



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Πολυτροπική Αναγνώριση Συναισθήματος

Διπλωματική Εργασία

Ασκητής Μάριος Χρυσόστομος

Λεονάρδος Φρέρης

Επιβλέπων: Ποταμιάνος Γεράσιμος

Φεβρουάριος 2024



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Πολυτροπική Αναγνώριση Συναισθήματος

Διπλωματική Εργασία

Ασκητής Μάριος Χρυσόστομος

Λεονάρδος Φρέρης

Επιβλέπων: Ποταμιάνος Γεράσιμος

Φεβρουάριος 2024



UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Multimodal Emotion Recognition

Diploma Thesis

Askitis Marios Chrysostomos

Leonardos Freris

Supervisor: Potamianos Gerasimos

February 2024

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων Ποταμιάνος Γεράσιμος

Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος Δασκαλοπούλου Ασπασία

Αναπληρώτρια Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος Αργυρίου Αντώνιος

Αναπληρωτής Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και
Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μας εργασίας, θα θέλαμε να εκφράσουμε τις θερμές μας ευχαριστίες σε όσους συνέβαλλαν στην εκπόνησή της.

Ευχαριστούμε θερμά τον επιβλέπων καθηγητή μας, Ποταμιάνο Γεράσιμο, για την εμπιστοσύνη που μας έδειξε εξ' αρχής αναθέτοντάς μας το συγκεκριμένο θέμα, την καθοδήγησή του και την άριστη επικοινωνία που είχαμε από την αρχή μέχρι το τέλος.

Επίσης ευχαριστούμε τα μέλη της εξεταστικής επιτροπής, Δασκαλοπούλου Ασπασία και Αργυρίου Αντώνιο.

Τέλος θέλαμε να εκφράσουμε την ευγνωμοσύνη μας στην οικογένειά μας και τους φίλους μας για όλη τη στήριξη, τη συμπαράσταση και την κατανόησή τους, καθ' όλη τη διάρκεια των σπουδών μας.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνουμε ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μας εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχουμε χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνουμε επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνουμε πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μας ανήκει διότι είναι προϊόν λογοκλοπής».

Οι Δηλώντες

Ασκητής Μάριος Χρυσόστομος και Λεονάρδος Φρέρης

Διπλωματική Εργασία
Πολυτροπική Αναγνώριση Συναισθήματος
Ασκητής Μάριος Χρυσόστομος
Λεονάρδος Φρέρης

Περίληψη

Το πεδίο της αναγνώρισης συναισθημάτων αποτελεί τα τελευταία χρόνια αντικείμενο μελέτης και ειδικά η πολυτροπική αναγνώριση συναισθήματος έχει συναντήσει αυξημένο ενδιαφέρον. Βρίσκει εφαρμογή σε θέματα ψυχικής υγείας, σε σύγχρονες διεπαφές ανθρώπου-μηχανής και σε θέματα ασφαλείας. Βασίζεται στο γεγονός ότι υπάρχει καθολικότητα των εκφράσεων σε όλους τους ανθρώπους όλων των πολιτισμών. Οι περισσότερες εκ των πρώτων εργασιών επικεντρώθηκαν σε δεδομένα μονοτροπικού τύπου, όπως μια ηχητική έκφραση ή μια έκφραση του προσώπου. Πιο πρόσφατες προσπάθειες έχουν επικεντρωθεί στην πολυτροπική συγχώνευση, καθώς το ανθρώπινο συναίσθημα εκφράζεται μέσω πολλαπλών τρόπων, συγκεκριμένα μέσω κειμένου, εκφράσεων του προσώπου και φωνής. Η παρούσα διπλωματική εργασία επικεντρώνεται στις διαδικασίες αναγνώρισης μέσω εικόνας και ήχου ώστε να προκύψει μια συνδυαστική πρόβλεψη από βίντεο, που θα καταστήσει το δύσκολο και περίπλοκο έργο της αυτοματοποιημένης πρόβλεψης πιο αποτελεσματικό.

Αρχικά μελετάται η σχετική θεωρία και η βιβλιογραφία με τις ανάλογες μεθόδους και προσεγγίσεις για κάθε έναν από τους δύο τρόπους. Έπειτα επιλέγεται το σύνολο δεδομένων RAVDESS ως το κατάλληλο μέσο υλοποίησης των πειραμάτων. Τα βήματα που ακολουθούνται στις μονοτροπικές μεθοδολογίες είναι η εξαγωγή των χαρακτηριστικών, η προεξεργασία των δεδομένων και η δοκιμή μιας μεγάλης ποικιλίας αρχιτεκτονικών. Αυτές βασίζονται στην χρήση τεχνικών Βαθιάς Μάθησης και συγκεκριμένα τα συνελεκτικά νευρωνικά δίκτυα ανέβασαν σημαντικά την απόδοση των μεθόδων ταξινόμησης και είναι η κύρια κατεύθυνση που εξετάζει η διπλωματική.

Στη συνέχεια, παρουσιάζεται το πεδίο της πολυτροπικής αναγνώρισης συναισθημάτων. Μια σημαντική πτυχή η οποία διερευνάται είναι η συγχώνευση των τρόπων, που κατά κύριο λόγο εκτελείται μέσω συγχώνευσης σε επίπεδο χαρακτηριστικών (Early Fusion) ή σε επίπεδο απόφασης (Late Fusion). Επιλέχθηκε η προσέγγιση σε επίπεδο απόφασης και βασίζεται στα μοντέλα μονοτροπικών μεθόδων που εμφάνισαν την υψηλότερη απόδοση. Έπειτα στοχεύει

στην συγχώνευση των διανυσμάτων των πιθανοτήτων τους για την τελική πρόβλεψη. Προκειμένου να διερευνηθεί η απόδοση των προτεινόμενων μοντέλων στην αναγνώριση συναισθημάτων, υλοποιούνται και αξιολογούνται σε μια ποικιλία δεδομένων δοκιμής του συνόλου RAVDESS. Από τα πειραματικά αποτελέσματα παρατηρείται ότι με την χρήση Συγχώνευσης με Μηχανισμό Προσοχής επιτυγχάνεται η υψηλότερη ακρίβεια. Έτσι αποδεικνύεται πως ο συνδυασμός συνελεκτικών νευρωνικών δικτύων με Mel-Φασματογράμματα ως χαρακτηριστικά ήχου και ασπρόμαυρα πλαίσια εστιασμένα στο πρόσωπο ως οπτικά χαρακτηριστικά από τα βίντεο, με Συγχώνευση με Μηχανισμό Προσοχής αποτελεί μία πολύ ισχυρή και αποδοτική προσέγγιση πολυτροπικής αναγνώρισης συναισθημάτων.

Τέλος, η προτεινόμενη προσέγγιση προσαρμόζεται σε ένα πιο πρακτικό περιβάλλον, υλοποιώντας ένα σύστημα διεπαφής του χρήστη που εμφανίζει τα αποτελέσματα της πολυτροπικής αναγνώρισης συναισθημάτων σε ένα τυχαίο βίντεο του RAVDESS για κάθε ένα συναίσθημα. Ο χρήστης έχει τη δυνατότητα να εισάγει ως είσοδο το συναίσθημα που επιθυμεί να εξετάσει και να λάβει την πρόβλεψη συναισθήματος τόσο των μονοτροπικών μεθόδων όσο και της τελικής απόφασης μέσω της προτεινόμενης προσέγγισης. Συνολικά, απεικονίζονται οι διάφορες πτυχές της ανάλυσης που εκτελούνται κατά την πολυτροπική αναγνώριση συναισθημάτων.

Λέξεις-κλειδιά:

Αναγνώριση Συναισθημάτων, Μηχανική Μάθηση, Βαθιά Μάθηση, Συνελικτικά Νευρωνικά Δίκτυα, Πολυτροπική Αναγνώριση Συναισθημάτων, Ηχητική Αναγνώριση Συναισθημάτων, Οπτική Αναγνώριση Συναισθημάτων, Συγχώνευση σε Επίπεδο Απόφασης, Μηχανισμός Προσοχής

Diploma Thesis
Multimodal Emotion Recognition
Askitis Marios Chrysostomos
Leonardos Freris

Abstract

The field of emotion recognition has been a subject of study in recent years, and especially multimodal emotion recognition has attracted increased interest. It finds application in mental health problems, in modern human-machine interfaces and in security problems. It is based on the fact that there is universality of expressions across people of all cultures. Most of the early work has focused on single modality data, such as audio or images. More recent efforts have focused on multimodal fusion, as human emotion is expressed through multiple modalities, namely text, facial expressions, and voice. This Thesis focuses on image and audio recognition approaches to derive a combined prediction from video, which will make the challenging task of automated prediction more successful.

First, the relevant theory and literature are studied along with the relevant methods and approaches for each of the two modalities. Then the RAVDESS dataset is selected as the appropriate database for implementation of our experiments. The steps followed in the audio and image based methodologies are feature extraction, data pre-processing, and testing of a wide variety of architectures. These are based on the use of Deep Learning techniques and in particular the use of convolutional neural networks, which has significantly increased the performance of classification methods and is the main direction addressed in the Thesis.

Next, the field of multimodal emotion recognition is presented. An important aspect that is explored is the fusion of modalities, primarily performed through feature-level (Early Fusion) or decision-level (Late Fusion) fusion. The decision-level approach was preferred and is based on the unimodal models that showed the highest performance. It then aims to merge their probability vectors for the final prediction. In order to investigate the performance of the proposed models in emotion recognition, they are implemented and evaluated on a variety of test data of the RAVDESS dataset. From the experimental results it is observed that using Attention Mechanism Fusion achieves the highest accuracy. Thus, it is shown that the combination of convolutional neural networks with Mel-Spectograms as audio features and

grey-scale face-centered frames as visual features from the videos with Attention Mechanism Fusion is a very powerful and efficient approach for multimodal emotion recognition.

Finally, the proposed approach is adapted to a more practical environment by implementing a user interface system that displays the results of multimodal emotion recognition on any RAVDESS video for each emotion. The user is able to choose as input the emotion he/she wishes to consider and obtains the emotion prediction of both the unimodal methods and the final decision through the proposed approach. Overall, the different aspects of the analysis performed during multimodal emotion recognition are illustrated.

Keywords:

Emotion Recognition, Machine Learning, Deep Learning, Convolutional Neural Networks, Multimodal Emotion Recognition, Speech Emotion Recognition, Visual Emotion Recognition, Late Fusion, Attention Mechanism Fusion

Πίνακας περιεχομένων

Ευχαριστίες	ix
Περίληψη	xii
Abstract	xiv
Πίνακας περιεχομένων	xvii
Κατάλογος σχημάτων	xxi
Κατάλογος πινάκων	xxv
Συντομογραφίες	xxvii
1 Εισαγωγή	1
1.1 Στόχος και αντικείμενο της διπλωματικής	3
1.2 Οργάνωση του τόμου	4
2 Θεωρητικό Υπόβαθρο	5
2.1 Εισαγωγή στην Τεχνητή Νοημοσύνη	5
2.2 Μηχανική Μάθηση	6
2.3 Βαθιά Μάθηση	8
2.4 Βαθιά Νευρωνικά Δίκτυα	10
2.5 Συνελικτικά Νευρωνικά Δίκτυα (CNN)	12
2.5.1 Convolutional Layer	13
2.5.2 Pooling Layer	15
2.5.3 Fully Connected Layer	15
2.6 Αναδρομικό Νευρωνικό Δίκτυο	16

2.6.1	Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM)	17
2.7	Συναρτήσεις Ενεργοποίησης	19
2.7.1	Sigmoid	19
2.7.2	ReLU (Rectified Linear Units)	20
2.7.3	Softmax	21
2.8	Συναρτήσεις Απώλειας	22
2.8.1	Cross-Entropy Loss Function	23
2.8.2	Categorical Cross-Entropy and Sparse Categorical Cross-Entropy .	24
2.9	Βελτιστοποίηση αλγορίθμων βαθιάς μάθησης	25
2.9.1	Βελτιστοποιητής Gradient Descent	26
2.9.2	Βελτιστοποιητής Adam	28
2.10	Μείωση υπερπροσαρμογής (Dropout)	30
2.11	Υπερπαράμετροι	31
2.11.1	Υπερπαράμετροι Βελτιστοποίησης	32
2.11.2	Συγκεκριμένοι υπερπαράμετροι μοντέλου	32
2.12	Μετρικές αξιολόγησης για πρόβλημα ταξινόμησης	33
3	Σχετικές Εργασίες	35
4	Δεδομένα και Εργαλεία ανάπτυξης	39
4.1	Σύνολο δεδομένων διπλωματικής	39
4.2	Εργαλεία ανάπτυξης διπλωματικής	42
5	Αναγνώριση Συναισθήματος από Ήχο	45
5.1	Εισαγωγή	45
5.2	Φασματικά ακουστικά χαρακτηριστικά	45
5.2.1	Mel-Frequency Cepstrum Coefficients (MFCC)	46
5.2.2	Mel-Spectrogram	47
5.2.3	Σύγκριση των MFCC και Mel-Spectrogram	48
5.3	Προ-επεξεργασία Δεδομένων	49
5.4	Μεθοδολογία	53
5.4.1	CNN1D-LSTM	54
5.4.2	CNN2D	57
5.4.3	CNN2D-Augmentation Data	61

5.4.4	Σύνοψη	63
6	Αναγνώριση Συναισθήματος από Εικόνες	67
6.1	Εισαγωγή	67
6.2	Εξαγωγή Frame	68
6.3	Προ-επεξεργασία των Frames	70
6.4	Μεθοδολογία	73
6.4.1	Έγχρωμα Frames	73
6.4.2	Ασπρόμαυρα Frames	78
6.4.3	Ασπρόμαυρα Frames εστιασμένα στο πρόσωπο	80
6.4.4	Σύνοψη	82
7	Πολυτροπική Αναγνώριση Συναισθήματος	83
7.1	Εισαγωγή	83
7.2	Τεχνικές Πολυτροπικής Συγχώνευσης	84
7.3	Μεθοδολογία	86
7.3.1	Επιλογή Αποδοτικότερων Μονοτροπικών Μοντέλων	87
7.4	Τεχνικές και Πειραματικά Αποτελέσματα	89
7.4.1	Geometric Mean Fusion	90
7.4.2	F1-score	91
7.4.3	Attention Mechanism	92
7.5	Σύνοψη	93
8	Εφαρμογή Πολυτροπικής Αναγνώρισης Συναισθημάτων	95
8.1	Εισαγωγή	95
8.2	Παρουσίαση Εφαρμογής	95
9	Συμπεράσματα	99
9.1	Σύνοψη	99
9.2	Μελλοντικές επεκτάσεις	100
	Βιβλιογραφία	101

Κατάλογος σχημάτων

1.1	Διάγραμμα οπτικοποίησης του πολυτροπικού AER συστήματος. Οι τρόποι που χρησιμοποιούνται είναι η φωνή, το κείμενο και η έκφραση του προσώπου και το αποτέλεσμα είναι η πρόβλεψη του συναισθήματος του ανθρώπου.	2
2.1	Διάγραμμα Venn της τεχνητής νοημοσύνης και των υποπεδίων της.	8
2.2	Διάγραμμα απόδοσης μηχανικής και βαθιάς μάθησης σε συνάρτηση με τον όγκο δεδομένων.	9
2.3	Η εξαγωγή των χαρακτηριστικών ανάμεσα στις μεθόδους μηχανικής και βαθιάς μάθησης.	10
2.4	Νευρωνικό δίκτυο (αριστερά) εναντίον βαθιού νευρωνικού δικτύου (δεξιά).	11
2.5	Τυπική αρχιτεκτονική ενός συνελικτικού νευρωνικού δικτύου (CNN).	13
2.6	Παράδειγμα συνέλιξης.	14
2.7	Παράδειγμα εφαρμογής pooling layer.	15
2.8	LSTM Διάγραμμα.	17
2.9	Γραφική αναπαράσταση της sigmoid συνάρτησης.	20
2.10	Γραφική αναπαράσταση της ReLU συνάρτησης.	21
2.11	Γραφική αναπαράσταση της Leaky ReLU συνάρτησης.	21
2.12	Το κόκκινο υποδηλώνει απώλεια μηδέν-ένα (zero-one loss), το μπλε υποδηλώνει απώλεια άρθρωσης (hinge loss), το πράσινο υποδηλώνει τετραγωνική απώλεια άρθρωσης (squared hinge loss).	23
2.13	Αναπαράσταση της cross-entropy loss με true label ίσο με 1.	24
2.14	Σύγκριση των ρυθμών εκμάθησης κατά την εκπαίδευση.	26
2.15	Γενικός αλγόριθμος του Stochastic Gradient Descent.	27
2.16	Περιγράμματα διαφορετικής κανονικοποίησης απωλειών για ποινή ίση με 1.	28
2.17	Γενικός αλγόριθμος του Adam.	30

2.18 Ένα νευρωνικό δίκτυο πριν και μετά την εφαρμογή Dropout	31
4.1 Περιγραφή της κωδικοποίησης των αρχείων του RAVDESS.	41
4.2 Οπτικά παραδείγματα των συναισθημάτων του RAVDESS.	41
5.1 Mel-scale.	46
5.2 Mel-Spectrogram vs MFCC	48
5.3 Αναπαράσταση του φασματογράμματος Mel για ένα τυχαίο δείγμα ήχου από το σύνολο των ηχητικών σημάτων.	51
5.4 Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.	56
5.5 Confusion matrix CNN1D-LSTM	57
5.6 Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.	59
5.7 Confusion matrix CNN2D	60
5.8 Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.	61
5.9 Confusion matrix CNN2D-Augmentation Data	63
5.10 Αναπαράσταση CNN2D.	64
6.1 Ένα δείγμα χαρακτηριστικών του Haar που χρησιμοποιούνται στο Πρωτότυπο Ερευνητικό Έγγραφο που εκδόθηκε από τους Viola and Jones.	69
6.2 Ένα δείγμα frames πριν και μετά την χρήση του Haar-Cascade.	69
6.3 Augmentation των ατόφιων frames.	71
6.4 Augmentation των ασπρόμαυρων ατόφιων frames.	72
6.5 Augmentation των ασπρόμαυρων frames που εστιάζουν στο πρόσωπο.	72
6.6 Η αρχιτεκτονική του Resnet-50.	74
6.7 Η αρχιτεκτονική του Xception.	75
6.8 Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.	77
6.9 Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.	79
6.10 Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.	81
7.1 Μέθοδοι πολυτροπικής συγχώνευσης.	85
7.2 Απεικόνιση της προτεινόμενης συγχώνευσης σε επίπεδο απόφασης, συνδυάζοντας ήχο και εικόνα για την εκτέλεση μιας πρόβλεψης.	86
7.3 Confusion Matrix στη CNN2D αρχιτεκτονική με επαυξημένα δεδομένα εκπαίδευσης, για τα αρχεία δοκιμής ήχου	87

7.4	Confusion Matrix στη CNN2D αρχιτεκτονική των ασπρόμαυρων εστιασμένων στο πρόσωπο frames για τα αρχεία δοκιμής.	88
7.5	Confusion Matrix για το Posterior Probability Sum	89
7.6	Confusion Matrix για το Geometric mean fusion.	90
7.7	Confusion Matrix για το f1 score fusion.	91
7.8	Confusion Matrix για το Attention Mechanism Fusion.	92
8.1	Αρχικό παράθυρο εφαρμογής.	96
8.2	Παράθυρο αποτελεσμάτων εφαρμογής.	97

Κατάλογος πινάκων

5.1	CNN1D-LSTM Model Architecture	55
5.2	Classification report CNN1D-LSTM	56
5.3	CNN2D Model Architecture	59
5.4	Classification report CNN1D-LSTM	60
5.5	Classification report CNN2D-Augmentation Data	62
5.6	Σύνοψη των μοντέλων για αναγνώριση συναισθημάτων από ήχο.	63
6.1	CNN2D Model Architecture	77
6.2	Classification report του CNN2D στα έγχρωμα ατόφια frames.	78
6.3	Classification report του CNN2D σε ασπρόμαυρα ατόφια frames.	79
6.4	Classification report του CNN2D για ασπρόμαυρα frames που εστιάζουν στο πρόσωπο.	81
6.5	Σύνοψη των μοντέλων για αναγνώριση συναισθημάτων από εικόνες.	82
7.1	Σύνοψη των fusion μοντέλων.	93

Συντομογραφίες

AER	Automatic Emotion Recognition
TN	Τεχνητή Νοημοσύνη
ML	Machine Learning
DL	Deep Learning
GPU	Graphics Processing Unit
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
RGB	Red Green Blue
LSTM	Long Short-Term Memory
FC	Fully Connected
SGD	Stochastic Gradient Descent
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
SVM	Support Vector Machine
HMM	Hidden Markov Model
FFT	Fast Fourier Transform
SER	Speech Emotion Recognition
FER	Facial Emotion Recognition
PCA	Principal Component Analysis
MER	Multimodal Emotion Recognition

Κεφάλαιο 1

Εισαγωγή

Τα συναισθήματα είναι ένα έμφυτο χαρακτηριστικό της ανθρώπινης δραστηριότητας και αποτελούν ένα σημαντικό παράγοντα στην διαδικασία λήψης αποφάσεων, στις γνωστικές διαδικασίες και στην αλληλεπίδραση με άλλους ανθρώπους. Η αναγνώριση συναισθημάτων οδηγεί πιο κοντά στην πλήρη αλληλεπίδραση μεταξύ ανθρώπου και μηχανής [1].

Για αυτόν τον λόγο, τις τελευταίες δεκαετίες το ενδιαφέρον στη διερεύνηση της συναισθηματικής νοημοσύνης έχει αυξηθεί στην επιστημονική κοινότητα, με συνέπεια την εφεύρεση του πεδίου αυτόματης αναγνώρισης συναισθημάτων ή όπως είναι γνωστό ως Automatic Emotion Recognition (AER). Ένας κύριος λόγος εφαρμογής του πεδίου εντοπίζεται στα περιβάλλοντα όπου οι μηχανές χρειάζεται να αλληλοεπιδράσουν ή να παρακολουθήσουν ανθρώπους. Πιο συγκεκριμένα παραδείγματα χρήσης αυτής της τεχνολογίας συναντώνται στην ανάπτυξη της ψυχικής υγείας, στις σύγχρονες διεπαφές ανθρώπου-μηχανής, όπως τα chat-bots και οι φωνητικοί βοηθοί, διαδικτυακά παιχνίδια, ψηφιακή διαφήμιση, ανίχνευση ρητορικής μίσους στα μέσα κοινωνικής δικτύωσης, συναισθηματικά συστήματα μάθησης κλπ. [2, 3, 4].

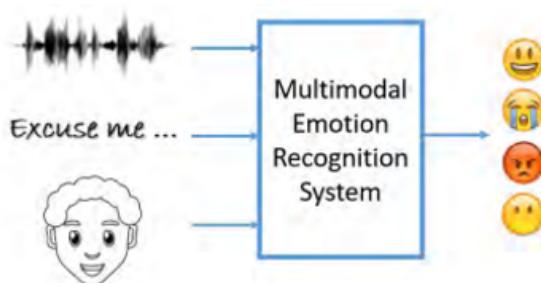
Τα συναισθήματα αυτά αποτελούν ένα υποσύνολο του Sentiment Analysis/Opinion Mining και διακρίνονται, όπως τα έχει καθορίσει ο Ekman [5], στα έξι βασικά συναισθήματα:

- Θυμός (Anger)
- Αηδία (Disgust)
- Φόβος (Fear)
- Χαρά (Happiness)

- Λύπη (Sadness)
- Έκπληξη (Surprise)

Ωστόσο, το έργο της αναγνώρισης συναισθημάτων αποτελεί πρόκληση επειδή το ανθρώπινο συναισθημα είναι πολύ σύνθετο από τη φύση του, κι αυτό γιατί δεν διαθέτει χρονικά όρια, είναι εξαιρετικά προσωπικό και διαφορετικά άτομα εκφράζουν το συναισθημα με διαφορετικούς τρόπους. Επίσης ο μεγάλος όγκος των εκφράσεων του προσώπου, των χαρακτηριστικών φωνής, των γλωσσικών περιεχομένων της λεκτικής επικοινωνίας και των στάσεων του σώματος που είναι διαθέσιμα στον Παγκόσμιο Ιστό, αποδεικνύουν ότι υπάρχει ανάγκη για αποτελεσματικές αναλύσεις για την αναγνώριση των συναισθημάτων. Κατά συνέπεια, δεδομένου ότι η αναγνώριση συναισθημάτων είναι ένα πολύ δύσκολο έργο ακόμη και για τους ανθρώπους, είναι ακόμη πιο πολύπλοκο για τις αυτοματοποιημένες μεθόδους.

Σύμφωνα με πρόσφατες αναφορές ένα πολυτροπικό AER εμφανίζεται πιο αποδοτικό, καθώς αυτό βρίσκεται πιο κοντά στο ανθρώπινο σύστημα συναισθημάτων [6, 7]. Συλλέγοντας και συνδυάζοντας πληροφορίες όπου η μία συμπληρώνει την άλλη από πολλαπλούς τρόπους, εξάγεται μια πιο ισχυρή πληροφορία και δημιουργείται ένα πιο αξιόπιστο σύστημα. Παρόλα αυτά η πολυτροπική αναγνώριση συναισθημάτων εμφανίζει την πρόκληση ποιων τρόπων να συνδυαστούν και πως. Αυτή τη στιγμή, υπάρχει έλλειψη κοινής παραδοχής στο ποιος μηχανισμός συνδυασμού είναι πιο αποδοτικός γνωστό ως fusion [8].



Σχήμα 1.1: Διάγραμμα οπτικοποίησης του πολυτροπικού AER συστήματος. Οι τρόποι που χρησιμοποιούνται είναι η φωνή, το κείμενο και η έκφραση του προσώπου και το αποτέλεσμα είναι η πρόβλεψη του συναισθήματος του ανθρώπου.

(Πηγή Σχήματος: [9])

1.1 Στόχος και αντικείμενο της διπλωματικής

Στην παρούσα διπλωματική εργασία μελετάται το πεδίο της Βαθιάς Μάθησης (Deep Learning) και της συγχώνευσης πληροφοριών σε βάθος, με σκοπό να επιτευχθεί η αναγνώριση συναισθημάτων και η επέκτασή της με τη χρήση πολλαπλών τρόπων. Συνολικά τα δεδομένα στον τομέα αυτό, τα οποία έχουν γνωρίσει μεγάλη αύξηση τόσο στο ενδιαφέρον της κοινότητας όσο και στην διαθεσιμότητά τους, περιλαμβάνουν σημαντική πληροφορία για να αναλυθεί και να εξαχθεί η γνώση που θα χρειαστεί στις αυτοματοποιημένες τεχνικές. Για παράδειγμα η ακριβής αναγνώριση των συναισθημάτων ξεχωριστά από εικόνες ή ήχο είναι κάτι αρκετά δύσκολο λόγω του περιορισμένου αριθμού διαθέσιμων δεδομένων [10].

Γι' αυτόν τον λόγο, κρίνεται αναγκαίος ο συνδυασμός δεδομένων ήχου και εικόνων μέσω υβριδικών τεχνικών. Κατ' αρχάς, πραγματεύονται οι μονοτροπικές διαδικασίες αναγνώρισης συναισθημάτων από ήχο και εικόνες. Στη συνέχεια, παρουσιάζονται οι πολυτροπικές μέθοδοι αναγνώρισης συναισθημάτων πραγματοποιώντας συγχώνευση των παραπάνω δύο ενοτήτων. Είναι σημαντικό να αναφερθεί ότι διερευνούνται διαφορετικοί μέθοδοι συγχώνευσης των δύο τρόπων (εικόνας-ήχου) και επεκτείνονται σύμφωνα με τις προκλήσεις που αντιμετωπίστηκαν. Ως αποτέλεσμα, εισάγεται η θεωρία πίσω από τα προαναφερθέντα πεδία, η σχετική βιβλιογραφία και το μαθηματικό υπόβαθρο των μοντέλων. Έπειτα θα παρουσιαστεί μια ποικιλία μοντέλων που κυμαίνονται από απλά μέχρι προσεγγίσεις τελευταίας τεχνολογίας, με στόχο την εφαρμογή τους σε δεδομένα πραγματικού κόσμου.

Συνοψίζοντας, σχεδιάζεται μια αλληλεπίδραση του χρήστη με γραφικό περιβάλλον το οποίο θα περιλαμβάνει τις πιο αποτελεσματικές μεθόδους αναγνώρισης συναισθήματος τόσο με τη χρήση των μονοτροπικών μεθόδων όσο και με του προτεινόμενου υβριδικού μοντέλου. Έτσι θα επισημανθεί η βελτιωμένη απόδοση στην αναγνώριση των συναισθημάτων μέσω της πολυτροπικής προσέγγισης, σε σύγκριση με αυτές που επιτυγχάνεται από τους δύο τρόπους ξεχωριστά.

1.2 Οργάνωση του τόμου

Το υπόλοιπο της διπλωματικής έχει την ακόλουθη διάρθρωση: Στο Κεφάλαιο 2 παρουσιάζονται οι ορισμοί και η θεωρία πίσω από τα Βαθιά Νευρωνικά Δίκτυα. Στο Κεφάλαιο 3 παρουσιάζονται ιστορικά θέματα, συναφείς εργασίες, μέθοδοι και προσεγγίσεις σχετικά με οπτικά, ηχητικά και πολυυτροπικά μοντέλα αναγνώρισης συναισθημάτων. Στο Κεφάλαιο 4 περιγράφονται τα χαρακτηριστικά και οι λόγοι επιλογής της βάσης δεδομένων αλλά και τα εργαλεία ανάπτυξης του προγραμματιστικού σκέλους της εργασίας. Στα Κεφάλαια 5 και 6 παραθέτονται οι προτεινόμενες μεθοδολογίες, η υλοποίηση αυτών και το pipeline για την επίτευξη της αναγνώρισης συναισθημάτων από τους ξεχωριστούς τρόπους (εικονα-ήχος). Στο Κεφάλαιο 7 ακολουθούν οι τρόποι συγχώνευσης (Fusion) των καλύτερων μοντέλων από τα κεφάλαια 5, 6 και η προτεινόμενη προσέγγιση. Τέλος στο Κεφάλαιο 8 υλοποιείται η εφαρμογή για την αλληλεπίδραση του χρήστη με όλα τα παραπάνω και στο Κεφάλαιο 9 παραθέτονται τα συμπεράσματα και οι μελλοντικοί στόχοι της εργασίας.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Αυτό το κεφάλαιο πραγματεύεται την τεχνητή νοημοσύνη αποδομώντας και αναλύοντας τα επιμέρους υποσύνολά της, με σκοπό την κατανόηση δομικών όρων και αλγορίθμων που χρησιμοποιούνται καθ' όλη τη διάρκεια της διπλωματικής και οδηγούν στην ολοκλήρωση του AER. Τα υποσύνολα αυτά ξεκινώντας από την Μηχανική Μάθηση (Machine Learning) και σε εξέλιξη της στην Βαθιά Μάθηση (Deep Learning) χωρίζονται σε μάθηση με επίβλεψη (Supervised Learning), μάθηση χωρίς επίβλεψη (Unsupervised Learning) και στην ενισχυτική μάθηση (Reinforcement Learning). Περισσότερο έμφαση δίνεται στην θεωρητική ανάλυση της εκπαίδευσης και στην βελτιστοποίηση των αλγορίθμων βαθιάς μάθησης.

2.1 Εισαγωγή στην Τεχνητή Νοημοσύνη

Η τεχνητή νοημοσύνη (TN) αποτελεί μέρος της επιστήμης των υπολογιστών, ενώ από μόνη της απαρτίζεται από έναν τεράστιο αριθμό επιμέρους τομέων. Μέσω αυτής υπάρχει η δυνατότητα δημιουργίας ευφυών μηχανών που μπορούν να συμπεριφέρονται και να σκέφτονται όπως οι άνθρωποι, καθώς και να λαμβάνουν αποφάσεις ανεξάρτητα. Ο ανθρώπινος παράγοντας είναι ζωτικής σημασίας στην TN, διότι η γλώσσα της “μηχανής” αποτελείται από το 0 και το 1. Έτσι, ο άνθρωπος είναι απαραίτητο να δημιουργήσει ένα κατάλληλα ειδικό περιβάλλον επικοινωνίας μεταξύ αυτού και αυτής για να είναι ικανή να παράγει σωστά αποτελέσματα. Επίσης είναι σημαντικό να αναφερθεί ότι η τεχνητή νοημοσύνη βασίζεται σε μεγάλο βαθμό στον όγκο και στην ανάλυση των παρεχόμενων δεδομένων, καθώς η ευφυΐα που θα αποκτήσει ο αλγόριθμος είναι ανάλογη με την ποσότητα και την ποιότητα των δεδομένων. Αυτό που πρέπει να κατανοηθεί είναι ότι η τεχνητή νοημοσύνη είναι μια τεχνολογία

που προσαρμόζεται στις ανάγκες των ανθρώπων και γενικά είναι σχεδιασμένη στην φιλοσοφία της ανθρώπινης εκμάθησης.

2.2 Μηχανική Μάθηση

Η γνωστή σε πολλούς έννοια του Machine Learning (ML) θεωρείται υποπεδίο της Τεχνητής Νοημοσύνης. Αρχικά, η Μηχανική Μάθηση είναι η μελέτη αλγορίθμων, υπολογιστικών μεθόδων και στατιστικών μοντέλων που χρησιμοποιούν προηγούμενη γνώση για την εκτέλεση μιας συγκεκριμένης εργασίας. Η βασική ιδέα είναι η δημιουργία ενός μοντέλου που μπορεί να να παράγει προβλέψεις και να λαμβάνει αποφάσεις με βάση ορισμένους αλγορίθμους χωρίς να είναι ρητά προγραμματισμένο από έναν χρήστη για μια συγκεκριμένη εργασία. Κατά μία έννοια ο στόχος είναι να “εκπαιδευτεί” το μοντέλο αυτό σε ένα σύνολο δεδομένων, ώστε να μπορεί να αποκτήσει και να μεταφέρει τις γνώσεις αυτές σε νέα δεδομένα εισόδου αργότερα. Οι αλγόριθμοι λοιπόν που χρησιμοποιούνται στη Μηχανική Μάθηση προέρχονται από τα πεδία της στατιστικής και των μαθηματικών με σκοπό να κατασκευάσουν ένα μοντέλο λαμβάνοντας δεδομένα δείγματος (“δεδομένα εκπαίδευσης”) ως είσοδο. Ένα από τα κύρια οφέλη είναι ότι η διαδικασία εξαγωγής γνώσης από τα δεδομένα είναι εντελώς αυτόματη. Υπάρχουν διαφορετικοί τρόποι προσέγγισης της μάθησης, ανάλογα με την φύση του προβλήματος, όπου διακρίνονται σε μάθηση με επίβλεψη, μάθηση χωρίς επίβλεψη, μάθηση με ημιεπίβλεψη και ενισχυτική μάθηση.

Μάθηση με επίβλεψη: Το χαρακτηριστικό της μάθησης με επίβλεψη είναι ότι οι περιπτώσεις στο σύνολο δεδομένων εκπαιδευσης είναι επισημασμένες (έχουν ετικέτες). Συνήθως οι ετικέτες είναι ετικέτες κλάσεων σε προβλήματα ταξινόμησης. Ο στόχος αυτού του τύπου μάθησης είναι η δημιουργία μιας συνάρτησης που αντιστοιχίζει τις εισόδους στις επιθυμητές εξόδους και επομένως να προκύψουν μοντέλα που μπορούν να χρησιμοποιηθούν για την ταξινόμηση άλλων δεδομένων χωρίς ετικέτες [11].

Μάθηση χωρίς επίβλεψη: Η μάθηση χωρίς επίβλεψη χρησιμοποιείται κυρίως για την ομαδοποίηση και την μείωση της διάστασης των χαρακτηριστικών. Σε αυτόν τον τύπο μάθησης

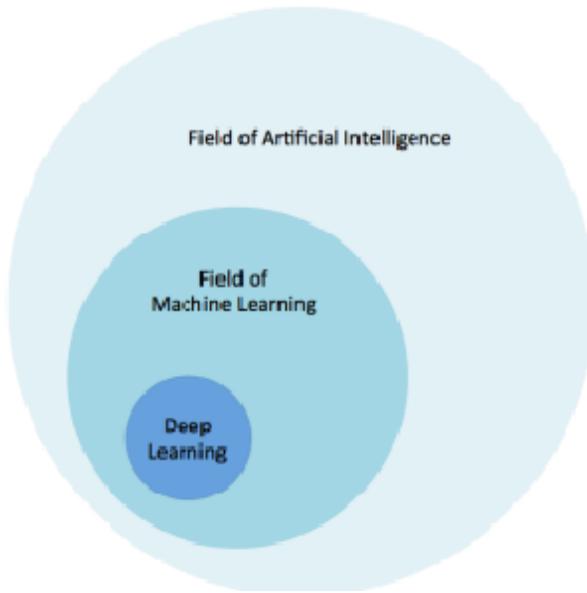
οι είσοδοι δεν είναι επισημασμένες (χωρίς ετικέτα) και ο στόχος είναι να βρεθούν κρυμμένα μοτίβα στα δεδομένα [12, 13].

Μάθηση με ημιεπίβλεψη: Αυτός ο τύπος μάθησης είναι ένας συνδυασμός επιβλεπόμενης και μη επιβλεπόμενης μάθησης και χρησιμοποιεί μια μίξη επισημασμένων και μη επισημασμένων εισόδων [12].

Ενισχυτική μάθηση: Στην ενισχυτική μάθηση, ο αλγόριθμος λαμβάνει ανατροφοδότηση από το περιβάλλον μόνο αφού επιλέξει μια έξοδο για κάθε παρατήρηση δεδομένων. Χρησιμοποιείται συνήθως σε διαδοχικά προβλήματα λήψης αποφάσεων [14].

Το ενδιαφέρον για τις μεθόδους μηχανικής μάθησης αυξήθηκε σημαντικά παράλληλα με την αύξηση του μεγέθους των δεδομένων, στα οποία έχει γίνει ευκολότερη η πρόσβαση. Τα κοινωνικά δίκτυα, οι ιστοσελίδες, τα φόροντα, οι διαδικτυακές μεγάλες βάσεις καθώς και άλλες εφαρμογές έχουν συναντήσει μεγάλο ενδιαφέρον και έτσι τα δεδομένα (κείμενα, φωτογραφίες, βίντεο κλπ.) που μπορούν να ληφθούν από αυτά είναι τόσο μεγάλα που αναφέρονται ως “μεγάλα δεδομένα” (Big Data). Ως αποτέλεσμα ένα νέο υποπεδίο της Μηχανικής Μάθησης έχει εμφανιστεί, που ονομάζεται Βαθιά Μάθηση (Deep Learning) και ξεπερνά τον προκάτοχό του επιτυγχάνοντας πολύ υψηλότερες επιδόσεις σε μεγάλα δεδομένα. Το DL χρησιμοποιεί αλγορίθμους με πολλαπλά επίπεδα, γνωστά ως νευρωνικά δίκτυα και ένα τυπικό μοντέλο έχει τουλάχιστον τρία επίπεδα. Κάθε επίπεδο δέχεται τις πληροφορίες από το προηγούμενο και τις μεταβιβάζει στο επόμενο. Τα ML & DL βρίσκονται στο προσκήνιο της επιστήμης των δεδομένων, των μεγάλων δεδομένων και των τεχνολογικών καινοτομιών.

Μία ακόμα προσέγγιση μηχανικής μάθησης που αξίζει να αναφερθεί είναι το Ensemble Learning. Η ιδέα πίσω από το Ensemble Learning είναι ότι υπάρχει πιθανότητα να αυξηθεί η ακρίβεια της ταξινόμησης μέσω συνδυασμού των αλγορίθμων μάθησης, σε σύγκριση με την απόδοση που θα επιτυγχάνονταν από το μεμονωμένα καλύτερο αλγόριθμο [15].

