2.4.2 Προεπεξεργασία ήχου

Οι ήχοι, ακόμη και αυτοί που ανήκουν στο ίδιο σύνολο δεδομένων, μπορεί να έχουν διαφορετικά γνωρίσματα που σχετίζονται με το πρωτογενές ακατέργαστο σήμα. Αυτά περιλαμβάνουν τη συχνότητα δειγματοληψίας, τον αριθμό των καναλιών και τη διάρκεια των αποσπασμάτων. Οι διαφορές αυτές μπορεί να επηρεάσουν την επακόλουθη επεξεργασία. Για παράδειγμα, ένα στερεοφωνικό σήμα μπορεί να παρέχει διπλάσιο αριθμό αναπαραστάσεων χρόνου-συχνότητας. Ένα πρώτο βήμα λοιπόν είναι η ομογενοποίηση των ήχων ως προς τη συχνότητα δειγματοληψίας και τον αριθμό των καναλιών. Συγκεκριμένα, η ίδια συχνότητα δειγματοληψίας, που τυπικά ανήκει στο χαμηλότερο εύρος τιμών (16 kHz), εφαρμόζεται σε όλα τα ηχητικά αποσπάσματα. Τα στερεοφωνικά (και πολυκαναλικά) σήματα ήχου μετατρέπονται σε μονοφωνικά. Το σύνολο δεδομένων χωρίζεται σε σύνολα εκπαίδευσης (training), επικύρωσης (validation) και ελέγχου (test), που αντιστοιχούν στο 60%, στο 20% και στο 20% των αρχείων, αντίστοιχα.

Εν συνεχεία, το πρωτογενές σήμα μετατρέπεται σε εικόνα. Το ηχητικό σήμα χωρίζεται σε επικαλυπτόμενα, περιοδικά παράθυρα Hanning μήκους 25 ms, με μήκος αναπήδησης (hop) 10 ms. Υπολογίζεται ο σύντομος μετασχηματισμός Fourier και το πλάτος των φασματικών τιμών διέρχεται μέσω συστοιχιών φίλτρων συχνότητας Mel 64 ζωνών, που εκτείνονται στο εύρος των 125-7500 Hz. Τα διαγράμματα κλίμακας-χρόνου παράγονται ως η απόλυτη τιμή των συντελεστών του Συνεχή Μετασχηματισμού της Περιβάλλουσας (Continuous Wavelet Transform - CWT) του ηχητικού σήματος.

Όσον αφορά τη μετατροπή σε εικόνες, υπάρχουν δύο επιλογές. (1) Ολόκληρο το ηχητικό απόσπασμα μετατρέπεται σε μια ενιαία εικόνα και τροφοδοτείται στον αλγόριθμο του ΣΝΔ για ταξινόμηση και συνάγεται μια μεμονωμένη απόφαση. (2) Το ηχητικό σήμα χωρίζεται σε τμήματα (των 0.96 ms) και κάθε ένα από αυτά τα αποσπάσματα μετατρέπεται σε εικόνα. Κάθε τμήμα αντιστοιχεί σε ένα από τα 96 μέρη του ηχητικού αποσπάσματος διάρκειας 10 ms, με αποτέλεσμα διάρκεια $96 * 10 = 960ms = 0.96s$. Λαμβάνοντας υπόψη το ποσοστό επικάλυψης (στο 50%), το μήκος αναπήδησης υπολογίζεται σε: $hop = 0.96/2 = 0.48\ ms$. Ο ήχος περιγράφεται από μία $96 \times 64 \times 1 \times k$ συστοιχία, όπου k είναι ο αριθμός των φασματογραμμάτων ο οποίος εξαρτάται από το μήκος του ήχου και το ποσοστό επικάλυψης μεταξύ τους. Ο αριθμός των φασματογραμμάτων (k) που εξετάζεται σε ένα ηχητικό απόσπασμα μπορεί να προσεγγιστεί μέσω του ακόλουθου τύπου:

$$(mod)\left(\frac{FL}{SL*FL} - \left(\frac{1}{OL} - 1\right)\right) \quad (2.1)$$

όπου, mod είναι το υπόλοιπο της διαίρεσης, FL είναι το πλήρες μήκος (Full Length) του ηχητικού

αποσπάσματος, SL είναι το μήκος του τμήματος (Segment Length), και OL ο λόγος επικάλυψης (Overlapping ratio), για τον οποίο έχει υποτεθεί θετική τιμή.

Για τα τρία ΣΝΔ Εικόνας, GoogleNet, SqueezeNet και ShuffleNet, τα διαγράμματα κλίμακας-χρόνου υποβλήθηκαν στην τεχνική ενίσχυσης (augmentation) δεδομένων για δύο λόγους: (1) για να ληφθούν οι απαιτούμενες διαστάσεις εικόνας που απαιτεί για είσοδο κάθε δίκτυο, και (2) για να αποφευχθεί η υπερπροσαρμογή. Η ενίσχυση αυτή πραγματοποιήθηκε μέσω μετασχηματισμών εικόνας, όπως π.χ. μετατόπιση, μεγέθυνση και περιστροφή.

Η προκύπτουσα αναπαράσταση βάσει εικόνας χρησιμοποιείται για την επανεκπαίδευση των προεκπαιδευμένων ΣΝΔ εικόνας και ήχου με τα υποσύνολα εκπαίδευσης και επικύρωσης. Κάθε εικόνα ελέγχου τροφοδοτείται στη συνέχεια στον αλγόριθμο ΣΝΔ και λαμβάνεται μια απόφαση ταξινόμησης για κάθε εικόνα (στην περίπτωση ενός μεμονωμένου, ανά ηχητικό απόσπασμα) ή για κάθε τμήμα (στην περίπτωση τμηματοποίησης). Για την τελευταία περίπτωση, η απόφαση για το απόσπασμα βασίζεται στη συγχώνευση των αποφάσεων των επιμέρους τμημάτων.

2.4.3 Υπερπαράμετροι επανεκπαίδευσης

Οι υπερπαράμετροι εκπαίδευσης σχετίζονται με τον υπολογισμό των βαρών για την ελαχιστοποίηση της συνάρτησης απωλειών (loss function) με τη λήψη διορθωτικών βημάτων με τη χρήση οπισθοδιάδοσης (back-propagation). Το σύνολο εκπαίδευσης διαιρείται με το μέγεθος του mini-batch. Αυτό το πηλίκο είναι ο αριθμός των επαναλήψεων που επεξεργάζεται το μοντέλο για τον υπολογισμό του σφάλματος πρόβλεψης και την ανάλογη ενημέρωση των βαρών. Το σύνολο επικύρωσης χρησιμοποιείται κατά τη διάρκεια της εκπαίδευσης για τον έλεγχο των ενδιάμεσων τιμών και την πραγματοποίηση των αντίστοιχων διορθωτικών βημάτων (learning rate) για την επιλογή των κατάλληλων βαρών. Μια εποχή (epoch) είναι ένα πλήρες πέρασμα από ολόκληρο το σύνολο εκπαίδευσης. Υπομονή επικύρωσης (validation patience) είναι ο αριθμός των επαναλήψεων που επιτρέπονται χωρίς αύξηση της ακρίβειας επικύρωσης (validation accuracy).

Η Στοχαστική Κλίση Καθόδου με Ορμή (Stochastic Gradient Descent with Momentum - SGDM) επιτρέπει τη συμβολή του προηγούμενου διορθωτικού βήματος στη βελτίωση της σύγκλισης [82], με τα δεδομένα εκπαίδευσης να ανακατασκευάζονται σε κάθε επανάληψη. Καθώς ο βελτιστοποιητής (optimizer) SGDM κλιμακώνει ομοιόμορφα τη διαβάθμιση σφάλματος, στην περίπτωση μη ομοιόμορφων συνόλων δεδομένων μπορεί να μειώσει την απόδοση. Οι προσαρμοστικοί αλγόριθμοι μπορούν να βελτιώσουν τα διορθωτικά βήματα. Ο αλγόριθμος της Προσαρμοστικής Εκτίμησης Ροπής (Adaptive Moment Estimation - Adam) [83] συνδυάζει τα οφέλη της μεθόδου της Μέσης Τετραγωνικής Ρίζας Διάδοσης (Root Mean Square Propagation - RMSProp) και του Αλγορίθμου Προσαρμοστικής Κλίσης (Adaptive Gradient Algorithm - AdaGrad). Στην [84], το SGDM και ο Adam συγκρίνονται ως προς την απόδοση, με το πρώτο να συγκλίνει πιο αργά αλλά να έχει καλύτερη απόδοση. Έχει γίνει έρευνα [85], [86] για το μέγεθος παρτίδας, τον ρυθμό εκμάθησης και τις επαναλήψεις που απαιτούνται για την αντιμετώπιση του επονομαζόμενου «χάσματος γενίκευσης» (generalization gap). Στη συνέχεια, εξετάστηκαν δύο σετ υπερπαραμέτρων για την επανεκπαίδευση των ΣΝΔ Εικόνας και Ήχου για να διερευνήσουμε την επιρροή τους στην ακρίβεια ταξινόμησης.

Στον Πίνακα 2.5 παρουσιάζονται οι παραλλαγές στις ρυθμίσεις των υπερπαραμέτρων που εξετάστηκαν όσον αφορά στα ΣΝΔ Εικόνας. Πρόκειται για δύο τύπους optimizers, τους SGDM και Adam, τρεις τιμές για το μέγεθος του mini-batch, της εποχής και του ρυθμού εκμάθησης εκάστης παραμέτρου. Στην συνέχεια αυτοί οι όροι θα αναφέρονται με την αγγλική τους ορολογία για να συνάδουν με τις εικόνες που ακολουθούν.

Πίνακας 2.5: Σετ τιμών των υπερπαραμέτρων που εξετάστηκαν για τα ΣΝΔ Εικόνας

Optimizer	**Mini-Batch Size**	**Epochs**	**Learning rate (×10^{-4})**
SGDM, Adam	8, 16, 32	6, 8, 10	0.5, 1, 2

Συνολικά εξετάστηκαν 54 συνδυασμοί ρυθμίσεων:

$$(\#Optimizers) * (\#Mini - Batch\ sizes) * (\#Epochs) * (\#Learning\ Rates) = 2 * 3 * 3 * 3 = 54 \quad (2.2)$$

Όλοι οι συνδυασμοί χρησιμοποιήθηκαν για την επανεκπαίδευση και την αξιολόγηση των ΣΝΔ της Εικόνας όσον αφορά την ακρίβεια ταξινόμησης και τους χρόνους εκπαίδευσης για τα τρία σύνολα δεδομένων, με αποτέλεσμα συνολικά $54 * 3 = 162$ δοκιμές.

Στον Πίνακα 2.6 παρουσιάζονται οι τιμές των υπερπαραμέτρων που επιλέχθηκαν για την επανεκπαίδευση των ΣΝΔ Ήχου. Καθώς τα ΣΝΔ Ήχου είναι πιο εκτεταμένα (με όρους αριθμού επιπέδων) από αυτά της Εικόνας, και για να αποφευχθούν σημαντικές ασυμμετρίες όσον αφορά τον απαιτούμενο χρόνο εκπαίδευσης, το μέγεθος του mini-batch ορίστηκε σε 64, 128 και 256. Ο Adam χρησιμοποιήθηκε ως optimizer καθώς ήταν σταθερά ο καλύτερος σε σύγκριση με τον SGDM. Κατά τη διάρκεια των πειραμάτων και της παρακολούθησης της προόδου της εκπαίδευσης, η ακρίβεια της επικύρωσης σταθεροποιήθηκε σε μια μέγιστη τιμή μετά από μερικές εποχές (λιγότερες από 10). Επομένως, ορίστηκε μέγιστος αριθμός εποχών (ίσος με 10), και validation patience (δηλαδή, όταν η ακρίβεια σταματά να αυξάνεται) ορίστηκε σε 2 εποχές, μειώνοντας σημαντικά τον αριθμό των συνδυασμών.

Πίνακας 2.6: Σετ τιμών των υπερπαραμέτρων που εξετάστηκαν για τα ΣΝΔ Ήχου.

Optimizer	**Mini-Batch Size**	**Maximum Epochs**	**Learning rate (×10^{-4})**
Adam	64, 128, 256	10	0.5, 1, 2

Συνολικά εξετάστηκαν εννέα συνδυασμοί ρυθμίσεων:

$$(\#Optimizers) * (\#Mini - Batch\ sizes) * (\#Epochs) * (\#Learning\ Rates) = 1 * 3 * 1 * 3 = 9 \quad (2.3)$$

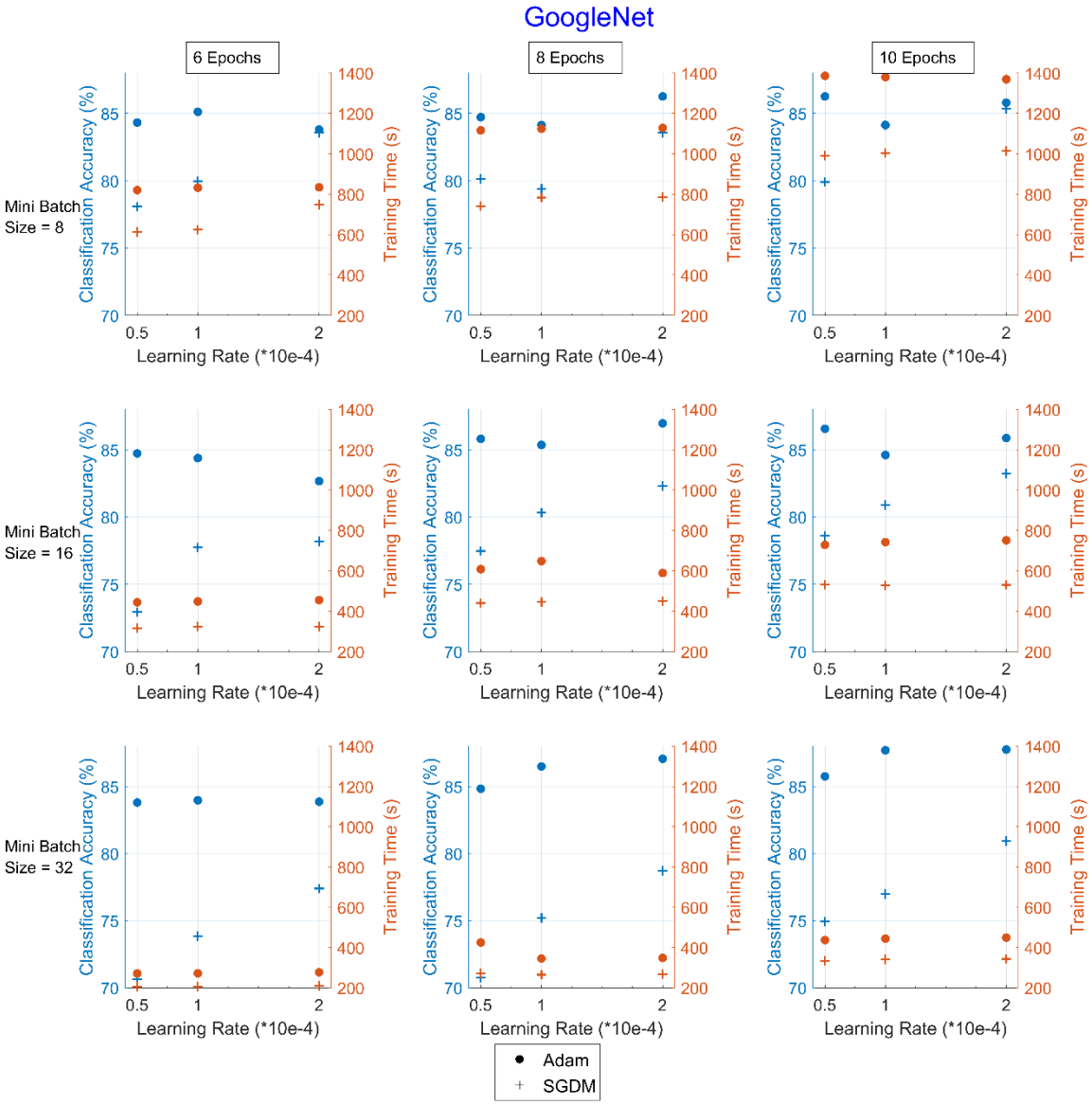
Ομοίως, αυτοί οι συνδυασμοί χρησιμοποιήθηκαν για την επανεκπαίδευση και αξιολόγηση των δύο ΣΝΔ Ήχου (VGGish και YAMNet) και για τρία σύνολα ήχου με αποτέλεσμα 27 δοκιμές.

Τα πειράματα έγιναν σε υπολογιστή με 32 GB RAM, με Intel Core Επεξεργαστής i7-10700K, οκτώ πυρήνων έως 3.8 GHz, με την κάρτα γραφικών NVIDIA GeForce RTX 3060, ενώ τα μοντέλα υλοποιήθηκαν στο Matlab R2021a.

2.5 Αποτελέσματα

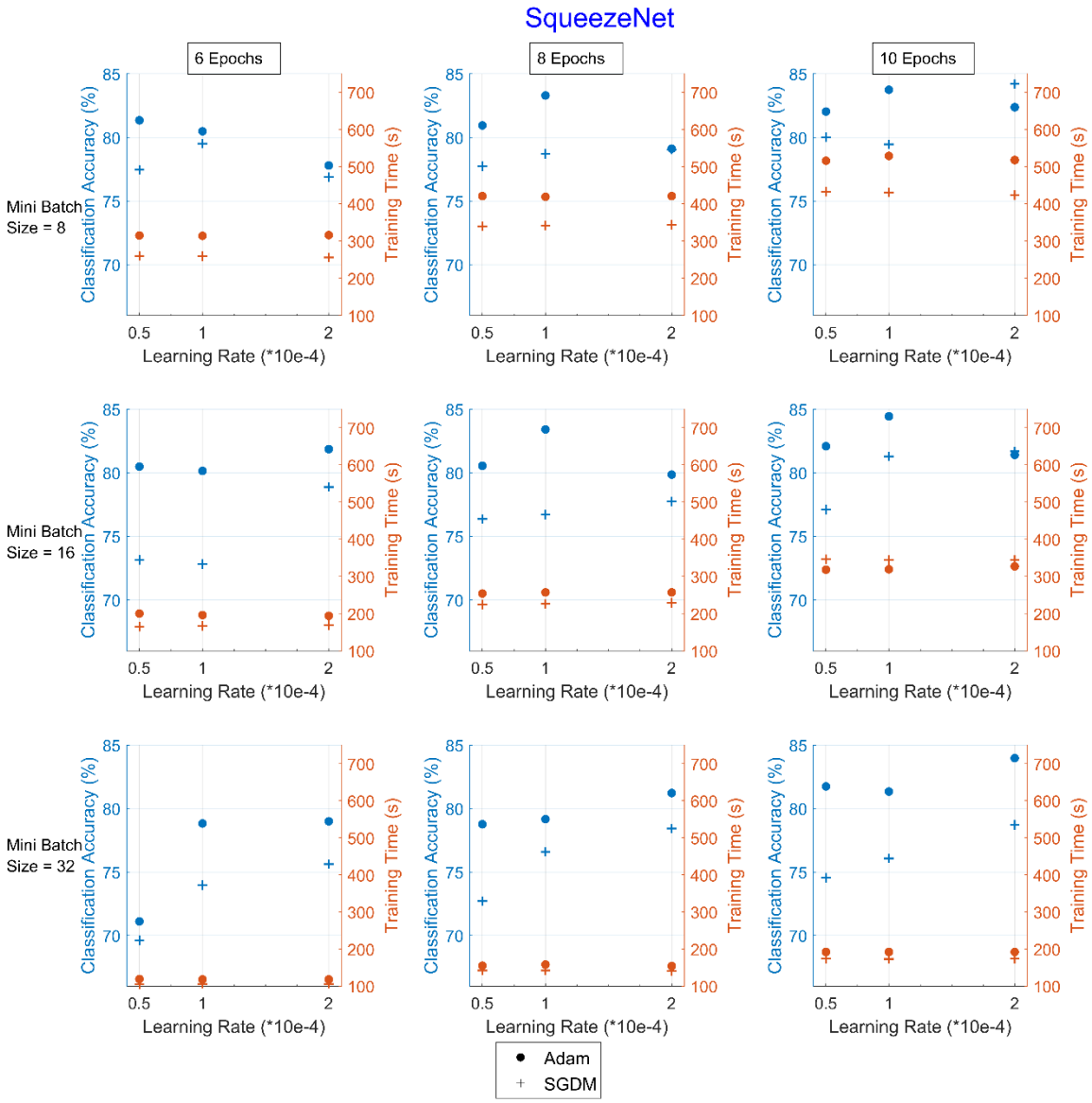
2.5.1 Απόδοση των ΣΝΔ Εικόνας

Τα τρία ΣΝΔ Εικόνας, GoogleNet, SqueezeNet και ShuffleNet, εκπαιδεύτηκαν χρησιμοποιώντας 54 συνδυασμούς τιμών υπερπαραμέτρων για τα τρία σύνολα ήχου. Η αξιολόγηση βασίστηκε στην ακρίβεια ταξινόμησης (classification accuracy) και τον χρόνο εκπαίδευσης (training time). Τα τρία σύνολα ήχου ακολούθησαν παρόμοια πορεία, για αυτόν τον λόγο στα Σχήματα 2.3-2.5 παρουσιάζονται τα αποτελέσματα μόνο για το UrbanSound8K.

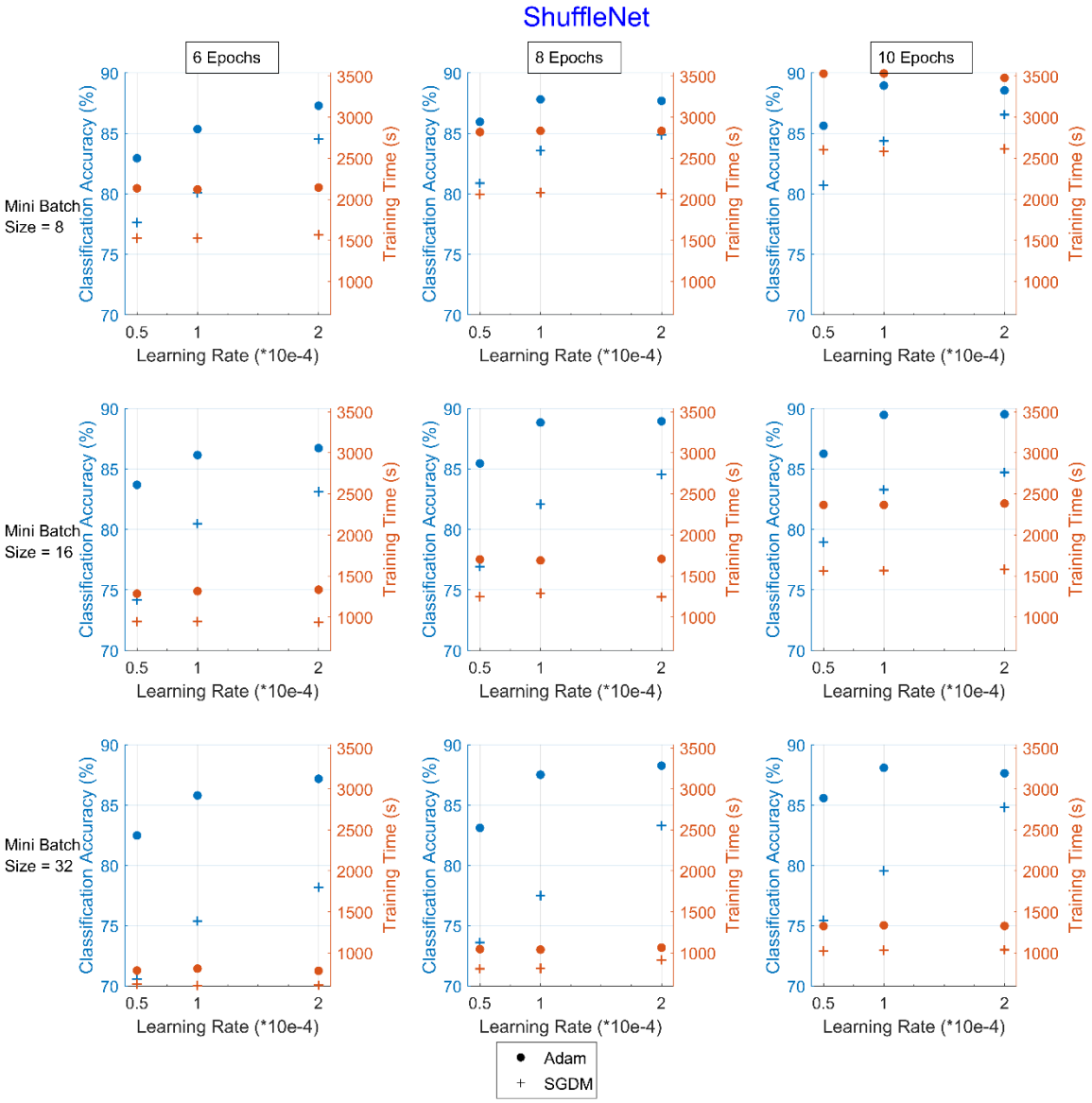


Σχήμα 2.3:Η ακρίβεια ταξινόμησης και ο χρόνος εκπαίδευσης για τους συνδυασμούς των υπερπαραμέτρων που εφαρμόζονται στο GoogleNet για το σύνολο δεδομένων UrbanSound8K. Η κουκίδα αντιστοιχεί στον optimizer Adam, ενώ ο σταυρός στον SGDM. Το μπλέ χρώμα αναφέρεται στην ακρίβεια ταξινόμησης (%) και το κόκκινο στον χρόνο εκπαίδευσης (s).

Ο Adam optimizer είχε καλύτερη απόδοση όσον αφορά την ακρίβεια ταξινόμησης (μπλε χρώμα), ενώ το SGDM είχε καλύτερες επιδόσεις όσον αφορά τον χρόνο εκπαίδευσης (κόκκινο χρώμα). Επιπλέον, καθώς αυξανόταν το μέγεθος της μίνι παρτίδας, ο χρόνος εκπαίδευσης μειωνόταν, ενώ, όπως ήταν αναμενόμενο ο χρόνος εκπαίδευσης αυξανόταν με αύξηση του αριθμού των εποχών. Ο ρυθμός εκμάθησης, ο οποίος επηρεάζει το πόσο γρήγορα μαθαίνει ένα δίκτυο, έχει άμεσο αντίκτυπο και στην ποιότητα της εκμάθησης, γεγονός που καθιστά αυτήν την υπερπαράμετρο την πιο ευαίσθητη στη ρύθμιση.



Σχήμα 2.4: Η ακρίβεια ταξινόμησης και ο χρόνος εκπαίδευσης για τους συνδυασμούς των υπερπαραμέτρων που εφαρμόζονται στο SqueezeNet για το σύνολο δεδομένων UrbanSound8K. Η κουκίδα αντιστοιχεί στον optimizer Adam, ενώ ο σταυρός στον SGDM. Το μπλέ χρώμα αναφέρεται στην ακρίβεια ταξινόμησης (%) και το κόκκινο στον χρόνο εκπαίδευσης (s).



Σχήμα 2.5: Η ακρίβεια ταξινόμησης και ο χρόνος εκπαίδευσης για τους συνδυασμούς των υπερπαραμέτρων που εφαρμόζονται στο ShuffleNet για το σύνολο δεδομένων UrbanSound8K. Η κουκίδα αντιστοιχεί στον optimizer Adam, ενώ ο σταυρός στον SGDM. Το μπλέ χρώμα αναφέρεται στην ακρίβεια ταξινόμησης (%) και το κόκκινο στον χρόνο εκπαίδευσης (s).

Σε όλα τα ΣΝΔ Εικόνας, ο Adam (μπλε κουκκίδα) πέτυχε καλύτερη απόδοση από τον SGDM (κόκκινη κουκκίδα). Στον Πίνακα 2.7 τα δίκτυα συγκρίνονται ως προς τις μέσες τιμές της ακρίβειας ταξινόμησης και του χρόνου εκπαίδευσης για κάθε optimizer για το σύνολο δεδομένων UrbanSound8K. Το SqueezeNet ήταν το δίκτυο με τον μικρότερο χρόνο εκπαίδευσης αλλά και με την χαμηλότερη ακρίβεια ταξινόμησης. Το ShuffleNet παρουσίασε τον υψηλότερο χρόνο εκπαίδευσης, περίπου 2.7 φορές μεγαλύτερο από το αντίστοιχο του GoogleNet.

Πίνακας 2.7: Μέσες τιμές της ακρίβειας ταξινόμησης και του χρόνου εκπαίδευσης για κάθε ΣΝΔ Εικόνας με τους δύο otpimizers για το σετ ήχου UrbanSound8K

CNN	Adam		SGDM	
	Ακρίβεια ταξινόμησης (%)	**Χρόνος εκπαίδευσης (s)**	**Ακρίβεια ταξινόμησης (%)**	**Χρόνος εκπαίδευσης (s)**
GoogleNet	85.28	691	78.60	505
SqueezeNet	80.98	277	77.22	268
ShuffleNet	86.70	1892	80.37	1381

Τα αντίστοιχα αποτελέσματα για τα σύνολα ήχου ESC-10 και Air Compressor παρουσιάζονται στους Πίνακες 2.8 και 2.9.

Πίνακας 2.8: Μέσες τιμές της ακρίβειας ταξινόμησης και του χρόνου εκπαίδευσης για κάθε ΣΝΔ Εικόνας με τους δύο optimizers για το σετ ήχου ESC-10

CNN	Adam		SGDM	
	Ακρίβεια ταξινόμησης (%)	**Χρόνος εκπαίδευσης (s)**	**Ακρίβεια ταξινόμησης (%)**	**Χρόνος εκπαίδευσης (s)**
GoogleNet	84.10	33	78.75	24
SqueezeNet	79.49	15	70.28	13
ShuffleNet	82.22	112	73.29	87

Πίνακας 2.9: Μέσες τιμές της ακρίβειας ταξινόμησης και του χρόνου εκπαίδευσης για κάθε ΣΝΔ Εικόνας με τους δύο optimizers για το σετ ήχου Air Compressor

CNN	Adam		SGDM	
	Ακρίβεια ταξινόμησης (%)	**Χρόνος εκπαίδευσης (s)**	**Ακρίβεια ταξινόμησης (%)**	**Χρόνος εκπαίδευσης (s)**
GoogleNet	95.92	202	86.61	101
SqueezeNet	94.00	62	87.49	48
ShuffleNet	96.01	368	89.15	279

2.5.2 Απόδοση των ΣΝΔ Ήχου

Εφαρμόζοντας τους εννέα συνδυασμούς υπερπαραμέτρων στα ΣΝΔ Ήχου, τόσο το VGGish όσο και το YAMNet ξεπέρασαν το 95% στην ακρίβεια ταξινόμησης για το σύνολο δεδομένων UrbanSound8K και έφτασαν το 100% για το σύνολο δεδομένων Air Compressor. Το VGGish έδειξε οριακά χαμηλότερη απόδοση από τα ΣΝΔ Εικόνας για το σύνολο δεδομένων ESC-10, ενώ το YAMNet πέτυχε καλύτερη απόδοση από όλα τα ΣΝΔ Εικόνας. Όσον αφορά τον χρόνο εκπαίδευσης, και τα δύο δίκτυα Ήχου παρουσίασαν σημαντική μείωση, κατά 80% κατά μέσο όρο. Οι μέσες τιμές της ακρίβειας ταξινόμησης και του χρόνου εκπαίδευσης για όλους τους

συνδυασμούς για το VGGish και το YAMNet για τα τρία σύνολα δεδομένων εμφανίζονται στον Πίνακα2.10.

Πίνακας 2.10: Μέσες τιμές της ακρίβειας ταξινόμησης και του χρόνου εκπαίδευσης για όλους τους συνδυασμούς υπερπαραμέτρων για τα ΣΝΔ Ήχου, για τα τρία σετ ήχου

CNN	Ακρίβεια ταξινόμησης (%)			Χρόνος εκπαίδευσης (s)		
	Urban Sound8K	**ESC-10**	**Air Compressor**	**Urban Sound8K**	**ESC-10**	**Air Compressor**
VGGish	95.68	82.36	99.97	253	18	45
YAMNet	96.24	88.06	100	797	55	151

2.5.3 Βέλτιστοι συνδυασμοί υπερπαραμέτρων

Για την επιλογή των καλύτερων συνδυασμών τιμών των υπερπαραμέτρων διαμορφώθηκαν δύο κριτήρια. Το πρώτο κριτήριο εξαρτάται από την απόδοση της ακρίβειας ταξινόμησης (AT) και τον χρόνο εκπαίδευσης (ΧΕ) και συγκεκριμένα απαιτείται το μοντέλο να έχει καλύτερες επιδόσεις σε αυτές τις τιμές κατά τουλάχιστον μία τυπική απόκλιση (std) από τις μέσες τιμές τους (avg). Προφανώς, η ΑΤ πρέπει να είναι μεγαλύτερη και ο ΧΕ μικρότερος κατά μία τυπική απόκλιση αντίστοιχα (Σχέση 2.4). Το δεύτερο κριτήριο χρησιμοποιεί τον σταθμισμένο μέσο όρο της (θετικής) διαφοράς αυτών των τιμών προς τις μέσες τιμές τους. Το βάρος K επέτρεψε την κανονικοποίηση και έδωσε προτεραιότητα στην ΑΤ με συντελεστή 2 (Σχέση 2.5).

$$XE < \left(XE_{avg} - XE_{std}\right) \quad \& \quad AT > \left(AT_{avg} - AT_{std}\right) \tag{2.4}$$

$$\left(XE_{avg} - XE\right) + K * \left(AT - AT_{avg}\right), \text{όπου } K = 2 * XE_{avg}/AT_{avg} \tag{2.5}$$

Σύμφωνα με τα δύο αυτά κριτήρια καθορίστηκαν οι πιο αποτελεσματικοί συνδυασμοί τιμών των υπερπαραμέτρων. Στον Πίνακα 2.11 παρουσιάζονται αυτοί για τα ΣΝΔ Ήχου, και στον Πίνακα 2.12 για τα ΣΝΔ Εικόνας.

Πίνακας 2.11: Οι αποτελεσματικότεροι συνδυασμοί τιμών των υπερπαραμέτρων για τα ΣΝΔ Ήχου, για κάθε σετ ήχων

ΣΝΔ Ήχου	**VGGish**				**YAMNet**			
Σύνολο ήχων	Optimizer	Mini Batch Size	Epochs	Learning Rate ($\times 10^{-4}$)	Optimizer	Mini Batch Size	Epochs	Learning Rate ($\times 10^{-4}$)
Urban Sound8K	Adam	256	4	2	Adam	256	4	2
ESC-10	Adam	256	5	2	Adam	256	5	2
Air Compressor	Adam	256	5	2	Adam	256	5	2

Πίνακας 2.12: Οι αποτελεσματικότεροι συνδυασμοί υπερπαραμέτρων για τα ΣΝΔ Εικόνας, για κάθε σετ ήχων

ΣΝΔ Εικόνας	**GoogleNet**				**SqueezeNet**				**ShuffleNet**			
Σύνολο ήχων	Optimizer	Mini Batch Size	Epochs	Learning Rate ($\times 10^{-4}$)	Optimizer	Mini Batch Size	Epochs	Learning Rate ($*10^{-4}$)	Optimizer	Mini Batch Size	Epochs	Learning Rate ($\times 10^{-4}$)
Urban Sound8K	Adam	32	8	2	Adam	32	8	2	Adam	32	6	2
ESC-10	Adam	32	8	2	Adam	32	6	2	Adam	32	8	2
Air Compressor	Adam	32	10	2	Adam	32	6	2	Adam	32	8	2

Η έρευνα έδειξε ότι τα ΣΝΔ Εικόνας ήταν ευθυγραμμισμένα ως προς τον optimizer (Adam), το μέγεθος του mini-batch (32) και τον ρυθμό μάθησης (2×10^{-4}) για τα τρία σύνολα δεδομένων. Ο αριθμός των εποχών που επηρέασαν τον αριθμό των επαναλήψεων στη διαδικασία εκπαίδευσης διέφερε μεταξύ των συνόλων δεδομένων, γεγονός που σχετίζεται άμεσα με την αρχιτεκτονική του δικτύου και την ποικιλομορφία των δεδομένων (π.χ. εγγραφές σε διάφορα περιβάλλοντα όπου ακούγονταν άλλοι ήχοι στο παρασκήνιο ή ο διαφορετικός τόνος).

Για τα ΣΝΔ Ήχου, επίσης τα VGGish και YAMNet ευθυγραμμίστηκαν όσον αφορά τον optimizer (Adam), το μέγεθος του mini-batch (256) και τον ρυθμό μάθησης (2×10^{-4}). Με βάση τους αποτελεσματικότερους συνδυασμούς, πραγματοποιήθηκε ταξινόμηση στο υποσύνολο ελέγχου. Προκειμένου να υπάρχει συνέπεια στην ανάκτηση των αποτελεσμάτων, επιλέχθηκε το ίδιο σύνολο ψευδοτυχαίων αρχείων ελέγχου για όλες τις περιπτώσεις. Για την περίπτωση των ΣΝΔ Εικόνας επιλέχθηκε ένα σύνολο εικόνων scalograms, ενώ για την περίπτωση των ΣΝΔ Ήχου το ίδιο υποσύνολο αρχείων ήχου μετατράπηκε σε spectrograms. Ο χρόνος εκπαίδευσης αντιστοιχεί στην εκπαίδευση του (60 + 20)% των αρχείων.

Ο χρόνος προεπεξεργασίας αφορά στον χρόνο για τη δημιουργία των διαγραμμάτων scalograms και των spectrograms για κάθε αρχείο ολόκληρου του συνόλου δεδομένων για τα ΣΝΔ Εικόνας και Ήχου, αντίστοιχα. Για τη δημιουργία των διαγραμμάτων scalograms ο χρόνος προεπεξεργασίας ήταν 11172 s για την περίπτωση του UrbanSound8K, 522 s για το ESC-10 και 432 s για το σύνολο δεδομένων Air Compressor. Τα αποτελέσματα της ακρίβειας ταξινόμησης και του χρόνου εκπαίδευσης για τα τρία ΣΝΔ Εικόνας παρουσιάζονται στον Πίνακα 2.13.

Πίνακας 2.13: Η απόδοση ταξινόμησης και ο χρόνος εκπαίδευσης για τα ΣΝΔ Εικόνας για τα τρία σύνολα ήχων

ΣΝΔ Εικόνας	**GoogleNet**		**SqueezeNet**		**ShuffleNet**	
Σύνολο ήχων	**AT (%)**	**XE (s)**	**AT (%)**	**XE (s)**	**AT (%)**	**XE (s)**
UrbanSound8K	87.06	348	83.97	192	87.18	780
ESC-10	86.25	17	83.75	7	87.50	53
Air Compressor	97.22	90	94.72	27	97.22	217

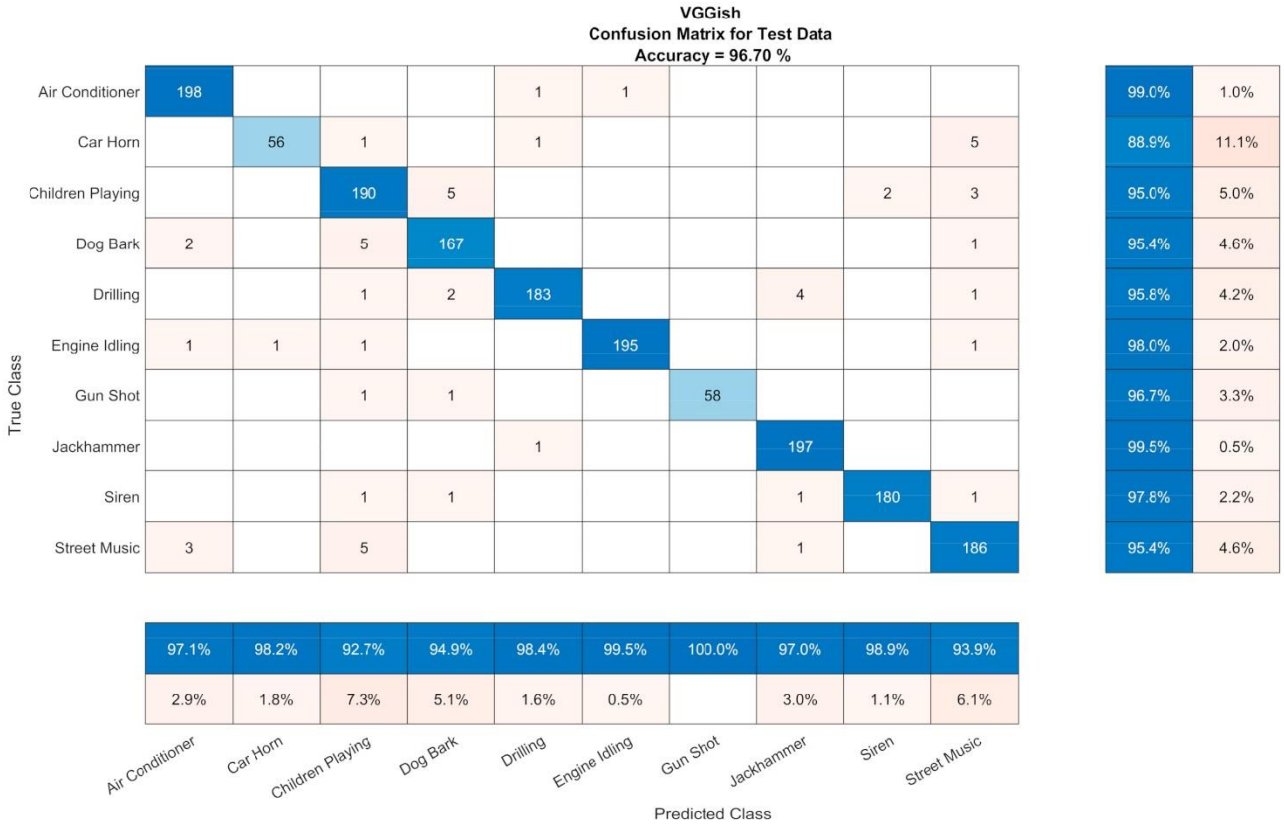
Ένα ακόμη έλεγχος που πραγματοποιήθηκε ήταν η τροφοδοσία των ΣΝΔ Εικόνας με τα Mel-spectrograms των αρχείων ήχου που αποθηκεύτηκαν ως εικόνες με διαστάσεις κατάλληλες για τα συγκεκριμένα δίκτυα. Αυτή η διαδικασία χρειάστηκε 6270 δευτερόλεπτα (χρόνος προεπεξεργασίας) για το UrbanSound8K και είχε ως αποτέλεσμα μια αύξηση της τάξεως του 2.3% στην ακρίβεια ταξινόμησης και 27.3% στον χρόνο εκπαίδευσης, κατά μέσο όρο.

Οι εικόνες εισόδου στα ΣΝΔ Ήχου είναι τα spectrograms. Για την δημιουργία αυτών ο χρόνος που χρειάστηκε (ο χρόνος προεπεξεργασίας) ήταν 783 s για το UrbanSound8K, 9s για το ESC-10 και 27s για το Air Compressor. Τα αποτελέσματα της ακρίβειας ταξινόμησης και του χρόνου εκπαίδευσης με χρήση των ΣΝΔ Ήχου συγκεντρώνονται στον Πίνακα 2.14.

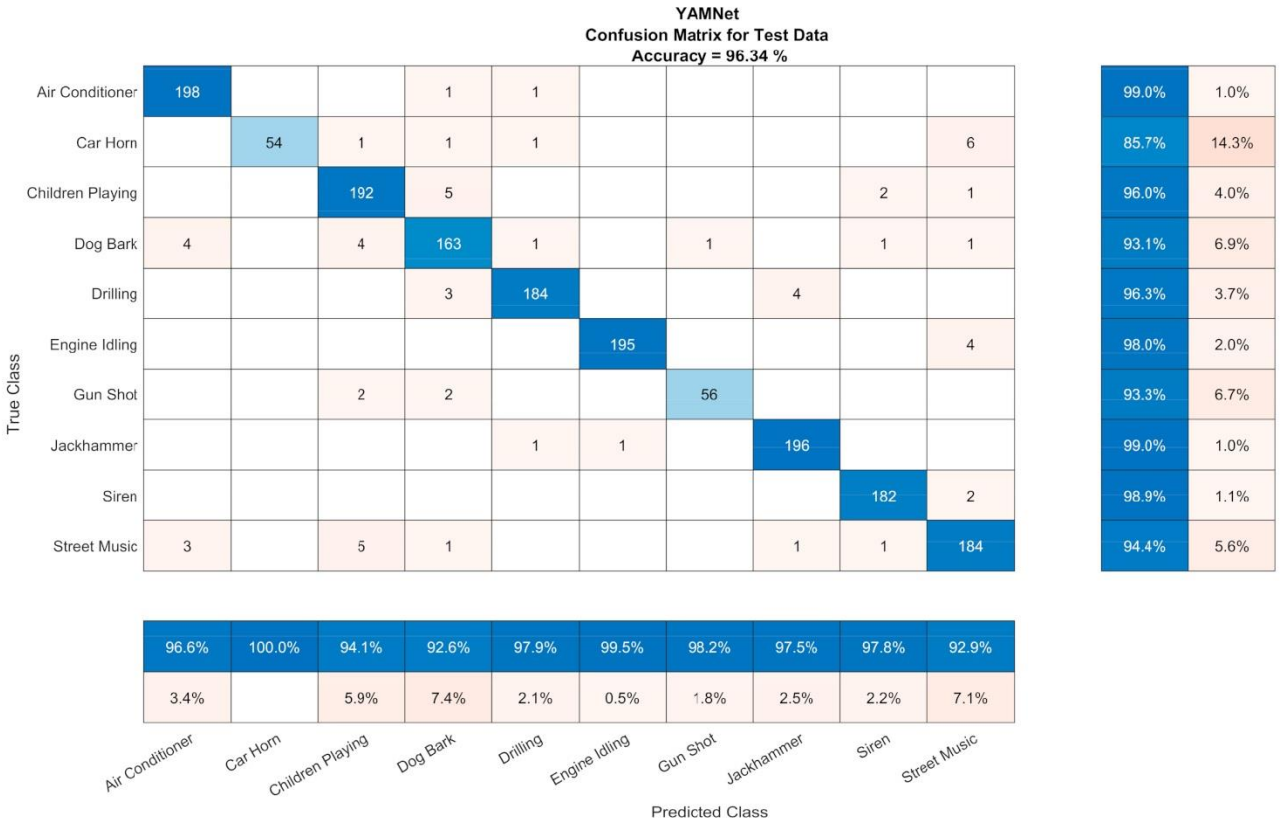
Πίνακας 2.14: Η ακρίβεια ταξινόμησης και ο χρόνος εκπαίδευσης των ΣΝΔ Ήχου για τα τρία σύνολα ήχων

ΣΝΔ Ήχου	**VGGish**		**YAMNet**	
Σύνολο ήχων	**AT (%)**	**XE (s)**	**AT (%)**	**XE (s)**
UrbanSound8K	96.70	210	96.16	603
ESC-10	91.25	16	91.25	36
Air Compressor	100	37	100	62

Τα αποτελέσματα της ακρίβεια ταξινόμησης για τα ΣΝΔ Ήχου για το σύνολο UrbanSound8K αφορούν το υποσύνολο των αρχείων ελέγχου τα οποία έχουν διάρκεια μεγαλύτερη από 0,96 s, που είναι το ελάχιστο τμήμα στην περίπτωση κατάτμησης του ηχητικού αρχείου (το ESC-10 και το Air Compressor είχαν αρχεία μήκους 5 και 3 δευτερολέπτων αντίστοιχα). Λαμβάνοντας υπόψη και τα υπόλοιπα αρχεία ελέγχου για το UrbanSound8K (δηλαδή, αυτά με διάρκεια μικρότερη από το ελάχιστο) ως λανθασμένες προβλέψεις η απόδοση ταξινόμησης φτάνει το 92.05% και 91.99% για το VGGish και το YAMNet, αντίστοιχα. Οι Πίνακες σύγχυσης των ΣΝΔ Ήχου, χρησιμοποιώντας τους αποτελεσματικότερους συνδυασμούς τιμών των υπερπαραμέτρων, παρουσιάζονται στα Σχήματα 2.6 και 2.7.

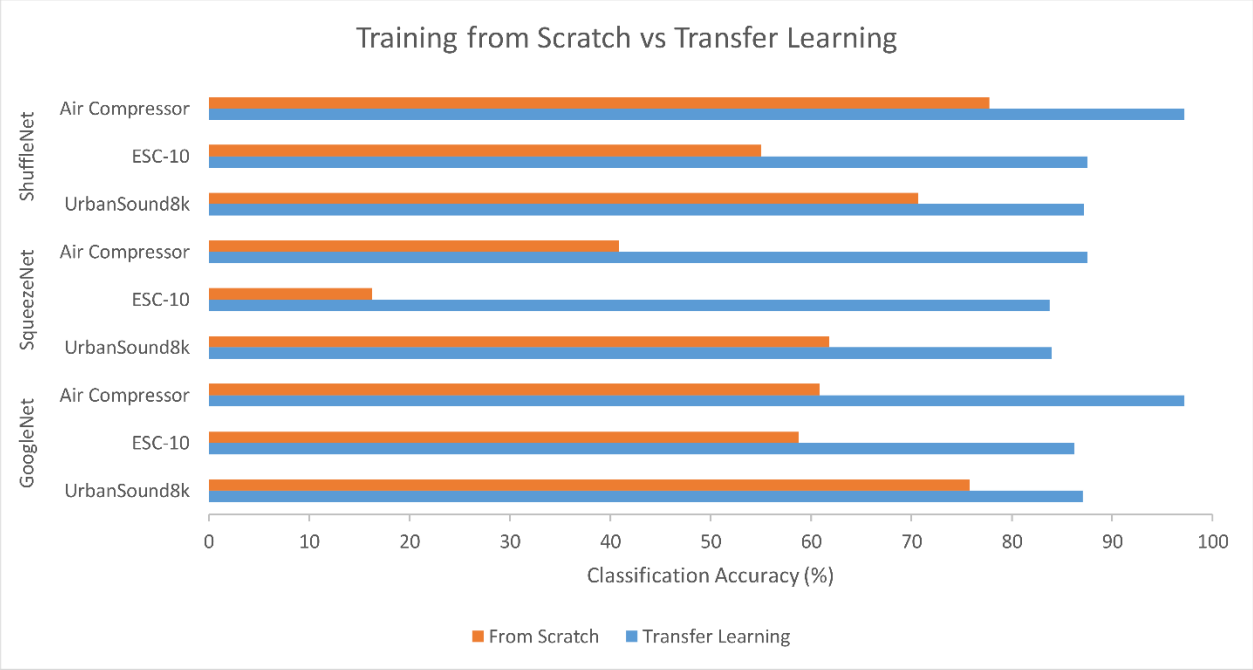


Σχήμα 2.6: Ο Πίνακας Σύγχυσης για το VGGish με τον αποτελεσματικότερο συνδυασμό τιμών υπερπαραμέτρων (Adam optimizer, 256 mini-batch size, 4 epochs, 2×10^{-4} learning rate).



Σχήμα 2.7: Ο Πίνακας Σύγχυσης για το YAMNet με τον αποτελεσματικότερο συνδυασμό τιμών υπερπαραμέτρων (Adam optimizer, 256 mini-batch size, 4 epochs, 2×10^{-4} learning rate).

Σε αυτό το σημείο, είναι σκόπιμο να συγκριθούν τα αποτελέσματα της μεταφοράς μάθησης με τα αντίστοιχα της εξαρχής εκπαίδευσης (training from the scratch). Αξιοποιώντας τους αποδοτικότερους συνδυασμούς τιμών των υπερπαραμέτρων (Πίνακας 2.12), εκπαιδεύτηκαν from the scratch τα ΣΝΔ Εικόνας στα τρία σύνολα δεδομένων ήχου. Η ακρίβεια ταξινόμησης που επιτεύχθηκε είναι μικρότερη από την αντίστοιχη με την μεταφορά μάθησης, όπως φαίνεται στο Σχήμα 2.8. Συγκεκριμένα, η ακρίβεια ταξινόμησης είναι κατά μέσο όρο μικρότερη κατά 27% για το GoogleNet, κατά 57% για το SqueezeNet και 25% για το ShuffleNet.



Σχήμα 2.8: Σύγκριση της ακρίβειας ταξινόμησης που επιτεύχθηκε με μεταφορά μάθησης και με training from the scratch για τα τρία σύνολα δεδομένων ήχου με ΣΝΔ Εικόνας.

