Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Αναγνώριση Μουσικού Είδους και Υποείδους με Τεχνικές Μηχανικής Μάθησης

Διπλωματικη Εργασια

της

ΛΕΜΠΕΣΗ Σ. ΠΑΡΑΣΚΕΥΗΣ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Αναγνώριση Μουσικού Είδους και Υποείδους με Τεχνικές Μηχανικής Μάθησης

Διπλωματικη Εργασια

της

ΛΕΜΠΕΣΗ Σ. ΠΑΡΑΣΚΕΥΗΣ

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 8η Ιουλίου 2021.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης	Γεώργιος Στάμου	Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

.....
Παρασκευή Λεμπέση

3η Ιουλίου 2021

Περίληψη

Η μουσική ήταν ανέκαθεν ένα αναπόσπαστο τμήμα του πολιτισμού του ανθρώπου, καθώς είναι άρρηκτα συνδεδεμένη με την έκφραση των συναισθημάτων, τη διασκέδαση και την κοινωνική ζωή. Με το πέρασμα του χρόνου και ανάλογα με τα κοινωνικά φαινόμενα, τη μόδα και τις καλλιτεχνικές επιρροές της κάθε εποχής, δημιουργούνται νέα μουσικά είδη, ενώ άλλα παλαιότερα σταδιακά εκλείπουν. Η ψηφιοποίηση της μουσικής βιομηχανίας έχει επιφέρει ριζικές αλλαγές στον τρόπο με τον οποίο η μουσική δημιουργείται, αναπαράγεται και διαδίδεται. Πολλές διαδικασίες έχουν απλοποιηθεί και αυτοματοποιηθεί, αλλά ταυτόχρονα έχουν προκύψει νέες ανάγκες και προκλήσεις. Μία από αυτές τις προκλήσεις είναι το ζήτημα της αυτόματης αναγνώρισης του μουσικού είδους.

Η ταξινόμηση μουσικής σε κατηγορίες, είτε ανά μουσικό είδος, είτε ανά καλλιτέχνη, είτε ανά διάθεση είναι ένα ενεργό πρόβλημα στον τομέα της Ανάκτησης Μουσικής Πληροφορίας (Music Information Retrieval - MIR). Πιο συγκεκριμένα, η ταξινόμηση ανά μουσικό είδος είναι ένα πεδίο το οποίο τα τελευταία χρόνια έχει γνωρίσει μεγάλη ανάπτυξη, καθώς έχουν πραγματοποιηθεί πολλές έρευνες με αξιοσημείωτα αποτελέσματα. Ταυτόχρονα, στο πεδίο της ταξινόμησης ανά μουσικό υποείδος, η έρευνα και οι προσπάθειες βελτίωσης συνεχίζονται ενεργά, καθώς η αυτόματη διάκριση μεταξύ πολύ παρόμοιων κατηγοριών είναι μία ακόμα μεγαλύτερη πρόκληση.

Το θέμα της παρούσας διπλωματικής εργασίας είναι αρχικά η δημιουργία ενός συστήματος ταξινόμησης μουσικής στα 10 βασικά είδη (hiphop, country, disco, metal, reggae, blues, rock, classical, jazz, pop). Στη συνέχεια γίνεται μεγαλύτερη εμβάθυνση, μέσω της υλοποίησης ενός συστήματος ταξινόμησης της ροκ μουσικής στα 10 υποείδη της (punk, post-rock, lo-fi, metal, psych-rock, indie-rock, industrial, garage, new wave, progressive). Και στις δύο αυτές περιπτώσεις, το θέμα προσεγγίζεται με τεχνικές Βαθιάς Μάθησης. Αρχικά, γίνεται η εξαγωγή χρήσιμων χαρακτηριστικών με μεθόδους Ψηφιακής Επεξεργασίας Σήματος. Στη συνέχεια, τα χαρακτηριστικά αυτά τροφοδοτούνται σε διάφορες αρχιτεκτονικές Νευρωνικών Δικτύων. Τέλος, πραγματοποιείται αξιολόγηση όλων των πειραμάτων και επιλογή των βέλτιστων χαρακτηριστικών και αρχιτεκτονικών.