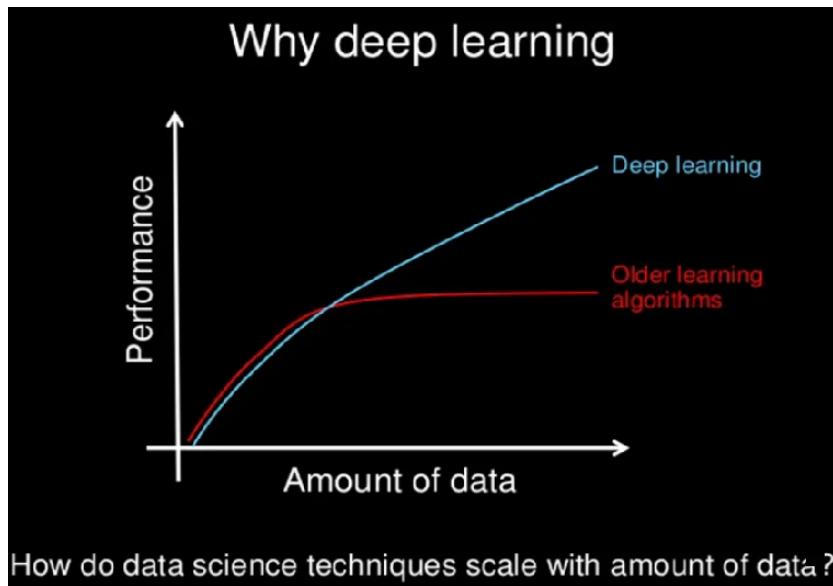


Σχήμα 2.1: Διάγραμμα Venn της τεχνητής νοημοσύνης και των υποπεδίων της.
(Πηγή Σχήματος: [16])

2.3 Βαθιά Μάθηση

Ο όγκος των πληροφοριών που εκθέτουν οι χρήστες εξαιτίας του παγκόσμιου ιστού την τελευταία δεκαετία έχει αυξηθεί εκθετικά. Επομένως, όπως αναφέρθηκε προηγουμένως χάρη στα κοινωνικά δίκτυα και τις εφαρμογές που στηρίζονται στον παγκόσμιο ιστό, έχουμε να κάνουμε πλέον με τα λεγόμενα μεγάλα δεδομένα (Big Data). Κατά συνέπεια, τα παραδοσιακά μοντέλα και οι τεχνικές μηχανικής μάθησης δεν επαρκούν για την αντιμετώπιση αυτής της πρόκλησης. Η δραματική αύξηση της διαθεσιμότητας των δεδομένων και η δραματική μείωση του υπολογιστικού κόστους (ισχυρές GPU) οδήγησαν στην εμφάνιση ενός νέου υποπεδίου της Μηχανικής Μάθησης, που ονομάζεται Βαθιά Μάθηση, και ξεπερνά τον προκάτοχό της επιτυγχάνοντας πολύ υψηλή ακρίβεια σε μεγάλα δεδομένα.

Η βαθιά μάθηση ασχολείται με αλγόριθμους εμπνευσμένους από τη δομή και τη λειτουργία του εγκεφάλου, που ονομάζονται τεχνητά νευρωνικά δίκτυα. Οι αλγόριθμοι βαθιάς μάθησης είναι παρόμοιοι με τον τρόπο δομής του νευρικού συστήματος όπου οι νευρώνες συνδέονται μεταξύ τους και μεταφέρουν την πληροφορία. Με άλλα λόγια, αντικατοπτρίζουν τη λειτουργία του εγκεφάλου. Έτσι έχει αποδειχθεί ότι αυτού του είδους οι αλγόριθμοι τείνουν να αποδίδουν καλύτερα με τον όγκο δεδομένων, σε αντίθεση με τα μοντέλα μηχανικής μάθησης που σταματούν να βελτιώνονται μετά από ένα σημείο κορεσμού [17].

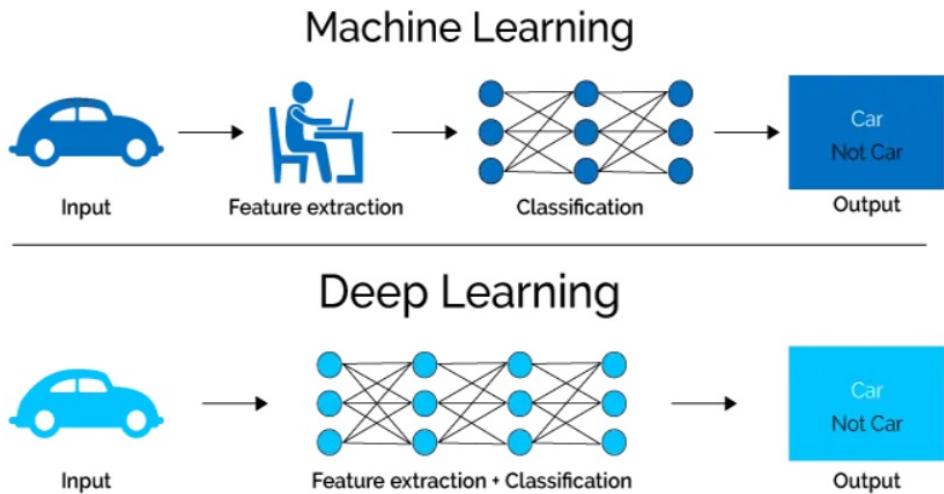


Σχήμα 2.2: Διάγραμμα απόδοσης μηχανικής και βαθιάς μάθησης σε συνάρτηση με τον όγκο δεδομένων.

(Πηγή Σχήματος: [17])

Οι αλγόριθμοι βαθιάς μάθησης έχουν μια αυξανόμενη ιεραρχία πολυπλοκότητας. Κάθε αλγόριθμος εφαρμόζει έναν μη γραμμικό μετασχηματισμό στην είσοδό του και μέσω αυτού “μαθαίνει” για να δημιουργήσει ένα στατιστικό μοντέλο ως έξοδο. Η διαδικασία επαναλαμβάνεται έως ότου η έξοδος να έχει ένα αποδεκτό επίπεδο ακρίβειας. Στη βαθιά μάθηση, το μοντέλο μαθαίνει από μόνο του ανακαλύπτοντας τα μοτίβα μεταξύ των μεταβλητών. Οι άνθρωποι δεν παρεμβαίνουν, το μοντέλο είναι ικανό να κάνει την επιλογή των χαρακτηριστικών από μόνο του. Μερικές φορές είναι περίπλοκο να εξάγονται υψηλού επιπέδου αφηρημένα χαρακτηριστικά από τα ακατέργαστα δεδομένα, γεγονός που αποτελούσε πρόβλημα στο ML, καθώς απαιτούσε την εξειδίκευση του μηχανικού από τον εκάστοτε τομέα που προέρχονταν τα δεδομένα. Η βαθιά μάθηση λύνει αυτό το πρόβλημα εκφράζοντας αυτά τα πολύπλοκα χαρακτηριστικά με όρους συνδυασμών απλούστερων χαρακτηριστικών.

Συμπεραίνουμε, λοιπόν, ότι μία ακόμα διαφορά ανάμεσα στα δύο αυτά υποπεδία της τεχνητής νοημοσύνης συναντάται στην εξαγωγή των χαρακτηριστικών. Η εξαγωγή χαρακτηριστικών γίνεται από τον άνθρωπο στη μηχανική μάθηση, ενώ στο μοντέλο βαθιάς μάθησης από μόνο του.



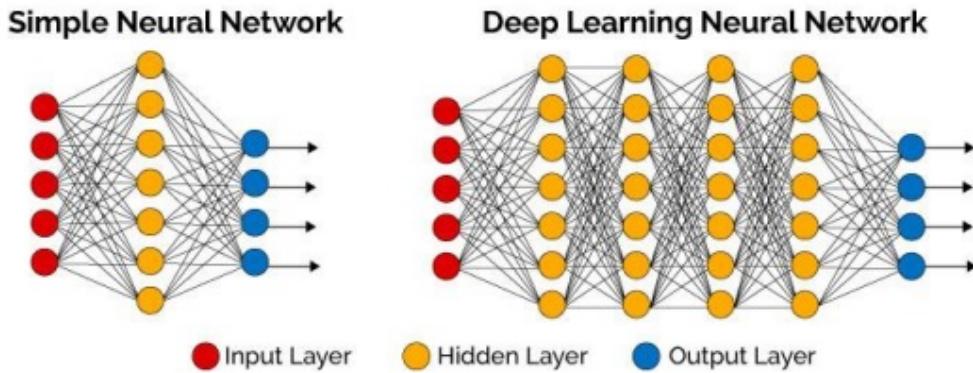
Σχήμα 2.3: Η εξαγωγή των χαρακτηριστικών ανάμεσα στις μεθόδους μηχανικής και βαθιάς μάθησης.

(Πηγή Σχήματος: [18])

Ως αποτέλεσμα, οι εφαρμογές είναι ατελείωτες και ορισμένοι τομείς στους οποίους η βαθιά μάθηση έχει μέγιστη απόδοση είναι η όραση υπολογιστών, η βιοπληροφορική και η αναγνώριση ομιλίας. Η πλειονότητα των διαγωνισμών και των προκλήσεων σε όλους τους τομείς κερδίζονται από μοντέλα Βαθιάς Μάθησης [19, 20]. Είναι σημαντικό να αναφερθεί ότι οι μέθοδοι DL χρησιμοποιούνται εκτενώς για την αναγνώριση συναισθημάτων. Για παράδειγμα, θα μπορούσε να εκτελεστεί ταξινόμηση ήχων ή εικόνων σε προκαθορισμένες κατηγορίες. Γεγονός που αποτελεί και το κύριο αντικείμενο της παρούσας εργασίας.

2.4 Βαθιά Νευρωνικά Δίκτυα

Τα κλασικά νευρωνικά δίκτυα είναι ένας αλγόριθμος Μηχανικής Μάθησης και όχι Βαθιάς Μάθησης. Ένα βαθύ νευρωνικό δίκτυο είναι ένα τεχνητό νευρωνικό δίκτυο με πολλαπλά στρώματα μεταξύ της εισόδου και εξόδου, τα οποία μπορούν να εκμεταλλευτούν τη δύναμη αναπαράστασης ενός νευρωνικού δικτύου σε μεγαλύτερο βαθμό.



Σχήμα 2.4: Νευρωνικό δίκτυο (αριστερά) εναντίον βαθιού νευρωνικού δικτύου (δεξιά).

(Πηγή Σχήματος: [21])

Κάθε στρώμα (layer) μπορεί να θεωρηθεί ως ένας μαθηματικός χειρισμός με τα στρώματα πιο κοντά στην είσοδο να μαθαίνουν απλά χαρακτηριστικά, ενώ τα επόμενα να μαθαίνουν πιο σύνθετα χαρακτηριστικά που προέρχονται από τα προηγούμενα. Τα σύνθετα νευρωνικά δίκτυα έχουν πολλά στρώματα, εξού και ο όρος “βαθιά” μάθηση/δίκτυο. Ο στόχος είναι το δίκτυο να εκπαιδευτεί αρκετά στα χαρακτηριστικά, ώστε να εντοπίζει τάσεις που υπάρχουν σε όλα τα δείγματα και στη συνέχεια να είναι σε θέση να ταξινομεί μη επισημασμένες περιπτώσεις. Τα βαθιά νευρωνικά δίκτυα μπορούν να μοντελοποιήσουν πολύπλοκες μη γραμμικές σχέσεις και τα επιπλέον layers επιτρέπουν τη σύνθεση χαρακτηριστικών από χαμηλότερα layers, μοντελοποιώντας δυνητικά πολύπλοκα δεδομένα με λιγότερες μονάδες (hidden units) σε σχέση με ένα ρηχό δίκτυο με παρόμοιες επιδόσεις [22].

Οι αρχιτεκτονικές στη βαθιά μάθηση έρχονται σε πολλές παραλλαγές ορισμένων βασικών προσεγγίσεων και κάθε αρχιτεκτονική έχει βρει επιτυχία σε συγκεκριμένους τομείς. Τυπικά, ένα βαθύ νευρωνικό δίκτυο είναι τροφοδοτούμενο, στο οποίο τα δεδομένα ρέουν από το layer εισόδου στο layer εξόδου. Αρχικά, δημιουργείται μια ένωση μεταξύ των νευρώνων και ανατίθενται τυχαία βάρη στις συνδέσεις μεταξύ τους. Αυτά τα βάρη παίρνουν τιμές μεταξύ 0 και 1 και προσαρμόζονται καθ' όλη τη διάρκεια εκτέλεσης του αλγορίθμου, ανάλογα με τα δεδομένα εκπαίδευσης. Παρόλο που τα βαθιά νευρωνικά δίκτυα έχουν πληθώρα πλεονεκτημάτων, αντιμετωπίζουν επίσης τις ακόλουθες προκλήσεις:

- Ερμηνευσιμότητα (Interpretability)
- Υπερπροσαρμογή (Overfitting)

- Χρόνος Υπολογισμού
- Πολλοί Παράμετροι
- Δυσκολία Προσθήκης Νέων Δεδομένων

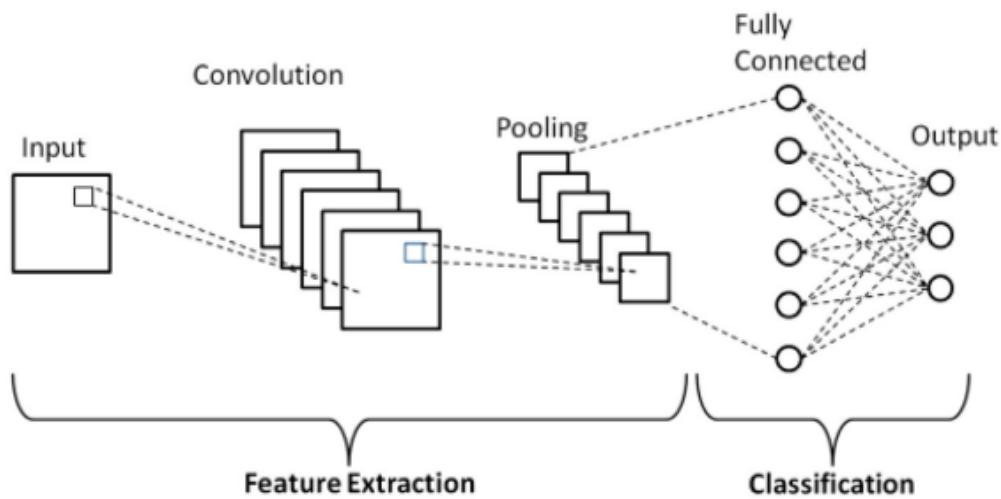
Πρώτον, υπάρχουν προβλήματα με την ερμηνευσιμότητα των αποφάσεών τους. Λειτουργούν ως μαύρο κουτί (black box) και ακόμη και οι σχεδιαστές τους και οι καλύτεροι μαθηματικοί δεν μπορούν να εξηγήσουν γιατί κατέληξαν σε μια συγκεκριμένη απόφαση. Αυτό είναι ένα ζήτημα επειδή ένα μοντέλο τεχνητής νοημοσύνης θα πρέπει να δίνει λεπτομερείς εξηγήσεις για τις ενέργειές του, που να μπορούν να κατανοήσουν οι άνθρωποι. Δεύτερον, είναι επιρρεπείς στην υπερπροσαρμογή λόγω των πρόσθετων επιπέδων αφαίρεσης, τα οποία τους επιτρέπουν να μοντελοποιούν σπάνιες εξαρτήσεις στα δεδομένα εκπαίδευσης. Οι τρόποι καταπολέμησης αυτού του φαινομένου είναι η μείωση του βάρους και η εγκατάλειψη νευρώνων (Dropout). Τρίτον, ο χρόνος υπολογισμού είναι σημαντικά υψηλότερος από οποιοδήποτε άλλη τεχνική μάθησης. Οι μονάδες (hidden units) επεξεργασίας με μεγάλες δυνατότητες μπορούν να επιταχύνουν την διαδικασία εκπαίδευσης. Τέταρτον, ένα βαθύ νευρωνικό δίκτυο συνοδεύεται από μεγάλο αριθμό παραμέτρων όπως ο αριθμός των επιπέδων, ο αριθμός των κόμβων ανά επίπεδο, ο ρυθμός μάθησης, τα αρχικά βάρη και άλλα. Σε αντίθεση με άλλες τεχνικές μηχανικής μάθησης, εδώ ο χρήστης πρέπει να αναζητήσει τις βέλτιστες παραμέτρους, το οποίο είναι δύσκολο λόγω του υψηλού χρόνου υπολογισμού. Τέλος, η εισαγωγή νέων δεδομένων ή γνώσεων σε ένα νευρωνικό δίκτυο που έχει ήδη εκπαιδευτεί αποτελεί πρόβλημα.

Η παρούσα εργασία εστιάζει στα νευρωνικά δίκτυα συνελικτικής μάθησης (CNN), δεδομένου ότι αυτός ο τύπος δικτύων είναι πολύ αποτελεσματικός για εργασίες ανίχνευσης και ταξινόμησης εικόνων. Αυτό το είδος νευρωνικού, συναντάται σε τεχνικές Μεταφοράς Μάθησης (Transfer Learning) και συνδυάζεται με Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM).

2.5 Συνελικτικά Νευρωνικά Δίκτυα (CNN)

Τα CNN είναι ένα από τα πιο συχνά χρησιμοποιούμενα βαθιά νευρωνικά δίκτυα. Τα συνελικτικά νευρωνικά δίκτυα είναι ένας συγκεκριμένος τύπος μοντέλων, με την έννοια ότι δέχονται δισδιάστατα δεδομένα ως είσοδο. Ως αποτέλεσμα, χρησιμοποιούνται κυρίως στην άραση υπολογιστών, αλλά μπορούν επίσης να χρησιμοποιηθούν για την επεξεργασία φυσι-

κής γλώσσας. Είναι εμπνευσμένα από μία βιολογική διαδικασία, την οργάνωση του οπτικού φλοιού των ζώων. Ο οπτικός φλοιός περιέχει μεμονωμένους φλοιώδεις νευρώνες οι οποίοι ανταποκρίνονται στο φως μόνο σε μια περιορισμένη περιοχή του οπτικού πεδίου, γνωστού ως δεκτικό πεδίο. Το όνομα προέρχεται από τη μαθηματική γραμμική πράξη της συνέλιξης, η οποία είναι πάντα παρούσα σε τουλάχιστον ένα από τα στρώματα του δικτύου. Η πιο τυπική λειτουργία συνέλιξης που χρησιμοποιείται στη Βαθιά Μάθηση είναι η 2D συνέλιξη μιας δισδιάστατης εικόνας με πυρήνα (kernel) 2 διαστάσεων [23].



Σχήμα 2.5: Τυπική αρχιτεκτονική ενός συνελικτικού νευρωνικού δικτύου (CNN).
(Πηγή Σχήματος: [24])

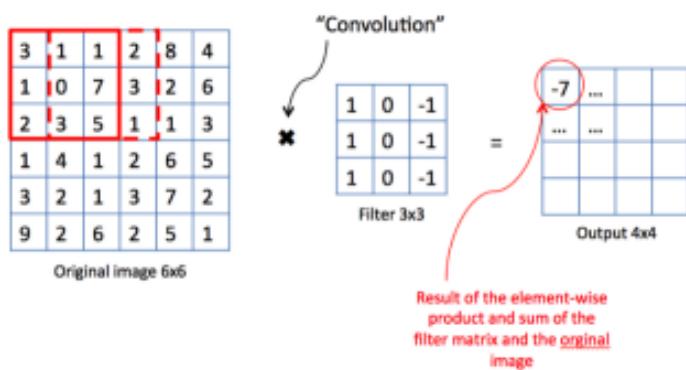
Τα 3 σημαντικά στρώματα του CNN, όπως φαίνεται στο Σχήμα 2.5 είναι το στρώμα εισόδου, τα κρυφά στρώματα και ένα στρώμα εξόδου αντίστοιχα γνωστά ως convolutional layer, pooling layer, και fully-connected layer. Το σχήμα δείχνει πως η είσοδος που δίνεται στο μοντέλο επεξεργάζεται από το CNN για να παράγει αξιολογημένες εξόδους μετά την εκπαίδευση του μοντέλου [25].

2.5.1 Convolutional Layer

Αυτά τα κρυφά στρώματα ασχολούνται με την εκμάθηση τοπικών μοτίβων σε μικρά δισδιάστατα παράθυρα. Πρόκειται για στρώματα που χρησιμεύουν στην ανίχνευση οπτικών χαρακτηριστικών στις εικόνες, όπως γραμμές, ακμές και έντονες αλλαγές χρωμάτων. Αυτά τα στρώματα μαθαίνουν μια ιδιότητα της εικόνας σε ένα συγκεκριμένο σημείο και είναι σε θέση να την αναγνωρίζουν οπουδήποτε μέσα στην εικόνα. Τα Convolutional Layers είναι σε

θέση να μαθαίνουν διαφορετικά στοιχεία, πιο σύνθετα, σύμφωνα με αυτά που μαθαίνονται στο προηγούμενο επίπεδο. Ένα παράδειγμα είναι ότι ένα στρώμα μαθαίνει διαφορετικού τύπους γραμμών που εμφανίζονται στην εικόνα και το επόμενο στρώμα είναι σε θέση να μάθει στοιχεία της εικόνας που αποτελούνται από διαφορετικές γραμμές που μαθαίνονται στο προηγούμενο στρώμα. Η ικανότητα για κάτι τέτοιο επιτρέπει αυτά τα δίκτυα να καταλαβαίνουν με αποτελεσματικό τρόπο οπτικές έννοιες που γίνονται όλο και πιο πολύπλοκες. Τα συνελικτικά νευρωνικά δίκτυα λειτουργούν σε 3 άξονες, πλάτος, ύψος και το κανάλι. Στην είσοδο, το κανάλι θα μπορούσε να είναι 1 εάν η εικόνα εισόδου είναι σε κλίμακα του γκρι ή 3 για μια έγχρωμη εικόνα RGB, ένα για κάθε χρώμα (κόκκινο, πράσινο και μπλε). Στο εσωτερικό του δικτύου, το βάθος είναι συνήθως μεγαλύτερο από 3 και ανάλογο του πλήθους των μονάδων (hidden units).

Η έξοδος είναι το αποτέλεσμα της εφαρμογής του φίλτρου (Kernel) στην αρχική εικόνα. Η μαθηματική πράξη της συνέλιξης γίνεται μεταξύ της εικόνας εισόδου και ενός φίλτρου ορισμένου μεγέθους MxM σε αυτό το στρώμα. Η συνέλιξη μεταξύ του φίλτρου και των τιμημάτων της εικόνας εισόδου πραγματοποιείται μέσω ολίσθησης του φίλτρου κατά μήκος της εικόνας εισόδου. Το μέγεθος εξόδου της εικόνας αλλάζει ανάλογα με το μέγεθος του φίλτρου. Για παράδειγμα, στην εικόνα 2.6 παρουσιάζεται ένα φίλτρο για την ανίχνευση οριζόντιων ακμών. Κάθε τιμή εξόδου είναι το αποτέλεσμα του γινομένου και του αθροίσματος των αρχικής εικόνας με το φίλτρο στην πρώτη εφαρμογή του φίλτρου.



Σχήμα 2.6: Παράδειγμα συνέλιξης.

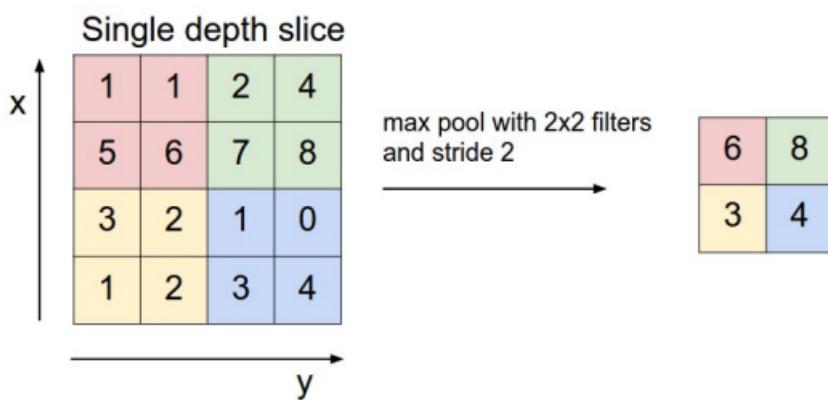
(Πηγή Σχήματος: [26])

Το αποτέλεσμα αυτής της πράξης σε όλο το μήκος της εικόνας οδηγεί στο Χάρτη Χαρακτηριστικών (Feature Map), το οποίο περιέχει πληροφορίες για την εικόνα όπως οι γωνίες και

οι ακμές της. Αυτό το Feature Map παρέχεται στη συνέχεια σε περαιτέρω επίπεδα, τα οποία μαθαίνουν μια ποικιλία διαφορετικών χαρακτηριστικών από την εικόνα εισόδου.

2.5.2 Pooling Layer

Το στρώμα συγκέντρωσης ακολουθεί ένα στρώμα συνελικτικής ανάλυσης. Ο κύριος στόχος αυτού του στρώματος είναι να μειώσει το μεγέθους του feature map με σύμπτυξη, προκειμένου να μειωθούν οι υπολογιστικές δαπάνες. Αυτό επιτυγχάνεται με τη μείωση των συνδέσεων μεταξύ των στρωμάτων και την ανεξάρτητη λειτουργία σε κάθε χάρτη χαρακτηριστικών. Υπάρχουν πολλά είδη διαδικασιών pooling, ανάλογα με τον χρησιμοποιούμενο μηχανισμό. Στο Max Pooling, το μεγαλύτερο στοιχείο λαμβάνεται από το feature map. Ο μέσος όρος των στοιχείων σε ένα τμήμα εικόνας προκαθορισμένου μεγέθους χρησιμοποιείται στο Average Pooling. Το μικρότερο στοιχείο λαμβάνεται από το feature map στο Min Pooling. Το Sum Pooling υπολογίζει το συνολικό άθροισμα των στοιχείων στο καθορισμένο τμήμα. Επιπλέον, μια συνηθισμένη χρήση του pooling είναι για τη σύνδεση του Convolutional Layer και του πλήρους συνδεδεμένου επιπέδου (Fully-Connected). Στην παρακάτω εικόνα 2.9 φαίνεται η εφαρμογή του max-pooling σε μία είσοδο.



Σχήμα 2.7: Παράδειγμα εφαρμογής pooling layer.

(Πηγή Σχηματος: [27])

2.5.3 Fully Connected Layer

Τα βάρη (weights) και τα biases, καθώς και οι νευρώνες, συνθέτουν το στρώμα FC, το οποίο χρησιμοποιείται για τη σύνδεση των νευρώνων μεταξύ δύο στρωμάτων. Τα τελευταία στρώματα μιας αρχιτεκτονικής CNN τοποθετούνται γενικά πριν από το στρώμα εξό-

δου. Επειδή υπάρχουν τόσα πολλά κρυμμένα στρώματα με ποικίλα βάρη για την έξοδο κάθε νευρώνα, είναι δύσκολο να ακολουθήσει κανείς τα δεδομένα μετά από αυτό το στάδιο. Εδώ λαμβάνει χώρα όλη η συλλογιστική και ο υπολογισμός των δεδομένων. Οι εικόνες εισόδου των προηγούμενων στρωμάτων ισοπεδώνονται (Flattened) και παρέχονται στο στρώμα FC σε αυτό το στάδιο. Στη συνέχεια, το ισοπεδωμένο διάνυσμα αποστέλλεται μέσω μερικών πρόσθετων επιπέδων FC, όπου εκτελούνται συνήθως οι μαθηματικές λειτουργικές πράξεις. Στο τέλος η softmax συνάρτηση χρησιμοποιείται στο FC για να παράγει το αποτέλεσμα το οποίο αποτελεί τις προβλεπόμενες πιθανότητες για το εκάστοτε label. Η ταξινόμηση αρχίζει σε αυτό το σημείο.

2.6 Αναδρομικό Νευρωνικό Δίκτυο

Τα αναδρομικά νευρωνικά δίκτυα σε αντίθεση με τα κλασικά νευρωνικά δίκτυα και τα πολυεπίπεδα Perceptrons είναι μοντέλα στα οποία η έξοδος είναι συνάρτηση όχι μόνο της τρέχουσας εισόδου αλλά εξαρτάται και από τις προηγούμενες εξόδους, χρησιμοποιώντας μια εσωτερική μνήμη (κρυφή κατάσταση) για την επεξεργασία της ακολουθίας των εισόδων. Αυτή η ικανότητα να θυμούνται προηγούμενες εξόδους/επεξεργασμένες πληροφορίες, επιτρέπει στα Recurrent Neural Networks να κωδικοποιούν πληροφορίες που υπάρχουν στην ίδια την ακολουθία, ενώ εξαρτώνται πάντα από όλους τους προηγούμενους υπολογισμούς. Κατά συνέπεια, τα RNN είναι πολύ χρήσιμα για την επεξεργασία διαδοχικών δεδομένων. Όσο μεγαλύτερη είναι η ακολουθία, τόσο περισσότερα επίπεδα θα έχει το μοντέλο.

Η μνήμη (κρυφή κατάσταση) ενημερώνεται από ένα μετασχηματισμό του τρέχουντος στρώματος εισόδου και του προηγούμενου κρυμμένου στρώματος, με βάση την ακόλουθη εξίσωση:

$$h_t = \sigma_h(W_h x_t + U_h h_{t-1}) \quad (2.1)$$

όπου t είναι το χρονικό βήμα, σ_h είναι η συνάρτηση ενεργοποίησης, W και U είναι οι πίνακες βαρών που καθορίζουν τη σημασία που δίνεται στην τρέχουσα είσοδο και στην προηγούμενη κατάσταση.

Η έξοδος y υπολογίζεται με βάση την ακόλουθη εξίσωση:

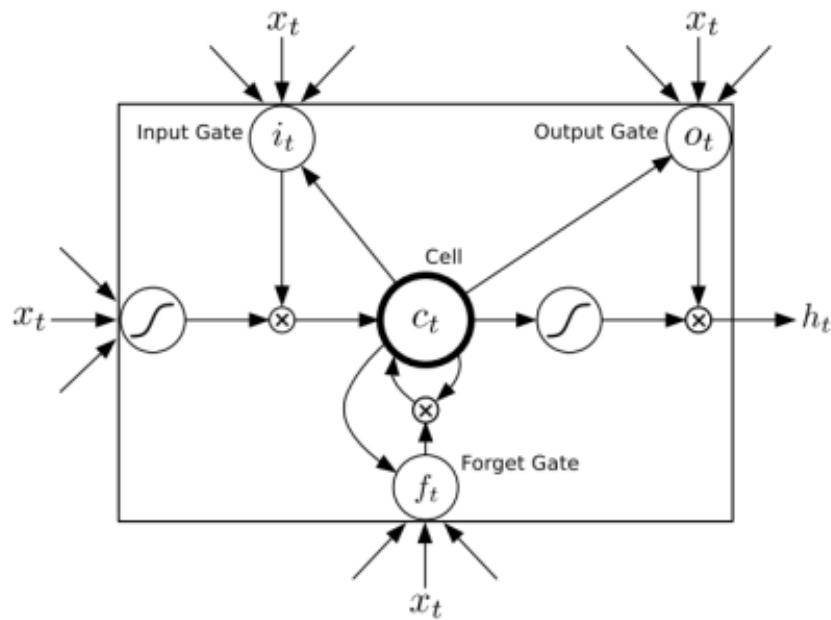
$$y_t = \sigma_y(W_y h_t) \quad (2.2)$$

όπου σ_y είναι η softmax συνάρτηση.

Σε αντίθεση με τα νευρωνικά δίκτυα τροφοδότησης, τα RNN μοιράζονται τις ίδιες παραμέτρους σε όλα τα επίπεδα και συνήθως εκπαιδεύονται με την τεχνική backpropagation. Ωστόσο, η χρήση αυτής της τεχνικής μπορεί να οδηγήσει σε προβλήματα όπου οι συσσωρευμένες κλίσεις για μεγάλες ακολουθίες γίνονται εξαιρετικά μεγάλες/μικρές [28].

2.6.1 Δίκτυα Μακράς Βραχύχρονης Μνήμης (LSTM)

Τα μοντέλα μακράς βραχυπρόθεσμης μνήμης (LSTM) βασίζονται στην αρχιτεκτονική RNN. Ξεπερνούν τις αδυναμίες του RNN όσον αφορά τις κλίσεις (gradients) και βελτιώνουν επίσης την ικανότητα μάθησης για δεδομένα ακολουθιών μεγάλου χρόνου. Οι διαφορές είναι ότι αντί να έχουν ένα μόνο επίπεδο νευρωνικού δικτύου, υπάρχουν τέσσερα στρώματα που αλληλεπιδρούν με συγκεκριμένο τρόπο. Εισάγεται επίσης ένα cell μνήμης που είναι σε θέση να διατηρήσει την κατάσταση σε μεγάλες χρονικές περιόδους. Τώρα υπάρχουν δύο καταστάσεις, μια κρυφή (hidden) και μια cell.



Σχήμα 2.8: LSTM Διάγραμμα.

(Πηγή Σχήματος: [29])

Μια κοινή μονάδα LSTM αποτελείται από ένα cell, μια πύλη εισόδου (input gate), μια πύλη εξόδου (output gate) και μια πύλη λήθης (forget gate). Το cell θυμάται τιμές για αυθαίρετα χρονικά διαστήματα και οι τρεις πύλες ρυθμίζουν τη ροή των πληροφοριών μέσα και έξω από το cell. Η πύλη εισόδου (η οποία αποτελείται από δύο μέρη) ελέγχει τη νέα πληροφορία και ποιο μέρος της θα αποθηκευτεί στην κατάσταση του κυττάρου με βάση μια συνάρτηση. Η πύλη λήθης αντιμετωπίζει το πρόβλημα της κλίσης. Μετά την προσθήκη της μεταβλητής bias, το αποτέλεσμα αποστέλλεται στη σιγμοειδή συνάρτηση για να αποφασιστεί πόση πληροφορία θα ξεχαστεί. Παρακάτω παραθέτονται οι εξισώσεις των πυλών ενός LSTM :

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2.3)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2.4)$$

$$\tilde{C}_t = \sigma(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (2.5)$$

Τέλος, το LSTM αποφασίζει για την έξοδο με τη χρήση μιας σιγμοειδούς συνάρτησης στην πύλη εξόδου. Στη συνέχεια, το LSTM πολλαπλασιάζει (γινόμενο Hadamard) την κατάσταση του cell με την πύλη sigmoid, ελέγχοντας έτσι πόση πληροφορία θα πρέπει να εξάγεται. Παρακάτω παραθέτονται οι εξισώσεις εξόδου ενός LSTM :

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (2.6)$$

$$h_t = o_t \times \tanh(C_t) \quad (2.7)$$

Συνολικά, οι πύλες ενεργούν με βάση τα σήματα που λαμβάνουν και μπλοκάρουν ή περνούν τις πληροφορίες με βάση τη δύναμη και τη σημασία τους χρησιμοποιώντας τα δικά τους σετ βαρών, τα οποία προσαρμόζονται σε όλη τη μαθησιακή διαδικασία. Μαθαίνουν πότε να επιτρέπουν/διαγράφουν δεδομένα με βάση το σφάλμα οπισθοδιάδοσης καθώς και άλλα στοιχεία [30].

2.7 Συναρτήσεις Ενεργοποίησης

Στα νευρωνικά δίκτυα, η Συνάρτηση Ενεργοποίησης ενός κόμβου καθορίζει την έξοδο του ανάλογα με την είσοδο ή το σύνολο των εισόδων. Για κάθε νευρώνα, η είσοδος υπολογίζεται με τις συναρτήσεις ενεργοποίησης και στη συνέχεια το αποτέλεσμα αποστέλλεται ως έξοδος. Ανάλογα με αυτή την έξοδο, η συνάρτηση ενεργοποίησης αποφασίζει αν οι συνδέσεις αυτού του νευρώνα ενεργοποιούνται ή όχι. Κάποιες από τις βασικές συναρτήσεις ενεργοποίησης είναι:

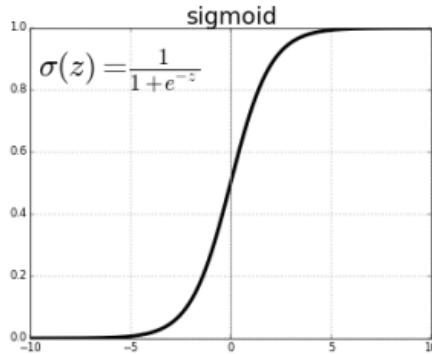
2.7.1 Sigmoid

Η μη γραμμική φύση της σιγμοειδούς συνάρτησης είναι ιδανική για νευρωνικά δίκτυα. Η σιγμοειδής συνάρτηση μετασχηματίζει τις εισαγόμενες τιμές σε μια κλίμακα (0,1), όπου οι υψηλές τιμές έχουν ασυμπτωτική πορεία προς το 1 και οι πολύ χαμηλές τιμές τείνουν ασυμπτωτικά προς το 0. Η εξίσωσή της είναι η ακόλουθη:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.8)$$

Τα χαρακτηριστικά της είναι:

- Αργή σύγκλιση
- Δεν είναι κεντραρισμένη στο 0 (προτιμάται στα νευρωνικά δίκτυα η είσοδος να είναι κεντραρισμένη στο 0 και να έχει κατά προσέγγιση ίδιο εύρος)
- Περιορίζεται μεταξύ του 0 και 1
- Καλή απόδοση στο τελευταίο στρώμα (χρησιμοποιείται στην έξοδο ενός νευρωνικού δικτύου για δυαδική ταξινόμηση)



Σχήμα 2.9: Γραφική αναπαράσταση της sigmoid συνάρτησης.

(Πηγή Σχήματος: [31])

2.7.2 ReLU (Rectified Linear Units)

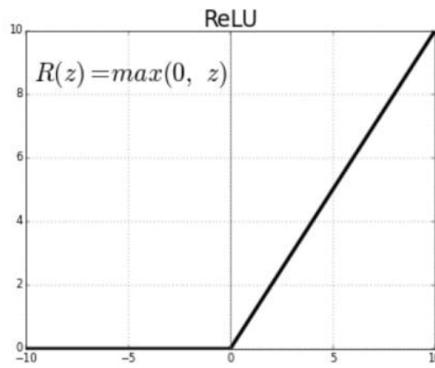
Τα τελευταία χρόνια, η ReLU είναι η πλέον χρησιμοποιούμενη συνάρτηση ενεργοποίησης λόγω της απλότητάς της. Η ReLU έχει αποδειχθεί ότι επιταχύνει την εκπαίδευση, σε σύγκριση με άλλες συναρτήσεις ενεργοποίησης. Επιστρέφει 0 εάν η είσοδος είναι μικρότερη από 0 και την ίδια τιμή της εισόδου εάν η είσοδος είναι μεγαλύτερη από 0.

Αν και μπορεί να φαίνεται από το όνομά της, αυτή η συνάρτηση ενεργοποίησης δεν είναι γραμμική, αφού όλες οι αρνητικές τιμές μετατοπίζονται στο 0. Η εξίσωση που την αντιπροσωπεύει είναι η ακόλουθη:

$$ReLU(x) = \max(0, x)$$

και η γραφική αναπαράστασή της φαίνεται στο Σχήμα 2.10.