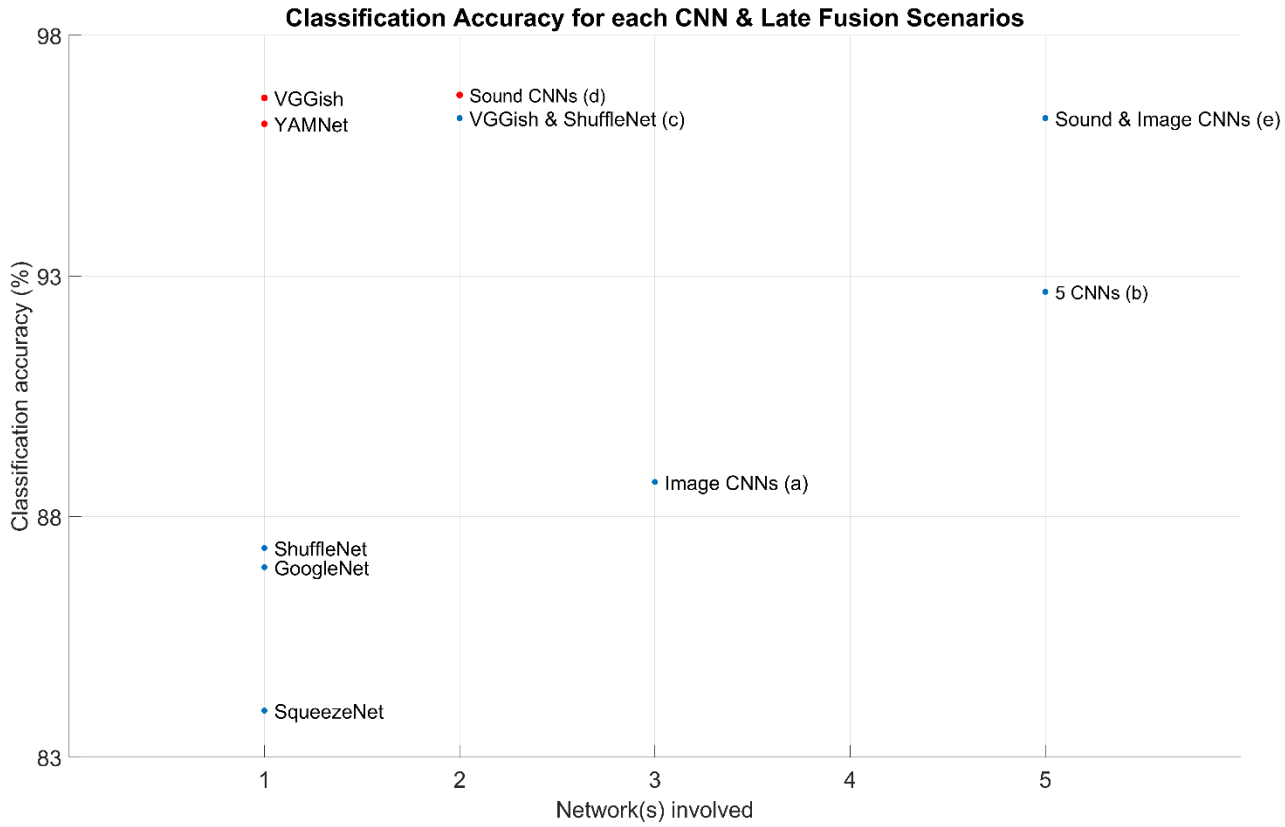
Τα αποτελέσματα αυτά οφείλονται στην απαίτηση της from the scratch εκπαίδευσης των δικτύων με μεγαλύτερα σύνολα δεδομένων αλλά και με περαιτέρω επαναλήψεις για τον υπολογισμό των βαρών. Η παρατήρηση αυτή ενισχύει την ελκυστικότητα της τεχνικής της μεταφοράς μάθησης.

2.5.4 Συγχώνευση μεθόδων

Η διαθεσιμότητα πολλαπλών επανεκπαιδευμένων δικτύων επιτρέπει τον συνδυασμό των αποτελεσμάτων (ύστερη συγχώνευση-late fusion). Εξετάστηκαν τα ακόλουθα σενάρια για την επιλογή του αποτελέσματος της ταξινόμησης (πρόβλεψης):

α. Επιλογή με βάση την πλειοψηφία των προβλέψεων, λαμβάνοντας υπόψη τα τρία ΣΝΔ εικόνας.
β. Επιλογή με βάση την πλειοψηφία των προβλέψεων, λαμβάνοντας υπόψη τα πέντε ΣΝΔ.
γ. Συγχώνευση των προβλέψεων του αποδοτικότερου (δηλαδή με την υψηλότερη ακρίβεια ταξινόμησης) ΣΝΔ Ήχου (VGGish) με τις αντίστοιχες του αποδοτικότερου ΣΝΔ Εικόνας (ShuffleNet), για τις περιπτώσεις που η διάρκεια των ηχητικών αποσπασμάτων ήταν μεγαλύτερη από το απαραίτητο όριο.
δ. Συγχώνευση των προβλέψεων των ΣΝΔ Ήχου μετά τον υπολογισμό του στάθμης εμπιστοσύνης (confidence level) για κάθε αποτέλεσμα ταξινόμησης. Η στάθμη εμπιστοσύνης υπολογίστηκε ως ο λόγος του αριθμού των επιλεγμένων ταξινομήσεων προς τον συνολικό αριθμό τμημάτων του ηχητικού αποσπάσματος.
ε. Συγχώνευση των προβλέψεων του ΣΝΔ Ήχου (χρησιμοποιώντας την στάθμη εμπιστοσύνης) σε συνδυασμό με την πρόβλεψη που βασίζεται στην πλειοψηφία των προβλέψεων των τριών ΣΝΔ Εικόνας για τα ηχητικά αποσπάσματα με διάρκεια μικρότερη από το όριο των 0.96 s.

Τα αποτελέσματα της ακρίβειας ταξινόμησης (%) ανά σενάριο συγχώνευσης αλλά και τα αντίστοιχα από κάθε ένα ΣΝΔ (Εικόνας ή Ήχου) για το σύνολο UrbanSound8K παρουσιάζονται στο Σχήμα 2.9.



Σχήμα 2.9: Η ακρίβεια ταξινόμησης για κάθε ένα ΣΝΔ και για κάθε σενάριο συγχώνευσης αποτελεσμάτων (α-ε). Οι μπλε κουκίδες αναφέρονται στην ακρίβεια ταξινόμησης λαμβάνοντας υπόψη ολόκληρο το σετ ελέγχου, ενώ οι κόκκινες κουκίδες αναφέρονται στην ακρίβεια ταξινόμησης στο σύνολο των αρχείων που έχουν διάρκεια μεγαλύτερη από το επιτρεπόμενο όριο.

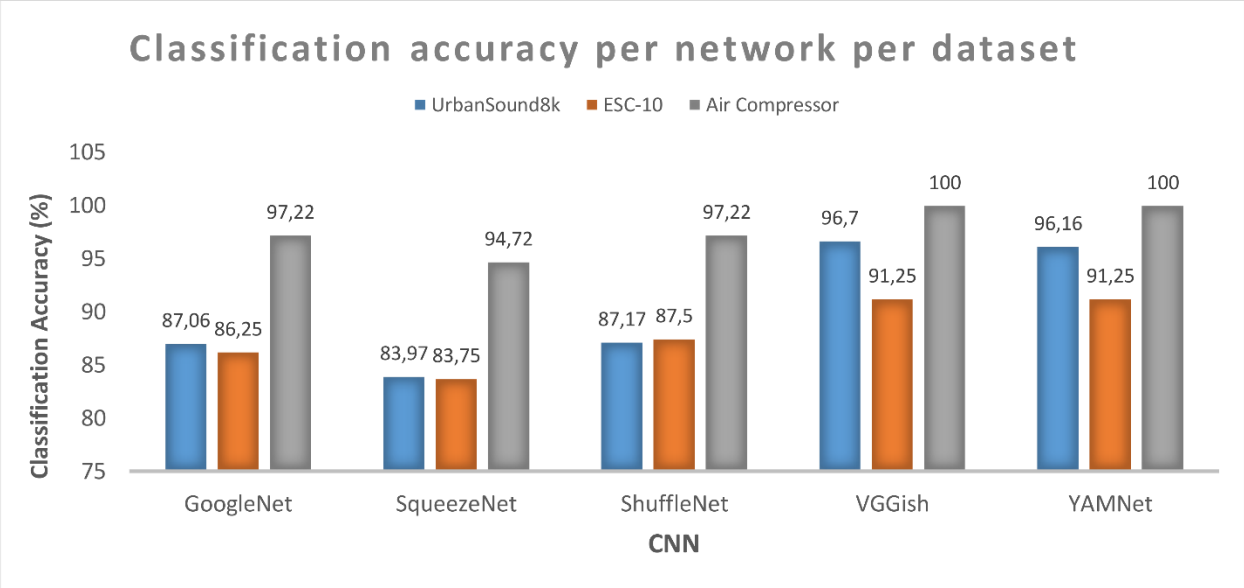
2.6 Συμπεράσματα

Το σύνολο των στόχων που τέθηκαν στην παρούσα φάση της έρευνας εκπληρώθηκαν. Συγκεκριμένα, η προκαθορισμένη χρήση των ΣΝΔ για ταξινόμηση εικόνων επεκτάθηκε στην ταξινόμηση του ήχου. Ο κοινός παρονομαστής των δύο πεδίων της έρευνας είναι η μετατροπή του ηχητικού σήματος σε εικόνες ώστε να τροφοδοτηθούν τα ΣΝΔ.

Για την μετατροπή του ηχητικού σήματος σε εικόνα χρησιμοποιήθηκαν ξεχωριστοί μετασχηματισμοί. Για τα ΣΝΔ Εικόνας τα ηχητικά αρχεία μετατράπηκαν σε διαγράμματα κλίμακας-χρόνου (scalograms), ενώ για τα ΣΝΔ Ήχου τα ηχητικά αρχεία μετατράπηκαν σε φασματογράμματα (spectrograms). Παράλληλα, υπάρχει η δυνατότητα κάθε ηχητικό απόσπασμα να μετατραπεί σε μία ενιαία εικόνα (για την περίπτωση των ΣΝΔ Εικόνας), ή σε πολλαπλές εικόνες οι οποίες αντιστοιχούν σε τμήματα του ηχητικού αποσπάσματος (για την περίπτωση των ΣΝΔ Ήχου). Η τελευταία δυνατότητα επιτρέπει μία πιο λεπτομερή αξιολόγηση της ταξινόμησης, ενώ ταυτόχρονα η ομοιογένεια των αποτελεσμάτων ταξινόμησης των τμημάτων μπορεί να μεταφραστεί ως η στάθμη εμπιστοσύνης για την ταξινόμηση του συνολικού ηχητικού αποσπάσματος.

Αξιοποιώντας τα ΣΝΔ για την ταξινόμηση του ήχου, αφενός δεν αποσαφηνίζεται το ηχητικό σήμα όπως με τις κλασσικές μεθόδους ΜΜ στις οποίες εξάγονται συγκεκριμένα χαρακτηριστικά, αφετέρου αυτή η αδιαφανής διαδικασία αντισταθμίζεται από την υψηλότερη απόδοση στην ακρίβεια ταξινόμησης. Ταυτόχρονα, η εκπαίδευση των ΣΝΔ απαιτεί μεγάλο όγκο δεδομένων και κατ' επέκταση την κατανάλωση υπολογιστικών πόρων. Η τεχνική της μεταφοράς μάθησης δίνει την δυνατότητα για μία πιο αποδοτική χρήση αυτών.

Η μεταφορά μάθησης έχει διάφορες παραλλαγές όσον αφορά τα τμήματα των ΣΝΔ που επαναχρησιμοποιούνται ή επανεκπαιδεύονται, ενώ παράλληλα οι παράμετροι της επανεκπαίδευσης (υπερπαράμετροι) επηρεάζουν τόσο την ακρίβεια της ταξινόμησης όσο και τον χρόνο της επανεκπαίδευσης. Λαμβάνοντας υπόψη αυτά τα δύο μεγέθη (ΑΤ και ΧΕ) διατυπώθηκαν δύο διαφορετικά κριτήρια για την επιλογή του καλύτερου συνδυασμού των τιμών των υπερπαραμέτρων. Ο καλύτερος συνδυασμός των υπερπαραμέτρων είναι αυτός που μεγιστοποιεί την απόδοση ενός δικτύου σύμφωνα με το υπό εξέταση κριτήριο. Τα αποτελέσματα των δοκιμών που πραγματοποιήθηκαν, χρησιμοποιώντας τρία σύνολα δεδομένων ήχου, υποδεικνύουν ότι τα ΣΝΔ Ήχου επιτυγχάνουν καλύτερες επιδόσεις από τα ΣΝΔ Εικόνας, με το VGGish να έχει την μέγιστη απόδοση και το YAMNet να ακολουθεί. Στο Σχήμα 2.10 συγκεντρώνονται τα αποτελέσματα της ακρίβειας ταξινόμησης ανά δίκτυο και σύνολο δεδομένων ήχου.



Σχήμα 2.10: Η υψηλότερη απόδοση της ακρίβειας ταξινόμησης που επιτεύχθηκε ανά ΣΝΔ και ανά σύνολο δεδομένων ήχου.

Από αυτή την σκοπιά επιτεύχθηκε ο εξορθολογισμός και η επικύρωση της ταξινόμησης του ήχου με μεθόδους που βασίζονται σε ΣΝΔ και της τεχνικής της μεταφοράς μάθησης. Επίσης, αναδείχθηκαν διαφορετικές δυνατότητες παραμετροποίησης της διαδικασίας της μεταφοράς μάθησης διερευνώντας τον αντίκτυπο που αυτές έχουν στην ακρίβεια ταξινόμησης.

Τα σενάρια συνδυασμού (fusion) των αποτελεσμάτων ταξινόμησης από τα διάφορα ΣΝΔ μπορούν επίσης να έχουν θετικά αποτελέσματα. Ο συνδυασμός των αποτελεσμάτων μπορεί να πραγματοποιηθεί είτε με έναν μόνο ταξινομητή μέσω τμηματοποίησης του ήχου, είτε με ανεξάρτητες

ταξινομήσεις ανά τμήμα ή με συγχώνευση των αποτελεσμάτων (για τα ΣΝΔ Ήχου). Επίσης μπορεί να πραγματοποιηθεί συνδυασμός αποτελεσμάτων πρόβλεψης, εάν είναι διαθέσιμοι πολλοί ταξινομητές, ως ταξινόμηση που βασίζεται στην πλειοψηφία.

Για το σύνολο δεδομένων UrbanSound8K, ο συνδυασμός των τριών ΣΝΔ Εικόνας ξεπέρασε την απόδοση των μεμονωμένων δικτύων. Η συγχώνευση και των πέντε δικτύων ήταν περίπου ο μέσος όρος του συνδυασμού των τριών δικτύων Εικόνας και του συνδυασμού των δύο δικτύων Ήχου. Τα σενάρια συγχώνευσης των αποτελεσμάτων των VGGish και YAMNet με τα αποτελέσματα της πλειοψηφίας των τριών ΣΝΔ Εικόνας οδήγησαν την ακρίβεια ταξινόμησης στο 96.28%.

Τέλος, η ακρίβεια ταξινόμησης των ίδιων ΣΝΔ σε διαφορετικά σύνολα δεδομένων επιτρέπει συγκριτικές παρατηρήσεις όσον αφορά την ποσότητα και την ποιότητα του συνόλου δεδομένων καθώς και της καθαρότητας των τύπων (ή της επικάλυψης πολλών ήχων). Όσον αφορά τα τρία σύνολα δεδομένων που χρησιμοποιήσαμε, το Air Compressor είχε την καλύτερη συμπεριφορά καθώς έδωσε την υψηλότερη ακρίβεια ταξινόμησης για όλα τα ΣΝΔ. Πρέπει να επισημανθεί όμως ότι αυτό το σύνολο δεδομένων είχε τον μικρότερο αριθμό κλάσεων καθώς και περιορισμένο αριθμό αρχείων ήχου.

Το σύνολο αυτής της έρευνας καθώς και τα πειραματικά αποτελέσματα των επιμέρους μεθόδων αντανακλώνται στην δημοσίευση [87].

3. Περαιτέρω διερεύνηση της ταξινόμησης του ήχου

3.1 Κατάτμηση και ταξινόμηση των ηχητικών ροών με βάση τα ΣΝΔ σε ρεαλιστικές συνθήκες

3.1.1 Εισαγωγή

Οι τεχνικές ΒΜ εφαρμόζονται ολοένα και περισσότερο σε προβλήματα ταξινόμησης του ήχου ως εναλλακτικές στις παραδοσιακές τεχνικές ΜΜ όπου εξάγονται οι τιμές των χαρακτηριστικών του ήχου. Υπάρχουν ΣΝΔ τα οποία έχουν σχεδιαστεί ειδικά για την ταξινόμηση των ήχων όπως το YAMNet, VGGish [88], OpenL3 [89] και Crepe [90]. Παράλληλα, δημιουργούνται σύνολα δεδομένων ήχου, τα οποία αποτελούνται από ήχους διαφόρων τύπων (που εκτείνονται από συνηθισμένους ήχους έως ήχους πιο εξειδικευμένων περιπτώσεων όπως π.χ. η λειτουργία βιομηχανικών μηχανών) και τα οποία χρησιμοποιούνται για την εκπαίδευση και τον έλεγχο των αλγορίθμων. Μία πρόκληση όσον αφορά τη χρήση των συνόλων δεδομένων ήχου οφείλεται στην έλλειψη μηχανισμού συσχέτισης μεταξύ των κλάσεων ήχου που περιλαμβάνονται στα ίδια ή σε διαφορετικά σύνολα δεδομένων. Η συσχέτιση σχετίζεται με την ομοιότητα των κλάσεων ως προς το πλαίσιο (π.χ. την προέλευση), την τεχνολογία (π.χ. ηχητικά χαρακτηριστικά) και τη τοποθέτηση ετικέτας του ήχου η οποία συνήθως πραγματοποιείται χειροκίνητα. Η τοποθέτηση ετικέτας μπορεί επίσης να έχει διαφορετικό επίπεδο λεπτομέρειας, π.χ. από τους γενικούς *ανθρώπινους ήχους* έως τους πιο ειδικούς π.χ. του *κλάματος μωρού*. Μία άλλη διαφορά μεταξύ των συνόλων δεδομένων ήχου σχετίζεται με την ποιότητα των ήχων. Μία περαιτέρω πτυχή είναι ότι, εκτός από τις ρεαλιστικές συνθήκες ταξινόμησης των ήχων, οι ήχοι εμφανίζονται με διαδοχικό ή ακόμα και ταυτόχρονο τρόπο με αποτέλεσμα να απαιτούνται ισχυρές τεχνικές διαχωρισμού. Αυτοί οι παράγοντες δημιουργούν μία τμηματική προσέγγιση στο τοπίο των ηχητικών συνόλων δεδομένων.

3.1.2 Στόχοι

Στην συγκεκριμένη μελέτη διερευνώνται δύο τύποι συστηματικών συσχετίσεων μεταξύ των κλάσεων των ήχων: α) η *σημασιολογική*, λαμβάνοντας υπόψη την προέλευση και β) την σύγκριση με βάση τα ηχητικά χαρακτηριστικά. Όσον αφορά το (α), οι κλάσεις μπορούν να συνδεθούν σημασιολογικά λαμβάνοντας υπόψη ένα ενοποιητικό γράφημα οντολογίας. Μία από τις πρώτες και πιο συστηματικές προσεγγίσεις έχει πραγματοποιηθεί στο [53] με την οντολογία AudioSet. Η οντολογία AudioSet περιλαμβάνει 632 κλάσεις συνδεδεμένες σε μία δενδροειδή δομή, με την συσχέτιση του ήχου να βασίζεται στην προέλευση του ήχου. Αυτό μπορεί να επιτρέψει την αντιστοίχιση υψηλού επιπέδου κλάσεων που ανήκουν σε διαφορετικά σύνολα δεδομένων, ιδίως για πανομοιότυπες ή παρόμοιες κλάσεις. Όσον αφορά το (β), η συσχέτιση των κλάσεων ήχου μπορεί να βασίζεται στην εξαγωγή χαρακτηριστικών ήχου και στον υπολογισμό της (Ευκλείδειας ή άλλης) απόστασης μεταξύ τους. Τα χαρακτηριστικά μπορούν να εξαχθούν και να σταθμιστούν σύμφωνα με τις υπάρχουσες μεθοδολογίες [91]. Τέτοιες μετρικές ομοιότητας (αποστάσεις) μπορεί να μην συμπίπτουν απαραίτητα με την σημασιολογική συγγένεια των ήχων.

Ο ήχος που προέρχεται από ρεαλιστικά περιβάλλοντα μπορεί να περιλαμβάνει πολλούς τύπους, συνδυασμένους με διαδοχικό ή/και επικαλυπτόμενο τρόπο. Σε τέτοιες περιπτώσεις, η ταξινόμηση και η ταυτοποίηση των ήχων θα πρέπει να εξετάζει επίσης τον διαχωρισμό των ήχων και την διαχείριση των ταυτοποιημένων ήχων. Αυτή η διαδικασία περιλαμβάνει αποφάσεις σχετικά με α) τον διαχωρισμό ή την συγχώνευση των ίδιων ηχητικών αποσπασμάτων ανάλογα με την χρονική τους γειτνίαση και β) με αποφάσεις που σχετίζονται με το κατώφλι, όπως π.χ. η ελάχιστη διάρκεια ώστε να λαμβάνεται υπόψη ένα αναγνωρισμένο ηχητικό απόσπασμα. Ο τομέας αυτός ερευνάται αποτελεσματικά με τεχνικές δυναμικής ανάλυσης και κατάτμησης (segmentation) των ηχητικών

χρονοσειρών, αλλά ταυτόχρονα παρουσιάζει προκλήσεις σε σενάρια συνεχούς ροής ήχου (δηλαδή χωρίς προκαθορισμένους διαλόγους).

Λαμβάνοντας υπόψη τα παραπάνω ο στόχοι της συγκεκριμένης δημοσίευσης συνοψίζονται στους εξής:

- Διερεύνηση της συσχέτισης μεταξύ των κλάσεων του ήχου, λαμβάνοντας υπόψη τόσο την σημασιολογική οπτική όσο και τις τιμές των ηχητικών χαρακτηριστικών.
- Αξιολόγηση της ταξινόμησης των ήχων με την χρήση σύγχρονων εργαλείων ΣΝΔ και συμβολή στην αντιστοίχιση κλάσεων ήχου οι οποίες ανήκουν στο ίδιο ή σε διαφορετικά σύνολα δεδομένων, με ποσοτικό τρόπο λαμβάνοντας υπόψη την σημασιολογική συγγένεια και την ομοιότητα βάσει των ηχητικών χαρακτηριστικών.
- Επέκταση των σεναρίων ταξινόμησης ήχων από ανεξάρτητες, διακριτές κατηγορίες ήχων προς σύνθετες ροές ήχων μεγαλύτερης διάρκειας, οι οποίες περιλαμβάνουν πολλαπλούς τύπους ήχων και σχεδιασμός αλγορίθμου κατάτμησης και ταξινόμησης ήχων σε τέτοια περιβάλλοντα.

3.1.3 Συναφής έρευνα

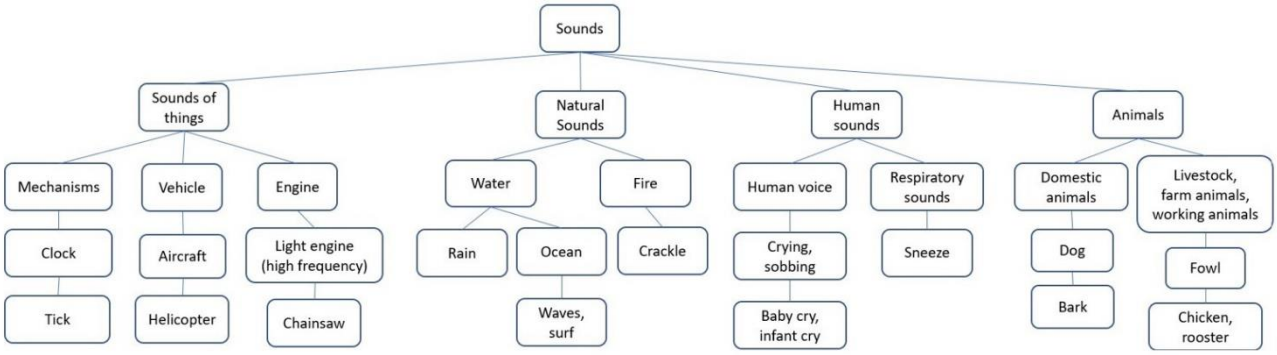
Η ταυτόχρονη παρουσία ήχων από διαφορετικές πηγές είτε σε εξωτερικούς χώρους (αστικά κέντρα ή περιβαλλοντικοί χώροι) [92], είτε σε εσωτερικούς χώρους (επιχειρήσεις, κατοικίες, εκπαιδευτικά κέντρα) [93] καθιστά την ταξινόμηση του ήχου ένα προκλητικό ερευνητικό πεδίο. Οι τεχνικές ταξινόμησης ήχων προσανατολίζονται όλο και περισσότερο σε μηχανισμούς ΜΜ [94], και ΒΜ [95], [96]. Στην πρώτη περίπτωση εξάγονται χαρακτηριστικά του ήχου και αξιολογούνται αυτά ως προς την περιγραφική τους δύναμη, χρησιμοποιώντας τεχνικές όπως η Relief-F [97] ή η PCA [98], [40]. Εν συνεχεία, αυτά τα επιλεγόμενα χαρακτηριστικά τροφοδοτούν τα μοντέλα ταξινόμησης. Στην ΒΜ η διαδικασία λαμβάνει χώρα εσωτερικά στα ΣΝΔ τα οποία εκπαιδεύονται σταδιακά από επίπεδο σε επίπεδο, με το τελευταίο να εκτελεί την ταξινόμηση [99]. Η μεταφορά μάθησης είναι μία ευρέως χρησιμοποιούμενη μέθοδος ταξινόμησης όπου, τα προεκπαιδευμένα ΣΝΔ σε ένα μεγάλο σύνολο εικόνων (π.χ. ImageNet) ή σε ένα μεγάλο σύνολο ήχων (π.χ. AudioSet), επανεκπαιδεύονται στο υπό εξέταση σύνολο δεδομένων. Η επανεκπαίδευση των ΣΝΔ έχει συνήθως οφέλη σε υπολογιστικούς πόρους και ακρίβεια ταξινόμησης [100].

Η κατάτμηση του ηχητικού σήματος αποτελεί ένα υποσύνολο της τμηματικής επεξεργασίας του ήχου για μη ντετερμινιστικά σήματα. Η κατάτμηση του σήματος βασίζεται στον εντοπισμό και την επεξεργασία αλλαγών στην συχνότητα και το πλάτος του σήματος, ενώ έχουν εφαρμοστεί και άλλες τεχνικές, όπως π.χ. στο [101] όπου έχει χρησιμοποιηθεί ο διακριτός μετασχηματισμός κυματιδίων (Discrete Wavelet Transform – DWT) για την αποσύνθεση σημάτων σε ορθοκανονικές χρονοσειρές με διαφορετικές ζώνες συχνοτήτων (π.χ. σε σήματα ηλεκτροεγκεφαλογραφήματος (electricoencephalography – EEG). Στην περίπτωση του ήχου, το Κριτήριο Πληροφορίας Bayesian (Bayesian Information Criterion) έχει χρησιμοποιηθεί για την κατάτμηση της ομιλίας [102].

3.1.4 Αντιστοίχιση τύπων ήχου και ομοιότητα

Τα σύνολα δεδομένων ήχου περιλαμβάνουν διαφορετικούς τύπους ήχων, οι οποίοι επιλέγονται αυθαίρετα. Η σημασιολογική συσχέτιση αυτών των τύπων μπορεί να επιτευχθεί μέσω της δενδροειδούς ιεραρχίας η οποία βασίζεται στο AudioSet. Αυτό επιβεβαιώνεται με το σύνολο δεδομένων ήχου ESC-10 [50] το οποίο περιλαμβάνει 10 κλάσεις: γάβγισμα σκύλου (dog), βροχή (rain), θαλάσσια κύματα (sea waves), κλάμα μωρού (baby cry), χτύπος ρολογιού (tick), φτέρνισμα (sneeze), ελικόπτερο (helicopter), αλυσοπρίονο (chainwaw), κόκορας (rooster), και ήχος φωτιάς (crackle). Η αντιστοίχιση είναι απλή, καθώς αυτές οι κλάσεις μπορούν να συσχετιστούν με κλάσεις

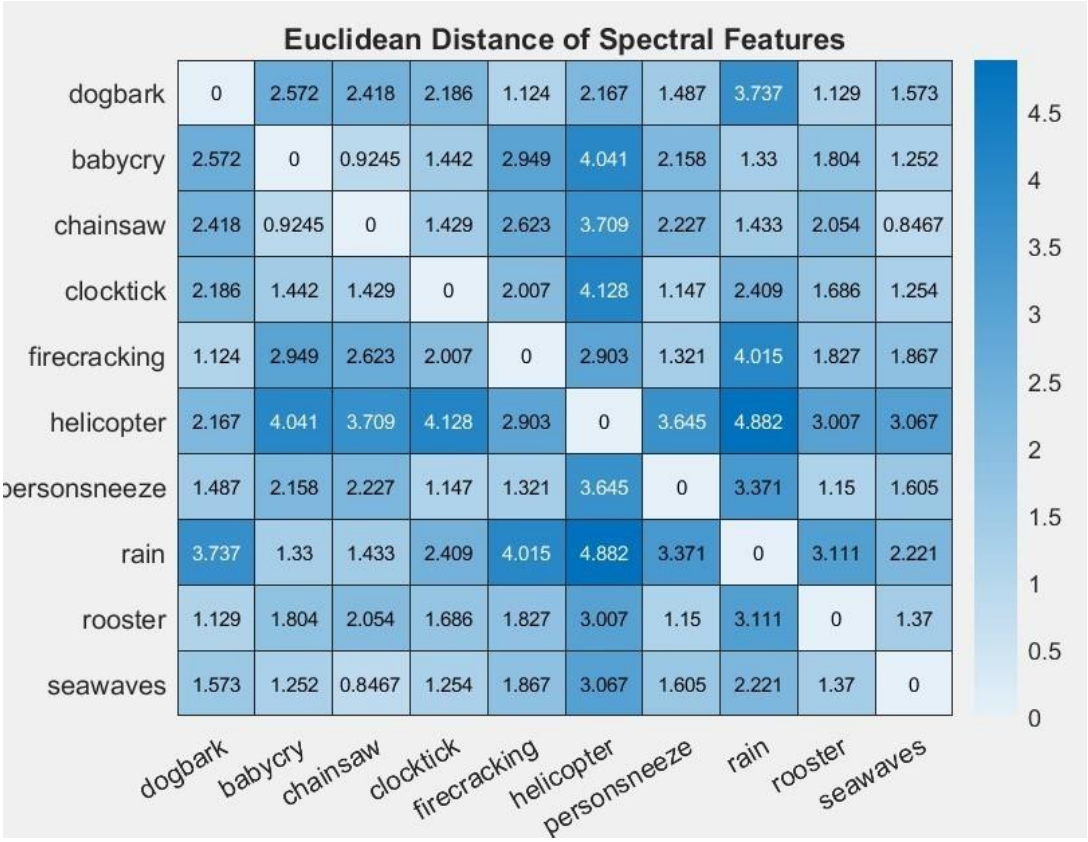
της οντολογίας Audioset όπως φαίνεται στο Σχήμα 3.1. Για παράδειγμα, το χτύπος του ρολογιού, το ελικόπτερο και το αλυσοπρίονο ανήκουν στο 2ο επίπεδο ήχων αντικειμένων, ενώ η βροχή και τα κύματα στο 3ο επίπεδο ήχων νερού.



Σχήμα 3.1: Υποσύνολο της οντολογίας AudioSet στο οποίο περιλαμβάνονται οι κλάσεις του συνόλου ESC-10.

Ένα άλλο κριτήριο για την συσχέτιση των κλάσεων είναι ο υπολογισμός της ομοιότητας βάσει των τιμών των ηχητικών χαρακτηριστικών. Η εξαγωγή ενός υποσυνόλου αντιπροσωπευτικών χαρακτηριστικών και ο υπολογισμός της Ευκλείδειας απόστασης των τιμών τους μπορεί να προσφέρει μία ποσοτική συσχέτιση «τεχνικής» ομοιότητας. Επιλέχθηκαν τα φασματικά Centroid, Flux και Roll-off με βάση την εργασία [40]. Για κάθε κλάση του ESC-10 δημιουργείται ένα αντιπροσωπευτικό αρχείο που αποτελείται από 20 συνδεδεμένα αποσπάσματα. Η προεπεξεργασία περιλαμβάνει την μετατροπή σε μονοφωνικό ήχο, στην περίπτωση που τα αρχικά δείγματα είναι στερεοφωνικά, τον ενιαίο ρυθμό δειγματοληψίας στα 44.1 kHz και το σχήμα κβαντισμού. Τα αρχεία αυτά υποβάλλονται σε επεξεργασία για την εξαγωγή των τιμών των ηχητικών χαρακτηριστικών χρησιμοποιώντας παράθυρο Hanning 3ms με μήκος επικάλυψης 2ms.

Αφού πραγματοποιηθεί ο υπολογισμός του μέσου όρου της κινητής διαμέσου για κάθε χαρακτηριστικό και οι τιμές κανονικοποιηθούν, υπολογίζεται η Ευκλείδεια απόσταση των τιμών των χαρακτηριστικών. Τα αποτελέσματα απεικονίζονται στο Σχήμα 3.2 με τον χάρτη χρωματικής συσχέτισης. Τα αποτελέσματα παρέχουν μία ποσοτική ένδειξη της «τεχνικής» ομοιότητας μεταξύ των κλάσεων. Για παράδειγμα, η απόσταση μεταξύ του γαβγίσματος και του κόκορα είναι πιο μικρή από αυτή μεταξύ της βροχής και του ελικόπτερου.



Σχήμα 3.2: Ομοιότητα κλάσεων ήχου με βάση την Ευκλείδεια απόσταση.

3.1.5 Ταξινόμηση και κατάτμηση ήχου

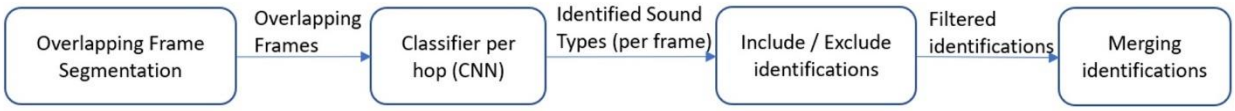
Η ταυτοποίηση του ήχου σε ρεαλιστικές συνθήκες περιλαμβάνει ηχητικές ροές οι οποίες αποτελούνται από πολλαπλούς, διαδοχικούς ήχους, ενώ σε ορισμένες περιπτώσεις οι ήχοι αυτοί μπορεί να είναι επικαλυπτόμενοι. Αυτό δημιουργεί την ανάγκη για έναν ευέλικτο και ισχυρό μηχανισμό κατάτμησης ο οποίος α) θα υποδεικνύει έναν τύπο ήχου μόνο όταν το επίπεδο εμπιστοσύνης (confidence level) (δηλαδή το επίπεδο αναγνώρισης) υπερβαίνει ένα κατώφλι και β) θα προσδιορίζει τα χρονικά όρια των ήχων. Επιπλέον, λαμβάνοντας υπόψη ότι ήχοι οι οποίοι είναι πολύ περιορισμένης διάρκειας (π.χ. δεκάδες χιλιοστά του δευτερολέπτου) μπορεί να έχουν μικρή αξία για πρακτικές εφαρμογές, ο μηχανισμός θα πρέπει να χειρίζεται την ελάχιστη διάρκεια των αναγνωρισμένων ήχων. Επίσης, ήχοι του ίδιου τύπου συνήθως χωρίζονται από μικρές περιόδους θορύβου ή ησυχίας, π.χ. μεταξύ διαδοχικών γαβγισμάτων σκύλων. Τέτοιες περιπτώσεις είναι πιθανό να έχει περισσότερο νόημα να θεωρηθούν ως μία ενιαία συνεχής περίοδος παρά ως πολλαπλές διακριτές περίοδοι του ίδιου τύπου ήχου. Με αυτόν τον τρόπο μπορούν να συγχωνευτούν επαρκώς παρακείμενοι ήχοι του ίδιου τύπου. Για την ποσοτική προσέγγιση αυτών των λειτουργιών θεωρήθηκε ένα σύνολο παραμέτρων, όπως περιγράφεται στον Πίνακα 3.1.

Πίνακας 3.1: Παράμετροι για κατάτμηση και την αναγνώριση ήχων σε ηχητικές ροές

Παράμετρος	**Περιγραφή**	**Τιμή**
Ελάχιστη διάρκεια ήχου (MINDUR)	MINDUR, είναι η ελάχιστη διάρκεια ενός ήχου, ώστε αυτός να αναγνωρίζεται	0.5 (s)
Εμπιστοσύνη αναγνώρισης (IDCONF)	IDCONF, χρησιμοποιείται για να συμπεριλάβει ή να εξαιρέσει αναγνωρισμένους ήχους	[0, 1]
Ελάχιστη χρονική απόσταση (MINSEP)	MINSEP, είναι η ελάχιστη χρονική απόσταση μεταξύ διαδοχικών ήχων της ίδιας κλάσης	0.25 (s)
Κλάσεις ήχου που περιλαμβάνονται	Υποσύνολο ήχων της οντολογίας του AudioSet οι οποίοι περιλαμβάνονται	Υποσύνολο κλάσεων
Κλάσεις ήχου που εξαιρούνται	Υποσύνολο ήχων της οντολογίας του AudioSet οι οποίοι εξαιρούνται	Υποσύνολο κλάσεων
Επίπεδο Ειδικότητας (Specificity level)	Το επίπεδο στο δενδροδιάγραμμα του AudioSet στο οποίο βρίσκεται ο αναγνωρισμένος ήχος	0, 1, 2
Μήκος ηχητικού παραθύρου (WINLEN)	WINLEN, είναι το χρονικό μήκος του παραθύρου στον οποίο εφαρμόζεται ο αλγόριθμος	0.5 to 1.5 (s)
Μήκος άλματος (HOPLEN)	HOPLEN, είναι το μήκος του άλματος που επιτρέπει την κατάτμηση κάθε ηχητικού αποσπάσματος	80 to 250 (ms)

Η διαδικασία που εφαρμόζεται, όπως απεικονίζεται στο Σχήμα 3.3, είναι η ακόλουθη: τα αρχεία περνούν από προεπεξεργασία για την ομογενοποίηση της συχνότητας δειγματοληψίας και του επιπέδου κβαντισμού. Κάθε ηχητικό απόσπασμα (μήκους FILELEN) χωρίζεται σε πλαίσια σταθερού μήκους (WINLEN) με επικάλυψη και μήκος άλματος ίσο με HOPLEN. Τα μήκη παραθύρου επικάλυψης μπορούν να σχετίζονται με το αντίστροφο του ρυθμού δειγματοληψίας και μεταξύ τους (π.χ. το HOPLEN μπορεί να είναι 12.5% έως 50% του μήκους παραθύρου WINLEN). Ο αριθμός των πλαισίων N προσεγγίζεται ως εξής:

$$N = (FILELEN - WINLEN)/(HOPLEN) + 1 \quad (3.1)$$



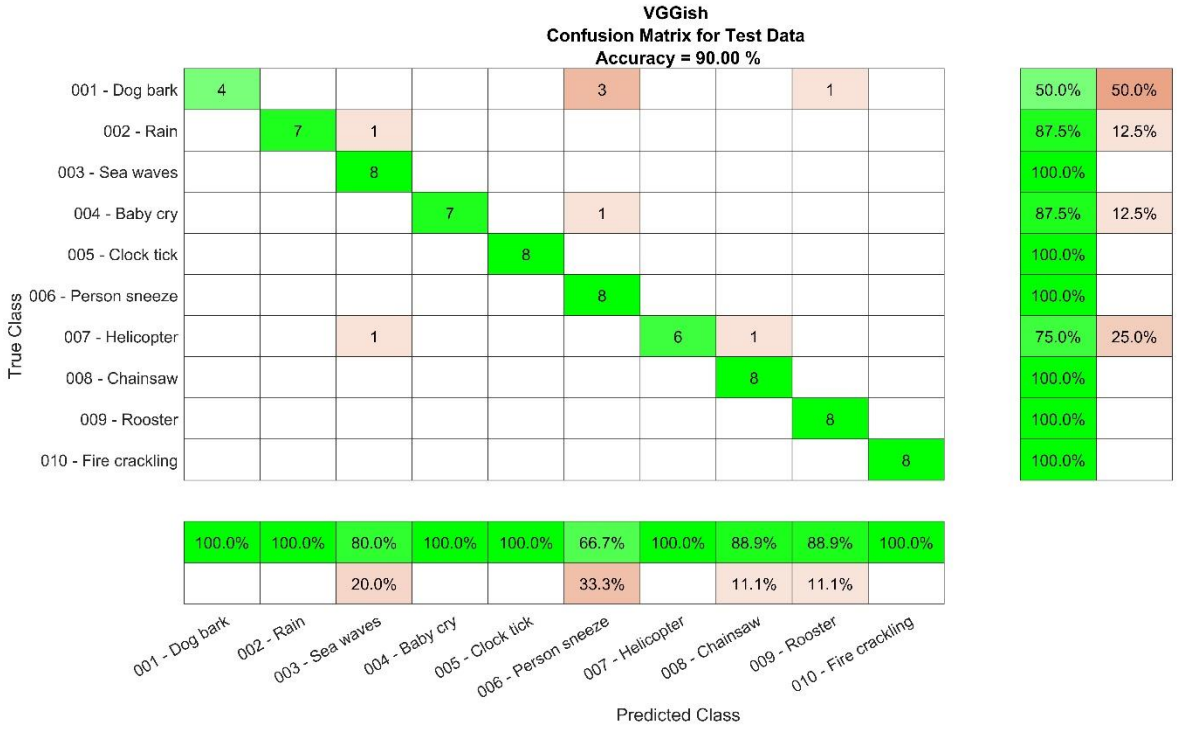
Σχήμα 3.3: Ροή εργασιών για την κατάτμηση ηχητικών ροών και την αναγνώριση του τύπου του ήχου.

Για κάθε πλαίσιο, υπολογίζεται το φασματογράφημα Mel με χρήση του STFT με αποτέλεσμα να προκύπτουν N στοιχεία. Τα N φασματογραφήματα Mel αναπαριστούν το αρχείο ήχου ως σύνολο σχημάτων και για κάθε ένα εκτιμάται ο τύπος του ήχου με την χρήση του ΣΝΔ. Καθώς αυτές οι εκτιμήσεις μπορεί να περιλαμβάνουν πολλαπλούς επικαλυπτόμενους τύπους ήχου, επιλέγεται η επικρατούσα ταυτοποίηση έτσι ώστε να υπάρχει μόνο μία ταυτοποίηση ανά βήμα.

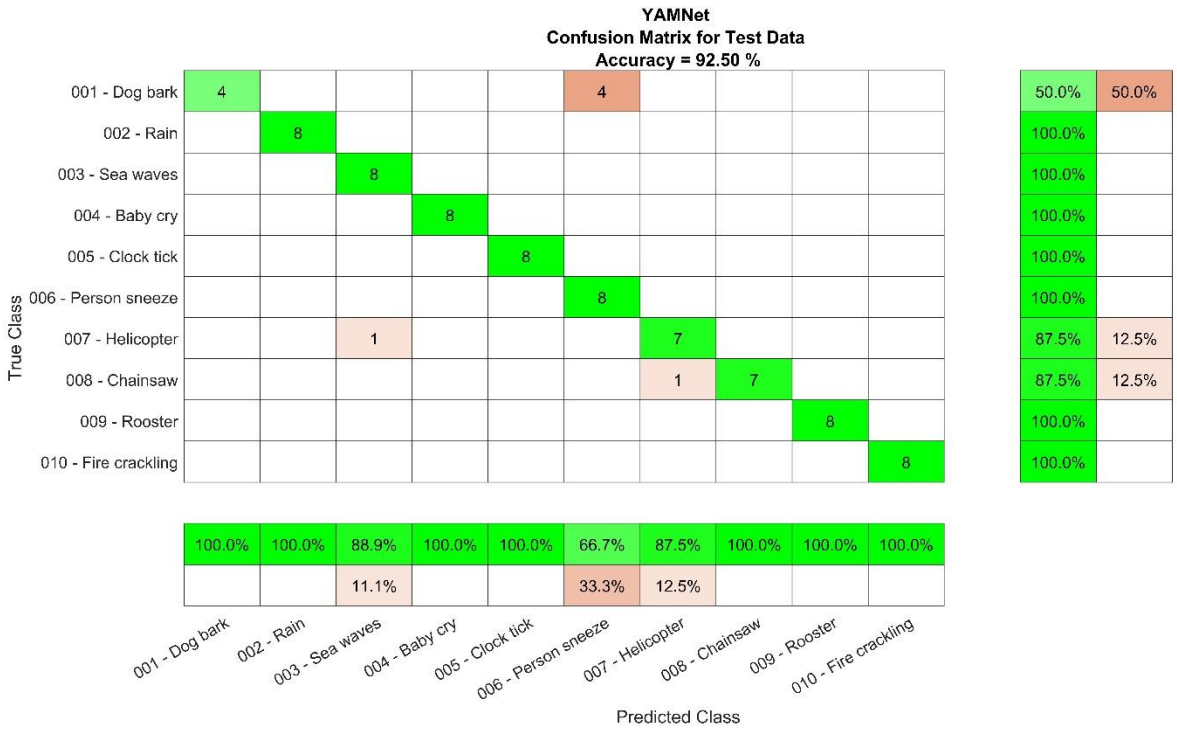
Με βάση το σύνολο των αναγνωρισμένων ήχων και των χρονικών ορίων τα αποτελέσματα φιλτράρονται περαιτέρω λαμβάνοντας υπόψη το επίπεδο αξιοπιστίας αναγνώρισης (μέσω σύγκρισης με το κατώφλι IDCONF). Εξετάζονται οι ήχοι που πρέπει να συμπεριληφθούν ή να αποκλειστούν. Εφαρμόζεται το μήκος της ελάχιστης διάρκειας (MINDUR) αφού μετατραπεί σε βήματα (hops) μέσω της διαίρεσής του με το HOPLEN και αφαιρούνται οι ταυτοποιήσεις ήχων με διάρκειες μικρότερες από την ελάχιστη αποδεκτή. Παράλληλα, αν το διάστημα μεταξύ δύο πανομοιότυπων ταυτοποιήσεων είναι μικρότερο ή ίσο με το MINSEP (το οποίο μεταφράζεται σε αριθμό βημάτων μέσω της διαίρεσής του με το HOPLEN), τα ηχητικά αποσπάσματα συγχωνεύονται.

3.1.6 Αποτελέσματα

Χρησιμοποιήθηκαν δύο ΣΝΔ Ήχου, το VGGish και το YAMNet. Το σύνολο δεδομένων χωρίστηκε σε σύνολο εκπαίδευσης, σύνολο επαλήθευσης και σύνολο ελέγχου με ποσοστά 60%, 20% και 20% αντίστοιχα του συνόλου των αρχείων. Η επανεκπαίδευση των δικτύων πραγματοποιήθηκε με την επιλογή των υπερπαραμέτρων σύμφωνα με την έρευνα στο [87], στην οποία έγινε αξιολόγηση πολλαπλών συνδυασμών των τιμών αυτών. Ως optimizer ορίστηκε ο Adam, το μέγεθος του mini-batch size ορίστηκε ίσο με 32, το learning rate ίσο με 0.5×10^{-4}, και ο μέγιστος αριθμός epochs ίσος με 10, και για τα δύο ΣΝΔ. Η ακρίβεια ταξινόμησης και οι λανθασμένες ταυτοποιήσεις ανά κλάση απεικονίζονται στους πίνακες σύγχυσης στα Σχήματα 3.4 και 3.5 για τα το VGGish και YAMNet αντίστοιχα.



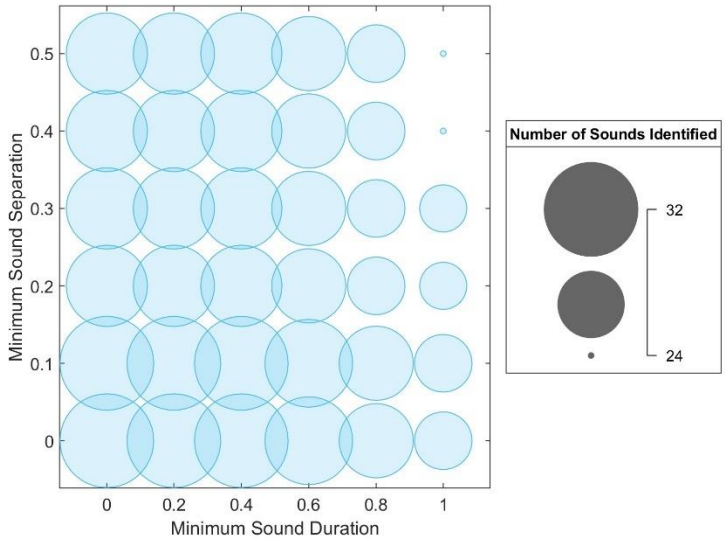
Σχήμα 3.4: Πίνακας Σύγχυσης της ταξινόμησης του ESC-10 με το VGGish.



Σχήμα 3.5: Πίνακας Σύγχυσης της ταξινόμησης του ESC-10 με το YAMNet.

Η ακρίβεια ταξινόμησης που επιτυγχάνει το YAMNet είναι υψηλότερη από την αντίστοιχη του VGGish (95% έναντι 90%), και με τα δύο όμως δίκτυα να επιδεικνύουν υψηλή απόδοση. Και στις δύο περιπτώσεις δικτύων, η κλάση στην οποία πραγματοποιούνται οι περισσότερες λανθασμένες ταξινομήσεις αφορά στο γάβγισμα του σκύλου.

Προκειμένου να επαληθευτεί η συμπεριφορά του αλγορίθμου ταξινόμησης και κατάτμησης, εξετάζονται τιμές για τις δύο βασικές παραμέτρους που σχετίζονται με την συγχώνευση και την κατάτμηση. Πρόκειται για την ελάχιστη διάρκεια ήχου (MINDUR) και της ελάχιστης χρονικής απόστασης (MINSEP). Οι παράμετροι παίρνουν τις τιμές {0, 0.2, 0.4, 0.6, 0.8, 1} και {0, 0.1, 0.2, 0.3, 0.4, 0.5} αντίστοιχα. Ο αλγόριθμος ταξινόμησης εκτελείται για καθέναν από τους 35 συνδυασμούς, με βάση μία ηχητική συνένωση αντιπροσωπευτική του συνόλου δεδομένων η οποία περιλαμβάνει δύο αποσπάσματα από κάθε κλάση. Στο Σχήμα 3.6 φαίνεται ο αριθμός των αναγνωρίσεων ήχου για κάθε έναν από τους συνδυασμούς, ανά διακριτές τιμές ελάχιστης διάρκειας στον οριζόντιο άξονα και ελάχιστου διαχωρισμού στον κατακόρυφο άξονα. Η ακτίνα των κύκλων είναι ανάλογη του αριθμού των ταυτοποιήσεων. Όταν αυτές οι παράμετροι παίρνουν την ελάχιστη τιμή τους (MINDUR = MINSEP = 0) η κατάτμηση του ήχου παίρνει την πιο λεπτομερή μορφή της και ο αριθμός των ταυτοποιήσεων είναι ίσος με 32. Ενώ, όταν οι παράμετροι παίρνουν τις μέγιστες τιμές τους, η κατάτμηση γίνεται ελάχιστα λεπτομερής, και ο αντίστοιχος αριθμός αναγνώρισης ήχου συγκλίνει στον αριθμό των συνυφασμένων διακριτών ηχητικών αποσπασμάτων, ίσος με 24. Αυτή η συμπεριφορά είναι ενδεικτική της σαφήνειας του συνόλου των δεδομένων, δηλαδή αν κάθε ηχητικό απόσπασμα αποτελείται από έναν ή πολλούς διαδοχικούς ή επικαλυπτόμενους ήχους.



Σχήμα 3.6: Αποτελέσματα της κατάτμησης και αναγνώρισης του ήχου για διαφορετικές τιμές των παραμέτρων της ελάχιστης διάρκειας και της ελάχιστης χρονικής απόστασης διαδοχικών ήχων.

Το σενάριο μπορεί να επεκταθεί αν θεωρηθούν αντιπροσωπευτικά αποσπάσματα ήχου για κάθε τύπο ήχου εντός του ίδιου ή διαφορετικού συνόλου δεδομένων. Για την περίπτωση του ESC-10 επαναλαμβάνεται το πείραμα για κάθε μία από τις κλάσεις, μετά από συνένωση 20 αρχείων που ανήκουν στην ίδια κλάση. Τα στατιστικά αποτελέσματα, όσον αφορά τον αριθμό των αναγνωρίσεων ήχου παρουσιάζονται στον Πίνακα 3.2. Αναφέρονται ο ελάχιστος, ο μέγιστος, ο μέσος όρος και η τυπική απόκλιση του αριθμού των ταυτοποιήσεων για τους συνδυασμούς των τιμών MINDUR και MINSEP.