Λέξεις Κλειδιά

Ταξινόμηση σε μουσικό είδος, Ταξινόμηση σε μουσικό υποείδος, Ροκ μουσική, Ψηφιακή Επεξεργασία Σήματος, Μηχανική Μάθηση, Βαθιά Μάθηση, Ανάκτηση Μουσικής Πληροφορίας, MIR

Abstract

Music has always been an integral part of human civilisation, as it is directly related to the expression of emotions, entertainment and social life. Over time, depending on the social phenomena, fashion and artistic influences of each era, new music genres are emerging, while older ones are gradually disappearing. The digitization of the music industry has brought about radical changes in the way with which music is created, reproduced and distributed. Many processes have been simplified and automated, but at the same time new needs and challenges have arisen. One of these challenges is the issue of automatic recognition of the musical genre.

Classifying music into categories, either by genre, artist, or mood is an ongoing problem in the field of Music Information Retrieval - MIR. More specifically, Music Genre Classification is an area which in the recent years has experienced great development, as many studies have been carried out with notable results. At the same time, in the area of Music Sub-genre Classification, research and improvement efforts actively continue, as the automatic distinction between very similar categories is considered an even bigger challenge.

The subject of this diploma thesis is, in a first stage, to create a system classifying music into the 10 basic genres (hiphop, country, disco, metal, reggae, blues, rock, classical, jazz, pop). Then we dive deeper, through the implementation of a system classifying rock music in its 10 sub-genres (punk, post-rock, lo-fi, metal, psych-rock, indie-rock, industrial, garage, new wave, progressive). In both cases, the topic is approached with Deep Learning techniques. At first, useful features are extracted using Digital Signal Processing methods. Then, these features are fed into various Neural Network architectures. Finally, all experiments are evaluated and the optimal features and architectures are selected

Keywords

Music genre classification, Music sub-genre classification, Rock music, Digital Signal Processing, Machine Learning, Deep Learning, Music Information Retrieval, MIR

στους γονείς μου

Ευχαριστίες

Ολοκληρώντας την εκπόνηση της διπλωματικής μου εργασίας και κλείνοντας έτσι τον κύκλο της φοίτησής μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Ηλεκτρονικών Υπολογιστών, θα ήθελα να ευχαριστήσω όλους εκείνους που συνέβαλαν σε αυτή μου την προσπάθεια και κατέστησαν εφικτό αυτό το αποτέλεσμα.

Καταρχάς θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή Ανδρέα Σταφυλοπάτη, ο οποίος μου έδωσε την ευκαιρία να ασχοληθώ με αυτό το ιδιαίτερα ενδιαφέρον ερευνητικό ζήτημα. Ακόμα, ευχαριστώ ιδιαίτερα τον Δρ. Γεώργιο Σιόλα για την καθοριστική καθοδήγηση και τις πολύτιμες συμβουλές που μου παρείχε σε όλη τη διάρκεια του εγχειρήματος της διπλωματικής εργασίας.

Θα ήθελα επίσης να ευχαριστήσω την ομάδα μου στην Orfium για την ευκαιρία που μου παρείχαν να εργαστώ δίπλα τους κατά τη διάρκεια της Πρακτικής Άσκησης. Αυτό υπήρξε το έναυσμα όχι μόνο για την επιλογή του συγκεκριμένου θέματος διπλωματικής εργασίας, αλλά και για τα πρώτα μου βήματα στην επαγγελματική μου σταδιοδρομία. Χωρίς τις συμβουλές των συναδέλφων μου κατά τη διάρκεια της έρευνας αλλά και την παροχή των απαραίτητων υπολογιστικών πόρων για τη διεξαγωγή των πειραμάτων, η εργασία αυτή δεν θα ήταν εφικτή.

Τέλος, θα ήθελα να ευχαριστήσω ειλικρινά και μέσα από την καρδιά μου την οικογένειά μου, του φίλους μου και τον Β. για την αγάπη, τη συμπαράσταση και την υπομονή τους σε όλη τη διάρκεια των σπουδών μου. Είμαι ευγνώμων που τους είχα κοντά μου.