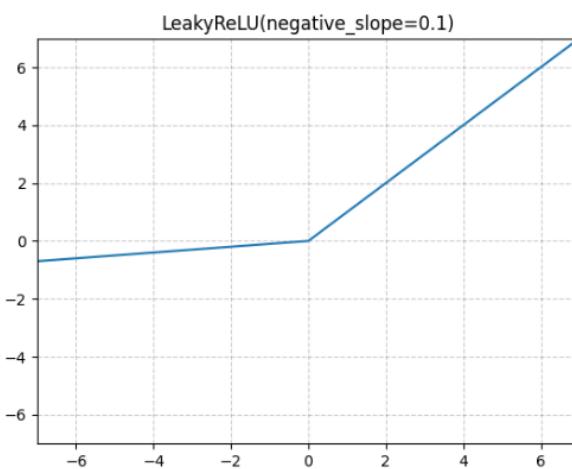
Ένας σημαντικός λόγος που η ReLU σε σχέση με την Sigmoid αποτελεί στο σύχρονο Deep Learning την πρώτη επιλογή ως συνάρτηση ενεργοποίησης είναι ότι ξεπερνάει το πρόβλημα της εξαφάνισης της κλίσης (vanishing gradient problem). Εάν η συνάρτηση ReLU χρησιμοποιείται για συνάρτηση ενεργοποίησης σε ένα νευρωνικό δίκτυο στη θέση μιας σιγμοειδούς, η τιμή της μερικής παραγώγου της απώλειας θα έχει τιμές 0 ή 1 που εμποδίζει την εξαφάνιση της κλίσης. Το πρόβλημα με τη χρήση της ReLU είναι όταν η μερική παράγωγος έχει τιμή 0. Σε τέτοιες περιπτώσεις, ο κόμβος θεωρείται ως νεκρός κόμβος αφού οι παλιές και οι νέες τιμές των βαρών παραμένουν οι ίδιες. Αυτή η κατάσταση μπορεί να αποφευχθεί με τη χρήση της Leaky ReLU που αποτρέπει την πτώση της μερική παραγώγου στη μηδενική τιμή,



Σχήμα 2.10: Γραφική αναπαράσταση της ReLU συνάρτησης.

(Πηγή Σχήματος: [31])

αφού έχει μικρή κλίση για αρνητικές τιμές αντί για επίπεδη. Παρακάτω φαίνεται η γραφική αναπαράσταση της Leaky ReLU, όπου για αρνητικές τιμές δεν έχει σταθερή παράγωγο.



Σχήμα 2.11: Γραφική αναπαράσταση της Leaky ReLU συνάρτησης.

(Πηγή Σχήματος: [32])

Συμπερασματικά, το πρόβλημα της εξαφάνισης της κλίσης προκύπτει από τη φύση της μερικής παραγώγου της συνάρτησης ενεργοποίησης που χρησιμοποιείται για τη δημιουργία του νευρωνικού δικτύου. Το πρόβλημα μπορεί να είναι χειρότερο σε βαθιά νευρωνικά δίκτυα που χρησιμοποιούν τη Sigmoid. Μπορεί να μειωθεί σημαντικά χρησιμοποιώντας συναρτήσεις ενεργοποίησης όπως η ReLU και η Leaky ReLU.

2.7.3 Softmax

Η softmax είναι μια συνάρτηση ενεργοποίησης που χρησιμοποιείται στην πολυταξική ή πολυωνυμική ταξινόμηση, η οποία κλιμακώνει τους αριθμούς (logits) σε πιθανότητες. Η έξο-

δος της Softmax είναι ένα διάνυσμα με πιθανότητες για κάθε πιθανή κλάση. Οι πιθανότητες στο διάνυσμα αθροίζονται σε ένα για όλα τα πιθανά αποτελέσματα ή κλάσεις.

Μαθηματικά, η Softmax ορίζεται ως εξής,

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (2.9)$$

με x_i να είναι ένα διάνυσμα εισόδου στην συνάρτηση Softmax, το οποίο αποτελείται από πλήθος στοιχείων ισάριθμο με το πλήθος των πιθανών κλάσεων. Ο παρανομαστής αποτελεί έναν όρο κανονικοποίησης. Οι έξοδοι της συνάρτησης παίρνουν τιμές στο διάστημα $[0, 1]$ και το άθροισμά τους ισούται με 1. Ως εκ τούτου, η Softmax μετατρέπει το διάνυσμα εισόδου σε κατανομή πιθανότητας.

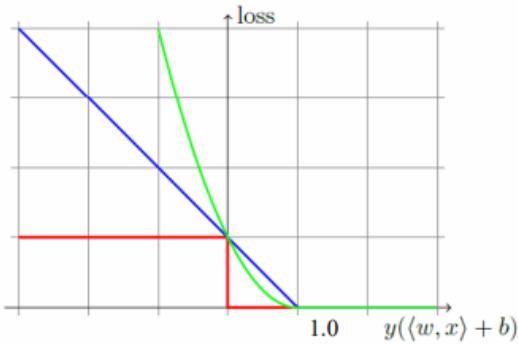
[33].

2.8 Συνάρτησεις Απώλειας

Μια συνάρτηση απώλειας είναι ένας μαθηματικός τρόπος μέτρησης του πόσο λάθος είναι οι προβλέψεις. Επιπλέον, είναι η συνάρτηση που επιδιώκεται να ελαχιστοποιηθεί καθ' όλη τη διάρκεια της διαδικασίας της εκπαίδευσης. Για παράδειγμα, η απώλεια άρθρωσης (hinge loss) είναι μια συνάρτηση απώλειας που χρησιμοποιείται συνήθως για την εκπαίδευση ταξινομητών μέγιστου περιθωρίου. Στη δυαδική ταξινόμηση, η πραγματική απώλεια που θα θέλαμε να ελαχιστοποιήσουμε είναι η λεγόμενη απώλεια μηδέν-ένα (zero-one loss):

$$l(x, y, z) = \begin{cases} 0 & \text{if } y(\langle w, x \rangle + b) \geq 1 \\ 1 & \text{otherwise} \end{cases} \quad (2.10)$$

Οστόσο, αυτή η απώλεια είναι δύσκολο να χρησιμοποιηθεί κι αυτό επειδή δεν είναι κυρτή. Η απώλεια άρθρωσης (hinge loss) είναι μια δημοφιλής επιλογή αντί της απώλειας μηδέν-ένα (zero-one loss). Η τετραγωνισμένη απώλεια άρθρωσης (squared hinge loss), σε αντίθεση με την απώλεια άρθρωσης, παραγοντοποιείται επιπλέον παντού.



Σχήμα 2.12: Το κόκκινο υποδηλώνει απώλεια μηδέν-ένα (zero-one loss), το μπλε υποδηλώνει απώλεια άρθρωσης (hinge loss), το πράσινο υποδηλώνει τετραγωνική απώλεια άρθρωσης (squared hinge loss).

(Πηγή Σχήματος: [34])

Στην ταξινόμηση πολλαπλών κατηγοριών κάθε δείγμα μπορεί να συσχετιστεί με πολλές πιθανές ετικέτες. Ο στόχος εδώ είναι ο υπολογισμός μιας βαθμολογίας συμβατότητας (compatibility score) μεταξύ ενός δείγματος και κάθε ετικέτας και να ανατεθεί η ετικέτα με την υψηλότερη βαθμολογία στο δείγμα. Ως αποτέλεσμα, πολλές διαφορετικές παραλλαγές της πολυταξικής αρθρωτής απώλειας έχουν προταθεί. Για παράδειγμα, θα μπορούσαν να “τιμωρηθούν” όλες οι ετικέτες που έχουν λανθασμένη ταξινόμηση σύμφωνα με την εξίσωση:

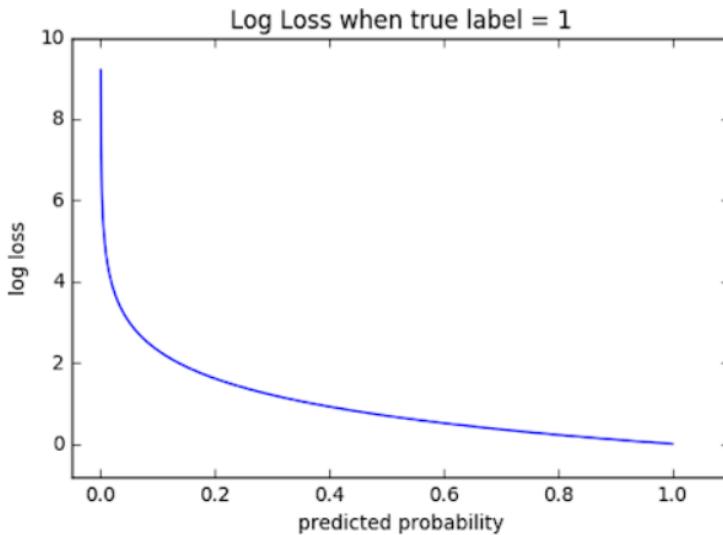
$$l(y) = \sum_{y \neq t} \max(0, 1 + W_y X - W_t X) \quad (2.11)$$

2.8.1 Cross-Entropy Loss Function

Είναι σημαντικό να αναφερθεί ότι είναι η πιο συνηθισμένη συνάρτηση απώλειας που χρησιμοποιείται σε προβλήματα ταξινόμησης και κυρίως σε προβλήματα πολυταξικής ταξινόμησης (multiclass classification). Ονομάζεται επίσης logarithmic loss, log loss ή logistic loss. Κάθε προβλεπόμενη πιθανότητα κλάσης συγκρίνεται με την πραγματική επιθυμητή έξοδο κλάσης 0 ή 1 και υπολογίζεται μια βαθμολογία/απώλεια (score/loss) που τιμωρεί την πιθανότητα με βάση το πόσο απέχει από την πραγματική αναμενόμενη τιμή. Η ποινή είναι λογαριθμικής φύσης και αποδίδει μεγάλο σκορ για μεγάλες διαφορές κοντά στο 1 και μικρό σκορ για μικρές διαφορές που τείνουν στο 0. Αυτό σημαίνει ότι η cross-entropy loss μειώνεται καθώς η προβλεπόμενη πιθανότητα συγκλίνει στην πραγματική ετικέτα.

Η απώλεια διασταυρούμενης εντροπίας (cross-entropy) χρησιμοποιείται κατά την προσαρμογή των βαρών του μοντέλου στην εκπαίδευση. Στόχος είναι η ελαχιστοποίηση της απώλειας, δηλαδή όσο μικρότερη είναι η απώλεια τόσο καλύτερο είναι το μοντέλο. Ένα τέλειο μοντέλο έχει απώλεια cross-entropy 0. Η εξίσωση απώλειας της διασταυρούμενης εντροπίας και η αναπαράσταση της έχουν ως εξής:

$$L_{CE} = - \sum_{i=1}^n t_i \log(p_i) \quad (2.12)$$



Σχήμα 2.13: Αναπαράσταση της cross-entropy loss με true label ίσο με 1.

(Πηγή Σχήματος: [35])

όπου το t_i είναι ο αριθμός των κλάσεων, το t_i είναι η πραγματική κλάση και p_i είναι η softmax πιθανότητα για την i_{th} κλάση.

2.8.2 Categorical Cross-Entropy and Sparse Categorical Cross-Entropy

Τόσο η Categorical Cross-Entropy όσο και η Sparse Categorical Cross-Entropy έχουν την ίδια συνάρτηση απώλειας όπως ορίζεται στην Εξίσωση 2.12. Η μόνη διαφορά μεταξύ των δύο είναι ο τρόπος με τον οποίο ορίζονται οι ετικέτες αληθειας.

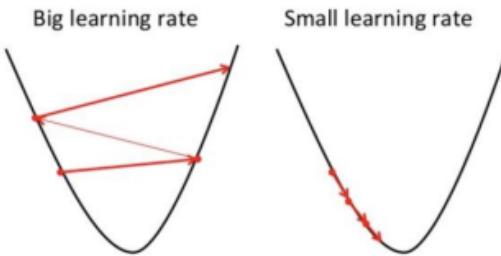
- Η Categorical Cross-Entropy χρησιμοποιείται όταν οι αληθείς ετικέτες είναι κωδικοποιημένες σε one hot, για παράδειγμα υπάρχουν οι ακόλουθες αληθείς τιμές για το πρόβλημα ταξινόμησης 3 κατηγοριών $[1,0,0]$, $[0,1,0]$ και $[0,0,1]$.

- Στην Sparse Categorical Cross-Entropy , οι ετικέτες αλήθειας είναι κωδικοποιημένες σε ακέραιο αριθμό από 0 μέχρι $n-1$ όπου n ο αριθμός των κλάσεων. Για παράδειγμα 0, 1 και 2 για πρόβλημα 3 κατηγοριών [35].

2.9 Βελτιστοποίηση αλγορίθμων βαθιάς μάθησης

Κατά την εκπαίδευση ενός Νευρωνικού Δικτύου, ένας αλγόριθμος βελτιστοποίησης χρησιμοποιείται για να παράξει μια ελαφρώς καλύτερη απόδοση. Αυτή η στρατηγική βελτιστοποίησης αποδυναμώνει και ενημερώνει τους παραμέτρους του μοντέλου, ως απόκριση στην έξοδο της συνάρτησης απώλειας. Έτσι γίνεται μία προσπάθεια να προσεγγιστεί η βέλτιστη λύση. Ειδικότερα, οι παράμετροι του μοντέλου μπορεί να είναι τα βάρη και το bias, παράμετροι δηλαδή που παίζουν πολύ σημαντικό ρόλο και μπορούν να οδηγήσουν σε εντελώς διαφορετικά αποτελέσματα. Ο ορισμός ενός αλγορίθμου βελτιστοποίησης έχει ως εξής: προσπαθεί να ελαχιστοποιήσει ή μεγιστοποιήσει - ελαχιστοποιήσει στη συγκεκριμένη περίπτωση - μιας αντικειμενικής συνάρτησης. Σε αυτό το σενάριο, η αντικειμενική συνάρτηση είναι μια συνάρτηση σφάλματος (error function) η οποία εξαρτάται από τους μαθησιακούς παραμέτρους του μοντέλου.

Η διαδικασία πίσω από τους βελτιστοποιητές θυμίζει τους αλγορίθμους “αναρρίχησης λόφου”, και με παρόμοιο τρόπο μπορεί να κολλήσει σε τοπικά ελάχιστα. Για να αποφευχθεί αυτό το φαινόμενο, προσαρμόζεται μια παράμετρος που ονομάζεται ρυθμός εκμάθησης (learning rate). Έτσι διασφαλίζεται πως τα βήματα που πραγματοποιούνται σε κάθε επανάληψη δεν είναι πολύ μεγάλα ή πολύ μικρά. Έτσι, οι κλίσεις (gradient) πολλαπλασιάζονται με τον ρυθμό εκμάθησης και κλιμακώνονται κατάλληλα. Παράλληλα με τον ρυθμό εκμάθησης, χρησιμοποιείται μια έννοια που ονομάζεται ορμή (momentum). Ο λόγος αυτής είναι η δυσκολία επίτευξης σύγκλισης. Η ορμή επιταχύνει την διαδικασία καθόδου, προσθέτοντας ένα καθορισμένο από τον χρήστη κλάσμα του προηγούμενου διανύσματος στο τρέχον. Το κλάσμα ωφελεί τις διαστάσεις των οποίων οι κλίσεις δείχνουν προς την ίδια κατεύθυνση και αποθαρρύνει τις ενημερώσεις των διαστάσεων των οποίων οι κλίσεις ποικίλουν ως προς την κατεύθυνση.



Σχήμα 2.14: Σύγκριση των ρυθμών εκμάθησης κατά την εκπαίδευση.

(Πηγή Σχήματος: [36])

2.9.1 Βελτιστοποιητής Gradient Descent

Ο πιο γνωστός βελτιστοποιητής είναι ο Gradient Descent, ένας αλγόριθμος που δεν περιορίζεται σε Νευρωνικά Δίκτυα, αλλά χρησιμοποιείται σε όλους τους τύπους προβλημάτων. Η βασική ιδέα πίσω από αυτόν είναι να υπολογίζει την επίδραση που έχουν οι αλλαγές σε κάθε μοναδικό βάρος της συνάρτησης απωλειών και να προσαρμόζει κάθε βάρος με βάση την κλίση του. Πρόκειται για μια επαναληπτική διαδικασία που βασίζεται σε μια συνάρτηση απωλειών και μπορεί να θεωρηθεί ως ένα ειδικό είδος “αλγορίθμου αναρρίχησης λόφου”. Μια κλίση (Gradient) είναι ένα διάνυσμα μιας παραγώγου που μετρά τον ρυθμό μεταβολής. Υπολογίζεται με τη βοήθεια της μερικής παραγώγου και είναι αυτό που συνδέει τη συνάρτηση απώλειας με τα βάρη. Συμπερασματικά, ο Gradient Descent βρίσκει τα ελάχιστα και ενημερώνει τις παραμέτρους του μοντέλου μέχρι να επιτευχθεί σύγκλιση. Ορισμένα από τα πλεονεκτήματά του περιλαμβάνουν την ταχύτητα και την στιβαρότητα (robustness).

Ο Stochastic Gradient Descent [37] βελτιώνει τον προηγούμενο αλγόριθμο με εκτέλεση μιας ενημέρωσης για υποσύνολο των δεδομένων εκπαίδευσης αντί μιας ενημέρωσης της κλίσης για το σύνολο των δεδομένων εκπαίδευσης, γεγονός που θα οδηγούσε σε προφανή προβλήματα. Ο στόχος είναι να μάθει μια γραμμική συνάρτηση βαθμολόγησης προκειμένου να γίνουν προβλέψεις. Ονομάζεται στοχαστικός επειδή τα δείγματα επιλέγονται τυχαία (ή ανακατεμένα) αντί για μια ενιαία ομάδα (όπως στα τυπικά Gradient Descent) ή με τη σειρά που εμφανίζονται στο σύνολο εκπαίδευσης. Ο Stochastic Gradient Descent είναι πολύ ταχύτερος και πιο αποτελεσματικός από τον Gradient Descent, ειδικά σε μεγάλα σύνολα δεδομένων. Το αποτέλεσμα της χρήσης αυτής της τεχνικής είναι η πολύ γρήγορη σύγκλιση αλλά και η

έλλειψη στιβαρότητας. Η ενημέρωση των παραμέτρων της SGD έχει τη μορφή:

$$\omega := \omega - n \nabla Q(\omega) = \omega - n \sum_{i=1}^n \nabla Q_i(\omega) / n_i \quad (2.13)$$

όπου n είναι ο ρυθμός εκμάθησης (learning rate).

$Q(\omega)$ είναι η συνάρτηση κόστους η οποία έχει τη μορφή ενός αθροίσματος:

$$Q(\omega) = \frac{1}{n} \sum_{j=1}^n Q_j(\omega) \quad (2.14)$$

με ω να είναι ο παράγοντας που ελαχιστοποιεί την $Q(\omega)$ και πρέπει να υπολογιστεί. Κάθε συνάρτηση αθροίσματος $Q_i(\omega)$ σχετίζεται με την i -οστή παρατήρηση στο σύνολο δεδομένων (χρησιμοποιείται για εκπαίδευση).

Ένας απλός, αλλά αποτελεσματικός κανόνας για την επιλογή της παραμέτρου συντονισμού του SGD είναι η επιλογή ενός μικρού υποσυνόλου δεδομένων, η δοκιμή διάφορων τιμών σε αυτό το υποσύνολο και η επιλογή αυτού που μειώνει περισσότερο την αντικειμενική συνάρτηση. Έχει αποδειχθεί ότι η SGD ασυμπτωτικά συγκλίνει στον πραγματικό ελαχιστοποιητή του εκτιμητή.

```

1: Input: Maximum iterations  $T$ , batch size  $k$ , and  $\tau$ 
2: Set  $t = 0$  and  $w_0 = 0$ 
3: while  $t < T$  do
4:   Choose a subset of  $k$  data points  $(x_i^t, y_i^t)$  and compute  $\nabla J_t(w_t)$ 
5:   Compute stepsize  $\eta_t = \sqrt{\frac{\tau}{\tau+t}}$ 
6:    $w_{t+1} = w_t - \eta_t \nabla J_t(w_t)$ 
7:    $t = t + 1$ 
8: end while
9: Return:  $w_T$ 
```

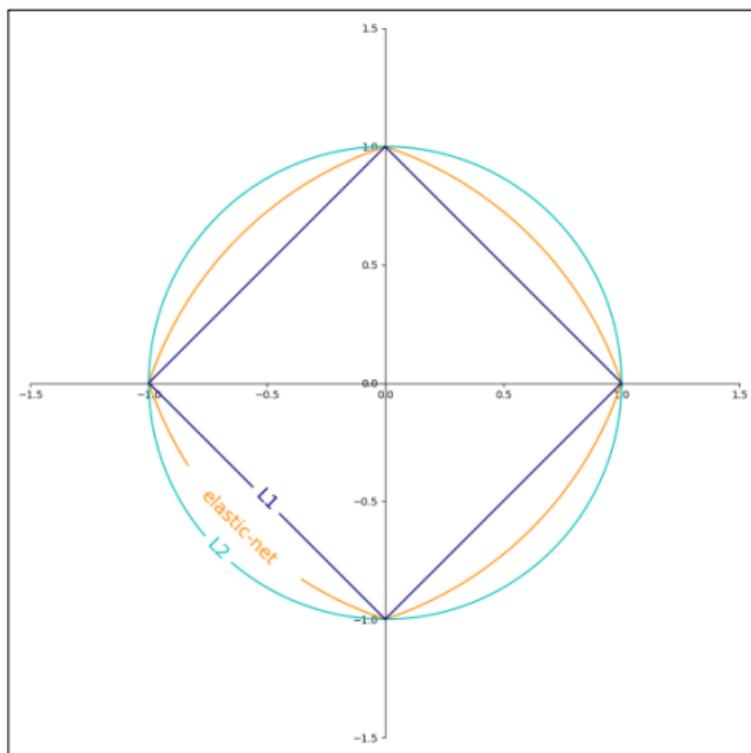
Σχήμα 2.15: Γενικός αλγόριθμος του Stochastic Gradient Descent.

(Πηγή Σχήματος: [38])

Οπως αναφέρθηκε, ο SGD έχει το πλεονέκτημα της ταχύτητας σύγκλισης και της αποτελεσματικότητας αλλά έχει επίσης τρία μειονεκτήματα: (1) απαιτεί αρκετές υπερπαραμέτρους, όπως η παράμετρος κανονικοποίησης και ο αριθμός των επαναλήψεων, (2) ίδιο ρυθμό εκμάθησης που εφαρμόζεται σε όλες τις ενημερώσεις των παραμέτρων, (3) είναι ευαίσθητος στην κλιμάκωση των χαρακτηριστικών. Επιπλέον, ο SGD κανονικοποιεί. Η διαδικασία της κανονικοποίησης (regularization) χρησιμοποιείται για να αποφευχθεί η υπερπροσαρμογή - υπερβολική εξειδίκευση στα δεδομένα εκπαίδευσης - με την “τιμωρία” των μεγάλων τιμών βαρών και, ως εκ τούτου, “τιμωρώντας” την πολυπλοκότητα του μοντέλου. Δημοφιλείς επιλογές για

την κανονικοποίηση είναι οι συναρτήσεις απώλειας L1- και L2-norm.

Η παρακάτω εικόνα παρουσιάζει τα περιγράμματα των διαφορετικών όρων κανονικοποίησης στο παραμετρικό χώρο για ποινή ίση με 1. Το ελαστικό δίκτυο είναι ένας κυρτός συνδυασμός των L2 και L1.



Σχήμα 2.16: Περιγράμματα διαφορετικής κανονικοποίησης απωλειών για ποινή ίση με 1.
(Πηγή Σχήματος: [39])

Μια καλή προσέγγιση για την εξισορρόπηση των αδυναμιών τόσο του Gradient Descent όσο και του SGD είναι η χρησιμοποίηση του Mini Batch Gradient Descent. Οι ενημερώσεις στον Mini Batch Gradient Descent πραγματοποιούνται για κάθε παρτίδα του αρχικού συνόλου δεδομένων που περιλαμβάνει σταθερό αριθμό δειγμάτων. Συνήθως πρόκειται για δύναμη του δύο. Ο SGD τείνει να χρησιμοποιείται στα νευρωνικά δίκτυα όταν το μέγεθος των δεδομένων εκπαίδευσης παρατηρείται να είναι μεγάλο.

2.9.2 Βελτιστοποιητής Adam

Ο αλγόριθμος βελτιστοποίησης Adam, συντομία της Προσαρμοστικής Εκτίμησης Βαρών (Adaptive Moment Estimation), αποτελεί επέκταση του SGD και έχει σχεδιαστεί για

την ενημέρωση των βαρών ενός νευρωνικού δικτύου κατά τη διάρκεια της εκπαίδευσης. Σε αντίθεση με τον SGD, ο οποίος διατηρεί έναν ενιαίο ρυθμό εκμάθησης καθ' όλη τη διάρκεια της εκπαίδευσης, ο βελτιστοποιητής Adam υπολογίζει τις τρέχουσες κλίσεις αποθηκεύοντας έναν φθίνοντα μέσο όρο των προηγούμενων κλίσεων και παράλληλα χρησιμοποιεί το momentum. Με την ενσωμάτωση τόσο της πρώτης στιγμής (μέσος όρος) όσο και της δεύτερης στιγμής (μη συγκεντρωμένη διακύμανση) των κλίσεων, ο βελτιστοποιητής Adam επιτυγχάνει έναν προσαρμοστικό ρυθμό μάθησης που μπορεί να πλοηγηθεί αποτελεσματικά στο τοπίο της βελτιστοποίησης κατά τη διάρκεια της εκπαίδευσης. Αυτή η προσαρμοστικότητα συμβάλλει στην ταχύτερη σύγκλιση και στη βελτίωση της απόδοσης του νευρωνικού δικτύου.

Ο βελτιστοποιητής Adam έχει πολλά πλεονεκτήματα, εξαιτίας των οποίων χρησιμοποιείται ευρέως. Προσαρμόζεται ως σημείο αναφοράς για εργασίες Βαθιάς Μάθησης και συνιστάται ως προεπιλεγμένος αλγόριθμος βελτιστοποίησης. Επιπλέον, ο αλγόριθμος είναι απλός στην υλοποίηση, έχει ταχύτερο χρόνο εκτέλεσης, χαμηλές απαιτήσεις μνήμης και απαιτεί λιγότερη ρύθμιση από οποιονδήποτε άλλο αλγόριθμο βελτιστοποίησης [40]. Για να υπολογιστούν οι στιγμές, ο Adam χρησιμοποιεί εκθετικά κινούμενους μέσους όρους.

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \left[\frac{\partial C}{\partial \omega_t} \right] u_t \quad (2.15)$$

$$u_t = \beta_2 u_{t-1} + (1 - \beta_2) \left[\frac{\partial C}{\partial \omega_t} \right]^2 \quad (2.16)$$

Οι παραπάνω τύποι αντιπροσωπεύουν τη λειτουργία του βελτιστοποιητή Adam. Τα m_t και u_t είναι οι κινούμενοι μέσοι όροι, $\frac{\partial C}{\partial \omega_t}$ είναι η κλίση και τα β_1 και β_2 αντιπροσωπεύουν το ρυθμό αποσύνθεσης του μέσου όρου των κλίσεων. Ο ψευδοκώδικας αυτής της μεθόδου απεικονίζεται ως:

```

while  $w$  not converged do
     $i \leftarrow i + 1$ 
     $m_i \leftarrow \beta_1 \cdot m_{i-1} + (1 - \beta_1) \cdot \frac{\partial C}{\partial w}(w_i)$ 
     $v_i \leftarrow \beta_2 \cdot v_{i-1} + (1 - \beta_2) \cdot \frac{\partial C}{\partial w}(w_i)^2$ 
     $\hat{m}_i \leftarrow m_i / (1 - \beta_1^i)$ 
     $\hat{v}_i \leftarrow v_i / (1 - \beta_2^i)$ 
     $w_{i+1} \leftarrow w_i - \eta \cdot \hat{m}_i / \left( \sqrt{\hat{v}_i} + \epsilon \right)$ 
end while
return  $w_i$  (Resulting parameters)

```

Σχήμα 2.17: Γενικός αλγόριθμος του Adam.

(Πηγή Σχήματος: [41])

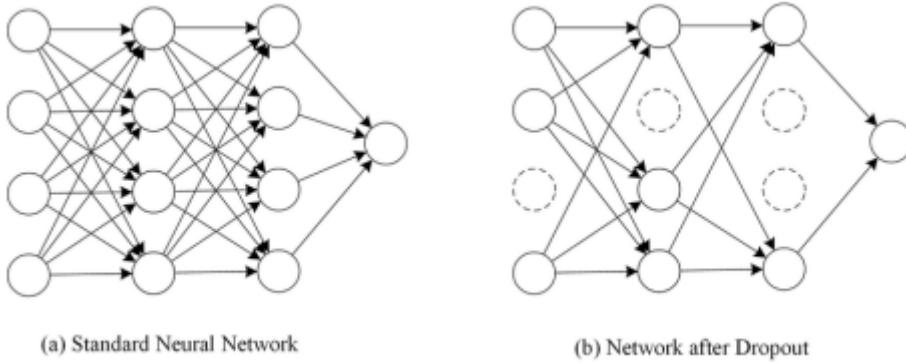
όπου n είναι ο ρυθμός εκμάθησης (learning rate), β_1 και β_2 ανήκουν στο διάστημα $[0, 1]$ και είναι τα ποσοστά εκθετικής αποσύνθεσης για τις εκτιμήσεις της στιγμής, $C(\omega)$ η συνάρτηση κόστους με w παραμέτρους και w_0 το αρχικό διάνυσμα παραμέτρων.

2.10 Μείωση υπερπροσαρμογής (Dropout)

Οι αρχιτεκτονικές βαθιάς μάθησης θα μπορούσαν να έχουν εκατομμύρια παραμέτρους, αφού κάθε στρώμα είναι συνδεδεμένο με συνδέσεις βάρους. Αυτά τα μοντέλα με πολλές παραμέτρους μπορούν εύκολα να προσαρμοστούν στα δεδομένα εκπαίδευσης, δηλαδή να αντιμετωπίσουν το πρόβλημα του overfitting. Η προσαρμογή συμβαίνει όταν ένα μοντέλο εξειδικεύεται στα δεδομένα εκπαίδευσης και τα διαμορφώνει πάρα πολύ καλά. Ως αποτέλεσμα, η απόδοση στα νέα δεδομένα επηρεάζεται αρνητικά, οδηγώντας σε πολύ υψηλή ακρίβεια στο σύνολο εκπαίδευσης και χαμηλή στο σύνολο δοκιμής.

Για να αποφευχθεί αυτό το πρόβλημα, υπάρχουν ορισμένες μέθοδοι που βοηθούν στη μείωσή του. Ένα από αυτά είναι το Dropout που αποτελεί μια δημοφιλής τεχνική κανονικοποίησης που χρησιμοποιείται στα Νευρωνικά Δίκτυα [42]. Ο όρος αναφέρεται στην εγκατάλειψη μονάδων (units) σε ένα νευρωνικό όπως φαίνεται στο σχήμα 2.18. Η εγκατάλειψη μονάδων είναι η αφαίρεση νευρώνων από το δίκτυο, συμπεριλαμβανομένων των συνδέσεων των βαρών τους. Το Dropout χρησιμοποιεί μια σταθερή πιθανότητα διατήρησης και η επιλογή των μονάδων που θα απορριφθούν είναι τυχαία. Τα νευρωνικά δίκτυα που χρησιμοποιούν

Dropout μπορούν να εκπαιδευτούν χρησιμοποιώντας μια μέθοδο προσαρμοστικής μάθησης. Η διαφορά έγκειται στο γεγονός πως για κάθε μίνι-πακέτο εκπαίδευσης, η εμπρόσθια και η οπισθοδιάδοση (forward and backpropagation) γίνονται σε ένα αραιωμένο δίκτυο που πρέπει να εγκαταλείψει μονάδες.



Σχήμα 2.18: Ένα νευρωνικό δίκτυο πριν και μετά την εφαρμογή Dropout.

(Πηγή Σχήματος: [43])

Συνεπώς, το Dropout δημιουργεί υποδίκτυα (παιδικά μοντέλα), με στόχο την ελαχιστοποίηση των αναμενόμενων απωλειών με την εγκατάλειψη μονάδων. Μειώνει την υπερπροσαρμογή σε ένα νευρωνικό δίκτυο αποτρέποντας πολύπλοκες προσαρμογές κατά την εκπαίδευση και αποκλείει σπάνιες εξαρτήσεις από το μοντέλο [44].

2.11 Υπερπαράμετροι

Οι υπερπαράμετροι στη βαθιά μάθηση είναι τιμές που χρησιμοποιούνται για την επιρροή της διαδικασίας μάθησης. Υπάρχουν δύο τύποι υπερπαραμέτρων. Ο πρώτος τύπος είναι οι υπερπαράμετροι του μοντέλου οι οποίοι δεν μπορούν να συναχθούν κατά την προσαρμογή της μηχανής στο σύνολο εκπαίδευσης, καθώς αναφέρονται στην εργασία επιλογής μοντέλου. Η τοπολογία και το μέγεθος ενός νευρωνικού δικτύου είναι παραδείγματα των υπερπαραμέτρων του μοντέλου. Ο δεύτερος τύπος είναι οι υπερπαράμετροι αλγορίθμου οι οποίοι δεν επηρεάζουν την απόδοση του μοντέλου, αλλά επηρεάζουν την ταχύτητα και την ποιότητα της διαδικασία μάθησης. Ο ρυθμός μάθησης, το batch size και το min-batch size είναι παραδείγματα υπερπαραμέτρων του αλγορίθμου. Ένα ολόκληρο δείγμα δεδομένων αναφέρεται ως batch size, ενώ ένα μικρότερο σύνολο δειγμάτων αναφέρεται min-batch size [45].

Διαφορετικοί υπερπαράμετροι μπορούν να χρησιμοποιηθούν για τη βελτίωση της απόδοσης του μοντέλου για διαφορετικούς τύπους συνόλων δεδομένων και μοντέλων. Γενικά, οι υπερπαράμετροι μπορούν να χωριστούν σε δύο ομάδες:

2.11.1 Υπερπαράμετροι Βελτιστοποίησης

Αυτές οι υπερπαράμετροι χρησιμοποιούνται για τη βελτιστοποίηση του μοντέλου, όπως υποδηλώνει και το όνομα τους [46].

1. Ρυθμός εκμάθησης (Learning Rate): Αυτή η υπερπαράμετρος καθορίζει πόσα πρόσφατα συλλεχθέντα δεδομένα θα παρακάμψουν τα προηγούμενα προσβάσιμα δεδομένα. Εάν η τιμή αυτής της υπερπαραμέτρου είναι μεγάλη, το μοντέλο δεν θα βελτιστοποιηθεί με επιτυχία, αφού υπάρχει πιθανότητα να παρακάμψει τα ελάχιστα (minimum). Από την άλλη πλευρά, εάν ο ρυθμός εκμάθησης είναι πολύ χαμηλός η σύγκλιση θα πάρει πολύ χρόνο.
2. Batch size: Το σύνολο εκπαίδευσης διαχωρίζεται σε πολλαπλά batches για να επιταχυνθεί η διαδικασία μάθησης. Εάν το batch size είναι μεγάλο, ο χρόνος εκμάθησης θα είναι μεγαλύτερος και θα απαιτηθεί περισσότερη μνήμη για την εκτέλεση του πολλαπλασιασμού του πίνακα. Αντίθετα, θα υπάρχει περισσότερος θόρυβος στον υπολογισμό των σφάλματος εάν το batch size είναι μικρότερο.
3. Number of Epochs: Στη Βαθιά Μάθηση, ένα epoch αντιπροσωπεύει έναν ολόκληρο κύκλο δεδομένων που πρέπει να εκπαιδευτεί. Στην επαναληπτική διαδικασία μάθησης, τα epochs είναι εξαιρετικά σημαντικά. Ο αριθμός των epochs μπορεί να αυξηθεί εφόσον το σφάλμα επικύρωσης (validation error) μειώνεται. Εάν το σφάλμα επικύρωσης (validation error) δεν βελτιωθεί μετά από έναν ορισμένο αριθμό epochs, είναι μια ένδειξη ότι πρέπει να μειωθεί. Μερικές φορές αναφέρεται ως “πρώιμη διακοπή (early stopping)”.

2.11.2 Συγκεκριμένοι υπερπαράμετροι μοντέλου

Η δομή του μοντέλου περιλαμβάνει επίσης διάφορες υπερπαραμέτρους [46].

1. Αριθμός των hidden units: Στα μοντέλα DL, αυτή η υπερπαράμετρος χρησιμοποιείται για να καθορίσει την ικανότητα μάθησης του μοντέλου. Πρέπει να οριστούν πολλά

hidden units για περίπλοκες συναρτήσεις, αλλά χρειάζεται προσοχή να μην υπερπροσαρμοστεί το μοντέλο.

2. Αριθμός των στρωμάτων (Layers): Στα μοντέλα CNN, η απόδοση του μοντέλου βελτιώνεται καθώς ο αριθμός των στρωμάτων (layers) αυξάνεται. Όσο πιο πολλά στρώματα, τόσο πιο βαθύ είναι το μοντέλο.

2.12 Μετρικές αξιολόγησης για πρόβλημα ταξινόμησης

Οι μετρικές αξιολόγησης είναι ποσοτικά μέτρα που χρησιμοποιούνται για την αξιολόγηση της απόδοσης και της αποτελεσματικότητας ενός στατιστικού μοντέλου ή ενός μοντέλου βαθιάς μάθησης. Αυτές οι μετρικές παρέχουν πληροφορίες σχετικά με την απόδοση του μοντέλου και βοηθούν στη σύγκριση διαφορετικών αλγορίθμων [47]. Μετρικές αξιολόγησης που χρησιμοποιούνται στην παρούσα εργασία είναι:

Πίνακας Σύγχυσης (Confusion Matrix)

Ο πίνακας σύγχυσης, επίσης γνωστός ως πίνακας σφαλμάτων, είναι μια δομή πίνακα που επιτρέπει την απόδειξη της απόδοσης ενός αλγορίθμου. Κάθε γραμμή του πίνακα περιέχει πραγματικές εμφανίσεις κλάσεων, ενώ κάθε στήλη αντιπροσωπεύει προβλεπόμενες περιπτώσεις κλάσεων, ή αντίστροφα. Χρησιμοποιείται για τον υπολογισμό των accuracy, precision, recall and F1 scores. Οι όροι που χρησιμοποιούνται στον πίνακα σύγχυσης έχουν ως εξής:

1. True Positive (TP): Είναι η περίπτωση κατά την οποία το πραγματικό και το προβλεπόμενο αποτέλεσμα τάξης ενός σημείου δεδομένων είναι και στα δύο 1.
2. True Negative (TN): Είναι η περίπτωση κατά την οποία το πραγματικό και το προβλεπόμενο αποτέλεσμα τάξης ενός σημείου δεδομένων είναι και στα δύο 0.
3. False Positive (FP): Είναι η περίπτωση κατά την οποία η πραγματική κλάση ενός σημείου δεδομένων είναι 0 και η προβλεπόμενη κλάση είναι 1.
4. False Negative (FN): Είναι η περίπτωση κατά την οποία η πραγματική κλάση ενός σημείου δεδομένων είναι 1 και η προβλεπόμενη κλάση είναι 0.