Πίνακας 3.2: Ελάχιστος, μέγιστος, μέσος όρος και τυπική απόκλιση αριθμού αναγνωρίσεων ήχων για κάθε κλάση

Κλάση	**Ελάχιστος αριθμός αναγνωρίσεων**	**Μέγιστος αριθμός αναγνωρίσεων**	**Μέσος όρος αναγνωρίσεων**	**Τυπική απόκλιση αριθμού αναγνωρίσεων**
Γάβγισμα	14	18	16.8	1.1
Βροχή	10	22	17.7	3.8
Κύματα θάλασσας	16	23	20.8	1.9
Κλάμα μωρού	13	33	22.5	6.1
Ήχος ρολογιού	5	13	9	2.4
Φτέρνισμα	27	46	36	6.7
Ελικόπτερο	10	16	13.1	1.6
Αλυσοπρίονο	17	33	27.2	5.4
Κόκορας	35	37	36.3	0.9
Ήχος φωτιάς	13	28	22.6	4.9

Λαμβάνοντας υπόψη ότι για κάθε κλάση ο αριθμός των διακριτών εμφανίσεων αυτού του τύπου ήχου είναι 20 και ότι όλες οι κλάσεις αναγνωρίζονται σωστά από τον αλγόριθμο ταξινόμησης, ο ελάχιστος αριθμός αναγνωρίσεων μπορεί να αποτελέσει ένδειξη της αναγνωρισιμότητας της κλάσης. Σύμφωνα με τον Πίνακα 3.2 και την στήλη του μέσου αριθμού αναγνωρίσεων, ο κόκορας και το φτέρνισμα είναι οι πιο αναγνωρίσιμες κλάσεις, ενώ ο ήχος του ρολογιού και το ελικόπτερο οι λιγότερο αναγνωρίσιμες.

3.1.7 Συμπεράσματα

Τα σύνολα δεδομένων ήχου υποστηρίζουν εφαρμογές ΜΜ μέσω της εκπαίδευσης των αλγορίθμων και των ΣΝΔ. Τέτοια σύνολα δεδομένων μπορεί να είναι ετερογενή όσον αφορά τις τεχνικά χαρακτηριστικά τους (ρυθμός δειγματοληψίας, κβάντιση) και κυρίως όσον αφορά τις κατηγορίες ήχου που περιλαμβάνουν. Στην παρούσα έρευνα πραγματοποιήθηκε διερεύνηση της συσχέτισης μεταξύ των κλάσεων του ήχου τόσο από την σημασιολογική σκοπιά όσο και από την σκοπιά των τιμών των ηχητικών χαρακτηριστικών. Η σημασιολογική συσχέτιση επιτεύχθηκε με την βοήθεια της οντολογίας AudioSet. Επιπλέον, η σύγκριση των τύπων ήχου πραγματοποιήθηκε με χρήση των τιμών των χαρακτηριστικών που εξάγονται και της μεταξύ τους απόστασης. Οι κλάσεις του ήχου που παρουσιάζουν μεγαλύτερη απόσταση στις τιμές των χαρακτηριστικών τους είναι εύκολα διαχωρίσιμες, ενώ αυτές που παρουσιάζουν γειτνίαση των τιμών τους θα χρειαστούν βελτίωση των μεθόδων διαχωρισμού μεταξύ τους. Η εκ των προτέρων γνώση της απόστασης, είτε της σημασιολογικής είτε με βάση τις τιμές των χαρακτηριστικών, για κάθε πιθανό ζεύγος κλάσεων μπορεί να οδηγήσει στην εκλέπτυνση των μοντέλων ταξινόμησης.

Παράλληλα, αξιολογήθηκαν τεχνικές ταξινόμησης ήχου με την χρήση των ΣΝΔ VGGish και YAMNet, επιτυγχάνοντας υψηλή ακρίβεια ταξινόμησης (μεγαλύτερη από 90%) στο σύνολο δεδομένων ESC-10. Ο αλγόριθμος ταξινόμησης επεκτάθηκε από διακριτά ηχητικά αποσπάσματα μεμονωμένων κλάσεων σε πιο σύνθετα και ρεαλιστικά αρχεία ήχου μεγαλύτερης διάρκειας, τα οποία περιλαμβάνουν πολλές κλάσεις και χρειάζονται διαχωρισμό. Ορίστηκε ένα σύνολο λειτουργικών παραμέτρων, μεταξύ των οποίων η ελάχιστη διάρκεια ήχου και τα ηχητικά διαστήματα μεταξύ των ήχων, τα οποία καθορίζουν την διαδικασία της κατάτμησης. Η εφαρμογή του αλγορίθμου κατάτμησης με διαφορετικές τιμές των παραμέτρων επαλήθευσε την δυνατότητα προσαρμογής της κατάτμησης από χονδροειδή σε λεπτομερή, με αποτελέσματα μικρότερο και μεγαλύτερο αριθμό ταυτοποιήσεων ήχου αντίστοιχα.

Η συνένωση ηχητικών αρχείων ίδιας κλάσης και η εφαρμογή του αλγορίθμου κατάτμησης στο ενιαίο αρχείο μπορεί να οδηγήσει σε συμπεράσματα σχετικά με την αναγνωρισιμότητα της κάθε κλάσης και κατά συνέπεια στην εκλέπτυνση της μεθοδολογίας για τις κλάσεις που εμφανίζονται δυσκολότερα αναγνωρίσιμες.

Αυτή η έρευνα παρουσιάστηκε στο 26ο Πανελλήνιο Συνέδριο Πληροφορικής και περιέχεται στην δημοσίευση [103].

3.2 Αξιολόγηση και πρόβλεψη της χρήσης υπολογιστικών πόρων για την ταξινόμηση με μεθόδους ΒΜ

3.2.1 Εισαγωγή

Οι πρόσφατες εξελίξεις στον τομέα της πληροφορικής, που σχετίζονται με την εφαρμογή τεχνικών ΒΜ σε μία ευρεία ποικιλία τομέων (π.χ. τεχνητή όραση και επεξεργασία ομιλίας), και η ενσωμάτωση των τεχνολογιών IoT και του edge computing επιβάλλουν την ανάπτυξη οικονομικά αποδοτικής φιλοξενίας πολλαπλών εφαρμογών σε ένα ασφαλές, προσαρμόσιμο υπολογιστικό περιβάλλον. Αυτό οδηγεί σε υψηλότερη αξιοποίηση των πόρων με μειωμένο κόστος σε όλα τα επίπεδα [104]. Πιο αναλυτικά, οι αλγόριθμοι ΒΜ θεωρούνται ως οι πλέον σύγχρονες τεχνικές για διάφορες υπολογιστικές εργασίες που σχετίζονται είτε με την επεξεργασία εικόνας με υψηλό επίπεδο κατανόησης των απαιτήσεων σημασιολογίας, όπως η ταξινόμηση εικόνας, η κατάτμηση και ομαδοποίηση αντικειμένων, η ανίχνευση ανωμαλιών ή ακόμα και σε περιπτώσεις όπου απαιτούνται εργασίες επεξεργασίας εικόνας χαμηλού επιπέδου. Παράλληλα, δημιουργούνται σύνολα δεδομένων για την εκπαίδευση και την επικύρωση των αλγορίθμων μάθησης και οι εργασίες μπορούν να πραγματοποιηθούν με συγκεντρωτικό τρόπο ή/και με κατανεμημένο, με την συμμετοχή συσκευών και κόμβων διαφορετικών πόρων και δυνατοτήτων. Για την υποστήριξη ολοένα και πιο ποικίλων πεδίων εφαρμογής της, τα μοντέλα ΒΜ εξελίσσονται σε μεγαλύτερα με περισσότερα επίπεδα και παραμέτρους. Αυτό έχει ως αποτέλεσμα την αύξηση του χρόνου εκπαίδευσης και την ανάγκη για αξιοποίηση πόρων. Παρά τις προόδους στα σχήματα μάθησης, οι χρήστες εξακολουθούν να αντιμετωπίζουν ζητήματα σχετικά με τη βέλτιστη διαμόρφωση των ρυθμίσεων εκτέλεσης της ΒΜ, των μηχανισμών που σχετίζονται με την κατανομή των υπολογιστικών πόρων και τη σχέση τους με τον χρόνο εκπαίδευσης και την ακρίβεια ταξινόμησης.

Οι τεχνικές διαχείρισης πόρων παίζουν αξιοσημείωτο ρόλο για όλους τους τύπους υπολογιστών και σχετίζονται με τις απαιτήσεις υψηλής εικονικοποίησης (virtualization), επεκτασιμότητας και διαφάνειας. Η παρακολούθηση των πόρων αποτελεί μέρος της διαχείρισης των πόρων για ένα σύγχρονο υπολογιστικό περιβάλλον όπου εφαρμόζεται η ΒΜ, η οποία παρέχει καλύτερη κατανόηση για την κατανομή των πόρων η οποία εξυπηρετεί την αύξηση της αποτελεσματικότητας των υπηρεσιών, όπως: προγραμματισμός εργασιών, προγραμματισμός χωρητικότητας, προληπτική κλιμάκωση και εξισορρόπηση φορτίου εργασίας, βελτίωση της απόδοσης των υπολογιστών και του δικτύου. Προϋπόθεση για την αποτελεσματική διαχείριση πόρων είναι η εκ των προτέρων εκτίμηση των απαιτούμενων πόρων.

Αυτή η εργασία δημιουργεί αρκετές προκλήσεις, καθώς δεν υπάρχει μηχανισμός συσχέτισης μεταξύ των διαδικασιών μάθησης (εκπαίδευση και εξαγωγή συμπερασμάτων), του όγκου των εμπλεκόμενων δικτύων και των απαιτούμενων πόρων με όρους επεξεργαστή και μνήμης. Προς την κατεύθυνση αυτή απαιτείται ακριβής και απλή εκτίμηση της χρήσης των πόρων, λαμβάνοντας υπόψη τον δυναμικό και χρονικά μεταβαλλόμενο φόρτο εργασίας των σύγχρονων υπολογιστικών περιβαλλόντων υπό διαφορετικούς περιορισμούς. Επιπλέον, οι σύγχρονες υπολογιστικές συσκευές φαίνονται ευέλικτες στην διαχείριση διαφορετικού φόρτου εργασίας σε περιβάλλον περιορισμένων πόρων [105]. Παρά τις νέες προσεγγίσεις σχεδιασμού σε αυτόν τον τομέα, η αξιολόγηση της κατανομής των πόρων και η παρακολούθηση της απόδοσης για εφαρμογές ΒΜ παραμένει ένα ενδιαφέρον ερευνητικό ερώτημα, ιδίως όταν τα μοντέλα ΒΜ πρέπει να αναπτυχθούν σε περιβάλλον περιορισμένων πόρων, όπως π.χ. στις κινητές συσκευές, στους αισθητήρες IoT, όπου απαιτείται η βέλτιστη διαμόρφωση των ρυθμίσεων εκτέλεσης.

3.2.2 Στόχος

Στην συγκεκριμένη έρευνα ο στόχος είναι η πραγματοποίηση μελέτης σχετικά με τους απαιτούμενους υπολογιστικούς πόρους αναλύοντας διαφορετικά σενάρια συστημάτων ΒΜ. Το χαρτοφυλάκιο των διαθέσιμων νευρωνικών δικτύων είναι ευρύ, αποτελούμενο από συμπαγή έως μεγάλα δίκτυα, με πολλαπλά επίπεδα, παραμέτρους και πολυπλοκότητα. Αυτή η ετερογένεια στην τοπολογία των δικτύων αντανακλάται, όχι απαραίτητα γραμμικά, στους απαιτούμενους υπολογιστικούς πόρους για την εκπαίδευσή τους και την εξαγωγή αποφάσεων. Από αυτή την άποψη, οι απαιτήσεις υπολογιστικών πόρων μπορούν να αποτελέσουν ένα από τα κριτήρια για την διαχείριση των πόρων και την επιλογή των νευρωνικών δικτύων.

Πιο συγκεκριμένα, επιδιώκεται μία πολύ-παραμετρική αξιολόγηση της αξιοποίησης των πόρων. Οι παράμετροι που διαφοροποιούν τα πειράματα είναι οι διαμορφώσεις των ΣΝΔ, οι τιμές των υπερπαραμέτρων εκπαίδευσης και τα δεδομένα εισόδου. Οι διαφοροποιήσεις αυτές σχετίζονται με τους υπολογιστικούς πόρους αλλά έχουν κοινό παρονομαστή σύγκρισης τον χρόνο εκπαίδευσης που απαιτείται.

Η συγκεκριμένη μελέτη αποσκοπεί στην εκ των προτέρων γνώση σχετικά με τις απαιτήσεις των πόρων για διάφορα σενάρια ταξινόμησης ήχου και εικόνας.

3.2.3 Συναφής έρευνα

Μία λεπτομερής επισκόπηση στον τομέα της αποτελεσματικής εξαγωγής συμπερασμάτων παρέχεται στο [106], όπου προσδιορίζονται και αναλύονται οι κύριες ερευνητικές προκλήσεις. Σε αυτή την εργασία, μελετάται η σημασία των περιορισμών στους πόρων και στην παρακολούθηση στο περιβάλλον του edge computing. Επιπλέον, παρουσιάζεται μία αντιπροσωπευτική ανάλυση έρευνας για μοντέλα ΒΜ η οποία παρέχει επίσης τις κατάλληλες μετρικές για την αξιολόγηση της απόδοσης. Μία ολοκληρωμένη ανάλυση των αρχιτεκτονικών βαθιών νευρωνικών δικτύων δίνεται στο [107]. Στην συγκεκριμένη εργασία περιγράφονται και αξιολογούνται διάφορες αρχιτεκτονικές νευρωνικών δικτύων με βάση πολλαπλούς δείκτες επιδόσεων, όπως η ακρίβεια αναγνώρισης, η πολυπλοκότητα του μοντέλου, η υπολογιστική πολυπλοκότητα, η χρήση μνήμης και ο χρόνος εξαγωγής αποτελεσμάτων. Η μελέτη αυτή παρέχει κατανόηση του αντίκτυπου των περιορισμών των πόρων, όταν αυτές οι αρχιτεκτονικές βρίσκονται στην διαδικασία πρακτικής ανάπτυξης. Πιο αναλυτικά, 44 διαφορετικά μοντέλα νευρωνικών δικτύων δοκιμάζονται και αξιολογούνται σε περιβάλλον περιορισμένων πόρων, δίνοντας πειραματικά αποτελέσματα της δοκιμής των δικτύων.

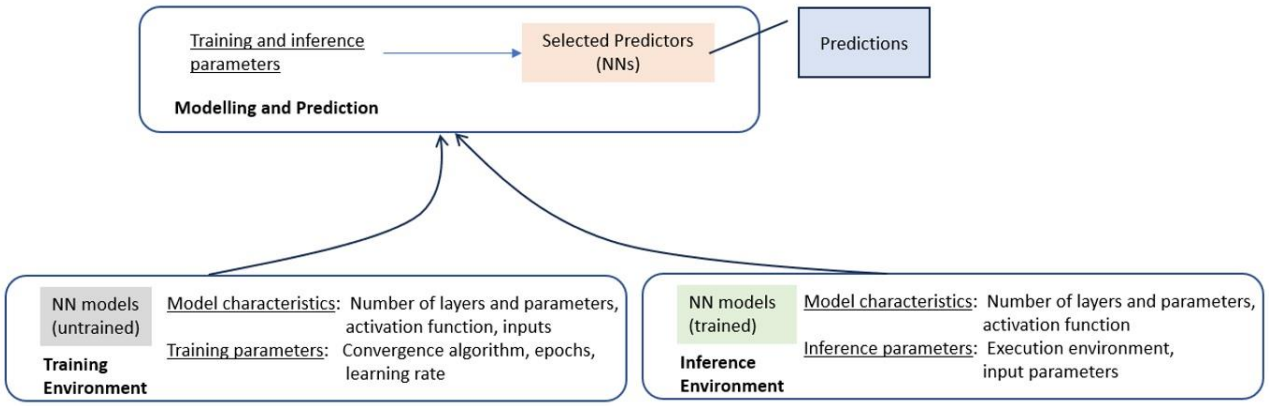
Στο [108] πραγματοποιείται ανάλυση της εφαρμογής προ-εκπαιδευμένων νευρωνικών δικτύων. Παρουσιάζεται η σημασία και η χρησιμότητα των προ-εκπαιδευμένων μοντέλων, αναδεικνύοντας την προσέγγιση της μεταφοράς μάθησης. Στην περίπτωση αυτή, παρουσιάζονται τα πλεονεκτήματα των προ-εκπαιδευμένων μοντέλων όσον αφορά τις υπολογιστικές απαιτήσεις και τους πόρους, επιτρέποντας ταχύτερη εξαγωγή συμπερασμάτων και μειώνοντας την πολυπλοκότητα.

Η εκτίμηση της χρήσης των πόρων και η ακριβής χρήση του κατάλληλου μοντέλου έχουν διερευνηθεί σε διάφορα πλαίσια. Στο [109], ένας ταξινομητής εκπαιδεύεται με τιμές χρήσης πόρων έτσι ώστε να διαμορφωθεί ένα μοντέλο πρόβλεψης χρήσης πόρων εντός συγκεκριμένων χρονικών διαστημάτων. Προς την κατεύθυνση αυτή, εφαρμόζονται διάφορες μέθοδοι πρόβλεψης προκειμένου να βελτιωθεί η ακρίβεια της πρόβλεψης των υπολογιστικών πόρων, δεδομένου ότι η πρόβλεψη των απαιτήσεων πόρων, όπως π.χ. η χρήση του επεξεργαστή, είναι ένα πολύπλοκο έργο το οποίο εξαρτάται από τους εισερχόμενες εργασίες. Στο [110], προτείνεται ένα μοντέλα ΒΜ και αξιολογείται σύμφωνα με την ακρίβεια πρόβλεψης και την ελαχιστοποίηση σφάλματος. Παρόμοιες μελέτες έχουν παρουσιαστεί με στόχο την πρόβλεψη των πόρων και την υποστήριξη του καταλληλότερου δικτύου με βάση τεχνικές

αξιολόγησης οι οποίες αναδεικνύουν την σημαντικότητα της εκ των προτέρων εκτίμησης απαραίτητων πόρων, όπως στο [111] όπου προτείνεται μία νέα στρατηγική. Πιο αναλυτικά, τα νευρωνικά δίκτυα γραφημάτων επιστρατεύονται για να προβλέψουν την κατανάλωση πόρων σε διαφορετικές περιπτώσεις φόρτου εργασίας, και η μεταφορά μάθησης μπορεί να αξιοποιηθεί περαιτέρω προκειμένου να επεκταθούν αυτά τα δίκτυα και να προσαρμοστούν σε διαφορετικά υπολογιστικά περιβάλλοντα. Επιπλέον, τεχνικές βελτιστοποίησης εφαρμόζονται στο [112], όπου εφαρμόζεται προσαρμοστική επιλογή μοντέλου με βάση την ΜΜ για την ανάπτυξη ενός χαμηλού κόστους μοντέλου πρόβλεψης, προκειμένου να επιλέγεται άμεσα εάν προ εκπαιδευμένο μοντέλο με βάση την απαιτούμενη ακρίβεια ταξινόμησης και τον χρόνο εξαγωγής αποτελεσμάτων.

3.2.4 Μεθοδολογία

Στην μελέτη αυτή παρατηρείται και μετριέται ο υπολογιστικός χρόνος για την εκπαίδευση και εξαγωγή αποτελεσμάτων ενός επιλεγμένου συνόλου ΣΝΔ με διαφορετικούς συνδυασμούς τιμών των υπερπαραμέτρων εκπαίδευσης και διαμορφώσεις. Τα συγκεκριμένα δίκτυα εκπαιδεύονται για την ταξινόμηση ήχου. Τα αρχεία του ήχου αφού μετατραπούν σε σήματα εικόνας εκπαιδεύουν τα δίκτυα. Όπως φαίνεται στο Σχήμα 3.7, οι δραστηριότητες εκπαίδευσης και εξαγωγής συμπερασμάτων (κατώτερο επίπεδο) εξαρτώνται από το υπολογιστικό περιβάλλον, τα εγγενή χαρακτηριστικά του δικτύου και τις τιμές των παραμέτρων εκπαίδευσης. Η διαμόρφωση και οι μετρήσεις που ανακτώνται από την εκπαίδευση και την εξαγωγή συμπερασμάτων παρέχονται για την μοντελοποίηση και πρόβλεψη του υπολογιστικού χρόνου (ανώτερο επίπεδο). Σε αυτό το επίπεδο ένα σύνολο μοντέλων εκπαιδεύεται με βάση αυτές τις παραμέτρους και τις αντίστοιχες μετρήσεις υπολογιστικού χρόνου ώστε να μπορέσουν να παρέχουν εκτιμήσεις και προβλέψεις (έξοδοι). Οι προβλέψεις αυτές αποτελούν τις εκτιμήσεις των απαιτούμενων πόρων για τις δραστηριότητες εκπαίδευσης και εξαγωγής αποτελεσμάτων για ένα αντιπροσωπευτικό σύνολο ΣΝΔ και μπορούν να υποστηρίξουν την αποτελεσματική διαχείριση των υπολογιστικών πόρων καθώς και την επιλογή των ΣΝΔ, λαμβάνοντας υπόψη και τις αντίστοιχες ακρίβειες ταξινόμησης.



Σχήμα 3.7: Πρόβλεψη απαιτούμενου υπολογιστικού χρόνου για την εκπαίδευση και εξαγωγή αποτελεσμάτων.

Τα ΣΝΔ που έχουν επιλεγεί για την εκπαίδευση και ταξινόμηση είναι τα GoogleNet, SqueezeNet, ShuffleNet, VGGish και YAMNet. Αυτά τα δίκτυα είναι γνωστά και χρησιμοποιούνται ευρέως σε ένα πλούσιο σύνολο εφαρμογών. Χρησιμοποιούνται για την αναγνώριση και ταξινόμηση εικόνων, με τις εικόνες αυτές να είναι οι εικονικές αναπαραστάσεις ηχητικών σημάτων. Είναι επαρκώς πολύπλοκα καθώς ο αριθμός των επιπέδων τους κυμαίνεται από 9 έως 50, επιτρέποντας την αποτελεσματική παραμετροποίηση των διαδικασιών εκπαίδευσής τους. Τα χαρακτηριστικά των δικτύων περιλαμβάνουν την αρχιτεκτονική και τον σχεδιασμό τους, τον αριθμό των επιπέδων τους, τον αριθμό των παραμέτρων, την συνάρτηση ενεργοποίησης και το μέγεθος που καταλαμβάνουν στην μνήμη. Η αναλυτική περιγραφή των συγκεκριμένων δικτύων έχει γίνει αναλυτικά στην Παράγραφο 2.2.1.

Η εκπαίδευση ενός νευρωνικού δικτύου είναι μία διαδικασία μεγάλου υπολογιστικού φόρτου η οποία μπορεί να απαιτεί εκτεταμένη χρονική διάρκεια. Μία λειτουργική επιλογή είναι η εφαρμογή της μεταφοράς μάθησης, δηλαδή η μεταφορά της γνώσης που έχει αποκτηθεί από έναν τομέα προέλευσης να μεταφέρεται σε άλλη εφαρμογή προορισμού. Πρακτικά, αυτό σημαίνει ότι χρησιμοποιείται το εκπαιδευμένο δίκτυο και ένα μέρος του (ή μέρη του) επανεκπαιδεύεται με βάση τα δεδομένα προορισμού. Εκτός από την μείωση των απαιτήσεων σε υπολογιστικό χρόνο και πόρους, ο μηχανισμός της μεταφοράς μάθησης αντιμετωπίζει επίσης την πιθανότητα μειωμένου συνόλου δεδομένων στον τομέα προορισμού. Η επέκταση τμήματος του δικτύου συνήθως αφορά μόνο το τμήμα του ταξινομητή. Σε ορισμένες περιπτώσεις το συνελικτικό τμήμα του δικτύου μπορεί να επανεκπαιδευτεί (με προσθήκη, αφαίρεση ή προσαρμογή των επιπέδων) μαζί με το επίπεδο ταξινόμησης. Σε σπάνιες περιπτώσεις, το πλήρες δίκτυο εκπαιδεύεται from the scratch, λαμβάνοντας υπόψη τα σύνολα δεδομένων του πεδίου ορισμού.

Η επιλογή εξαρτάται από κάθε συγκεκριμένο πρόβλημα καθώς και από την συγγένεια των περιοχών προέλευσης και προορισμού. Οι πιο τυπικές περιπτώσεις περιλαμβάνουν την προσαρμογή (επανεκπαίδευση) των τμημάτων ταξινόμησης των δικτύων, διατηρώντας την συνολική αρχιτεκτονική και τον αριθμό των επιπέδων. Από αυτή την άποψη, οι μετρήσεις που έγιναν σε αυτή τη μελέτη αναφέρονται στην διάρκεια επανεκπαίδευσης (χρόνος επεξεργαστή) του τμήματος ταξινόμησης των δικτύων που αναφέρθηκαν. Ο χρόνος αυτός επηρεάζεται από τις τιμές των αντίστοιχων παραμέτρων εκπαίδευσης (optimizer, mini-batch size, epochs, και learning rate).

Επίσης, οι παράμετροι του συνόλου δεδομένων εισόδου παίζουν επίσης ρόλο. Αυτές αφορούν τον αριθμό των αρχείων που χρησιμοποιούνται για την εκπαίδευση, την ανάλυση και την μορφή τους (π.χ. jpg ή tiff). Το μέγεθος των δεδομένων εισόδου καθορίζεται από την αρχιτεκτονική των ίδιων των δικτύων.

Χρησιμοποιήθηκαν τα σύνολα δεδομένων UrbanSound8K, ESC-10 και Air Compressor των οποίων ο αριθμός των αρχείων και οι κλάσεις μαζί με τις υπόλοιπες παραμέτρους των δικτύων περιγράφονται στον Πίνακα 3.3.

Πίνακας 3.3: Παράμετροι δικτύων, εκπαίδευσης και συνόλων δεδομένων

Παράμετροι	**Τιμές**				
ΣΝΔ	**GogLeNet**	**SqueezeNet**	**ShuffleNet**	**VGGish**	**YAMNet**
# Επίπεδα	22	18	50	9	28
Παράμετροι (εκατομμύρια)	7	1.24	1.4	72.1	3.75
Μέγεθος στην μνήμη (MB)	27	5.2	5.4	289	15.5
Ανάλυση εικόνων εισόδου	224x224x3	227x227x3	224x224x3	96x64x1	96x64x1
Τιμές παραμέτρων εκπαίδευσης	**Optimizer**	**Mini-batch size**	**Epochs**	**Learning rate**	
	Adam, SGDM	8, 16, 32 (Image CNNs)	6, 8, 10	0.5, 1, 2 ($x10^{-3}$)	
		64, 128, 256 (Sound CNNs)			
Σύνολα δεδομένων	**UrbanSound8K**	**ESC-10**	**Air Compressor**		
# Αρχεία	8732	400	1800		
# Κλάσεις	10	10	8		

3.2.5 Πρόβλεψη με βάση την παλινδρόμηση

Το σύνολο δεδομένων περιλαμβάνει τις μετρήσεις του χρόνου επεξεργαστή για τους συνδυασμούς που προκύπτουν σύμφωνα με τις τιμές των παραμέτρων του Πίνακα 3.3. Οι συνδυασμοί αυτοί είναι:

$[\#\Sigma N\Delta] \times [\#\Sigma υνόλων\ δεδομένων] \times [\#Optimizers] \times [\#Mini-batch\ sizes] \times [\#Epochs] \times [\#Learning\ rates]$

οι οποίοι οδήγησαν σε 504 συνδυασμούς και μετρήσεις

Ο χρόνος του επεξεργαστή μετριέται χρησιμοποιώντας το περιβάλλον στο οποίο υλοποιήθηκαν οι αλγόριθμοι εκπαίδευσης και ελέγχου το οποίο είναι το Matlab 2023b. Εκτός από τους χρόνους εκπαίδευσης, μετρήθηκαν και οι αντίστοιχοι χρόνοι για την εξαγωγή συμπερασμάτων για ένα υποσύνολο των εκπαιδευμένων δικτύων (ένα από κάθε αρχιτεκτονική) και διερευνήθηκε η σχέση (συσχέτιση) με τους χρόνους εκπαίδευσης. Όλα τα σενάρια εκπαίδευσης και ελέγχου πραγματοποιήθηκαν με την ίδια υπολογιστική υποδομή και εφαρμόστηκε κανονικοποίηση. Συγκεκριμένα, χρησιμοποιήθηκε επιτραπέζιος υπολογιστής με μνήμη 32GB RAM, επεξεργαστή Intel Core i7-10700K, οκτώ πυρήνων έως 3.8GHz, με την κάρτα γραφικών NVIDIA GeForce RTX 3060.

Πριν από την εκπαίδευση πραγματοποιήθηκε μία διαδικασία επιλογής μεταξύ των διαθέσιμων μοντέλων παλινδρόμησης. Η διαδικασία περιλαμβάνει την εκπαίδευση των υποψήφιων μοντέλων, συμπεριλαμβανόμενης της γραμμικής παλινδρόμησης, των μηχανισμών διανυσμάτων υποστήριξης (SVM), των δέντρων παλινδρόμησης και συνδυασμού αυτών, της προσέγγισης kernel καθώς και των νευρωνικών δικτύων. Η επιλογή βασίστηκε στην ελαχιστοποίηση του σφάλματος επικύρωσης, το οποίο υπολογίζεται ως το μέσο τετραγωνικό σφάλμα ρίζας (Root Mean Square Error-RMSE).

Όλα τα επιλεγμένα μοντέλα βασίζονται στα νευρωνικά δίκτυα και συγκεκριμένα περιλαμβάνουν το Narrow, το Medium, το Wide με ένα επίπεδο, τα Bilayered και Trilayered με 2 και 3 επίπεδα αντίστοιχα. Όλα τα δίκτυα χρησιμοποιούν την συνάρτηση ενεργοποίησης ReLu.

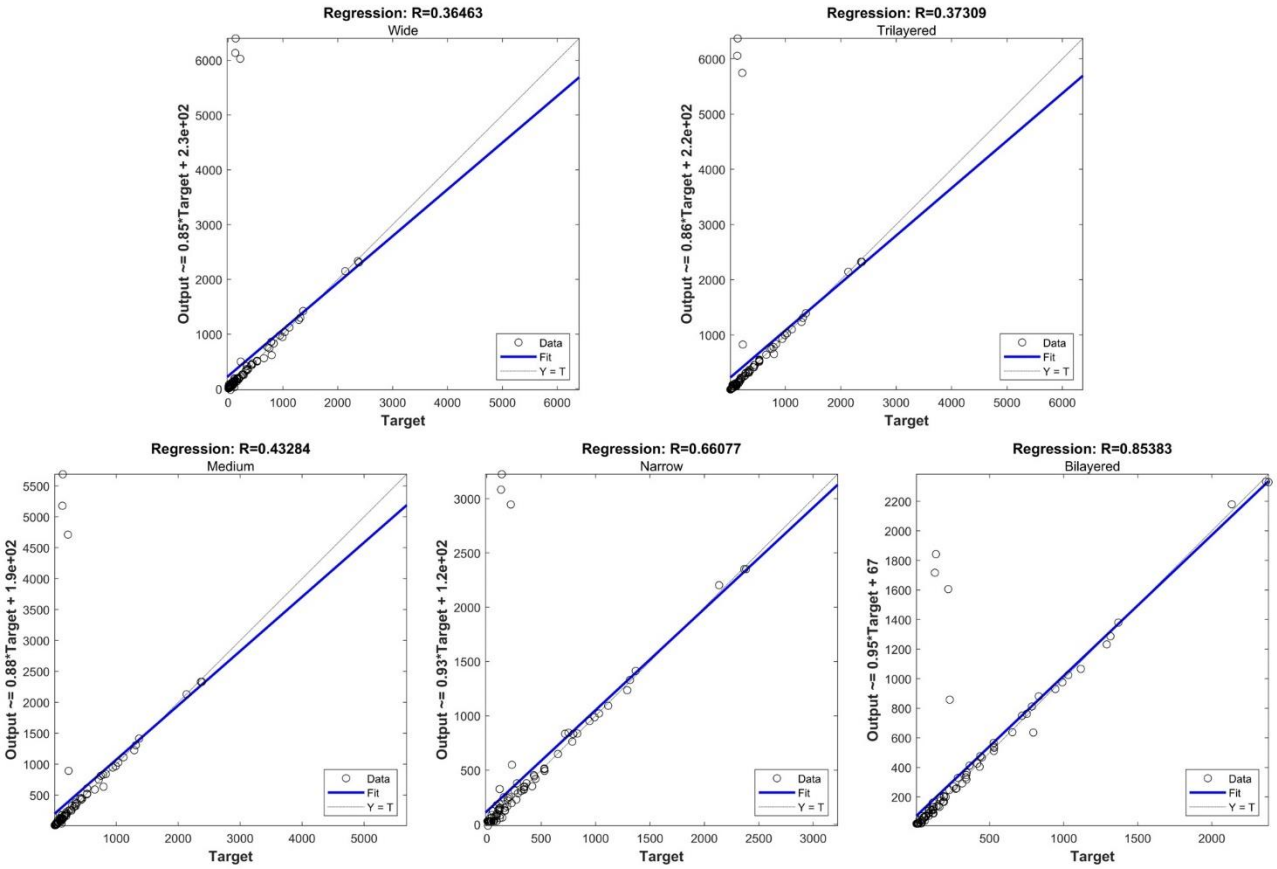
Για την εκπαίδευση και τον έλεγχο χρησιμοποιήθηκαν το 80% και 20% αντίστοιχα του συνόλου δεδομένων και τα αποτελέσματα της πρόβλεψης για τους χρόνους εκπαίδευσης του επεξεργαστή παρουσιάζονται στο Σχήμα 3.8. Στο σχήμα αυτό φαίνονται τα πέντε επιμέρους σχήματα τα οποία παρουσιάζουν τα αποτελέσματα καθενός από τα πέντε νευρωνικά δίκτυα πρόβλεψης.

Η αξιολόγηση των διαθέσιμων μοντέλων πραγματοποιήθηκε λαμβάνοντας υπόψη το $RMSE$ και τον συντελεστή προσδιορισμού (coefficient of determination), R^2. Οι τύποι υπολογισμού παρατίθενται παρακάτω:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{load(\iota)} - load(i))^2} \tag{3.2}$$

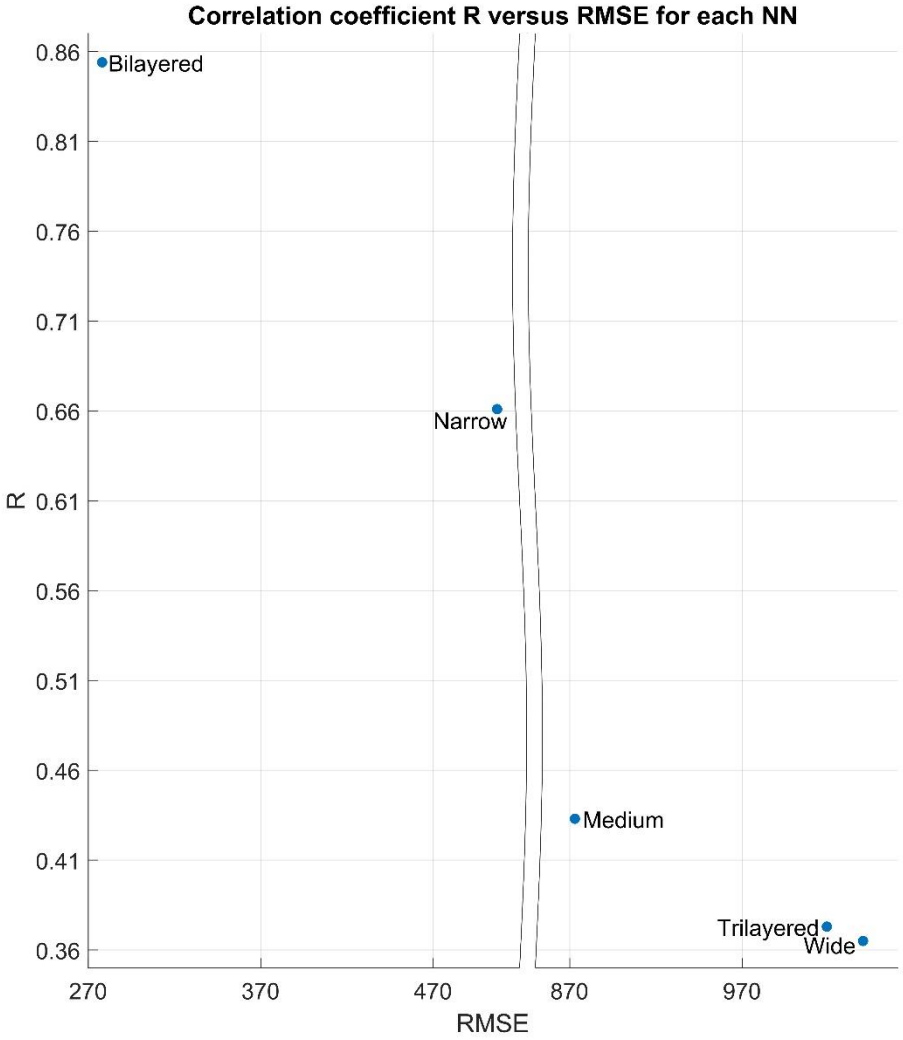
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{N}(\widehat{load(\iota)} - load(i))^2}{\sum_{i=1}^{N}(load(i)) - \frac{1}{N}\sum_{i=1}^{N}(load(i))} \tag{3.3}$$

Όπου, $\widehat{load(\iota)}$ είναι η προβλεπόμενη τιμή και $load(i)$ η πραγματική τιμή. Επίσης, SST είναι το άθροισμα των τετραγώνων συνολικά (sum of squares total), SSR είναι το άθροισμα των τετραγώνων παλινδρόμησης (sum of squares regression) και SSE το άθροισμα των τετραγώνων σφάλματος (sum of squares error).



Σχήμα 3.8: Αποτελέσματα πρόβλεψης.

Οι συντελεστές συσχέτισης R σε σχέση με το $RMSE$ για κάθε νευρωνικό δίκτυο παρουσιάζονται στο Σχήμα 3.9, στον κατακόρυφο και οριζόντιο άξονα αντίστοιχα. Όπως φαίνεται στο Σχήμα 3.9, το μοντέλο που παρέχει την ακριβέστερη πρόβλεψη είναι το δίκτυο των δύο επιπέδων (Bilayered), το οποίο επιτυγχάνει την μεγαλύτερη τιμή R $(= 0.85)$ και την μικρότερη τιμή $RMSE$ $(= 279)$. Ακολουθούν το Narrow δίκτυο και το Medium δίκτυο. Το δίκτυο των τριών επιπέδων (Trilayered) και το Wide παρουσιάζουν τις πιο αδύναμες επιδόσεις.



Σχήμα 3.9: Ο συντελεστής συσχέτισης R σε σχέση με το μέσο τετραγωνικό σφάλμα ρίζας RMSE.

3.2.6 Χρόνος εξαγωγής συμπερασμάτων

Μετρήθηκε ο χρόνος εξαγωγής αποτελεσμάτων (inference time) για ένα σύνολο των εκπαιδευμένων δικτύων, ένα ανά αρχιτεκτονική, και σε όλα τα δίκτυα οι διαμορφώσεις εκπαίδευσης ρυθμίστηκαν στις πιο τυπικές τιμές. Συγκεκριμένα, οι τιμές των υπερπαραμέτρων που ορίστηκαν για όλα τα ΣΝΔ ήταν, ο Adam optimizer, ο αριθμός των epochs ίσος με 8, το learning rate ίσο με 2×10^{-3}. Για τα ΣΝΔ GoogleNet, SqueezeNet και ShuffleNet το mini-batch size ορίστηκε ίσο με 32, ενώ για τα ΣΝΔ VGGish και YAMNet ίσο με 256, καθώς στην φάση της προεπεξεργασίας κάθε αρχείο χωρίζεται κατά 0.96. Για την εξαγωγή των αποτελεσμάτων χρησιμοποιήθηκε το 20% των αρχείων του UrbanSound8K, το οποίο αντιστοιχεί σε 1747 αρχεία ήχου.

Οι χρόνοι εκπαίδευσης και εξαγωγής αποτελεσμάτων φαίνονται στον Πίνακα 3.4.

Πίνακας 3.4: Οι χρόνοι εκπαίδευσης και εξαγωγής αποτελεσμάτων για κάθε ΣΝΔ

CNN	**Χρόνος εκπαίδευσης (s)**	**Χρόνος εξαγωγής αποτελεσμάτων (ms)**
GoogleNet	342	2.04
SqueezeNet	154	1.28
ShuffleNet	1063	2.00
VGGish	373	2.94
YAMNet	590	26.67

Το ShuffleNet παρουσιάζει τον μεγαλύτερο χρόνο εκπαίδευσης, γεγονός το οποίο εξηγείται εν μέρει από την αρχιτεκτονική του, ενώ ο χρόνος εξαγωγής αποτελεσμάτων είναι ο μικρότερος. Το YAMNet εμφανίζει σχετικά μεγάλο χρόνο εκπαίδευσης και τον μεγαλύτερο χρόνο εξαγωγής αποτελεσμάτων. Το SqueezeNet έχει την μικρότερη διάρκεια εξαγωγής αποτελεσμάτων και αυτό το καθιστά το πιο κατάλληλο για περιβάλλοντα με πιο περιορισμένους υπολογιστικούς πόρους.

Μία ακόμη ενδιαφέρουσα παρατήρηση είναι ότι οι διάρκειες εξαγωγής αποτελεσμάτων και οι αντίστοιχες διάρκειες εκπαίδευσης δεν συσχετίζονται (ο δείκτης συσχέτισης είναι ίσος με 0.15).

Ενώ σε μία συγκεκριμένη διάταξη η εκπαίδευση και η εξαγωγή αποτελεσμάτων μπορεί να πραγματοποιούνται ανεξάρτητα, ακόμα και με την χρήση διαφορετικών περιβαλλόντων (εκπαίδευση στον κεντρικό κόμβο και εξαγωγή αποτελεσμάτων στους περιφερειακούς κόμβους), σε πιο κατανεμημένες και αποκεντρωμένες ρυθμίσεις και οι δύο λειτουργίες πραγματοποιούνται στον ίδιο κόμβο και με διαφορετική συχνότητα. Από αυτή την άποψη, μπορεί να θεωρηθεί ένας γραμμικός συνδυασμός των δύο διαρκειών και να σχηματιστεί ένας δείκτης, όπως π.χ.:

$$I_{train,inf} = a_1 T_{tr} + a_2 T_{inf} \tag{3.4}$$

Οι συντελεστές a_1 και a_2 μπορούν να προσεγγιστούν μέσω της εκτίμησης της σχετικής συχνότητας της εκπαίδευσης και συμπερασματολογίας αντίστοιχα.

3.2.6 Συμπεράσματα

Η εκπαίδευση και η επανεκπαίδευση των ΣΝΔ μπορεί να αποτελέσει πρόκληση από την άποψη των υπολογιστικών πόρων, ενώ η εξαγωγή αποτελεσμάτων απαιτεί επίσης ένα μέρος των πόρων αυτών. Δεδομένου ενός συγκεκριμένου υπολογιστικού περιβάλλοντος, οι απαιτήσεις σε πόρους σε συνδυασμό με τις επιδόσεις και την ακρίβεια που επιτυγχάνονται, μπορούν να αποτελέσουν ένα κριτήριο για την επιλογή των μοντέλων ΒΜ. Για τον σκοπό αυτό απαιτούνται μηχανισμοί και εργαλεία για την εκτίμηση των απαιτούμενων πόρων.

Στην εργασία αυτή εξετάστηκε ένα σύνολο πέντε ΣΝΔ τα οποία αποτελούν ένα αντιπροσωπευτικό δείγμα αρχιτεκτονικής και πολυπλοκότητας. Τα μοντέλα αυτά επανεκπαιδεύτηκαν με διαφορετικές διαμορφώσεις και για κάθε συνδυασμό μετρήθηκε ο χρόνος του επεξεργαστή για την εκπαίδευση τους. Το αποτέλεσμα ήταν η δημιουργία ενός συνόλου δεδομένων με περισσότερες από 500 τιμές.

Στην συνέχεια, εκπαιδεύτηκε ένα σύνολο πέντε νευρωνικών δικτύων παλινδρόμησης για την εκτίμηση του χρόνου εκπαίδευσης. Τα δίκτυα αυτά αξιολογήθηκαν από την άποψη του μέσου

τετραγωνικού σφάλματος ρίζας (RMSE) και του συντελεστή συσχέτισης (R). Από αυτά το δίκτυο των δύο επιπέδων πέτυχε την υψηλότερη απόδοση πρόβλεψης αποδίδοντας τον μεγαλύτερο συντελεστή συσχέτισης R και το μικρότερο σφάλμα RMSE.

Μία ακόμη παρατήρηση σχετίζεται με τις μετρήσεις των χρόνων εξαγωγής αποτελεσμάτων για κάθε ένα από τα εκπαιδευμένα μοντέλα. Ενώ για τα τέσσερα από τα πέντε μοντέλα η διάρκεια εξαγωγής αποτελεσμάτων είναι ομοιογενής (από 1.28 έως 2.94 χιλιοστά του δευτερολέπτου), το YAMNet παρουσιάζει πολύ μεγαλύτερο χρόνο εξαγωγής συμπερασμάτων. Είναι επίσης ενδιαφέρον ότι η διάρκεια εκπαίδευσης δεν συσχετίζεται με την διάρκεια εξαγωγής αποτελεσμάτων. Για να συνδυαστούν αυτές οι δύο μετρικές προτάθηκε ένας απλός γραμμικός συνδυασμός των δύο δραστηριοτήτων με βάση τις σχετικές συχνότητες που αυτές πραγματοποιούνται.

Ο αρχικός στόχος της πολυ-παραμετρικής αξιολόγησης του υπολογιστικού χρόνου επιτεύχθηκε. Συγκεκριμένα, με νευρωνικό δίκτυο παλινδρόμησης δύο επιπέδων είναι δυνατή η εκτίμηση του χρόνου εκπαίδευσης πέντε γνωστών ΣΝΔ με συγκεκριμένες ρυθμίσεις των τιμών των υπερπαραμέτρων εκπαίδευσης για το σύνολο δεδομένων UrbanSound8K. Η μελέτη αυτή μπορεί να επεκταθεί και σε άλλα σύνολα δεδομένων και ΣΝΔ. Το σύνολο της συγκεκριμένης μελέτης παρουσιάστηκε στο 27ο Πανελλήνιο Συνέδριο Υπολογιστών και Πληροφορικής και περιέχεται στην δημοσίευση [113].

4. Ταξινόμησης της εικόνας

4.1 Εισαγωγή

Η εικόνα αποτελεί το κατεξοχήν σήμα αναγνώρισης του χώρου, των προσώπων τα οποία συμμετέχουν και των δραστηριοτήτων που λαμβάνουν χώρα σε οποιοδήποτε περιβάλλον. Ακόμα και χωρίς την ύπαρξη άλλου σήματος είναι δυνατή η αντίληψη του γεγονότος που συντελείται. Ο συνδυασμός του σήματος της εικόνας με το ηχητικό σήμα μπορεί να οδηγήσει στην κατανόηση των εκάστοτε ενεργειών και καταστάσεων σχεδόν εξ ολοκλήρου. Αναφορικά με την εκπαιδευτική διαδικασία, οι ηχητικές κλάσεις είναι ικανές σε μεγάλο βαθμό να ταυτοποιήσουν το συγκεκριμένο στάδιο αυτής. Η συνοδεία του ήχου από τις αντίστοιχες εικόνες της εκπαιδευτικής αίθουσας μπορούν να αποσαφηνίσουν περαιτέρω τις συνθήκες και την εξέλιξη της πορείας της συνολικής εκπαιδευτικής διεργασίας.

Η αναγνώριση του χώρου (αίθουσα ή εργαστήριο), του τύπου του μαθήματος (διάλεξη ή πείραμα) και η καταμέτρηση των συμμετεχόντων μπορεί να συντελεστεί με τον εντοπισμό αντίστοιχων αντικειμένων και των προσώπων που υπάρχουν στον χώρο. Στην παρούσα έρευνα γίνεται διεξοδικότερη ανάλυση της εικόνας με στόχο την αναγνώριση των συναισθημάτων των προσώπων. Η αναγνώριση των συναισθημάτων του προσώπου (Facial Emotion Recognition – FER) αποτελεί μέρος της ευρύτερης τεχνολογίας συναισθηματικής υπολογιστικής (affective computing) [114], ένα πεδίο έρευνας για την αλληλεπίδραση μεταξύ ανθρώπων και υπολογιστών το οποίο βασίζεται σε τεχνολογίες τεχνητής νοημοσύνης. Πρόσφατα, η αναγνώριση των συναισθημάτων του προσώπου έχει αποδειχθεί ότι αποτελεί σημαντικό εργαλείο στους τομείς της ιατρικής [115], [116], της υγειονομικής περίθαλψης [117], της έξυπνης διαβίωσης [118], της οδικής κυκλοφορίας [119] και ότι μπορεί να αξιοποιηθεί σε πολλές ακόμη εφαρμογές.

Οι εκφράσεις του προσώπου συνδέονται με τον συνδυασμό των στάσεων των μυών του προσώπου έτσι ώστε κάθε συνδυασμός να αντιστοιχεί σε κάποιο συναίσθημα. Η αναγνώριση των συναισθημάτων του προσώπου έχει στηριχθεί κυρίως στο σύστημα κωδικοποίησης δράσης του προσώπου (Facial Action Coding System) [120] στο οποίο συγκεκριμένες μονάδες δράσης (Action Units – AU) (μύες) αναλύουν τις εκφράσεις του προσώπου. Αποδομώντας κάθε έκφραση στις AU μπορεί να κωδικοποιηθεί κάθε συνδυασμός αυτών που σχηματίζει την έκφραση του προσώπου. Με αυτόν τον τρόπο μπορούν να ληφθούν αποφάσεις συμπεριλαμβανομένης της αναγνώρισης των βασικών συναισθημάτων. Τα βασικά συναισθήματα είναι επτά και σε αυτά συμπεριλαμβάνονται η χαρά, η λύπη, η έκπληξη, ο φόβος, ο θυμός, η αηδία και η περιφρόνηση.

Σε αντιστοιχία με την έρευνα σχετικά με το σήμα του ήχου η ταξινόμηση των εικόνων πραγματοποιείται με (α) εξαγωγή των χαρακτηριστικών της εικόνας με χειροκίνητες μεθόδους (handcrafted methods) και (β) με τεχνικές ΒΜ και συγκεκριμένα με ΣΝΔ.

4.2 Στόχοι

Η εξαγωγή των χαρακτηριστικών της εικόνας είναι ένα κρίσιμο βήμα στην διαδικασίας της ταξινόμησης εικόνων. Το πληροφοριακό περιεχόμενο αυτών των χαρακτηριστικών καθορίζει την ακρίβεια της ταξινόμησης. Η εξαγωγή των χαρακτηριστικών μπορεί να επιτευχθεί είτε με handcrafted μεθόδους είτε με μεθόδους που βασίζονται στα ΣΝΔ. Στην πρώτη περίπτωση καθορίζεται ο μετασχηματισμός που εφαρμόζεται στην εικόνα και οι πληροφορίες που εξάγονται είναι ορισμένες και γνωστές (π.χ. ανάλυση της υφής, περιγραφή των ακμών και γωνιών). Οι handcrafted μέθοδοι αποτελούσαν μέχρι πρότινος το κατεξοχήν εργαλείο εξαγωγής χαρακτηριστικών των εικόνων. Την τελευταία δεκαετία η χρήση των ΣΝΔ έχει αναπτυχθεί σημαντικά. Η χρήση των ΣΝΔ για την εξαγωγή

των χαρακτηριστικών της εικόνας έχει εγείρει ερωτήματα σχετικά με α) το επίπεδο βάθους του δικτύου από το οποίο είναι καταλληλότερο να εξάγονται τα χαρακτηριστικά και β) την ανάγκη εκπαίδευσης των ΣΝΔ είτε from the scratch (η οποία προϋποθέτει σημαντικούς υπολογιστικούς πόρους και μεγάλο σύνολο δεδομένων), είτε την χρήση προ-εκπαιδευμένων δικτύων με τεχνικές μεταφοράς μάθησης.

Στην παρούσα διατριβή επιδιώκεται η διερεύνηση των πτυχών των μεθόδων εξαγωγής των χαρακτηριστικών εικόνων για FER. Η ερευνητική στρατηγική περιλαμβάνει

- την επιλογή μεθόδων εξαγωγής των χαρακτηριστικών εικόνων,
- τον προσδιορισμό των βάσεων δεδομένων εικόνων FER, και
- την αξιολόγηση των μεθόδων ανά βάση δεδομένων διερευνώντας ταυτόχρονα τις δυνατότητες προσαρμογής και εξατομίκευσης.