Αθήνα, Ιούλιος 2021

Παρασκευή Λεμπέση

Περιεχόμενα

Περίληψη 1

Abstract 3

Ευχαριστίες 7

1 Εισαγωγή 17
1.1 Αντικείμενο της Διπλωματικής . . . 17
1.2 Παρεμφερείς Εργασίες . . . 18
1.3 Οργάνωση του Τόμου . . . 19

I Θεωρητικό Μέρος 21

2 Θεωρητικό Υπόβαθρο 23
2.1 Τεχνητή Νοημοσύνη . . . 23
2.2 Μηχανική Μάθηση . . . 23
2.2.1 Επιβλεπόμενη Μάθηση . . . 24
2.2.2 Μη Επιβλεπόμενη Μάθηση . . . 25
2.2.3 Ενισχυτική Μάθηση . . . 25
2.3 Νευρωνικά Δίκτυα . . . 25
2.3.1 Τεχνητά Νευρωνικά Δίκτυα . . . 25
2.3.2 Ο Αλγόριθμος Perceptron . . . 26
2.3.3 Πολυεπίπεδα Νευρωνικά Δίκτυα . . . 27
2.4 Συνελικτικά Νευρωνικά Δίκτυα . . . 28
2.4.1 Επίπεδα Επεξεργασίας . . . 29
2.5 Αναδρομικά Νευρωνικά Δίκτυα . . . 32
2.5.1 Νευρωνικά Δίκτυα Long Short-Term Memory . . . 33
2.5.2 Νευρωνικά Δίκτυα Gated Recurrent Unit . . . 34
2.6 Εκπαίδευση Νευρωνικών Δικτύων . . . 36
2.6.1 Συνάρτηση Κόστους . . . 36
2.6.2 Αλγόριθμος Οπισθοδιάδοσης (Backpropagation) . . . 36
2.6.3 Αλγόριθμος Κατάβασης Κλίσης (Gradient Descent) . . . 37
2.6.4 Αλγόριθμοι Βελτιστοποίησης . . . 38
2.6.5 Συνάρτηση Ενεργοποίησης . . . 39
2.7 Αξιολόγηση Επιδόσεων . . . 41

2.7.1 Μετρικές Αξιολόγησης 42

II Πρακτικό Μέρος Α΄ - Ταξινόμηση σε μουσικό είδος 47

3 Δεδομένα και Προεπεξεργασία 49
3.1 Σύνολο Δεδομένων 49
3.2 Προεπεξεργασία και Εξαγωγή Χρήσιμων Χαρακτηριστικών 50
3.2.1 Διαίρεση Κομματιού 50
3.2.2 Μετασχηματισμοί Ηχητικού Σήματος 51
3.2.3 Διαχωρισμός Συνόλου Δεδομένων 55

4 Αρχιτεκτονικές Νευρωνικών Δικτύων 57
4.1 Αρχιτεκτονική Νευρωνικού Α1 57
4.1.1 Σχήμα και Διαστάσεις των Πυρήνων των Συνελίξεων 58
4.1.2 Σύνδεση Συντόμευσης (Shortcut Connection - Residual Block) . . . 58
4.1.3 Κανονικοποίηση Δέσμης (Batch Normalization) 59
4.1.4 Στρώματα Εγκατάλειψης (Dropout Layers) 59
4.1.5 Ομαλοποίηση Πυρήνα (Kernel Regularizer) 59
4.2 Αρχιτεκτονική Νευρωνικού Α2 60
4.2.1 Είσοδος STFT 60
4.2.2 Είσοδος Chroma STFT 61
4.2.3 Είσοδος Mel Spectrogram 61
4.2.4 Είσοδος MFCC 63
4.2.5 Είσοδος MFCC & delta και CQT 63

5 Αποτελέσματα και Αξιολόγηση 65
5.1 Πειράματα Αρχιτεκτονικής Νευρωνικού Α1 65
5.2 Πειράματα Αρχιτεκτονικής Νευρωνικού Α2 66
5.2.1 Αποτελέσματα Πειραμάτων για Διαφορετικές Εισόδους 67
5.2.2 Αποτελέσματα Πειραμάτων Συνδυασμού Εισόδων 69

III Πρακτικό Μέρος Β΄ - Ταξινόμηση σε μουσικό υποείδος 73

6 Δεδομένα και Προεπεξεργασία 75
6.1 Σύνολο Δεδομένων 75
6.1.1 Σύνολο FMA 75
6.1.2 Υποσύνολο του FMA για την Ταξινόμηση σε Μουσικό Υποείδος . . . 76
6.2 Προεπεξεργασία και Εξαγωγή Χρήσιμων Χαρακτηριστικών 77
6.2.1 Διαίρεση Κομματιού 77
6.2.2 Εξαγωγή Μετασχηματισμού STFT και Κανονικοποίηση 78
6.2.3 Υποδειγματοληψία 79
6.2.4 Επαύξηση Δεδομένων (Data Augmentation) 80
6.2.5 Διαχωρισμός Συνόλου Δεδομένων 80