Accuracy: Ο λόγος των πραγματικών προβλεπόμενων τιμών προς τις συνολικές προβλεπόμενες τιμές είναι γνωστός ως ακρίβεια. Όσο υψηλότερη είναι η ακρίβεια, τόσο καλύτερη

είναι η απόδοση του μοντέλου.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.17)$$

Precision: Το precision χρησιμοποιείται για τη μέτρηση των μοτίβων που προβλέπονται σωστά από το σύνολο των προβλεπόμενων μοτίβων σε μια θετική κλάση. Το υψηλό precision υποδεικνύει πως το μοντέλο ορθά προβλέπει την επιθυμητή κλάση τις περισσότερες φορές.

$$Precision = \frac{TP}{TP + FP} \quad (2.18)$$

Recall: Το recall χρησιμοποιείται για τη μέτρηση των σωστά προβλεπόμενων μοτίβων από το σύνολο των πραγματικών μοτίβων σε μια θετική κλάση.

$$Recall = \frac{TP}{TP + FN} \quad (2.19)$$

F1-score: Το F1-score είναι ο αρμονικός μέσος όρος των precision και recall.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2.20)$$

Loss: Η απώλεια που προέκυψε κατά τη διαδικασία εκπαίδευσης, η οποία ονομάζεται επίσης απώλεια εκπαίδευσης. Είναι ο συνολικός αριθμός των σφαλμάτων που προέκυψαν κατά την εκπαίδευση μοντέλου.

Test-loss: Υποδεικνύει το σφάλμα στο σύνολο δεδομένων δοκιμής. Είναι ο συνολικός αριθμός σφαλμάτων που προέκυψαν για το σύνολο δοκιμών μετά από εκπαίδευση του μοντέλου με το σύνολο εκπαίδευσης.

Κεφάλαιο 3

Σχετικές Εργασίες

Η απόδοση των μοντέλων αναγνώρισης προτύπων έχει βελτιωθεί σημαντικά με την συμβολή της βαθιάς μάθησης. Πρόσφατα, “αναζωογονήθηκαν” διάφορες νέες αρχιτεκτονικές νευρωνικών δικτύων, όπως τα autoencoder networks, CNN, LSTM. Τα μοντέλα αυτά έχουν χρησιμοποιηθεί με διάφορους τρόπους για πολυτροπικές εργασίες αναγνώρισης. Με βάση τις αναπαραστάσεις των χαρακτηριστικών, τα συστήματα αναγνώρισης συναισθημάτων που χρησιμοποιούν μόνο οπτική είσοδο, δηλαδή καρέ βίντεο, μπορούν να χωριστούν σε στατικές και δυναμικές μεθόδους. Στις στατικές μεθόδους, τα χαρακτηριστικά κωδικοποιούνται με χωρικές πληροφορίες από μεμονωμένα καρέ χωρίς να λαμβάνεται υπόψη η χρονική έκταση, ενώ οι δυναμικές μέθοδοι λαμβάνουν υπόψη τη χρονική σχέση μεταξύ συνεχών καρέ στην ακολουθία εισόδου.

Έχουν προταθεί υπερσύγχρονα σχέδια βαθιών νευρωνικών δικτύων (π. χ. VGG [48], ResNet [49]) για εξαγωγή χαρακτηριστικών σε στατικές προσεγγίσεις, ενώ η ταξινόμηση σε κατηγορίες συναισθημάτων πραγματοποιείται με τη χρήση ταξινομητή Support Vector Machine (SVM). Κατά την τελευταία δεκαετία, η αναγνώριση συναισθημάτων από ήχο ήταν ένα πολυσυζητημένο ερευνητικό θέμα. Μία από τις πιο επιτυχημένες προσεγγίσεις για την εξαγωγή χαρακτηριστικών ήχου και την ταξινόμηση ομιλίας ήταν η εργαλειοθήκη openSMILE, η οποία χρησιμοποιείται ευρέως στην αυτόματη αναγνώριση συναισθημάτων από ήχο. Ομοίως, οι φασματογραφικές αναπαραστάσεις της συναισθηματικής ομιλίας ήταν επίσης επιτυχείς στην αυτόματη αναγνώριση συναισθημάτων ομιλίας. Δεδομένης της επιτυχίας τους σε πολλές εργασίες αναγνώρισης από εικόνες, τα συνελικτικά νευρωνικά δίκτυα είναι ικανά να συλλάβουν αναπαραστάσεις υψηλού επιπέδου στο χωρικό πεδίο, ώστε να παρέχουν

λύσεις για πολυάριθμες εργασίες που σχετίζονται με τις προκλήσεις της επεξεργασίας ήχου, όπως η αρχιτεκτονική ResNet που χρησιμοποιείται σε ένα μοντέλο για την αναγνώριση συναισθημάτων και ομιλητών. Τα CNN έχουν χρησιμοποιηθεί για την εξαγωγή σημαντικών χαρακτηριστικών από φασματογραφήματα στην αναγνώριση συναισθημάτων με βάση την ομιλία ή παράλληλα με μια Αμφίδρομη Μονάδα LSTM Βασισμένη στη Προσοχή (Attention-Based bidirectional LSTM module).

Αν και πολλοί αλγόριθμοι αναγνώρισης συναισθημάτων εστιάζουν στη χρήση μόνο ηχητικών δεδομένων για την αναγνώριση συναισθημάτων, η αποτελεσματικότητά τους είναι περιορισμένη λόγω του μικρού όγκου δεδομένων στον τομέα του ήχου. Ένας τρόπος για να βελτιωθεί η απόδοση αυτών των συστημάτων είναι να μεταφερθεί η γνώση από τα καρέ βίντεο στον ετερογενή τομέα του ήχου με ετικέτες. Υπό αυτή την έννοια, η μέθοδος που παρουσιάζεται στο [50] χρησιμεύει ως τεχνική επαύξησης δεδομένων που χρησιμοποιεί ένα μεγάλο σύνολο οπτικών δεδομένων με ετικέτες για να αυξήσει τον όγκο των δεδομένων αναγνώρισης συναισθημάτων με βάση τον ήχο. Με τον ίδιο τρόπο, οι εκφράσεις του προσώπου από βίντεο μπορούν να χρησιμοποιηθούν για να ενισχύσουν την επίγνωση και την παρακολούθηση της πρόβλεψης των συναισθημάτων σε δεδομένα ήχου, οδηγώντας σε μια μεταφορά γνώσης μεταξύ των ηχητικών και των μορφών του προσώπου στο συναισθηματικό πλαίσιο [51].

Ο συνδυασμός πληροφοριών και από τις δύο μορφές, ήχο και βίντεο, οδηγεί σε αυξημένη απόδοση αναγνώρισης συναισθημάτων. Τα πολυτροπικά συστήματα απαιτούν συχνά μηχανισμούς συγχώνευσης που συνδυάζουν αποτελεσματικά τα χαρακτηριστικά που εξάγονται από διαφορετικές λειτουργίες, προκειμένου να παράγουν μια συνολική απόφαση. Αυτές οι στρατηγικές συγχώνευσης για το συνδυασμό ηχητικών και οπτικών τρόπων δίνουν έμφαση στα πιο σημαντικά πλαίσια που αποκαλύπτουν το συναίσθημα του υποκειμένου. Για παράδειγμα, οι Lim κ.α. [52], αφού μετασχημάτισαν τα δεδομένα χρησιμοποιώντας μετασχηματισμό Fourier μικρού χρόνου, χρησιμοποίησαν CNN για να εξάγουν χαρακτηριστικά υψηλού επιπέδου. Για την αποτύπωση της χρονικής δομής χρησιμοποιήθηκαν LSTM. Σε μια παρόμοια εργασία, οι Trigeorgis κ.α. [53] πρότειναν ένα ολοκληρωμένο μοντέλο που χρησιμοποιεί ένα CNN για την εξαγωγή χαρακτηριστικών από το ακατέργαστο σήμα και στη συνέχεια μια σειρά από LSTM για την καταγραφή των πληροφοριών πλαισίου στα δεδομένα. Άλλες εργασίες προσπαθούν να επιλύσουν το έργο της αναγνώρισης συναισθημάτων

χρησιμοποιώντας πληροφορίες για το πρόσωπο με DNN. Για παράδειγμα, οι Ebrahimi κ.α. [54] συνδύασαν CNN και RNN για την αναγνώριση κατηγορικών συναισθημάτων σε βίντεο. Ένα CNN εκπαιδεύτηκε αρχικά για να ταξινομήσει στατικές εικόνες που περιέχουν συναισθήματα. Στη συνέχεια, τα χαρακτηριστικά που εξήγησαν από το CNN χρησιμοποιήθηκαν για την εκπαίδευση ενός RNN για την παραγωγή ενός συναισθήματος για ολόκληρο το βίντεο.

Ο συνδυασμός ακουστικών και οπτικών μέσων έχει μεγάλη επιτυχία για την αναγνώριση συναισθημάτων. Οι Kim κ.α. [55] πρότειναν τέσσερις διαφορετικές αρχιτεκτονικές DNN, με μία από αυτές να είναι μια βασική DNN 2 επιπέδων και οι υπόλοιπες να αποτελούν παραλλαγές της. Η βασική αρχιτεκτονική μαθαίνει πρώτα τα χαρακτηριστικά του ήχου και του βίντεο ξεχωριστά και, στη συνέχεια συνδυάζει αυτά τα χαρακτηριστικά από τις δύο λειτουργίες και τα χρησιμοποιεί για την εκμάθηση του δεύτερου επιπέδου. Τα χαρακτηριστικά αξιολογήθηκαν με τη χρήση SVM. Σε μια άλλη μελέτη, οι Kahou κ.α. [56] πρότειναν να συνδυάσουν DNN που σχετίζονται με συγκεκριμένες λειτουργίες για την αναγνώριση κατηγορικών συναισθημάτων σε βίντεο. Χρησιμοποιήθηκε ένα CNN για την ανάλυση των καρέ βίντεο, ένα DNN για την καταγραφή των πληροφοριών ήχου και ένας autoencoder για τη μοντελοποίηση των ανθρώπινων ενεργειών που απεικονίζονται σε ολόκληρη τη σκηνή. Στο τέλος, ένα δίκτυο CNN εφαρμόστηκε για την εξαγωγή χαρακτηριστικών από το στόμα του ανθρώπου. Για την εξαγωγή της τελικής πρόβλεψης χρησιμοποίησαν το μέσο όρο των προβλέψεων.

Οι Zhang κ.α. [57] χρησιμοποίησαν ένα πολυτροπικό CNN για την ταξινόμηση συναισθημάτων με ηχητικές και οπτικές λειτουργίες. Το μοντέλο εκπαιδεύεται σε δύο φάσεις. Στην πρώτη φάση, τα δύο CNN προ-εκπαιδεύτηκαν σε μεγάλα σύνολα δεδομένων εικόνων και ρυθμίστηκαν ώστε να εκτελέσουν αναγνώριση συναισθημάτων. Το ηχητικό CNN λάμβανε ως είσοδο το τμήμα του ηχητικού σήματος και το CNN για το βίντεο λάμβανε το πρόσωπο. Στη δεύτερη φάση, εκπαιδεύτηκε ένα DNN που αποτελούταν από έναν αριθμό πλήρως συνδεδεμένων στρωμάτων. Η συνένωση των χαρακτηριστικών που εξήγησαν από τα δύο CNN ήταν η είσοδος. Σε μια άλλη μελέτη, οι Han κ.α. [58] προτείνουν ένα πλαίσιο μοντελοποίησης ισχύος, το οποίο μπορεί να εφαρμοστεί ως στρατηγική συγχώνευσης σε επίπεδο χαρακτηριστικών και σε επίπεδο απόφασης και αποτελείται από δύο μοντέλα παλινδρόμησης. Οι προβλέψεις του πρώτου μοντέλου συνενώθηκαν με το αρχικό διάνυσμα χαρακτηριστικών

και τροφοδοτήθηκαν στο δεύτερο μοντέλο παλινδρόμησης για την τελική πρόβλεψη.

Συνοψίζοντας, η βαθιά μάθηση έχει βελτιώσει σημαντικά την απόδοση των μοντέλων αναγνώρισης συναισθημάτων. Ο συνδυασμός πληροφοριών από ήχο και εικόνες, μαζί με στρατηγικές συγχώνευσης, έχει αποδειχθεί ο πλεον αποτελεσματικός στην απόδοση των συστημάτων αναγνώρισης συναισθημάτων. Η χρήση διαφόρων αρχιτεκτονικών νευρωνικών δικτύων και τεχνικών μεταφοράς γνώσης έχει συμβάλει στην πρόοδο του τομέα και μέχρι σήμερα συνεχίζει να εντυπωσιάζει με νέες επιτυχίες.

Κεφάλαιο 4

Δεδομένα και Εργαλεία ανάπτυξης

4.1 Σύνολο δεδομένων διπλωματικής

Το σύνολο δεδομένων που χρησιμοποιείται για την ανάπτυξη αυτής της εργασίας είναι η βάση δεδομένων Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Πρόκειται για ένα δωρεάν σώμα αναφοράς που προορίζεται στην επιστημονική κοινότητα για την αναγνώριση συναισθημάτων. Ενώ υπάρχουν πολλά σύνολα δεδομένων συναισθημάτων που χρησιμοποιούνται σήμερα, η επιλογή του συνόλου δεδομένων βασίζεται στον στόχο της σύγκρισης των αποτελεσμάτων με προηγούμενες εργασίες. Το RAVDESS περιλαμβάνει ηχογραφήσεις και βιντεοσκοπήσεις των ηθοποιών. Αυτές οι οπτικοακουστικές πληροφορίες μπορούν να είναι χρήσιμες για εργασίες, όπως η πολυτροπική αναγνώριση συναισθημάτων, που περιλαμβάνουν την ανάλυση εκφράσεων του προσώπου και χειρονομιών εκτός από την ομιλία.

Μερικοί από τους άλλους λόγους για τους οποίους επιλέγεται το σύνολο δεδομένων RAVDESS έναντι άλλων για στην παρούσα εργασία, είναι το ευρύ φάσμα συναισθημάτων που επιτρέπει ολοκληρωμένες αναλύσεις για την αναγνώρισης συναισθημάτων. Είναι αξιόπιστη και συγκρίσιμη με διάφορες μελέτες ή εφαρμογές και περιλαμβάνει ηχογραφήσεις από πολλούς ηθοποιούς, άνδρες και γυναίκες. Αυτή η ποικιλομορφία όσον αφορά το φύλο βοηθά στην ανάπτυξη μοντέλων που δεν είναι προκατειλημμένα προς συγκεκριμένα άτομα. Μεταξύ των πλεονεκτημάτων της, επικρατεί ο ισορροπημένος αριθμός αρχείων ανά συναίσθημα, γεγονός που αποτρέπει τα προβλήματα που προκύπτουν από την εκπαίδευση αλγορίθμων με μη ισορροπημένα δεδομένα. Έτσι έχει αποκτήσει δημοτικότητα στην ερευνητική κοινότητα, πράγμα

που σημαίνει ότι υπάρχουν διαθέσιμοι πολυάριθμοι πόροι, έγγραφα και προ-εκπαιδευμένα μοντέλα [59].

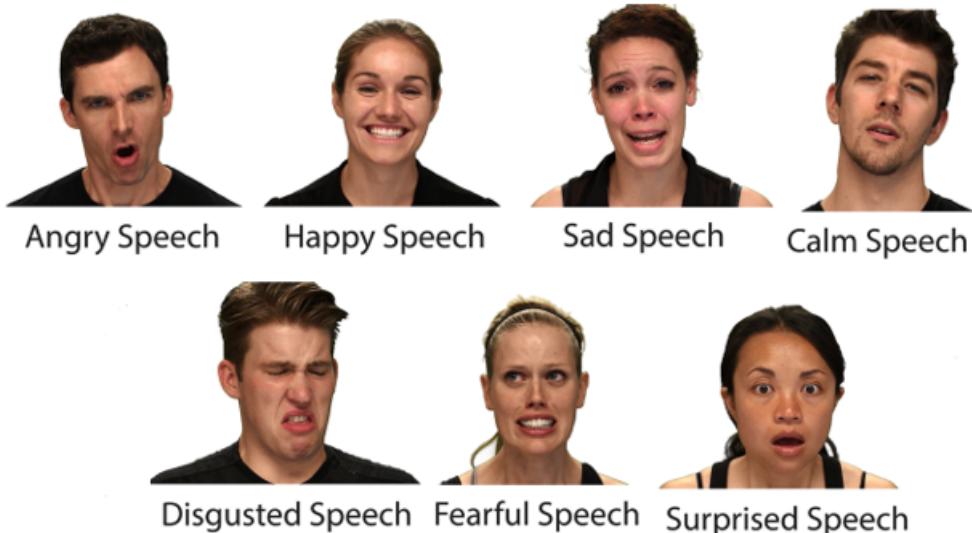
Το RAVDESS περιέχει 7356 εγγραφές με ενεργητικό-συναισθηματικό περιεχόμενο. Τα αρχεία αυτά χωρίζονται σε τρεις μορφές (πλήρες βίντεο, βίντεο μόνο με εικόνες και μόνο ήχος) και σε δύο φωνητικά κανάλια (ομιλία και τραγούδι). Κάθε αρχείο περιέχει έναν ηθοποιό που αντιπροσωπεύει ένα συναίσθημα που μπορεί να είναι μία από τις οκτώ ακόλουθες κατηγορίες: ηρεμία, ουδετερότητα, χαρά, λύπη, θυμός, φόβος, έκπληξη και αηδία. Οι εκφράσεις αυτές παράγονται σε δύο επίπεδα συναισθηματικής έντασης (κανονική και ισχυρή) εκτός από το ουδέτερο συναίσθημα που περιέχει μόνο κανονική ένταση. Στην παρούσα εργασία διατηρούνται τα επτά συναισθήματα (εκτός από το ουδέτερο) για να διατηρηθεί η εννοιολόγηση της εργασίας. Κατά την πειραματική διαδικασία χρησιμοποιείται μόνο το κανάλι ομιλίας, δεδομένου ότι επικεντρώνεται στο έργο της οπτικοακουστικής αναγνώρισης συναισθημάτων στην ομιλία αλλά όχι στα τραγούδια. Η επιλογή αυτή μείωσε τον αριθμό των αρχείων σε 1440 βίντεο που έχουν μέγιστη και ελάχιστη διάρκεια 5,31 και 2,99 δευτερόλεπτα, αντίστοιχα. Τα βίντεο ανήκουν σε 24 ηθοποιούς με ισορροπημένο φύλο, οι οποίοι εκφράζουν μόνο δύο λεξιλογικά ταιριαστές δηλώσεις με ουδέτερη βορειοαμερικανική προφορά, καθιστώντας τα κατάλληλα για τη μελέτη της παραγλωσσολογίας που σχετίζεται με τα συναισθήματα. Επιπλέον, μειώνεται η μεροληψία στις συναισθηματικές εκφράσεις που μπορεί να προκαλέσει η κουλτούρα.

Κάθε αρχείο RAVDESS που χρησιμοποιείται έχει ένα μοναδικό όνομα αρχείου. Το όνομα αρχείου αποτελείται από επτά διψήφια αριθμητικά αναγνωριστικά, τα οποία χωρίζονται με παύλες (π.χ. 02-01-06-01-02-01-12.mp4). Κάθε διψήφιο αριθμητικό αναγνωριστικό ορίζει το επίπεδο ενός διαφορετικού πειραματικού παράγοντα. Τα αναγνωριστικά είναι διατεταγμένα: Modality-Channel-Emotion-Intensity-Statement-Repetition-Actor.mp4 ή .wav. Η αριθμητική κωδικοποίηση των επιπέδων περιγράφεται στον πίνακα 1. Για παράδειγμα, το όνομα αρχείου "02-01-06-01-02-01-12. mp4" αναφέρεται σε: Μόνο βίντεο (02) - Ομιλία (01) - Φόβος (06) - Ένταση κανονική (01) - Δήλωση "dogs" (02) - Πρώτη επανάληψη (01) - Δωδέκατος ηθοποιός, γυναίκα (12). Το φύλο του ηθοποιού κωδικοποιείται από τον αριθμό του ηθοποιού, όπου οι ηθοποιοί με μονό αριθμό είναι άνδρες, ενώ οι ηθοποιοί με ζυγό αριθμό είναι γυναίκες.

Identifier	Coding description of factor levels
Modality	01 = Audio-video, 02 = Video-only, 03 = Audio-only
Channel	01 = Speech, 02 = Song
Emotion	01 = Neutral, 02 = Calm, 03 = Happy, 04 = Sad, 05 = Angry, 06 = Fearful, 07 = Disgust, 08 = Surprised
Intensity	01 = Normal, 02 = Strong
Statement	01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"
Repetition	01 = First repetition, 02 = Second repetition
Actor	01 = First actor, . . . , 24 = Twenty-fourth actor

Σχήμα 4.1: Περιγραφή της κωδικοποίησης των αρχείων του RAVDESS.
(Πηγή Σχήματος: [60])

Η επικύρωση του RAVDESS πραγματοποιήθηκε με 247 βαθμολογητές από τη Βόρεια Αμερική. Η εγκυρότητα αναφερόταν στην ακρίβεια με την οποία οι συμμετέχοντες αναγνώριζαν σωστά τα επιθυμητά συναισθήματα των ηθοποιών. Οι συνολικές βαθμολογίες ήταν υψηλές, επιτυγχάνοντας 80% για τον ήχο-βίντεο. Ένα επιπλέον σύνολο 72 συμμετεχόντων παρείχε δεδομένα δοκιμής-επαναληπτικής δοκιμής. Η αξιοπιστία αναφερόταν στην πιθανότητα οι συμμετέχοντες να επιλέξουν την ίδια συναισθηματική κατηγορία για ένα συγκεκριμένο ερέθισμα που παρουσιάστηκε δύο φορές. Συλλογικά, τα αποτελέσματα αυτά επιβεβαιώνουν ότι το RAVDESS έχει καλή εγκυρότητα και αξιοπιστία δοκιμής-επαναληπτικής εξέτασης [61].



Σχήμα 4.2: Οπτικά παραδείγματα των συναισθημάτων του RAVDESS.
(Πηγή Σχήματος: [62])

4.2 Εργαλεία ανάπτυξης διπλωματικής

Για την πολυτροπική αναγνώριση συναισθήματος με τεχνικές αυτόματης μάθησης χρησιμοποιούνται διαφορετικά εργαλεία. Τα εργαλεία είναι:

Python: Γλώσσα προγραμματισμού που χρησιμοποιείται για την εφαρμογή αυτών των τεχνικών. Έκδοση 3.10.4.

Pip: Εργαλείο που χρησιμοποιείται για την εγκατάσταση των βιβλιοθηκών Python με πιο άνετο, ταχύτερο και αποτελεσματικό τρόπο.

Compute Unified Device Architecture (CUDA): Πλατφόρμα υπολογιστών που περιλαμβάνει έναν μεταγλωττιστή και ένα σύνολο εργαλείων ανάπτυξης που δημιουργήθηκαν από την NVIDIA και επιτρέπει στο χρήστη να κωδικοποιεί αλγορίθμους σε NVIDIA GPU.

Keras: Βιβλιοθήκη νευρωνικών δικτύων γραμμένη σε Python. Χρησιμοποιείται για την εφαρμογή τεχνικών βαθιάς μάθησης.

Tensorflow & Tensorflow-gpu: Σύστημα backend που επιτρέπει τη μεταγλώττιση κώδικα βαθιών νευρωνικών δικτύων. Με την εφαρμογή της επέκτασης GPU, παρέχεται στον υπολογιστή η δυνατότητα να επεξεργαστεί μέσω της μονάδας επεξεργασίας γραφικών της κάρτας γραφικών.

OpenCV: Δωρεάν βιβλιοθήκη τεχνητής όρασης που αναπτύχθηκε από την Intel.

Numpy: Βιβλιοθήκη μαθηματικών συναρτήσεων υψηλού επιπέδου για την εργασία με μεγάλα διανύσματα και πίνακες.

Pandas: Βιβλιοθήκη λογισμικού γραμμένη σε Python για την επεξεργασία και ανάλυση δεδομένων. Προσφέρει δομές δεδομένων και λειτουργίες για το χειρισμό αριθμητικών πινάκων και χρονοσειρών.

Seaborn: Βιβλιοθήκη οπτικοποίησης δεδομένων Python. Παρέχει μια διεπαφή υψηλού επι-

πέδου για τη σχεδίαση ελκυστικών και κατατοπιστικών στατιστικών γραφικών.

Librosa: Πακέτο Python για την ανάλυση μουσικής και ήχου. Παρέχει τα απαραίτητα δομικά στοιχεία για τη δημιουργία συστημάτων ανάκτησης μουσικών πληροφοριών.

Sklearn: Βιβλιοθήκη μηχανικής μάθησης ελεύθερου λογισμικού στη γλώσσα προγραμματισμού Python. Διαθέτει διάφορους αλγορίθμους ταξινόμησης, παλινδρόμησης και ομαδοποίησης.

audiomentations: Βιβλιοθήκη Python για την επαύξηση δεδομένων ήχου (audio data augmentation).

LabelEncoder: Συνάρτηση της sklearn που αποτελεί μία τεχνική που χρησιμοποιείται για τη μετατροπή κατηγορικών στηλών σε αριθμητικές.

Κεφάλαιο 5

Αναγνώριση Συναισθήματος από Ήχο

5.1 Εισαγωγή

Η αναγνώριση συναισθημάτων ομιλίας (Speech Emotion Recognition) είναι η διαδικασία πρόβλεψης των ανθρώπινων συναισθημάτων από ηχητικά σήματα χρησιμοποιώντας τεχνικές τεχνητής νοημοσύνης. Στις προκλήσεις αναγνώρισης συναισθημάτων με βάση ηχητικά δεδομένα έχουν χρησιμοποιηθεί παραδοσιακοί αλγόριθμοι μηχανικής μάθησης, όπως τα κρυφά μοντέλα Markov (HMM), οι μηχανές διανυσμάτων υποστήριξης (SVM) και οι μέθοδοι που βασίζονται σε δέντρα αποφάσεων [63]. Οι ερευνητές ανέπτυξαν πρόσφατα διάφορες αρχιτεκτονικές βασισμένες σε νευρωνικά δίκτυα για να αυξήσουν την απόδοση της αναγνώρισης συναισθημάτων ομιλίας. Με την πρόοδο των προτεγγίσεων βαθιάς μάθησης έχουν προταθεί πιο σύνθετες αρχιτεκτονικές βασισμένες σε νευρωνικά συστήματα. Οι πληροφορίες συλλέγονται μέσω φασματογραφημάτων (spectograms) ή χαρακτηριστικών ήχου από τα ακατέργαστα ηχητικά σήματα. Τέτοια χαρακτηριστικά αποτελούν τα Mel-Frequency Cepstral Coefficients (MFCC) και τα Mel-Spectrograms που χρησιμοποιούνται για την εκπαίδευση μοντέλων ήχου βασισμένων σε CNN [64].

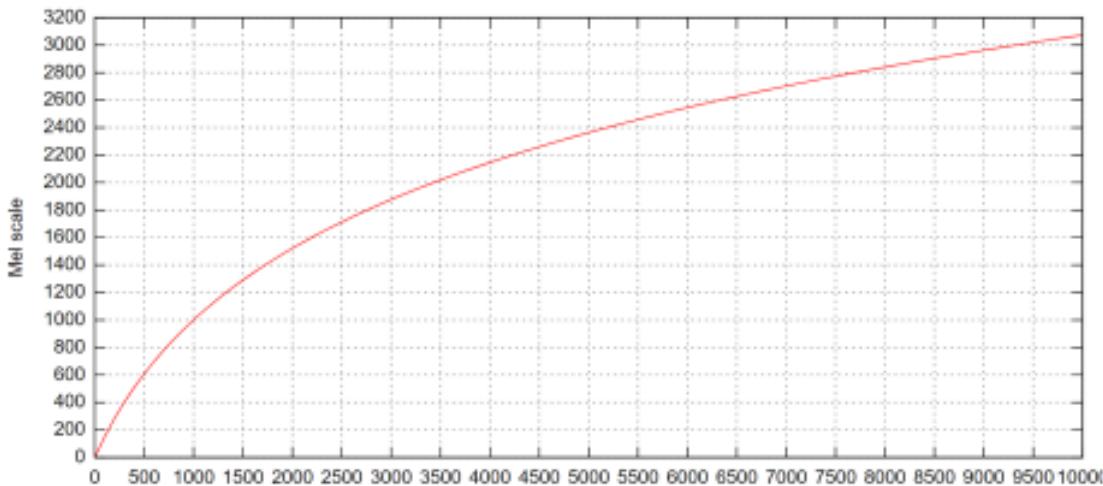
5.2 Φασματικά ακουστικά χαρακτηριστικά

Τα εξαγόμενα χαρακτηριστικά ήχου πρέπει να είναι ανθεκτικά σε ποικίλους παράγοντες όπως ο θόρυβος, η γλώσσα ή η κατάσταση του ομιλητή (π.χ. κρύο, άγχος, κούραση, μέθη κλπ.). Μπορούν να σχετίζονται με ποικίλα στοιχεία, όπως η διάρκεια των συνιστώσων του σήματος, η θεμελιώδης συχνότητα, η ποιότητα της φωνής, η ένταση, ο ρυθμός και η συχνό-

τητα. Τα ηχητικά χαρακτηριστικά που σχετίζονται με τη συχνότητα ή τα φασματικά χαρακτηριστικά διαδραματίζουν σημαντικό ρόλο στη αναγνώριση συναισθημάτων από ομιλία. Εξάγονται από τη φασματική πληροφορία του σήματος ομιλίας και συχνά αποκαλύπτουν χρήσιμες πληροφορίες σχετικά με τη συναισθηματική κατάσταση του ομιλητή. Στους τομείς της μουσικής και της αναγνώρισης συναισθημάτων ομιλίας, είναι σύνηθες να χρησιμοποιείται η κλίμακα Mel (Mel-scale) [65] για την αναπαράσταση της διάστασης της συχνότητας. Για την υλοποίησή της απαιτείται η εφαρμογή του παρακάτω τύπου στις τιμές της συχνότητας:

$$\text{Mel scaled frequency} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (5.1)$$

Εάν το φασματικό περιεχόμενο του χαρακτηριστικού ενός ηχητικού σήματος αναπαρίσταται ως συνάρτηση του χρόνου, τότε αποτελεί ηχητικό χαρακτηριστικό 1D. Η θεμελιώδης διαφορά μεταξύ 1D και 2D ηχητικών χαρακτηριστικών είναι ότι τα 2D χαρακτηριστικά αναπριστούν το φασματικό περιεχόμενο ως συνάρτηση τόσο του χρόνου όσο και της συχνότητας.



Σχήμα 5.1: Mel-scale.

(Πηγή Σχήματος: [66])

5.2.1 Mel-Frequency Cepstrum Coefficients (MFCC)

Τα MFCC είναι δημοφιλής για την υψηλή τους απόδοση και την ομοιότητά τους με το σύστημα της ανθρώπινης φωνής. Η εξαγωγή τους περιλαμβάνει τα ακόλουθα στάδια [67]:

1. Υπολογισμός του FFT της κυματομορφής ομιλίας.
2. Χαρτογράφηση στο Mel-scale χρησιμοποιώντας μια κατασκευασμένη τράπεζα φίλτρων Mel (Mel filter bank).
3. Μετατροπή του λογαριθμικού φάσματος Mel (Log Mel-Spectrum) πίσω στο χρόνο.

Τα πλάτη που προκύπτουν είναι τα MFCC. Συνήθως, μόνο οι πρώτοι συντελεστές απαιτούνται, καθώς παρέχουν επαρκή πληροφορία, αλλά αυτό εξαρτάται από τη συγκεκριμένη εφαρμογή στην εκάστοτε περίπτωση. Επιπλέον, ο αλγόριθμος εξαγωγής των MFCC μπορεί να διαφέρει σε διάφορες εφαρμογές και υλοποιήσεις. Έτσι, εξάγονται MFCCs κατά μήκος του άξονα του χρόνου και να οργανώνονται σε μια εικόνα 2D. Αυτή η εικόνα μπορεί στη συνέχεια να τροφοδοτηθεί σε ένα CNN για τον εντοπισμό των συναισθηματικού περιεχομένου που αποτυπώνεται στην εικόνα.

Τα MFCCs, ως συμπαγής αναπαράσταση του ηχητικού σήματος, χρησιμοποιούνται συχνά όταν υπάρχει ανάγκη μείωσης της διάστασης. Είναι κατάλληλα για εργασίες όπου το νευρωνικό δίκτυο πρέπει να εστιάζει στα πιο σημαντικά χαρακτηριστικά, απορρίπτοντας τις λιγότερο σημαντικές πληροφορίες. Λόγω της διαδοχικής φύσης τους (καθώς υπολογίζονται σε μικρά χρονικά παράθυρα) τα MFCC παρουσιάζουν μεγαλύτερη απόδοση ως είσοδοι σε επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) ή άλλες αρχιτεκτονικές που μπορούν να αξιοποιήσουν τις διαδοχικές εξαρτήσεις, σε σχέση με τα συνελικτικά νευρωνικά δίκτυα.

5.2.2 Mel-Spectrogram

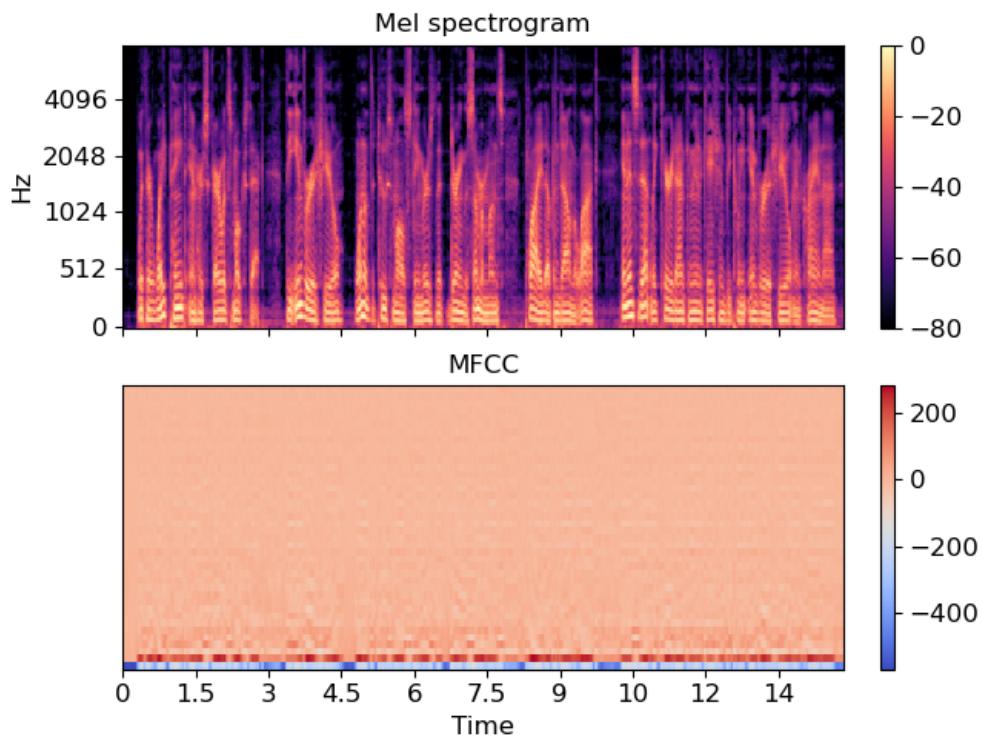
Ένα άλλο δημοφιλές χαρακτηριστικό στην επεξεργασία ήχου είναι το φασματογράμμα Mel (Mel-Spectrogram). Η απλουστευμένη διαδικασία υπολογισμού περιλαμβάνει τα ακόλουθα βήματα [68]:

1. Υπολογισμός του φασματογραφήματος (magnitude spectrogram).
2. Χαρτογράφηση στο Mel-scale χρησιμοποιώντας μια κατασκευασμένη τράπεζα φίλτρων Mel (Mel filter bank).

Ένα φασματόγραμμα Mel παρέχει μια οπτική αναπαράσταση του φασματικού περιεχομένου ενός σήματος ομιλίας με την πάροδο του χρόνου. Μπορεί να χρησιμοποιηθεί ως είσοδος για

διάφορες μεθόδους ανάλυσης ομιλίας και εργασίες αναγνώρισης συναισθημάτων.

Τα φασματογράμματα καταγράφουν το περιεχόμενο συχνότητας με την πάροδο του χρόνου, παρέχοντας μια δισδιάστατη αναπαράσταση του ηχητικού σήματος, η οποία μπορεί να είναι κατάλληλη για τα CNN που έχουν σχεδιαστεί για να εργάζονται με δεδομένα που μοιάζουν με εικόνες. Σε αυτή την προσέγγιση, το CNN μπορεί να μάθει ιεραρχικά χαρακτηριστικά από το φασματόγραμμα, καταγράφοντας δυνητικά τόσο τοπικά όσο και καθολικά μοτίβα.



Σχήμα 5.2: Mel-Spectrogram vs MFCC

(Πηγή Σχήματος: [69])

5.2.3 Σύγκριση των MFCC και Mel-Spectrogram

Στο πλαίσιο της αναγνώρισης συναισθημάτων ανθρώπινης ομιλίας, έχουν χρησιμοποιηθεί με επιτυχία τόσο τα φασματογράμματα Mel όσο και τα MFCC, και η επιλογή εξαρτάται συχνά από τα ειδικά χαρακτηριστικά της εργασίας και των δεδομένων. Ωστόσο, σε πρόσφατες έρευνες και για το σύνολο δεδομένων RAVDESS έχει παρατηρηθεί ότι τα φασματογράμματα Mel χρησιμοποιούνται συνήθως και μπορούν να είναι πιο αποτελεσματικά για εργασίες αναγνώρισης συναισθημάτων ομιλίας.

Τα συναισθήματα, στην ομιλία εκφράζονται συχνά μέσω αλλαγών στον τόνο και τον ρυθμό, οι οποίες μπορούν να παρατηρηθούν μέσω των δεδομένων. Τα φασματογράμματα Mel διατηρούν λεπτομέρειες που επιτρέπουν στο μοντέλο να συλλάβει τις διακυμάνσεις, στο ύψος, την ενέργεια και τα άλλα σημαντικά χαρακτηριστικά που σχετίζονται με το συναίσθημα.

Τα φασματογράμματα Mel διατηρούν πλούσια πληροφορία συχνότητας, επιτρέποντας στα CNN να διακρίνουν διακυμάνσεις στο φασματικό πεδίο. Σε ορισμένες ηχητικές εργασίες, ιδίως σε εκείνες όπου το λεπτομερές φασματικό περιεχόμενο είναι ζωτικής σημασίας, τα φασματογράμματα Mel μπορεί να είναι πιο κατατοπιστικά από τη συμπαγή αναπαράσταση που παρέχουν τα MFCC. Σε αντίθεση με τα MFCCs, τα φασματογράμματα Mel δεν συνεπάγονται απώλεια πληροφοριών λόγω μετασχηματισμού και διακριτού μετασχηματισμού συνημιτόνου. Αν και αυτή η απώλεια πληροφοριών μπορεί να είναι σκόπιμη για εργασίες, όπως η μείωση της διάστασης και η εστίαση στα σχετικά χαρακτηριστικά, μπορεί να μην είναι επιθυμητή για εργασίες όπου οι λεπτομερείς φασματικές πληροφορίες παίζουν κρίσιμο ρόλο.

Συνοψίζοντας, για την αναγνώριση συναισθημάτων ομιλίας, τα φασματογράμματα Mel συχνά προτιμώνται λόγω της ικανότητάς τους να διατηρούν λεπτομερείς φασματικές πληροφορίες και της συμβατότητάς τους με τα νευρωνικά δίκτυα συνελικτικής ανάλυσης.