Συγκεκριμένα, στόχος είναι η αξιολόγηση των δύο μεθόδων εξαγωγής χαρακτηριστικών και η καταλληλόλητα των εξαγόμενων χαρακτηριστικών ως προς την ακρίβεια ταξινόμησης που επιτυγχάνεται. Οι μέθοδοι που επιστρατεύονται περιλαμβάνουν δύο χειροκίνητες μεθόδους εξαγωγής χαρακτηριστικών και πέντε ΣΝΔ. Οι χειροκίνητες μέθοδοι διερευνώνται ως προς (α) το μέγεθος των χαρακτηριστικών που εξάγουν ανάλογα με τις εσωτερικές τους παραμέτρους και (β) την ακρίβεια ταξινόμησης που επιτυγχάνουν. Τα ΣΝΔ εξετάζονται ως προς (α) το επίπεδο βάθους τους από το οποίο εξάγονται τα χαρακτηριστικά, (β) την εξαγωγή των χαρακτηριστικών εφαρμόζοντας την τεχνική της μεταφοράς μάθησης, και (γ) την ακρίβεια ταξινόμησης που επιτυγχάνουν.

Παράλληλα και οι δύο τύποι μεθόδων αξιολογούνται ως προς την ανθεκτικότητά τους σε δύο είδη θορύβων. Στόχος είναι η σύγκριση της απόδοσης κάθε μεθόδου στις εφαρμογές FER ώστε να είναι δυνατή η επιλογή της κατάλληλης μεθόδου αναλόγως των απαιτήσεων, των υπολογιστικών πόρων και των περιορισμών.

Χρησιμοποιούνται τρεις δημόσια διαθέσιμες βάσεις δεδομένων εικόνων οι οποίες περιέχουν διαφορετικό αριθμό εικόνων και με διαφορετικά χαρακτηριστικά, όπως π.χ. το χρώμα, ο αριθμός των κλάσεων και οι πόζες ανά κλάση. Στην συνέχεια γίνεται περιγραφή αυτών καθώς και των υπόλοιπων μηχανισμών που συμμετείχαν στην έρευνα.

4.3 Συναφής έρευνα

4.2.1 Handcrafted μέθοδοι εξαγωγής χαρακτηριστικών

Ο αλγόριθμος Harris-Stephens (1988) [121] βασίζεται στην ανίχνευση γωνιών του Movarec και λαμβάνει υπόψη την κατεύθυνση της αλλαγής της έντασης, καθιστώντας την διάκριση μεταξύ γωνιών και ακμών πιο λεπτομερή. Ενώ η μέθοδος Harris ήταν αμετάβλητη ως προς την περιστροφή, δεν ήταν αμετάβλητη ως προς την κλίμακα. Ο αλγόριθμος Scale-Invariant Feature Transform (SIFT, 2004) [122] ήταν ανεκτικός στην κλιμάκωση αλλάζοντας το μέγεθος του παραθύρου ανάλογα με την κλιμάκωση της εικόνας. Το 2005 προτάθηκε ο έλεγχος Features from Accelerated Segment Test (FAST) [123] για την αντιμετώπιση περιπτώσεων εφαρμογών πραγματικού χρόνου με την προσαρμογή μίας γρήγορης ανίχνευσης χαρακτηριστικών. Σε αυτόν τον έλεγχο, ένα χαρακτηριστικό ανιχνεύεται σε ένα εικονοστοιχείο (pixel) p εάν τα pixels στα καρδιακά σημεία ενός κύκλου ακτίνας 16 pixel με κέντρο αυτό το pixel p έχουν όλα εντάσεις μεγαλύτερες ή μικρότερες από την ένταση του p. Ο αλγόριθμος Speed-Up Robust Feature (SURF, 2006) [124], όπως υποδηλώνει και το όνομά του, είναι η επιταχυνόμενη έκδοση του SIFT. Σε αυτόν τον αλγόριθμο, η Laplacian των Gaussian φίλτρων προσεγγίζεται με BisFilters, τα οποία μπορούν να υπολογιστούν για διαφορετικές κλίμακες ταυτόχρονα. Οι αλγόριθμοι SIFT και SURF έχουν το μειονέκτημα των μεγάλων διανυσμάτων

χαρακτηριστικών, τα οποία είναι επιζήμια από άποψη μνήμης. Το πρόβλημα αυτό επιλύεται με την μέθοδο Binary Robust Independent Elementary Features (BRIEF, 2010) [125], η οποία δίνει δυαδικές συμβολοσειρές συγκρίνοντας τις εντάσεις των ζευγών pixels. Η μέθοδος αυτή δεν ανιχνεύει τα χαρακτηριστικά και επομένως χρειάζεται να προηγηθεί αλγόριθμος ανίχνευσης. Η μέθοδος Oriented Fast and Rotated Binary Robust Independent Elementary Features (ORB, 2011) [126] είναι ένας συνδυασμός των αλγορίθμων FAST, SIFT και SURF ο οποίος διατίθεται ελεύθερα από τα εργαστήρια OpenCV. Τα χαρακτηριστικά KAZE (2012) [127], αντιμετωπίζουν την gaussian θολούρα ανιχνεύοντας και περιγράφοντας δισδιάστατα χαρακτηριστικά σε χώρο μη γραμμικής κλίμακας. Οι τεχνικές διαχωρισμού προσθετικών τελεστών έχουν ως αποτέλεσμα τη μείωση του θορύβου με ταυτόχρονη διατήρηση των ορίων των αντικειμένων. Επιπλέον, τα Local Binary Patters (LBP, 1996) [128] και το Histogram of Oriented Gradients (HOG, 2005) [129] είναι δύο πρακτικοί και ευρέως χρησιμοποιούμενοι αλγόριθμοι, οι οποίοι χρησιμοποιούνται στην παρούσα έρευνα και αναλύονται εκτενώς στην ενότητα 4.4.1. Επίσης έχουν πραγματοποιηθεί μελέτες οι οποίες συγκρίνουν τις παραπάνω μεθόδους [130].

4.3.2 Σύγκριση χαρακτηριστικών που βασίζονται σε handcrafted μεθόδους και σε ΣΝΔ για την ταξινόμηση εικόνας

Η αρχιτεκτονική βαθιάς μάθησης που διαθέτουν τα νευρωνικά δίκτυα συνελικτικής μάθησης τους επιτρέπει να μαθαίνουν απευθείας από τα δεδομένα και να παρέχουν εξαιρετικά ακριβή αποτελέσματα αναγνώρισης. Τα ΣΝΔ μπορούν να εξάγουν και να επεξεργάζονται εσωτερικά χαρακτηριστικά για την εκτέλεση εργασιών όπως η ταξινόμηση εικόνων, ο εντοπισμός αντικειμένων και η αναγνώριση.

Έχουν πραγματοποιηθεί συγκρίσεις των αποτελεσμάτων ταξινόμησης με τις handcrafted μεθόδους και αυτών που βασίζονται στα ΣΝΔ. Στο [131], οι χειροκίνητες μέθοδοι LBP και HOG συγκρίνονται με βαθιά χαρακτηριστικά για την ταξινόμηση ιστοπαθολογικών εικόνων, με την μέθοδο LBP να δίνει τα καλύτερα αποτελέσματα. Από την άλλη πλευρά στο [132], η ακρίβεια ταξινόμησης που επιτεύχθηκε με τη χρήση νευρωνικών δικτύων ήταν 22% υψηλότερη από εκείνη που επιτεύχθηκε με την χρήση διάφορων handcrafted μεθόδων. Στο [133], η συγχώνευση των χαρακτηριστικών που προέρχονται από τις δύο μεθόδους φαίνεται να αποδίδει καλύτερα από κάθε μεμονωμένη περίπτωση για την αναγνώριση εικόνων ήπατος. Στο [134], 18 σύνολα δεδομένων που περιέχουν εικόνες από διάφορες κατηγορίες, οι οποίες εκτείνονται από ιατρικές και υποκυτταρικές μέχρι είδη πεταλούδας, υλικά, χλωρίδα, εικόνες καπνού, πίνακες ζωγραφικής κ.λπ. ταξινομούνται χρησιμοποιώντας τόσο χαρακτηριστικά που βασίζονται στην BM όσο και στις handcrafted μεθόδους. Στην περίπτωση BM, περιλαμβάνεται το δίκτυο ανάλυσης κύριων συνιστωσών (PCAN) και τον συμπαγή δυαδικό περιγραφέα (Compact Binary Descriptor – CBD), καθώς και μεθόδους μεταφοράς μάθησης. Στην περίπτωση των handcrafted μεθόδων, περιλαμβάνεται η μέθοδος LBP και οκτώ παραλλαγές αυτής, η μέθοδος Local Ternary Pattern (LTP) και η Local Phase Quantization (LPQ). Η μείωση της διαστασιολόγησης των χαρακτηριστικών τα οποία εξάγονται από τα ΣΝΔ πραγματοποιήθηκε επίσης με τις μεθόδους του διακριτού μετασχηματισμού συνημίτονου (Discrete Cosine Transform – DCT) και της μεθόδου PCA. Η σύγκριση μεταξύ των handcrafted, των χαρακτηριστικών που εξάγονται από ΣΝΔ, και του συνδυασμού τους έδειξε ότι οι δύο μέθοδοι εξαγωγής χαρακτηριστικών παρέχουν διαφορετικές πληροφορίες και, ως εκ τούτου, η συγχώνευση των χαρακτηριστικών των δύο μεθόδων υπερτερεί των τυπικών προσεγγίσεων. Τα αποτελέσματα της ακρίβειας ταξινόμησης του μοντέλου bag-of-visual-words (BoVW), των χαρακτηριστικών που βασίζονται στα ΣΝΔ και της μεταφοράς μάθησης στο AlexNet συγκρίνονται στο [135], με την τελευταία να υπερέχει. Τα ποσοστά σφάλματος ταξινόμησης σε εικόνες δακτυλικών αποτυπωμάτων με τις δύο μεθόδους εξαγωγής χαρακτηριστικών

μελετήθηκαν στο [136]. Τα handcrafted χαρακτηριστικά είχαν καλύτερη απόδοση στο πλαίσιο του συνόλου δεδομένων, ενώ κατά την αξιολόγηση των αισθητήρων τα χαρακτηριστικά BM επέδειξαν υψηλότερη ακρίβεια ενώ τα handcrafted μικρότερο ποσοστό εσφαλμένων ταξινομήσεων.

4.3.3 Σύγκριση αποτελεσμάτων που βασίζονται σε handcrafted μεθόδους και σε ΣΝΔ για εφαρμογές FER

Ειδικά για τις εφαρμογές αναγνώρισης συναισθήματος προσώπου, έχουν πραγματοποιηθεί λίγες συγκρίσεις μεταξύ των χαρακτηριστικών των δύο μεθόδων. Στο [137], οι συγγραφείς παρέχουν μία επισκόπηση των πρόσφατων εξελίξεων στην αναγνώριση συναισθημάτων με χρήση πολυτροπικών (multimodal) σημάτων, όπου και οι δύο τρόποι εξαγωγής χαρακτηριστικών έχουν χρησιμοποιηθεί. Στο [138], ένας συνδυασμός χαρακτηριστικών από ΣΝΔ από την αρχιτεκτονική VGG με handcrafted χαρακτηριστικά που υπολογίζονται από το μοντέλο BoVW επιτυγχάνει ακρίβεια ταξινόμησης 75.42% στο σύνολο δεδομένων FER-2013 και 87.76% στο σύνολο δεδομένων FER+.

Στις εφαρμογές FER οι μέθοδοι που βασίζονται στα ΣΝΔ κυριαρχούν. Στο [139], η υποδομή του δικτύου ResNet50 χρησιμοποιείται για την εξαγωγή χαρακτηριστικών και την αναγνώριση εκφράσεων του προσώπου, επιτυγχάνοντας ακρίβεια 95% σε ένα σύνολο δεδομένων που δημιουργήθηκε από τους συγγραφείς και το οποίο αποτελείται από 700 εικόνες και επτά διαφορετικές κατηγορίες συναισθημάτων. Στο [140], προτείνεται μία μέθοδος που βασίζεται σε ΣΝΔ και ανίχνευση ακμών εικόνας, επιτυγχάνοντας μέσο ποσοστό αναγνώρισης 88.56% για βάση δεδομένων που έχει προκύψει με συνδυασμό του συνόλου FER-2013 και του συνόλου Labeled Faces in the Wild (LFW). Στο [141], τα ΣΝΔ χρησιμοποιούνται για την αναγνώριση της έκφρασης του προσώπου. Οι συγγραφείς δημιούργησαν μία συλλογή δεδομένων με εικόνες από διάφορα σύνολα για να αποφύγουν την προκατάληψη σε οποιοδήποτε σύνολο. Η επαύξηση των εικόνων επέτρεψε ακρίβεια επικύρωσης 96.24%. Η δυσκολία αναγνώρισης συναισθημάτων από εκφράσεις προσώπου οι οποίες απεικονίζονται σε εικόνες που έχουν ληφθεί σε πραγματικό περιβάλλον αντιμετωπίζεται στο [142] με την χρήση ασύμμετρων πυραμιδικών δικτύων με πυρήνες πολλαπλών κλιμάκων και την υιοθέτηση της στοχαστικής κλίσης καθόδου με optimizer τον SGDM. Η μέθοδος αυτή πέτυχε ακρίβεια ταξινόμησης 74.1% για την βάση δεδομένων FER-2013, 98.50% για την βάση δεδομένων CK+ και 99.80% για την βάση δεδομένων JAFFE. Στο [143], προβάλλονται στρατηγικές περικοπή και περιστροφής προσώπου, απλοποίησης του ΣΝΔ επιτυγχάνοντας ακρίβεια ταξινόμησης 97.38% και 97.18% στις βάσεις δεδομένων CK+ και JAFFE αντίστοιχα.

4.4 Περιγραφή των βάσεων δεδομένων εικόνων

Οι τρεις δημόσια διαθέσιμες βιβλιοθήκες εικόνων που χρησιμοποιήθηκαν περιλαμβάνουν εικόνες οι οποίες συλλέχθηκαν υπό ελεγχόμενες συνθήκες λήψης και τα άτομα πόζαραν με συγκεκριμένες εκφράσεις προσώπου. Αυτές είναι οι εξής:

- **Karolinska Directed Emotional Faces (KDEF)**, η οποία αποτελείται από 4900 εικόνες που κατανέμονται εξίσου σε επτά συναισθήματα προσώπου (κλάσεις) υπό πέντε διαφορετικές γωνίες λήψης (πόζες). Οι συμμετέχοντες είναι άνδρες και γυναίκες, σε ίσα ποσοστά συμμετοχής, ηλικίας από 20 έως 30 ετών, και οι οποίοι δεν φορούν γυαλιά ή κοσμήματα και δεν έχουν γένια ή μουστάκι. Οι εικόνες είναι διαστάσεων 567×762 pixels, με χρωματικές τιμές 24-bit, σε μορφή jpeg [144].
- **Japanese Female Facial Expression (JAFFE)**, η οποία αποτελείται από 213 πρόσωπα γυναικών που εκφράζουν επτά συναισθήματα. Η γωνία λήψης είναι κατά μέτωπο (ανφάς). Οι εικόνες είναι διαστάσεων 256×256 pixels, με χρωματικές τιμές 8-bit της κλίμακας του γκρι,

σε μορφή tiff [145]. Αυτή η βάση δεδομένων επιλέχθηκε για να εξεταστεί η απόδοση των αλγορίθμων σε μικρά σύνολα με εικόνες χαμηλής ανάλυσης.

- **Radboud Faces Database (RaFD)**, αποτελείται από 8040 εικόνες οι οποίες κατανέμονται εξίσου σε οκτώ συναισθήματα προσώπου, υπό πέντε γωνίες λήψης. Οι συμμετέχοντες είναι λευκοί ενήλικες, τόσο άνδρες (30%) όσο και γυναίκες (28%), παιδιά, με τα αγόρια να συμμετέχουν σε ποσοστό 6% και τα κορίτσια σε ποσοστό 9% και μαύροι ενήλικες άνδρες με ποσοστό συμμετοχής 27%. Οι εικόνες είναι διαστάσεων 681×1024 pixels, με χρωματικές τιμές 24-bit, σε μορφή jpeg [146]. Η συγκεκριμένη βάση δεδομένων παρουσιάζει την μεγαλύτερη ποικιλία στα χαρακτηριστικά των συμμετεχόντων.

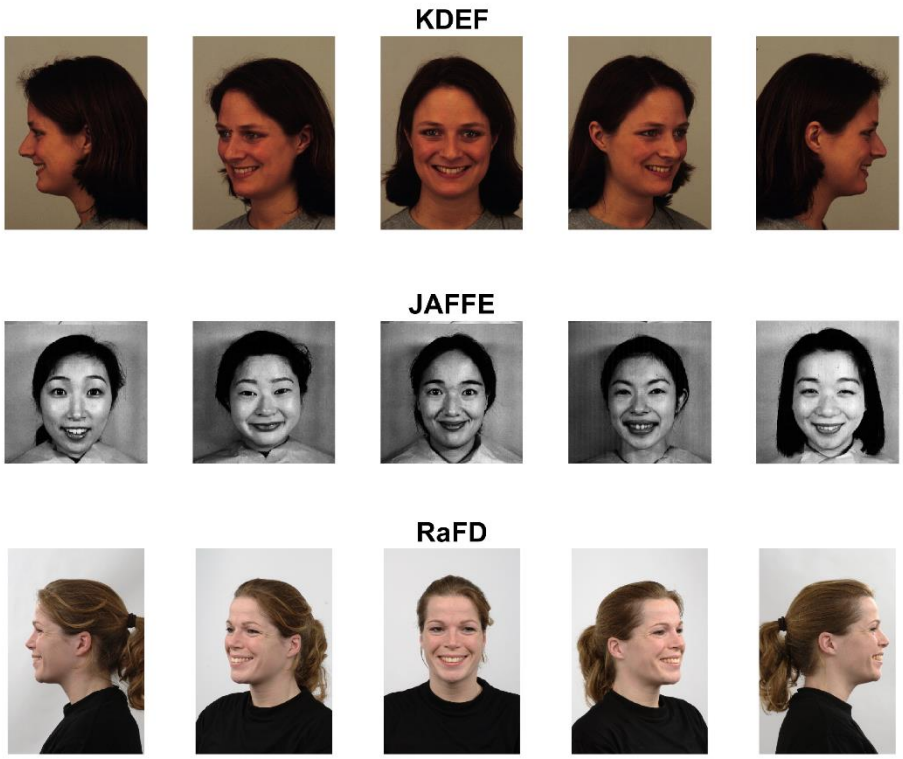
Οι βάσεις δεδομένων KDEF και JAFFE περιέχουν τα ίδια επτά συναισθήματα: θυμός, αηδία, φόβος, χαρά, ουδέτερο, θλίψη και έκπληξη, ενώ η βάση RaFD περιέχει μία ακόμα κατηγορία, την περιφρόνηση. Τα χαρακτηριστικά της κάθε βάσης εικόνων παρουσιάζονται στον Πίνακα 4.1 :

Πίνακας 4.1: Οι βάσεις δεδομένων με τον αριθμό των αρχείων, τις κλάσεις, τις πόζες, τις διαστάσεις και τον τύπο των εικόνων

Βάση δεδομένων	**Πλήθος αρχείων**	**Κλάσεις**	**Πόζες**	**Χρώμα**	**Διαστάσεις (πλάτος × ύψος) pixels**	**Τύπος**
KDEF	4900	7	5	Αληθινό (true color)	562 x 762	jpeg
JAFFE	231	7	1	Κλίμακα του γκρι	256 x 256	tiff
RaFD	8040	8	5	Αληθινό (true color)	681 x 1024	jpeg

Οι βάσεις δεδομένων διαφέρουν ως προς (α) τον αριθμό των εικόνων, (β) τα χαρακτηριστικά αυτών (ανάλυση, βάθος χρώματος, τύπος) και (γ) τον αριθμό των κλάσεων (για την βάση RaFD). Οι βάσεις RaFD και KDEF έχουν το μεγαλύτερο πλήθος εικόνων, γεγονός που υποστηρίζει την εκπαίδευση των μοντέλων αλλά ταυτόχρονα η ταξινόμηση είναι δύσκολη εξαιτίας των διαφορετικών γωνιών λήψης.

Στο Σχήμα 4.1 παρουσιάζεται ένα δείγμα από την κλάση «χαρά» από κάθε βάση δεδομένων.



Σχήμα 4.1: Δείγματα εικόνων από τις τρεις βάσεις δεδομένων που αναπαριστούν το συναίσθημα της χαράς.

4.5 Εξαγωγή χαρακτηριστικών εικόνας

4.5.1 Μέθοδοι handcrafted

Τα χαρακτηριστικά των εικόνων αναγνωρίζονται με τους αλγόριθμους ανίχνευσης χαρακτηριστικών και η ανάλυσή τους πραγματοποιείται με τους αλγόριθμους περιγραφής χαρακτηριστικών. Λόγω της ευρείας χρήσης τους σε εφαρμογές αναγνώρισης προσώπου [147]-[149] στην παρούσα έρευνα διερευνώνται οι αλγόριθμοι Τοπικών Δυαδικών Μοτίβων (Local Binary Patterns – LBP) και Ιστογράμματος Προσανατολισμένων Κλίσεων (Histogram of Oriented Gradients – HOG). Εφεξής οι αλγόριθμοι θα αναφέρονται με την αγγλική τους ορολογία.

(A) Local Binary Patterns

Ο αλγόριθμος LBP κωδικοποιεί τις πληροφορίες υφής μίας εικόνας σε κλίμακα του γκρι συγκρίνοντας την διαφορά στην ένταση κάθε pixel με τα γειτονικά του pixels. Αρχικά, το χρώμα της εικόνας μετατρέπεται σε κλίμακα του γκρι. Στην συνέχεια, η εικόνα διαιρείται σε ορθογώνια κελιά $[k \times k]$. Κάθε pixel i στο κελί συγκρίνεται, ως προς την ένταση, με τα γειτονικά pixels τα οποία βρίσκονται σε κύκλο με κέντρο το pixel i και ακτίνα r. Στο [128] τα γειτονικά pixels είναι 8 και η ακτίνα είναι 1. Θέτοντας την τιμή του κεντρικού pixel ως κατώφλι (που κυμαίνεται από 0 έως 255), τα γειτονικά στοιχεία λαμβάνουν δυαδικές τιμές ως εξής: όσα έχουν τιμή ίση ή μεγαλύτερη από το κατώφλι λαμβάνουν την τιμή 1 και όσα έχουν τιμή μικρότερη από το κατώφλι λαμβάνουν την τιμή 0. Αυτές οι δυαδικές τιμές μετατρέπονται σε δεκαδικές πολλαπλασιάζοντάς τες με δυνάμεις του 2

(διατηρώντας την ίδια κατεύθυνση) και αθροίζοντάς τες. Η διαδικασία επαναλαμβάνεται με κάθε pixel που ανήκει στα 9 διαφορετικά κελιά $[3 \times 3]$ ώστε να μπορεί να λάβει 2^8 διαφορετικές τιμές. Το ιστόγραμμα των $2^8 = 256$ πεδίων της συχνότητας των τιμών που λαμβάνει κάθε pixel αποτελεί το διάνυσμα χαρακτηριστικών 256 διαστάσεων.

Το γεγονός ότι ορισμένα από τα δυαδικά μοτίβα εμφανίζονται συχνότερα από άλλα οδήγησε σε μία βελτιωμένη εκδοχή αυτού του αλγορίθμου η οποία ονομάζεται ομοιόμορφο (uniform) [150]. Ένα μοτίβο είναι ομοιόμορφο όταν έχει το πολύ δύο μεταβάσεις $0 \rightarrow 1$ ή $1 \rightarrow 0$. Το ιστόγραμμα σε αυτή την περίπτωση έχει μία στήλη (bin) για κάθε ομοιόμορφο μοτίβο και μία στήλη για όλα τα μη ομοιόμορφα. Οι στήλες είναι ίσες με $P(P-1)+3$, όπου P είναι ο αριθμός των γειτονικών pixels. Για 8 γειτονικά pixels το ιστόγραμμα, των 256 διαστάσεων, μετατρέπεται σε ιστόγραμμα 59 διαστάσεων, οδηγώντας σε μείωση του μεγέθους των χαρακτηριστικών κατά 77%. Για μία εικόνα $[M \times N]$, το μέγεθος του χαρακτηριστικού δίνεται από την Σχέση (4.1):

$$LBP = \left[floor\left(\frac{M}{k}\right) \times floor\left(\frac{N}{k}\right)\right] \times [P(P-1)+3] \quad (4.1)$$

όπου, $floor$ είναι το ακέραιο μέρος του πηλίκου.

(B) Histogram of Oriented Gradients

Η τεχνική HOG περιγράφει τις ακμές και τις γωνίες ενός αντικειμένου μέσω της κατανομής των τοπικών κλίσεων της έντασης. Για μία εικόνα $[M \times N]$ υπολογίζονται οι κλίσεις κάθε pixel σε πολική μορφή. Οι τιμές αυτές δημιουργούν τους αντίστοιχους πίνακες πλάτους και γωνίας οι οποίοι έχουν τις ίδιες διαστάσεις. Οι πίνακες αυτοί χωρίζονται σε ορθογώνια κελιά $[k \times k]$. Για όλες τις k^2 τιμές υπολογίζεται ένα ιστόγραμμα 9 σημείων, έτσι ώστε κάθε σημείο να έχει εύρος 20 μοιρών στο διάστημα από 0 έως τις 160 μοίρες. Οι θέσεις στο ιστόγραμμα επιλέγονται σύμφωνα με την γωνία της κλίσης και οι τιμές σε κάθε στήλη προκύπτουν από το ποσοστό του αντίστοιχου πλάτους. Αυτά τα ιστογράμματα των 9-σημείων ομαδοποιούνται σε μπλοκ των τεσσάρων $[2 \times 2]$ δημιουργώντας ένα διάνυσμα χαρακτηριστικών 36 διαστάσεων. Η ομαδοποίηση πραγματοποιείται με επικάλυψη k pixel. Το μέγεθος του χαρακτηριστικού με την μέθοδο HOG δίνεται από την Σχέση (4.2):

$$HOG = \left[floor\left(\frac{M}{k}-1\right) \times floor\left(\frac{N}{k}-1\right)\right] \times 36 \quad (4.2)$$

4.5.2 Μέθοδοι βασισμένες σε ΣΝΔ

Τα νευρωνικά δίκτυα αποτελούνται εν γένει από το επίπεδο εισόδου, πολλαπλά κρυφά επίπεδα και το επίπεδο εξόδου (ή ταξινόμησης). Τα κρυφά επίπεδο είναι τριών τύπων: τα συνελικτικά (convolutional), τα συγκεντρωτικά (pooling) και τα πλήρως συνδεδεμένα (fully connected). Στα πρώτα, όπως υποδηλώνει και το όνομά τους, πραγματοποιείται η πράξη της συνέλιξης μεταξύ των τιμών των pixels και του πίνακα των βαρών (kernel filter). Αφού ολοκληρωθεί η σάρωση της εικόνας δημιουργείται ο χάρτης των χαρακτηριστικών (feature map) ή ενεργοποίησης (activation map). Στο επίπεδο συγκέντρωσης επίσης σαρώνεται η εικόνα με την διαφορά ότι σε αυτή τη φάση το φίλτρο δεν έχει βάρη. Το φίλτρο kernel εφαρμόζει μία συνάρτηση συνάθροισης (aggregation function) στις τιμές των pixels δημιουργώντας τον πίνακα εξόδου. Η συνάρτηση συνάθροισης μπορεί να χρησιμοποιεί είτε το pixel με την μέγιστη τιμή (max pooling), είτε την μέση τιμή αυτών (average pooling). Με αυτό τον τρόπο στα συγκεντρωτικά επίπεδα ενώ χάνεται μεγάλη ποσότητα πληροφορίας, επιτυγχάνεται μείωση της πολυπλοκότητας με αποτέλεσμα την βελτίωση της αποδοτικότητας. Τέλος, το πλήρως συνδεδεμένο επίπεδο εκτελεί την ταξινόμηση με βάση τα χαρακτηριστικά που εξάγονται από τα προηγούμενα επίπεδα. Το πρώτο κρυφό επίπεδο ανιχνεύει στοιχειώδη στοιχεία της εικόνας, όπως οι ακμές, τα οποία τροφοδοτούνται στο επόμενο επίπεδο το

οποίο ανιχνεύει πιο σύνθετα στοιχεία, όπως η υφή, κ.ο.κ. Αυτή η διαδικασία συνεχίζεται με αποτέλεσμα το βαθύτερο επίπεδο να ανιχνεύει τα χαρακτηριστικά του υψηλότερου επιπέδου.

Τις τελευταίες δεκαετίες έχουν αναπτυχθεί πολλές αρχιτεκτονικές και τεχνικές που οδήγησαν στην ανάπτυξη των ΣΝΔ. Στην παρούσα εργασία επιλέχθηκαν ΣΝΔ τα οποία έχουν εφαρμόσει διαφορετικές τεχνικές βελτίωσης της ακρίβειας ταξινόμησης: (α) την αρχιτεκτονική ResNet, (β) την αρχιτεκτονική Inception, και (γ) την αρχιτεκτονική Efficient. Συγκεκριμένα, χρησιμοποιούνται τρία δίκτυα από την οικογένεια ResNet, με αυξανόμενο βάθος, προκειμένου να διερευνηθεί αν το βάθος του δικτύου επηρεάζει την ακρίβεια ταξινόμησης. Από τις άλλες δύο αρχιτεκτονικές επιλέχθηκαν το Inception_v3 και το EfficientNet-B0. Όλα τα επιλεγμένα ΣΝΔ έχουν βάθος μέχρι περίπου 100 επίπεδα και λιγότερες από 45 εκατομμύρια παραμέτρους.

- ResNets

Η ιδέα πίσω από την ανάπτυξη της αρχιτεκτονικής της οικογένειας Residual Networks προέρχεται από τη διαίσθηση ότι όσο περισσότερα επίπεδα προστίθενται σε ένα δίκτυο, τόσο πιο σύνθετα προβλήματα μπορεί να επιλύσει και τόσο καλύτερη ακρίβεια θα επιτύχει. Ιδέα η οποία έχει διαψευστεί. Καθώς το βάθος των ΣΝΔ αυξάνεται με την προσθήκη επιπέδων, χάνεται η πληροφορία της κλίσης, με αποτέλεσμα πρώτα τον κορεσμό της απόδοσης και στη συνέχεια την υποβάθμισή της [65]. Η αρχιτεκτονική του ResNet βασίζεται σε συνδέσεις συντόμευσης με χαρτογράφηση της ταυτότητας. Η έξοδος της συντόμευσης προστίθεται στην έξοδο των στοιβαγμένων επιπέδων (αυτών που έχουν παρακαμφθεί), έτσι ώστε εάν κάποιο επίπεδο υποβαθμίζει την ακρίβεια, να παραλείπεται. Τα ΣΝΔ αυτής της οικογένειας αρχιτεκτονικής που χρησιμοποιούνται στην παρούσα μελέτη είναι τα ResNet18, ResNet50 και ResNet101, με τον αριθμό να υποδηλώνει το αντίστοιχο βάθος στρωμάτων.

- Inception_v3

Το γεγονός ότι το αντικείμενο ενδιαφέροντος μπορεί να καταλαμβάνει ένα αυθαίρετο τμήμα της εικόνας οδήγησε στην αρχιτεκτονική Inception. Στο Inception_v1, φίλτρα kernel διαφορετικών διαστάσεων ($1 \times 1, 3 \times 3$ και 5×5) εφαρμόζονται στο ίδιο επίπεδο και οι έξοδοι τους συνενώνονται σε μια ενιαία έξοδο (inception module), σχηματίζοντας ένα δίκτυο που είναι ευρύτερο παρά βαθύτερο. Εφαρμόστηκαν διάφορες τεχνικές βελτίωσης (όπως η παραγοντοποίηση της συνέλιξης 5×5 σε δύο συνελίξεις 3×3 και η χρήση ενός βοηθητικού ταξινομητή (auxiliary classifier)) και οδήγησαν στην προηγμένη έκδοση του Inception_v2. Το Inception_v3 είναι ένα δίκτυο βάθους 48 στρωμάτων στο οποίο, εκτός από τις τεχνικές του Inception_v2, εφαρμόζονται η παραγοντοποίηση της συνέλιξης 7×7 σε τρεις ασύμμετρες συνελίξεις 3×3, η κανονικοποίηση δέσμης (batch normalization) στους βοηθητικούς ταξινομητές και η συστηματοποίηση εξομάλυνσης ετικετών (label-smoothing regularization) [64].

- EfficientNet

Η διαίσθηση ότι όσο υψηλότερη είναι η ανάλυση μιας εικόνας, τόσο μεγαλύτερο πρέπει να είναι το βάθος και το πλάτος του δικτύου, ώστε τα μεγαλύτερα δεκτικά πεδία να μπορούν να ανιχνεύουν χαρακτηριστικά περισσότερων εικονοστοιχείων, οδήγησε στην υλοποίηση των EfficientNets. Στην αρχιτεκτονική των EfficientNets, αντί να επεκτείνεται μία από τις διαστάσεις των δικτύων (βάθος, πλάτος ή ανάλυση), η τεχνική που εφαρμόζεται είναι η ομοιόμορφη κλιμάκωση και των τριών διαστάσεων με ένα σύνολο σταθερών συντελεστών κλιμάκωσης, η επονομαζόμενη μέθοδος σύνθετης κλιμάκωσης. Το EfficientNet-B0, που επιλέχθηκε στην παρούσα μελέτη, έχει βάθος 82 επιπέδων, το οποίο είναι συγκρίσιμο με τα άλλα ΣΝΔ της παρούσας μελέτης [76].

4.6 Ταξινομητής

Η επιβλεπόμενη μηχανική μάθηση περιλαμβάνει δύο κατηγορίες: τους παραδοσιακούς αλγορίθμους ταξινόμησης (δηλ. μη-ΣΝΔ) και τα νευρωνικά δίκτυα. Οι παραδοσιακοί αλγόριθμοι ταξινόμησης, όπως οι Μηχανές Διανυσμάτων Υποστήριξης (SVM), η Γραμμική Διακριτική Ανάλυση (LDA), οι k-κοντινότεροι γείτονες (kNN), ο Naive Bayes και πολλοί άλλοι, χρησιμοποιούνται ευρέως εδώ και χρόνια και έχουν συγκριθεί ως προς την απόδοσή τους σε διάφορες εφαρμογές ταξινόμησης [40]. Στην παρούσα έρευνα, η ακρίβεια ταξινόμησης δοκιμάστηκε αρχικά με τέσσερις διαφορετικούς αλγορίθμους, τον SVM, τον LDA, τον kNN και τον Random Forest, χρησιμοποιώντας διάφορους συνδυασμούς βάθους και νευρωνικών δικτύων, και φάνηκε ότι όλοι έδωσαν παρόμοια αποτελέσματα, με τον SVM να υπερτερεί έναντι των άλλων (+2% περίπου). Ως αποτέλεσμα, ο SVM με την τεχνική "ένα εναντίον ενός" επιλέχθηκε ως ο αντιπροσωπευτικός των παραδοσιακών ταξινομητών.

4.7 Πειραματικά σενάρια

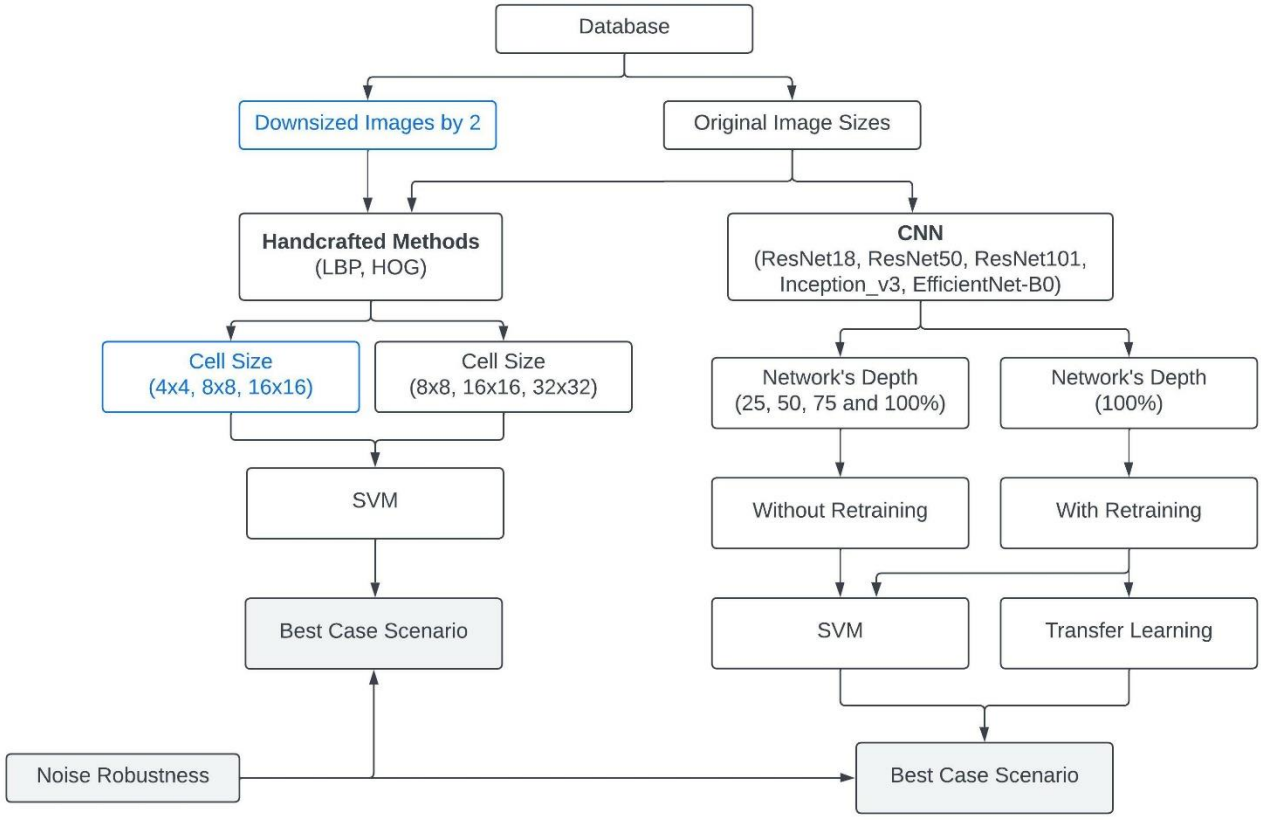
Η ροή εργασιών της παρούσας έρευνας απεικονίζεται στο Σχήμα 4.2. Κάθε μία από τις βάσεις δεδομένων έχει χωριστεί στο σύνολο εκπαίδευσης που περιέχει το 80% των αρχείων και στο σύνολο δοκιμής που περιέχει το υπόλοιπο 20%. Στη συνέχεια, τα χαρακτηριστικά εξάγονται με δύο μεθόδους: (A) την χειροκίνητη εξαγωγή χαρακτηριστικών και (B) εξαγωγή με βάση τα ΣΝΔ.

(A) Για την χειροκίνητη εξαγωγή χαρακτηριστικών, χρησιμοποιήθηκαν οι αλγόριθμοι εξαγωγής χαρακτηριστικών LBP και HOG με τις εικόνες των βάσεων δεδομένων να έχουν τις αρχικές τους διαστάσεις και με σμίκρυνση κατά δύο. Κατά τη διάρκεια των δοκιμών, η μέγιστη ακρίβεια επιτεύχθηκε με διαφορετικά μεγέθη χαρακτηριστικών. Καθώς το μέγεθος των χαρακτηριστικών βασίζεται στην ανάλυση της εικόνας, διερευνήθηκε η ύπαρξη μιας αναλογίας μεταξύ της εικόνας και των μεγεθών των χαρακτηριστικών. Για το σκοπό αυτό, έγινε σμίκρυνση των εικόνων με συντελεστή δύο και κατά τέσσερα, προκειμένου να επαληθευτεί αυτή την αναλογία.

Εφαρμόστηκαν τρία μεγέθη κελιών στις εικόνες με τα αρχικά τους μεγέθη: 8×8, 16×16 και 32×32. Στις εικόνες που έχουν μειωθεί οι διαστάσεις τους κατά δύο, εφαρμόστηκαν μεγέθη κελιών που παράγουν χαρακτηριστικά του ίδιου μεγέθους με τα αρχικά μεγέθη, δηλαδή 4×4, 8×8 και 16×16. Για κάθε συνδυασμό μεγέθους εικόνας, αλγόριθμου εξαγωγής χαρακτηριστικών και μεγέθους κελιού, τα εξαγόμενα χαρακτηριστικά τροφοδοτούν έναν ταξινομητή SVM για τον προσδιορισμό του καλύτερου συνδυασμού όσον αφορά την ακρίβεια ταξινόμησης για κάθε βάση δεδομένων. Επίσης δύο τύποι θορύβου, Gaussian και Salt & Pepper, επιβλήθηκαν στα αρχεία του συνόλου δοκιμών, στον καλύτερο συνδυασμό (μέγεθος εικόνας-μέγεθος κελιού-αλγόριθμος) ώστε να διερευνηθεί η επίδραση του κάθε τύπου στην ακρίβεια ταξινόμησης.

(B) Στην εξαγωγή χαρακτηριστικών με βάση τα ΣΝΔ, εφαρμόστηκαν δύο τρόποι: (i) τα χαρακτηριστικά που προκύπτουν από τα ΣΝΔ όπως αυτά έχουν εκπαιδευτεί με τις εικόνες του ImageNet, και (ii) τα χαρακτηριστικά που προκύπτουν αφού τα ΣΝΔ επανεκπαιδευτούν στις νέες βάσεις δεδομένων. Για τα προ-εκπαιδευμένα ΣΝΔ, η εξαγωγή χαρακτηριστικών πραγματοποιείται από τέσσερα διαφορετικά επίπεδα βάθους, δηλαδή 25%, 50%, 75% και 100% του βάθους τους, και τροφοδοτούν τον ταξινομητή SVM. Για τα επανεκπαιδευμένα ΣΝΔ, τα χαρακτηριστικά εξάγονται από το τελευταίο επίπεδο (καθώς τα ενδιάμεσα βάθη δεν οδήγησαν σε βελτιωμένα αποτελέσματα όσον αφορά την ακρίβεια ταξινόμησης και τον υπολογιστικό χρόνο). Τα εξαγόμενα χαρακτηριστικά τροφοδοτούν τον ταξινομητή SVM και το πλήρως συνδεδεμένο επίπεδο του ταξινομητή του ΣΝΔ (μεταφορά μάθησης). Δεν εφαρμόστηκε σμίκρυνση των εικόνων, καθώς οι εικόνες προσαρμόζονται στις απαιτούμενες από το δίκτυο

διαστάσεις. Η σύγκριση όσον αφορά την ακρίβεια ταξινόμησης δίνει τον καλύτερο συνδυασμό. Για αυτόν τον συνδυασμό, εξετάζεται η επίδραση των δύο τύπων θορύβου που επιβάλλονται στα αρχεία του συνόλου δοκιμών.



Σχήμα 4.2: Ροή εργασιών.

Οι αλγόριθμοι υλοποιήθηκαν με το Matlab R2022a και εκτελέστηκαν σε επιτραπέζιο υπολογιστή με 32 GB RAM, με επεξεργαστή Intel Core i7-10700K με οκτώ πυρήνες, 3.8 GHz, με την κάρτα γραφικών NVIDIA GeForce RTX 3060.

4.7.1 Εξαγωγή χαρακτηριστικών με τις handcrafted μεθόδους

Σε αυτό το σενάριο, τα χαρακτηριστικά της εικόνας εξάγονται με τις μεθόδους LBP και HOG και τροφοδοτούν τον ταξινομητή SVM. Σε κάθε μέθοδο, διερευνάται το μέγεθος των κελιών και, επομένως, το αντίστοιχο μέγεθος των χαρακτηριστικών, το οποίο δίνει την υψηλότερη ακρίβεια ταξινόμησης για κάθε βάση δεδομένων. Και οι δύο αλγόριθμοι μετατρέπουν τις εικόνες σε κλίμακα του γκρι.

Αρχικά εξάγονται τα χαρακτηριστικά από τις εικόνες στο αρχικό τους μέγεθος. Για αυτές τις διαστάσεις, χρησιμοποιούνται τετραγωνικά κελιά με διαστάσεις δυνάμεις του δύο, που κυμαίνονται από 2×2 έως 64×64. Οι Πίνακες 4.2 και 4.3 παρουσιάζουν τα αποτελέσματα με τις μεθόδους LBP και HOG αντίστοιχα, με μεγέθη κελιών 8×8, 16×16 και 32×32, καθώς τα αποτελέσματα για μικρότερα ή μεγαλύτερα μεγέθη κελιών είναι κατώτερα. Το μέγεθος των χαρακτηριστικών και στις δύο περιπτώσεις προκύπτει από τις μαθηματικές σχέσεις (4.1) και (4.2) αντίστοιχα.

Πίνακας 4.2: Η ακρίβεια ταξινόμησης (AT) και το μέγεθος χαρακτηριστικών με τη μέθοδο LBP που εφαρμόζεται στις αρχικές διαστάσεις της εικόνας

LBP	KDEF		JAFFE		RaFD	
Μέγεθος κελιού	**Μέγεθος χαρακτηριστικού**	**AT (%)**	**Μέγεθος χαρακτηριστικού**	**AT (%)**	**Μέγεθος χαρακτηριστικού**	**AT (%)**
8x8	392352	68.84	60416	78.57	641920	88.50
16x16	97055	73.65	15104	76.16	158592	92.66
32x32	23069	72.32	3776	50.00	39648	94.40

Πίνακας 4.3: Η ακρίβεια ταξινόμησης (AT) και το μέγεθος χαρακτηριστικών με τη μέθοδο HOG που εφαρμόζεται στις αρχικές διαστάσεις της εικόνας

HOG	KDEF		JAFFE		RaFD	
Μέγεθος κελιού	**Μέγεθος χαρακτηριστικού**	**AT (%)**	**Μέγεθος χαρακτηριστικού**	**AT (%)**	**Μέγεθος χαρακτηριστικού**	**AT (%)**
8x8	233496	55.57	34596	80.95	384048	69.84
16x16	56304	59.65	8100	80.95	92988	74.63
32x32	12672	56.49	1764	83.33	22320	76.12

Για τα σύνολα εικόνων KDEF και RaFD, οι δύο μέθοδοι παρέχουν τη μέγιστη ακρίβεια ταξινόμησης για το ίδιο μέγεθος κελιών (16×16 για την KDEF και 32×32 για την RaFD). Το μέγεθος του κελιού θα πρέπει να δίνει επαρκείς πληροφορίες με το μικρότερο δυνατό μέγεθος διανύσματος χαρακτηριστικών. Ωστόσο, η μέθοδος πληροφοριών υφής (LPB) αποδίδει σημαντικά υψηλότερα ποσοστά επιτυχίας, και συγκεκριμένα κατά 14% και 18% για τις KDEF και RaFD, αντίστοιχα, σε σύγκριση με τη μέθοδο HOG. Η JAFFE αποτελεί εξαίρεση σε αυτές τις παρατηρήσεις, καθώς παρουσιάζει ελαφρώς καλύτερη ακρίβεια ταξινόμησης με τα χαρακτηριστικά HOG. Ενώ και οι δύο μέθοδοι χρησιμοποιούν την κλίση της έντασης (μέγεθος και κατεύθυνση) ως πληροφορία γύρω από κάθε pixel, η μέθοδος LBP χρησιμοποιεί τα οκτώ γειτονικά pixels για την ανίχνευση τοπικών μοτίβων, ενώ η μέθοδος HOG χρησιμοποιεί μία κατεύθυνση για κάθε pixel. Αυτή η διαφορά καθιστά τη μέθοδο LBP πιο αποτελεσματική στις βάσεις δεδομένων με πολλαπλές γωνίες προσώπου (KDEF και RaFD), ενώ τη μέθοδο HOG στη βάση δεδομένων με εικόνες μόνο μετωπικής πόζας (JAFFE).

Στη συνέχεια οι εικόνες υπόκεινται σε σμίκρυνση με συντελεστή δύο, διατηρώντας την αρχική αναλογία διαστάσεων, προκειμένου να ελεγχθεί ο ρόλος της ανάλυσης της εικόνας στην ακρίβεια της ταξινόμησης και η συσχέτιση μεταξύ του μεγέθους των κελιών και των διαστάσεων της εικόνας. Σύμφωνα με τους τύπους (4.1) και (4.2), όταν οι διαστάσεις των εικόνων υποδιαιρούνται, το ίδιο μέγεθος χαρακτηριστικού προκύπτει και για το υποδιαιρεμένο μέγεθος κελιού. Δηλαδή, το μέγεθος του χαρακτηριστικού με τις αρχικές διαστάσεις της εικόνας για μέγεθος κελιού, π.χ. 8×8 είναι ίσο με εκείνο που προκύπτει για μια εικόνα υποδιαιρεμένη κατά δύο για μέγεθος κελιού 4×4. Επομένως, για να συμπεριληφθεί το χαρακτηριστικό με μέγεθος κελιού 8×8, όταν υποδιαιρούμε τις διαστάσεις των εικόνων, συμπεριλαμβάνεται το μέγεθος κελιού 4×4 και παραλείπεται το μέγεθος κελιού 32×32. Ο Πίνακας 4.4 περιέχει τα αποτελέσματα

ακρίβειας ταξινόμησης και για τις δύο μεθόδους με τις διαστάσεις των εικόνων μειωμένες κατά δύο.

Πίνακας 4.4: Η ακρίβεια ταξινόμησης (%) με τις μεθόδους LBP και HOG που εφαρμόστηκαν στις εικόνες διαστάσεων μειωμένων με συντελεστή δύο

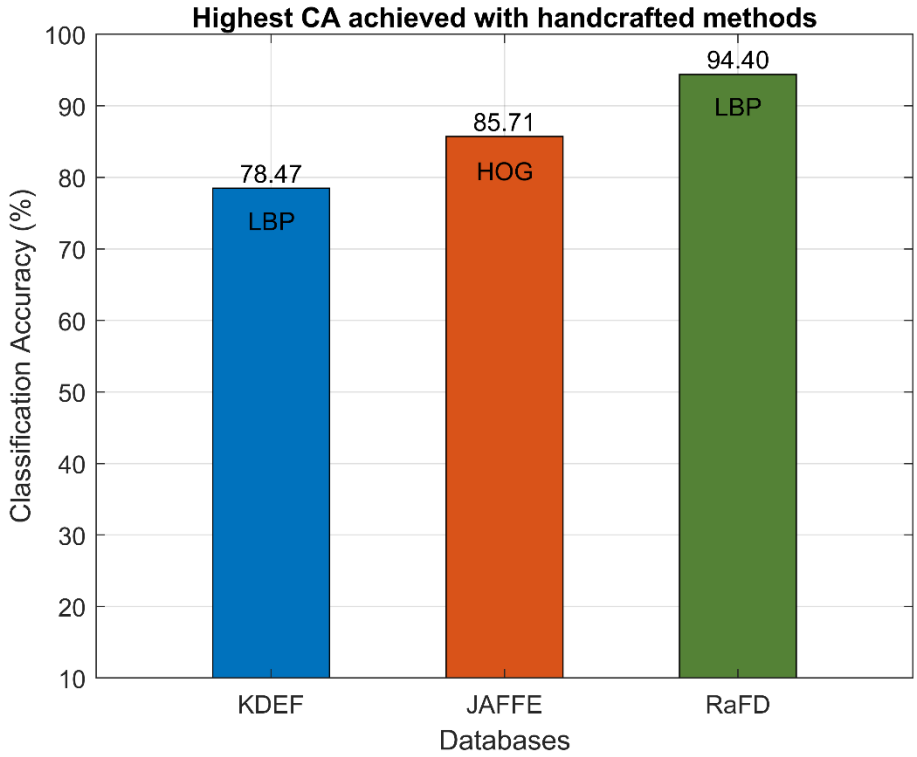
	KDEF		**JAFFE**		**RaFD**	
Μέγεθος κελιού	**LBP**	**HOG**	**LBP**	**HOG**	**LBP**	**HOG**
4x4	70.07%	56.65%	78.57%	84.21%	85.76%	73.82%
8x8	78.47%	69.92%	77.57%	84.71%	92.16%	77.67%
16x16	76.20%	61.08%	76.16%	85.71%	94.40%	78.30%

Η σμίκρυνση των εικόνων με συντελεστή δύο επηρεάζει θετικά την ακρίβεια ταξινόμησης για τις εικόνες KDEF: αυξάνεται κατά 6.3% με τη μέθοδο LBP και κατά 3.3% με τη μέθοδο HOG, κατά μέσο όρο. Το ίδιο ισχύει και για το σύνολο εικόνων JAFFE, με αντίστοιχη βελτίωση κατά 3.1% και 9.5%. Για το σύνολο εικόνων RaFD η σμίκρυνση κατά δύο φαίνεται να αυξάνει την ακρίβεια ταξινόμησης κατά 3.1% κατά μέσο όρο μόνο με τη μέθοδο HOG, ενώ με τη μέθοδο LBP έχουμε μείωση της ακρίβειας ταξινόμησης κατά 1.1% κατά μέσο όρο (δηλαδή παραμένει σχεδόν ανεπηρέαστη η ακρίβεια ταξινόμησης).