7 Αρχιτεκτονικές Νευρωνικών Δικτύων **81**
7.1 Αρχιτεκτονική Νευρωνικού B1 81
7.2 Αρχιτεκτονική Νευρωνικού B2 82
7.3 Αρχιτεκτονική Νευρωνικού B3 83
7.4 Αρχιτεκτονική Νευρωνικού B4 83
7.5 Αρχιτεκτονική Νευρωνικού B5 84

8 Αποτελέσματα και Αξιολόγηση **89**
8.1 Πειράματα με Διαφορετικές Αρχιτεκτονικές 89
8.2 Πειράματα με Διαφορετικά Αναδρομικά Επίπεδα 91
8.3 Πειράματα με Διαφορετικές Μεθόδους Σύμπτυξης 93
8.4 Πειράματα με Διαφορετικές Μεθόδους Συνδυασμού Παράλληλων Σκελών . . 95

IV Επίλογος **99**

9 Σύνοψη και Μελλοντικές Επεκτάσεις **101**
9.1 Σύνοψη 101
9.2 Μελλοντικές Επεκτάσεις 102

Βιβλιογραφία **108**

Κατάλογος Σχημάτων

2.1 Η Δομή ενός Τεχνητού Νευρώνα . . . 27
2.2 Πολυεπίπεδο Νευρωνικό Δίκτυο . . . 29
2.3 Συνελικτικό Νευρωνικό Δίκτυο . . . 30
2.4 Συνελικτικό Επίπεδο . . . 31
2.5 Συγκεντρωτικό Επίπεδο . . . 32
2.6 Αναδρομικό νευρωνικό δίκτυο σε συμπτυγμένη και αναπτυγμένη μορφή. . . . 33
2.7 Κελί LSTM . . . 34
2.8 Κελί GRU . . . 35
2.9 Αλγόριθμος κατάβασης κλίσης. . . . 37
2.10 Επίδραση του ρυθμού μάθησης στον αλγόριθμο κατάβασης κλίσης. . . . 38
2.11 Σιγμοειδής συνάρτηση. . . . 40
2.12 Συνάρτηση υπερβολικής εφαπτομένης. . . . 40
2.13 Συνάρτηση Rectified Linear Unit - ReLU. . . . 41
2.14 Συνάρτηση Rectified Linear Unit - ReLU. . . . 41
2.15 Συνάρτηση Softmax. . . . 42
2.16 Πίνακας Σύγχυσης (Confusion Matrix) - Ορισμός. . . . 44
2.17 Πίνακας Σύγχυσης (Confusion Matrix) - Παράδειγμα. . . . 44

3.1 Κυματομορφή ηχητικού σήματος . . . 50
3.2 Φασματογράφημα μετασχηματισμού STFT . . . 51
3.3 Φασματογράφημα μετασχηματισμού Chroma STFT . . . 53
3.4 Φασματογράφημα μετασχηματισμού Mel Spectrogram . . . 53
3.5 Διαδικασία Εξαγωγής μετασχηματισμού MFCC . . . 54
3.6 Φασματογράφημα μετασχηματισμού CQT . . . 55

4.1 Αρχιτεκτονική Νευρωνικού A1 . . . 57
4.2 Εφαρμογή τεχνικής dropout . . . 60
4.3 Αρχιτεκτονική Νευρωνικού A2 - Είσοδος STFT . . . 61
4.4 Αρχιτεκτονική Νευρωνικού A2 - Είσοδος Chroma STFT . . . 62
4.5 Αρχιτεκτονική Νευρωνικού A2 - Είσοδος Mel Spectrogram . . . 62
4.6 Αρχιτεκτονική Νευρωνικού A2 - Είσοδος MFCC . . . 63
4.7 Αρχιτεκτονική Νευρωνικού A2 - Είσοδος MFCC & delta . . . 64
4.8 Αρχιτεκτονική Νευρωνικού A2 - Είσοδος CQT . . . 64