5.3 Προ-επεξεργασία Δεδομένων

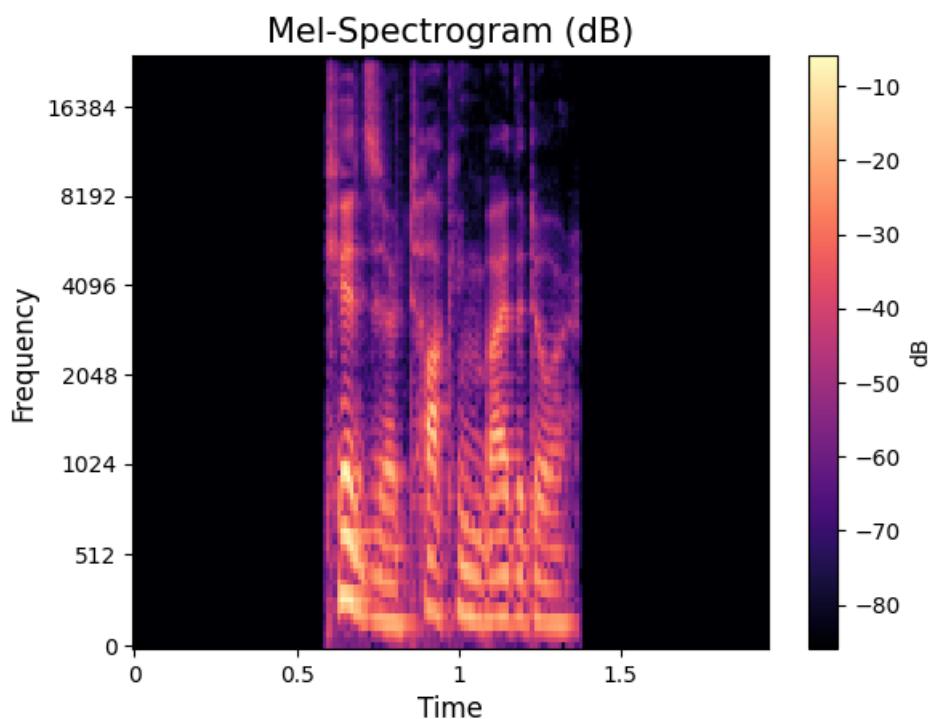
Για την αναγνώριση συναισθήματος ομιλίας χρησιμοποιείται η συλλογή δεδομένων RAVDESS που αναλύθηκε στο Κεφάλαιο 4 και συγκεκριμένα τα ηχητικά δεδομένα των ηθοποιών. Έπειτα, με την χρήση της συνάρτησης “load” της βιβλιοθήκης librosa που λαμβάνει ως είσοδο τη διαδρομή του αρχείου .wav αντλείται το ηχητικό σήμα και ο ρυθμός δειγματοληψίας. Έπειτα, ελέγχεται η βάση δεδομένων να είναι ισορροπημένη, δηλαδή αν για κάθε ηθοποιό υπάρχει ίδιος αριθμός αρχείων ήχου. Έτσι διαπιστώνεται ότι κάθε ένας από τους 24 ηθοποιούς έχει 60 αρχεία. Αυτά διαιρούνται αναλόγως της ετικέτας τους σε 7 κατηγορίες συναισθήματος (χαρά, λύπη, ηρεμία, θυμός, φόβος, αηδία, έκπληξη) και εξαιρείται το ουδέτερο συναίσθημα λόγο της ομοιότητας του με την ηρεμία.

Για κάθε αρχείο ήχου της συλλογής πραγματοποιείται η παρουσίαση της κυματομορφής του, εντοπίζεται ο αρχικός θόρυβος εξαιτίας της αναπνοής πριν την ομιλία, ανιχνεύεται η μέση διάρκεια των αρχείων να είναι περίπου 3 δευτερόλεπτα και οι στιγμές σιγής πριν και στο τέλος κάθε ηχητικού. Για την αντιμετώπιση αυτών των προβλημάτων δίνεται έμφαση στα πιο κεντρικά και ισχυρά μέρη των κυματομορφών. Για το σκοπό αυτό επιλέγεται η τεχνική του “τριμαρίσματος” με την χρήση της συνάρτησης “trim” της librosa με κατώφλι (threshold) τα 30db. Ως αποτέλεσμα απομακρύνεται μεγάλο μέρος της σιγής και του θορύβου. Παρόλα αυτά δεν έχει αντιμετωπιστεί το πρόβλημα της ίδιας διάρκειας για όλα τα ηχητικά κύματα (waveforms), κάτι που είναι αναγκαίο για την είσοδο των δεδομένων σε μοντέλα βαθιάς μάθησης. Για την εξισορρόπηση των κυματομορφών εντοπίζεται αυτό με την μέγιστη διάρκεια μετατρέποντας και τα υπόλοιπα στην ίδια, κάνοντας padding με μηδενικά.

Μετά την προετοιμασία των δεδομένων, ακολουθεί η εξαγωγή των χαρακτηριστικών και συγκεκριμένα των Mel-spectograms που όπως αναφέρθηλε είναι ικανοί είσοδοι για τα νευρωνικά δίκτυα που ακολουθούν. Αρχικά δημιουργείται και οπτικοποιείται ένα φασματόγραμμα Mel για ένα συγκεκριμένο δείγμα ήχου. Η διαδικασία περιλαμβάνει τη μετατροπή των ακατέργαστων δεδομένων ήχου σε ένα φασματόγραμμα Mel χρησιμοποιώντας τη βιβλιοθήκη Librosa. Για τον σκοπό αυτό επιλέγεται η συνάρτηση “Melspectrogram” με παραμέτρους τον ρυθμό δειγματοληψίας, τον αριθμό των δειγμάτων σε κάθε frame για τον FFT, τον αριθμό των Mels και τη μέγιστη και ελάχιστη συχνότητα που ορίζουν το εύρος της συχνότητας εφαρμογής της συνάρτησης. Στην παρούσα εργασία χρησιμοποιείται ως ρυθμός δειγματοληψίας τα 48000, ως αριθμός δειγμάτων σε κάθε frame τα 1024 και ως αριθμός των Mels τα 128 (αποτελεί μια συνηθισμένη επιλογή και παρέχει μια ισορροπία μεταξύ της καταγραφής λεπτομερών φασματικών πληροφοριών και της αποφυγής υπερβολικού υπολογιστικού κόστους). Επιπλέον, ως ελάχιστη και μέγιστη συχνότητα επιλέγονται τα 50 και 2400 hz αντίστοιχα. Το φασματόγραμμα Mel που παράγεται στη συνέχεια απεικονίζεται γραφικά με τη χρήση της Matplotlib όπως φαίνεται στην εικόνα 5.3, παρέχοντας μια οπτική αναπάρασταση του συχνοτικού περιεχομένου του σήματος με την πάροδο του χρόνου.

Για να την δημιουργία μιας πιο συστηματικής προσέγγισης στην εξαγωγή χαρακτηριστικών σε διάφορα δείγματα ήχου, υλοποιούνται δύο συναρτήσεις. Αυτές οι συναρτήσεις δέχονται

ως εισόδους τα ηχητικά σήματα και υπολογίζουν το φασματόγραμμα Mel χρησιμοποιώντας την συνάρτηση “Melspectrogram” της librosa πού περιγράφηκε στο προηγούμενο βήμα. Η διαφορά των δύο συναρτήσεων είναι ότι η μια υπολογίζει τη μέση τιμή κατά μήκος του άξονα του χρόνου, προσφέροντας μια συμπυκνωμένη αναπαράσταση των χρονικών χαρακτηριστικών του ηχητικού σήματος. Έτσι μέσω αυτών των δύο συναρτήσεων γίνεται η μαζική εξαγωγή των χαρακτηριστικών σε όλα τα ηχητικά σήματα της βάσης. Τελικά έχουμε ως αποτέλεσμα δύο είδη χαρακτηριστικών. Το πρώτο είδος περιέχει 2D χαρακτηριστικά που αποτελούν μια δισδιάστατη αναπαράσταση του φασματικού περιεχομένου του ηχητικού σήματος με την πάροδο του χρόνου. Οι διαστάσεις των εικόνων είναι (128,184), δηλαδή παίρνονται 128 mel για 184 στιγμές χρόνου. Οι εικόνες αυτές θα αποτελέσουν την είσοδο για ένα 2D CNN. Το δεύτερο είδος χαρακτηριστικών είναι 1D που περιέχουν μια τιμή συχνότητας (την μέση τιμή) με την πάροδο του χρόνου. Δηλαδή παίρνονται 184 μέσες τιμές συχνοτήτων, μια για κάθε στιγμή του χρόνου. Αυτές θα αποτελέσουν την είσοδο για ένα 1D CNN. Από αυτά τα χαρακτηριστικά διατηρείται ένα μέρος, της τάξεως περίπου του 17%, ως δεδομένα δοκιμής μεγέθους 184x1 στα 1D και μεγέθους 128x184x1 στα 2D.



Σχήμα 5.3: Αναπαράσταση του φασματογράμματος Mel για ένα τυχαίο δείγμα ήχου από το σύνολο των ηχητικών σημάτων.

Τα υπόλοιπα χαρακτηριστικά αποτελούν τα δεδομένα εκπαίδευσης για τα μοντέλα αναγνώρισης συναισθήματος από ομιλία. Στον τομέα της αναγνώρισης συναισθημάτων από ήχο, η προσθήκη δεδομένων εκπαίδευσης είναι ζωτικής σημασίας για τη βελτίωση της ισχύος και της γενίκευσης των μοντέλων. Η παρούσα εργασία διερευνά μια προσέγγιση, για την επαύξηση (Augmentation) των δεδομένων εκπαίδευσης SER χρησιμοποιώντας τεχνικές μετασχηματισμού του αρχικού συνόλου δεδομένων. Η διαδικασία επαύξησης περιλαμβάνει την προσθήκη θορύβου (Add Noise) για την προσομοίωση των πραγματικών παραλλαγών στην ακουστική, την προσαρμογή του τόνου (Pitch Scaling) ώστε να ληφθούν υπόψη τα εκφραστικά μοτίβα ομιλίας, την επιμήκυνση του χρόνου (Time Stretch) για να ληφθούν υπόψη οι παραμορφώσεις και τη μετατόπιση του χρόνου (Time Shift) ώστε να αντιμετωπιστούν οι παραλλαγές στην ευθυγράμμιση του σήματος. Τα σύνολα δεδομένων που προκύπτουν δημιουργούν μια συλλογή εκπαίδευσης που αποτυπώνει τις αποχρώσεις της ομιλίας. Το Augmentation παίζει ρόλο στη μείωση της υπερπροσαρμογής (overfitting), ενισχύοντας την απόδοση του μοντέλου και αυξάνοντας την προσαρμοστικότητα των συστημάτων SER για την αντιμετώπιση των διακυμάνσεων της συναισθηματικής έκφρασης. Πιο αναλυτικά για τις τεχνικές Augmentation έχουμε:

Add Noise: Αυτή η διαταραχή προσθέτει θόρυβο στο ηχητικό σήμα εισόδου σε έναν απαιτούμενο λόγο σήματος προς θόρυβο (SNR). Το SNR είναι ο λόγος του μέσου τετραγωνικού πλάτους (RMS) του σήματος προς το RMS πλάτος του θορύβου, στο τετράγωνο, το οποίο αποτέλεσε τη βάση της εφαρμογής. Στην παρούσα περίπτωση, χρησιμοποιείται Gaussian Noise για το είδος του θορύβου που προστίθεται στο αρχικό ηχητικό σήμα.

Pitch Scaling: Αυτός ο μετασχηματισμός αλλάζει το ύψος του ήχου εισόδου χωρίς να αλλάζει τη διάρκεια. Δηλαδή μετατοπίζει την συχνότητα (pitch) χωρίς να επηρεάζει την ταχύτητα του ηχητικού σήματος.

Time Stretching: Αυτή η τεχνική αλλάζει τη χρονική διάρκεια του ηχητικού σήματος εισόδου. Ανάλογα με την απαίτηση, μπορεί να τεντώσει ή να συμπιέσει χωρίς να αλλάξει το ύψος του ήχου. Πιο συγκεκριμένα αλλάζει την ταχύτητα του ηχητικού σήματος χωρίς να επηρεάζει το pitch.

Time Shift: Αυτός ο μετασχηματισμός παίρνει συγκεκριμένα κομμάτια του ηχητικού σήματος και τα μετατοπίζει με αποτέλεσμα τα ηχητικά σήματα να ακούγονται “μπροστά” (forwards) ή από το βάθος (backwards).

Ως εκ τούτου, ανακτούνται πλέον συνολικά 5600 δείγματα εκπαίδευσης ηχητικών σημάτων τόσο 1D διάστασης (184) όσο και 2D διαστάσεων (128, 184). Τα δείγματα αυτά όμως, πριν την εισαγωγή τους σε ένα μοντέλο αναγνώρισης ήχου βαθιάς μάθησης, προϋποθέτουν να έχουν μέση τιμή 0 και τυπική απόκλιση ίση με 1. Αυτό βοηθάει να επιτευχθεί πιο γρήγορη σύγκλιση του μοντέλου και να αποφευχθεί το πρόβλημα του vanishing point. Αυτό εφαρμόζεται μέσω της συνάρτησης “StandardScaler”. Επιπλέον, οι ετικέτες των συναισθημάτων πρώτα μετασχηματίζονται σε αριθμούς (0-Anger, 1-Calm, 2-Disgust, 3-Fear, 4-Happy, 5-Sad, 6-Surprise) μέσω του “LabelEncoder” και στη συνέχεια σε one-hot διανύσματα που προϋποθέτει η συνάρτηση απώλειας “Categorical Cross Entropy”. Αυτός ο μετασχηματισμός οδηγεί στην δημιουργία ετικετών διανυσμάτων μεγέθους 5600x7 (αριθμός συναισθημάτων), τα οποία αποτελούν τα διανύσματα εξόδου στα μοντέλα βαθιάς μάθησης. Τέλος, τα δεδομένα εκπαίδευσης αλλά και δοκιμής για να είναι συμβατά με την αρχιτεκτονική του CNN επιτάσσουν μια προσθήκη νέας διάστασης. Αυτή η διάσταση αντιπροσωπεύει τη διάσταση του καναλιού και στην συγκεκριμένη περίπτωση είναι ίση με 1.

5.4 Μεθοδολογία

Σε αυτή την ενότητα, παρουσιάζονται και συγκρίνονται μια σειρά από αρχιτεκτονικές βαθιάς μάθησης για το έργο ταξινόμησης συναισθημάτων ομιλίας. Για τη δημιουργία των μοντέλων, χρησιμοποιείται η γλώσσα προγραμματισμού Python. Όλοι οι αλγόριθμοι και οι αρχιτεκτονικές περιγράφονται λεπτομερώς στα παρακάτω υποκεφάλαια. Για την εφαρμογή των νευρωνικά δικτύων, εφαρμόζεται ένα υπάρχον πλαίσιο βαθιάς μάθησης που ονομάζεται Tensorflow. Το Tensorflow υποστηρίζει υπολογισμούς GPU, επιταχύνοντας σημαντικά την διαδικασία κατάρτισης. Επιπρόσθετα στην ενότητα παρουσιάζονται η πειραματική διάταξη και τα αποτελέσματα.

5.4.1 CNN1D-LSTM

Ως πρωταρχικός στόχος είναι η παραγωγή αποτελεσμάτων για τα αρχικά δεδομένα, δηλαδή γι' αυτά που δεν έχουν υποστεί Augmentation. Αυτά τα δεδομένα στο πρόγραμμα ονομάζονται “Vanilla”. Η χρήση αυτών είναι να παρθεί μια αρχική παρατήρηση για το ποσοστό της ακρίβειας της εκπαίδευσης (train) και επαλήθευσης (validation). Επομένως, απομονώνονται τα πρώτα 1120 από τα 5600 δεδομένα εκ των οποίων τα πρώτα 896 είναι τα δεδομένα εκπαίδευσης μεγέθους $896 \times 184 \times 1$ και τα υπόλοιπα αποτελούν τα δεδομένα επαλήθευσης (validation) μεγέθους $224 \times 184 \times 1$.

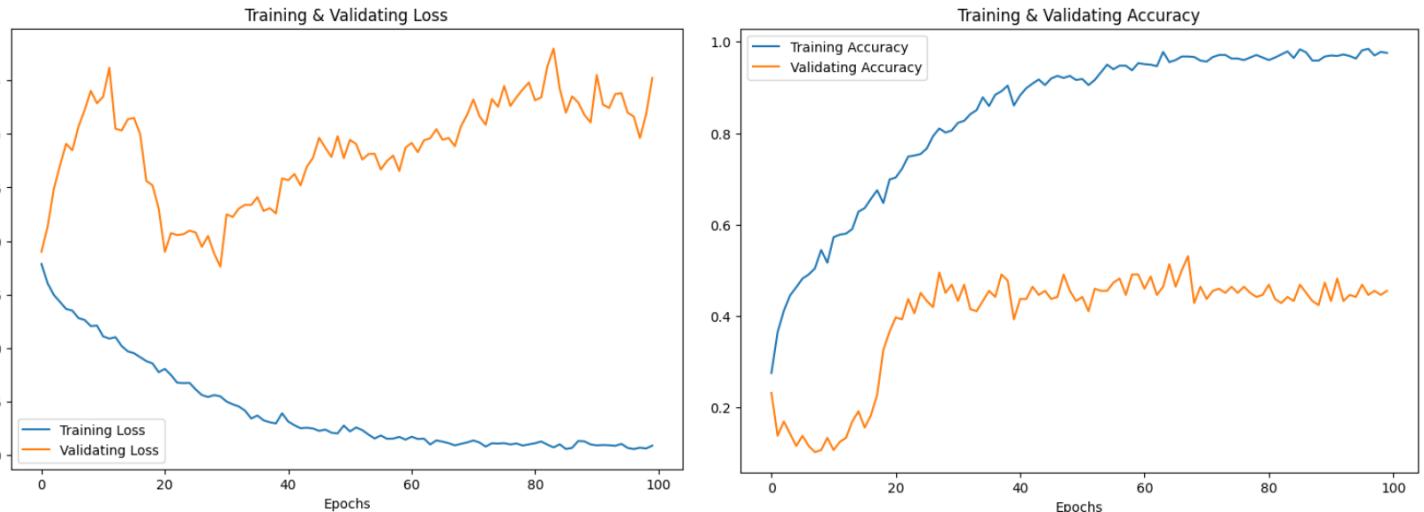
Η πρώτη αρχιτεκτονική που εφαρμόζεται λοιπόν, είναι ο συνδυασμός CNN και LSTM, με το CNN να παίζει το ρόλο του εξαγωγέα χαρακτηριστικών υψηλού επιπέδου και το LSTM να μοντελοποιεί τις χρονικές πληροφορίες και εξαρτήσεις με πλήρως συνδεδεμένα επίπεδα για την παραγωγή των πιθανοτήτων κλάσης. Για κάθε αρχείο ήχου στα δεδομένα εκπαίδευσης (896) δημιουργήθηκαν 184 features που παρέχονται ως είσοδος. Κάθε μία από τις 184 τιμές αντιπροσωπεύουν μια συνοπτική αριθμητική μορφή (μέσο όρο) ενός πλαισίου ήχου (Mel-φασματογράμματος) των 128 Mels. Έτσι η παρεχόμενη είσοδος για το πρώτο συνελικτικό στρώμα έχει μέγεθος 896 (αριθμός αρχείων εκπαίδευσης) $\times 184 \times 1$. Γι' αυτό το στρώμα το μέγεθος του kernel είναι 3, ο αριθμός των φίλτρων (feature maps) είναι 64 και η συνάρτηση ενεργοποίησης είναι η relu. Στην συνέχεια τα νέα δεδομένα εισέρχονται στο BatchNormalization Layer, το οποίο χρησιμοποιείται συνήθως μετά από κάθε CNN layer για να διασφαλιστεί ότι τα δεδομένα που περνούν σε υψηλότερα στρώματα κανονικοποιούνται. Αυτό το γεγονός ενισχύει τη σταθερότητα και την απόδοση καθώς επιταχύνει τη διαδικασία εκπαίδευσης [70, 71]. Έπειτα ακολουθεί το Pooling Layer και συγκεκριμένα εφαρμόζεται το Max-pooling που βοηθά το μοντέλο να εστιάζει μόνο στα κύρια χαρακτηριστικά κάθε κομματιού δεδομένων και δεν αλλάζει ανάλογα με τη θέση του.

Η ίδια διαδικασία εκτελείται ξανά τρεις φορές αλλάζοντας όμως τον αριθμό των φίλτρων στα συνελικτικά στρώματα σε 128, 128 και 64 αντίστοιχα. Οι τιμές αυτές επιλέχθηκαν έπειτα από δοκιμές, ως οι πιο αποδοτικές. Μετά από κάθε Maxpooling, εκτός του πρώτου, εφαρμόζεται Dropout με πιθανότητα εγκατάλειψης 20%. Η απομάκρυνση χρησιμοποιείται για να ξεπεραστεί το πρόβλημα της υπερπροσαρμογής όπως αναλύθηκε στο Κεφάλαιο 2.10. Για την καταγραφή των χρονικών εξαρτήσεων χρησιμοποιούνται δύο αρχιτεκτονικές LSTM με 64

units αντίστοιχα. Η είσοδος για το LSTM απαιτεί διανύσματα χαρακτηριστικών με χρονικά βήματα και ως εκ τούτου τα πλαίσια των χαρακτηριστικών χρησιμοποιούνται ως χρονικά βήματα. Έτσι η είσοδος του πρώτου LSTM αποτελεί την έξοδο του τελευταίου συνελικτικού στρώματος με μέγεθος 16×64 , όπου συμβολίζει 16 χρονικά βήματα των 64 features το καθένα. Τελικά το τελευταίο χρονικό βήμα που παρέχεται από το δεύτερο LSTM διακατέχει όλη τη πληροφορία από όλα τα προηγούμενα χρονικά βήματα και εισέρχεται ως είσοδος σε ένα Fully Connected μεγέθους 128 units με συνάρτηση ενεργοποίησης την relu. Αυτό με την σειρά του συνδέεται με ένα άλλο μικρότερου μεγέθους των 7 units (όσο και ο αριθμός των συναισθημάτων) και μία softmax για την παραγωγή των πιθανοτήτων κάθε κλάσης. Όλα τα Layers εκτός του Dropout και του BatchNormalization φαίνονται αναλυτικά στο Πίνακα 5.1.

layer	type	input	filter/units	kernel	stride	activation	output
data	input	184x1	N/A	N/A	N/A	N/A	184x1
conv1d_1	convolution 1D	184x1	64	3	1	relu	182x64
pool1d_1	max pooling 1D	182x64	N/A	3	2	N/A	91x64
conv1d_2	convolution 1D	91x64	128	3	1	relu	89x128
pool1d_2	max pooling 1D	89x128	N/A	3	2	N/A	45x128
conv1d_3	convolution 1D	45x128	128	3	1	relu	43x128
pool1d_3	max pooling 1D	43x128	N/A	3	2	N/A	22x128
conv1d_4	convolution 1D	22x128	64	3	1	relu	20x64
pool1d_4	max pooling 1D	20x64	N/A	3	2	N/A	10x64
lstm_1	recurrent	10x64	64	N/A	N/A	N/A	10x64
lstm_2	recurrent	10x64	64	N/A	N/A	N/A	64
dense_1	fully connected	64	128	N/A	N/A	relu	128
dense_2	fully connected	128	7	N/A	N/A	softmax	7

Πίνακας 5.1: CNN1D-LSTM Model Architecture



Σχήμα 5.4: Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.

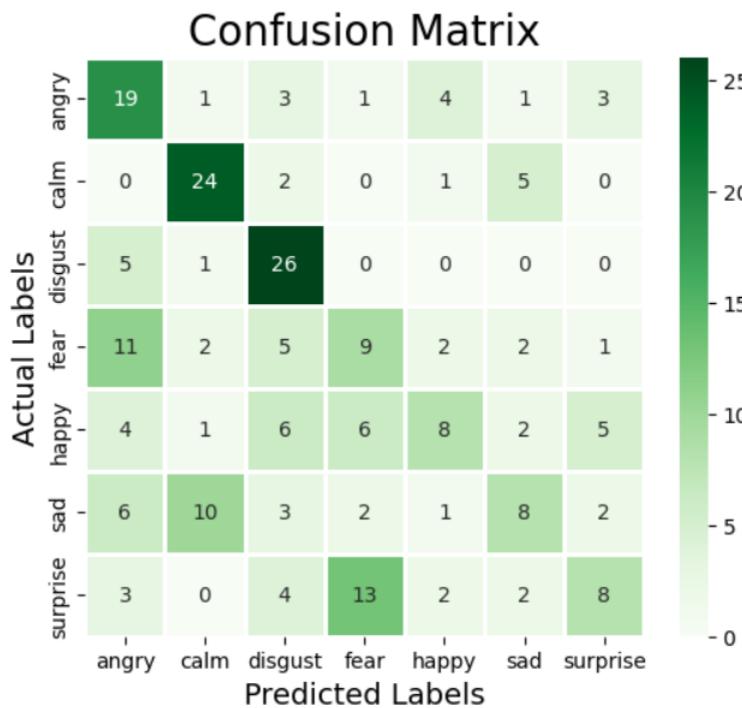
Για την διαδικασία εκπαίδευσης εφαρμόζεται ο βελτιστοποιητής Adam με learning rate 0.001, ως συνάρτηση απώλειας την Categorical Cross Entropy, ως μετρική η Categorical Accuracy και για τους hyperparameters batch size και epoch χρησιμοποιούνται οι τιμές 32 και 100 αντίστοιχα.

	precision	recall	f1-score
0	0.40	0.59	0.40
1	0.62	0.75	0.68
2	0.53	0.81	0.64
3	0.29	0.28	0.29
4	0.44	0.25	0.32
5	0.40	0.25	0.31
6	0.42	0.25	0.31
accuracy			0.46
macro avg	0.44	0.46	0.43
weighted avg	0.44	0.46	0.43

Πίνακας 5.2: Classification report CNN1D-LSTM

Η οπτική αναπαράσταση του Confusion matrix επιτυγχάνεται μέσω heatmap, επιτρέποντας τη γρήγορη κατανόηση της απόδοσης του μοντέλου. Κάθε κελί στο heatmap αντιπροσω-

πεύει τον αριθμό των περιπτώσεων, όπου η προβλεπόμενη ετικέτα και η πραγματική ετικέτα συμπίπτουν.



Σχήμα 5.5: Confusion matrix CNN1D-LSTM

Επομένως, όπως παρατηρείται και στον πίνακα 5.2 η αρχιτεκτονική αυτή οδηγεί σε 46% ακρίβεια στα δεδομένα επαλήθευσης (validation data). Τα αποτελέσματα όμως διαμορφώνονται ανομοιόμορφα όπως αποτυπώνεται στο Confusion Matrix 5.5. Δηλαδή σε κάποια συναισθήματα, όπως για την αηδία και την ηρεμία, η ακρίβεια είναι φανερά μεγαλύτερη από ότι σε άλλα που δεν υπάρχει το επιθυμητό αποτέλεσμα. Για παράδειγμα η συγκεκριμένη αρχιτεκτονική εμφανίζει ελάττωμα στη διάκριση των συναισθημάτων φόβου και έκπληξης. Για το σκοπό αυτό, είναι σημαντικό να εντοπιστεί μια αρχιτεκτονική με υψηλότερη και ομοιόμορφη ακρίβεια ανάμεσα στα συναισθήματα προτού προχωρήσουμε στην βελτιστοποίηση μέσω του augmentation data.

5.4.2 CNN2D

Με το ίδιο τρόπο σε αυτό το μοντέλο χρησιμοποιούνται τα αρχικά δεδομένα χωρίς να έχουν υποστεί Augmentation. Έτσι κι σε αυτήν την περίπτωση απομονώνονται τα πρώτα 1120 δεδομένα, με την διαφορά ότι εδώ έχουμε την πλήρη αναπαράσταση του Mel φασματογράμματος μεγέθους 128x184. Τελικά εμφανίζονται 896 δεδομένα εκπαίδευσης μεγέθους

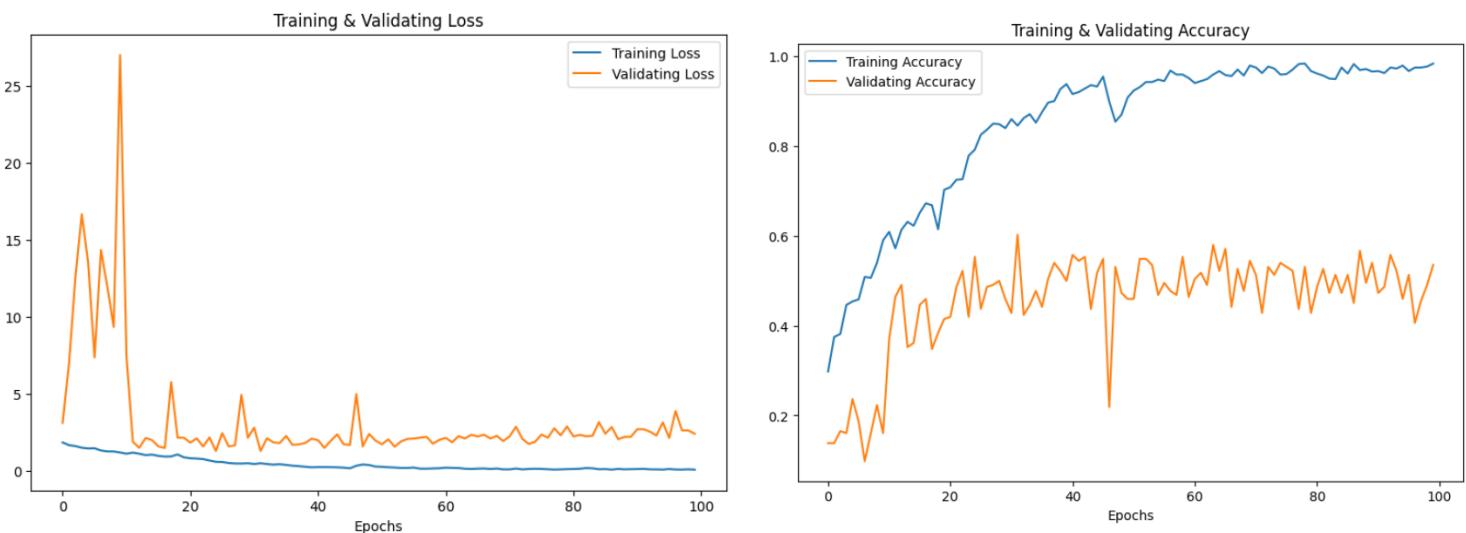
896x128x184x1 και 224 δεδομένα επαλήθευσης (validation) μεγέθους 224x128x184x1.

Σε αυτήν την αρχιτεκτονική εφαρμόζεται μία σειρά από 2D συνελεκτικά στρώματα για την εξαγωγή υψηλού επιπέδου χαρακτηριστικών από την 2D αναπαράσταση των Mel spectrograms, τα οποία με την σειρά τους τροφοδοτούνται σε ένα Fully-Connected στρώμα (classifier) για την παραγωγή των πιθανοτήτων κλάσης. Για κάθε αρχείο ήχου στα δεδομένα εκπαίδευσης (896) δημιουργήθηκαν 128x184 2D features που παρέχονται ως είσοδος στο πρώτο συνελικτικό στρώμα. Σε αυτό το στρώμα το μέγεθος του kernel είναι 3x3, ο αριθμός των φίλτρων (feature maps) είναι 64 και η συνάρτηση ενεργοποίησης είναι η relu. Τα νέα δεδομένα εισέρχονται στο BatchNormalization Layer για να διασφαλίσει ότι τα δεδομένα που περνούν σε υψηλότερα στρώματα κανονικοποιούνται. Στη συνέχεια, εφαρμόζεται ένα Maxpooling Layer με kernel διάστασης 3x3 και βήμα 2x2 με padding "same" όπου χρειάζεται.

Ξανά η ίδια διαδικασία επαναλαμβάνεται τρεις φορές, αλλάζοντας όμως τον αριθμό των φίλτρων στα συνελικτικά στρώματα σε 128, 128 και 64 αντίστοιχα. Οι τιμές αυτές επιλέχθηκαν έπειτα από δοκιμές ως οι πιο αποδοτικές. Μετά από κάθε Maxpooling εκτός του πρώτου εφαρμόζεται Dropout με πιθανότητα εγκατάλειψης 20%. Για να εισέλθουν τα 2D δεδομένα στο Fully-Connected στρώμα θα περάσουν πρώτα από το Flatten για να μετατραπούν σε 1D διανύσματα. Έτσι πλέον τα δεδομένα είναι στην κατάλληλη μορφή για να μεταφερθούν στο Fully-Connected στρώμα μεγέθους 64 units, με συνάρτηση ενεργοποίησης την relu. Αυτό με την σειρά του συνδέεται με ένα άλλο μικρότερου μεγέθους των 7 units (όσο και ο αριθμός των συναισθημάτων) και μία softmax για την παραγωγή των πιθανοτήτων κάθε κλάσης. Όλα τα layers εκτός του Dropout και του BatchNormalization φαίνονται αναλυτικά στο Πίνακα 5.3. Αυτή η προσέγγιση οδηγεί στα αποτελέσματα:

layer	type	input	filter/units	kernel	stride	activation	output
data	input	128x184x1	N/A	N/A	N/A	N/A	128x184x1
conv2d_1	convolution 2D	128x184x1	64	3x3	1x1	relu	126x182x64
pool2d_1	max pooling 2D	126x182x64	N/A	3x3	2x2	N/A	63x91x64
conv2d_2	convolution 2D	63x91x64	128	3x3	1x1	relu	61x89x128
pool2d_2	max pooling 2D	61x89x128	N/A	3x3	2x2	N/A	31x45x128
conv2d_3	convolution 2D	31x45x128	128	3x3	1x1	relu	29x43x128
pool2d_3	max pooling 2D	29x43x128	N/A	3x3	2x2	N/A	15x22x128
conv2d_4	convolution 2D	15x22x128	64	3x3	1x1	relu	13x20x64
pool2d_4	max pooling 2D	13x20x64	N/A	3x3	2x2	N/A	7x10x64
flatten	N/A	7x10x64	N/A	N/A	N/A	N/A	4480
dense_1	fully connected	4480	64	N/A	N/A	relu	64
dense_2	fully connected	64	7	N/A	N/A	softmax	7

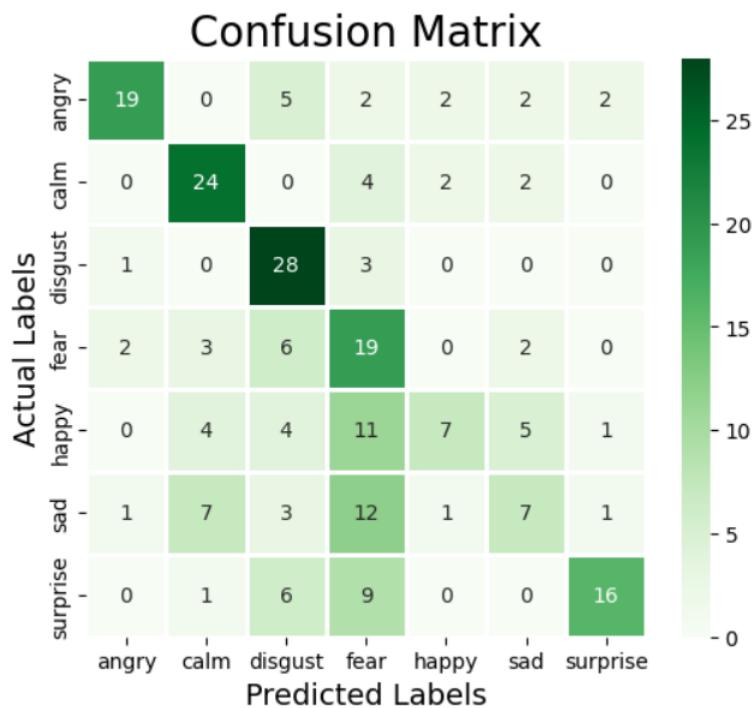
Πίνακας 5.3: CNN2D Model Architecture



Σχήμα 5.6: Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.

	precision	recall	f1-score
0	0.83	0.59	0.69
1	0.62	0.75	0.68
2	0.54	0.88	0.67
3	0.32	0.59	0.41
4	0.58	0.22	0.32
5	0.39	0.22	0.28
6	0.80	0.50	0.62
accuracy			0.54
macro avg	0.58	0.54	0.52
weighted avg	0.58	0.54	0.52

Πίνακας 5.4: Classification report CNN1D-LSTM



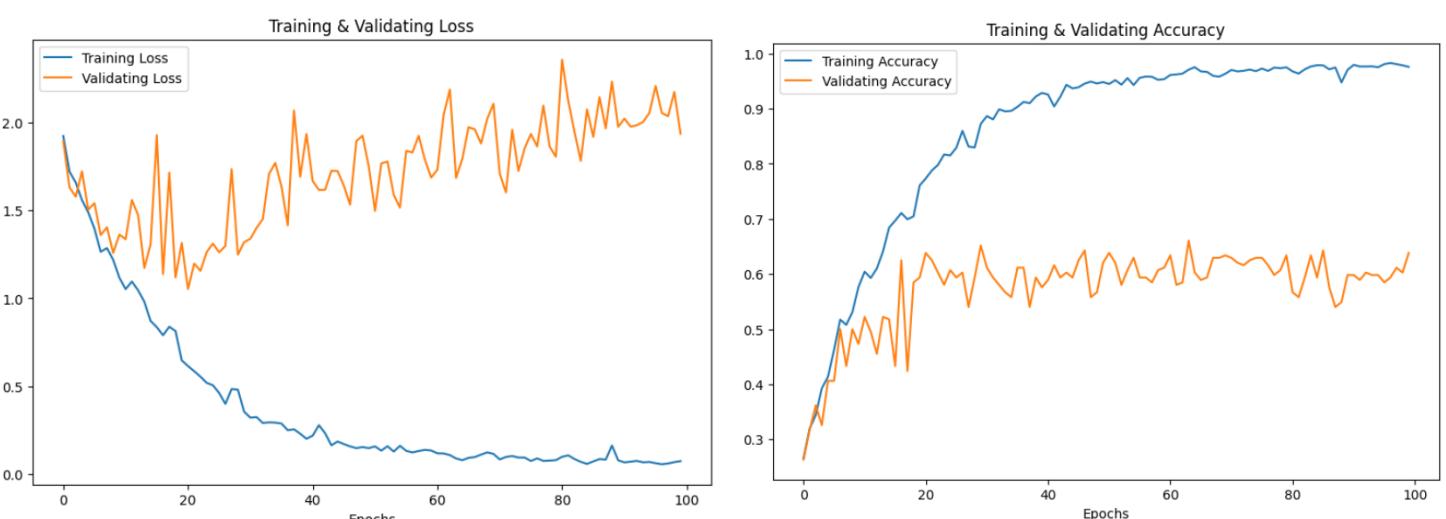
Σχήμα 5.7: Confusion matrix CNN2D

Όμοια εφαρμόστηκε ο βελτιστοποιητής Adam με learning rate 0.001, η συνάρτηση απώλειας Categorical Cross Entropy, η μετρική Categorical Accuracy ενώ για τους hyperparameters batch size και epoch χρησιμοποιήθηκαν οι τιμές 16 και 100 αντίστοιχα.