Τα ίδια αποτελέσματα εξετάστηκαν επίσης, με τις διαστάσεις των εικόνων μειωμένες κατά τέσσερα. Η ακρίβεια ταξινόμησης, σε αυτή την περίπτωση, φάνηκε να είναι περίπου 2% χαμηλότερη από ό,τι στην περίπτωση της σμίκρυνσης κατά δύο, γι' αυτό αυτά τα αποτελέσματα παραλείπονται. Οι παρατηρήσεις από τους Πίνακες 4.2 και 4.3 σε σύγκριση με τον Πίνακα 4.4 συνοψίζονται στα εξής:

- Η τεχνική LBP παρέχει σημαντικά βελτιωμένη ακρίβεια ταξινόμησης σε σύγκριση με την HOG σε όλες τις βάσεις δεδομένων, εκτός από τη βάση δεδομένων JAFFE.
- Η υψηλότερη ακρίβεια ταξινόμησης για κάθε τεχνική και βάση δεδομένων επιτυγχάνεται με το ίδιο μέγεθος χαρακτηριστικών.
- Η σμίκρυνση των εικόνων και των κελιών (έτσι ώστε το χαρακτηριστικό γνώρισμα να έχει το ίδιο μέγεθος) βελτιώνει την ακρίβεια ταξινόμησης με σμίκρυνση έως και κατά δύο για όλες τις βάσεις δεδομένων. Συγκεκριμένα, η μείωση των διαστάσεων των εικόνων κατά δύο φορές οδήγησε σε βελτιωμένα αποτελέσματα ταξινόμησης με τη μέθοδο HOG για όλες τις βάσεις δεδομένων. Με το αλγόριθμο LBP τα αποτελέσματα βελτιώνονται μόνο για τις βάσεις δεδομένων KDEF και JAFFE.

Στην Σχήμα 4.3 φαίνεται η υψηλότερη ακρίβεια ταξινόμησης που επιτεύχθηκε για κάθε βάση δεδομένων με τις τεχνικές που εφαρμόστηκαν μέχρι τώρα. Σε κάθε περίπτωση, οι διαστάσεις των εικόνων έχουν μειωθεί κατά δύο. Η μέθοδος LBP αποδείχθηκε πιο αποτελεσματική για τις βάσεις δεδομένων KDEF και RaFD, ενώ η μέθοδος HOG για τις εικόνες ευθείας πόζας της JAFFE. Για την KDEF, το βέλτιστο μέγεθος κελιού είναι 8×8 και για τις βάσεις δεδομένων JAFFE και RaFD είναι 16×16.



Σχήμα 4.3: Η υψηλότερη ακρίβεια ταξινόμησης που επιτεύχθηκε για κάθε βάση δεδομένων με τις handcrafted μεθόδους.

4.7.2 Εξαγωγή χαρακτηριστικών από ΣΝΔ

Για την συνέχεια της πειραματικής διερεύνησης έχει οριστεί να χρησιμοποιηθούν τα ίδια αρχεία εικόνων στις φάσεις της εκπαίδευσης και ελέγχου προκειμένου να υπάρχει ομοιογένεια στην σύγκριση των αποτελεσμάτων.

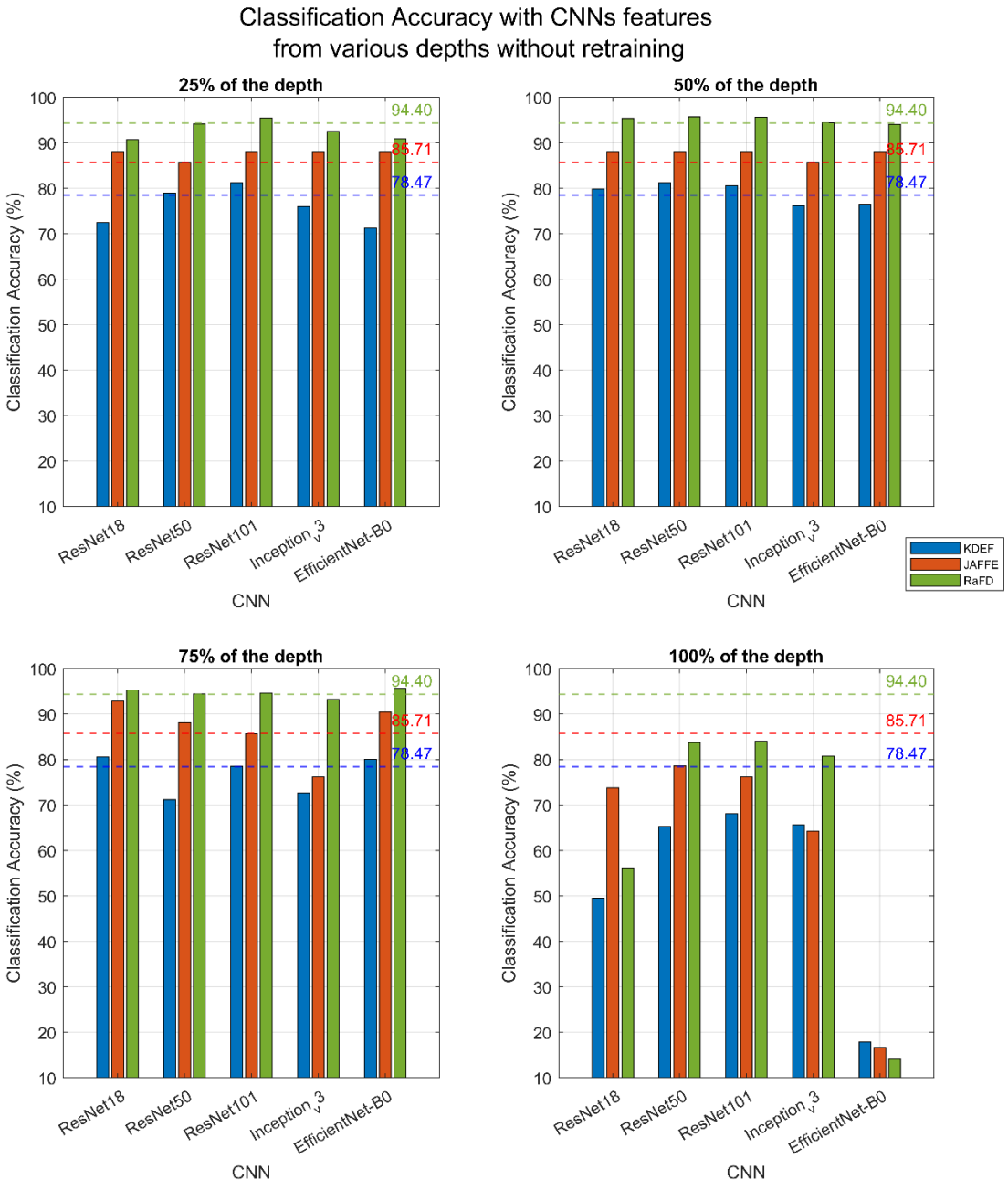
4.7.2.1 Εξαγωγή χαρακτηριστικών χωρίς επανεκπαίδευση των ΣΝΔ

Ανάλογα με το βάθος του επιπέδου, τα ΣΝΔ εξάγουν χαρακτηριστικά διαφορετικού μεγέθους και χωρικής ανάλυσης. Στην πειραματική διαδικασία χρησιμοποιούνται τα χαρακτηριστικά που εξάγονται από τέσσερα διαφορετικά βάθη των ΣΝΔ, συγκεκριμένα από το 25%, 50%, 75% και 100% του βάθους τους, για να εξεταστεί αν τα χαρακτηριστικά που εξάγονται από ρηχότερα επίπεδα είναι πιο κατάλληλα από ό,τι από το επίπεδο εξόδου.

Στο Σχήμα 4.4 απεικονίζονται τα αποτελέσματα της ακρίβειας ταξινόμησης ανά σύνολο (μπλε μπάρα για το KDEF, κόκκινη μπάρα για το JAFFE και πράσινη μπάρα για το RaFD), ανά δίκτυο και ανά βάθος. Η εξαγωγή χαρακτηριστικών από το τελευταίο και βαθύτερο επίπεδο των δικτύων (δηλ. από το 100% του βάθους) οδηγεί στα χειρότερα αποτελέσματα συγκριτικά, όσον αφορά την ακρίβεια ταξινόμησης, με τα επίπεδα επιτυχίας να είναι πολύ χαμηλότερα από εκείνα των handcrafted μεθόδων. Αντίθετα, στο 25%, 50% και 75% του βάθους των δικτύων, η ακρίβεια ταξινόμησης φτάνει ή υπερβαίνει την ακρίβεια ταξινόμησης των handcrafted μεθόδων ανάλογα με το δίκτυο που χρησιμοποιείται.

Στο 25% του βάθους, το μεγαλύτερο σε βάθος δίκτυο, το ResNet101, αποδίδει καλύτερα από τα άλλα δίκτυα, δίνοντας καλύτερα αποτελέσματα και στις τρεις βάσεις δεδομένων. Στο 50% του

βάθους, η τεχνική της οικογένειας ResNet αποδίδει καλύτερα, και συγκεκριμένα, το ResNet50 δίνει παρόμοια αποτελέσματα με αυτά του ResNet101 στο 25%, κάτι που ήταν αναμενόμενο, καθώς η διαφορά στο ποσοστό βάθους σε συνδυασμό με το συνολικό βάθος των δικτύων δίνει χαρακτηριστικά του ίδιου μεγέθους. Στο 75% του βάθους, τα αποτελέσματα είναι κατώτερα από τα προηγούμενα, εκτός από το σύνολο δεδομένων JAFFE, για το οποίο έχουμε την υψηλότερη απόδοση με το ResNet18.



Σχήμα 4.4: Η ακρίβεια ταξινόμησης ανά ΣΝΔ και ανά βάθος δικτύου για κάθε βάση δεδομένων. Τα αποτελέσματα αυτά προκύπτουν χωρίς επανεκπαίδευση των δικτύων με ταξινομητή τον SVM. Οι διακεκομμένες γραμμές απεικονίζουν την υψηλότερη ακρίβεια ταξινόμησης που επιτεύχθηκε με τις handcrafted μεθόδους.

Στον Πίνακα 4.5 συνοψίζονται η υψηλότερη ακρίβεια ταξινόμησης ανά μέθοδο με τον αντίστοιχο υπολογιστικό χρόνο που χρειάστηκε για την εξαγωγή των χαρακτηριστικών και την ταξινόμηση του συνόλου δοκιμών. Για τις περιπτώσεις που προέκυψε το ίδιο ποσοστό ταξινόμησης με διαφορετικά δίκτυα και βάθη, η επιλογή έγινε με βάση τον συντομότερο χρόνο εξαγωγής των αποτελεσμάτων.

Πίνακας 4.5: Η υψηλότερη ακρίβεια ταξινόμησης και ο αντίστοιχος συνολικός χρόνος εξαγωγής αποτελεσμάτων, ανά βάση δεδομένων και μέθοδο εξαγωγής χαρακτηριστικών

	Handcrafted μέθοδοι			**ΣΝΔ**		
Βάση δεδομένων	**ΑΤ (%)**	**Τεχνική**	**Χρόνος (s)**	**ΑΤ (%)**	**ΣΝΔ και βάθος**	**Χρόνος (s)**
KDEF	78.47	LBP (8x8)	1623	81.21	ResNet50, 50% βάθος	1214
JAFFE	85.71	HOG (16x16)	5	92.86	ResNet18, 75% βάθος	55
RaFD	94.40	LBP (16x16)	3746	95.71	ResNet50, 50% βάθος	2988

Μέχρι αυτό το σημείο της πειραματικής διαδικασίας, η εξαγωγή χαρακτηριστικών από το ResNet50 χωρίς επανεκπαίδευση στις νέες βάσεις δεδομένων από το 50% του βάθους του αποδίδει καλύτερα σε σύγκριση με τις χειροκίνητες μεθόδους όσον αφορά το χρόνο εκτέλεσης και την ακρίβεια ταξινόμησης για τα σύνολα εικόνων KDEF και RaFD. Για το σύνολο JAFFE, η ακρίβεια βελτιώνεται επίσης σημαντικά συνοδευόμενη από μία μικρή αύξηση στον χρόνο εκτέλεσης που απαιτείται από το ResNet18 στο 75% του βάθους του.

4.7.2.2 Εξαγωγή χαρακτηριστικών από επανεκπαιδευμένα ΣΝΔ

Η επιλογή των τιμών των υπερπαραμέτρων επανεκπαίδευσης των ΣΝΔ μπορεί να επηρεάζει την ακρίβεια ταξινόμησης, όπως φάνηκε και στην αντίστοιχη πειραματική διαδικασία ταξινόμησης του ήχου στην Παράγραφο 2.3.3. Λαμβάνοντας υπόψη την διερεύνηση που έχει γίνει για τον ήχο, οι τιμές των υπερπαραμέτρων για την ταξινόμηση εικόνων παίρνουν τις εξής τιμές:

- Ο optimizer ρυθμίζεται στον αλγόριθμο Στοχαστικής Κλίσης Καθόδου με Ορμή (SGDM) για την ελαχιστοποίηση της συνάρτησης απώλειας. Στο [84], ο SGDM φαίνεται να συγκλίνει πιο αργά αλλά γενικεύει καλύτερα από τον αλγόριθμο Adam.
- Ο ρυθμός μάθησης είναι ίσος με 0.001, ρύθμιση που συνεπάγεται ότι σε κάθε επανάληψη πραγματοποιούνται μικρά βήματα διόρθωσης.
- Δεδομένου ότι το σύνολο δεδομένων JAFFE είναι σχετικά μικρό (213 εικόνες), το μέγεθος του mini-batch ορίστηκε ίσο με 10, ώστε να υπάρχει επαρκής αριθμός επαναλήψεων για τον υπολογισμό του βάρους.
- Ο μέγιστος αριθμός εποχών ορίστηκε σε 15, ώστε σε συνδυασμό με,
- Την υπομονή επικύρωσης, η οποία ορίστηκε σε 2, να ελεγχθούν οι ενδιάμεσες τιμές των εποχών που είναι επαρκείς για επανεκπαίδευση. Ειδικά για το σύνολο JAFFE, δεν ενεργοποιήθηκε η υπερπαράμετρος της υπομονής επικύρωσης, καθώς πρόκειται για ένα μικρό σύνολο, επιτρέποντας στην εκπαίδευση να εκτελεστεί και για τις 15 εποχές.

Επιπλέον, εφαρμόστηκαν πρόσθετες λειτουργίες επαύξησης των συνόλων δεδομένων, με μετασχηματισμό των εικόνων μέσω μετατόπισης, περιστροφής και κλιμάκωσης αυτών, ώστε να αποφευχθεί η υπερπροσαρμογή.

Μετά την επανεκπαίδευση των ΣΝΔ, η ταξινόμηση των εικόνων στο σύνολο δοκιμών πραγματοποιείται με δύο τρόπους:

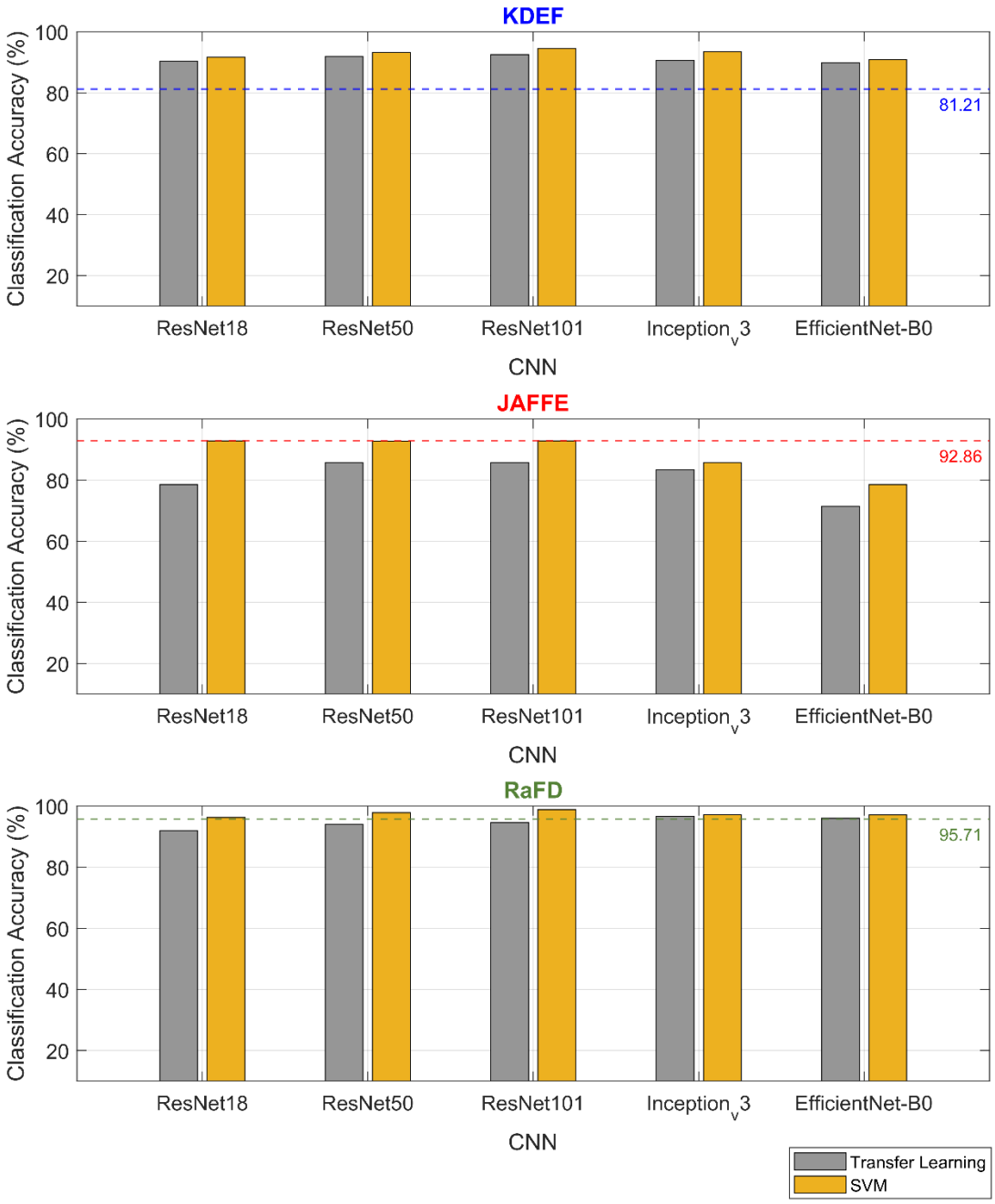
(Α) Εξάγοντας τα χαρακτηριστικά της εικόνας από το τελευταίο (βαθύτερο) επίπεδο και τροφοδοτώντας με αυτά έναν ταξινομητή SVM. Με αυτόν τον τρόπο, συγκρίνουμε την απόδοση ταξινόμησης σε σχέση με αυτήν που προέκυψε με χαρακτηριστικά χωρίς την επανεκπαίδευση των δικτύων που πραγματοποιήθηκε στην Παράγραφο 4.6.2.1. (την προηγούμενη).

(Β) Με τη μεταφοράς μάθησης. Δηλαδή, μετά τη λεπτομερή ρύθμιση των τελευταίων επιπέδων κάθε δικτύου και την αντικατάσταση των εξόδων με τις κλάσεις κάθε βάσης δεδομένων. Η ταξινόμηση πραγματοποιείται από το ίδιο το δίκτυο (πλήρως συνδεδεμένο επίπεδο). Με αυτόν τον τρόπο, συγκρίνονται οι ταξινομητές, δηλαδή ο SVM και ο ενσωματωμένος στο ΣΝΔ.

Τα νέα αποτελέσματα παρουσιάζονται στο Σχήμα 4.5. Οι διακεκομμένες γραμμές σηματοδοτούν τα προηγούμενα μέγιστα επίπεδα ακρίβειας ταξινόμησης που επιτεύχθηκαν από τα ΣΝΔ χωρίς επανεκπαίδευση.

Τα συμπεράσματα τα οποία εξάγονται είναι τα εξής:

- Όσον αφορά τα χαρακτηριστικά, η επανεκπαίδευση έχει νόημα σε σύνολα με πολυάριθμα αρχεία. Ειδικά για το σύνολο KDEF, παρατηρούμε σημαντική αύξηση της ακρίβειας ταξινόμησης, ενώ για το RaFD, η ακρίβεια ταξινόμησης, η οποία ήταν ήδη υψηλή, αυξήθηκε ελαφρώς. Για το σύνολο JAFFE, πρόκειται για μια μικρή βάση δεδομένων, παρατηρείται ότι η επανεκπαίδευση των δικτύων δεν ωφελεί, καθώς μόνο τα ResNets φτάνουν το προηγούμενο μέγιστο (με ταξινομητή SVM).
- Όσον αφορά τα δίκτυα και τις αντίστοιχες αρχιτεκτονικές και μεθόδους τους, τα ResNets παρέχουν υψηλότερα ποσοστά ταξινόμησης και όσο βαθύτερο είναι το δίκτυο, τόσο υψηλότερη είναι η ακρίβεια ταξινόμησης. Ακολουθεί η αρχιτεκτονική Inception_v3 με αποτελέσματα παρόμοια με εκείνα του ResNet50 για τις βάσεις δεδομένων KDEF και RaFD. Τέλος, το EfficientNet-B0 έχει καλές επιδόσεις μόνο στην πιο εκτεταμένη βάση δεδομένων RaFD, ενώ στην μικρότερη βάση δεδομένων, JAFFE, παρατηρείται η χαμηλότερη ακρίβεια ταξινόμησης από όλα τα δίκτυα.
- Όσον αφορά τους ταξινομητές, σε όλες τις περιπτώσεις, ο SVM δίνει καλύτερα αποτελέσματα από τον ενσωματωμένο ταξινομητή του ΣΝΔ.



Σχήμα 4.5:Η ακρίβεια ταξινόμησης μετά την επανεκπαίδευση των ΣΝΔ στην αντίστοιχη βάση δεδομένων. Οι διακεκομμένες γραμμές σηματοδοτούν τα προηγούμενα μέγιστα αποτελέσματα (χωρίς επανεκπαίδευση).

Ο υπολογιστικός χρόνος που απαιτείται για την συνολικό χρόνο εκτέλεσης (επανεκπαίδευση ΣΝΔ και εξαγωγή αποτελεσμάτων) απεικονίζεται στον Πίνακα 4.6.

Πίνακας 4.6: Ο συνολικός χρόνος (s) για την επανεκπαίδευση των ΣΝΔ με το 80% των αρχείων, και την εξαγωγή αποτελεσμάτων για το 20% των αρχείων κάθε βάσης δεδομένων

Συνολικός χρόνος εκτέλεσης (s)					
Βάση δεδομένων	**ResNet18**	**ResNet50**	**ResNet101**	**Inception_v3**	**EfficientNet-B0**
KDEF	2953	6642	8802	6764	19808
JAFFE	131	327	683	638	1031
RaFD	3604	10030	22347	13014	32897

Το EfficientNet-B0, εκτός από τα παρόμοια ή χαμηλότερα αποτελέσματα στην ακρίβεια ταξινόμησης, απαιτεί επίσης τον μεγαλύτερο συνολικό χρόνο. Ακολουθεί το ResNet101 με τον μεγαλύτερο απαιτούμενο χρόνο, κάτι το οποίο είναι αναμενόμενο δεδομένου ότι είναι το μεγαλύτερο από τα δίκτυα που συμμετέχουν στην σύγκριση. Η σύγκριση μεταξύ των ResNet50 και Inception_v3 υποδεικνύει ότι τα δύο αυτά δίκτυα είναι συγκρίσιμα όσον αφορά τον υπολογιστικό χρόνο. Τέλος, το μικρότερο δίκτυο ResNet18 απαιτεί τον μικρότερο χρόνο, ενώ τα αποτελέσματα της ακρίβειας ταξινόμησής του είναι ελαφρώς χαμηλότερα από αυτά των ResNet50 και Inception_v3. Συνολικά, η επιλογή του καταλληλότερου δικτύου προσανατολίζεται σε κάποιο ΣΝΔ της αρχιτεκτονικής οικογένειας ResNet, με το ResNet50 να αποτελεί τη μέση λύση (απόδοση ακρίβειας ταξινόμησης/χρόνου εκτέλεσης).

Μετά την επανεκπαίδευση των δικτύων, η μέγιστη ακρίβεια ταξινόμησης για κάθε σύνολο εικόνων διαμορφώνεται ως εξής:

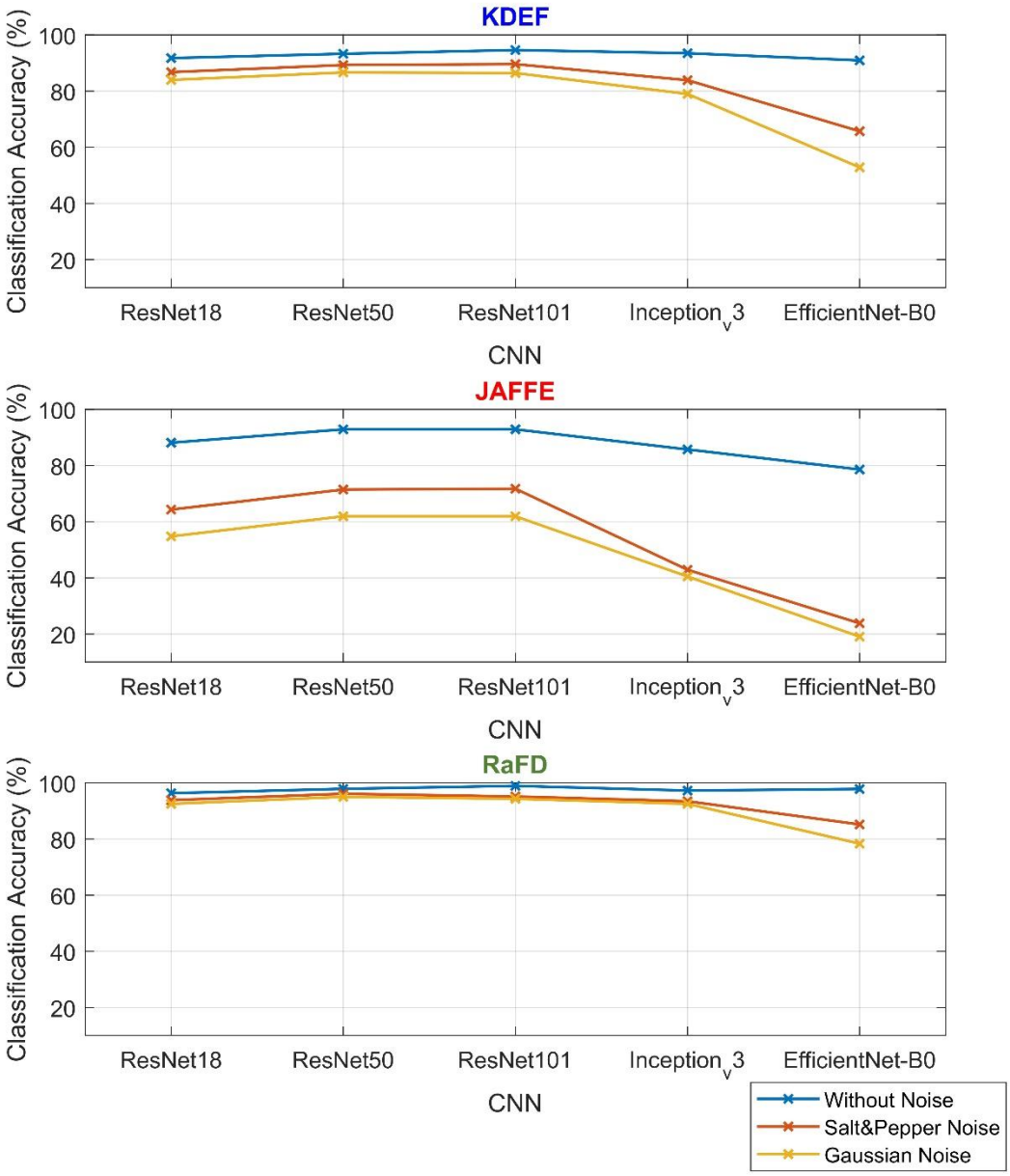
- Για τη βάση δεδομένων KDEF, η ακρίβεια ταξινόμησης έφτασε το 94.59% με το ReseNet101 (βελτίωση κατά 13.4%). Τα ResNet50 και Inception_v3 σημείωσαν επίσης σημαντική βελτίωση στο ποσοστό ταξινόμησης, κατά 12.1% και 12.3%, αντίστοιχα, σε μικρότερο χρόνο εκτέλεσης.
- Για τη βάση δεδομένων JAFFE, η επανεκπαίδευση των δικτύων δεν οδήγησε σε υψηλότερα αποτελέσματα. Το ποσοστό ταξινόμησης έφτασε στο προηγούμενο επίπεδο του 92.86%, αλλά με αυξημένο χρόνο που απαιτήθηκε για τη διαδικασία επανεκπαίδευσης.
- Για τη βάση δεδομένων RaFD, η ακρίβεια ταξινόμησης αυξήθηκε κατά 3.17%, φτάνοντας το 98.88%, με το ResNet101 να έχει την υψηλότερη απόδοση (όσον αφορά την ακρίβεια ταξινόμησης) από όλα τα δίκτυα.

4.7.3 Ανθεκτικότητα στον θόρυβο

Διαφορετικοί τύποι θορύβου μπορούν να επηρεάσουν τις αρχικές εικόνες. Διερευνάται η ανθεκτικότητα των παραπάνω σεναρίων σε δύο τύπους θορύβου.

Gaussian θόρυβος: ο οποίος εμφανίζεται κατά τη λήψη των εικόνων λόγω του θερμικού θορύβου του αισθητήρα και των κυκλωμάτων που συνδέονται με αυτόν. Αυτός ο θόρυβος είναι προσθετικός, ανεξάρτητος και ανεξάρτητος από την ένταση κάθε εικονοστοιχείου με κανονική κατανομή πυκνότητας πιθανότητας, ο οποίος αλλοιώνει κάθε εικονοστοιχείο [151]. **Salt & Pepper θόρυβος:** ο οποίος προκύπτει συνήθως από εσφαλμένα bit κατά τη μετάδοση και την ψηφιοποίηση της εικόνας. Σε αυτή την περίπτωση θορύβου, φωτεινά (salt) ή σκοτεινά (pepper) εικονοστοιχεία είναι διάσπαρτα σε όλη την εικόνα [152]. Η ισχύς του Gaussian θορύβου καθορίζεται από την μέση τιμή και την διακύμανση, ενώ η ισχύς του θορύβου Salt & Pepper από

το ποσοστό των θολωμένων pixels [153]. Ορίσαμε αυτές τις τιμές έτσι ώστε και οι δύο θόρυβοι να έχουν ίσο Μέσο Μέγιστο Λόγο Σήματος προς Θόρυβο (Peak Signal-to-Noise Ratio - PSNR), και περίπου 15 dB. Συγκεκριμένα, εξετάζεται η επίδραση του θορύβου στα υψηλότερα ποσοστά επιτυχίας που επιτυγχάνονται και με τις δύο μεθόδους (είτε με χειροκίνητες είτε με τις βασισμένες στα ΣΝΔ μεθόδους). Στο Σχήμα 4.6 παρουσιάζεται η απόδοση ταξινόμησης σε αλλοιωμένες, από τους δύο τύπους θορύβου, εικόνες. Η εκπαίδευση των ΣΝΔ πραγματοποιήθηκε σε μη αλλοιωμένες εικόνες προκειμένου να εξετάσουμε την χειρότερη περίπτωση απόδοσης. Η ταξινόμηση πραγματοποιείται με τον ταξινομητή SVM.



Σχήμα 4.6: Η ακρίβεια ταξινόμησης ανά ΣΝΔ για κάθε βάση δεδομένων. Η μπλε γραμμή αντιστοιχεί σε καθαρές εικόνες, η κόκκινη εικόνες αλλοιωμένες με Salt & Pepper, και η κίτρινη σε εικόνες αλλοιωμένες με Gaussian θόρυβο. Η εκπαίδευση πραγματοποιήθηκε με καθαρές εικόνες και η ταξινόμηση με SVM.

Οι παρατηρήσεις που προκύπτουν από το Σχήμα 4.6 είναι οι εξής:

1. Όλα τα ΣΝΔ είναι πιο ανθεκτικά (δηλαδή, η ακρίβεια ταξινόμησής τους επηρεάζεται λιγότερο) στον θόρυβο Salt & Pepper από ό,τι στον Gaussian θόρυβο.
2. Η απόδοση μεταξύ των δικτύων διατηρεί την ίδια τάση σε όλες τις περιπτώσεις των βάσεων δεδομένων.
3. Οι εικόνες του συνόλου JAFFE οι οποίες είναι χαμηλής ανάλυσης και με χρωματική κλίμακα του γκρι επηρεάζονται περισσότερο από ό,τι οι έγχρωμες και υψηλότερης ανάλυσης εικόνες των συνόλων KDEF και RaFD. Οι εικόνες της υψηλότερης ανάλυσης του συνόλου (RaFD) επηρεάζονται το λιγότερο από τον θόρυβο.
4. Τα ΣΝΔ που επηρεάζονται περισσότερο από τις αλλοιωμένες εικόνες είναι τα EfficientNet-B0 και Inception_v3, με το πρώτο να είναι το λιγότερο ανθεκτικό.
5. Το πιο εύρωστο δίκτυο, δηλαδή αυτό που σε όλες τις περιπτώσεις των βάσεων δεδομένων, η απόσταση των αποτελεσμάτων της ακρίβειας ταξινόμησης μεταξύ καθαρών και αλλοιωμένων εικόνων είναι η μικρότερη, είναι το ResNet50.

Η ευρωστία των δικτύων στον θόρυβο αποτελεί ένα πεδίο έρευνας και μελέτης το οποίο έχει εξελιχθεί σε σημαντικό βαθμό τα τελευταία χρόνια. Οι συγγραφείς στο [154] διερευνούν την απόδοση προσεγγίσεων που βασίζονται σε βαθιά ΣΝΔ για εφαρμογές αναγνώρισης προσώπου κάτω από πολλές αλλοιώσεις της εικόνας, συμπεριλαμβανομένων των θορύβων Gaussian και Salt & Pepper. Επίσης, στο [155], επισημαίνεται ότι δεν μπορεί να προβλεφθεί εκ των προτέρων πώς θα συμπεριφερθεί το ΣΝΔ με δεδομένα αλλοιωμένα από θόρυβο. Και οι δύο προαναφερθείσες μελέτες προτείνουν την προσθήκη κάποιου θορύβου στο σετ εκπαίδευσης. Στην παρούσα έρευνα, εκπαιδεύτηκαν τα ΣΝΔ με εικόνες αλλοιωμένες (από τους δύο συγκεκριμένους τύπους θορύβου) και η ακρίβεια ταξινόμησης που προέκυψε είναι μεταξύ αυτών που είχαν και στα δύο σετ (εκπαίδευσης και ελέγχου) καθαρές εικόνες και εκείνων στις οποίες τα ΣΝΔ εκπαιδεύτηκαν με καθαρές εικόνες και ελέγχθηκαν σε σετ αλλοιωμένων εικόνων.

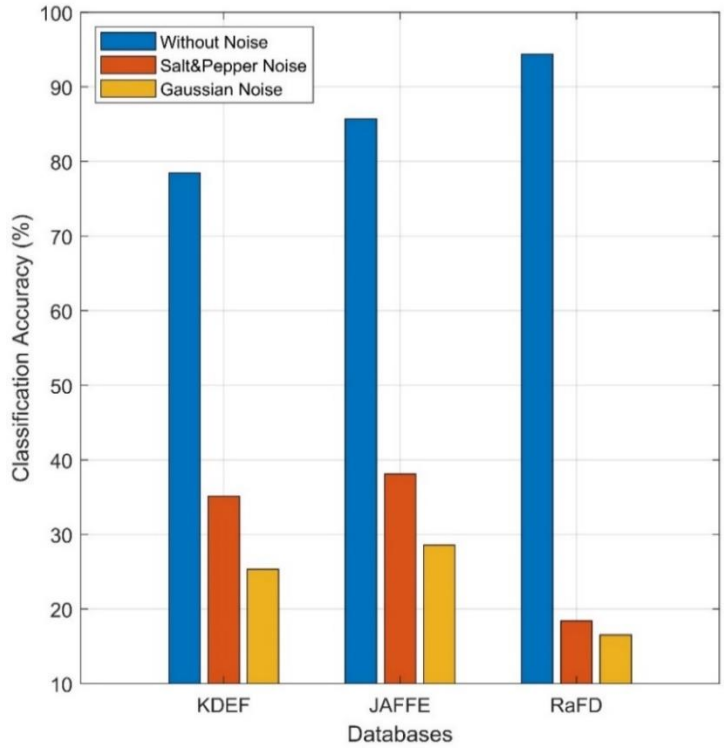
Στον Πίνακα 4.7 παρουσιάζονται οι ποσοστιαίες αποκλίσεις στην ακρίβεια ταξινόμησης που καταγράφηκαν με αλλοιωμένες εικόνες ελέγχου σε σχέση με την ακρίβεια ταξινόμησης με καθαρές εικόνες σε κάθε περίπτωση βάσης δεδομένων και σε κάθε περίπτωση ΣΝΔ.

Πίνακας 4.7: Ποσοστιαία απόκλιση της ακρίβειας ταξινόμησης με αλλοιωμένες εικόνες ελέγχου (σε σχέση με τις καθαρές εικόνες) για κάθε ΣΝΔ και βάση δεδομένων

ΣΝΔ	**ResNet18**		**ResNet50**		**ResNet101**		**Inception_v3**		**EfficientNet-B0**	
	Salt & Pepper	**Gaussian**	**Salt & Pepper**	**Gaussian**	**Salt & Pepper**	**Gaussian**	**Salt & Pepper**	**Gaussian**	**Salt & Pepper**	**Gaussian**
KDEF	5.46	8.47	4.28	7.12	5.30	8.65	10.27	15.51	27.75	41.91
JAFFE	27.03	37.84	23.08	33.34	23.08	33.34	49.99	52.77	69.70	75.75
RaFD	2.59	3.94	1.78	2.92	3.83	4.65	3.90	4.86	12.91	19.91

Σύγκριση της ανθεκτικότητας των αλγορίθμων HOG και LBP έναντι των παραμορφώσεων της εικόνας, συμπεριλαμβανομένων αυτών των δύο εν λόγω τύπων θορύβου, πραγματοποιήθηκε στο [156]. Τα αποτελέσματα δείχνουν ότι ο Gaussian θόρυβος έχει αρνητική επίδραση και στις δύο μεθόδους επειδή επηρεάζεται η πληροφορία των ακμών και η απότομη αλλαγή της κλίσης μπορεί να εκληφθεί σαν ψεύτικη ακμή. Για τον θόρυβο Salt & Pepper, οι υψηλοί ή χαμηλοί παλμοί οδηγούν σε κλίσεις με μεγαλύτερο πλάτος, και η κατεύθυνση

θα δείχνει προς αυτά τα pixels. Στην παρούσα εργασία, η επίδραση του θορύβου στο αντίστοιχο σενάριο από το οποίο προέκυψε η υψηλότερη ακρίβεια ταξινόμησης, με χειροκίνητα μοντέλα παρουσιάζεται στο Σχήμα 4.7. Δηλαδή, τα αποτελέσματα που παρουσιάζονται αφορούν τις αντίστοιχες μεθόδους (LBP για τις βάσεις δεδομένων KDEF και RaFD, και HOG για τη βάση δεδομένων JAFFE) και τα μεγέθη κελιών που κάθε βάση δεδομένων παρουσίασε την υψηλότερη ακρίβεια ταξινόμησης (8×8 για την KDEF, 16×16 για την JAFFE και την RaFD).



Σχήμα 4.7: Η απόδοση ταξινόμησης για κάθε βάση δεδομένων με handcrafted μεθόδους. Η μπλε μπάρα αντιστοιχεί στα αποτελέσματα για καθαρές εικόνες, η κόκκινη για εικόνες ελέγχου με Salt & Pepper, και η κίτρινη μπάρα για εικόνες ελέγχου με Gaussian θόρυβο. Σε κάθε περίπτωση η εκπαίδευση πραγματοποιήθηκε με αναλλοίωτες εικόνες.

Και στην περίπτωση των handcrafted μεθόδων, ο Gaussian θόρυβος υποβαθμίζει την ακρίβεια ταξινόμησης περισσότερο από ό,τι ο Salt & Pepper. Τέλος, είναι αξιοσημείωτο ότι και οι δύο τύποι θορύβου έχουν τα πιο καταστροφικά αποτελέσματα στην αρτιότερη, από άποψη ποιότητας και ποσότητας εικόνων, βάσης δεδομένων, δηλαδή στην RaFD. Ο Πίνακας 4.8 παρουσιάζει τις αντίστοιχες αποκλίσεις στην ακρίβεια ταξινόμησης από τις αντίστοιχες με αλλοιωμένες εικόνες ελέγχου.

Πίνακας 4.8 Ποσοστιαία απόκλιση της ακρίβειας ταξινόμησης με αλλοιωμένες εικόνες ελέγχου (σε σχέση με τις καθαρές εικόνες) για τις handcrafted μεθόδους ανά βάση δεδομένων

Βάση δεδομένων	**Salt & Pepper Noise**	**Gaussian Noise**
KDEF	55.22	67.72
JAFFE	55.55	66.67
RaFD	80.50	82.48

Είναι εμφανές ότι η επίδραση των συγκεκριμένων τύπων θορύβου, οι οποίοι εξετάστηκαν στην παρούσα μελέτη, είναι εντονότερη στην περίπτωση των χειροκίνητων μεθόδων.

4.8 Συμπεράσματα

Η παρούσα μελέτη διερευνά την ακρίβεια ταξινόμησης καθώς και τον υπολογιστικό χρόνο που απαιτείται για την ταξινόμηση συναισθηματικών εκφράσεων του προσώπου. Συγκεκριμένα εξετάστηκαν α) οι χειροκίνητες μέθοδοι εξαγωγής χαρακτηριστικών των εικόνων LBP και HOG, και β) η εξαγωγή χαρακτηριστικών των εικόνων με βάση τα ΣΝΔ. Χρησιμοποιήθηκαν τρεις βάσεις δεδομένων, η KDEF, η RaFD και η JAFFE. Η χρήση των ΣΝΔ ήταν διττή. Αρχικά, τα χαρακτηριστικά εξήχθησαν χωρίς να επανεκπαιδευτούν τα δίκτυα στα νέα δεδομένα, από το 25%, 50%, 75% και 100% του βάθους τους. Το συμπέρασμα που προέκυψε από αυτή την έρευνα έδειξε ότι, η εξαγωγή των χαρακτηριστικών από τα ρηχότερα επίπεδα είναι σημαντικά αποτελεσματικότερη εάν οι νέες εικόνες είναι διαφορετικές από εκείνες στις οποίες τα δίκτυα έχουν αρχικά εκπαιδευτεί. Η δεύτερη χρήση των ΣΝΔ ήταν η εξαγωγή των χαρακτηριστικών εικόνων μετά την επανεκπαίδευσή τους στα νέα δεδομένα (μεταφορά μάθησης).

Ο Πίνακας 4.9 συνοψίζει τα υψηλότερα αποτελέσματα των τριών μεθόδων που εφαρμόστηκαν για τις τρεις βάσεις δεδομένων. Όσον αφορά τις χειροκίνητες μεθόδους, η LBP δίνει υψηλότερα ποσοστά επιτυχίας στις εικόνες υψηλής ανάλυσης (βάσεις δεδομένων KDEF και RaFD), ενώ αντίθετα η HOG αποδίδει καλύτερα στις εικόνες χαμηλότερης ανάλυσης (JAFFE). Τα αποτελέσματα της ταξινόμησης εμφανίζονται βελτιωμένα με την άμεση εξαγωγή των χαρακτηριστικών από ρηχά επίπεδα των δικτύων της αρχιτεκτονικής residual. Επιπλέον, παρατηρήθηκε μείωση του υπολογιστικού χρόνου για τις μεγάλες βάσεις δεδομένων σε σύγκριση με τις χειροκίνητες μεθόδους. Τέλος, η μεταφορά μάθησης βελτιώνει την ακρίβεια ταξινόμησης για τις μεγάλες βάσεις δεδομένων, επιβαρύνοντας όμως τον υπολογιστικό χρόνο. Για το μικρότερο σε πλήθος εικόνων σύνολο δεδομένων (JAFFE) η ακρίβεια δεν βελτιώθηκε, με το υψηλότερο ποσοστό ταξινόμησης να παραμένει στο 92.86%. Ο ταξινομητής SVM επέδειξε καλύτερες επιδόσεις από τον ενσωματωμένο ταξινομητή των ΣΝΔ.

Πίνακας 4.9: Συγκεντρωτικός πίνακας των υψηλότερων αποτελεσμάτων της ακρίβειας ταξινόμησης και του συνολικού υπολογιστικού χρόνου ανά μέθοδο και βάση δεδομένων

Βάση δεδομένων	**Μέθοδος**	**Απόδοση ταξινόμησης (%)**	**Υπολογιστικός χρόνος (s)**
KDEF	LBP	78.47	1623
	Χωρίς επανεκπαίδευση-50% ResNet50	81.21	1214
	Μεταφορά μάθησης-ResNet101	94.59	8802
JAFFE	HOG	85.71	5
	Χωρίς επανεκπαίδευση-75% ResNet18	92.86	55
	Μεταφορά μάθησης-ResNet18,50,101	92.86	131,327,683
RaFD	LBP	94.40	3746
	Χωρίς επανεκπαίδευση-50% ResNet50	95.71	2988
	Μεταφορά μάθησης-ResNet101	98.88	22347

Σύμφωνα με αυτά τα αποτελέσματα δημιουργήθηκε ένα πλαίσιο αποφάσεων για την υποστήριξη της κατάλληλης επιλογής με βάση τις προδιαγραφές κάθε εφαρμογής. Ένα τέτοιο πλαίσιο παρουσιάζεται στον Πίνακα 4.10.

Πίνακας 4.10: Πλαίσιο υποστήριξης απόφασης της επιλογής μεθόδου εξαγωγής χαρακτηριστικών

Τύπος βάσης δεδομένων	**Κριτήριο**	**Μέθοδος**
Μικρό πλήθος εικόνων	Υψηλή ακρίβεια ταξινόμησης	Χωρίς επανεκπαίδευση-75% ResNet18
Χαμηλή ποιότητα		
Ευθείες πόζες	Μικρός υπολογιστικός χρόνος	HOG
Μέτριο πλήθος εικόνων	Υψηλή ακρίβεια ταξινόμησης	Μεταφορά μάθησης-ResNet101
Υψηλή ποιότητα		
Πολλαπλές γωνίες πόζας	Μικρός υπολογιστικός χρόνος	Χωρίς επανεκπαίδευση-50% ResNet50
Μεγάλο πλήθος εικόνων	Υψηλή ακρίβεια ταξινόμησης	Μεταφορά μάθησης-ResNet101
Υψηλή ποιότητα		
Πολλαπλές γωνίες πόζας	Μικρός υπολογιστικός χρόνος	Χωρίς επανεκπαίδευση-50% ResNet50

Συνολικά, τα χαρακτηριστικά που εξάγονται με χειροκίνητες μεθόδους εδώ και δεκαετίες δεν φτάνουν τις επιδόσεις των χαρακτηριστικών των εξαγόμενων από ΣΝΔ. Η χρυσή τομή μεταξύ της απόδοσης ταξινόμησης και του υπολογιστικού χρόνου βρίσκεται στην εξαγωγή χαρακτηριστικών από επίπεδο ενδιάμεσου βάθους αναλόγως του πλήθους και της ποιότητας των εικόνων, χωρίς επανεκπαίδευση των δικτύων. Αν στην εκάστοτε εφαρμογή δίνεται προτεραιότητα στην υψηλή απόδοση ταξινόμησης, απαιτείται μεγάλο πλήθος από εικόνες

υψηλής ποιότητας για την επανεκπαίδευση των δικτύων (μεταφορά μάθησης). Μεταξύ των αρχιτεκτονικών που εξετάστηκαν η αρχιτεκτονική residual αποδείχθηκε η πιο αποτελεσματική. Οι δύο τύποι θορύβου που εξετάστηκαν φαίνεται να έχουν μεγαλύτερη επίδραση στις handcrafted μεθόδους, ενώ το ΣΝΔ ResNet50 αποδείχθηκε ότι είναι το πιο ανθεκτικό σε αυτούς.

Το σύνολο αυτής της έρευνας περιλαμβάνεται στην δημοσίευση [157] μέχρι την συγγραφή της οποίας δεν είχε πραγματοποιηθεί λεπτομερής σύγκριση σε σενάρια εξαγωγής χαρακτηριστικών εικόνας από διαφορετικά επίπεδα του βάθους των ΣΝΔ ως προς (α) την ακρίβεια ταξινόμησης, (β) τους υπολογιστικούς πόρους, με όρους χρόνου εκτέλεσης των αλγορίθμων, και (γ) την ανθεκτικότητα στον θόρυβο.

5. Επέκταση σε άλλες περιοχές εφαρμογών

Η διεξοδική έρευνα που πραγματοποιήθηκε, τόσο για το σήμα του ήχου όσο και για το σήμα της εικόνας, επέτρεψε την εφαρμογή και την αξιολόγηση των αλγορίθμων σε άλλες θεματικές περιοχές. Συγκεκριμένα, μετά το πέρας της μελέτης κάθε σήματος, οι αλγόριθμοι και οι μεθοδολογίες εφαρμόστηκαν σε μελέτες περιβαλλοντικού πλαισίου. Το αποτέλεσμα ήταν η συμμετοχή και η διάκριση δύο δημοσιεύσεων στο ετήσιο διεθνές συνέδριο Τεχνολογιών και Υλικών για Ανανεώσιμες Πηγές Ενέργειας, το Περιβάλλον και την Αειφορία (Technologies and Materials for Renewable Energy, Environment and Sustainability – TMREES) και η δημοσίευση των άρθρων στο επιστημονικό περιοδικό Energy Reports, Elsevier. Με αυτές τις δύο δημοσιεύσεις αναδεικνύεται η επεκτατική αξία της έρευνας που προηγήθηκε, καθιστώντας και άλλους χώρους «έξυπνους» μέσω της παρακολούθησης και διαχείρισης αυτών των δύο βασικών σημάτων. Στην συνέχεια παρατίθενται αυτές οι δύο μελέτες για κάθε ένα από τα σήματα. Η πρώτη αφορά τον περιβαλλοντικό θόρυβο και την ηχορρύπανση και η δεύτερη την έγκαιρη ανίχνευση πυρκαγιών και την ανάλυση κινδύνου αναλόγως των υποδομών που υπάρχουν στην εν λόγω περιοχή.

5.1 Παρακολούθηση, σκιαγράφηση και ταξινόμηση του αστικού περιβαλλοντικού θορύβου με χρήση ηχητικών χαρακτηριστικών και του αλγορίθμου k-NN

5.1.1 Εισαγωγή

Ο περιβαλλοντικός θόρυβος αποτελεί βασικό παράγοντα ο οποίος επηρεάζει την ποιότητα της ζωής στις σύγχρονες κοινωνίες, καθώς επηρεάζει ένα ευρύ σύνολο δραστηριοτήτων. Οι ανεπιθύμητοι ή ενοχλητικοί ήχοι, που συνήθως χαρακτηρίζονται ως *θόρυβος*, μπορεί να είναι πολλών ειδών και ποικίλλουν ως προς τις επιπτώσεις τους και τους τρόπους αντιμετώπισης. Οι επιπτώσεις του θορύβου έχουν μελετηθεί λεπτομερώς στο παρελθόν [158], [159], εντοπίζοντας μία σειρά από αρνητικές επιπτώσεις, συμπεριλαμβανομένων της εξασθένησης της ακοής, της παρεμβολής στην επικοινωνία, της διαταραχής του ύπνου, των επιπτώσεων στην ψυχική υγεία και την συμπεριφορά. Ο θόρυβος, ακόμα και χαμηλής έντασης έχει αρνητικές συνέπειες στις περιπτώσεις όπου εκτελείται πνευματική εργασία ή επιδιώκεται ηρεμία όπως π.χ. σε εκπαιδευτικά ιδρύματα ή σε νοσοκομεία. Στην Ευρώπη ο περιβαλλοντικός θόρυβος αναγνωρίζεται ως ένα από τα κυριότερα περιβαλλοντικά προβλήματα [160] και ο Παγκόσμιος Οργανισμός Υγείας έχει ορίσει τα επίπεδα της μέτριας και της σοβαρής ενόχλησης για τους υπαίθριους χώρους στα 50dB και 55dB αντίστοιχα. Οι αντίστοιχες τιμές για τους εσωτερικούς χώρους είναι τα 30dB και 35dB [158]. Οι τύποι και τα χαρακτηριστικά της ηχορρύπανσης ποικίλουν και θόρυβοι παρόμοιας έντασης έχουν διαφορετικό αντίκτυπο καθώς και διαφορετική προέλευση. Ενώ τα υπάρχοντα πλαίσια τα οποία σχετίζονται με τον ήχο εξετάζουν κυρίως το επίπεδο της έντασης του θορύβου, η περαιτέρω επεξεργασία και κατανόηση των τύπων θορύβου βρίσκεται ακόμα υπό διερεύνηση. Η κατηγοριοποίηση του θορύβου μπορεί να επιτρέψει την αποτελεσματικότερη διαχείριση και λήψη αποφάσεων.