5.1 Πίνακες Σύγχυσης Αρχιτεκτονικής A2 - Διαφορετικές Είσοδοι . . . 71

6.1 Κατανομή συνόλου FMA medium 76
6.2 Ιεραρχία Υποειδών Rock 77
6.3 Κατανομή στα υποείδη της Rock 78
6.4 Κατανομή στα υποείδη της Rock πριν και μετά την Υποδειγματοληψία 79

7.1 Αρχιτεκτονική Νευρωνικού B1 82
7.2 Αρχιτεκτονική Νευρωνικού B2 84
7.3 Αρχιτεκτονική Νευρωνικού B3 85
7.4 Αρχιτεκτονική Νευρωνικού B4 86
7.5 Αρχιτεκτονική Νευρωνικού B5 87

8.1 Διαφοροποιήσεις Αρχιτεκτονικών 90
8.2 Διαφορετικές Αρχιτεκτονικές - Πίνακες Σύγχυσης 92
8.3 Διαφορετικά Αναδρομικά Επίπεδα - Πίνακες Σύγχυσης 94
8.4 Διαφορετικές Μέθοδοι Σύμπτυξης - Πίνακες Σύγχυσης 96
8.5 Διαφορετικές Μέθοδοι Συνδυασμού Παράλληλων Σκελών - Πίνακες Σύγχυσης 97

Κατάλογος Πινάκων

5.1 Αξιολόγηση Πειραμάτων Αρχιτεκτονικής A1 66
5.2 Αξιολόγηση Πειραμάτων Αρχιτεκτονικής A2 68
5.3 Αξιολόγηση Μεθόδων Συνδυασμού Μοντέλων 70

7.1 Πλήθος παραμέτρων Αρχιτεκτονικών Μέρους Β′ 81

8.1 Αποτελέσματα πειραμάτων διαφορετικών Αρχιτεκτονικών 91
8.2 Αποτελέσματα πειραμάτων διαφορετικών Αναδρομικών Επιπέδων 93
8.3 Αποτελέσματα πειραμάτων διαφορετικών Μεθόδων Σύμπτυξης 95
8.4 Αποτελέσματα πειραμάτων διαφορετικών Μεθόδων Συνδυασμού Παράλληλων Σκελών 96

Κεφάλαιο 1

Εισαγωγή

Η ολοένα και ταχύτερη πρόοδος της τεχνολογίας τα τελευταία χρόνια έχει απλοποιήσει σημαντικά τη ζωή των ανθρώπων και έχει επιφέρει ως αποτέλεσμα τη διευκόλυνση πολλών πτυχών της καθημερινότητας. Φυσικά, για να επέλθει αυτό το αποτέλεσμα, έχει προηγηθεί ένας αναγκαίος, και ίσως δύσκολος και χρονοβόρος, ανασχηματισμός πολλών τομέων της ανθρώπινης δραστηριότητας, ώστε να είναι πλέον συμβατοί με αυτό που πολλοί ονομάζουν ως "ψηφιακή εποχή". Η μουσική βιομηχανία, ως ένας από τους κύριους πυλώνες της διασκέδασης και του πολιτισμού, δεν θα μπορούσε να μείνει ανεπηρέαστη από αυτή την συνθήκη. Η επικράτηση της ψηφιακής διανομής της μουσικής έχει καταστήσει την ριζική αλλαγή της μουσικής βιομηχανίας απαραίτητη για την επιβίωση και την αποδοτική της λειτουργία.

Μέρος του ανασχηματισμού της μουσικής βιομηχανίας είναι η αυτοματοποίηση πολλών διαδικασιών η οποίας παλαιότερα απαιτούσαν την ανθρώπινη παρέμβαση. Μία από αυτές, είναι και η αναγνώριση του μουσικού είδους. Η πληροφορία αναφορικά με το είδος στο οποίο κατάσσεται ένα μουσικό κομμάτι είναι ιδιαίτερα χρήσιμη, διότι διευκολύνει την διαχείριση και την οργάνωση του ομολογουμένως τεράστιου όγκου δεδομένων μουσικής. Μπορεί επίσης να χρησιμοποιηθεί ως επιμέρους συνιστώσα για την δημιουργία συστημάτων προτάσεων μουσικής, για την πρόβλεψη της επιτυχίας ενός νέου μουσικού κομματιού, την αυτόματη εύρεση διασκευής ενός ήδη υπάρχοντος μουσικού κομματιού ή για την εύρεση παρόμοιων καλλιτεχνών. Μάλιστα, ο μεγάλος όγκος δεδομένων μουσικής ευνοεί την αξιοποίηση της Μηχανικής Μάθησης για την αντιμετώπιση του συγκεκριμένου προβλήματος.