Και σε αυτήν την περίπτωση παρατηρείται από το Classification Report 5.4 ότι η ακρίβεια της αρχιτεκτονική αυτής αγγίζει το 54% στα δεδομένα επαλήθευσης (validation data). Η αύξηση της ακρίβειας εύκολα μπορεί να δικαιολογηθεί από το Confusion Matrix 5.7, το οποίο οδηγεί στην αντίληψη της ομοιόμορφης αντιστοίχησης συναισθημάτων στα πλαίσια αυτής της ακρίβειας. Σε αντίθεση με το CNN1D-LSTM, κατά πλειοψηφία σημειώνονται περισσότερες σωστές προβλέψεις για όλα τα συναισθήματα. Δηλαδή για το προβλεπόμενο συναίσθημα του φόβου, έχουμε περισσότερες ακριβείς προβλέψεις από ότι λανθασμένες σε σχέση με την προηγούμενη αρχιτεκτονική που εμφάνιζε περισσότερες στο συναίσθημα της έκπληξης. Μεγάλη επιτυχία υπάρχει και στα συναισθήματα ηρεμίας, θυμού και έκπληξης. Παρόλα αυτά, ακόμα σημειώνονται προβλήματα στη διάκριση των συναισθημάτων λύπης και χαράς. Βάση όλων των παραπάνω, η επιλογή της αρχιτεκτονικής CNN2D κρίνεται πιο ιδανική για την αναγνώριση συναισθημάτων μέσω ομιλίας. Έτσι η βελτίωση του αλγορίθμου θα προχωρήσει με την εφαρμογή του augmentation data.

5.4.3 CNN2D-Augmentation Data

Αυτή η μέθοδος βασίζεται στην ακριβώς προηγούμενη αρχιτεκτονική του πίνακα 5.3, του CNN2D, με τη μόνη διαφορά πως τα εισερχόμενα δεδομένα έχουν εμπλουντιστεί με τα επιπλέον επαυξημένα (Augmentation) δεδομένα, όπως παρουσιάστηκε ενδελεχώς στην Ενότητα 5.3.



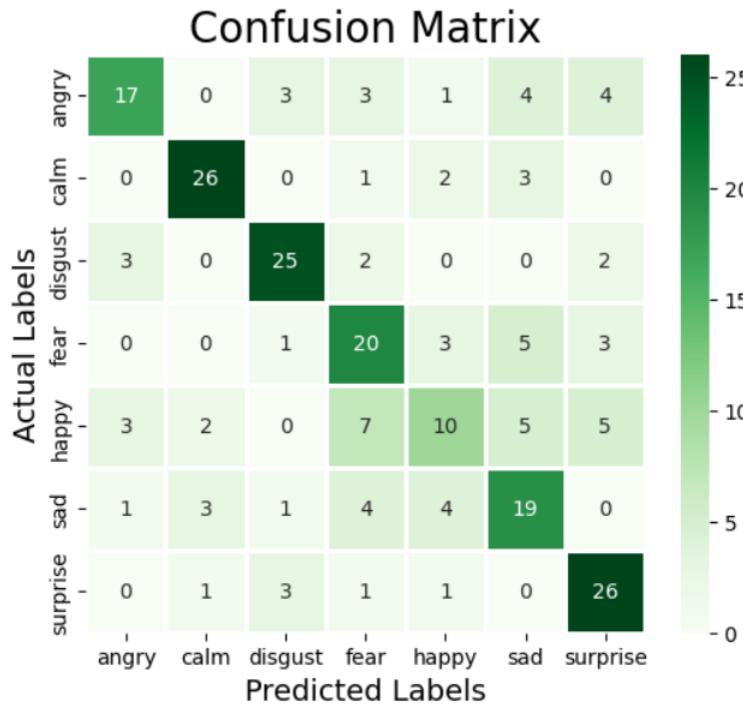
Σχήμα 5.8: Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.

Σε αυτήν την περίπτωση, για τα ίδια ακριβώς χαρακτηριστικά αρχιτεκτονικής του πίνακα 5.3 υπάρχουν 4480x128x184x1 (data+data-augmentation) δεδομένα εκπαίδευσης και 224 δεδομένα επαλήθευσης (validation) μεγέθους 224x128x184x1. Παρότι συνολικά διατίθενται 5600 επαυξημένα δεδομένα, στα δεδομένα επαλήθευσης συμπεριλμβάνονται μόνο τα “Vanilla” validation data καθώς επιθυμείται η αναγνώριση μόνο των αρχικών ομιλών. Τα επαυξημένα βιοηθούν αποκλειστικά για την εκπαίδευση του μοντέλου. Επιπλέον, είναι σημαντικό να αναφερθεί πως και σε αυτά τα δεδομένα έγινε κανονικοποίηση όπως ακριβώς και στα αρχικά. Επομένως, παίρνονται τα παραπάνω διαγράμματα ακρίβειας και απώλειας για τα δεδομένα επαλήθευσης και εκπαίδευσης.

Όμοια εφαρμόστηκε ο βελτιστοποιητής Adam με learning rate 0.001, η συνάρτηση απώλειας Categorical Cross Entropy, η μετρική Categorical Accuracy ενώ για τους hyperparameters batch size και epoch χρησιμοποιήθηκαν οι τιμές 16 και 100 αντίστοιχα.

	precision	recall	f1-score
0	0.71	0.53	0.61
1	0.81	0.81	0.81
2	0.76	0.78	0.77
3	0.53	0.62	0.77
4	0.48	0.31	0.38
5	0.53	0.59	0.56
6	0.65	0.81	0.72
accuracy			0.64
macro avg	0.64	0.64	0.63
weighted avg	0.64	0.64	0.63

Πίνακας 5.5: Classification report CNN2D-Augmentation Data



Σχήμα 5.9: Confusion matrix CNN2D-Augmentation Data

Παρατηρώντας το Classification report του πίνακα 5.5 σημειώνεται σημαντική αύξηση της ακρίβειας σε ποσοστό του 64%. Επίσης αυτή η αύξηση είναι εμφανής και από τον Confusion Matrix του σχήματος 5.9, καθώς στην συντριπτική πλειοψηφίας τους τα προβλεπόμενα συναισθήματα ταξινομούνται σωστά. Κανένα συναίσθημα δεν εμφανίζει σημαντικό αριθμό λανθασμένων προβλέψεων, αφού οι πιο έντονες περιοχές του Confusion Matrix εντοπίζονται στην κύρια διαγώνιο. Είναι σημαντικό να αναφερθεί ότι οι αναμενόμενες αποτυχίες δεν οδηγούν σε σύγχυση συγκεκριμένου συναισθήματος αλλά είναι κατανεμημένες ομοιόμορφα μεταξύ όλων.

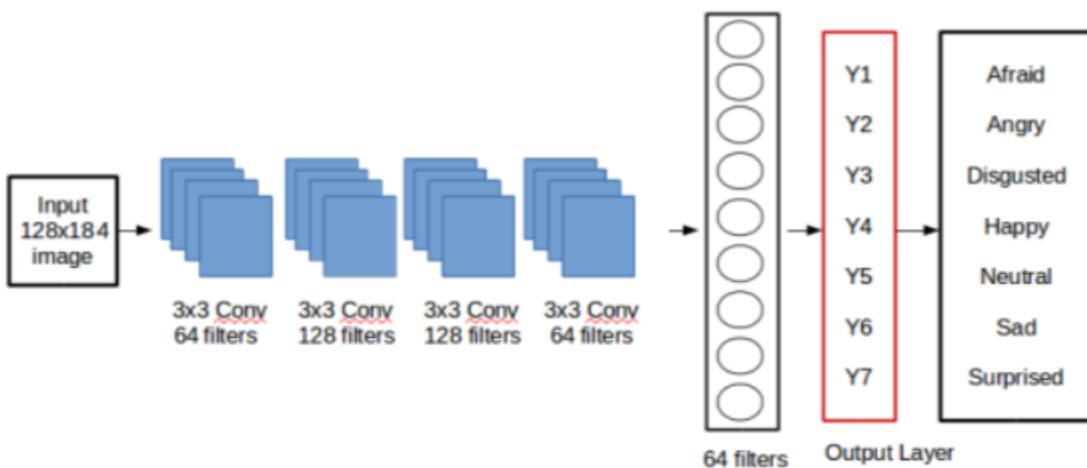
5.4.4 Σύνοψη

Model	Train Accuracy	Validation Accuracy
CNN1D-LSTM	0.98	0.46
CNN2D	0.99	0.54
CNN2D-AUGMENTATION DATA	0.98	0.64

Πίνακας 5.6: Σύνοψη των μοντέλων για αναγνώριση συναισθημάτων από ήχο.

Η σύνοψη όλων των παραπάνω διαδικασιών μπορεί εύκολα να απεικονιστεί στον Πίνακα 5.6, που επικεντρώνεται κυρίως στα σημαντικά αποτελέσματα που παράγονται. Δηλαδή περιέχει τα ποσοστά ακρίβειας στα δεδομένα εκπαίδευσης και επαλήθευσης. Σημαντικότερος παράγοντας όμως, αποτελεί η ακρίβεια επαλήθευσης, καθώς δείχνει πως ανταποκρίνεται το μοντέλο σε νέα δεδομένα γι' αυτό. Βάσει αυτού και όλων όσον προηγήθηκαν, επιλέγεται το CNN2D μοντέλο με επαυξημένα δεδομένα στην εκπαίδευση, χαρακτηριστικά και υπερπαραμέτρους που αναλύθηκαν στο υποκεφάλαιο 5.4.3 ως ο ιδανικός αλγόριθμος για την αναγνώριση συναισθημάτων μέσω ομιλίας.

Ιδιαίτερα, πρέπει να εξαχθούν ορισμένα συμπεράσματα σχετικά με την απόδοση. Τα CNNs έχουν καλές επιδόσεις, κερδίζοντας τα παραδοσιακά LSTMs. Γι αυτό η πειραματική ερεύνα διεξάγεται γύρω από ένα CNN2D εναντίον ενός CNN1D-LSTM στα δεδομένα εικόνων Mel-spectograms που παράγονται από τα αρχικά δεδομένα RAVDESS της Ενότητας 5.3. Επιπλέον, η χρησιμοποίηση του Dropout αυξάνει σχεδόν πάντα την απόδοση, δεδομένης της επαρκούς ρύθμισης των παραμέτρων. Όσο αφορά τον αριθμό των νευρώνων στο επίπεδο του νευρωνικού δικτύου είναι αρκετά απλός: όσο περισσότερους νευρώνες χρησιμοποιούνται τόσο βαθύτερο είναι το νευρωνικό δίκτυο, επομένως εξάγει πιο περίπλοκα μοτίβα από τα δεδομένα και άρα μεγαλύτερη απόδοση. Επιπλέον, οι συναρτήσεις ενεργοποίησης με την καλύτερη επίδοση είναι η “relu”. Ο βελτιστοποιητής με την καλύτερη επίδοση είναι ο Adam. Τέλος τα μοντέλα που βασίζονται στην επαύξηση των δεδομένων τους μπορούν να εγγυηθούν κέρδη στην ακρίβεια.



Σχήμα 5.10: Αναπαράσταση CNN2D.

Με βάση τα αποτελέσματα των δοκιμών, μπορούν να εξαχθούν διάφορα συμπεράσματα. Η υψηλότερη ακρίβεια στη δοκιμή των δεδομένων SER επιτυγχάνεται όταν χρησιμοποιούνται τα αρχικά δεδομένα σε συνδυασμό με τα επαυξημένα κατά την διάρκεια της εκπαίδευσης με ακρίβεια 64%. Η δεύτερη καλύτερη ακρίβεια εμφανίζεται όταν χρησιμοποιείται η αρχιτεκτονική του δισδιάστατου CNN χωρίς την χρήση των επαυξημένων δεδομένων στην διαδικασία της εκπαίδευσης. Αυτό είναι λογικό καθώς εισάγονται τα Mels-spectograms ως οπτική αναπαράσταση 2D διαστάσεων, κάτι που παρέχει όλη την απαραίτητη και σημαντική πληροφορία του σήματος φωνής. Ενώ αυτό με την χαμηλότερη ακρίβεια εμφανίζεται όταν εφαρμόζεται η CNN1D-LSTM αρχιτεκτονική, με ακρίβεια 46%. Μπορεί εύκολα να κατανοηθεί ότι τα 1D χαρακτηριστικά που περιέχουν μια τιμή συχνότητας (την μέση τιμή), με την πάροδο του χρόνου έχουν χάσει αρκετή σημαντική πληροφορία που είναι ικανή να ανεβάσει το ποσοστό της ακρίβειας. Λόγο αυτού εφαρμόστηκαν κατά σειρά δύο LSTM για να εξαχθεί η χρονική εξάρτηση μεταξύ των δεδομένων, κάτι που όπως αποδείχθηκε δεν ήταν ικανό να ξεπεράσει την απόδοση του CNN2D και της δισδιάστατης αναπαράστασης των δεδομένων.

Κεφάλαιο 6

Αναγνώριση Συναισθήματος από Εικόνες

6.1 Εισαγωγή

Η αναγνώριση συναισθημάτων προσώπου (Facial Emotion Recognition) είναι η τεχνολογία που αναλύει τις εκφράσεις του προσώπου από στατικές φωτογραφίες και βίντεο για να αποκαλύψει πληροφορίες σχετικά με τη συναισθηματική κατάσταση ενός ατόμου. Αρκετές μελέτες δείχνουν ότι οι εκφράσεις του προσώπου έχουν άμεση σχέση με τα συναισθήματα. Η ανθρώπινη ικανότητα αναγνώρισης συναισθήματος είναι αυτό που επιτρέπει σε δύο ανθρώπους να κατανοούν ο ένας τον άλλον. Το ανθρώπινο πρόσωπο είναι αναμφισβήτητα το πιο μελετημένο αντικείμενο στην όραση υπολογιστών. Το FER είναι μια αρκετά πολύπλοκη και κουραστική εργασία, αλλά έχει πολλές εφαρμογές σε διάφορους τομείς. Στην ακαδημαϊκή ιστορία, ο τομέας της αναγνώρισης συναισθημάτων προσώπου έχει περάσει από διάφορες φάσεις, από τα χειροποίητα χαρακτηριστικά (κλασσικοί μέθοδοι όρασης υπολογιστών) μέχρι τη χρήση τεχνικών Μηχανικής Μάθησης και Νευρωνικών Δικτύων. Η έρευνα μπορεί να χαρακτηριστεί σε δύο τύπους προσεγγίσεων [72]:

- Βασισμένες σε χαρακτηριστικά
- Ολιστική

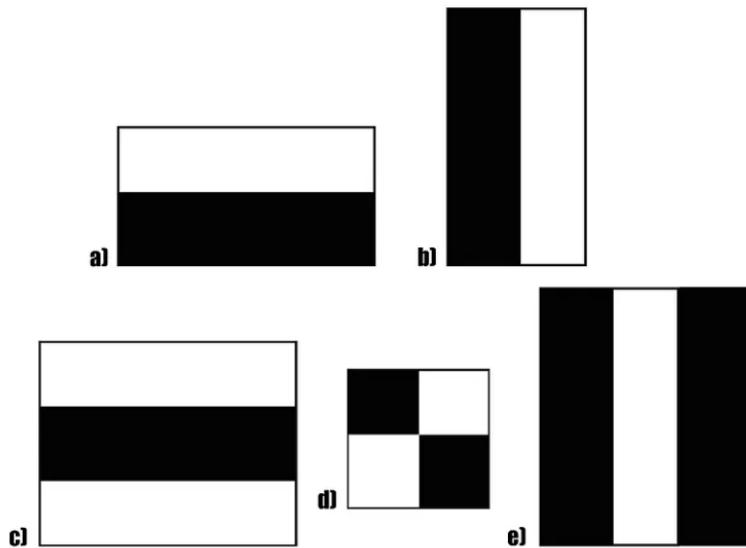
Οι πρώιμες εργασίες στον τομέα επικεντρώθηκαν εξ ολοκλήρου σε μεθόδους βασισμένες σε χαρακτηριστικά και προσπάθησαν ρητά να ορίσουν ένα χώρο αναπαράστασης χαμηλής διάστασης με βάση τις αναλογίες των γωνιών και των αποστάσεων. Αυτό είναι αρκετά περιοριστικό στην πράξη, επειδή η αυστηρά καθορισμένη από τον άνθρωπο αναπαράσταση

εννοιών και χώρων δεν είναι πάντα ορθή ούτε ακριβής. Οι εργασίες στα επόμενα χρόνια επικεντρώθηκαν σε ολιστικές προσεγγίσεις που προέρχονται από τα πεδία των μαθηματικών, της στατιστικής και της τεχνητής νοημοσύνης, οι οποίες λειτουργούν σε σύνολα δεδομένων προσώπου. Ένα παράδειγμα μιας τέτοιας μεθόδου είναι η ανάλυση κύριων συνιστωσών (PCA) και τα νευρωνικά δίκτυα. Η τελική “έκρηξη” στον τομέα σημειώθηκε τη δεκαετία του 2000, με τη χρήση των συνελικτικών νευρωνικών Δικτύων (CNN), τα οποία γνώρισαν τεράστια αύξηση όσον αφορά το ενδιαφέρον λόγω των ιδιαίτερων δυνατοτήτων και της ακρίβειας τους σε εργασίες που σχετίζονται με εικόνες.

6.2 Εξαγωγή Frame

Η εξαγωγή frames από αρχεία δεδομένων βίντεο είναι ένα θεμελιώδες βήμα που δίνει την δυνατότητα σε αυτήν την έρευνα να ενεργοποιήσει την χρήση τεχνικών Βαθιάς Μάθησης. Αρχικά, για κάθε βίντεο (κάθε ηθοποιό και κάθε συναίσθημα) εξάγονται frames διαστάσεων (224x398) ανά καθορισμένο διάστημα και αποθηκεύονται ως μεμονωμένες εικόνες png σε έναν νέο φάκελο που θα χρησιμοποιηθεί ως είσοδος στα μοντέλα που εφαρμόστηκαν. Πιο συγκεκριμένα το διάστημα ορίστηκε ίσο με τρία, δηλαδή παραβλέπονται τρία συνεχόμενα frames και αποθηκεύεται το επόμενο προς αποφυγή της επανάληψης πολλαπλών όμοιων στιγμιότυπων στα δεδομένα. Αυτή η διαδικασία διευκολύνει την προεπεξεργασία δεδομένων των βίντεο για μετέπειτα ανάλυση ή εφαρμογές, προσφέροντας ευελιξία στην εξαγωγή και αποθήκευση των καρέ για περαιτέρω επεξεργασία.

Στην συνέχεια πραγματοποιείται η ίδια προσέγγιση για την εξαγωγή frames διαστάσεων (224x224) που επικεντρώνονται αποκλειστικά στο πρόσωπο, δηλαδή στα σημεία που εμφανίζονται σε μεγαλύτερο βαθμό τα συναισθήματα. Η ανίχνευση ενός προσώπου μπορεί να είναι αρκετά πολύπλοκη, καθώς το ανθρώπινο πρόσωπο έχει διαφορετικά μεγέθη και σχήματα. Ανατρέχοντας στη βιβλιογραφία του παρελθόντος, ο πρώτος πρακτικός καθώς και ιδιαίτερα δημοφιλής αλγόριθμος ανίχνευσης προσώπου εμφανίζεται με την ονομασία Viola Jones Haar-cascade [73]. Διαχωρίζει τα πρόσωπα από τα μη πρόσωπα. Υπάρχουν δύο περιοχές, οι μαύρες σκιασμένες και οι λευκές σκιασμένες περιοχές.



Σχήμα 6.1: Ένα δείγμα χαρακτηριστικών του Haar που χρησιμοποιούνται στο Πρωτότυπο Ερευνητικό Έγγραφο που εκδόθηκε από τους Viola and Jones.

(Πηγή Σχήματος: [74])

Η πρώτη συμβολή στην έρευνα ήταν η εισαγωγή των χαρακτηριστικών Haar που φαίνονται στο προηγούμενο Σχήμα 6.1. Αυτά τα χαρακτηριστικά στην εικόνα διευκολύνουν τον εντοπισμό των άκρων ή των γραμμών στην εικόνα ή την επιλογή περιοχών όπου υπάρχει μια ξαφνική αλλαγή στις εντάσεις των pixel.

Το αποθετήριο έχει τα μοντέλα αποθηκευμένα σε αρχεία XML και μπορεί να διαβαστεί με τις μεθόδους OpenCV. Αυτά περιλαμβάνουν μοντέλα για ανίχνευση προσώπου, ανίχνευση ματιών, ανίχνευση άνω και κάτω σώματος, ανίχνευση πινακίδων κυκλοφορίας κλπ.



Σχήμα 6.2: Ένα δείγμα frames πριν και μετά την χρήση του Haar-Cascade.

6.3 Προ-επεξεργασία των Frames

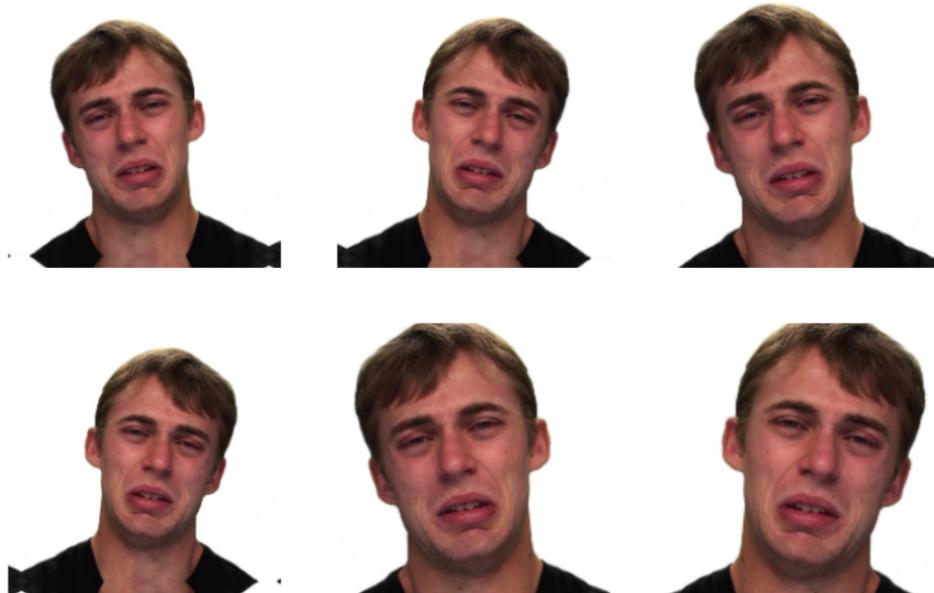
Η διαδικασία προ-επεξεργασίας των frames πρόκειται για μια κοινή διαδικασία και για τα δύο είδη frames ανά βίντεο που παράχθηκαν, δηλαδή για τα απλά και τα εστιασμένα στο πρόσωπο frames. Πιο συγκεκριμένα περιλαμβάνει τον διαχωρισμό των δεδομένων σε δεδομένα εκπαίδευσης και επαλήθευσης. Ως δεδομένα εκπαίδευσης χρησιμοποιούνται τα frames των ηθοποιών 1 έως 18 και αντίστοιχα επαλήθευσης των 19 και 20. Έπειτα, τα δεδομένα αντιστοιχίζονται με τις κατάλληλες ετικέτες συναισθημάτων.

Αφότου συλλεχθούν τα δεδομένα εισόδου-εξόδου που εφαρμόζονται στα μοντέλα Βαθιάς Μάθησης, δημιουργείται η ανάγκη εξισορρόπησης του αριθμού των πρώτων. Αυτό επιτυγχάνεται μέσω μίας συνάρτησης, η οποία πρώτα ελέγχει τον ελάχιστο αριθμό των frames ανά βίντεο και στην συνέχεια βάσει αυτού γίνεται τυχαία επιλογή ίδιου αριθμού frames και στα υπόλοιπα βίντεο. Έτσι, παράγονται 23 frames ανά βίντεο στα δεδομένα εκπαίδευσης δίνοντας σύνολο 23.184 frames και 24 frames ανά βίντεο στα δεδομένα επαλήθευσης δίνοντας σύνολο 2.688 frames. Τελικά τα δεδομένα τροποποιούνται στην κατάλληλη μορφή, δηλαδή χωρίζονται σε batch size των 64 με συνολικό αριθμό 362 για την εκπαίδευση και 42 για την επαλήθευση μέσω των κατάλληλων συναρτήσεων της tensorflow.

Τα Βαθιά νευρωνικά δίκτυα απαιτούν επαρκή δεδομένα εκπαίδευσης για να εξασφαλιστεί η γενίκευση σε μια δεδομένη εργασία αναγνώρισης. Ωστόσο, οι περισσότερες βάσεις δεδομένων που είναι διαθέσιμες στο κοινό, δεν διαθέτουν επαρκή ποσότητα και ποιότητα εικόνων για εκπαίδευση. Γεγονός που παρατηρείται και σε αυτήν την περίπτωση, όπως επιβεβαιώνεται και με τον συνολικό αριθμό των frames που παρουσιάστηκε παραπάνω, αλλά και από την επιβεβαίωση πως οι ασπρόμαυρες εικόνες είναι πιο αποδοτικές στα μοντέλα εξαιτίας του μικρότερου αριθμού πράξεων και περισσότερης έμφασης στην σημαντική πληροφορία. Επομένως, η αύξηση του αριθμού και η παραλλαγή χρώματος-φωτεινότητας των δεδομένων είναι ένα ζωτικό βήμα για ένα βαθύ FER. Παρόλα αυτά βάσει δοκιμών, διαπιστώνεται πως ανάλογα του είδους των frames απαιτείται διαφορετικός συνδυασμός τεχνικής Augmentation για να επιτευχθεί η μεγαλύτερη ακρίβεια στην εκάστοτε περίπτωση. Έτσι για τα ατόφια δεδομένα, αφού πρώτα εφαρμοστεί μια αποκοπή των διαστάσεων τους σε 224x224, ύστερα ενδεικνύονται σε τεχνικές Mirroring και Zooming. Το Mirroring αποτελεί το καθρέφτισμα των εικόνων με βάση τον κάθετο άξονα. Δεδομένου επίσης ότι τα βίντεο έχουν γυριστεί από

κάμερες που έχουν τοποθετηθεί σε διαφορετικά μέρη και ο ηθοποιός μπορεί να έχει κουνήσει το κεφάλι του κατά τη διάρκεια της ομιλίας, είναι ενδιαφέρον να υπάρχουν διαφορετικές οπτικές για την αναγνώριση. Γι' αυτό τον λόγο, εφαρμόστηκε τυχαία μεγέθυνση ή σμίκρυνση των εικόνων με παράγοντα ανάμεσα στις τιμές -0.2 έως 0.2.

disgust



Σχήμα 6.3: Augmentation των ατόφιων frames.

Για την περίπτωση ασπρόμαυρων εικόνων στον ίδιο φάκελο απαιτείται η μείωση των διαστάσεων των εικόνων σε 112x112 και η χρήση ίδιων τεχνικών Augmentation με το συνδυασμό της τυχαίας αποκοπής πληροφορίας. Αυτό πραγματοποιείται μέσω της συνάρτησης “RandomCutout” της keras_cv με τυχαία τοποθέτηση πλαισίου χρώματος μαύρου (τιμής 0) μεταβαλλόμενου μεγέθους.

surprise



Σχήμα 6.4: Augmentation των ασπρόμαυρων ατόφιων frames.

Ενώ στην περίπτωση των frames που εστιάζουν στο πρόσωπο, οι εικόνες επικεντρώνονται περισσότερο στην επιθυμητή πληροφορία. Αυτά αφού πρώτα μετατραπούν σε ασπρόμαυρα με διαστάσεις 112x112, μιας και με αυτή την τεχνική υπήρξε εμφανής βελτίωση στην απόδοση των μοντέλων, εφαρμόζονται σε αυτά οι ίδιες τεχνικές Augmentation. Μόνη διαφοροποίηση με την προηγούμενη περίπτωση, αποτελεί η μη εφαρμογή του Zooming.

calm



Σχήμα 6.5: Augmentation των ασπρόμαυρων frames που εστιάζουν στο πρόσωπο.

6.4 Μεθοδολογία

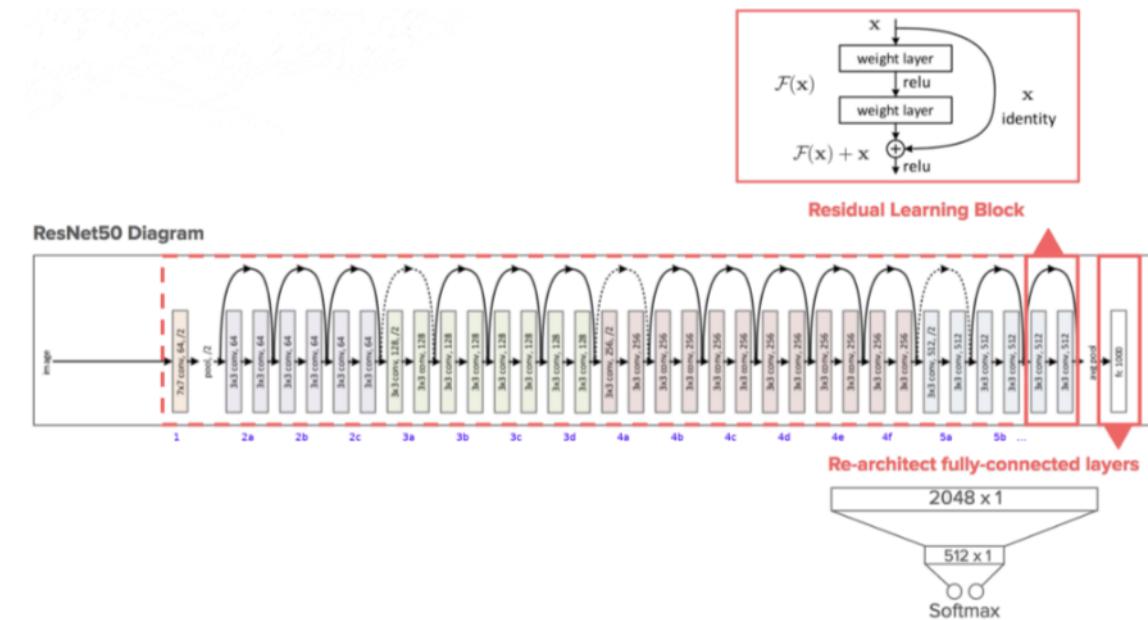
Σε αυτήν την ενότητα παρουσιάζονται και συγκρίνονται οι αρχιτεκτονικές Βαθιάς Μάθησης που εφαρμόζονται για την οπτική αναγνώριση συναισθημάτων. Όμοια με την αναγνώριση συναισθήματος από φωνή στην Ενότητα 5, δημιουργούνται με τα ίδια μέσα οι αρχιτεκτονικές Βαθιάς Μάθησης (CNNs) που όπως αποδείχθηκε αποδίδουν καλύτερα στην συγκεκριμένη βάση δεδομένων. Εντούτοις, παρόμοιες έρευνες κατέληξαν στο συμπέρασμα πως για μικρές βάσεις δεδομένων, όταν αυτές σχετίζονται με εικόνες και συνελικτικά μοντέλα Βαθιάς Μάθησης, είναι συνετό να δοκιμαστούν και τεχνικές Transfer Learning. Το Transfer Learning είναι μια τεχνική στη Βαθιά Μάθηση κατά την οποία η γνώση που αποκτάται από ένα προ-εκπαιδευμένο μοντέλο μπορεί να επαναχρησιμοποιηθεί σε μια άλλη, σχετική με αυτήν εργασία, προκειμένου να αυξηθεί η απόδοσή της. Συνεπώς τα μοντέλα Resnet-50 και Xception εμφανίζονται ως κατάλληλα για δοκιμή στην συγκεκριμένη εργασία.

Η μεθοδολογία και τα αποτελέσματα που παράγονται από τις πειραματικές διαδικασίες για τις κατηγορίες έγχρωμων και ασπρόμαυρων frames που εστιάζουν ή οχι στο πρόσωπο, θα αναλυθούν και θα συγκριθούν συγκεντρωτικά στη συνέχεια.

6.4.1 Έγχρωμα Frames

Πρωταρχικός στόχος είναι να ελεγχθούν τα αποτελέσματα του Transfer Learning για τα αρχικά δεδομένα, δηλαδή τα έγχρωμα frames όπως πάρθηκαν από τα βίντεο του RAVDESS. Το ResNet-50 είναι μια αρχιτεκτονική συνελικτικού νευρωνικού δικτύου (CNN) που έχει σχεδιαστεί για να υποστηρίζει εκατοντάδες ή χιλιάδες συνελικτικά επίπεδα. Οι προηγούμενες αρχιτεκτονικές του CNN δεν μπορούσαν να κλιμακωθούν σε μεγάλο αριθμό επιπέδων, γεγονός που είχε ως αποτέλεσμα περιορισμένη απόδοση. Ωστόσο, κατά την προσθήκη περισσότερων επιπέδων, οι ερευνητές αντιμετώπιζαν το πρόβλημα «εξαφανιζόμενης κλίσης». Το ResNet-50 παρέχει μια καινοτόμο λύση στο πρόβλημα της εξαφάνισης της κλίσης, που είναι γνωστό ως «παράλειψη συνδέσεων». Στοιβάζει πολλαπλά συνελικτικά επίπεδα που δεν κάνουν τίποτα στην αρχή, παρακάμπτει αυτά τα επίπεδα και επαναχρησιμοποιεί τις ενεργοποιήσεις του προηγούμενου επιπέδου. Η παράλειψη επιταχύνει την αρχική εκπαίδευση συμπιέζοντας το δίκτυο σε λιγότερα επίπεδα. Στη συνέχεια, όταν το δίκτυο επανεκπαιδευτεί, όλα τα επίπεδα επεκτείνονται και τα υπόλοιπα μέρη του δικτύου - γνωστά ως υπολειπόμενα

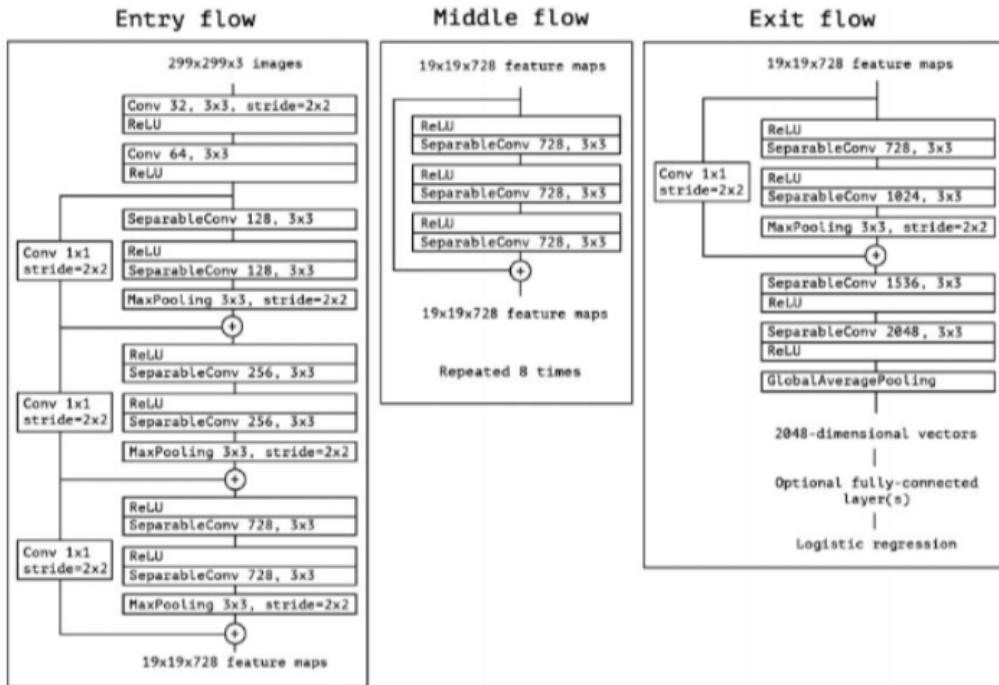
μέρη (Residual blocks) - επιτρέπεται να εξερευνήσουν περισσότερο τον χώρο χαρακτηριστικών της εικόνας εισόδου. Το ResNet-50 συγκεκριμένα είναι ένα συνελικτικό νευρωνικό δίκτυο 50 επιπέδων (48 συνελικτικά επίπεδα, ένα επίπεδο MaxPool και ένα επίπεδο μέσου όρου) [75].



Σχήμα 6.6: Η αρχιτεκτονική του Resnet-50.

(Πηγή Σχήματος: [76])

Επίσης η αρχιτεκτονική του Xception είναι αποδοτική και σε αρκετές περιπτώσεις ξεπερνάει κι αυτήν του Resnet-50. Αυτή βασίζεται σε δύο κύρια σημεία: την σύνδεση συντόμευσης μεταξύ των συνελικτικών επιπέδων όπως στο ResNet-50 και την διαχωρίσιμη κατά βάθος συνέλιξη (Depthwise Separable Convolution). Αυτές οι συνέλιξεις είναι εναλλακτικές των κλασσικών συνελίξεων και εμφανίζονται να είναι πιο αποδοτικές από την άποψη του χρόνου υπολογισμού. Αυτές με την σειρά τους ορίζονται σε δύο βασικά βήματα, τη συνέλιξη κατά βάθος (Depthwise Convolution) και τη σημειακή συνέλιξη (Pointwise Convolution). Το Depthwise Convolution είναι ένα πρώτο βήμα στο οποίο δεν πραγματοποιείται ο υπολογισμός της συνέλιξης σε όλα τα κανάλια, αλλά μόνο 1 προς 1 τη φορά. Μέχρι στιγμής, έχει υλοποιηθεί η λειτουργία συνέλιξης μόνο για 1 πυρήνα /φίλτρο της και όχι για όλα. Αφού ακολουθηθεί η ίδια διαδικασία και για τα υπόλοιπα φίλτρα, οδηγείται στο δεύτερο βήμα, τη σημειακή συνέλιξη (Pointwise Convolution). Αυτή λειτουργεί τη συνολική συνέλιξη για την παραγωγή του ίδιου αποτελέσματος [77].



Σχήμα 6.7: Η αρχιτεκτονική του Xception.

(Πηγή Σχήματος: [78])

Η χρήση του Transfer Learning μέσω των μοντέλων Resnet50 και Xception υποδεικνύονταν σε προηγούμενες σχετικές εργασίες ως υποσχόμενη. Παρόλα αυτά τα αποτελέσματα που παράγονται στην παρούσα εργασία είναι απογοητευτικά. Ως εκ τούτου τα μοντέλα για την αναγνώριση συναισθημάτων από εικόνες εκπαιδεύτηκαν από την αρχή με CNN αρχιτεκτονικές χρησιμοποιώντας τεχνικές Augmentation.

Πιο συγκεκριμένα, σε αυτήν την αρχιτεκτονική εφαρμόζεται μία σειρά από 2D συνελικτικά στρώματα για την εξαγωγή υψηλού επιπέδου χαρακτηριστικών από τα έγχρωμα ατόφια frames, τα οποία με την σειρά τους θα τροφοδοτηθούν σε ένα Fully-Connected στρώμα (classifier) για την παραγωγή των πιθανοτήτων κλάσης.

Για κάθε frame τριών καναλιών στα δεδομένα εκπαίδευσης (23.184) δημιουργούνται 224x224 2D features που παρέχονται ως είσοδο στο πρώτο συνελικτικό στρώμα. Η αρχικοποίηση των βαρών πραγματοποιείται μέσω kernel_initializer, και συγκεκριμένα του he_normal, ο οποίος ύστερα από δοκιμές αποδεικνύει πως αντιμετωπίζει καλύτερα το πρόβλημα του vanishing point. Σε αυτό το στρώμα το μέγεθος του kernel είναι 3x3, ο αριθμός των φίλτρων (feature maps) είναι 32 και με padding “same” όπου χρειάζεται.