5.1.2 Στόχοι

Η ικανότητα καλύτερης κατανόησης των χαρακτηριστικών του θορύβου είναι σημαντική για την αποτελεσματικότερη αντιμετώπιση και λήψη αποφάσεων. Οι στόχοι της παρούσας εργασίας περιλαμβάνουν:

α. Την διερεύνηση των χαρακτηριστικών επιλεγμένων θορύβων του αστικού περιβάλλοντος και την στατιστική καταγραφή του θορύβου με βάση ένα σύνολο ιδιοτήτων (χαρακτηριστικών) που σχετίζονται με τον ήχο.

β. Την επιλογή των κατάλληλων χαρακτηριστικών οι τιμές των οποίων πρέπει να εξαχθούν για οκτώ τύπους (κλάσεις) θορύβου, προκειμένου να πραγματοποιηθεί ταξινόμηση των θορύβων με χρήση του αλγόριθμου MM k-NN.
γ. Την αξιολόγηση των αποτελεσμάτων με την χρήση δειγμάτων θορύβου από δημόσια διαθέσιμη βάση δεδομένων ήχων.

5.1.3 Συναφής έρευνα

Για την αξιολόγηση και εν συνεχεία διαχείριση του περιβαλλοντικού θορύβου η Ευρωπαϊκή Επιτροπή εξέδωσε την οδηγία 2002/49/EC [160]. Σύμφωνα με αυτό το πλαίσιο, οι κύριες πόλεις αναμένεται να μετρούν την έκθεση στον θόρυβο και να παρέχουν ακριβείς χαρτογραφήσεις των επιπέδων της ηχορρύπανσης για την υποστήριξη της λήψης αποφάσεων και την ενεργοποίηση περαιτέρω δράσεων [161], [162]. Η στρατηγική χαρτογράφηση του θορύβου έχει εγκριθεί από τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ). Ο προσδιορισμός των επιπέδων θορύβου βασίστηκε στην μέγιστη ένταση για την αξιολόγηση α) της ενόχλησης και β) της διαταραχής ύπνου. Η ένταση έχει υπολογιστεί από την στάθμη της ηχητικής πίεσης και έχει εναρμονιστεί με τις τιμές αναφοράς. Η χαρτογράφηση του θορύβου περιλαμβάνει την χαρτογράφηση του περιγράμματος του θορύβου και την εκτίμηση της έκθεσης.

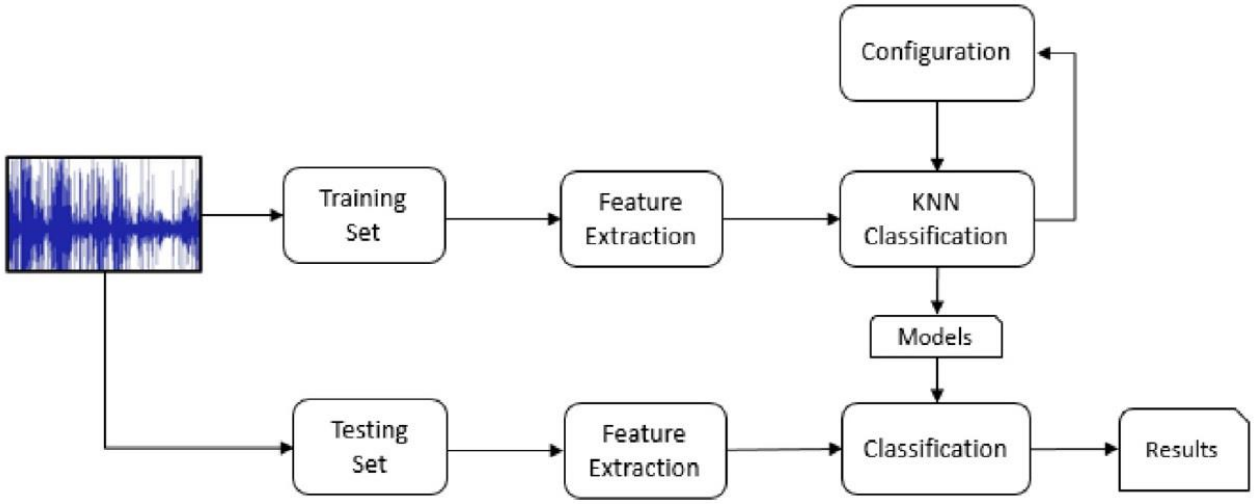
Η ένταση του θορύβου εξάγεται και αναπαρίσταται με απλό τρόπο, αλλά θεωρείται μόνο ένα βασικό χαρακτηριστικό όσον αφορά της ψυχο-ακουστική ενόχληση καθώς δεν περιλαμβάνονται χαρακτηριστικά συχνότητας. Θόρυβος παρόμοιας έντασης μπορεί να οδηγήσει σε διαφορετική αντίληψη αποτυγχάνοντας να παρέχει πληροφορίες οι οποίες σχετίζονται με την υποκειμενική ενόχληση και τις ψυχοακουστικές τους ιδιότητες [15]. Προς αυτή την κατεύθυνση, το μοντέλο ενόχληση Zwicker λαμβάνει υπόψη πρόσθετους παράγοντες, όπως η ένταση, η οξύτητα, η ένταση της διακύμανσης και η τραχύτητα [163], [164]. Από αυτή την άποψη, απαιτείται μία λεπτομερής ανάλυση των χαρακτηριστικών του ήχου η οποία να περιλαμβάνει ένα εκτεταμένο και καλά οργανωμένο σύνολο ηχητικών χαρακτηριστικών, τα οποία να επιτρέπουν περαιτέρω κατανόηση των ιδιοτήτων του ήχου.

Στο πλαίσιο των προσπαθειών για έξυπνες πόλεις, έχουν επιδιωχθεί υποδομές για την ανάκτηση μετρήσεων θορύβου με βάση το ασύρματο δίκτυο ακουστικών αισθητήρων (Wireless Acoustic Sensor Network – WASN). Η επεξεργασία των μετρήσεων επεκτείνεται όλο και περισσότερο πέρα από την παρακολούθηση της έντασης σε πιο λεπτομερή ανάλυση και αναγνώριση των τύπων θορύβου. Στο [15], οι ήχοι ανακτήθηκαν χρησιμοποιώντας Raspberry Pi Zero W και πραγματοποιήθηκε ταξινόμηση χρησιμοποιώντας τα χαρακτηριστικά MFCC τα οποία τροφοδότησαν τους αλγορίθμους επιβλεπόμενης μάθησης SVM και kNN. Στο [8], αναπτύχθηκε ταξινόμηση ήχου δύο επιπέδων βασισμένη σε νευρωνικά δίκτυα, για την παρακολούθηση του ήχου στο πλαίσιο της διαχείρισης της πολιτιστικής κληρονομιάς.

Νευρωνικά δίκτυα συνελικτικού τύπου, δύο πλήρως συνδεδεμένων επιπέδων, χρησιμοποιήθηκαν στο [163] για την ταξινόμηση σύντομων ηχητικών αποσπασμάτων περιβαλλοντικών ήχων. Η εκπαίδευση πραγματοποιήθηκε σε χαμηλού-επιπέδου αναπαραστάσεις spectrograms ηχητικών δεδομένων. Κατάτμηση χρόνου-συχνότητας πραγματοποιήθηκε στο [165] και ταξινόμηση με ΣΝΔ. Η ανίχνευση ηχητικών συμβάντων σε θορυβώδη περιβάλλοντα με ένταση συγκρίσιμη με τα ηχητικά συμβάντα πραγματοποιήθηκε με την χρήση της προσέγγισης bag of words στο [164]. Σε αυτές τις προσπάθειες, η ποσοτική αξιολόγηση της ακρίβειας ανίχνευσης των συστημάτων ανάλυσης ηχητικών συμβάντων πραγματοποιείται μέσω της σύγκρισης της εξόδου του συστήματος με δεδομένα ελέγχου αναφοράς [166], [167].

5.1.4 Μεθοδολογία

Η μεθοδολογία περιλαμβάνει τα ακόλουθα βήματα: αρχικά επιλέγονται οι τύποι του θορύβου, δηλαδή ήχοι που συναντώνται σε αστικά περιβάλλοντα. Τα ηχητικά αποσπάσματα με ετικέτα υπόκεινται σε προεπεξεργασία (συμπεριλαμβανομένης της κανονικοποίησης και της κατάτμησης) και διαιρούνται σε σύνολα εκπαίδευσης και ελέγχου. Στην συνέχεια, επιλέγονται τα ηχητικά χαρακτηριστικά που εξάγονται από τα αρχεία εκπαίδευσης και μετά την τυποποίησή τους, εισάγονται στον αλγόριθμο ταξινόμησης kNN. Ο αλγόριθμος ταξινόμησης δημιουργεί ένα σύνολο μοντέλων ταξινόμησης, επιτρέποντας την λεπτομερή ρύθμισή του όσον αφορά την μετρική της απόστασης και τον αριθμό των γειτόνων που χρησιμοποιούνται. Χρησιμοποιώντας το μοντέλο πραγματοποιείται ταξινόμηση στα αρχεία ελέγχου. Η αξιολόγηση των αποτελεσμάτων πραγματοποιείται με την σύγκριση των πραγματικών ετικετών. Η ροή εργασιών παρουσιάζεται στο Σχήμα 5.1.



Σχήμα 5.1: Ροή εργασιών.

5.1.5 Κατηγορίες θορύβου και προετοιμασία των συνόλων εκπαίδευσης και ελέγχου

Η υψηλού επιπέδου ταξινόμηση του περιβαλλοντικού θορύβου σε διακριτούς ήχους εξαρτάται από πολλαπλά κριτήρια. Ο ρυθμός μεταβολής της έντασης επιτρέπει την διάκριση σε *συνεχείς* θορύβους, δηλαδή σχετικά σταθερής έντασης, *διακοπτόμενους*, όπου η ένταση αυξάνεται και μειώνεται γρήγορα, *παλμικούς*, που χαρακτηρίζονται από απότομη μεταβολή της έντασης και θορύβους *χαμηλής συχνότητας*. Ένα άλλο κριτήριο είναι η προέλευση, δηλαδή αν ο ήχος προέρχεται από ανθρωπογενείς ή άλλες δραστηριότητες.

Στην παρούσα εργασία εξετάζονται οκτώ διαφορετικοί τύποι θορύβου, οι οποίοι εμφανίζονται συχνά στις αστικές περιοχές και επιβαρύνουν την αστική ηχορρύπανση (Πίνακας 5.1).

Πίνακας 5.1: Οι κλάσεις των εξεταζόμενων θορύβων, ο ρυθμός μεταβολής της έντασής τους και η προέλευσή τους

Κλάσεις	**Όνομα κλάσης**	**Ρυθμός αλλαγής έντασης**	**Δραστηριότητα/Προέλευση**
Κόρνα αυτοκινήτου	1	Παλμικός	Μετακίνηση
Παιδιά που παίζουν	2	Διακοπτόμενος	Ανθρώπινη δραστηριότητα
Γάβγισμα σκύλου	3	Διακοπτόμενος	Ζώα
Τρυπάνι	4	Συνεχής	Βιομηχανική
Ρελαντί κινητήρα	5	Συνεχής	Μετακίνηση
Κομπρεσέρ	6	Διακοπτόμενος	Βιομηχανική
Σειρήνα	7	Διακοπτόμενος	Μετακίνηση
Μουσική δρόμου	8	Διακοπτόμενος	Ανθρώπινη δραστηριότητα

Χρησιμοποιήθηκε το σύνολο δεδομένων ήχου UrbanSound8K [49], το οποίο αποτελεί υποσύνολο της βάσης δεδομένων Freesound [168] με περισσότερους από 400000 ήχους. Οι επιλεγμένοι ήχοι είναι ηχογραφήσεις σκηνών δρόμου με διάφορα επίπεδα κυκλοφορίας και ανθρώπινης δραστηριότητας. Τα ηχητικά αποσπάσματα έχουν ηχογραφηθεί σε διαφορετικές υπαίθριες τοποθεσίες σε κατοικημένες περιοχές και στο κέντρο της πόλης.

Με βάση τα σύνολα ήχου εκπαίδευσης δημιουργήθηκε ένα ενιαίο ηχητικό αρχείο διάρκειας 64 λεπτών, με κάθε κλάση ήχου να διαρκεί συνολικά οκτώ λεπτά. Ο ρυθμός δειγματοληψίας είναι τα 44.1 kHz. Οι ηχογραφήσεις κάθε κλάσεις έγιναν υπό διαφορετικές συνθήκες. Το 80% του ηχητικού αρχείου αποτελεί το αρχείο εκπαίδευσης και το υπόλοιπο 20% το αρχείο ελέγχου. Οι κλάσεις ήχου ανακατεύθηκαν ανά δευτερόλεπτο ώστε να μην εμφανίζεται η ίδια κλάση σε διαδοχικά δευτερόλεπτα.

5.1.6 Προφίλ ηχητικών συμβάντων

Η ανάλυση περιεχομένου του ήχου [29] βασίζεται στην εξαγωγή περιγραφικών χαρακτηριστικών ήχου για την αυτόματη αναγνώριση. Τα περιγραφικά χαρακτηριστικά είναι φασματικά, χρονικά και αντιληπτά (σχετιζόμενα με τον ανθρώπινο ακουστικό σύστημα). Για την ανάλυση του ήχου σε αυτή την μελέτη, το σύνολο των χαρακτηριστικών αποτελείται από 8 χρονικά χαρακτηριστικά, 11 φασματικά και 4 αντιληπτά. Τα χρονικά χαρακτηριστικά περιλαμβάνουν το Time-RMS, το Standard Deviation, και το Zero Crossing Rate [169]. Τα φασματικά χαρακτηριστικά περιλαμβάνουν το Slope (την κλίση της μείωσης της φασματικής περιβάλλουσας), το Skewness [31], [170], το Centroid [32], το Crest και το Spread [171].

Θεωρήθηκε επίσης το ηχητικό φάσμα στην περιοχή συχνοτήτων [133, 6854]Hz ως 40 Mel Frequency Cepstral Coefficients (MFCC). Η κλίμακα συχνοτήτων Mel μοντελοποιεί το ανθρώπινο ακουστικό σύστημα αντιστοιχίζοντας την πραγματική συχνότητα στο αντιληπτό τονικό ύψος [159], [170], [172]. Για παράδειγμα το Mel-3 ο οποίος περιλαμβάνεται στον Πίνακα 5.2 αφορά τους συντελεστές της κλίμακας Mel για το εύρος συχνοτήτων [267, 400]Hz. Συνολικά εξήχθησαν 62 τιμές

χαρακτηριστικών για μη επικαλυπτόμενα παράθυρα μήκους ενός δευτερολέπτου. Στον Πίνακα 5.2 παρουσιάζονται οι μέσες τιμές και οι συντελεστές διακύμανσης (coefficient of variance – CoV) για 10 χαρακτηριστικά ανά κλάση. Ο συντελεστής διακύμανσης (επίσης γνωστός και ως σχετική τυπική απόκλιση) ορίζεται ως ο λόγος της τυπικής απόκλισης προς την μέση τιμή (Σχέση 5.1). Η μέση τιμή κάθε χαρακτηριστικού έχει υπολογιστεί πριν την κανονικοποίηση. Οι συντελεστές CoV υπολογίζονται για την σύγκριση των κατανομή των τιμών [173].

$$CoV = \frac{s}{|\bar{x}|}\text{, όπου } s \text{ είναι η τυπική απόκλιση και } |\bar{x}| \text{ η απόλυτη τιμή του μέσου όρου} \qquad (5.1)$$

Οι μεγάλες τιμές του CoV υποδηλώνουν διασπορά γύρω από την μέση τιμή και οφείλονται στην επιλογή μεγάλης ποικιλίας ήχων ανά κλάση (π.χ. ηχογραφήσεις πολλών διαφορετικών σειρήνων, κόρνες κ.λπ.).

Πίνακας 5.2: Η μέση τιμή και ο συντελεστής διακύμανσης 10 χαρακτηριστικών ανά κλάση

Κλάσεις		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Χαρακτηριστικό									
Skewness	$\|\bar{x}\|$	4.84	5.12	6.25	3.76	7.30	2.76	5.27	5.77
	CoV	21.5	16.5	19.6	19.4	24.4	17.7	25.8	18.4
Centroid	$\|\bar{x}\|$	1081	965	825	1668	303	3985	1169	543
	CoV	59.5	50.4	54.8	36.9	91.2	6.5	30.9	60.7
Time ZCR	$\|\bar{x}\|$	0.06	0.07	0.05	0.09	0.04	0.20	0.07	0.04
	CoV	40.7	24.9	43.0	20.6	57.3	8.2	31.4	53.6
Crest	$\|\bar{x}\|$	0.15	0.15	0.22	0.10	0.26	0.06	0.16	0.19
	CoV	34.3	24.2	30.3	21.3	37.8	16.5	48.5	33.8
Kurtosis	$\|\bar{x}\|$	29	32	45	17	63	10	34	40
	CoV	48.7	34.1	39.4	50.1	40.4	40.9	49.8	32.9
Spread	$\|\bar{x}\|$	767	817	473	861	598	2370	801	598
	CoV	30.7	23.4	66.1	17.4	57.5	8.5	58.7	59.6
Decrease	$\|\bar{x}\|$	-0.31	-0.29	-0.31	-0.09	-1.78	-0.01	-0.06	-0.58
	CoV	199	150	297	267	62	111	246	109
Time std	$\|\bar{x}\|$	0.15	0.12	0.14	0.14	0.19	0.07	0.15	0.15

	CoV	28.6	27.3	31.1	10.1	31.8	13.3	29.3	29.9
Time RMS	$\|\bar{x}\|$	-16.9	-19.5	-19.3	-17.2	-15.1	-23.6	-17.4	-17.3
	CoV	16.2	17.4	16.1	10.1	21.6	5.0	18.9	15.2
Mel 3	$\|\|\bar{x}\|$	2.06	1.84	3.15	0.99	2.11	-0.59	2.27	3.13
	CoV	46.5	40.8	38.9	115	10.9	32.7	29.5	41.2

5.1.7 Μηχανική μάθηση για ταξινόμηση

Ο αλγόριθμος MM kNN είναι μία ευρέως χρησιμοποιούμενη μέθοδος ταξινόμησης η οποία βασίζεται σε μετρικές απόστασης. Ένα σημείο δοκιμής αναπαρίσταται ως ένα διάνυσμα n διαστάσεων, όπου n είναι ο αριθμός των παραμέτρων που το χαρακτηρίζουν. Η κλάση του αποφασίζεται με βάση την απόσταση από τους γείτονές του. Για την διαμόρφωση του αλγορίθμου χρειάζεται να αποφασιστούν οι ακόλουθες παράμετροι:

- Η μετρική της απόστασης που θα εφαρμοστεί (ενδεικτικά, Ευκλείδεια, Chebyshev και Συνημίτονου)
- Ο αριθμός των γειτόνων που θα ληφθούν υπόψη

Η Ευκλείδεια απόσταση υπολογίζεται από το άθροισμα των αποστάσεων των επιμέρους συντεταγμένων, ενώ η απόσταση Chebyshev είναι το μέγιστο αυτών των επιμέρους αποστάσεων. Η απόσταση του Συνημίτονου υπολογίζεται από το συνημίτονο της γωνίας μεταξύ των επιμέρους διανυσμάτων μέσω του εσωτερικού γινομένου και αποτελεί μέτρο ομοιότητας όταν τα διανύσματα είναι μεταξύ τους παράλληλα ή αντίστροφα μέτρο ανομοιότητας όταν τα διανύσματα είναι μεταξύ τους κάθετα [174].

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{n}(r_i - t_i)^2} \tag{5.2}$$

$$Chebyshev\ distance = max_i\{|r_i - t_i|\} \tag{5.3}$$

$$Cosine\ distance = \left(1 - \frac{\vec{r_i}-\vec{t_i}}{|\vec{r_i}||\vec{t_i}|}\right) \tag{5.4}$$

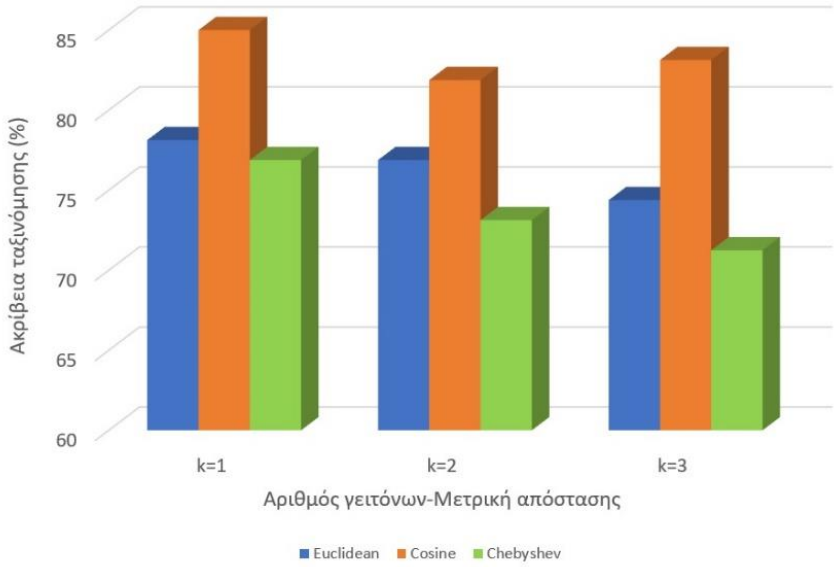
Όπου r_i και t_i είναι το διάνυσμα της εκπαίδευσης και το διάνυσμα ελέγχου αντίστοιχα σε κάθε χρονικό παράθυρο i. Κάθε σημείο εκπαίδευσης και δοκιμής έχει n -συντεταγμένες $(x_1, x_2, x_3, \dots, x_n)$ οι οποίες προκύπτουν από τις τιμές των εξαγόμενων χαρακτηριστικών. Στην προκειμένη περίπτωση μελέτης είναι $n = 23$.

Η κλάση που προβλέπεται καθορίζεται σύμφωνα με την πιο συχνά εμφανιζόμενη κλάση των γειτόνων, όταν αυτοί είναι περισσότεροι από ένας. Στην εργασία αυτή ο αλγόριθμος ρυθμίστηκε ώστε να χρησιμοποιεί και τις τρεις μετρικές απόστασης και $k = 1, 2$ και 3 γείτονες, με αποτέλεσμα συνολικά εννέα μοντέλα.

5.1.8 Αποτελέσματα

Τα αποτελέσματα συνοψίζονται στο Σχήμα 5.2. Το υψηλότερο ποσοστό πρόβλεψης (85%) επιτεύχθηκε με την χρήση $k = 1$ γείτονα με την μετρική της απόστασης να είναι αυτή του

Συνημίτονου. Η μετρικής του Συνημίτονου αποδίδει καλύτερα σε όλες τις περιπτώσεις αριθμού των γειτόνων, ενώ συγκρίνοντας τον αριθμό των γειτόνων η χρήση ενός γείτονα δίνει τα καλύτερα αποτελέσματα.



Σχήμα 5.2: Συγκριτική απεικόνιση της ακρίβειας ταξινόμησης για κάθε μετρική απόστασης και αριθμό γειτόνων.

Για την αξιολόγηση των αποτελεσμάτων λαμβάνουμε υπόψη το *αληθώς θετικό* ποσοστό, όπου ο ταξινομητής αναγνωρίζει σωστά την κλάση και το *αληθώς αρνητικό* ποσοστό, όπου ο ταξινομητής απορρίπτει σωστά το δείγμα. Στο *ψευδώς θετικό* ο ταξινομητής αναγνωρίζει λανθασμένα το δείγμα ως μία κλάση και στο *ψευδώς αρνητικό* ο ταξινομητής απορρίπτει λανθασμένα την κλάση. Ο πίνακας σύγχυσης (confusion matrix) για το μοντέλο με την υψηλότερη απόδοση (1 γείτονας και μετρική απόστασης αυτή του συνημίτονου) παρουσιάζεται στο Σχήμα 5.3.

		Predicted Class									
		Car horn	Children playing	Dog bark	Drilling	Engine idling	Jack hammer	Siren	Street music	True Positive	False Negative
True Class	Car horn	100%								100%	
	Children playing		100%							100%	
	Dog bark			80%		10%			10%	80%	20%
	Drilling				45%		55%			45%	55%
	Engine idling					100%				100%	
	Jack hammer						100%			100%	
	Siren							100%		100%	
	Street music	10%	40%						50%	50%	50%

Σχήμα 5.3: Πίνακας Σύγχυσης για το μοντέλο με έναν γείτονα και μετρική απόστασης του συνημίτονου.

Για τις κλάσεις κόρνα αυτοκινήτου, παιδικά παιχνίδια, ρελαντί κινητήρα, κομπρεσέρ και σειρήνα προέκυψαν 100% αληθώς θετικά αποτελέσματα, ενώ για το τρυπάνι και τη μουσική δρόμου ταξινομούνται λανθασμένα ως κομπρεσέρ και παιδικά παιχνίδια αντίστοιχα.

5.1.8 Συμπεράσματα

Στην εργασία αυτή διερευνήθηκε ο αστικός θόρυβος ως παράγοντας περιβαλλοντικής ρύπανσης. Αναλύθηκε ένα σύνολο οκτώ διακριτών τύπων θορύβου οι οποίοι είναι συχνά απαντώμενοι στα αστικά τοπία, όπως π.χ. η κόρνα, ο ήχος από κινητήρα οχημάτων, ο ήχος σειρήνας κ.ά. Ξεκινώντας από την τυπική μέτρηση που σχετίζεται με τον θόρυβο, δηλαδή την ένταση που εκφράζεται σε dB, εξετάστηκε ένα σύνολο χαρακτηριστικών οι οποίες βασίζονται στο πεδίο του χρόνου, της συχνότητας και της αντίληψης. Με βάση την εργασία της δημοσίευσης [40] επιλέχθηκαν 23 χαρακτηριστικά. Αναπτύχθηκε αλγόριθμος εξαγωγής των τιμών των ηχητικών χαρακτηριστικών οι οποίες αποτελούν τα δεδομένα εκπαίδευσης του αλγορίθμου MM kNN.

Ο αλγόριθμος kNN ορίζεται με βάση των αριθμό των γειτόνων που λαμβάνονται υπόψη στην απόφαση της ταξινόμησης, και με βάση την μετρική της απόστασης. Διερευνήθηκαν τρεις τιμές αριθμού γειτόνων και τρεις μετρικές απόστασης. Ο συνδυασμός αυτών των τιμών οδήγησε στην δημιουργία συνολικά εννέα μοντέλων τα οποία αξιολογήθηκαν με βάση την ακρίβεια ταξινόμησης που πέτυχαν. Η ακρίβεια ταξινόμησης κυμάνθηκε από 70% έως 85%. Την υψηλότερη απόδοση επέδειξε το μοντέλο που λαμβάνει υπόψη την κλάση ενός γείτονα χρησιμοποιώντας την μετρική απόστασης του συνημίτονου.

Το σύνολο της εργασίας παρουσιάστηκε στο ετήσιο διεθνές συνέδριο TMREES 2020 όπου και διακρίθηκε με τον τίτλο “Best Paper Award”. Η έρευνα αυτή επίσης παρουσιάζεται στην δημοσίευση [175].

5.2 Ανίχνευση πυρκαγιάς και σημασιολογική κατάτμηση σε πραγματικό χρόνο με χρήση ΣΝΔ σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων

5.2.1 Εισαγωγή

Η έρευνα που αναπτύχθηκε σχετικά με την ποιότητα των χαρακτηριστικών εικόνας τα οποία εξάγονται είτε με handcrafted μεθόδους είτε από ΣΝΔ, ως προς την απόδοσης ταξινόμησης σε συνδυασμό με τον απαιτούμενο υπολογιστικό χρόνο, μπορεί να έχει άμεσο αντίκτυπο σε πλήθος εφαρμογών και κυρίως σε αυτές του πραγματικού χρόνου. Μία από αυτές σχετίζεται με την έντονη κλιματική αλλαγή που συντελείται στις μέρες μας, και η οποία συνδέεται άμεσα με την αυξανόμενη συχνότητα εμφάνισης πυρκαγιών. Αυτή η αυξανόμενη εμφάνιση πυρκαγιών, σε συνδυασμό με τη δυσκολία αντιμετώπισης αυτών σε περιπτώσεις απομακρυσμένων, υπαίθριων και δασικών περιοχών οδηγούν σε τρόπους έγκαιρης ανίχνευσης (ιδανικά σε πραγματικό χρόνο) αυτών. Παράλληλα, η ανάλυση κινδύνου με δεδομένα τα οποία σχετίζονται με τις υποδομές του περιβάλλοντα χώρου μπορεί να επιτρέψει την λήψη έξυπνων αποφάσεων. Ωστόσο, η ανίχνευση πυρκαγιάς ή/και καπνού αποτελεί ένα δύσκολο έργο επεξεργασίας εικόνας δεδομένης της δυναμικής φύσης αυτών των φαινομένων. Προς αυτή την κατεύθυνση διατίθενται νέες προσεγγίσεις τηλεπισκόπησης, συμπεριλαμβανομένων συστημάτων που βασίζονται σε αισθητήρες εδάφους, συστημάτων, επανδρωμένων και μη, εναέριων οχημάτων και συστημάτων που βασίζονται σε εικόνες που λαμβάνονται από δορυφόρους. Ενώ ένα πλούσιο σύνολο αισθητήρων (θερμοκρασίας, καπνού, υπέρυθρου φωτός κ.λπ.) μπορεί να χρησιμοποιηθεί αποτελεσματικά σε περιορισμένα περιβάλλοντα, η ανάγκη για εκτεταμένη κάλυψη δημιουργεί προκλήσεις σε περιπτώσεις απομακρυσμένων δασικών περιοχών. Στην συνέχεια, λοιπόν, μελέτης της συγκεκριμένης περίπτωσης, θεωρείται ότι τα συστήματα ανίχνευσης είναι εξοπλισμένα με

υπολογιστικούς πόρους που επιτρέπουν την επιτόπια επεξεργασία του διαθέσιμου οπτικού υλικού.

5.2.2 Στόχοι

Η έγκαιρη ανίχνευση και αναγνώριση πυρκαγιάς ανήκει στην οικογένεια κρίσιμων εφαρμογών οι οποίες περιλαμβάνουν την λήψη αποφάσεων σε πραγματικό χρόνο. Η δυναμική φύση του φαινομένου έχει σαν αποτέλεσμα οι εικόνες να λαμβάνονται ακανόνιστα στον χρόνο, η ποιότητά τους να ποικίλλει λόγω του μεταβαλλόμενου οπτικού πεδίου που επηρεάζεται από την απόσταση, τυχών εμποδίων ή ακόμα και από την εφαρμοζόμενη συμπίεση των εικόνων. Επίσης τα ίδια τα συστήματα λήψης εικόνων (π.χ. μη επανδρωμένα εναέρια οχήματα, drones κ.λπ.) έχουν τον κατάλληλο εξοπλισμό ώστε να επεξεργάζονται τις εικόνες. Θεωρείται λοιπόν, ότι οι εικόνες που λαμβάνονται είναι μέτριας ή και χαμηλής ποιότητας και ότι τα συστήματα τα οποία λαμβάνουν και επεξεργάζονται τις εικόνες διαθέτουν περιορισμένους υπολογιστικούς πόρους.

Οι ιδιαιτερότητες του συγκεκριμένου προβλήματος αφορούν α) την ύπαρξη περιορισμένου αριθμού βάσεων δεδομένων (ιδίως αν ληφθεί υπόψη ότι τα χαρακτηριστικά των συνόλων δεδομένων μπορούν να προσανατολίσουν ή και να καθορίσουν την προσέγγιση για την επίλυση του προβλήματος, αλλά και να επηρεάσουν άμεσα την απόδοση των αλγορίθμων) και β) την ανάγκη περιορισμού των απαιτούμενων υπολογιστικών πόρων. Οι στόχοι αυτής της εργασίας είναι:

- Η εξέταση ενός ευρύ συνόλου δεδομένων εικόνων τα οποία περιλαμβάνουν εικόνες διαφορετικής ανάλυσης και συμπίεσης, και διαφορετικής έκτασης πυρκαγιάς
- Η ανάπτυξη αλγορίθμων ταξινόμησης εικόνας με ελαφριά ΣΝΔ, δεδομένου ότι η διαδικασία εκτελείται από συστήματα περιορισμένων υπολογιστικών πόρων
- Η αξιολόγηση των αποτελεσμάτων αυτών των ΣΝΔ ως προς την ακρίβεια ταξινόμησης και η σύγκριση των αντίστοιχων με μεγαλύτερο ΣΝΔ
- Η αξιολόγηση των συστημάτων σε εικόνες που δεν έχουν συμπεριληφθεί στην εκπαίδευση των δικτύων (cross-dataset evaluation) προκειμένου να εκτιμηθεί η ευελιξία αυτών σε διάφορες περιπτώσεις εικόνων
- Η αξιολόγηση της ευρωστίας των συστημάτων σε εικόνες αλλοιωμένες από θόρυβο
- Η σημασιολογική κατάτμηση των εικόνων για την ανάλυση του κινδύνου δεδομένης της υπάρχουσας υποδομής και πρόσβασης στον εκάστοτε χώρο

5.2.3 Συναφής έρευνα

Τα συστήματα ανίχνευσης πυρκαγιάς, τα οποία βασίζονται στην όραση, έχουν χρησιμοποιήσει μία ποικιλία εξοπλισμού, μηχανισμών και τεχνικών, όπως κάμερες (π.χ. οπτικές κάμερες [176]), ευφυείς τεχνικές (π.χ. νευρωνικά δίκτυα [177]), βελτιστοποίηση σμήνους σωματιδίων [178], ασαφή λογική [179], ασύρματα δίκτυα αισθητήρων και δορυφορικά συστήματα [176], και ρομποτικά συστήματα (π.χ. μη επανδρωμένα εναέρια συστήματα [180], [181]). Η ταξινόμηση εικόνων στην ανίχνευση πυρκαγιών μπορεί να πραγματοποιηθεί με την χρήση παραδοσιακών μεθόδων ή και με την χρήση νευρωνικών δικτύων. Στην πρώτη περίπτωση χρησιμοποιούνται τα χαρακτηριστικά της εικόνας, όπως το χρώμα, η υφή και το σχήμα του καπνού και της φωτιάς. Στην δεύτερη περίπτωση τα νευρωνικά δίκτυα και συγκεκριμένα τα ΣΝΔ, μπορούν να επεξεργαστούν δεδομένα και να επιτύχουν υψηλή ακρίβεια ταξινόμησης [182], [157], [183]. Στο [184], παρουσιάστηκε ένα σύστημα ανίχνευσης δασικών πυρκαγιών με βάση την ΒΜ, χρησιμοποιώντας έναν ταξινομητή πυρκαγιών πλήρους εικόνας και λεπτομερών κηλίδων. Στο [185], παρουσιάζεται

ένα σύστημα ανίχνευσης πυρκαγιών το οποίο εφαρμόζει μία τεχνική επεξεργασίας εικόνας με βασισμένη σε κανόνες και στην χρονική μεταβολή. Στο [186], το σύστημα ανίχνευσης πυρκαγιάς βασίζεται σε βαθύ ΣΝΔ και εφαρμόζεται σε βίντεο. Στο [187] ανέπτυξαν ανιχνευτή πυρκαγιάς ο οποίος ανιχνεύει ακόμη και μικρές σπίθες και ηχεί συναγερμό εντός οκτώ δευτερολέπτων από την εκδήλωση της πυρκαγιάς. Αναπτύχθηκε ένα νέο νευρωνικό δίκτυο συνελικτικού τύπου για την ανίχνευση περιοχών πυρκαγιών με την χρήση ενός βελτιωμένου δικτύου You Only Look Once (YOLO). Οι συγγραφείς στο [188] προτείνουν ανίχνευση σε πολλαπλούς προσανατολισμούς με βάση μία μονάδα μηχανισμού μετατροπής τιμών και Mixed Network Management System (Mixed-NMS). Στο [189], παρουσιάζεται ένας ελαφρύς αλγόριθμος ανίχνευσης πυρκαγιάς, Light-YOLO, επιτυγχάνοντας ταχύτητα ανίχνευσης 91.1 fps. Στο [190], σχεδιάστηκε μία αρχιτεκτονική νευρωνικού δικτύου για την ανίχνευση και αναγνώριση δασικής πυρκαγιάς με βάση το Attention U-Net και το SqueezeNet (ATT Squeeze U-Net), εκτελώντας τμηματοποίηση για την εξαγωγή σχήματος της πυρκαγιάς και ταξινόμηση για την επαλήθευση της ανιχνευμένης περιοχής. Στο [191], πραγματοποιείται ανίχνευση πυρκαγιάς χρησιμοποιώντας την πρόταση ύποπτης περιοχής που γίνεται από ένα ελαφρύ ΣΝΔ το οποίο βασίζεται σε superpixel (συγκεκριμένα το Expanded Squeeze-and Excitation ShuffleNet – ESE-ShuffleNet). Η χρήση των ΣΝΔ εμφανίζεται ως ο κοινός παρονομαστής των πρόσφατων λύσεων που συνυπολογίζουν τον εξορθολογισμό του υπολογιστικού κόστους. Οι ιδιαιτερότητες που σχετίζονται με πολύπλοκα και δυναμικά μεταβαλλόμενα περιβάλλοντα δημιουργούν προκλήσεις, ειδικά λόγω του γεγονότος ότι τα σύνολα δεδομένων έχουν συγκεντρωθεί κάτω από συγκεκριμένες ρυθμίσεις.

5.2.4 Επιλογή ΣΝΔ και Μεταφορά Μάθησης

Η αρχιτεκτονική των ΣΝΔ ορίζεται, μεταξύ άλλων, μέσω του αριθμού των επιπέδων και των συνδέσεων μεταξύ τους. Η επιλογή των κατάλληλων ΣΝΔ αποτελεί ένα δύσκολο πρόβλημα, το οποίο εξαρτάται από την προκειμένη περίπτωση μελέτης, τις εκάστοτε απαιτήσεις και περιορισμούς. Στην προκειμένη περίπτωση μελέτης της ανίχνευσης και αναγνώρισης πυρκαγιάς, οι βασικές απαιτήσεις είναι η ανάγκη για υψηλή ακρίβεια ταξινόμησης, οι δυνατότητες γενίκευσης σε νέα δεδομένα, εξαιτίας του δυναμικά μεταβαλλόμενου γεωγραφικού πλαισίου, και οι περιορισμένοι υπολογιστικοί πόροι. Υπό αυτό το πρίσμα, επιλέχθηκαν τέσσερα ΣΝΔ και συγκεκριμένα α) το SqueezeNet [66] ως απλοποιημένη έκδοση του AlexNet, το οποίο επιτυγχάνει παρόμοια ακρίβεια ταξινόμησης με 50 φορές λιγότερες παραμέτρους, β) το ShuffleNet [74] το οποίο λειτουργεί σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων χρησιμοποιώντας τεχνικές σημειακής ομαδικής συνέλιξης (pointwise group convolution) και ανακατέματος καναλιών (channel shuffle), γ) το MobileNet_v2 [73] το οποίο μπορεί να εφαρμοστεί σε κινητές συσκευές χρησιμοποιώντας τεχνικές διαχωρισμού συνέλιξης κατά βάθος (depth-wise separable convolution) και δ) το ResNet50 [65] ως εκπρόσωπος των μεγαλύτερων ΣΝΔ. Στον Πίνακα 5.3 περιγράφονται τα χαρακτηριστικά των επιλεγμένων ΣΝΔ.

Πίνακας 5.3: Χαρακτηριστικά των επιλεγμένων ΣΝΔ

ΣΝΔ	**SqueezeNet**	**ShuffleNet**	**MobiletNet_v2**	**ResNet50**
Βάθος	18	50	53	50
Παράμετροι (εκατομμύρια)	1.24	1.4	3.5	25.6

Λαμβάνοντας υπόψη τους περιορισμούς στον αριθμό και την ποιότητα των δεδομένων με ετικέτα

(labelled data), το οποίο αποτελεί ένα τυπικό εμπόδιο στην επιβλεπόμενη μάθηση [192], εφαρμόστηκε η τεχνική της μεταφοράς μάθησης για την ανίχνευση πυρκαγιάς και εν συνεχεία για την σημασιολογική κατάτμηση και ανίχνευση αντικειμένων ενδιαφέροντος. Τα επιλεγμένα ΣΝΔ (τα οποία είναι προεκπαιδευμένα στο ImageNet) επανεκπαιδεύονται με τον συγκεκριμένο στόχο της ανίχνευσης πυρκαγιάς. Κατά την επανεκπαίδευση υπολογίζονται η απώλεια εκπαίδευσης (η αρνητική λογαριθμική πιθανότητα – negative log likelihood-NLL), καθώς και οι κλίσεις ανά παράμετρο του μοντέλου (backpropagation) και χρησιμοποιούνται για την ενημέρωση των παραμέτρων με τον βελτιστοποιητή (optimizer). Οι παράμετροι επανεκπαίδευσης είναι ένα ανοιχτό ζήτημα που συμπεριλαμβάνει τον αλγόριθμο του βελτιστοποιητή, τον αριθμό των εποχών, τον ρυθμό μάθησης κ.ά.

Για να μπορέσει το πλαίσιο να εφαρμοστεί σε κάθετες εφαρμογές, όπως π.χ. η προστασία των ενεργειακών υποδομών, πρέπει να αναγνωριστούν σχετικά αντικείμενα όπως κτίρια, δρόμοι, γραμμές μεταφοράς ηλεκτρικής ενέργειας. Η ανίχνευση αντικειμένων δεν μπορεί να εφαρμόζεται συνεχώς αλλά μόνο στην περίπτωση που η αναγνώριση πυρκαγιάς είναι θετική. Τα δίκτυα Faster R-CNN, You Only Look Once (YOLO) v2-4 και Single Shot Detection (SSD) εστιάζουν στην σημασιολογική κατάτμηση. Στην παρούσα εργασία χρησιμοποιήθηκε το δίκτυο Deeplab v3+ [193] το οποίο είναι προεκπαιδευμένο στην βάση δεδομένων CamVid [194] και επιλέχθηκαν ένα σύνολο από ενδιαφέροντα αντικείμενα.

5.2.5 Βάσεις δεδομένων

Τα σύνολα δεδομένων διαφέρουν ως προς το είδος και τον αριθμό των εικόνων που περιέχουν, την προέλευση των εικόνων, την προκατάληψη (bias) και την ομοιογένειά τους (εάν για παράδειγμα το υλικό έχει δημιουργηθεί για τον συγκεκριμένο σκοπό ή εάν έχουν αναζητηθεί από υπάρχουσες βάσεις δεδομένων), τα τεχνικά χαρακτηριστικά και την επιμέλεια των εικόνων. Η έρευνα για σύνολα δεδομένων εικόνων που σχετίζονται με την αναγνώριση πυρκαγιών απέδωσε σχετικά περιορισμένα αποτελέσματα, καθώς στα υπάρχοντα σύνολα δεδομένων η απεικόνιση πυρκαγιών μεγάλης κλίμακας καθιστά την αναγνώριση εύκολη και επιπλέον δεν είναι κατάλληλη για την εκπαίδευση ανίχνευσης πυρκαγιών οι οποίες βρίσκονται σε πρώιμο στάδιο. Επί παραδείγματι, η βάση δεδομένων FLAME [195] εξαιτίας της σαφούς απεικόνισης της φωτιάς η ακρίβεια ταξινόμησης μεγιστοποιείται. Από αυτή την σκοπιά, λοιπόν, επιλέχθηκαν τα σύνολα εικόνων α) Forest-Fire dataset [196] που περιέχει 1900 εικόνες από υπάρχουσες βιβλιοθήκες, δύο κλάσεων (φωτιά, χωρίς-φωτιά), β) Fire-Flame dataset [197] που περιέχει 3000 εικόνες, τριών κλάσεων (φωτιά, χωρίς-φωτιά, καπνός) που για λόγους ομοιομορφίας της σύγκρισης η κλάση του καπνού εξαιρέθηκε και γ) Άγνωστες εικόνες, το οποίο είναι ένα σύνολο εικόνων που δεν έχουν συμμετάσχει στην διαδικασία της εκπαίδευσης των δικτύων και έχουν συλλεχθεί από διάφορες δημόσια διαθέσιμες πηγές με 100 εικόνες για την κάθε κλάση. Στον Πίνακα 5.4 παρουσιάζονται τα χαρακτηριστικά των επιλεγμένων συνόλων εικόνων. Το Σχήμα 5.4 έχει δείγματα των εικόνων από κάθε βάση δεδομένων.

Πίνακας 5.4: Τα σύνολα εικόνων και τα χαρακτηριστικά τους

Σύνολο δεδομένων	**Forest-Fire**	**Fire-Flame**	**Άγνωστες εικόνες**
Κλάσεις	Φωτιά – Χωρίς φωτιά	Φωτιά – Χωρίς φωτιά	Φωτιά – Χωρίς φωτιά
Μέσα λήψης	Επίγεια και εναέρια	Επίγεια και εναέρια	Επίγεια και εναέρια
Τοποθεσία	Δάσος	Παντού	Παντού

Ανάλυση	250×250	Ανομοιογενής	Ανομοιογενής
Προβολή	Μπροστά και πάνω	Μπροστά	Μπροστά και πάνω
Αριθμός εικόνων	1900	2000	200
Κατανομή στις κλάσεις	Ομοιόμορφη	Ομοιόμορφη	Ομοιόμορφη

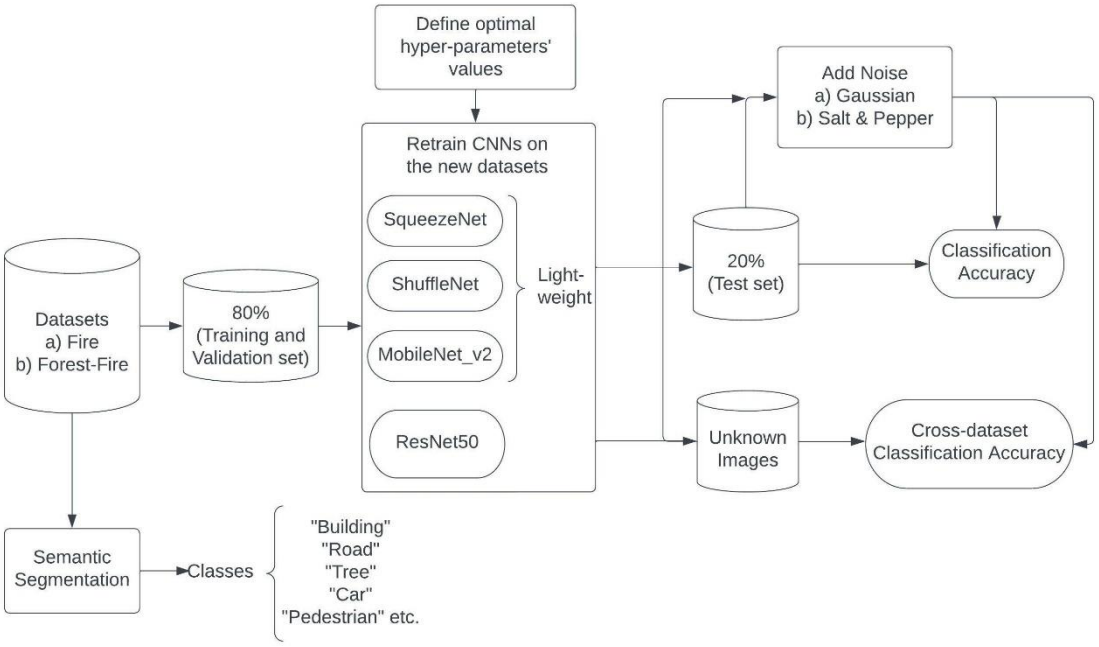
Σχήμα 5.4: Δείγματα εικόνων από τα τρία σύνολα δεδομένων.

5.2.6 Υλοποίηση

Με βάση τα δύο επιλεγμένα σύνολα εικόνων (Forest-Fire και Fire-Flame) και χρησιμοποιώντας τα προεκπαιδευμένα νευρωνικά δίκτυα που αναφέρονται στην Ενότητα 5.2.3 (SqueezeNet, ShuffleNet, MobileNet_v2 και ResNet50) εκτελέστηκαν πειράματα με βάση την μεταφορά μάθησης. Κάθε σύνολο χωρίζεται σε σύνολο εκπαίδευσης και σύνολο ελέγχου σε ποσοστά 80% και 20% αντίστοιχα. Στο πρώτο πείραμα διερευνώνται οι τιμές των παραμέτρων επανεκπαίδευσης αξιολογώντας την ακρίβεια ταξινόμησης. Αφού προσδιοριστούν οι βέλτιστες τιμές των παραμέτρων και το αντίστοιχο μέγιστο της ακρίβειας ταξινόμησης, οι εικόνες υποβάλλονται σε θόρυβο Gaussian και Salt & Pepper, σε δύο διαφορετικά επίπεδα σηματοθορυβικού λόγου. Έτσι αναγνωρίζεται ποιο δίκτυο είναι πιο ανθεκτικό στον θόρυβο και ποιος θόρυβος επιδρά πιο καταστροφικά στα αποτελέσματα.

Για να προσεγγιστούν σενάρια πραγματικού χρόνου, πραγματοποιούνται επίσης αξιολογήσεις σε διασταυρούμενα σύνολα δεδομένων: τα επανεκπαιδευμένα δίκτυα ελέγχονται σε σύνολο εικόνων οι οποίες δεν συμμετείχαν στην εκπαίδευση των δικτύων. Το σύνολο των εικόνων που δεν συμπεριλαμβάνονται στο σετ εκπαίδευσης (δηλαδή οι «Άγνωστες Εικόνες») χρησιμοποιείται ως σύνολο ελέγχου. Στις εικόνες αυτού του συνόλου η φωτιά καταλαμβάνει διαφορετικό τμήμα της εικόνας συμπεριλαμβανομένης της περίπτωσης που η πυρκαγιά βρίσκεται σε πρώιμο στάδιο (μικρή έκταση). Επίσης, οι εικόνες ελέγχονται αφού αλλοιωθούν με θόρυβο. Παράλληλα, οι εικόνες

αναλύονται με σημασιολογική κατάτμηση συσχετίζοντας κάθε pixel της εικόνας με μία κατηγόρια αντικειμένου εντοπίζοντας με αυτόν τον τρόπο περιοχές ενδιαφέροντος (όπως π.χ. δρόμοι, κτίρια, άνθρωποι, οχήματα κ.λπ.). Η ροή εργασιών περιγράφεται στο Σχήμα 5.5.



Σχήμα 5.5: Ροή εργασιών.

5.2.7 Τιμές παραμέτρων επανεκπαίδευσης

Ο καθορισμός των τιμών των υπερπαραμέτρων στην μεταφορά μάθησης σχετίζεται με την προσαρμογή των βαρών με την χρήση της οπισθοδιάδοσης (back-propagation) ώστε να μεγιστοποιηθεί η ακρίβεια ταξινόμησης και να ελαχιστοποιηθεί η συνάρτηση απώλειας. Εξετάστηκαν οι αλγόριθμοι βελτιστοποίησης (optimizers) SGDM [82] και Adam [83] με τον πρώτο να επιτυγχάνει καλύτερα αποτελέσματα. Στην συνέχεια, διερευνήθηκαν τέσσερις τιμές για το μέγεθος του mini-batch (50, 100, 200, και 300) το οποίο είναι ο διαιρέτης του αριθμού των αρχείων του συνόλου εκπαίδευσης και το αντίστοιχο πηλίκο που προκύπτει είναι ο αριθμός των επαναλήψεων που πραγματοποιούνται για την προσαρμογή των βαρών. Ο μέγιστος αριθμός των epochs είναι τα πλήρη περάσματα από το σύνολο εκπαίδευσης και το validation patience είναι ο αριθμός των epochs μετά τον οποίο η απώλεια σταματά να μειώνεται. Ο ρυθμός μάθησης (learning rate) αναφέρεται στο μέγεθος των διορθωτικών βημάτων στην διαδικασία ενημέρωσης των βαρών. Οι συνδυασμοί που προκύπτουν από τους δύο optimizers και τις τέσσερις τιμές του mini-batch size (συνολικά οχτώ συνδυασμοί) δοκιμάστηκαν επαναληπτικά, ενώ ο μέγιστος αριθμός των epochs ορίστηκε σε μία ευέλικτη τιμή η οποία συνδυάζεται με την τιμή του validation patience. Το learning rate ορίστηκε σε χαμηλή τιμή έτσι ώστε να πραγματοποιούνται μικρά βήματα ενημέρωσης των βαρών, το οποίο καθυστερεί μεν την εκπαίδευση αλλά οδηγεί σε βέλτιστες τιμές των βαρών.