1.1 Αντικείμενο της Διπλωματικής

Το αντικείμενο της παρούσα διπλωματικής εργασίας θα μπορούσε να διαχωριστεί σε δύο μέρη:

1. Το πρώτο μέρος αφορά στην δημιουργία ενός συστήματος ταξινόμησης της μουσικής σε 10 βασικά είδη. Το πρόβλημα αυτό πρόκειται για ένα ζήτημα που έχει μελετηθεί αρκετά από την ερευνητική κοινότητα, επομένως η πρόκληση εδώ έγκειται στο να πετύχουμε ένα σχετικά υψηλό ποσοστό ακρίβειας. Για το σκοπό αυτό είναι αναγκαίο ένα σύνολο δεδομένων το οποίο θα έχει χρησιμοποιηθεί από άλλες έρευνες ώστε να μπορέσουμε να το αξιοποιήσουμε ως σημείο αναφοράς για την αξιολόγηση του δικού μας συστήματος.

2. Το δεύτερο μέρος πρόκειται ουσιαστικά για την περαιτέρω εμβάθυνση του πρώτου

μέρους, με τη δημιουργία ενός συστήματος ταξινόμησης συγκεκριμένα της ροκ μουσικής στα επιμέρους 10 υποείδη της. Τέτοιου είδους προβλήματα, όπου οι κλάσεις στις οποίες προσπαθούμε να ταξινομήσουμε τα δεδομένα μας είναι πολύ παρόμοιες μεταξύ τους, ερευνώνται ενεργά από την επιστημονική κοινότητα και θεωρούνται ως μια αρκετά απαιτητική περίπτωση ταξινόμησης της μουσικής.

Τόσο για το πρώτο, όσο και για το δεύτερο μέρος, είναι απαραίτητο ένα επισημασμένο (labeled) σύνολο δεδομένων, ώστε να μπορέσουμε να εφαρμόσουμε τεχνικές Επιβλεπόμενης Μηχανικής Μάθησης. Τα σύνολα δεδομένων που χρησιμοποιούμε αποτελούνται από δεδομένα ήχου (audio data) από τα οποία εξάγουμε χρήσιμα χαρακτηριστικά αξιοποιώντας την θεωρία της Ψηφιακής Επεξεργασίας Σήματος. Στη συνέχεια αυτά τα χαρακτηριστικά τροφοδοτούνται σε διάφορες αρχιτεκτονικές Νευρωνικών Δικτύων και γίνεται προσπάθεια, μέσω πολλαπλών πειραματισμών, για την εύρεση του συνδυασμού χρήσιμων χαρακτηριστικών και αρχιτεκτονικής που δίνει τα βέλτιστα δυνατά αποτελέσματα.

1.2 Παρεμφερείς Εργασίες

Το πρόβλημα της Αναγνώρισης Μουσικού Είδους εισήχθη για πρώτη φορά το 2002 από τους Tzanetakis και Cook [1] ως ένα ζήτημα αναγνώρισης προτύπων. Έκτοτε, παραμένει ένα ενεργό ερευνητικό πεδίο στον τομέα της Ανάκτησης Μουσικής Πληροφορίας (Music Information Retrieval - MIR) και έχουν υπάρξει πολλές διαφορετικές προσεγγίσεις για την επίλυσή του. Στην πλειοψηφία των προσεγγίσεων αυτών, το ζήτημα διαχωρίζεται σε δύο σκέλη: την εξαγωγή κατάλληλων χαρακτηριστικών από το ηχητικό σήμα και την τροφοδότηση αυτών των χαρακτηριστικών σε έναν ταξινομητή. Δυστυχώς όμως, η χειροκίνητη εξαγωγή χαρακτηριστικών είναι μία περίπλοκη διαδικασία η οποία απαιτεί την εξειδίκευση των ερευνητών στον τομέα της μουσικής. Επιπλέον, ενέχει και σημαντικούς κινδύνους, καθώς τα χαρακτηριστικά αυτά στερούνται καθολικότητας, επομένως ενδέχεται να υπάρχουν μεγάλες διακυμάνσεις στην απόδοση από περίπτωση σε περίπτωση.