Η συνάρτηση ενεργοποίησης που εφαρμόζεται είναι η elu. Η συνάρτηση ενεργοποίησης elu έχει ίδια αντιμετώπιση των θετικών βαρών κατά την εκπαίδευση με την relu, αλλά διαφοροποιείται στις αρνητικές τιμές καθώς αντί για 0 τις ανανεώνει σύμφωνα με τον όρο $a(e^x - 1)$, με a έναν θετικό παράγοντα. Αυτή η συνάρτηση επιτρέπει τη μέση τιμή των βαρών να είναι κοντά στο 0, γεγονός που απαιτείται στα μοντέλα Βαθιάς Μάθησης. Η elu αποτελεί εξέλιξη της relu μιας και οδηγεί συνήθως σε γρηγορότερη εκπαίδευση και βελτιστοποίηση της απόδοσης του μοντέλου. Αυτό αποδεικνύεται και στο τρέχον υποκεφάλαιο, όποτε προτιμάται αντί της relu.

Τα νέα δεδομένα εισέρχονται στο BatchNormalization Layer, για να διασφαλιστεί η κανονικοποίηση των δεδομένων που περνούν σε υψηλότερα στρώματα. Στη συνέχεια εφαρμόζεται ένα Maxpooling Layer με kernel διάστασης 2x2 και βήμα 1x1.

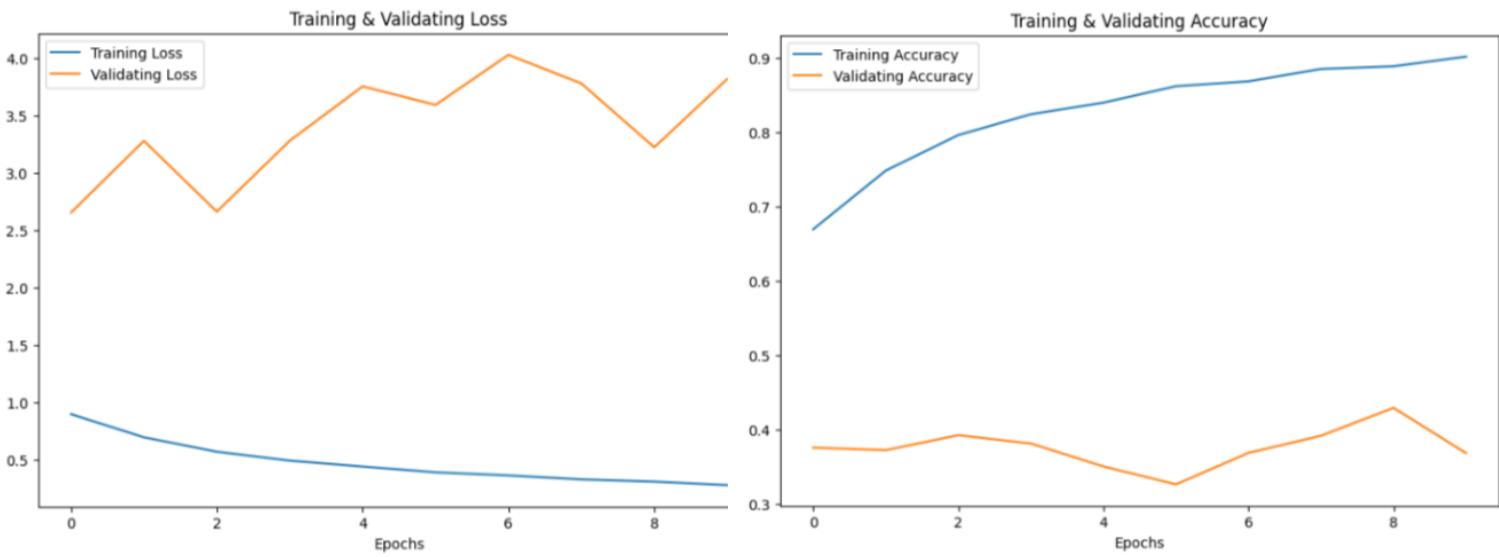
Ξανά η ίδια διαδικασία επαναλαμβάνεται τρεις φορές, αλλάζοντας μόνο τον αριθμό των φίλτρων στα συνελικτικά στρώματα σε 64, 128 και 128 αντίστοιχα. Οι τιμές αυτές επιλέγονται έπειτα από δοκιμές ως οι πιο αποδοτικές. Μετά από κάθε Maxpooling, εκτός του πρώτου, εφαρμόζεται Dropout με πιθανότητα εγκατάλειψης 20%, 20% και 30% αντίστοιχα.

Για να εισέλθουν τα 2D δεδομένα στο Fully-Connected στρώμα περνούν πρώτα από το Flatten για να μετατραπούν σε 1D διανύσματα. Έτσι πλέον τα δεδομένα είναι στην κατάλληλη μορφή για να μεταφερθούν στο Fully-Connected στρώμα μεγέθους 128 units, μια συνάρτηση ενεργοποίησης την elu, ένα στρώμα BatchNormalization και ένα Dropout με πιθανότητα εγκατάλειψης 40% κατά σειρά. Αυτό έπειτα συνδέεται με ένα άλλο μικρότερου μεγέθους των 7 units και μία softmax για την παραγωγή των πιθανοτήτων κάθε κλάσης. Όλα τα layers εκτός του Dropout και του BatchNormalization φαίνονται αναλυτικά στο Πίνακα 6.1.

layer	type	input	filter/units	kernel	stride	activation	output
data	input	224x224x3	N/A	N/A	N/A	N/A	224x224x3
conv2d_1	convolution 2D	224x224x3	32	3x3	1x1	elu	224x224x32
pool2d_1	max pooling 2D	224x224x32	N/A	2x2	1x1	N/A	112x112x32
conv2d_2	convolution 2D	112x112x32	64	3x3	1x1	elu	112x112x64
pool2d_2	max pooling 2D	112x112x64	N/A	2x2	1x1	N/A	56x56x64
conv2d_3	convolution 2D	56x56x64	128	3x3	1x1	elu	56x56x128
pool2d_3	max pooling 2D	56x56x128	N/A	2x2	1x1	N/A	28x28x128
conv2d_4	convolution 2D	28x28x128	128	3x3	1x1	elu	28x28x128
pool2d_4	max pooling 2D	28x28x128	N/A	2x2	1x1	N/A	14x14x128
flatten	N/A	14x14x128	N/A	N/A	N/A	N/A	25088
dense_1	fully connected	25088	128	N/A	N/A	elu	128
dense_2	fully connected	128	7	N/A	N/A	softmax	7

Πίνακας 6.1: CNN2D Model Architecture

Παρακάτω παραθέτονται επίσης οι πορείες της ακριβείας και της απόλειας ανά epoch και το Classification Report που προκύπτει μετά το τέλος της εκπαίδευσης του CNN2D των έγχρωμων ατόφιων frames.



Σχήμα 6.8: Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.

	precision	recall	f1-score
0	0.60	0.11	0.18
1	0.50	0.34	0.41
2	0.48	0.64	0.55
3	0.43	0.33	0.37
4	0.33	0.74	0.45
5	1.0	0.0	0.01
6	0.24	0.42	0.30
accuracy			0.37
macro avg	0.51	0.37	0.32
weighted avg	0.51	0.37	0.32

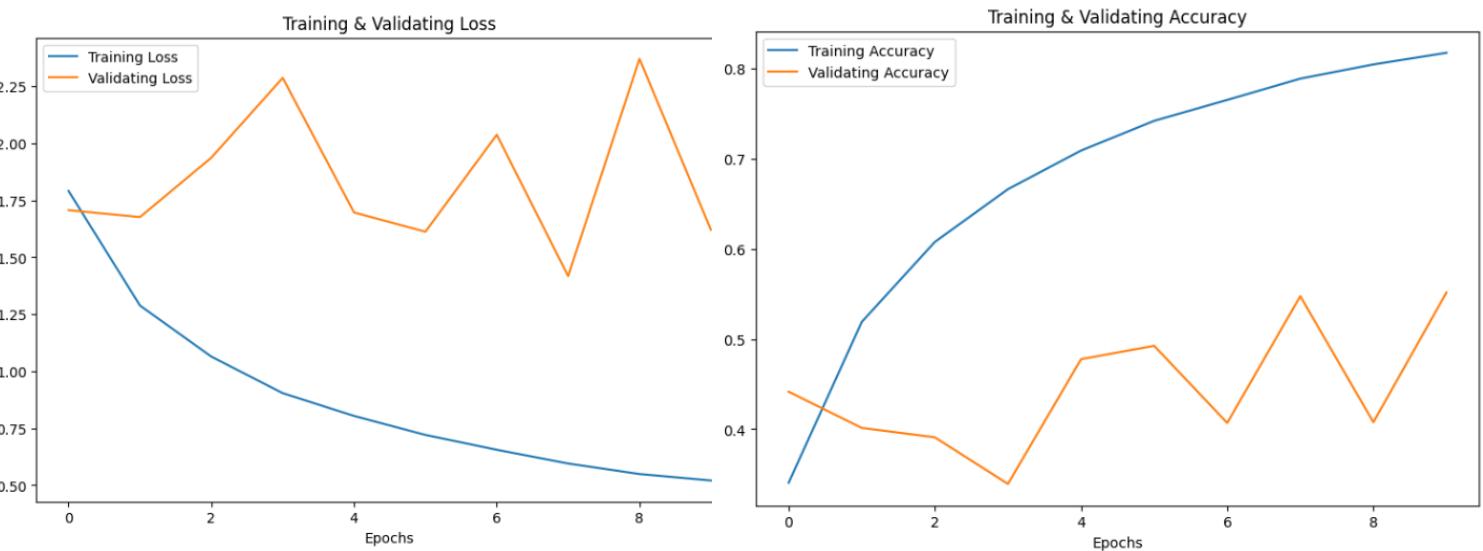
Πίνακας 6.2: Classification report του CNN2D στα έγχρωμα ατόφια frames.

Επομένως, μετά το τέλος της εκπαίδευσης το μοντέλο καταλήγει σε ποσοστό ακρίβειας 37%. Παρόλα αυτά απαιτείται υψηλότερη ακρίβεια, προκειμένου να πλησιάσει αυτή του ήχου για την καλύτερη απόδοση στην πολυτροπική προσέγγιση.

6.4.2 Ασπρόμαυρα Frames

Σκοπός αυτής της υποενότητας είναι να αποδείξει πως η εισαγωγή των ασπρόμαυρων frames αντί των έγχρωμων παράγει καλύτερα αποτελέσματα στην απόδοση. Είναι γνωστό ότι οι ασπρόμαυρες εικόνες έχουν μόνο ένα κανάλι το οποίο αντιπροσωπεύει την ένταση του φωτός, αντίθετα οι έγχρωμες εικόνες διαθέτουν τρία κανάλια. Χρησιμοποιώντας τις ασπρόμαυρες εικόνες μειώνεται η διάσταση των δεδομένων εισαγωγής, με αποτέλεσμα να απλοποιείται το μοντέλο και να μειώνονται οι υπολογιστικές απαιτήσεις. Αυτό οδηγεί σε γρηγορότερες εκπαιδεύσεις και μπορεί να είναι αρκετά σημαντικό για μεγάλα δεδομένα ή περιορισμένους υπολογιστικούς πόρους. Επιπλέον οι ασπρόμαυρες εικόνες είναι λιγότερο ευαίσθητες σε διακυμάνσεις των συνθηκών φωτισμού και σε παραμορφώσεις των χρωμάτων, καθιστώντας τις πιο ανθεκτικές στο θόρυβο και στις παραλλαγές των δεδομένα εισόδου. Επομένως εστιάζουν στις πληροφορίες φωτεινότητας ή έντασης, δίνοντας δυνητικά έμφαση στα κύρια χαρακτηριστικά και μειώνοντας την πολυπλοκότητα που εισάγει το χρώμα.

Ως εκ τούτου ακολουθείται η ίδια τακτική με την προηγούμενη υποενότητα 6.4.1. Και σε αυτή την περίπτωση το Transfer Learning αποδεικνύεται λιγότερο αποδοτικό σε σχέση με την CNN αρχιτεκτονική του Πίνακα 6.1, η οποία είναι η πιο αποδοτική στα έγχρωμα frames. Παρακάτω απεικονίζονται οι πορείες της ακρίβειας και της απώλειας ανά epoch και του Classification Report της μεθόδου αυτής για ατόφια frames με ένα κανάλι όμως.



Σχήμα 6.9: Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.

	precision	recall	f1-score
0	0.68	0.58	0.62
1	0.70	0.84	0.76
2	0.42	0.60	0.50
3	0.37	0.37	0.37
4	0.54	0.79	0.64
5	0.64	0.53	0.58
6	0.59	0.16	0.26
accuracy			0.55
macro avg	0.56	0.55	0.53
weighted avg	0.56	0.55	0.53

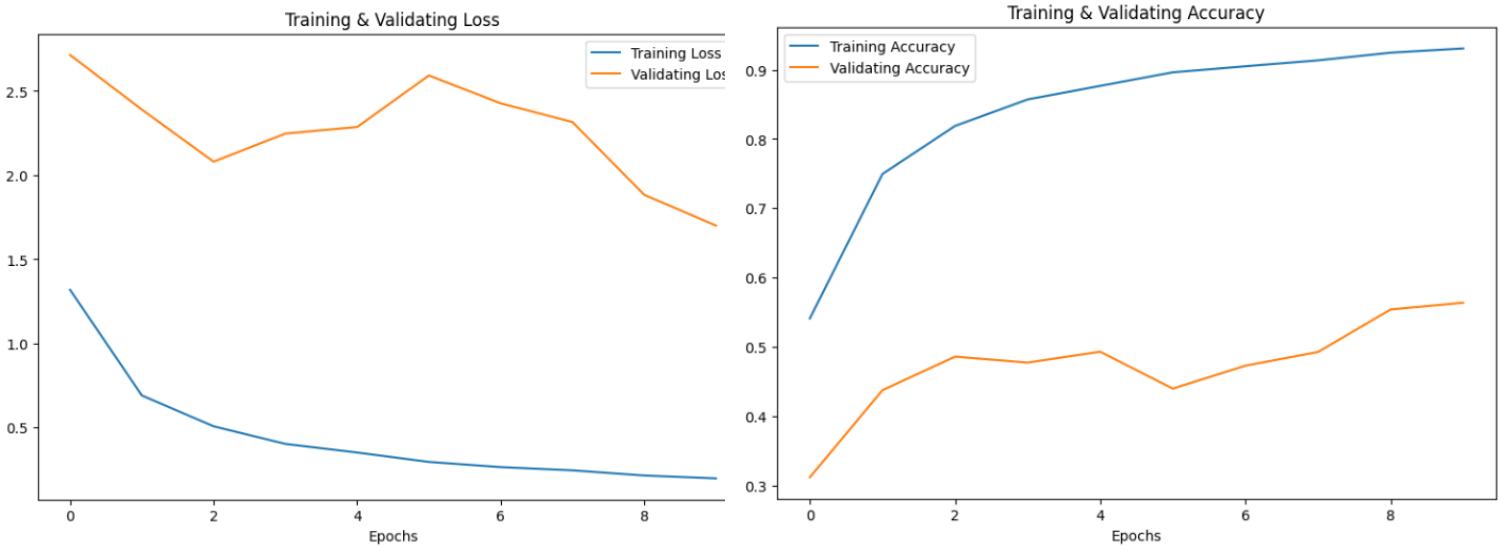
Πίνακας 6.3: Classification report του CNN2D σε ασπρόμαυρα ατόφια frames.

Παρατηρείται λοιπόν, αισθητή βελτίωση στην ακρίβεια με την μετατροπή των frames σε ασπρόμαυρα και συγκεκριμένα σε 55%. Η άνοδος αυτή κατά 18% οφείλεται στην απλοποίηση της πληροφορίας των εικόνων καθιστώντας τις πιο ανθεκτικές στο θόρυβο και τις διακυμάνσεις των συνθηκών φωτισμού.

6.4.3 Ασπρόμαυρα Frames εστιασμένα στο πρόσωπο

Δεδομένου της υποενότητας 6.4.2, τα ασπρόμαυρα frames καταλήγουν σε καλύτερα αποτελέσματα σε σχέση με τα έγχρωμα. Επομένως ενδείκνυται η δοκιμή για ασπρόμαυρες εικόνες στα frames που εστιάζουν στο πρόσωπο. Η λογική αυτής της δοκιμής ελλοχεύει σε ακόμη μεγαλύτερη απόδοση αφού frames τα οποία εστιάζουν στο πρόσωπο δίνουν περισσότερη έμφαση στην ουσιαστική πληροφορία (pattern) της εικόνας. Ο θεμελιώδης στόχος της οπτικής αναγνώρισης συναισθήματος είναι ο εντοπισμός της έκφρασης του προσώπου και κάθε διεργασία ανίχνευσης χαρακτηριστικών προσώπου είναι καθοριστική για την τελική απόφαση. Αυτό εν γένει οδηγεί στο συμπέρασμα πως τα εστιασμένα στο πρόσωπο frames θα οδηγήσουν σε καλύτερα αποτελέσματα διότι το μοντέλο θα επικεντρωθεί στην εξαγωγή αποδοτικότερων features για την αναγνώριση των συγκεκριμένων εκφράσεων του προσώπου. Επιπλέον με την περικοπή των εικόνων στο πρόσωπο μειώνονται σε μεγαλύτερο ακόμα βαθμό η πολυπλοκότητα και το μέγεθος των υπολογιστικών απαιτήσεων, απλοποιώντας την εκμάθηση των δεδομένων από το μοντέλο.

Βάσει όλων των παραπάνω, ακολουθείται εκ νέου η CNN αρχιτεκτονική του Πίνακα 6.1, η οποία είναι εξίσου αποδοτική στα έγχρωμα και ασπρόμαυρα frames. Η εκπαίδευση της υποενότητας αυτής πραγματοποιείται με augmented δεδομένα που δημιουργούνται μέσω των τεχνικών της ενότητας 6.3. Τα αποτελέσματα έπειτα από την εφαρμογή του CNN2D για τα ασπρόμαυρα frames που εστιάζουν στο πρόσωπο απεικονίζονται παρακάτω.



Σχήμα 6.10: Διαγράμματα Loss/Accuracy στα δεδομένα εκπαίδευσης και επαλήθευσης.

	precision	recall	f1-score
0	0.50	0.73	0.60
1	0.53	0.86	0.66
2	0.63	0.65	0.64
3	0.34	0.22	0.27
4	0.72	0.64	0.68
5	0.73	0.42	0.54
6	0.52	0.43	0.47
accuracy			0.57
macro avg	0.57	0.56	0.55
weighted avg	0.57	0.56	0.55

Πίνακας 6.4: Classification report του CNN2D για ασπρόμαυρα frames που εστιάζουν στο πρόσωπο.

Τέλος η βελτίωση στην ακρίβεια με την μετατροπή των frames σε ασπρόμαυρα και εστιασμένα στο πρόσωπο δεν ήταν σημαντική αφού υπήρξε άνοδος μόνο 2%. Αυτό είναι αναμενόμενο μιας και η βάση δεδομένων RAVDESS δημιουργήθηκε εξ αρχής με έμφαση στο πρόσωπο και το φόντο ήταν λευκό χωρίς κάποια επιρροή στη πληροφορία.

6.4.4 Σύνοψη

Model	Accuracy
CNN2D-COLOR FRAMES	0.37
CNN2D-BW FRAMES	0.55
CNN2D-BW FACE FRAMES	0.57

Πίνακας 6.5: Σύνοψη των μοντέλων για αναγνώριση συναισθημάτων από εικόνες.

Η σύνοψη όλων των παραπάνω διαδικασιών παρατίθεται στον Πίνακα 6.5, που επικεντρώνεται στα σημαντικά αποτελέσματα αυτών, δηλαδή το ποσοστό ακρίβειας στα δεδομένα επαλήθευσης μετά το πέρας της εκπαίδευσης. Βάσει αυτού και όλων όσων προηγούνται, επιλέγεται η προσέγγιση με CNN2D στα ασπρόμαυρα frames που εστιάζουν στο πρόσωπο ως ο ιδανικός αλγόριθμος για την οπτική αναγνώριση συναισθημάτων.

Τα αποτελέσματα των δοκιμών, υποδηλώνουν μεγάλη επίδραση των ασπρόμαυρων frames στην απόδοση της οπτικής αναγνώρισης. Υψηλότερη ακρίβεια στη δοκιμή των δεδομένων FER επιτυγχάνεται όταν χρησιμοποιούνται ως δεδομένα εκπαίδευσης τα ασπρόμαυρα frames καθώς επιτυγχάνουν ακρίβεια 55%. Επιπρόσθετα η εστίαση στο πρόσωπο αν και αποτελεί στην οπτική αναγνώριση την πλέον σημαντική τεχνική, δεν παρουσιάζει μεγάλη διαφορά στην ακρίβεια. Εν τούτοις βελτιώνει το FER κατά 2% ώστε να φτάσει ποσοστό ακρίβειας στα δεδομένα επαλήθευσης της τάξεως του 57%. Το ποσοστό αυτό είναι το υψηλότερο που επιτυγχάνεται και άρα η προσέγγιση με CNN2D στα ασπρόμαυρα frames επιλέγεται ως το ιδανικό μοντέλο για να συνδυαστεί με το αντίστοιχο του ήχο κατά την πολυτροπική διαδικασία της αναγνώρισης συναισθημάτων.

Κεφάλαιο 7

Πολυτροπική Αναγνώριση Συναισθήματος

7.1 Εισαγωγή

Με βάση τη θεωρία και την ανάλυση των Κεφαλαίων 5 και 6, εξάγεται το συμπέρασμα πως ο συνδυασμός ήχου και οπτικών δεδομένων στη Πολυτροπική Αναγνώριση Συναισθημάτων (Multimodal Emotion Recognition) συμβάλει στη δημιουργία ενός πιο ισχυρού μοντέλου αναγνώρισης συναισθημάτων. Οι περισσότερες εργασίες στην αναγνώριση συναισθημάτων επικεντρώνονται σε δεδομένα ενός μόνο τρόπου, όπως μια κριτική φωνής ή μια έκφραση προσώπου. Πιο πρόσφατες προσπάθειες στρέφονται στην πολυτροπική συγχώνευση, δεδομένου ότι το ανθρώπινο συναίσθημα εκφράζεται μέσω πολλαπλών τρόπων. Με την πρόοδο των μέσων κοινωνικής δικτύωσης και του περιεχομένου που δημιουργείται από τους χρήστες, ένας μεγάλος όγκος δεδομένων διαμορφώνεται από αυτούς. Τέτοια δεδομένα αποτελούνται από κείμενα (π.χ. Facebook), εικόνες (π.χ. Instagram), ήχο (π.χ. podcasts) και βίντεο (π.χ. YouTube). Όπως μπορεί κανείς να φανταστεί, η αυτόματη αναγνώριση της συναισθηματικής κατάστασης ενός ατόμου μπορεί να είναι ένα πολύ δύσκολο έργο και ως εκ τούτου η αποτελεσματική πολυτροπική αναγνώριση απαιτείται. Οι εκφράσεις του προσώπου και οι φωνητικές διαφοροποιήσεις που επιδεικνύει ο χρήστης, παρέχουν σημαντικές ενδείξεις για την καλύτερη αναγνώριση του συναισθήματος.

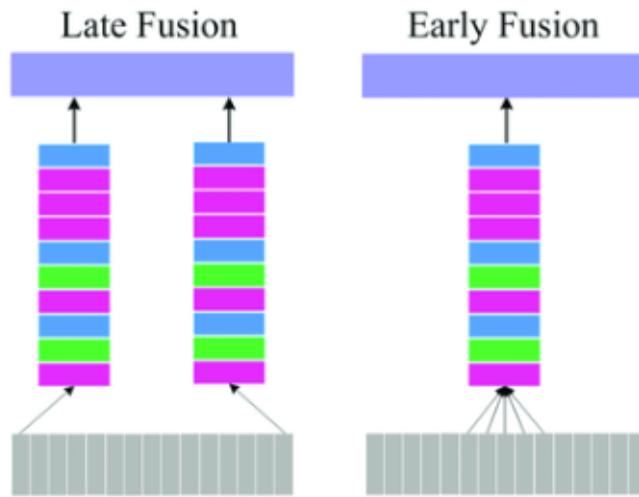
Τα πολυτροπικά συστήματα χρησιμοποιούν διαφορετικά κανάλια πληροφοριών με βάση το πρόσωπο, τη φωνή, καθώς και άλλες πηγές ταυτόχρονα. Μια εφαρμογή για παράδειγμα τέ-

τοιων μοντέλων είναι οι διεπαφές ανθρώπου-μηχανής σε ευφυή συστήματα που είναι κατασκευασμένα για να κατανοούν και να προσαρμόζονται στα ανθρώπινα συναισθήματα. Εάν το μοντέλο μπορεί να λάβει μια απόφαση με βάση τις παρεχόμενες πληροφορίες με επιτυχία συνδυάζοντας διάφορες πτυχές της συναισθηματικής έκφρασης, μπορεί επίσης να βοηθήσει σημαντικά στην περίπτωση της αλληλεπίδρασης ανθρώπου-υπολογιστή [79]. Η χρήση τεχνικών Βαθιάς Μάθησης στον τομέα εκτοξεύει τις επιδόσεις των μεθόδων ταξινόμησης και αποτελούν την κύρια κατεύθυνση που ακολουθείται σήμερα από ερευνητές, εισάγοντας ποικίλες προκλήσεις. Ο κύριος στόχος είναι η υλοποίηση ενός pipeline Βαθιάς Μάθησης, προκειμένου να αντιμετωπιστεί το πρόβλημα της κατανόησης των ανθρώπινων συναισθημάτων και να αυξηθεί η ακρίβεια των μοντέλων. Μια σημαντική πτυχή του πεδίου που διερευνάται είναι η συγχώνευση των τρόπων, η οποία συχνά πραγματοποιείται μέσω της συγχώνευσης σε επίπεδο χαρακτηριστικών (Early Fusion) ή και αποφάσεων (Late Fusion).

7.2 Τεχνικές Πολυτροπικής Συγχώνευσης

Σε αντίθεση με τις παραδοσιακές μονοτροπικές εργασίες, η πολυτροπική ανάλυση συναισθήματος υφίσταται μια διαδικασία συγχώνευσης κατά την οποία δεδομένα από διαφορετικές μορφές συγχωνεύονται. Προκειμένου να συγχωνευθούν οι πληροφορίες που εξάγονται από διαφορετικές λειτουργίες, έχει αναπτυχθεί ένα σύνολο διαφορετικών τεχνικών, γνωστές και ως Τεχνικές Πολυτροπικής Συγχώνευσης. Οι υπάρχουσες προσεγγίσεις μπορούν να ομαδοποιηθούν σε δύο κύριες κατηγορίες: τη συγχώνευση σε επίπεδο χαρακτηριστικών και σε επίπεδο απόφασης.

Η συγχώνευση σε επίπεδο χαρακτηριστικών (γνωστή ως Early Fusion) συγκεντρώνει όλα τα χαρακτηριστικά από κάθε τρόπο και τα ενώνει σε ένα ενιαίο διάνυσμα χαρακτηριστικών, το οποίο τελικά τροφοδοτείται σε έναν αλγόριθμο ταξινόμησης. Αυτός ο αλγόριθμος ταξινόμησης συνήθως είναι ένα βαθύ νευρωνικό δίκτυο. Η συγχώνευση σε επίπεδο απόφασης (γνωστή και Late Fusion) τροφοδοτεί τα δεδομένα που παράγονται από τους μονοτροπικούς μεθόδους και τα συνδέει στον δικό της αλγόριθμο ταξινόμησης. Έτσι, λαμβάνεται το τελικό συναίσθημα ταξινόμησης από τη συγχώνευση κάθε αποτελέσματος σε ένα ενιαίο διάνυσμα πιθανότητας.



Σχήμα 7.1: Μέθοδοι πολυτροπικής συγχώνευσης.

(Πηγή Σχήματος: [80])

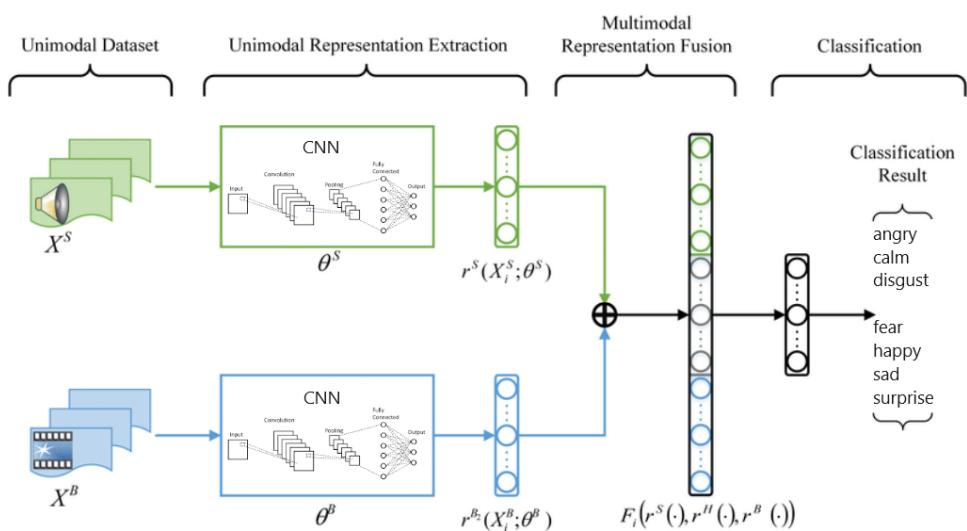
Ένα πλεονέκτημα της συγχώνευσης σε επίπεδο απόφασης είναι η εξάλειψη της ανάγκης συγχώνευσης ετερογενών δεδομένων από διαφορετικές πηγές. Εντούτοις κάθε μέσο μπορεί να χρησιμοποιήσει τον καταλληλότερο αλγόριθμο ταξινόμησης και να βελτιώνει ξεχωριστά κάθε μονοτροπική τεχνική στο μέγιστο δυνατό βαθμό. Κάθε πηγή μπορεί να επεξεργαστεί ξεχωριστά και η συγχώνευση πραγματοποιείται αργότερα στη διαδικασία της λήψης αποφάσεων. Αυτή η ευελιξία πλεονεκτεί όταν πρόκειται για διαφορετικές και δυναμικές πηγές δεδομένων. Αντίθετα, κατά τη συγχώνευση σε επίπεδο χαρακτηριστικών απαιτείται η ενσωμάτωση των πληροφοριών σε πρώιμο στάδιο. Γεγονός που μπορεί να περιορίσει την προσαρμοστικότητα σε διαφορετικούς τύπους δεδομένων ή σε μεταβαλλόμενες συνθήκες. Είναι συχνά πιο επεκτάσιμη, ιδίως όταν πρόκειται για μεγάλο αριθμό ετερογενών πηγών δεδομένων. Η συγχώνευση σε επίπεδο αποφάσεων επιτρέπει την εύκολη ενσωμάτωση νέων πηγών δεδομένων χωρίς να χρειάζεται να τροποποιηθεί ολόκληρη η διαδικασία της συγχώνευσης, καθώς επιτρέπει την ανεξάρτητη επεξεργασία πληροφοριών από διαφορετικές πηγές. Η κλιμάκωση με Early Fusion μπορεί να είναι πιο δύσκολη, καθώς κάθε νέα πηγή δεδομένων απαιτεί ενσωμάτωση στο αρχικό στάδιο, οδηγώντας ενδεχομένως σε πολύπλοκα και άκαμπτα συστήματα. Γι αυτό το λόγο, είναι πιθανόν να απαιτούνται τροποποιήσεις στη διαδικασία της σύντηξης, οδηγώντας σε ουσιαστικό επανασχεδιασμό. Ακόμη, απαιτείται μεγαλύτερη υπολογιστική ισχύ μιας και διαχειρίζεται ταυτόχρονα περισσότερο όγκο και είδη δεδομένων. Αντίθετα αυτό δεν συμβαίνει στην Late Fusion προσέγγιση, η οποία διαχειρίζεται μεμονωμένα τις περιπτώσεις. Αυτό οδηγεί πολλές φορές στο να συνιστάται για δυναμικά

περιβάλλοντα όπου τα χαρακτηριστικά των πηγών δεδομένων μπορεί να αλλάζουν με την πάροδο του χρόνου, αφού επιτρέπει εύκολες τροποποιήσεις και προσθήκες στη διαδικασία συγχώνευσης [81].

Κρίνεται σκόπιμο, λοιπόν, η πολυτροπική συγχώνευση να κατευθυνθεί σε τεχνικές συγχώνευσης σε επίπεδο αποφάσεων (Late Fusion) και να εξεταστούν οι βέλτιστες και πιο ταιριαστές λύσεις στο πρόβλημα.

7.3 Μεθοδολογία

Σε αυτή την ενότητα, παρουσιάζεται ένας αγωγός (pipeline) βαθιού νευρωνικού δικτύου για την πολυτροπική αναγνώριση συναισθημάτων και ιδίως για μια εργασία ταξινόμησης συναισθημάτων. Αυτή η τελική πρόταση συστήματος συνδυάζει τα μοντέλα με την υψηλότερη ακρίβεια των Κεφαλαίων 5 και 6. Όσον αφορά την διαδικασία εκπαίδευσης, τις υπερπαραμέτρους και τις λεπτομέρειες υλοποίησης των μοντέλων βλέπε τις Ενότητες 5.4 και 6.4. Η βασική ιδέα είναι να εφαρμοστούν και να συνδυαστούν τα διανύσματα προβλέψεων των πιο αποδοτικών μεθόδων, προκειμένου να επιλυθεί το πολυτροπικό έργο και να εκτελεστεί η σύντηξη σε επίπεδο απόφασης.



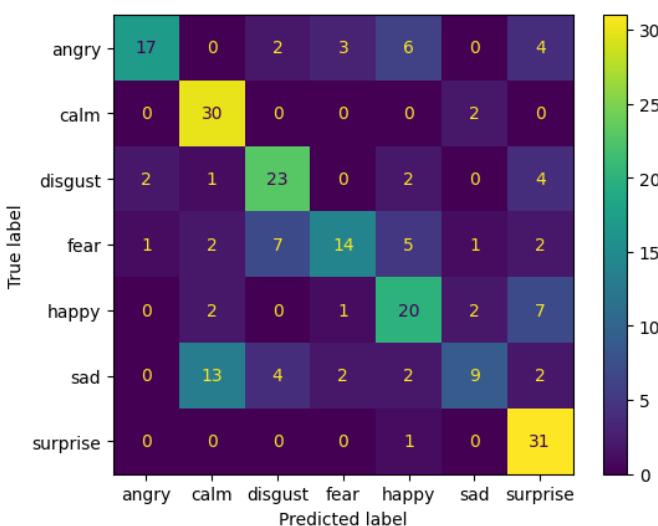
Σχήμα 7.2: Απεικόνιση της προτεινόμενης συγχώνευσης σε επίπεδο απόφασης, συνδυάζοντας ήχο και εικόνα για την εκτέλεση μιας πρόβλεψης.

(Πηγή Σχήματος: [82])

7.3.1 Επιλογή Αποδοτικότερων Μονοτροπικών Μοντέλων

Σκοπός του υποκεφαλαίου αποτελεί η εύρεση των μοντέλων με την υψηλότερη ακρίβεια σε αρχεία δοκιμής. Με την σειρά τους παράγουν τα διανύσματα πιθανοτήτων που τροφοδοτούνται στο Fusion σκέλος της αναγνώρισης ώστε να παρθεί η τελική απόφαση. Ειδικότερα, ως δεδομένα δοκιμής χρησιμοποιούνται τα αρχεία των ηθοποιών 21 μέχρι 24 και για τα δύο είδη μονοτροπικών μεθόδων. Τα αποτελέσματα που παρουσιάστηκαν στα δεδομένα επαλήθευσης και τα συμπεράσματα που προήλθαν μέσω αυτών είναι αναγκαίο να δοκιμαστούν και σε νέα δεδομένα έτσι ώστε να διασφαλιστεί η ακεραιότητα αυτών. Αυτό είναι κρίσιμο για την παραγωγή και την επιλογή ενός έμπιστου μοντέλου δίνοντας την ικανότητα μιας καλής γενίκευσης και πιο ασφαλών προβλέψεων σε εφαρμογές πέρα της βάσης δεδομένων που χρησιμοποιείται.

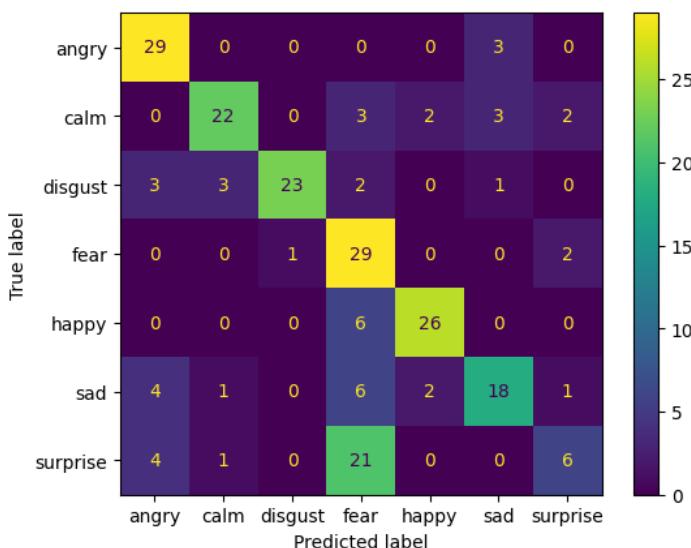
Για την εφαρμογή των δοκιμών είναι απαραίτητη η επανάληψη της προ-επεξεργασίας και των δύο μορφών αρχείων, όπως πραγματοποιήθηκαν στις υποενότητες 5.3 και 6.3 αντίστοιχα. Αρχικά στην περίπτωση του ήχου εφαρμόζονται τα δεδομένα δοκιμής στη CNN2D αρχιτεκτονική που φαίνεται στον πίνακα 5.3 με επαυξημένα ηχητικά δεδομένα εκπαίδευσης, η οποία διακρίνονταν με το υψηλότερο Validation Accuracy, όπως φαίνεται από τον Πίνακα 5.6. Το Test Accuracy που προκύπτει είναι της τάξεως του 64%. Το οποίο συμφωνεί με το Validation Accuracy και άρα το μοντέλο κρίνεται ως το πλέον κατάλληλο για την επιλογή του στο Fusion σκέλος της εργασίας.



Σχήμα 7.3: Confusion Matrix στη CNN2D αρχιτεκτονική με επαυξημένα δεδομένα εκπαίδευσης, για τα αρχεία δοκιμής ήχου

Από τον Confusion Matrix κατανοούται πως το μοντέλο ταξινομεί με παρόμοια επιτυχία τα συναισθήματα στα αρχεία δοκιμής, οπότε είναι ασφαλές το συμπέρασμα πως το μοντέλο θα συνεχίσει να παρουσιάζει παρόμοιες συμπεριφορές για νέα δεδομένα.

Στην περίπτωση των βίντεο εφαρμόζονται ξανά τα αρχεία δοκιμής στη CNN2D αρχιτεκτονική 6.1 για τα ασπρόμαυρα frames που είναι εστιασμένα στο πρόσωπο. Αυτή κατέχει το υψηλότερο Validation Accuracy, όπως φαίνεται και στον συνοπτικό Πίνακα 6.5. Προκύπτει Test Accuracy της τάξεως 57% ανά frame το οποίο έρχεται και σε αυτήν την περίπτωση σε συμφωνία με το Validation Accuracy. Παρόλα αυτά η τελική απόφαση ως προς τη δοκιμή πρέπει να αφορά ολόκληρο το βίντεο και όχι για κάθε frame του ξεχωριστά. Επομένως αφού υπολογιστεί ο μέσος όρος των τιμών των πιθανοτήτων για τα συναισθήματα που προκύπτουν από τη διαδικασία του μοντέλου για τα frames, καταλήγει σε ακρίβεια της τάξεως του 68% για ολόκληρα τα βίντεο. Είναι αξιοσημείωτη και λογική η αύξηση του accuracy, καθώς ο μέσος όρος των πιθανοτήτων των συναισθημάτων ανά frame για κάθε βίντεο οδηγεί στην τελική απόφαση να είναι ανεπηρέαστη από αποκλίνουσες τιμές και να επηρεάζεται από τα frames που είναι σε μεγαλύτερη συμφωνία μεταξύ τους.