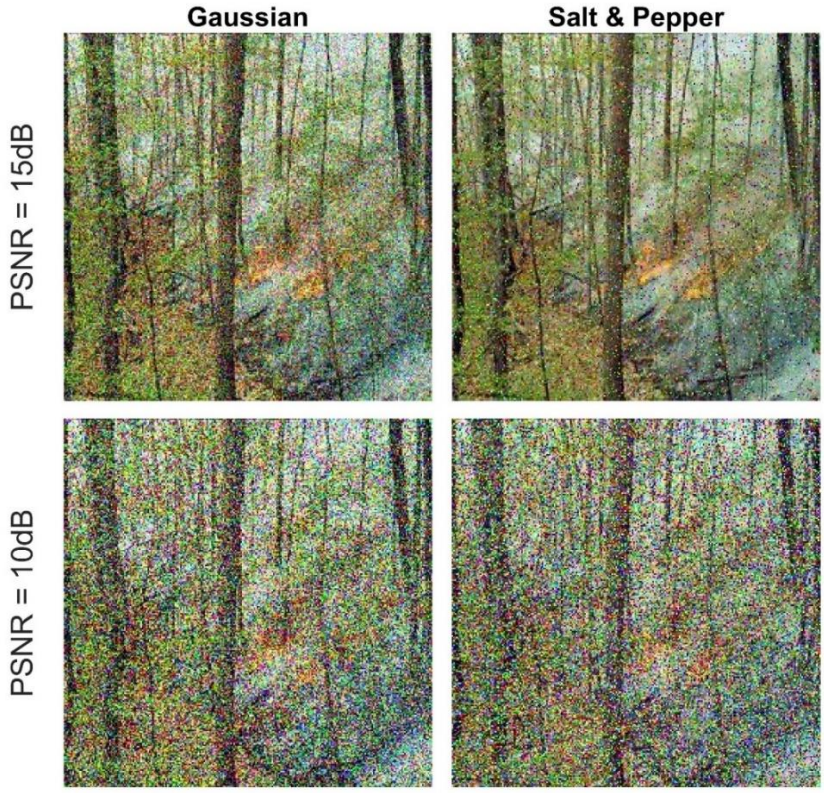
Συγκεκριμένα, οι τιμές των υπερπαραμέτρων ορίστηκαν ως εξής:

Optimizer: SGDM, Mini-batch size: 100, Maximum Epochs: 10 (σε όλες τις δοκιμές η διαδικασία εκπαίδευσης τερματίστηκε πριν το 10^{ο} epoch), Validation Patience: 2, Learning Rate: 2×10^{-3}. Επιπρόσθετα, οι εικόνες ανακατεύτηκαν ώστε σε κάθε epoch η ενημέρωση των βαρών να λαμβάνει υπόψη διαφορετικό σύνολο εικόνων καθώς επίσης πραγματοποιήθηκαν και πράξεις επαύξησης

προκειμένου να αποφευχθεί η υπερπροσαρμογή.

5.2.8 Θόρυβος

Οι εναέριες και επίγειες εικόνες είναι πιθανό να επηρεαστούν από θόρυβο κατά την λήψη (π.χ. λόγω του χρησιμοποιούμενου διαφράγματος και των παραμέτρων της λήψης), την συμπίεση και την μετάδοση. Αυτό σημαίνει ότι η αναγνώριση της πυρκαγιάς μπορεί να χρειαστεί να γίνει σε εικόνες χαμηλής ποιότητας ή αλλοιωμένες από θόρυβο. Τύποι συχνά απαντώμενων θορύβων σε τέτοιες περιπτώσεις είναι ο Gaussian [151], ο παλμικός (impulse) [152], θόρυβος κβαντοποίησης (quantization noise), έλλειψης δειγμάτων (missing image samples), απώλεια πακέτων κατά την μετάδοση (packet loss in transmission), παραποιημένες εικόνες (tampered images) κ.ά. Το ανθρώπινο οπτικό σύστημα μπορεί να αγνοήσει αυτούς του θορύβους, αλλά η απόδοση των ΣΝΔ επηρεάζεται. Λαμβάνοντας υπόψη ότι όταν τα μοντέλα ταξινομούν δεδομένα της ίδιας ποιότητας με τα δεδομένα εκπαίδευσης επιτυγχάνουν υψηλότερη ακρίβεια, επιλέχθηκε να δοκιμαστούν τα μοντέλα σε αλλοιωμένες από θόρυβο εικόνες, ενώ έχουν εκπαιδευτεί σε καθαρές προκειμένου να εκτιμηθεί το κατώτερο όριο της ακρίβειας ταξινόμησης. Οι εικόνες υποβλήθηκαν σε Gaussian και Salt & Pepper θορύβους ρυθμίζοντας τις παραμέτρους του κάθε τύπου ώστε να προκύψει ίδιος σηματοθορυβικός λόγος (PSNR) στα 15dB και10dB. Η επίδραση των δύο τύπων θορύβου στα δύο επίπεδα PSNR παρουσιάζεται στο Σχήμα 5.6.



Σχήμα 5.6: Απεικόνιση της επίδρασης των δύο τύπων θορύβου σε δύο επίπεδα PSNR.

5.2.9 Σημασιολογική κατάτμηση

Η σημασιολογική κατάτμηση είναι μία πρόσθετη διαδικασία η οποία σε συνδυασμό με την ανίχνευση της πυρκαγιάς υποστηρίζει τον προσδιορισμού του επιπέδου του κινδύνου (risk analysis) ανάλογα με την περιοχή στην οποία βρίσκεται η πυρκαγιά. Η ανίχνευση αντικειμένων ενδιαφέροντος μπορεί να δώσει μία ένδειξη σχετικά με τον τύπο της περιοχής και τις υποδομές οι οποίες βρίσκονται σε κίνδυνο. Χρησιμοποιήθηκε το δίκτυο Deeplab v3+ με βάση το ResNet18 (δηλαδή τα αρχικά του βάρη είναι ίδια με αυτά του ResNet18) το οποίο επανεκπαιδεύτηκε στο σύνολο δεδομένων CamVid. Αυτό το σύνολο δεδομένων αποτελείται από εικόνες σε επίπεδο γης με ετικέτες εικονοστοιχείων (pixel-level labels) 32 κατηγοριών, όπως π.χ. κτίριο, πεζός, δέντρο, αυτοκίνητο κ.λπ. Σύμφωνα με τον αριθμό των pixels που περιλαμβάνουν τα σημεία ενδιαφέροντος μπορούν να εξαχθούν τα αντίστοιχα συμπεράσματα. Στο Σχήμα 5.7 παρουσιάζεται ένα παράδειγμα σημασιολογικής κατάτμησης σε μία εικόνα που ταξινομήθηκε επιτυχώς. Παρόλο που η συγκεκριμένη εικόνα δεν απεικονίζει δασική πυρκαγιά επιλέχθηκε για να αναδειχθεί η λειτουργία της σημασιολογικής κατάτμησης δεδομένα ότι περιέχει αρκετά σημεία ενδιαφέροντος.

Σχήμα 5.7: Σημασιολογική κατάτμηση σε εικόνα με φωτιά.

Σύμφωνα με την χρωματική μπάρα στο πλάι, ο πεζός, το κτίριο, ο πυλώνας ηλεκτροδότησης, , δρόμος και τα φυτά ανιχνεύονται επιτυχώς. Αυτές οι κατηγορίες μπορούν να καθορίσουν το επίπεδο της αμεσότητας της επέμβασης αλλά και τον τρόπο αυτής (αφού υπάρχει δρόμος).

5.2.10 Αποτελέσματα

Η ακρίβεια ταξινόμησης και ο χρόνος εκπαίδευσης για κάθε σύνολο δεδομένων και κάθε ΣΝΔ παρουσιάζονται στον Πίνακα 5.5. Οι τιμές των υπερπαραμέτρων έχουν οριστεί στις τιμές που μεγιστοποιείται η απόδοση ταξινόμησης, και σε κάθε περίπτωση, η εκπαίδευση και ο έλεγχος πραγματοποιούνται σε υποσύνολα του ίδιου συνόλου εικόνων. Και τα τέσσερα ΣΝΔ επιτυγχάνουν υψηλά ποσοστά ακρίβειας τα οποία υπερβαίνουν το 95%. Ο χρόνος εκπαίδευσης κάθε δικτύου εξαρτάται από το βάθος του και τον αριθμό των εικόνων που χρησιμοποιούνται στην διαδικασία εκπαίδευσης. Το μεγαλύτερο δίκτυο MobileNet_v2 απαιτεί τον μεγαλύτερο χρόνο εκπαίδευσης, αλλά σε αντάλλαγμα αποδίδει την υψηλότερη ακρίβεια ταξινόμησης και για τα δύο σύνολα εικόνων. Λαμβάνοντας υπόψη ότι η διαδικασία εκπαίδευσης πραγματοποιείται μία φορά, σε αυτό το σημείο της σύγκρισης το MobileNet_v2 αποτελεί την καλύτερη επιλογή. Όσον αφορά τα δύο σύνολα εικόνων, τα αποτελέσματα ταξινόμησης είναι αρκετά παρόμοια με ελαφρώς καλύτερα αποτελέσματα για το σύνολο εικόνων Forest-Fire.

Πίνακας 5.5: Η ακρίβεια ταξινόμησης (%) και χρόνος εκπαίδευσης (s) ανά σύνολο εικόνων και ΣΝΔ με τις βέλτιστες επιλογές παραμέτρων εκπαίδευσης

	Forest-Fire		**Fire-Flame**	
ΣΝΔ	Ακρίβεια ταξινόμησης (%)	Χρόνος εκπαίδευσης (s)	Ακρίβεια ταξινόμησης (%)	Χρόνος εκπαίδευσης (s)
SqueezeNet	97.11	45	95.00	77
ShuffleNet	97.89	68	96.00	90
MobileNet_v2	98.95	151	97.50	164
ResNet50	97.63	166	96.00	175

Τα αποτελέσματα της ακρίβειας ταξινόμησης για την περίπτωση που ο θόρυβος που αλλοιώνει τις εικόνες του σετ ελέγχου είναι ήπιας έντασης φαίνονται στον Πίνακα 5.6 (στις παρενθέσεις αναφέρεται η μείωση της ακρίβειας ταξινόμησης σε σχέση με τις τιμές του Πίνακα 5.5). Ορίστηκε και για τους δύο τύπους θορύβου PSNR περίπου ίσο με 15dB.

Πίνακας 5.6: Η ακρίβεια ταξινόμησης (%) ανά σύνολο εικόνων και ΣΝΔ με αλλοιωμένες εικόνες ελέγχου και επίπεδο θορύβου PSNR = 15dB

(PSNR = 15dB)	**Forest-Fire**		**Fire-Flame**	
ΣΝΔ	Gaussian	Salt & Pepper	Gaussian	Salt & Pepper
SqueezeNet	76.58 (-20.53)	80.53 (-11.06)	87.00 (-8.00)	87.00 (-8.00)
ShuffleNet	80.53 (-17.36)	87.63 (-10.26)	90.00 (-6.00)	91.00 (-5.00)
MobileNet_v2	67.37 (-10.26)	85.26 (-13.69)	82.50 (-15.00)	89.50 (-8.00)
ResNet50	91.20 (-6.43)	94.58 (-3.05)	94.00 (-2.00)	94.50 (-1.50)

Όσον αφορά τους δύο τύπους θορύβου, ο Gaussian προκαλεί μεγαλύτερη μείωση στην ακρίβεια ταξινόμησης από ό,τι ο Salt & Pepper. Όσον αφορά τα δύο σύνολα εικόνων, το Forest-Fire έχει μία μέση μείωση της ακρίβειας ταξινόμησης γύρω στο 19% με Gaussian θόρυβο και γύρω στο 9.7% με Salt & Pepper. Το σύνολο Fire-Flame επηρεάζεται λιγότερο με μέση μείωση της ακρίβειας ταξινόμησης περίπου 7.8% και 5.4% με Gaussian και Salt & Pepper αντίστοιχα. Όσον αφορά τα ΣΝΔ, το ResNet50 είναι πιο ανθεκτικό στον θόρυβο με μέση μείωση της ακρίβειας ταξινόμησης γύρω στο 3.2%, ακολουθεί το ShuffleNet με μέση μείωση γύρω στο 9.7, ενώ το πιο ευάλωτο δίκτυο εμφανίζεται το MobileNet_v2 με μέση μείωση γύρω στο 17.1%. Επομένως, λαμβάνοντας υπόψη τα αποτελέσματα του Πίνακα 5.5 όπου η ακρίβεια ταξινόμησης για τα δίκτυα ShuffleNet, MobileNet_v2 και ResNet50 φαίνεται παραπλήσια, αλλά και την ανοχή στον θόρυβο με τα αποτελέσματα του Πίνακα 5.6, τα δίκτυα ResNet50 και ShuffleNet αποδεικνύονται οι καλύτερες επιλογές.

Τα αποτελέσματα της ακρίβειας ταξινόμησης για υψηλό επίπεδο θορύβου (PSNR = 10dB) παρουσιάζονται στον Πίνακα 5.7. Για τα ελαφριά ΣΝΔ η ακρίβεια ταξινόμησης μειώνεται ραγδαία και κυμαίνεται γύρω στο 50%. Τα αποτελέσματα για το ResNet50 είναι ελαφρώς καλύτερα για την περίπτωση του συνόλου Forest-Fire, ενώ στο σύνολο Fire-Flame ο Gaussian θόρυβος έχει

μικρότερη επίδραση (80.8%) από ό,τι ο Salt & Pepper θόρυβος (72%).

Πίνακας 5.7: Η ακρίβεια ταξινόμησης (%) ανά σύνολο εικόνων και ΣΝΔ με αλλοιωμένες εικόνες ελέγχου και επίπεδο θορύβου PSNR = 10dB

(PSNR = 10dB)	**Forest-Fire**		**Fire-Flame**	
ΣΝΔ	Gaussian	Salt & Pepper	Gaussian	Salt & Pepper
SqueezeNet	50.00 (-47.11)	51.50 (-45.61)	49.50 (-45.50)	55.50 (-39.50)
ShuffleNet	50.26 (-47.63)	53.42 (-44.47)	57.00 (-39.00)	72.50 (-23.50)
MobileNet_v2	50.00 (-48.95)	50.00 (-48.95)	49.00 (-48.50)	49.00 (-48.50)
ResNet50	63.20 (-34.43)	54.10 (-45.53)	80.80 (-15.20)	72.00 (-24.00)

Τα επανεκπαιδευμένα ΣΝΔ αξιολογήθηκαν και σε εικόνες που δεν έλαβαν μέρος στην διαδικασία της εκπαίδευσης (cross-dataset evaluation). Το σύνολο «Άγνωστες Εικόνες» χρησιμοποιείται αρχικά με καθαρές εικόνες και στην συνέχεια με αλλοιωμένες εικόνες με σηματοθορυβικό λόγο PSNR = 10dB. Με αυτόν τον τρόπο εξετάζεται α) αν το ShuffleNet παραμένει η κυρίαρχη επιλογή μεταξύ των ελαφριών ΣΝΔ και β) αν κάποιο από τα δύο σύνολα εικόνων που χρησιμοποιήθηκαν για την εκπαίδευση των δικτύων είναι καταλληλότερο για την εκπαίδευση (δηλαδή περιέχει πιο έντονα διαφοροποιημένες εικόνες). Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.8.

Πίνακας 5.8: Απόδοση ταξινόμησης (%) στο σετ ελέγχου «Άγνωστες Εικόνες» (cross-dataset evaluation) για καθαρές και αλλοιωμένες εικόνες

	Σετ εκπαίδευσης	**Forest-Fire**	**Fire-Flame**
ΣΝΔ	Σετ ελέγχου	Άγνωστες Εικόνες	Άγνωστες Εικόνες
SqueezeNet	Καθαρές εικόνες	88.50	85.50
	Gaussian	77.00	72.00
	Salt & Pepper	67.00	61.00
ShuffleNet	Καθαρές εικόνες	90.00	90.00
	Gaussian	71.50	71.00
	Salt & Pepper	65.50	65.00
MobileNet_v2	Καθαρές εικόνες	86.50	86.00
	Gaussian	71.50	71.00
	Salt & Pepper	65.50	65.00
ResNet50	Καθαρές εικόνες	85.50	85.50
	Gaussian	82.50	79.00
	Salt & Pepper	73.00	71.00

Σύμφωνα με τις τιμές του Πίνακα 5.8, το ShuffleNet αποδίδει την υψηλότερη απόδοση ταξινόμησης για την δοκιμή των διασταυρούμενων συνόλων. Επιβεβαιώνεται επίσης ότι το

ResNet50 χειρίζεται καλύτερα την επίδραση των θορύβων στις εικόνες. Σχετικά με το ποιο σύνολο είναι καταλληλότερο για εκπαίδευση, τα αποτελέσματα είναι συγκρίσιμα με ελαφρώς αυξημένα ποσοστά στην περίπτωση που τα δίκτυα έχουν εκπαιδευτεί με το σύνολο Forest-Fire.

5.2.11 Συμπεράσματα

Οι πυρκαγιές αποτελούν σημαντικό παράγοντα κινδύνου για την υποβάθμιση του περιβάλλοντος και έχουν σοβαρό αντίκτυπο στις ανθρώπινες ζωές και δραστηριότητες. Ο κύριος στόχος είναι η πρόληψη και η έγκαιρη ανίχνευση των πυρκαγιών και για τον σκοπό αυτό έχει σημειωθεί σημαντική πρόοδος όσον αφορά α) τα ηλεκτρομηχανικά μέσα που είναι σε θέση να ανακτούν εικόνες από γεωγραφικά απομακρυσμένες περιοχές και β) τα μέσα μηχανικής και βαθιάς μάθησης (και κυρίως με ΣΝΔ) τα οποία είναι σε θέση να επεξεργάζονται στοιχεία και να εξάγουν πληροφορίες. Αυτές οι προοπτικές μπορούν να συνδυαστούν και να προσφέρουν δυνατότητες σχεδόν σε πραγματικό χρόνο και μάλιστα με περιορισμένους υπολογιστικούς πόρους.

Για τον σκοπό αυτό σχεδιάστηκε ένα σύστημα βασισμένο σε ΣΝΔ ικανό να επεξεργάζεται εικόνες για τον εντοπισμό των πυρκαγιών και να εξάγει πληροφορίες που βασίζονται στο πλαίσιο των σημασιολογικών κατατμήσεων. Καθώς ο βασικός περιορισμός είναι ο περιορισμός στους υπολογιστικούς πόρους, επιλέχθηκαν τρία ελαφριά ΣΝΔ (SqueezeNet, ShuffleNet, και MobileNet_v2) και ένα μεγαλύτερο (ResNet50) για σκοπούς σύγκρισης.

Καθώς μία βασική ιδιαιτερότητα της εφαρμογής που βασίζεται σε μηχανισμούς ΒΜ ήταν η περιορισμένη διαθεσιμότητα συνόλων σχετικών εικόνων, πραγματοποιήθηκε μία σειρά δοκιμών που αφορούν αλλοιωμένες εικόνες (σε δύο επίπεδα PSNR) καθώς και σενάρια διασταυρούμενων συνόλων δεδομένων. Για ήπιο θόρυβο έως τα 15dB, τα ελαφριά ΣΝΔ (και κυρίως το ShuffleNet) φαίνονται προτιμότερα, καθώς επιτυγχάνουν υψηλή ακρίβεια ταξινόμησης και δεν απαιτούν πολλούς υπολογιστικούς πόρους. Εάν ο θόρυβος είναι υψηλότερου επιπέδου και υποθέτοντας ότι οι μηχανισμοί αφαίρεσης του θορύβου μπορεί να βελτιώσουν την ποιότητα της εικόνας για την ανθρώπινη αντίληψη αλλά ο θόρυβος εξακολουθεί να επηρεάζει τους αλγόριθμους ΜΜ, πρέπει να χρησιμοποιηθούν μεγαλύτερα δίκτυα (όσον αφορά τον αριθμό των παραμέτρων). Παρατηρήθηκε επίσης, ότι ο θόρυβος Gaussian σε μέτριο επίπεδο επηρεάζει περισσότερο την ακρίβεια ταξινόμησης, ενώ όταν το επίπεδο θορύβου αυξάνεται τότε ο Salt & Pepper θόρυβος είναι πιο επιδραστικός στην ακρίβεια ταξινόμησης. Σχετικά με την σημασιολογική κατάτμηση των εικόνων, επαληθεύτηκε ότι τα αντικείμενα τα οποία σχετίζονται με τις υποδομές (π.χ. σταθμοί ηλεκτρικής ενέργειας, ηλεκτρικοί σύνδεσμοι, στύλοι και μετασχηματιστές), υπάρχουν στα κύρια σύνολα ανίχνευσης αντικειμένων (CamVid και COCO [198]) και η χρήση ΣΝΔ τα οποία έχουν εκπαιδευτεί για την ανίχνευση αντικειμένων (όπως το ResNet18) απέδωσε ικανοποιητικά αποτελέσματα.

Οι στόχοι που τέθηκαν στην αρχή αυτής της εργασίας ικανοποιήθηκαν στο σύνολό τους. Η εργασία αυτή παρουσιάστηκε στο ετήσιο διεθνές συνέδριο TMREES 2023 και διακρίθηκε με τον τίτλο “Best Paper Award”. Η έρευνα αυτή παρουσιάζεται στην δημοσίευση [199].

6. ΣΥΜΠΕΡΑΣΜΑΤΑ - ΘΕΜΑΤΑ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

Στο κεφάλαιο αυτό παρουσιάζονται συγκεντρωτικά τα συμπεράσματα που εξάγονται από το σύνολο της έρευνας, καθώς και τα ερευνητικά θέματα που ανακύπτουν για περαιτέρω διερεύνηση στο μέλλον. Συγκεκριμένα, στην παράγραφο 6.1 εκθέτονται τα πειραματικά αποτελέσματα των αλγορίθμων που υλοποιήθηκαν και τα επακόλουθα συμπεράσματα αυτών. Στην παράγραφο 6.2, παρατίθενται τα ερευνητικά πεδία που αναφύονται με την παρούσα έρευνα.

6.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα έρευνα πραγματοποιήθηκε ενδελεχής μελέτη των δύο βασικών σημάτων τα οποία μπορούν να αποσαφηνίσουν την διεργασία που συντελείται σε έναν χώρο και ταυτόχρονα να αποτελέσουν δείκτες της ποιότητά της, δίνοντας την δυνατότητα της κατάλληλης διαχείρισης αυτής. Το πρώτο σήμα που μελετήθηκε ήταν αυτό του ήχου. Μελετήθηκε μία ευρεία γκάμα ηχητικών χαρακτηριστικών και υλοποιήθηκαν αλγόριθμοι εξαγωγής των τιμών αυτών. Το μεγάλο πλήθος των χαρακτηριστικών του ήχου αφενός δρα θετικά στην ακρίβεια ταξινόμησης των ηχητικών κλάσεων, αφετέρου δρα αρνητικά στον υπολογιστικό φόρτο των αλγορίθμων. Παράλληλα η χρήση περιττών, αναλόγως της περίπτωσης εφαρμογής, χαρακτηριστικών μπορεί να υποβαθμίσει την ακρίβεια της ταξινόμησης λόγω υπερπροσαρμογής. Από αυτή την σκοπιά προτάθηκε μία νέα μέθοδος ιεράρχησης των χαρακτηριστικών σύμφωνα με την περιγραφική τους ικανότητα βασισμένη στην Ανάλυση Κύριων Συνιστωσών. Η μέθοδος αυτή εφαρμόστηκε σε ένα σύνολο 143 ηχητικών χαρακτηριστικών τα οποία προέρχονται από το πεδίο του χρόνου, από το πεδίο του φάσματος και από το πεδίο της ανθρώπινης αντίληψης. Χρησιμοποιώντας αυξανόμενο αριθμό ηχητικών χαρακτηριστικών υλοποιήθηκαν αλγόριθμοι μηχανικής μάθησης πέντε διαφορετικών τεχνικών. Ονομαστικά, υλοποιήθηκαν μοντέλα Linear Discriminant Analysis, Quadratic Support Vector Machine, k-Nearest Neighbors, Boosted Trees και Random Forest. Τα μοντέλα αυτά εκπαιδεύτηκαν και αξιολογήθηκαν ως προς την αποδιδόμενη ακρίβεια ταξινόμησης, χρησιμοποιώντας την προτεινόμενη κατάταξη των χαρακτηριστικών. Τα αποτελέσματα αυτά συγκρίθηκαν με τα αντίστοιχα της γνωστής μεθόδου κατάταξης χαρακτηριστικών Relief-F. Η ακρίβεια ταξινόμησης με την χρήση των δύο τεχνικών μείωσης της διαστασιολόγησης είναι εν γένει συγκρίσιμα, με αυτά της προτεινόμενης μεθόδου να επιτυγχάνει σταθερότερη ακρίβεια ταξινόμησης σε σχέση με τον αριθμό των χαρακτηριστικών. Ταυτόχρονα, επιβεβαιώνεται η χρησιμότητα των μεθόδων κατάταξης των χαρακτηριστικών με την ακρίβεια ταξινόμησης να ξεπερνά το 92% κάνοντας χρήση περίπου του ενός τρίτου του συνόλου των χαρακτηριστικών (50 χαρακτηριστικά), μειώνοντας έτσι σημαντικό τον υπολογιστικό φόρτο.

Εν συνεχεία, χρησιμοποιήθηκαν αλγόριθμοι βαθιάς μάθησης και συγκεκριμένα Συνελικτικά Νευρωνικά Δίκτυα για την ταξινόμηση του ήχου. Με αυτόν τον τρόπο πραγματοποιήθηκε σύγκριση των αποτελεσμάτων ταξινόμησης με τις κλασσικές μεθόδους μηχανικής μάθησης και με όρους ακρίβειας ταξινόμησης αλλά και με όρους υπολογιστικού φόρτου. Δεδομένου ότι οι μηχανισμοί βαθιάς μάθησης απαιτούν μεγάλο όγκο δεδομένων για την εκπαίδευση των δικτύων και τον καθορισμό των βαρών των ενδιάμεσων επιπέδων, χρησιμοποιήθηκαν γνωστές, καθιερωμένες αρχιτεκτονικές. Επιλέχθηκαν δίκτυα τα οποία έχουν εκπαιδευτεί στην βάση δεδομένων 14 εκατομμυρίων εικόνων ImageNet και έγινε χρήση αυτών μέσω μεταφοράς μάθησης. Το βασικό όφελος της τεχνικής της μεταφοράς μάθησης είναι η εξοικονόμηση πόρων καθώς ως έναν βαθμό τα μοντέλα επαναχρησιμοποιούνται, και επανεκπαιδεύονται σε ένα μέρος τους με τα νέα σύνολα δεδομένων. Η απόδοση της τεχνικής της μεταφοράς μάθησης εξαρτάται σε μεγάλο βαθμό από τις τιμές των εσωτερικών παραμέτρων (υπερπαραμέτρων) επανεκπαίδευσης των δικτύων. Πραγματοποιήθηκε διεξοδική έρευνα των τιμών των υπερπαραμέτρων και των μεταξύ τους

συνδυασμών προκειμένου να μεγιστοποιηθεί η ακρίβεια ταξινόμησης και να ελαχιστοποιηθεί ο υπολογιστικός χρόνος. Οι υπερπαράμετροι που διερευνήθηκαν αφορούν: (α) στην μέθοδο οπισθοδιάδοσης (back-propagation) προκειμένου να ελαχιστοποιηθεί η συνάρτηση απώλειας, ο επονομαζόμενος optimizer, (β) στο μέγεθος του mini-batch το οποίο καθορίζει των αριθμό των επαναλήψεων για τον υπολογισμό του σφάλματος και την ανάλογη ενημέρωση των βαρών, (γ) το μέγεθος των διορθωτικών βημάτων, learning rate, (δ) τον αριθμό των epochs, δηλαδή των πλήρων περασμάτων από το σύνολο των αρχείων εκπαίδευσης, και (ε) το validation patience που αποτελεί τον αριθμό των epochs που δεν βελτιώνεται η ακρίβεια επικύρωσης και επομένως σταματά η εκπαίδευση. Προκειμένου να επιλεχθούν βέλτιστοι συνδυασμοί τιμών των υπερπαραμέτρων, διαμορφώθηκαν δύο κριτήρια. Το πρώτο κριτήριο (υπό όρους) απαιτεί το μοντέλο να έχει καλύτερες επιδόσεις σε όρους ακρίβειας ταξινόμησης και χρόνου εκπαίδευσης από τις μέσες αντίστοιχες τιμές κατά μία τυπική απόκλιση. Το δεύτερο κριτήριο (με βάρος) χρησιμοποίησε έναν σταθμισμένο μέσο όρο της (θετικής) διαφοράς αυτών (της ακρίβειας ταξινόμησης και χρόνου εκπαίδευσης) από τις αντίστοιχες μέσες τιμές. Το βάρος K επέτρεψε την κανονικοποίηση και έδωσε προτεραιότητα στην ακρίβεια ταξινόμησης (με συντελεστή 2). Το σύνολο των πειραμάτων πραγματοποιήθηκε σε τρεις δημόσια διαθέσιμες βάσεις δεδομένων ήχου: το UrbanSound8K, το ESC-10 και το Air Compressor. Επίσης, τα πειράματα πραγματοποιήθηκαν σε πέντε ΣΝΔ διαφορετικής αρχιτεκτονικής: τα GoogleNet, SqueezeNet και ShuffleNet με αρχική χρήση την ταξινόμηση εικόνων και τα VGGish και YAMNet με αρχική χρήση την ταξινόμηση ήχων. Και για τις δύο περιπτώσεις δικτύων (Εικόνας και Ήχου) προηγήθηκε κατάλληλη μετατροπή του ήχου σε εικονική αναπαράσταση, scalograms και spectrograms, αντίστοιχα. Η έρευνα έδειξε ότι όλα τα ΣΝΔ ευθυγραμμίστηκαν ως προς τον optimizer, με την καταλληλότερη επιλογή να είναι ο Adam, και το learning rate ίσο με 2×10^{-4}. Το μέγεθος του mini-batch και ο αριθμός των epochs και του validation patience εξαρτάται άμεσα από το πλήθος των αρχείων εκπαίδευσης, επομένως διαφοροποιείται για κάθε βάση δεδομένων. Και γι' αυτές όμως τις υπερπαραμέτρους οριοθετήθηκαν τιμές διερεύνησης αναλόγως του τύπου του ΣΝΔ. Τέλος, διερευνήθηκε η συγχώνευση των αποτελεσμάτων των επιμέρους δικτύων, με τα αποτελέσματα της πλειοψηφικής λογικής των τριών ΣΝΔ Εικόνας να υπερτερούν από του κάθε δικτύου ξεχωριστά. Συνολικά, η καλύτερη απόδοση ταξινόμησης επιτεύχθηκε από το VGGish.

Το πεδίο έρευνας που αφορά στην ταξινόμηση του ήχου έχει αναπτυχθεί ταχέως τις τελευταίες δεκαετίες με αποτέλεσμα την δημιουργία συνόλων δεδομένων ήχου, τα οποία περιλαμβάνουν διαφορετικές, αυθαίρετα επιλεγμένες κλάσεις ήχου για διαφορετικές μελέτες περιπτώσεων. Από αυτή την άποψη, διερευνήθηκαν δύο τύποι συστηματικών συσχετίσεων μεταξύ των κλάσεων του ήχου: α) η σημασιολογική και β) η σύγκριση με βάση τις τιμές των ηχητικών χαρακτηριστικά. Όσον αφορά την πρώτη συσχέτιση, οι κλάσεις των ήχων συνδέονται σημασιολογικά λαμβάνοντας υπόψη την ενοποιητική οντολογία AudioSet, με τις κλάσεις του ήχου να συσχετίζονται με βάση την προέλευση (πηγή) του ήχου. Όσον αφορά την δεύτερη συσχέτιση, αυτή βασίζεται στον υπολογισμό της απόστασης των τιμών των ηχητικών χαρακτηριστικών.

Παράλληλα, οι ήχοι που προέρχονται από ρεαλιστικά περιβάλλοντα περιλαμβάνουν κλάσεις οι οποίες συνδυάζονται με διαδοχικό ή/και επικαλυπτόμενο τρόπο. Σε αυτές τις περιπτώσεις είναι απαραίτητος ο διαχωρισμός (κατάτμηση) των ηχητικών χρονοσειρών προκειμένου να επιτευχθεί η ταξινόμηση κάθε ηχητικού τμήματος. Ορίστηκε ένα σύνολο παραμέτρων, όπως η ελάχιστη διάρκεια του ήχου και η χρονική απόσταση ήχων ίδιας κλάσης, το μήκος του παραθύρου και του άλματος ώστε να επιτευχθεί η διαδικασία κατάτμησης των ηχητικών ροών. Η ρύθμιση των παραμέτρων αυτών οδηγεί σε περισσότερο ή λιγότερο λεπτομερή κατάτμηση, και αντίστοιχα ταξινόμηση με αποτέλεσμα μεγαλύτερο ή μικρότερο αντίστοιχα αριθμό ηχητικών αναγνωρίσεων.

Το δεύτερο σήμα που μελετήθηκε ήταν αυτό της εικόνας. Δεδομένης της εξέλιξης των μεθοδολογιών και των εργαλείων που έχουν αναπτυχθεί τα τελευταία χρόνια, η ανίχνευση και αναγνώριση προσώπων και αντικειμένων έχει καταστεί τετριμμένη διαδικασία. Από αυτή την σκοπιά, η έρευνα αυτής της διατριβής εστίαση σε μία πιο εκλεπτυσμένη ανάλυση εικόνων που αφορά στις εκφράσεις συναισθημάτων προσώπων. Η αναγνώριση συναισθήματος του προσώπου (Facial Emotion Recognition) ανήκει στην κατηγορία των τεχνολογιών της συναισθηματικής υπολογιστικής με στόχο την αναγνώριση και ερμηνεία των συναισθημάτων. Αυτή η ταξινόμηση μπορεί να αποτελέσει δείκτη της ποιότητας της εξελισσόμενης διαδικασίας και κριτήριο για την αντίστοιχη διαχείριση αυτής. Η έρευνα έγινε στις βάσεις δεδομένων (α) KDEF, του πανεπιστημίου Karolinska του τμήματος Κλινικής Νευροεπιστήμης (Clinical Neuroscience), (β) RaFD, του Πανεπιστημίου Nijmegen του Ινστιτούτου Επιστήμης της Συμπεριφοράς (Behavioural Science Institute) αφού αποκτήθηκε πρόσβαση σε αυτές και (γ) της δημόσια διαθέσιμης μικρής βιβλιοθήκης JAFFE.

Η επεξεργασία των εικόνων πραγματοποιήθηκε με εξαγωγή των χαρακτηριστικών των εικόνων με χειροκίνητες μεθόδους (handcrafted methods) αλλά και με την χρήση ΣΝΔ. Συγκεκριμένα υλοποιήθηκαν οι αλγόριθμοι εξαγωγής χαρακτηριστικών με τις handcrafted μεθόδους Local Binary Patterns (LBP) και Histogram of Oriented Gradients (HOG). Η μέθοδος LBP κωδικοποιεί πληροφορίες της υφής μιας εικόνας σε κλίμακα του γκρι, συγκρίνοντας την διαφορά στην ένταση κάθε pixel με τα γειτονικά του. Η μέθοδος HOG περιγράφει τις ακμές και τις γωνίες ενός αντικειμένου μέσω των τοπικών κλίσεων της έντασης. Και για τις δύο περιπτώσεις μεθόδων πραγματοποιήθηκε διερεύνηση των αποτελεσμάτων ταξινόμησης, αλλάζοντας τις τιμές των εσωτερικών τους παραμέτρων αλλά και το μέγεθος της εικόνας. Η σμίκρυνση των εικόνων με έναν συντελεστή 2 επηρεάζει θετικά την ακρίβεια ταξινόμηση και στις δύο μεθόδους. Η μέθοδος LBP αποδείχθηκε πιο αποδοτική στις βάσεις δεδομένων που περιέχουν εικόνες υψηλότερης ανάλυσης (KDEF και RaFD), ενώ η μέθοδος HOG στις εικόνες χαμηλότερης ποιότητας της βάσης JAFFE. Η υψηλότερη ακρίβεια ταξινόμησης επιτυγχάνεται για κάθε τεχνική και βάση δεδομένων με χαρακτηριστικά ίσων διαστάσεων.

Τα ΣΝΔ που επιλέχθηκαν για την ταξινόμηση της εικόνας είναι τριών διαφορετικών μεταξύ τους αρχιτεκτονικών, και διαφορετικά από αυτά που χρησιμοποιήθηκαν για την ταξινόμηση του ήχου, προκειμένου να αποκτηθεί μία εποπτική αντίληψη της πλειοψηφίας των διαθέσιμων δικτύων. Συγκεκριμένα επιλέχθηκαν τρία δίκτυα διαβαθμισμένου βάθους της αρχιτεκτονικής ResNet (ResNet18, ResNet50 και ResNet101), προκειμένου να εξεταστεί αν το βάθος του δικτύου επηρεάζει τα αποτελέσματα της ταξινόμησης, και τα δίκτυα της αρχιτεκτονικής Inception (version 3) και Efficient (B0). Η χρήση των ΣΝΔ για την επεξεργασία FER ήταν διττή. Αρχικά, εξετάστηκε η απόδοση ταξινόμησης χωρίς την επανεκπαίδευση των δικτύων στα νέα δεδομένα. Αυτό πραγματοποιήθηκε εξάγοντας τα χαρακτηριστικά της εικόνας από επίπεδα διαφορετικού βάθους των δικτύων (από το 25%, 50%, 75% και 100% αυτών), και τα χαρακτηριστικά αυτά τροφοδότησαν ταξινομητή SVM, αφού επιβεβαιώθηκε ότι ο συγκεκριμένος αλγόριθμος αποδίδει καλύτερα αποτελέσματα από άλλους ταξινομητές μηχανικής μάθησης. Σε αυτό το σημείο, προέκυψε ένα σημαντικό αποτέλεσμα: η εξαγωγή των χαρακτηριστικών της εικόνας από το τελευταίο και βαθύτερο επίπεδο (δηλαδή από το 100% του βάθους των δικτύων) οδηγούσε στα χαμηλότερα αποτελέσματα της ακρίβειας ταξινόμησης σε σύγκριση και με τις handcrafted μεθόδους αλλά και με τα χαρακτηριστικά ρηχότερων επιπέδων. Αντίθετα, τα εξαγόμενα χαρακτηριστικά από το 25%, 50% και 75% του βάθους των δικτύων φτάνει ή και υπερβαίνει την ακρίβεια ταξινόμησης των handcrafted μεθόδων. Συγκεκριμένα, με τα χαρακτηριστικά εξαγόμενα από το 50% του ResNet50 και από το

25% του ResNet101 παρουσιάζεται η υψηλότερη απόδοση ταξινόμησης σε όλες τις περιπτώσεις των μέχρι εδώ μεθόδων, και σε όλες τις περιπτώσεις ποιότητας εικόνων.

Η δεύτερη χρήση των ΣΝΔ υλοποιήθηκε με μεταφορά μάθησης. Τα ΣΝΔ επανεκπαιδεύτηκαν στα νέα δεδομένα, κάνοντας χρήση της προηγούμενης διερεύνησης των τιμών των υπερπαραμέτρων. Η ταξινόμηση πραγματοποιήθηκε από (α) τον ενσωματωμένο ταξινομητή των δικτύων (πλήρως συνδεδεμένο επίπεδο), και (β) από ταξινομητή SVM. Και σε αυτή την περίπτωση τα πειραματικά αποτελέσματα ήταν σημαντικά. Καταρχάς, η μεταφορά μάθησης έχει νόημα όταν οι νέες βάσεις δεδομένων έχουν μεγάλο πλήθος αρχείων. Στην περίπτωση του συνόλου JAFFE δεν παρατηρήθηκε αύξηση της ακρίβειας ταξινόμησης, σε αντίθεση με τα άλλα δύο σύνολα. Κατά δεύτερον, το βάθος των δικτύων επηρεάζει θετικά την ακρίβεια ταξινόμησης, με το ResNet101 να αποδίδει την υψηλότερη ακρίβεια ταξινόμησης. Τέλος, ο ταξινομητής SVM αποδίδει καλύτερα αποτελέσματα από τον ενσωματωμένο ταξινομητή των δικτύων.

Σε όλες τις περιπτώσεις μεθόδων διερευνήθηκε εκτός από την ακρίβεια ταξινόμησης και ο υπολογιστικός χρόνος. Η μικρότερη υπολογιστική διάρκεια επιτεύχθηκε με την εξαγωγή χαρακτηριστικών της εικόνας από ενδιάμεσο βάθος των δικτύων. Επίσης, διερευνήθηκε η ανθεκτικότητα των επιμέρους μεθόδων σε δύο τύπους θορύβου: τον Gaussian και τον Salt & Pepper. Προκειμένου να εξεταστεί το χειρότερο σενάριο, εξετάστηκε η περίπτωση που η εκπαίδευση έχει πραγματοποιηθεί με καθαρές εικόνες και ο έλεγχος πραγματοποιείται σε αλλοιωμένες εικόνες. Τα ΣΝΔ αποδεικνύονται πιο ανθεκτικά στον θόρυβο Salt & Pepper από ότι στον Gaussian. Οι χαμηλής ποιότητας εικόνες (όπως του συνόλου JAFFE) επηρεάζονται περισσότερο, όπως ήταν αναμενόμενο. Οι handcrafted μέθοδοι επηρεάζονται έντονα από τους δύο τύπους θορύβου καθώς παραποιείται η πληροφορία ακμών εξαιτίας του Gaussian θορύβου, ενώ με τον Salt & Pepper οδηγεί σε κλίσεις με μεγαλύτερο μέγεθος και λανθασμένη κατεύθυνση.

Κατόπιν αυτής της διερεύνησης δημιουργήθηκε ένα πλαίσιο αποφάσεων για την υποστήριξη της κατάλληλης επιλογής με βάση τις προδιαγραφές κάθε εφαρμογής.

Η ετερογένεια των διαθέσιμων αλγορίθμων για την ταξινόμηση σημάτων αντανακλάται στους απαιτούμενους υπολογιστικούς πόρους για την εκπαίδευση των μοντέλων και την εξαγωγή των αποτελεσμάτων. Από αυτή την σκοπιά, οι απαιτήσεις υπολογιστικών πόρων μπορούν να αποτελέσουν ένα από τα κριτήρια για την επιλογή της μεθόδου ταξινόμησης. Διαμορφώθηκε λοιπόν ένα σύνολο τιμών χρόνου εκπαίδευσης για διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων, με διαφορετικές διαμορφώσεις εκπαίδευσης και για διαφορετικά σύνολα δεδομένων. Εκπαιδεύτηκαν πέντε μοντέλα παλινδρόμησης με βάση τα νευρωνικά δίκτυα για την εκτίμηση του χρόνου εκπαίδευσης. Τα μοντέλα αξιολογήθηκαν με βάση τον συντελεστή συσχέτισης και το μέσο τετραγωνικό σφάλμα. Το αποτέλεσμα ήταν ότι το νευρωνικό δίκτυο δύο επιπέδων (Bilayered) απέδωσε τον υψηλότερο συντελεστή συσχέτισης με το μικρότερο σφάλμα, γεγονός που υποδεικνύει ότι με αυτό το μοντέλο μπορεί να επιτευχθεί καλή προσέγγιση του υπολογιστικού χρόνου που θα χρειαστεί μία περίπτωση παραπλήσιων δεδομένων.

Η αξία αυτής της διδακτορικής διατριβής αναδεικνύεται από την εφαρμογή των αλγορίθμων και των πλαισίων επιλογής σε κάθε περίπτωση σήματος, σε διαφορετικά πεδία έρευνας. Για το σήμα του ήχου πραγματοποιήθηκε ανάλυση του περιβαλλοντικού θορύβου, με επιλογή των ηχητικών χαρακτηριστικών με βάση την προτεινόμενη μέθοδο που βασίζεται στην Ανάλυση Κύριων Συνιστωσών. Αναπτύχθηκε μοντέλο μηχανικής μάθησης kNN με το οποίο επιτεύχθηκε ταξινόμηση οκτώ κλάσεων ήχου που επιβαρύνουν το αστικό περιβάλλον σε ποσοστό 85%. Με αυτόν τον τρόπο είναι δυνατή η κατανόηση της προέλευσης του ήχου και της λήψης αντίστοιχων μέτρων προκειμένου να ανακουφιστεί το περιβάλλον από την ηχορρύπανση. Για το σήμα της εικόνας, το πεδίο εφαρμογής

ήταν η έγκαιρη ανίχνευση πυρκαγιάς μέσω ταξινόμησης εικόνων που περιλαμβάνουν φωτιά σε διαφορετικά στάδια εξέλιξης αυτής. Ταυτόχρονα, με χρήση ΣΝΔ και εκπαίδευσης αυτών στην ανίχνευση συγκεκριμένων αντικειμένων ενδιαφέροντος, επιτεύχθηκε η σημασιολογική κατάτμηση του περιεχομένου των εικόνων σε όρους των υποδομών που βρίσκονται στον χώρο της πυρκαγιάς. Με αυτόν τον τρόπο είναι δυνατή η δημιουργία πλαισίου ανάλυσης κινδύνου αναλόγως των υποδομών που κινδυνεύουν αλλά και επιλογής τρόπου επέμβασης αναλόγως της πρόσβασης σε αυτές.

Το έργο της διατριβής αυτής μπορεί να συνοψιστεί στα εξής:

- Μελέτη Ήχου
  - Μηχανισμοί Μηχανικής Μάθησης
    - Προσδιορισμός 143 ηχητικών χαρακτηριστικών και ανάπτυξη αλγορίθμων για την εξαγωγή των τιμών αυτών
    - Ανάπτυξη μεθόδου ιεράρχησης των ηχητικών χαρακτηριστικών με βάση την περιγραφική τους ικανότητα
    - Ανάπτυξη πέντε αλγορίθμων Μηχανικής Μάθησης για την ταξινόμηση του ήχου χρησιμοποιώντας αυξανόμενο αριθμό ηχητικών χαρακτηριστικών
    - Αξιολόγηση των μοντέλων ταξινόμησης και της μεθόδου κατάταξης των χαρακτηριστικών ως προς την ακρίβεια ταξινόμησης
  - Μηχανισμοί Βαθιάς Μάθησης
    - Επιλογή τριών ΣΝΔ διαφορετικής αρχιτεκτονικής
    - Διερεύνηση των τιμών των υπερπαραμέτρων επανεκπαίδευσης των ΣΝΔ
    - Ανάπτυξη κριτηρίων για την επιλογή του βέλτιστου συνδυασμού των τιμών αυτών
  - Σημασιολογική και τεχνική συσχέτιση των κλάσεων ήχων ίδιας ή/και διαφορετικής βάσης δεδομένων
  - Κατάτμηση ηχητικών ροών ανά ηχητική κλάση
- Μελέτη Εικόνας
  - Ανάπτυξη αλγορίθμων για την εξαγωγή χαρακτηριστικών εικόνας με δύο handcrafted μεθόδους
  - Χρήση πέντε ΣΝΔ, τριών της αρχιτεκτονικής residual διαβαθμισμένου βάθους και δύο διαφορετικής αρχιτεκτονικής
    - Εξαγωγή χαρακτηριστικών χωρίς την εκπαίδευση των δικτύων στα νέα δεδομένα από διαφορετικά επίπεδα του βάθους τους
    - Εξαγωγή των χαρακτηριστικών μετά την επανεκπαίδευση των δικτύων στα νέα δεδομένα μέσω μεταφοράς μάθησης
  - Αξιολόγηση των επιμέρους μεθόδων ως προς την ακρίβεια ταξινόμησης και τον απαιτούμενο υπολογιστικό χρόνο
  - Διερεύνηση της επίδρασης δύο τύπων θορύβου σε κάθε περίπτωση χαρακτηριστικών εικόνας
  - Ανάπτυξη πλαισίου επιλογής κατάλληλης επιλογής εξαγωγής χαρακτηριστικών αναλόγως των προδιαγραφών και απαιτήσεων της εφαρμογής

- Ανάπτυξη μοντέλων παλινδρόμησης με βάση νευρωνικά δίκτυα για την εκτίμηση του χρόνου εκπαίδευσης για διαφορετικές διαμορφώσεις εκπαίδευσης και διαφορετικά σύνολα δεδομένων
- Επέκταση του πεδίου εφαρμογών σε περιπτώσεις μελέτης ανοιχτών χώρων για την ταξινόμηση του ήχου και της εικόνας

6.2 ΘΕΜΑΤΑ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

Τις τελευταίες δεκαετίες έχει υπάρξει μία τεράστια αύξηση των συσκευών IoT, με εκτιμήσεις να κάνουν λόγο για 75 δισεκατομμύρια συσκευές μέχρι το 2025 [200]. Με την ανάπτυξη του IoT, των εφαρμογών Μεγάλων Δεδομένων (Big Data), της επεξεργασίας δεδομένων και των εφαρμογών Μηχανικής Μάθησης, το ζητούμενο είναι να μπορούν να υλοποιηθούν τα μοντέλα αναγνώρισης και ταξινόμησης με συσκευές μικρές όσον αφορά το μέγεθος και την κατανάλωση πόρων, αλλά μεγάλης απόδοσης. Η επιλογή των αισθητήρων εξαρτάται από τον χώρο και τις ενέργειες που πραγματοποιούνται σε αυτόν. Ο στόχος είναι η συλλογή δεδομένων ικανών να περιγράψουν τις συνθήκες και τις δραστηριότητες που συμβαίνουν προκειμένου να αυτοματοποιηθούν οι διαδικασίες και να ληφθούν κατάλληλες αποφάσεις ώστε να προσαρμοστούν και βελτιστοποιηθούν οι επιμέρους διεργασίες. Επομένως, το πλαίσιο που χαρακτηρίζει τον χώρο και τις ενέργειες που συμβαίνουν σε αυτόν πρέπει να συμπεριληφθούν στην προτεινόμενη αρχιτεκτονική. Ένα πρώτο θέμα που προκύπτει είναι η επιλογή των συσκευών που συλλέγουν τα σήματα και επεξεργάζονται τα δεδομένα, όπως οι εμπορικές συσκευές all-in-one αισθητήρων και η δημιουργία κόμβων αυτών, προκειμένου να είναι εφικτή η ανάκτηση και μετάδοση των επιμέρους σημάτων.

Ο τελικός στόχος της παρούσας έρευνας είναι η αναβάθμιση ενός χώρου από διάφορες πτυχές με χρήση των νέων τεχνολογιών προκειμένου να καταστεί αυτός «έξυπνος». Τα βασικά σημεία της έννοιας των ευφυών χώρων είναι:

- Η διαθεσιμότητα των συνδεδεμένων συσκευών και αισθητήρων
- Η δυνατότητα συλλογής και επεξεργασίας των δεδομένων
- Η εξαγωγή συμπερασμάτων από την επεξεργασία των δεδομένων
- Η υποστήριξη αποφάσεων

Όμως, η εκπαίδευση και η ανάπτυξη των μοντέλων ταξινόμησης είναι μία υπολογιστικά δαπανηρή διαδικασία, γεγονός που ενισχύει την ανάγκη εύρεσης λύσεων προκειμένου τα μοντέλα μηχανικής μάθησης να γίνουν πιο λειτουργικά. Οι μέχρι τώρα λύσεις βασίζονται στο cloud computing. Προκειμένου όμως να υλοποιηθούν ρεαλιστικά σενάρια πραγματικού χρόνου είναι απαραίτητη η εύρεση βελτιωμένων προσεγγίσεων.

Επιπρόσθετα, προς αυτή την κατεύθυνση σχεδιάστηκαν τα μοντέλα Transformers. Η πρόσφατη έρευνα έχει δείξει ότι οι Transformers αποδίδουν καλύτερα από τα ΣΝΔ τόσο σε εφαρμογές εικόνας όσο και σε εφαρμογές ήχου. Η αρχιτεκτονική αυτών των δικτύων βασίζεται σε μηχανισμούς προσοχής (attention mechanisms) με αποτέλεσμα να μπορούν να εκτελούνται παράλληλα υπολογισμοί ενσωματώνοντας το γενικό πλαίσιο των χαρακτηριστικών, με αποτέλεσμα την επίτευξη αξιόπιστων αποτελεσμάτων. Μία από τις πρώτες μελλοντικές έρευνες είναι η μελέτη και χρήση μοντέλων Transformers στις περιπτώσεις μελέτης που πραγματοποιήθηκαν σε αυτή την διατριβή προκειμένου να υπάρξει μία σύγκριση των αποτελεσμάτων, και ως προς την απόδοση ταξινόμησης και ως προς τους υπολογιστικούς πόρους. Σε αυτό το σημείο θα είχε νόημα να διερευνηθεί η απόδοση των μοντέλων αυτών κάνοντας συγκριτική αξιολόγηση της μεταφοράς μάθησης με διάφορα σετ τιμών των υπερπαραμέτρων.