Οι επιδόσεις των συστημάτων Αναγνώρισης Μουσικού Είδους εκτοξεύτηκαν τα τελευταία χρόνια, με την πρόοδο στον τομέα της Ταξινόμησης Εικόνων με τη χρήση συνελικτικών νευρωνικών δικτύων (Convolutional Neural Networks - CNNs). Ταυτόχρονα, οι Dieleman και Schrauwen [2] απέδειξαν πως τα φασματογραφήματα ηχητικών σημάτων μπορούν να αντιμετωπιστούν ως εικόνες και να έχουν αρκετά ικανοποιητικές επιδόσεις με τη χρήση συνελικτικών νευρωνικών δικτύων. Υπό τις συνθήκες αυτές, άρχισε να αναπτύσσεται ραγδαία η τεχνική της εκμάθησης αναπαραστάσεων χαρακτηριστικών από τα φασματογραφήματα με τη χρήση συνελικτικών νευρωνικών δικτύων και τεχνικών βαθιάς μηχανικής μάθησης. Σημαντικό πλεονέκτημα σε αυτή την προσέγγιση είναι πως δεν απαιτείται πρότερη γνώση επί του πεδίου της μουσικής. Επιπλέον, στα πλαίσια αυτά, οι ερευνητικές προσπάθειες στον τομέα της Αναγνώρισης Μουσικού Είδους άρχισαν να επηρεάζονται από ερευνητικές προσπάθειες στον τομέα της Ταξινόμησης Εικόνων, δεδομένης της συνάφειας μεταξύ τους. Σημαντικό παράδειγμα αποτελεί η προσέγγιση των Zhang et al. [3], η οποία είναι εμπνευσμένη από τη λογική των συνδέσεων συντόμευσης [4] στα συνελικτικά δίκτυα για Ταξινόμηση Εικόνας.

Παρόλα αυτά, όπως επισημάνθηκε από τους J.Pons, T.Lidy και X.Serra [5], ‘‘μια συχνή

επίκριση σε βάρος των τεχνικών βαθιάς μάθησης είναι η δυσκολία αντίληψης των υποβόσκουσων σχέσων οι οποίες μαθαίνονται από τα νευρωνικά δίκτυα, και επομένως η λειτουργία τους ως ένα μαύρο κουτί''. Συνεπώς, πειραματιζόμενοι με τα σχήματα των πυρήνων των συνελίξεων, απέδειξαν ότι ο συνδυασμός ενός συνελικτικού δικτύου που ενσωματώνει χρονική πληροφορία και ένα δεύτερο που ενσωματώνει τονική πληροφορία είναι μια ελπιδοφόρα προσέγγιση στον τομέα της Αναγνώρισης Μουσικού Είδους.

Παρά τα ιδιαίτερα ενθαρρυντικά αποτελέσματα της χρήσης συνελικτικών νευρωνικών δικτύων, διαπιστώθηκε πως τα φασματογραφήματα, σε αντίθεση με τις κοινές εικόνες, εμπεριέχουν ακολουθιακές χρονικές σχέσεις οι οποίες δεν μπορούν να μοντελοποιηθούν επαρκώς με τη συγκεκριμένη τεχνική. Λαμβάνοντας υπόψιν τη συγκεκριμένη παρατήρηση, οι Choi et al. [6] σχεδίασαν ένα υβριδικό μοντέλο, αποτελούμενο από δύο αναδρομικά επίπεδα τα οποία ακολουθούν τη συνελικτική δομή και λειτουργούν συνοψίζοντας τη χρονική πληροφορία. Ωστόσο, δεδομένου ότι τα αναδρομικά επίπεδα βρίσκονται μετά από τα συνελικτικά, ένα μεγάλο μέρος των χρονικών σχέσεων ενδέχεται να χαθεί λόγω των επαναλαμβανόμενων συνελίξεων και να μην μπορεί να γίνει αντιληπτό στο τέλος του δικτύου. Στο πλαίσιο αυτής της διαπίστωσης, οι Yang et al. [7] πρότειναν ένα υβριδικό μοντέλο αποτελούμενο από ένα συνελικτικό δίκτυο συνδεδεμένο παράλληλα με ένα αναδρομικό, δίνοντας ιδιαίτερα ενθαρρυντικά αποτελέσματα.