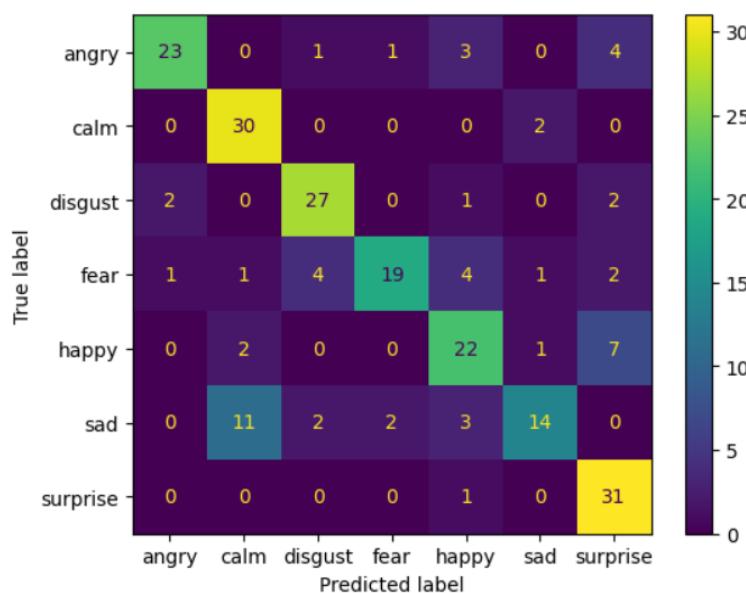


Σχήμα 7.4: Confusion Matrix στη CNN2D αρχιτεκτονική των ασπρόμαυρων εστιασμένων στο πρόσωπο frames για τα αρχεία δοκιμής.

Είναι αισθητή η αύξηση της ακρίβειας συνολικά για κάθε βίντεο και καταλήγει σε ένα αρκετά ικανοποιητικό επίπεδο, πολύ κοντά και με τα αντίστοιχα αποτελέσματα μέσω αρχείων ήχου.

7.4 Τεχνικές και Πειραματικά Αποτελέσματα

Η τεχνική του Late Fusion επιλέγεται ως προσέγγιση η οποία θα επιδιώξει να ενισχύσει την ακρίβεια της Πολυτροπικής Αναγνώρισης Συναίσθημάτος. Κατά το Late Fusion συνδυάζονται οι πιθανότητες του CNN2D με επαυξημένα ηχητικά δεδομένα εκπαίδευσης και του CNN2D των ασπρόμαυρων frames που εστιάζουν στο πρόσωπο. Με αυτό τον τρόπο συμβάλει αποτελεσματικά η πληροφορία τόσο του ήχου όσο και των εικόνων. Η εργασία όμως εμβαθύνει περαιτέρω στην επιρροή του συνδυασμού των πληροφοριών από το κλασικό Late Fusion με το μεταγενέστερο άθροισμα πιθανοτήτων (Posterior Probability Sum). Κατά το Posterior Probability Sum αθροίζονται οι πιθανότητες για κάθε συναίσθημα από τα δύο μονοτροπικά συστήματα και προκύπτει ακρίβεια της τάξεως του 74%.

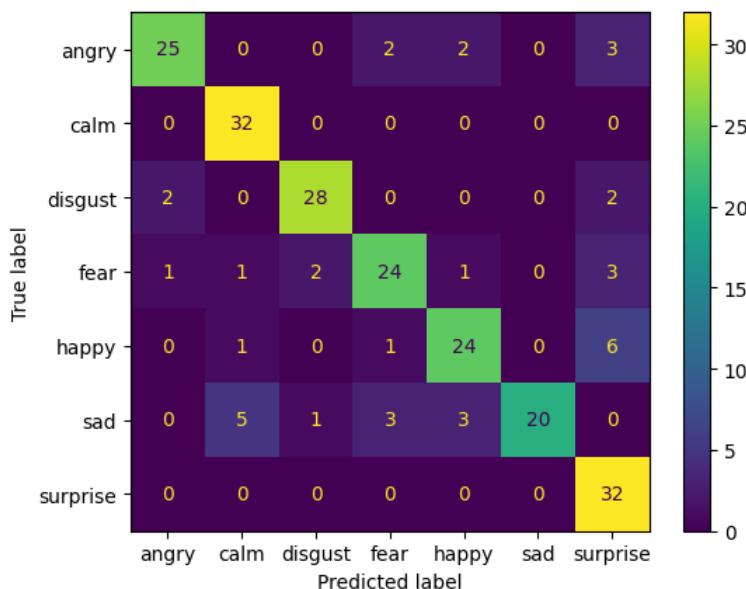


Σχήμα 7.5: Confusion Matrix για το Posterior Probability Sum

Σύμφωνα με το Confusion Matrix για το Posterior Probability Sum, παρατηρείται βελτίωση της πρόβλεψης σε όλο τον πίνακα 7.5. Αναμενόμενο εξαιτίας της αύξησης του accuracy στο 74%. Κυρίως εντοπίζεται βελτίωση στην κύρια διαγώνιο του confusion matrix πάνω στην οποία υπάρχει συμφωνία των προβλεπόμενων-πραγματικών συναίσθημάτων. Παρόλα αυτά εντοπίζεται η αδυναμία της σύγχυσης των συναίσθημάτων ηρεμίας και λύπης γεγονός που αντιμετωπίζουν και τα μονοτροπικά μοντέλα. Είναι αναγκαία λοιπόν η αναζήτηση βελτίωσης αυτού του προβλήματος. Για το σκοπό αυτό θα επεκταθούν οι τεχνικές του Late Fusion δίνοντας ισοδύναμη βαρύτητα στα διανύσματα πιθανοτήτων ήχου και εικόνας.

7.4.1 Geometric Mean Fusion

Σε αυτή την προσέγγιση, οι προβλέψεις τόσο από το μονοτροπικό μοντέλο ήχου όσο και από το μονοτροπικό μοντέλο εικόνων πολλαπλασιάζονται στοιχειωδώς για κάθε κλάση (συναισθήμα) και στη συνέχεια λαμβάνεται η τετραγωνική ρίζα των τιμών που προκύπτουν. Αυτή η διαδικασία εξασφαλίζει ότι η συγχωνευμένη πρόβλεψη διατηρεί τις πολλαπλασιαστικές σχέσεις μεταξύ των πιθανοτήτων, ενώ συνδυάζει τις πληροφορίες και από τις δύο λειτουργίες. Η τετραγωνική ρίζα χρησιμοποιείται για την αναγωγή των τιμών, καθώς το γινόμενο των πιθανοτήτων μπορεί να γίνει αρκετά μεγάλο. Η σύντηξη του Geometric Mean χρησιμοποιείται συχνά όταν πρόκειται για πιθανότητες για τη δημιουργία μιας ισορροπημένης σύντηξης που λαμβάνει υπόψη τη δύναμη των προβλέψεων και από τις δύο λειτουργίες. Η ακρίβεια αυτής της μεθόδου είναι της τάξεως του 82%.

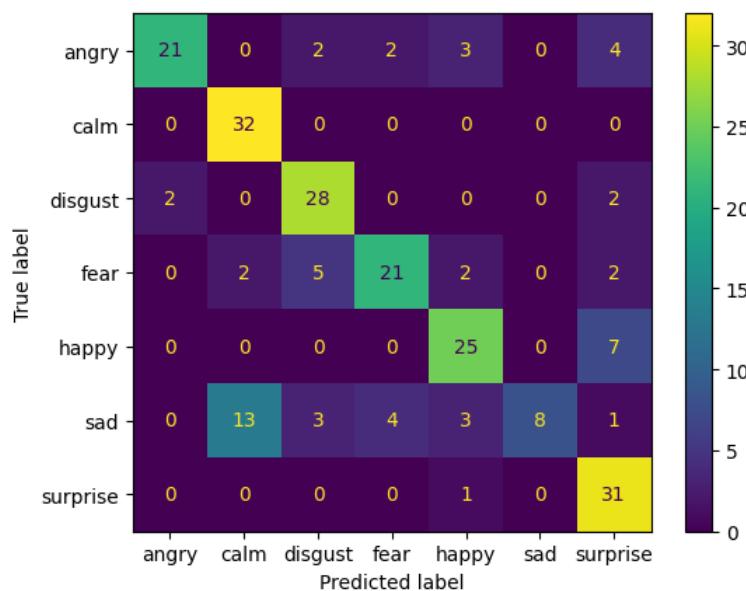


Σχήμα 7.6: Confusion Matrix για το Geometric mean fusion.

Με αυτή τη τεχνική σημειώνεται σημαντική βελτίωση στις προβλέψεις του μοντέλου σε σχέση με της προηγούμενης, καθώς οι τιμές του confusion matrix 7.6 στην διαγώνιο έχουν αυξηθεί έντονα και εντοπίζεται μεγάλη μείωση στην σύγχυση των συναισθημάτων της χαράς και της ηρεμίας. Ιδιαίτερα δεν μπορεί να αναφερθεί κάποια προβληματική σύγχυση συναισθημάτων μέσω του Confusion Matrix. Δεδομένου του αρκετά υψηλού ποσοστού ακρίβειας, η Geometric Mean Fusion καθίσταται ως μια πολύ ισχυρή τεχνική πολυτροπικής αναγνώρισης.

7.4.2 F1-score

Σε αυτή την προσέγγιση πολυτροπικής αναγνώρισης συναισθημάτων, τα αποτελέσματα F1 υπολογίζονται αρχικά για κάθε ετικέτα από τα τελικά μονοτροπικά μοντέλα κάθε είδους. Στη συνέχεια, τα αποτελέσματα F1 κανονικοποιούνται ώστε το άθροισμά τους να είναι 1, μετατρέποντάς τα σε βάρη που αντικατοπτρίζουν την απόδοση κάθε τρόπου (ήχου και βίντεο) για κάθε κατηγορία συναισθήματος. Αυτές οι κανονικοποιημένες F1-Scores, εφαρμόζονται στη συνέχεια ως βάρη στις αντίστοιχες προβλεπόμενες πιθανότητες για κάθε ετικέτα (συναίσθημα) στον ήχο όσο και στο βίντεο. Το τελικό βήμα περιλαμβάνει το συνδυασμό αυτών των σταθμισμένων πιθανοτήτων μέσω του Geometric Mean που εμφάνισε βελτίωση στην απόδοση. Ως αποτέλεσμα λαμβάνεται μια συγχωνευμένη πρόβλεψη, η οποία εξαρτάται από τις επιδόσεις των επιμέρους τρόπων. Ουσιαστικά, η μέθοδος αυτή διασφαλίζει ότι η συμβολή του ήχου και των εικόνων στην τελική πρόβλεψη σταθμίζεται με βάση την ιστορική του απόδοση για τις συγκεκριμένες συναισθηματικές κατηγορίες. Έτσι, παρέχεται μια ισορροπημένη και τεκμηριωμένη συγχώνευση πληροφοριών από πηγές ήχου και εικόνων. Τέλος, επιτυγχάνεται εκ' νέου μια απόδοση ίδιας επιτυχίας με αυτής του αρχικού posterior probability sum, δηλαδή 74%.



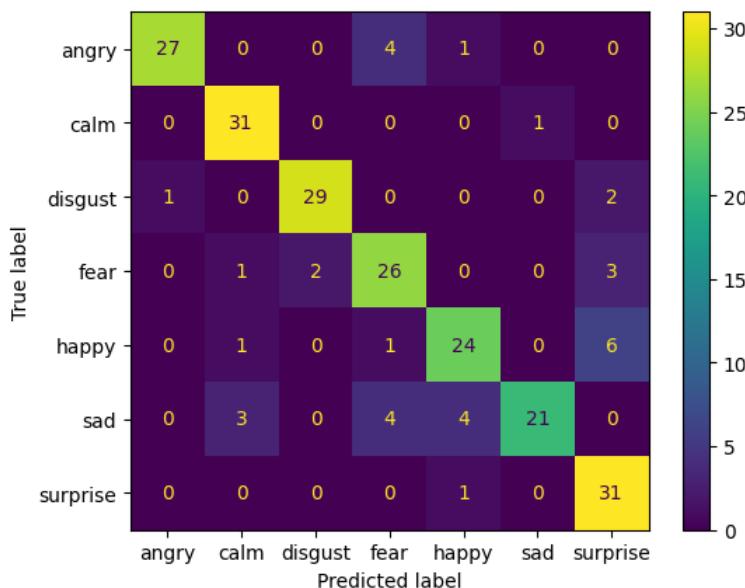
Σχήμα 7.7: Confusion Matrix για το f1 score fusion.

Με την συμβολή του F1-Score αντανακλάται η συμπεριφορά του Posterior Probability Sum, με το πρόβλημα της σύγχυσης των συναισθημάτων λύπης και ηρεμίας. Αυτό καθιστά αυτή τη τεχνική μη συγκρίσιμη με αυτή του Geometric Mean Fusion και απορρίπτεται ως μια αδύ-

ναμη επιλογή για τη τελική πολυτροπική προσέγγιση.

7.4.3 Attention Mechanism

Σε αυτό το πλαίσιο Late fusion εφαρμόζεται Μηχανισμός Προσοχής (Attention Mechanism), με τα βάρη προσοχής να υπολογίζονται για κάθε περίπτωση ή χρονικό βήμα τόσο στον ήχο όσο και στο βίντεο. Τα βάρη προσοχής προκύπτουν από τον μέσο όρο των προβλέψεων σε όλες τις περιπτώσεις, παρέχοντας ένα μέτρο σημαντικότητας για κάθε χρονικό τμήμα. Για να εξασφαλιστεί η ουσιαστική στάθμιση, εφαρμόζεται μια συνάρτηση softmax στα βάρη προσοχής, ομαλοποιώντας τα σε μια κατανομή πιθανότητας. Το επόμενο βήμα περιλαμβάνει τον στοιχειομετρικό πολλαπλασιασμό των προβλέψεων, με τα σταθμισμένα βάρη προσοχής προβλέψεων τόσο από τον ήχο όσο και από το βίντεο. Ως αποτέλεσμα, λαμβάνεται μια σύνθετη πιθανότητα για κάθε είδος. Τέλος η πρόβλεψη γίνεται μέσω του Geometric Mean που αποτελεί τον πιο αποδοτικό τρόπο συγχώνευσης των πιθανοτήτων. Ο Attention Mechanism λοιπόν παρέχει σε συνδυασμό με το Geometric Mean ακόμη μεγαλύτερη ακρίβεια από ότι παρουσίασε μόνο του. Έτσι παρατηρείται επιτυχία της τάξεως του 84% και συνιστά την τελική επιλογή για τους μηχανισμούς Fusion της πολυτροπικής αναγνώρισης.



Σχήμα 7.8: Confusion Matrix για το Attetion Mechanism Fusion.

Το συμπέρασμα αντικατροπτίζεται και από την εικόνα 7.8 που παρέχει το τελικό Confusion Matrix. Παρατηρείται η πιο ιδανική συμπεριφορά κατά την αντιστοίχηση των προβλεπόμε-

νων και πραγματικών συναισθημάτων, σε συνδυασμό με τον πλέον ικανοποιητικό βαθμό ακριβείας. Επιπλέον δεν εμφανίζει εμφανή σύγχυση μεταξύ συγκεκριμένων συναισθημάτων αλλά μόνο τυχαίες αστοχίες οι οποίες δεν ακολουθούν κάποιο μοτίβο και όπως αναμένεται δεν μπορούν να αποφευχθούν.

7.5 Σύνοψη

Model	Accuracy
Posterior Probability Sum Fusion	0.74
Geometric Mean Fusion	0.82
F1-score Fusion	0.74
Attention Mechanism Fusion	0.84

Πίνακας 7.1: Σύνοψη των fusion μοντέλων.

Καταγράφεται η συνολική εικόνα για όλες τις τεχνικές Late Fusion που εφαρμόζονται στον πίνακα 7.1, μαζί με τα ποσοστό accuracy που παράγουν. Συνεπώς εύκολα παρατηρείται η μεθοδική και σταδιακή βελτίωση του μοντέλου, μέχρι την επίτευξη της μέγιστης δυνατής ακρίβειας για τα δεδομένα δοκιμής. Αυτή επιτυγχάνεται με την τεχνική του Attention Mechanism Fusion ακρίβειας 84% και η οποία έχει την δυνατότητα να ταξινομεί με εξίσου καλή απόδοση όλα τα συναισθήματα.

Συμπερασματικά, η επιλογή του Multimodal Emotion Recognition με Attention Mechanism συνιστά την πιο ισχυρή τεχνική καθώς αξιοποιεί την πολύ καλή συγχώνευση των πιθανοτήτων που παρέχει ο Geometric Mean. Επίσης αποτελεί μια στρατηγική επιλογή που όχι μόνο ενισχύει την ακρίβεια της αναγνώρισης συναισθημάτων, αλλά παρέχει επίσης μια λεπτομερή κατανόηση των λεπτών αποχρώσεων εντός των διαφόρων τρόπων. Προχωρώντας ένα βήμα παραπέρα, δημιουργείται μια εφαρμογή που χρησιμεύει ως παρουσίαση των ερευνητικών ευρημάτων. Αυτή η εφαρμογή στοχεύει στους χρήστες, προσφέροντάς τους μια άμεση διεπαφή για να αντιλαμβάνονται και να ερμηνεύονται τα αποτελέσματα που προκύπτουν από τις περίπλοκες διαδικασίες που χρησιμοποιούνται. Μέσω του φιλικού προς το χρήστη σχεδιασμού, η εφαρμογή γίνεται γέφυρα μεταξύ της έρευνας και των πραγματικών εφαρμογών,

προιωθώντας μια βαθύτερη σύνδεση μεταξύ της τεχνολογίας και της ανθρώπινης κατανόησης.

Κεφάλαιο 8

Εφαρμογή Πολυτροπικής Αναγνώρισης Συναισθημάτων

8.1 Εισαγωγή

Ως επιστέγασμα της εκτενούς έρευνας που διεξάγεται για την πολυτροπική αναγνώριση συναισθημάτων, μια πρακτική επίδειξη που συνοψίζει τα βήματα και τα αποτελέσματα αυτής της εργασίας παρουσιάζεται σε αυτό το κεφάλαιο. Η εφαρμογή βασίζεται σε επτά βίντεο, τα οποία αντιστοιχούν σε κάθε ένα από τα επτά συναισθήματα. Ως απόδειξη της ευρωστίας και της εφαρμοσιμότητας του προτεινόμενου πλαισίου, έχει αναπτυχθεί μια πλήρως λειτουργική εφαρμογή. Αυτή χρησιμεύει ως επίδειξη των δυνατοτήτων της προτεινόμενης προσέγγισης στην αναγνώριση συναισθημάτων από εικόνες και ήχο, προσφέροντας μια πρακτική εμπειρία για την επικύρωση της αποτελεσματικότητάς του. Η συγχώνευση αυτών των τρόπων στοχεύει να ενισχύσει την ακρίβεια και το βάθος της αναγνώρισης συναισθημάτων. Μέσω αυτής της διαδραστικής επίδειξης, ο αντίκτυπος της πολυτροπικής αναγνώρισης συναισθημάτων γίνεται αισθητός, παρέχοντας μια απτή παρουσίαση των βημάτων που ακολουθούνται κατά την επίτευξη της πολυτροπικής πρόβλεψης σε σύγκριση με το πραγματικό συναίσθημα του ηθοποιού που επιλέγεται.

8.2 Παρουσίαση Εφαρμογής

Η γραφική διεπαφή χρήστη που παρουσιάζεται σε αυτό το κεφάλαιο αξιοποιεί τη βιβλιοθήκη Tkinter στην Python για να παρουσιάσει την πρακτική εφαρμογή του πολυτροπικού

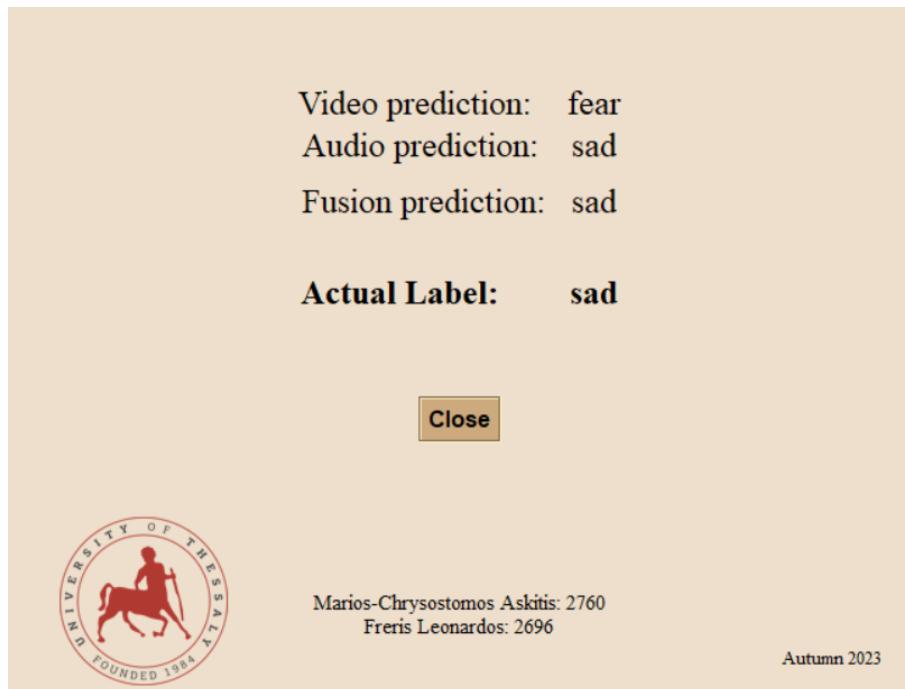
μοντέλου αναγνώρισης συναισθημάτων. Το κύριο παράθυρο διαθέτει τίτλο “App For Multi-modal Emotion Recognition”, υπότιτλο “Demonstration of fusion model (video + audio) for each of 7 emotions” και μια φιλική προς το χρήστη διεπαφή, που επιτρέπει σε αυτούς να εισάγουν μια αριθμητική τιμή που αντιστοιχεί σε ένα από τα επτά συναισθήματα. Οι αριθμητικές τιμές διακρίνονται σε 0,1,2,3,4,5,6 και αντιπροσωπεύουν την ηρεμία, την χαρά, την λύπη, τον θυμό, τον φόβο, την αηδία και την έκπληξη αντίστοιχα. Οι ετικέτες δηλαδή παρέχουν καθοδήγηση για τη χαρτογράφηση συναισθημάτων-αριθμών. Το κουμπί “Select Emotion” ενεργοποιεί την εμφάνιση του επιλεγμένου βίντεο συναισθήματος. Το αρχικό παράθυρο λοιπόν εμφανίζεται ως:



Σχήμα 8.1: Αρχικό παράθυρο εφαρμογής.

Κατόπιν ακολουθείται η αλγορίθμική διαδικασία για την πρόβλεψη τόσο των μονοτροπικών μεθόδων ξεχωριστά όσο και του πολυτροπικού πλαισίου που υποδεικνύεται στο Κεφάλαιο 7. Ειδικότερα για την ανάλυση του βίντεο πραγματοποιείται εξαγωγή των frames για την παραγωγή ασπρόμαυρων στιγμιοτύπων μεγέθους 112x112. Στην συνέχεια απομονώνεται ο ήχος από το βίντεο του συναισθήματος που επιλέγεται και τροποποιείται στο κατάλληλο μήκος για να ικανοποιεί τις απαιτήσεις του μοντέλου με σκοπό να εξαχθούν τα Mels-spectrogram. Οι αναλύσεις αυτές εισάγονται με την σειρά τους στα μονοτροπικά μοντέλα που αποδείχθηκαν πιο αποτελεσματικά στα Κεφάλαια 5.3 και 6.3, παράγοντας έτσι τα διανύσματα προβλέ-

ψεων. Επιπλέον αυτά συμβάλλουν στο πολυτροπικό μοντέλο, αφού συνδυάζονται σύμφωνα με το Attention Mechanism Fusion που επιτυγχάνει την υψηλότερη ακρίβεια όπως φαίνεται στον Πίνακα 7.1. Τέλος εμφανίζεται το παράθυρο αποτελεσμάτων το οποίο περιέχει τέσσερις βασικές πληροφορίες: τα προβλεπόμενα συναισθήματα από το οπτικό μοντέλο, το ηχητικό μοντέλο, το μοντέλο σύντηξης (Fusion) και την πραγματική ετικέτα που λαμβάνεται από το συναίσθημα του βίντεο. Κάθε πρόβλεψη παρουσιάζεται με σαφή και ευανάγνωστη μορφή, ενώ η «Πραγματική ετικέτα» επισημαίνεται με έντονη γραφή. Ένα κουμπί “Close” είναι ενσωματωμένο για διευκόλυνση του χρήστη, επιτρέποντας τον τερματισμό της εμφάνισης των αποτελεσμάτων. Το παράθυρο που εμφανίζεται λοιπόν κατά την ολοκλήρωση της περιήγησης του χρήστη στην εφαρμογή είναι:



Σχήμα 8.2: Παράθυρο αποτελεσμάτων εφαρμογής.

Κεφάλαιο 9

Συμπεράσματα

9.1 Σύνοψη

Η αναγνώριση συναισθημάτων αποτελεί ένα πολυσυζητημένο τομέα της τεχνολογίας και δη η πολυτροπική αναγνώριση συναισθημάτων έχει αποκτήσει σημαντική δημοφιλία τα τελευταία χρόνια. Σε αυτή την εργασία προτάσσεται μια ποικιλία μοντέλων Βαθιάς Μάθησης με διαφορετικές προσεγγίσεις, χρησιμοποιώντας αρχικά μονοτροπικές μεθόδους και κατ' επέκταση πολυτροπικές. Οι μονοτροπικές προσεγγίσεις επικεντρώθηκαν στην CNN αρχιτεκτονική όπως Transfer Learning (Resnet50, Xception), CNN1D-LSTM και CNN2D. Η μεγαλύτερη ακρίβεια επιτεύχθηκε εφαρμόζοντας την CNN2D μαζί με τη σωστή ρύθμιση των υπερπαραμέτρων (Dropout, Maxpooling, Kernel_initializer, Batchnormalize). Συγκεκριμένα, τα πειραματικά αποτελέσματα υποδεικνύουν ενδιαφέροντα ευρήματα σχετικά με την απόδοση όλων των μοντέλων.

Όσον αφορά τη αναγνώριση συναισθήματος από ήχο, ο πίνακας 5.6 των αποτελεσμάτων αξιολόγησης παρουσιάζει μια σύγκριση πολλών διαφορετικών μοντέλων νευρωνικών δικτύων που υλοποιήθηκαν. Συγκεκριμένα, το εκπαιδευμένο με επαυξημένα ηχητικά δεδομένα CNN2D μοντέλο του πίνακα 5.3 αποτελεί το πιο αποδοτικό με ακρίβεια 64%.

Σχετικά με την αναγνώριση συναισθήματος από εικόνες, ο αντίστοιχος πίνακας 6.5 προβάλει τη CNN2D αρχιτεκτονική του πίνακα 6.1 για ασπρόμαυρα frames εστιασμένα στο πρόσωπο ως το πιο αποδοτικό με ακρίβεια 56%.

Έπειτα, η εργασία επικεντρώνεται στην αναγνώριση συναισθημάτων με την σύντηξη των δύο προηγούμενων τεχνικών για την επίτευξη ακόμη υψηλότερου ποσοστού επιτυχίας. Ειδικότερα δημιουργήθηκαν Late Fusion μοντέλα που συνδύασαν τα αποτελέσματα πιθανοτήτων των δύο μονοτροπικών προσεγγίσεων. Τεχνικές όπως Posterior Probability Sum Fusion, Geometric Mean Sum, F1-score Fusion και Attention Mechanism Fusion καταλήγουν σε αξησημείωτες υψηλές επιδόσεις για το σύνολο δεδομένων RAVDESS. Είναι σημαντικό να αναφερθεί πως τα Geometric Mean Sum και F1-score Fusion παρουσίασαν παρόμοια συμπεριφορά, με ποσοστά επιτυχίας της τάξης του 80%. Ξεχωρίζει, όμως, το μοντέλο με Μηχανισμό Προσοχής. Αποδεικνύει πως είναι μια ισχυρή τεχνική, που οδηγεί σε μεγάλα κέρδη στην απόδοση των νευρωνικών δικτύων έως και 3% όσον αφορά την ακρίβεια, σε σύγκριση με τα υπόλοιπα.

9.2 Μελλοντικές επεκτάσεις

Υπάρχουν διάφορες κατευθύνσεις που θα μπορούσαν να εξεταστούν σε μελλοντικές εργασίες, σε μια προσπάθεια να συμπληρωθούν υπάρχουσες ανάγκες και κενά. Σε μελλοντικές εργασίες μπορούν να επεκταθούν τα πειράματα με περισσότερο όγκο δεδομένων και με μεγαλύτερη ποικιλία βάσεων δεδομένων.

Συγκεκριμένα οι εργασίες αναγνώρισης συναισθημάτων αποτελούν πρόκληση, καθώς το συναίσθημα είναι εξαιρετικά προσωπικό αλλά προσωπικά δεδομένα είναι δύσκολο να ληφθούν. Οι ερευνητές εργάζονται για την εξισορρόπηση της εξατομίκευσης και της γενίκευσης των συναισθημάτων. Κάποιοι προτείνουν τη χρήση δημογραφικών στοιχείων ατόμων όπως π.χ φύλο, ηλικία, επάγγελμα και στη συνέχεια με χρήση μεθόδων Μεταφοράς Μάθησης (Transfer Learning). Πιθανώς η Μεταφορά Μάθησης να φανεί πιο αποδοτική, σε αντίθεση με τη παρούσα εργασία που παρουσιάζεται ως μη αποδοτική. Επιπλέον έχει αποδειχθεί ότι τα κριτήρια που χρησιμοποιούνται για τον καθορισμό των πτυχών της διασταυρούμενης-επικύρωσης (cross-validation) έχουν μεγάλη επίδραση στα αποτελέσματα, αφού επί του παρόντος τα δεδομένα επικύρωσης επιλέχθηκαν κατά τύχη. Τέλος, είναι σημαντική η εστίαση στη δημιουργία ερμηνεύσιμων μοντέλων βαθιάς μάθησης. Χρειάζεται μια καλύτερη κατανόηση της υποκείμενης συμπεριφοράς και δυναμικής αυτών των μοντέλων, η οποία μπορεί με τη σειρά της να οδηγήσει στην ανάπτυξη ακόμη καλύτερων και ισχυρότερων μοντέλων.

Βιβλιογραφία

- [1] M Sreeshakthy and J Preethi. Classification of emotion from EEG using hybrid radial basis function networks with elitist PSO. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 6(3-4):60–73, 2016.
- [2] Min Chen, Yin Zhang, Meikang Qiu, Nadra Guizani, and Yixue Hao. SPHA: Smart personal health advisor based on deep analytics. *IEEE Communications Magazine*, 56(3):164–169, 2018.
- [3] Faiyaz Doctor, Charalampos Karyotis, Rahat Iqbal, and Anne James. An intelligent framework for emotion aware e-healthcare support systems. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE, 2016.
- [4] Kai Lin, Fuzhen Xia, Wenjian Wang, Dixin Tian, and Jeungeun Song. System design for big data application in emotion-aware healthcare. *IEEE Access*, 4:6901–6909, 2016.
- [5] Paul Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.
- [6] Maja Pantic, Nicu Sebe, Jeffrey F Cohn, and Thomas Huang. Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676, 2005.
- [7] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2):211–223, 2011.
- [8] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1359–1367, 2020.

- [9] Wen Wu. Multimodal emotion recognition. *University of Cambridge*, 2020.
- [10] Lukasz Augustyniak, Tomasz Kajdanowicz, Piotr Szymański, Włodzimierz Tuliglowicz, Przemysław Kazienko, Reda Alhajj, and Bolesław Szymański. Simpler is better? Lexicon-based ensemble sentiment classification beats supervised methods. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 924–929. IEEE, 2014.
- [11] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. In *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49. Springer, 2008.
- [12] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).*, 9(1):381–386, 2020.
- [13] Muhammad Usama, Junaid Qadir, Aunn Raza, Hunain Arif, Kok-Lim Alvin Yau, Yehia Elkhatib, Amir Hussain, and Ala Al-Fuqaha. Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, 7:65579–65615, 2019.
- [14] Osvaldo Simeone. A very brief introduction to machine learning with applications to communication systems. *IEEE Transactions on Cognitive Communications and Networking*, 4(4):648–664, 2018.
- [15] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.
- [16] <https://siddhithakkar.com/2020/02/24/ai-vs-ml-vs-dl-whats-the-difference/>.
- [17] <https://machinelearningmastery.com/what-is-deep-learning/>.
- [18] <https://towardsdatascience.com/cnn-application-on-structured-data-automated-feature-extraction-8f2cd28d9a7e>.
- [19] Dan Ciresan, Alessandro Giusti, Luca Gambardella, and Jürgen Schmidhuber. Deep neural networks segment neuronal membranes in electron microscopy images. *Advances in neural information processing systems*, 25, 2012.

- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [21] Vinay Williams, Vasileios Argyriou, Peter Shaw, Christop Montag, Georg Herdrich, Aaron Knoll, and Maximilian Moertl. Development of PPTNet a neural network for the rapid prototyping of pulsed plasma thrusters. In *Proceedings of the 36th International Electric Propulsion Conference*. Electric Rocket Propulsion Soc., 2019.
- [22] Yoshua Bengio et al. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.
- [23] Adrian Carrio, Carlos Sampedro, Alejandro Rodriguez-Ramos, Pascual Campoy, et al. A review of deep learning methods and applications for unmanned aerial vehicles. *Journal of Sensors*, 2017, 2017.
- [24] <https://www.upgrad.com/blog/basic-cnn-architecture/>.
- [25] Arohan Ajit, Koustav Acharya, and Abhishek Samanta. A review of convolutional neural networks. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–5, 2020.
- [26] <https://medium.com/machine-learning-bites/deeplearning-series-convolutional-neural-networks-a9c2f2ee1524>.
- [27] <https://www.geeksforgeeks.org/apply-a-2d-max-pooling-in-pytorch/>.
- [28] Charalampos Tzamos. *VentusNet: Deep Learning for Wind Speed prediction*. PhD thesis, Dept Informatics & Telecoms, University of Athens, 2019.
- [29] <https://blog.mlreview.com/understanding-lstm-and-its-diagrams-37e2f46f1714>.
- [30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [31] <https://medium.com/analytics-vidhya/activation-function-c762b22fd4da>.

- [32] <https://pytorch.org/docs/stable/generated/torch.nn.LeakyReLU.html>.
- [33] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316, 2017.
- [34] Sandhya Tripathi and N Hemachandra. Scalable linear classifiers based on exponential loss function. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 190–200, 2018.
- [35] <https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e>.
- [36] <https://mohitmishra786687.medium.com/the-learning-rate-a-hyperparameter-that-matters-b2f3b68324ab>.
- [37] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010: 19th International Conference on Computational Statistics Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.
- [38] <https://nvnhcus.medium.com/asynchronous-stochastic-gradient-descent-with-diminishing-stepsize-dcffda3e24e5>.
- [39] Mahmoud Said ElSayed, Nhien-An Le-Khac, Marwan Ali Albahar, and Anca Jurcut. A novel hybrid model for intrusion detection systems in SDNs based on CNN and a new regularization technique. *Journal of Network and Computer Applications*, 191:103160, 2021.
- [40] <https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-deep-learning-optimizers/>.
- [41] <https://towardsdatascience.com/adam-optimization-algorithm-1cdc9b12724a>.
- [42] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

- [43] Amine Ben Khalifa and Hichem Frigui. Multiple instance fuzzy inference neural networks. *arXiv preprint arXiv:1610.04973*, 2016.
- [44] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [45] Frank Hutter, Jörg Lücke, and Lars Schmidt-Thieme. Beyond manual tuning of hyperparameters. *KI-Künstliche Intelligenz*, 29:329–337, 2015.
- [46] Lokesh Bejjagam and Reshma Chakradhara. Facial Emotion Recognition using Convolutional Neural Network with Multiclass Classification and Bayesian Optimization for Hyper Parameter Tuning. *DVGDT Bachelor Qualification Plan in Computer Science*, 2022.
- [47] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.
- [48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [50] Gewen He, Xiaofeng Liu, Fangfang Fan, and Jane You. Classification-aware semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 964–965, 2020.
- [51] Christos Athanasiadis, Enrique Hortal, and Stylianos Asteriadis. Audio–visual domain adaptation using conditional semi-supervised generative adversarial networks. *Neurocomputing*, 397:331–344, 2020.
- [52] Wootaeck Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, pages 1–4. IEEE, 2016.

- [53] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE, 2016.
- [54] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 467–474, 2015.
- [55] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 3687–3691. IEEE, 2013.
- [56] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. EmoNets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10:99–111, 2016.
- [57] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Multimodal deep convolutional neural network for audio-visual emotion recognition. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 281–284, 2016.
- [58] Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, and Björn Schuller. Strength modelling for real-world automatic continuous affect recognition from audio-visual signals. *Image and Vision Computing*, 65:76–86, 2017.
- [59] Cristina Luna-Jiménez, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan M Montero, and Fernando Fernández-Martínez. Multimodal emotion recognition on RAVDESS dataset using transfer learning. *Sensors*, 21(22):7665, 2021.
- [60] <https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0196391.t001>.

- [61] Steven R Livingstone and Frank A Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS one*, 13(5):e0196391, 2018.
- [62] <https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0196391.g003>.
- [63] Thapanee Seehapoch and Sartra Wongthanavasu. Speech emotion recognition using support vector machines. In *2013 5th international conference on Knowledge and smart technology (KST)*, pages 86–91. IEEE, 2013.
- [64] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE, 2017.
- [65] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.
- [66] https://en.wikipedia.org/wiki/Mel_scale.
- [67] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al. Speaker identification using Mel frequency cepstral coefficients. *Variations*, 1(4):565–568, 2004.
- [68] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. Audio and music signal analysis in Python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.
- [69] <https://librosa.org/doc/main/generated/librosa.feature.mfcc.html>.
- [70] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.
- [71] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control*, 47:312–323, 2019.

- [72] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. OpenFace: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*, 6(2):20, 2016.
- [73] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57:137–154, 2004.
- [74] <https://www.analyticsvidhya.com/blog/2022/04/object-detection-using-haar-cascade-opencv/>.
- [75] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [76] <https://medium.com/@siddheshb008/resnet-architecture-explained-47309ea9283d>.
- [77] Chollet Fran et al. Deep learning with depth wise separable convolutions. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.
- [78] <https://maelfabien.github.io/deeplearning/xception/>.
- [79] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.
- [80] Min Peng, Chongyang Wang, Tong Chen, Guangyuan Liu, and Xiaolan Fu. Dual temporal scale Convolutional Neural Network for micro-expression recognition. *Frontiers in psychology*, 8:1745, 2017.
- [81] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*, 37:98–125, 2017.
- [82] Hongyi Liu, Tongtong Fang, Tianyu Zhou, and Lihui Wang. Towards robust human-robot collaborative manufacturing: Multimodal Fusion. *IEEE Access*, 6:74762–74771, 2018.