Ωστόσο, η έως τώρα έρευνα στην βιβλιογραφία των Transformers υποδεικνύει ότι και αυτά τα μοντέλα απαιτούν μεγάλο πλήθος δεδομένων. Δεδομένου ότι αυτό αποτέλεσε εμπόδιο και στην μέχρι τώρα έρευνα, μία ακόμη μελλοντική εργασία θα είναι η διερεύνηση τεχνικών ενίσχυσης των υπαρχόντων συνόλων δεδομένων, όπως παραδείγματος χάριν με την τεχνική Random Erasing Data Augmentation κ.ά. [201]-[203].

Επίσης, πρόσφατα έγινε διαθέσιμο ένα υποσύνολο 67000 ηχητικών αρχείων του AudioSet με ισχυρές ετικέτες ακρίβειας ανάλυσης 0.1 sec, σε σύγκριση με το σύνολο των 1.8 εκατομμυρίων ηχητικών αρχείων με ετικέτες ανάλυσης 10 sec [198]. Μπορεί έτσι να πραγματοποιηθεί μία πιο εκλεπτυσμένη ταξινόμηση ήχων αλλά και ηχητικών ροών. Προς αυτή την κατεύθυνση και σε αντιστοιχία με το FER που μελετήθηκε σε αυτή την διατριβή, είναι δυνατή η αναγνώριση συναισθημάτων μέσω ηχητικών σημάτων (Speech Emotion Recognition – SER).

Το σήμα της εικόνας μπορεί επίσης να αξιοποιηθεί περαιτέρω για την απόκτηση μίας πιο εποπτικής ανάλυσης του χώρου και των δραστηριοτήτων. Είναι δυνατόν να γίνει σημασιολογική αναγνώριση του χώρου, καθώς και αναγνώριση της στάσης του σώματος και των χειρονομιών.

Η συγχώνευση των αποτελεσμάτων που προκύπτουν από την ταξινόμηση του ήχου και την ταξινόμηση της εικόνας, είναι η κύρια μελλοντική εργασία. Ο στόχος είναι η αποσαφήνιση των δραστηριοτήτων και της ροής της συνολικής διεργασίας προκειμένου να βελτιωθούν και να αναβαθμιστούν οι χρησιμοποιούμενες μέθοδοι για την εκάστοτε περίπτωση. Για τον σκοπό αυτό, μπορούν να συμπεριληφθούν και άλλα σήματα ή τιμές παραμέτρων περιβάλλοντος.

Τέλος, η αναζήτηση περιοχών που μπορεί να εφαρμοστούν οι αλγόριθμοι που αναπτύχθηκαν με κατάλληλη προσαρμογή αυτών είναι μία από τις μελλοντικές κινήσεις. Η αξιοποίηση της μέχρι τώρα έρευνας στο πεδίο της βιομηχανίας, της ιατρικής, της μεταφοράς, της ασφάλειας ή ακόμα και σε πιο εξειδικευμένους χώρους συνύπαρξης και συνεργασίας ατόμων αποτελεί πρόκληση.

ΑΝΑΦΟΡΕΣ

[1] O. Lartillot and P. Toiviainen, "A Matlab toolbox for musical feature extraction from audio", *Int'l Conf. on Digital Audio Effects (DAFx)*, 2007, pp.237-244. https://www.dafx.de/paper-archive/2007/Papers/p237.pdf

[2] O. Brdiczka, M. Langet, J. Maisonnasse, and J.L. Crowley, Detecting human behavior models from multimodal observation in a smart home, *IEEE Transactions on Automation Science and Engineering*, vol. 6, issue 4, Oct. 2009, pp. 588-597. https://doi.org/10.1109/TASE.2008.2004965

[3] P. Lipar, J. Prezelj, P. Šteblaj, M. Čudina, and F. Mihelič, "Identification of machinery sounds", *2013 21st Telecommunications Forum Telfor (TELFOR)*, IEEE, 2013, pp. 470-473. https://doi.org/10.1109/TELFOR.2013.6716269

[4] J.P. Bello, C. Mydlarz, and J. Salamon, Sound analysis in smart cities, *Computational Analysis of Sound Scenes and Events*, Sept. 2017, pp.373-397. https://doi.org/10.1007/978-3-319-63450-0_13

[5] J. Cao, M. Cao, J. Wang, C. Yin, D. Wang, and P.P. Vidal, Urban noise recognition with convolutional neural network, *Multimedia Tools and Applications*, vol. 78, July 2018, pp. 29021-29041. https://doi.org/10.1007/s11042-018-6295-8

[6] C. Mydlarz, M.S. Sharma, Y. Lockerman, B. Steers, C.T. Silva, and J.P. Bello, The Life of a New York City Noise Sensor Network, *Sensors*, vol. 19, issue 6, 1415, Mar. 2019, pp.1415. https://doi.org/10.3390/s19061415

[7] J. Shen, L. Nie, and T.S. Chua, "Smart ambient sound analysis via structured statistical modeling", *Proc. MultiMedia Modeling: 22nd Int'l Conf. (MMM 2016)*, Springer, Cham, 2016, pp. 231-243. https://doi.org/10.1007/978-3-319-27674-8_21

[8] S.A. Mitilineos, S.M. Potirakis, N.A. Tatlas, and M. Rangoussi, A Two-Level Sound Classification Platform for Environmental Monitoring, *Journal of Sensors*, June 2018. https://doi.org/10.1155/2018/5828074

[9] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley, Detection and classification of acoustic scenes and events, *IEEE Transactions on Multimedia*, vol. 17, issue 10, Oct. 2015, pp. 1733-1746. https://doi.org/10.1109/TMM.2015.2428998

[10] J. Ye, T. Kobayashi, and M. Murakawa, Urban sound event classification based on local and global features aggregation, *Applied Acoustics*, vol. 117, Feb. 2017, pp. 246-256. https://doi.org/10.1016/j.apacoust.2016.08.002

[11] M. Rossi, G. Troster, and O. Amft, "Recognizing daily life context using web-collected audio data", *2012 16th Int'l Symposium on Wearable Computers (ISWC)*, IEEE, 2012, pp.25-28. https://doi.org/10.1109/ISWC.2012.12

[12] J. Ramirez, J.C. Segura, C. Benitez, A. De La Torre, and A. Rubio, Efficient voice activity detection algorithms using long-term speech information, *Speech Communication*, vol. 42 issues 3-4, Apr. 2004, pp. 271-287. https://doi.org/10.1016/j.specom.2003.10.002

[13] G. Liu, Y. Lei, J. Hansen, and H.L. John, "Robust feature front-end for speaker identification", *2012 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2012, pp. 4233-4236. https://doi.org/10.1109/ICASSP.2012.6288853

[14] M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges", *Proc. of the IEEE*, IAEE, vol. 96, issue 4, 2008, pp. 668-696. https://doi.org/10.1109/JPROC.2008.916370

[15] Y. Alsouda, S. Pllana, and A. Kurti, A machine learning driven IoT solution for noise classification in smart cities, *arXiv preprint*, Sept. 2018. https://doi.org/10.48550/arXiv.1809.00238

[16] P.J. Donnelly, N. Blanchard, A.M. Olney, S. Kelly, M. Nystrand, and S.K. D'Mello, "Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context", *Proc. 7th Int'l Learning Analytics & Knowledge Conference (LAK17)*, ACM, 2017, pp.218-227. https://doi.org/10.1145/3027385.3027417

[17] R. Cosbey, A. Wusterbarth, and B. Hutchinson, "Deep learning for classroom activity detection from audio", *ICASSP 2019-2019 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 3727-3731. https://doi.org/10.1109/ICASSP.2019.8683365

[18] M.S. Khan and I. Zualkernan, "Using Convolutional Neural Networks for Smart Classroom Observation", *2020 Int'l Conf. on Artificial Intelligence in Information and Communication (ICAIIC)*, IEEE, 2020, pp. 608-612. https://doi.org/10.1109/ICAIIC48513.2020.9065260

[19] A. James, Y.H.V. Chua, T. Maszczyk, A.M. Núñez, R. Bull, K. Lee, and J. Dauwels, "Automated classification of classroom climate by audio analysis", *9th Int'l Workshop on Spoken Dialogue System Technology*, Springer, Singapore, 2019, pp. 41-49. https://doi.org/10.1007/978-981-13-9443-0_4

[20] M. Robnik-Šikonja and I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, *Machine Learning*, vol. 53, Issues 1-2, Oct. 2003, pp. 23-69. https://doi.org/10.1023/A:1025667309714

[21] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, and J.H. Moore, Relief-based feature selection: Introduction and review, *Journal of Biomedical Informatics*, vol. 85, Sept. 2018, pp. 189-203. https://doi.org/10.1016/j.jbi.2018.07.014

[22] S. Fakhraei, H. Soltanian-Zadeh, and F. Fotouhi , Bias and stability of single variable classifiers for feature ranking and selection, *Expert Systems with Applications*, vol. 41, issue 15, Nov. 2014, pp. 6945-6958. https://doi.org/10.1016/j.eswa.2014.05.007

[23] U. Moorthy and U.D. Gandhi, Forest optimization algorithm-based feature selection using classifier ensemble, *Computational Intelligence*, vol. 36, issue 4, Nov. 2020, pp. 1445-1462. https://doi.org/10.1111/coin.12265

[24] L.V.D. Maaten and G. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research*, vol. 9, Nov. 2008, pp. 2579-2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl

[25] M. Crocco, M. Cristani, A. Trucco and V. Murino, Audio surveillance: A systematic review, *ACM Computing Surveys (CSUR)*, vol. 48, issue 4, Feb. 2016, pp. 1-46. https://doi.org/10.1145/2871183

[26] D.J. Mashao, "Comparing SVM and GMM on parametric feature-sets", *Proc. 14th Annual Symposium of the Pattern Recognition Association of South Africa*, Citeseer, 2003, pp.15-20.https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4ab18f3f96d4d2ac2d0b667dc293159cf254724e#page=25

[27] S.S. Keerthi and C.J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, *Neural Computation*, vol. 15 issue 7, July 2003, pp. 1667-1689. https://doi.org/10.1162/089976603321891855

[28] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, and T. Virtanen, DCASE 2017 challenge setup: Tasks, datasets and baseline system, *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, Nov.2017. https://inria.hal.science/hal-01627981/document

[29] A. Lerch, *An introduction to audio content analysis: Applications in signal processing and music informatics*, Wiley-IEEE Press, 2012. https://dl.acm.org/doi/abs/10.5555/2392638

[30] F. Alías, J.C. Socoró, and X. Sevillano, A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds, *Applied Sciences*, vol. 6 issue 5, May 2016, https://doi.org/10.3390/app6050143

[31] R. Serizel, V. Bisot, S. Essid, and G. Richard, Acoustic features for environmental sound analysis, *Computational Analysis of Sound Scenes and Events*, Sept. 2017, pp. 71-101. https://doi.org/10.1007/978-3-319-63450-0_4

[32] G. Peeters, B.L. Giordano, P. Susini, N. Misdariis, and S. McAdams, The timbre toolbox: Extracting audio descriptors from musical signals, *The Journal of the Acoustical Society of America*, vol. 130, issue 5, Nov. 2011, pp. 2902-2916. https://doi.org/10.1121/1.3642604

[33] L. Muda, M. Begam, and I. Elamvazuthi , Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques, *arXiv preprint*, Mar. 2010. https://doi.org/10.48550/arXiv.1003.4083

[34] S. Kumar, S. Bhattacharya, and P. Patel, "A new pitch detection scheme based on ACF and AMDF", *2014 IEEE Int'l Conf. on Advanced Communications, Control and Computing Technologies (ICACCCT)*, IEEE, 2014, pp. 1235-1240. https://doi.org/10.1109/ICACCCT.2014.7019296

[35] G. Chandrashekar and F. Sahin, A survey on feature selection methods, *Computers & Electrical Engineering*, vol. 40, issue 1, Jan. 2014, pp. 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[36] K. Kira and L.A. Rendell, "A practical approach to feature selection", *Proc. 9th Int'l Workshop on Machine Learning (ML '92)*, Morgan Kaufmann, 1992, pp.249-256. https://doi.org/10.1016/B978-1-55860-247-2.50037-1

[37] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, Springer New York, 2013. https://doi.org/10.1007/978-1-0716-1418-1

[38] R.E. Schapire, "A brief introduction to boosting", *Proc. 16th Int'l Joint Conference on Artificial Intelligence (IJCAI)*, 1999, pp. 1401-1406. https://www.ijcai.org/Proceedings/99-2/Papers/103.pdf

[39] L. Breiman, Random Forests, *Machine Learning*, vol. 45, issue 1, Oct. 2001, pp. 5-32. https://doi.org/10.1023/A:1010933404324

[40] E. Tsalera, A. Papadakis, and M. Samarakou, Novel principal component analysis-based feature selection mechanism for classroom sound classification, *Computational Intelligence*, vol. 37, issue 4, Nov. 2021, pp. 1827-1843. https://doi.org/10.1111/coin.12468

[41] https://www.image-net.org/ [Προσπελάστηκε 1/2/2022]

[42] S. Chachada and C.C.J. Kuo, Environmental sound recognition: A survey, *APSIPA Transactions on Signal and Information Processing*, vol. 3, Dec. 2014. https://doi.org/10.1017/ATSIP.2014.12

[43] H. Wang, Y. Zou, D. Chong, and W. Wang, Environmental Sound Classification with Parallel Temporal-Spectral Attention, *arXiv preprint*, May 2020. https://doi.org/10.48550/arXiv.1912.06808

[44] R.M. Alsina-Pagès, J. Navarro, F. Alías, and M. Hervás, homeSound: Real-Time Audio Event Detection Based on High Performance Computing for Behaviour and Surveillance Remote Monitoring, *Sensors*, vol. 17, issue 4, 854, Apr. 2017. https://doi.org/10.3390/s17040854

[45] M.K. Saini and N. Goel, How smart are smart classrooms? A review of smart classroom technologies, *ACM Computing Surveys (CSUR)*, vol. 52, issue 6, Dec. 2019, pp. 1-28. https://doi.org/10.1145/3365757

[46] R. Togneri and D. Pullella, An Overview of Speaker Identification: Accuracy and Robustness Issues, *IEEE Circuits and Systems Magazine*, vol. 11, issue 2, June 2011, pp. 23-61. https://doi.org/10.1109/mcas.2011.941079

[47] I. Vatolkin, P. Ginsel, and G. Rudolph, "Advancements in the Music Information Retrieval Framework AMUSE over the Last Decade", *Proc. 44th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'21),* 2021, pp. 2383-2389. https://doi.org/10.1145/3404835.3463252.

[48] F.G. Encinas, L.A. Silva, A.S. Mendes, G.V. Gonzalez, V.R.Q. Leithardt, and J.F.D.P. Santana, Singular Spectrum Analysis for Source Separation in Drone-Based Audio Recording, *IEEE Access*, vol. 9, Mar. 2021, pp. 43444 – 43457. https://doi.org/10.1109/access.2021.3065775

[49] J. Salamon, C. Jacoby, and J.P. Bello, "A Dataset and Taxonomy for Urban Sound Research", *Proc. 22nd ACM Int'l Conf. on Multimedia (MM'14)*, 2014, pp. 1041-1044. https://doi.org/10.1145/2647868.2655045

[50] K.J. Piczak, "ESC: Dataset for environmental sound classification", *Proc. 23rd ACM Int'l Conf. on Multimedia (MM'15)*, 2015, pp. 1015–1018. https://doi.org/10.1145/2733373.2806390

[51] N.K. Verma, R.K. Sevakula, S. Dixit, and A. Salour, Intelligent Condition Based Monitoring Using Acoustic Signals for Air Compressors, *IEEE Transactions on Reliability*, vol. 65, issue 1, Mar. 2016, pp. 291-309. https://doi.org/10.1109/tr.2015.2459684

[52] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, Youtube-8m: A large-scale video classification benchmark, *arXiv preprint*, Sept. 2016. https://doi.org/10.48550/arXiv.1609.08675

[53] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, Features for Content-Based Audio Retrieval, *Advances in Computers*, vol. 78, 2010, pp. 71-150. https://doi.org/10.1016/s0065-2458(10)78003-7.

[54] Z. Wu, X. Wang, and B. Jiang, Fault Diagnosis for Wind Turbines Based on ReliefF and eXtreme Gradient Boosting, *Applied Sciences*, vol. 10, issue 9, May 2020. https://doi.org/10.3390/app10093258

[55] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, Classifying environmental sounds using image recognition networks, *Procedia Computer Science*, vol. 112, 2017, pp. 2048-2056. https://doi.org/10.1016/j.procs.2017.08.250

[56] L. Hertel, H. Phan, and A. Mertins, "Comparing time and frequency domain for audio event recognition using deep learning", *2016 Int'l Joint Conf.on Neural Networks (IJCNN)* , IEEE, 2016, pp. 3407-3411. https://doi.org/10.1109/IJCNN.2016.7727635

[57] R. Sharan, H. Xiong, and S. Berkovsky, Benchmarking Audio Signal Representation Techniques for Classification with Convolutional Neural Networks, *Sensors*, vol. 21, issue 10, 3434, May 2021. https://doi.org/10.3390/s21103434

[58] N.S. Neto, S. Stefenon, L. Meyer, R. Bruns, A. Nied, L. Seman, G. Gonzalez, V. Leithardt, and K.C. Yow, A Study of Multilayer Perceptron Networks Applied to Classification of Ceramic Insulators Using Ultrasound, *Applied Sciences*, vol. 11, issue 4, Feb. 2021. https://doi.org/10.3390/app11041592

[59] https://www.image-net.org/challenges/LSVRC/2012/, [Προσπελάστηκε 1/2/2022]

[60] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network", *Proc. 2nd Int'l Conf. on Neural Information Processing Systems (NIPS 89),*1989, pp. 396–404. https://dl.acm.org/doi/10.5555/2969830.2969879

[61] B.A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012, pp. 1097–1105. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[62] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv preprint*, Apr. 2015. https://doi.org/10.48550/arXiv.1409.1556

[63] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2015, pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[64] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision", *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[66] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, and K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size, *arXiv preprint*, Nov. 2016, https://doi.org/10.48550/arXiv.1602.07360

[67] F. Chollet, "Xception: Deep learning with depthwise separable convolutions", *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 1251–1258. https://doi.org/10.1109/CVPR.2017.195

[68] C. Szegedy, S. Ioffe, V. Vanhoucke, and A.A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning", *Proc. 31st AAAI Conf. on Artificial Intelligence (AAAI 17)*, AAAI Press,vol. 31, 1, 2017. https://doi.org/10.1609/aaai.v31i1.11231

[69] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, "Densely Connected Convolutional Networks", *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 4700-4708. https://doi.org/10.1109/cvpr.2017.243

[70] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger", *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2017, pp. 7263–7271. https://doi.org/10.1109/cvpr.2017.690

[71] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, and H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint*, Apr. 2017. https://doi.org/10.48550/arXiv.1704.04861

[72] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices", *Proc. 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, (CVPR)*, IEEE, 2018, p. 6848-6856. https://doi.org/10.48550/arXiv.1707.01083

[73] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks", *Proc. 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, (CVPR)*, IEEE, 2018, pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

[74] N. Ma, X. Zhang, H.T. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design", *Proc. European Conf. on Computer Vision (ECCV 2018)*, Springer, Cham, 2018, pp 122–138. https://doi.org/10.1007/978-3-030-01264-9_8

[75] A. Howard, M. Sandler, G. Chu, L.C. Chen, B. Chen, M. Tan, and H. Adam, "Searching for mobilenetv3", *Proc. 2019 IEEE/CVF Int'l Conf. on Computer Vision, (ICCV)*, IEEE, 2019, pp. 1314–1324. https://doi.org/10.1109/ICCV.2019.00140

[76] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks", *Proc. 36th Int'l Conf. Machine Learning (PMLR)*, 2019, pp. 6105-6114. https://proceedings.mlr.press/v97/tan19a/tan19a.pdf

[77] E. Cakir, G. Parascandolo, T.K. Heittola, H. Huttunen, and T. Virtanen, Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, issue 6, June 2017, pp. 1291-1303. https://doi.org/10.1109/taslp.2017.2690575

[78] S.H. Khan, M. Hayat, M. Bennamoun, F. Sohel, and R. Togneri, Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data, *IEEE transactions on Neural Networks and Learning Systems*, vol. 29, issue 8, Aug. 2018, pp. 3573 – 3587. https://doi.org/10.1109/tnnls.2017.2732482

[79] S.J. Pan and Q. Yang, A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, issue 10, Oct. 2010, pp. 1345-1359. https://doi.org/10.1109/tkde.2009.191

[80] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks", *Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2014, pp. 1717-1724. https://doi.org/10.1109/cvpr.2014.222

[81] N. Becherer, J. Pecarina, S. Nykl, and K. Hopkinson, Improving optimization of convolutional neural networks through parameter fine-tuning, *Neural Computing and Applications*, vol. 31, Aug. 2019, pp. 3469-3479. https://doi.org/10.1007/s00521-017-3285-0

[82] A. Ramezani-Kebrya, A. Khisti, and B. Liang, On the Generalization of Stochastic Gradient Descent with Momentum, *arXiv preprint*, Sept. 2021. https://doi.org/10.48550/arXiv.1809.04564

[83] N.S. Keskar, and R. Socher, Improving generalization performance by switching from adam to sgd, *arXiv preprint*, Dec. 2017. https://doi.org/10.48550/arXiv.1712.07628

[84] P. Zhou, J. Feng, C. Ma, C. Xiong, and S. Hoi, Towards theoretically understanding why sgd generalizes better than adam in deep learning, *arXiv preprint*, Nov. 2021. https://doi.org/10.48550/arXiv.2010.05627

[85] S.L. Smith, P.J. Kindermans, C. Ying, and Q.V. Le, Don't decay the learning rate, increase the batch size, *arXiv preprint*, Feb. 2018. https://doi.org/10.48550/arXiv.1711.00489

[86] E. Hoffer, I. Hubara, and D.Soudry, Train longer, generalize better: Closing the generalization gap in large batch training of neural networks, *arXiv preprint*, Jan. 2018. https://doi.org/10.48550/arXiv.1705.08741

[87] E. Tsalera, A. Papadakis, and M. Samarakou, Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning, *Journal of Sensor and Actuator Networks*, vol. 10, issue 4, Dec. 2021. https://doi.org/10.3390/jsan10040072

[88] S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R. Channing Moore, M. Plakal, D. Platt, R.A. Saurus, B. Seybold, M. Slaney, R.J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification", *Proc. 2017 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2017, pp. 131-135. https://doi.org/10.1109/ICASSP.2017.7952132

[89] J. Cramer, H.H Wu, J. Salamon, and J.P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings", *Proc. 2019 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp.3852-3856. https://doi.org/10.1109/ICASSP.2019.8682475

[90] J.W. Kim, J. Salamon, P. Li, and J.P. Bello, "Crepe: A Convolutional Representation for Pitch Estimation", *Proc. 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 161-165. https://doi.org/10.1109/ICASSP.2018.8461329

[91] C.O.S. Sorzano, J. Vargas, and A.P. Montano, A survey of dimensionality reduction techniques, *arXiv preprint*, Mar. 2014. https://doi.org/10.48550/arXiv.1403.2877

[92] H. Wang, Y. Zou, D. Chong, and W. Wang, Environmental Sound Classification with Parallel Temporal-Spectral Attention, *arXiv preprint*, May 2020. https://doi.org/10.48550/arXiv.1912.06808

[93] I.M. Pires, G. Marques, N.M. Garcia, F. Flórez-Revuelta, M.C. Teixeira, E. Zdravevski, and S. Spinsante, Recognition of Activities of Daily Living Based on a Mobile Data Source Framework, *Bio-inspired Neurocomputing*, 2021, pp.321-335. https://doi.org/10.1007/978-981-15-5495-7_18

[94] P. Wei, F. He, L. Li ,and J. Li, Research on sound classification based on SVM, *Neural Computing and Applications*, vol. 32, issue 6, Mar. 2020, pp. 1593-1607. https://doi.org/10.1007/s00521-019-04182-0

[95] A. Khamparia, D. Gupta, N.G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network, *IEEE Access*, vol. 7, Jan. 2019, pp. 7717-7727. https://doi.org/10.1109/ACCESS.2018.2888882

[96] Z. Mushtaq and S.F. Su, Environmental sound classification using regularized deep convolutional neural network with data augmentation, *Applied Acoustics*, vol. 167, Oct. 2020, pp. 107389. https://doi.org/10.1016/j.apacoust.2020.107389

[97] I. Kononenko, E. Simec, and M.Robnik-Sikonja, Overcoming the myopia of inductive learning algorithms with RELIEFF, *Applied Intelligence*, vol. 7, Jan. 1997, pp. 39-55. https://doi.org/10.1023/A:1008280620621

[98] H. Hotelling, Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, vol. 24, issue 6, Sept. 1933, pp. 417. https://doi.org/10.1037/h0071325

[99] L. Alzubaidi, J. Zhang, A.J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaria, M.A. Fadhel, M. Al-Amidie, and L. Farhan, Review of deep learning; concepts, CNN architectures, challenges, applications, future directions, *Journal of big Data*, vol. 8, Mar. 2021, pp. 1-74. https://doi.org/10.1186/s40537-021-00444-8

[100] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning", *Proc. IEEE*, vol. 109, issue 1, Jan. 2021, pp. 43-76. https://doi.org/10.1109/JPROC.2020.3004555

[101] H. Azami, H. Hassanpour, J. Escudero, and S.Sanei, An intelligent approach for variable size segmentation of non-stationary signals, *Journal of Advanced Research*, vol. 6, issue 5. Sep. 2015, pp. 687-698. https://doi.org/10.1016/j.jare.2014.03.004

[102] G. Almpanidis and C. Kotropoulos, Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion, *Speech Communication*, vol. 50, issue 1, Jan. 2008, pp. 38-55. https://doi.org/10.1016/j.specom.2007.06.005

[103] E. Tsalera, A. Papadakis, M. Samarakou, and I. Voyiatzis, "CNN-based Segmentation and Classification of Sound Streams under realistic conditions", *Proc. 26th Pan-Hellenic Conf. on Informatics* (PCI), 2022, pp. 373-378. https://doi.org/10.1145/3575879.3576020

[104] D. Preuveneers, I. Tsingenopoulos, and W.Joosen, Resource Usage and Performance Trade-offs for Machine Learning Models in Smart Environments, *Sensors*, vol. 20, issue 4, 1176, Feb. 2020. https://doi.org/10.3390/s20041176

[105] J. Han, L. Xu, M.M. Rafique, A.R. Butt , and S.H. Lim, "A Quantitative Study of Deep Learning Training on Heterogeneous Supercomputers", *2019 IEEE Int'l Conf. on Cluster Computing (CLUSTER)*, IEEE, 2019, pp. 1-12. https://doi.org/10.1109/CLUSTER.2019.8890993

[106] M.M.H. Shuvo, S.K. Islam, J. Cheng, and B.I. Morshed, "Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review", *Proc. IEEE*, vol. 111, issue 1, 2023, pp. 42-91. https://doi.org/10.1109/JPROC.2022.3226481

[107] S. Bianco, R. Cadene, L. Celona and P. Napoletano, Benchmark Analysis of Representative Deep Neural Network Architectures, *IEEE Access*, vol. 6, Oct. 2018, pp. 64270-64277. https://doi.org/10.1109/ACCESS.2018.2877890

[108] M. Lamrini, M.Y. Chkouri, and A.Touhafi, Evaluating the Performance of Pre-Trained Convolutional Neural Network for Audio Classification on Embedded Systems for Anomaly Detection in Smart Cities, *Sensors*, vol. 23, issue 13, 6227, July 2023. https://doi.org/10.3390/s23136227

[109] S.-U.-R. Baig, W. Iqbal, J.L. Berral, A. Erradi, and D. Carrera, Adaptive Prediction Models for Data Center Resources Utilization Estimation, *IEEE Transactions on Network and Service Management*, vol. 16, issue 4, Dec. 2019, pp. 1681-1693. https://doi.org/10.1109/TNSM.2019.2932840

[110] M.S. Al-Asaly, M.A. Bencherif, A. Alsanad, and M.M. Hassan, A deep learning-based resource usage prediction model for resource provisioning in an automatic cloud computing environment, *Neural Computing and Applications*, vol. 34, issue 13, Jul. 2022, pp. 10211-10228. https://doi.org/10.1007/s00521-021-06665-5

[111] G. Yang, C. Shin, J. Lee, Y. Yoo, and C. Yoo, "Prediction of the Resource Consumption of Distributed Deep Learning Systems", *Proc. ACM on Measurement and Analysis of Computing Systems*, ACM, 2022, vol. 6, issue 2, pp. 1-25. https://doi.org/10.1145/3530895

[112] V.S. Marco, B. Taylor, Z. Wang, and Y. Elkhatib, Optimizing Deep Learning Inference on Embedded Systems Through Adaptive Model Selection, *ACM Transactions on Embedded Computing Systems*, vol. 19, issue 1, Feb. 2020, pp. 1-28. https://doi.org/10.1145/3371154

[113] E. Tsalera, D. Stratogiannis, A. Papadakis, I. Voyiatzis, and M. Samarakou, "Evaluation and Prediction of Resource Usage for multi-parametric Deep Learning training and inference", *Proc. 27th Pan-Hellenic Conf. on Progress on Computing and Informatics*, 2023. https://doi.org/10.1145/3635059.3635070

[114] R. W. Picard, "Affective Computing for HCI", *Proc. 8th Int'l Conf. on Human-Computer Interaction: Ergonomics and User Interfaces (HCI)*, L. Erlbaum Associates Inc., vol. 1, 1999, pp. 829-833. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=bdf665dadcf9ec730be5340a9ca2653f41e3ad5c

[115] B. Sonawane, and P. Sharma, Review of automated emotion-based quantification of facial expression in Parkinson's patients, *The Visual Computer*, vol. 37, issue 5, May 2021, pp. 1151-1167. https://doi.org/10.1007/s00371-020-01859-9

[116] G. Mattavelli, E. Barvas, C. Longo, F. Zappini, D. Ottaviani, M.C. Malaguti, and C. Papagno, Facial expressions recognition and discrimination in Parkinson's disease, *Journal of Neuropsychology*, vol. 15, issue 1, Mar. 2021, pp. 46-68. https://doi.org/10.1111/jnp.12209

[117] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion recognition for healthcare surveillance systems using neural networks: A survey", *2021 Int'l Wireless Communications and Mobile Computing (IWCMC)*, IEEE, 2021, pp. 681-687. https://doi.org/10.1109/IWCMC51323.2021.9498861

[118] H. Kaushik, T. Kumar, and K. Bhalla, iSecureHome: A deep fusion framework for surveillance of smart homes using real-time emotion recognition, *Applied Soft Computing*, vol. 122, 108788, June 2022. https://doi.org/10.1016/j.asoc.2022.108788

[119] G. Du, Z. Wang, B. Gao, S. Mumtaz, K.M. Abualnaja, and C. Du, A convolution bidirectional long short-term memory neural network for driver emotion recognition, *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, issue 7, July 2021, pp. 4570-4578. https://doi.org10.1109/TITS.2020.3007357

[120] P. Ekman and W. V. Friesen, Facial action coding system, *Environmental Psychology & Nonverbal Behavior*, 1978. https://doi.org/10.1037/t27734-000

[121] C. Harris and M. Stephens, "A Combined Corner and Edge Detector", *Proc. 4th Alvey Vision Conf.*, 1988, pp. 147-151. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=88cdfbeb78058e0eb2613e79d1818c567f0920e2

[122] D.G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, vol. 60, issue 2, Nov. 2004, pp. 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[123] E. Rosten and T. Drummond, "Fusing Points and Lines for High Performance Tracking", *Proc. 2005 10th IEEE International Conference on Computer Vision (ICCV 05)*, IEEE, vol.1, Dec. 2005, pp. 1508–1515. https://doi.org/10.1109/ICCV.2005.104

[124] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, Speeded Up Robust Features (SURF), *Computer Vision and Image Understanding*, vol. 110, issue 3, June 2008, pp. 346-359. https://doi.org/10.1016/j.cviu.2007.09.014

[125] M. Calonder,V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features", *Proc. 11th European Conf. on Computer Vision (ECCV 2010)*, Springer, Berlin, 2010, pp.778-792. https://doi.org/10.1007/978-3-642-15561-1_56

[126] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF", *Proc. 2011 Int'l Conf. on Computer Vision (ICCV 2011)*, IEEE, 2011, pp. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544

[127] P.F. Alcantarilla, A. Bartoli, and A.J. Davison, "KAZE Features", *Proc. 12th European Conf. Computer Vision (ECCV 2012)*, Springer, vol. 7577, 2012, pp. 214-227. http://dx.doi.org/10.1007/978-3-642-33783-3_16

[128] T. Ojala, M. Pietikäinen, and D. Harwood, A comparative study of texture measures with classification based on featured distributions, *Pattern Recognition*, vol. 29, issue 1, Jan. 1996, pp. 51-59. https://doi.org/10.1016/0031-3203(95)00067-4

[129] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", *Proc. 2005 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR '05)*, IEEE, vol. 1, 2005, pp. 886-893. https://doi.org/10.1109/CVPR.2005.177

[130] S.A.K. Tareen and Z. Saleem, "A comparative analysis of sift, surf, kaze, akaze, orb, and brisk", *2018 Int'l Conf. on Computing, Mathematics and Engineering Technologies (iCoMET)*, IEEE, 2018, pp. 1-10. https://doi.org/10.1109/ICOMET.2018.8346440

[131] T.J. Alhindi, S. Kalra, K.H. Ng, A. Afrin, and H.R. Tizhoosh, "Comparing LBP, HOG and deep features for classification of histopathology images", *Int'l Joint Conf. on Neural Networks (IJCNN)*, IEEE, 2018, pp. 1-7. https://doi.org/10.1109/IJCNN.2018.8489329

[132] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, Handcrafted versus CNN features for ear recognition, *Symmetry*, vol. 11, issue 12, 1493, Dec. 2019. https://doi.org/10.3390/sym11121493

[133] W. Lin, K. Hasenstab, G. Moura Cunha, and A. Schwartzman, Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment, *Scientific Reports*, vol. 10, issue 1, 20336, Nov. 2020. https://doi.org/10.1038/s41598-020-77264-y

[134] L. Nanni, S. Ghidoni, and S. Brahnam, Handcrafted vs. non-handcrafted features for computer vision classification, *Pattern Recognition*, vol. 71, Nov. 2017, pp. 158-172. https://doi.org/10.1016/j.patcog.2017.05.025

[135] M.R. Zare, D.O. Alebiosu, and S.L. Lee, "Comparison of handcrafted features and deep learning in classification of medical x-ray images", *2018 4th Int'l Conf. on Information Retrieval and Knowledge Management (CAMP)*, IEEE, 2018, pp. 1-5. https://doi.org/10.1109/INFRKM.2018.8464688

[136] S. Agarwal, A. Rattani, and C.R. Chowdary, A comparative study on handcrafted features v/s deep features for open-set fingerprint liveness detection, *Pattern Recognition Letters*, vol. 147, July 2021, pp. 34-40. https://doi.org/10.1016/j.patrec.2021.03.032

[137] S.M.S.A. Abdullah, S.Y.A. Ameen, M.A. Sadeeq, and S. Zeebaree, Multimodal emotion recognition using deep learning, *Journal of Applied Science and Technology Trends*, vol. 2, issue 2, Apr. 2021, pp. 52-58. https://doi.org/10.38094/jastt20291

[138] M.I. Georgescu, R.T. Ionescu, and M. Popescu, Local learning with deep and handcrafted features for facial expression recognition, *IEEE Access*, vol. 7, May 2019, pp. 64827-64836. https://doi.org10.1109/ACCESS.2019.2917266

[139] B. Li, and D. Lima, Facial expression recognition via ResNet-50, *International Journal of Cognitive Computing in Engineering*, vol. 2, June 2021, pp. 57-64. https://doi.org/10.1016/j.ijcce.2021.02.002

[140] H. Zhang, A. Jolfaei, and M. Alazab, A face emotion recognition method using convolutional neural network and image edge computing, *IEEE Access*, vol. 7, Oct. 2019, pp. 159081-159089. https://doi.org/10.1109/ACCESS.2019.2949741

[141] T.U. Ahmed, S. Hossain, M.S. Hossain, R. ul Islam, and K. Andersson, "Facial expression recognition using convolutional neural network with data augmentation", *2019 Joint 8th Int'l Conf. on Informatics, Electronics & Vision (ICIEV) and 2019 3rd Int'l Conf. on Imaging, Vision and Pattern Recognition (icIVPR)*, IEEE, 2019, pp. 336-341. https://doi.org/10.1109/ICIEV.2019.8858529

[142] H. Zang, S.Y. Foo, S. Bernadin, and A. Meyer-Baese, Facial Emotion Recognition Using Asymmetric Pyramidal Networks With Gradient Centralization, *IEEE Access*, vol. 9, 2021, pp. 64487-64498. https://doi.org/10.1109/ACCESS.2021.3075389

[143] K. Li, Y. Jin, M.W. Akram, R. Han, and J. Chen, Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy, *The Visual Computer*, vol. 36, issue 2, Feb. 2020, pp. 391-404. https://doi.org/10.1007/s00371-019-01627-4

[144] D. Lundqvist, A. Flykt, and A. Öhman, Karolinska directed emotional faces, *APA PsycTESTS Dataset*, 1998, https://doi.org/10.1037/t27732-000

[145] M. J. Lyons, M. Kamachi, and J. Gyoba, Coding facial expressions with Gabor wavelets (IVC special issue), *arXiv preprint*, Sept. 2020, https://doi.org/10.48550/arXiv.2009.05938

[146] O. Langner, R. Dotsch, G. Bijlstra, D.H.J. Wigboldus, S.T. Hawk, and A. van Knippenberg, Presentation and validation of the Radboud Faces Database, *Cognition & Emotion*, vol. 24, issue 8, Nov. 2010, pp. 1377-1388. https://doi.org/10.1080/02699930903485076

[147] A. Adouani, W.M.B. Henia, and Z. Lachiri, "Comparison of Haar-like, HOG and LBP approaches for face detection in video sequences", *2019 16th Int'l Multi-Conf. on Systems, Signals & Devices (SSD)*, IEEE, 2019, pp. 266-271. https://doi.org/10.1109/SSD.2019.8893214

[148] T. Chen, T. Gao, S. Li, X. Zhang, J. Cao, D. Yao, and Y. Li, A novel face recognition method based on fusion of LBP and HOG, *IET Image Processing*, vol. 15, issue 14, Dec. 2021, pp. 3559-3572. https://doi.org/10.1049/ipr2.12192

[149] M. Sun and D. Li, Smart face identification via improved LBP and HOG features, *Internet Technology Letters*, vol. 4, issue 3, June 2021, pp. e229. https://doi.org/10.1002/itl2.229

[150] T. Ojala, M. Pietikainen, and T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, issue 7, July 2002, pp. 971-987. https://doi.org/10.1109/TPAMI.2002.1017623

[151] S.C. Kumain, M. Singh, N. Singh, and K. Kumar, "An efficient Gaussian noise reduction technique for noisy images using optimized filter approach", 2018 *Proc. 1st Int'l Conf. Secure Cyber Computing and Communication (ICSCCC)*, IEEE, 2018, pp.246-248. https://doi.org/10.1109/ICSCCC.2018.8703305

[152] B. Fu, X. Zhao, C. Song, X. Li, and X. Wang, A salt and pepper noise image denoising method based on the generative classification, *Multimedia Tools and Applications*, vol. 78, issue 9, May 2019, pp. 12043-12053. https://doi.org/10.1007/s11042-018-6732-8

[153] A. Awad, Denoising images corrupted with impulse, Gaussian, or a mixture of impulse and Gaussian noise, *Engineering Science and Technology, an International Journal*, vol. 22, issue 3, June 2019, pp. 746-753. https://doi.org/10.1016/j.jestch.2019.01.012

[154] S. Karahan, M.K. Yildirum, K. Kirtac, F.S. Rende, G. Butun, and H.K. Ekenel, "How image degradations affect deep CNN-based face recognition?", *Int'l Conf. of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2016, pp. 1-5. https://doi.org/10.1109/BIOSIG.2016.7736924

[155] V. Ziyadinov, and M. Tereshonok, Noise immunity and robustness study of image recognition using a convolutional neural network, *Sensors*, vol. 22, issue 3, 1241, Feb. 2022. https://doi.org/10.3390/s22031241

[156] H. Ren, A comprehensive study on robustness of HOG and LBP towards image distortions, *Journal of Physics: Conference Series*, vol. 1325, 012012, Nov. 2019. https://doi.org/10.1088/1742-6596/1325/1/012012

[157] E. Tsalera, A. Papadakis, M. Samarakou and I.Voyiatzis, Feature Extraction with Handcrafted Methods and Convolutional Neural Networks for Facial Emotion Recognition, *Applied Sciences*, vol. 12, issue 17, 8455, Aug. 2022. https://doi.org/10.3390/app12178455

[158] B. Berglund, T. Lindvall, and D.H. Schwela, ed., *Guidelines for community noise*, World Health Organization Occupational and Environmental Health Team, 1999. https://www.who.int/publications/i/item/a68672

[159] B.J. Mohan., N.B. Ramesh, "Speech recognition using MFCC and DTW", *2014 Int'l Conf. on Advances in Electrical Engineering (ICAEE)*, IEEE, 2014, pp. 1-4. https://doi.org/10.1109/ICAEE.2014.6838564

[160] Commission of the European Communities, *Future noise policy*, European Commission Green paper, Brussels 1996. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:51996DC0540&from=PT

[161] E. Murphy and E.A. King, Strategic environmental noise mapping: Methodological issues concerning the implementation of the EU Environmental Noise Directive and their policy implications, *Environment International*, vol. 36, issue 3, Apr. 2010, pp. 290–298. https://doi.org/10.1016/j.envint.2009.11.006

[162] S. Kephalopoulos, M. Paviotti, F. Anfosso-Lédée, D. Van Maercke, S. Shilton, and N. Jones, Advances in the development of common noise assessment methods in Europe: The CNOSSOS-EU framework for strategic environmental noise mapping, *Science of The Total Environment*, vols. 482-483, June 2014, pp. 400-410. https://doi.org/10.1016/j.scitotenv.2014.02.031

[163] K.J. Piczak, "Environmental sound classification with convolutional neural networks", 2015 *IEEE 25th Int'l Workshop on Machine Learning for Signal Processing (MLSP)*, IEEE, 2015, pp. 1–6. https://doi.org/10.1109/MLSP.2015.7324337

[164] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, Reliable detection of audio events in highly noisy environments, *Pattern Recognition Letters*, vol. 65, Nov. 2015, pp. 22-28. https://doi.org/10.1016/j.patrec.2015.06.026

[165] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M.D. Plumbley, Sound event detection and time–frequency segmentation from weakly labeled data, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, issue 4, Apr. 2019, pp. 777-787. https://doi.org/10.1109/TASLP.2019.2895254

[166] A. Mesaros, T. Heittola, and T. Virtanen, Metrics for polyphonic sound event detection, *Applied Sciences*, vol. 6, issue 6, 162, May 2016. https://doi.org/10.3390/app6060162

[167] A. Papadakis, E. Tsalera, and M. Samarakou, Survey on sound and video analysis methods for monitoring face-to-face module delivery, *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, issue 8, Apr. 2019, pp. 229–240. https://doi.org/10.3991/ijet.v14i08.9813

[168] https://freesound.org/ [Προσπελάστηκε 2/5/2020]

[169] R.G. Bachu, S. Kopparthi, B. Adapa, and B.D. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal", *2008 Proc. American Society for Engineering Education (ASEE)*, ASEE, 2008, pp. 1-7.

[170] N. Dave, Feature extraction methods LPC, PLP and MFCC in speech recognition, *International Journal for Advance Research in Engineering and Technology*, vol.1, issue 6, July 2013, pp. 1-5.

[171] G. Peeters, A Large set of audio features for sound description (similarity and classification) in the CUIDADO project, *CUIDADO 1st Project Report*, vol. 54, Apr. 2004, http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf

[172] M. Xu, L.Y. Duan, J. Cai, L.T. Chia, C. Xu, and Q. Tian, "HMM-Based audio keyword generation", *Pacific-Rim Conf. on Multimedia (PCM 2004)*, Springer, Berlin, 2004, pp. 566-574. https://doi.org/10.1007/978-3-540-30543-9_71

[173] C. Heumann, M. Schomaker, and Shalabh, *Introduction to Statistics and Data Analysis*, Springer, Cham, 2016. https://doi.org/10.1007/978-3-319-46162-5

[174] K.H. Rosen, ed., *Handbook of discrete and combinatorial mathematics*, CRC press, 1999.

[175] E. Tsalera, A. Papadakis, and M. Samarakou , Monitoring, Profiling and Classification of Urban Environmental Noise using Sound Characteristics and the KNN algorithm, *Energy Reports*, vol. 6, suppl. 6, Nov. 2020, pp. 223-230. https://doi.org/10.1016/j.egyr.2020.08.045

[176] T. Bouwmans, S. Javed, M. Sultana, and S.K. Jung, Deep neural network concepts for background subtraction: A systematic review and comparative evaluation, *Neural Networks*, vol. 117, Sept. 2019, pp. 8-66. https://doi.org/10.1016/j.neunet.2019.04.024

[177] Q. Zhang, G. Lin, Y. Zhang, G. Xu, and J. Wang, Wildland forest fire smoke detection based on faster R-CNN using synthetic smoke images, *Procedia Engineering*, vol. 211, 2018, pp. 441-446. https://doi.org/10.1016/j.proeng.2017.12.034

[178] A. Khatami, S. Mirghasemi, , A. Khosravi, C.P. Lim, and S. Nahavandi, A new PSO-based approach to fire flame detection using K-Medoids clustering, *Expert Systems with Applications*, vol. 68, Feb. 2017, pp. 69-80. https://doi.org/10.1016/j.eswa.2016.09.021

[179] B.C. Ko, J.H. Jung, and J.Y. Nam, Fire detection and 3D surface reconstruction based on stereoscopic pictures and probabilistic fuzzy logic, *Fire Safety Journal*, vol.68, Aug.2014, pp. 61-70. https://doi.org/10.1016/j.firesaf.2014.05.015

[180] C. Yuan, , Z. Liu, and Y. Zhang, "Vision-based forest fire detection in aerial images for firefighting using UAVs", *2016 Proc. Int'l Conf. on Unmanned Aircraft Systems (ICUAS)*, IEEE, 2016, pp. 1200-1205. https://doi.org/10.1109/ICUAS.2016.7502546

[181] Y. Zhao, J. Ma, X. Li, and J. Zhang, Saliency detection and deep learning-based wildfire identification in UAV imagery, *Sensors*, vol.18, issue 3, 712, Feb. 2018. https://doi.org/10.3390/s18030712

[182] A. Gómez-Ríos, S.Tabik, J. Luengo, A.S.M. Shihavuddin, B. Krawczyk, and F.Herrera, Towards highly accurate coral texture images classification using deep convolutional neural networks and data augmentation, *Expert Systems with Applications*, vol. 118, Mar. 2019, pp. 315-328. https://doi.org/10.1016/j.eswa.2018.10.010

[183] F. Bu and M.S. Gharajeh, , Intelligent and vision-based fire detection systems: A survey, *Image and Vision Computing Journal*, vol. 91, 103803, Nov. 2019. https://doi.org/10.1016/j.imavis.2019.08.007

[184] Q. Zhang, J. Xu, L. Xu, and H. Guo, , "Deep convolutional neural networks for forest fire detection", *2016 Proc. Int'l Forum on Management, Education and Information Technology Application (IFMEITA)*, Atlantis Press, 2016, pp. 568–575. https://doi.org/10.2991/ifmeita-16.2016.105

[185] M.A.I Mahmoud and H. Ren, Forest fire detection using a rule-based image processing algorithm and temporal variation, *Mathematical Problems in Engineering*, vol. 2018, Oct. 2018. https://doi.org/10.1155/2018/7612487

[186] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S.W. Baik, Efficient deep CNN-based fire detection and localization in video surveillance applications, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, IEEE, vol. 49, issue 7, July 2019, pp. 1419-1434. https://doi.org10.1109/TSMC.2018.2830099

[187] K. Avazov, M. Mukhiddinov, F. Makhmudov, and Y.I Cho, Fire Detection Method in Smart City Environments Using a Deep-Learning-Based Approach, *Electronics*, vol. 11, issue 1, 73, Dec. 2021. https://doi.org/10.3390/electronics11010073

[188] Y. Hu, J. Zhan, G. Zhou, A. Chen, W. Cai, K. Guo, Y. Hu, and L. Li, Fast forest fire smoke detection using MVMNet, *Knowledge-Based Systems*, vol. 241, 108219, Apr. 2022. https://doi.org/10.1016/j.knosys.2022.108219

[189] H. Xu, B. Li, and F. Zhong, Light-YOLOv5: A Lightweight Algorithm for Improved YOLOv5 in Complex Fire Scenarios, *Applied Sciences*, vol. 12, issue 23, 12312, Dec. 2022. https://doi.org/10.3390/app122312312

[190] J. Zhang, H. Zhu, P. Wang, and X. Ling, ATT squeeze U-Net: A lightweight network for forest fire detection and recognition, *IEEE Access*, vol. 9, Jan. 2021, pp. 10858- 10870. https://doi.org/10.1109/ACCESS.2021.3050628

[191] P. Wang, J. Zhang, and H. Zhu, Fire detection in video surveillance using superpixel-based region proposal and ESE-ShuffleNet, *Multimedia Tools and Applications*, vol. 82, Sept. 2021, pp. 1-28. https://doi.org/10.1007/s11042-021-11261-9

[192] A. Kensert, P.J. Harrison, and O. Spjuth, Transfer learning with deep convolutional neural networks for classifying cellular morphological changes, *SLAS Discovery: Advancing Life Sciences R&D*, vol. 24, issue 4, Jan. 2019, pp. 466-475. https://doi.org/10.1177/2472555218818756

[193] L.C Chen, Y. Zhu, G. Papandreou, F. Schroff and A. Hartwig, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", *2018 Proc. 18th European Conf. on Computer Vision (ECCV 2018)*, Springer, Cham, 2018, pp.833-851. https://doi.org/10.1007/978-3-030-01234-2_49

[194] G.J. Brostow, J. Fauqueur, and R. Cipolla, Semantic object classes in video: A high-definition ground truth database, *Pattern Recognition Letters*, vol. 30, issue 2, Jan. 2009, pp. 88-97. https://doi.org/10.1016/j.patrec.2008.04.005

[195] FLAME, https://ieee-dataport.org/open-access/flame-dataset-aerial-imagery-pile-burn-detection-using-drones-uavs [Προσπελάστηκε 6/12/2022]

[196] Khan, Ali, Hassan, Bilal, Khan, Somaiya, Ahmed, Ramsha, and Adnan, Abuassba, DeepFire: A Novel Dataset and Deep Transfer Learning Benchmark for Forest Fire Detection, *Mobile Information Systems*, vol. 2022, Apr. 2022. https://doi.org/10.1155/2022/5358359

[197] Fire-Flame, https://github.com/DeepQuestAI/Fire-Smoke-Dataset [Προσπελάστηκε 6/12/2022]

[198] COCO, https://cocodataset.org/#home [Προσπελάστηκε 6/01/2023]

[199] E. Tsalera, A. Papadakis, I. Voyiatzis, and M. Samarakou, CNN-based, contextualized, real-time fire detection in computational resource-constrained environments, *Energy Reports*, vol. 9, suppl. 9, Sept. 2023, pp. 247-257. https://doi.org/10.1016/j.egyr.2023.05.260

[200] M. Merenda, C. Porcaro, and D. Iero, Edge Machine Learning for AI-enabled IoT Devices: A Review, *Sensors*, vol. 20, issue 9, 2533, Apr. 2020. https://doi.org/10.3390/s20092533

[201] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random Erasing Data Augmentation", *2020 Proc. 34th AAAI Conf. on Artificial Intelligence (AAAI 20)*, AAAI Press, vol. 34, 7, 2020, pp. 13001-13008. https://doi.org/10.1609/aaai.v34i07.7000

[202] C. Shorten and T.M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data*, vol. 6, 60, July 2019, pp. 1-48. https://doi.org/10.1186/s40537-019-0197-0

[203] C. Summers and M.J. Dinneen, "Improved Mixed-Example Data Augmentation", *2019 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, IEEE, 2019, pp. 1262-1270. https://doi.org/10.1109/WACV.2019.00139