1.3 Οργάνωση του Τόμου

Η παρούσα διπλωματική εργασία χωρίζεται σε 4 μέρη. Στο πρώτο μέρος, το οποίο περιλαμβάνει το Κεφάλαιο 2, αναπτύσσονται αναλυτικά θεωρητικές έννοιες που αφορούν τη Μηχανική Μάθηση και τα Νευρωνικά Δίκτυα. Όσον αφορά τα δεύτερα, παρουσιάζεται το Perceptron, τα Συνελικτικά Νευρωνικά Δίκτυα, τα Αναδρομικά Νευρωνικά Δίκτυα καθώς και οι αλγόριθμοι που χρησιμοποιούνται για την εκπαίδευσή τους. Στα επόμενα 2 μέρη της εργασίας περιγράφεται η πρακτική προσέγγιση του ζητήματος της Αναγνώρισης Μουσικού Είδους και Μουσικού Υποείδους αντίστοιχα. Πιο συγκεκριμένα, στο Κεφάλαιο 3 παρουσιάζεται το σύνολο δεδομένων που χρησιμοποιήθηκε για την εκπαίδευση του συστήματος Αναγνώρισης Μουσικού Είδους καθώς και οι τεχνικές που εφαρμόστηκαν για την προεπεξεργασία των δεδομένων αυτών και την εξαγωγή χρήσιμων χαρακτηριστικών. Ακολούθως, στο Κεφάλαιο 4 περιγράφονται οι αρχιτεκτονικές των νευρωνικών δικτύων πάνω στις οποίες διεξάχθηκαν πειράματα, ενώ στο Κεφάλαιο 5 γίνεται αναλυτική παρουσίαση των πειραμάτων αυτών καθώς και των αποτελεσμάτων που προέκυψαν. Κατ΄ αντιστοιχία, στα Κεφάλαια 6, 7 και 8 καταγράφεται η προσέγγιση που ακολουθήθηκε για τη δημιουργία ενός συστήματος Αναγνώρισης Μουσικού Υποείδους. Συγκεκριμένα, το Κεφάλαιο 6 πραγματεύεται το σύνολο δεδομένων και τις μεθόδους προεπεξεργασίας του, το Κεφάλαιο 7 τις προτεινόμενες αρχιτεκτονικές, ενώ το Κεφάλαιο 8 περιλαμβάνει τα αποτελέσματα των επιμέρους πειραμάτων. Τέλος, το τέταρτο μέρος του τόμου, το οποίο αποτελείται από το Κεφάλαιο 9, αποτελεί μια σύνοψη της δουλειάς της εργασίας αυτής, ενώ επίσης περιέχει προτάσεις και ιδέες για επιπλέον μελλοντική έρευνα.

Μέρος I

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Ο στόχος του παρόντος κεφαλαίου είναι η παρουσίαση των θεωρητικών εννοιών και του επιστημονικού υποβάθρου για την καλύτερη κατανόηση της διπλωματικής εργασίας. Θα γίνει αναφορά στους ορισμούς και στις βασικές αρχές που διέπουν την λειτουργία των συστημάτων Τεχνητής Νοημοσύνης, Μηχανικής Μάθησης και των Νευρωνικών Δικτύων. Επίσης, θα παρουσιαστούν οι επιμέρους κατηγορίες Μηχανικής Μάθησης και Νευρωνικών Δικτύων αλλά και οι τρόποι με τους οποίους οι προαναφερθείσες τεχνολογίες αξιολογούνται από την επιστημονική κοινότητα.

2.1 Τεχνητή Νοημοσύνη

Η Τεχνητή Νοημοσύνη είναι ο κλάδος της Επιστήμης των Υπολογιστών ο οποίος, σύμφωνα με τον John McCarthy, εστιάζει στη δημιουργία έξυπνων μηχανών και πιο συγκεκριμένα έξυπνων υπολογιστικών προγραμμάτων [8]. Υπάρχουν αρκετοί εναλλακτικοί ορισμοί της Τεχνητής Νοημοσύνης, μέρος εκ των οποίων εστιάζουν στις διαδικασίες σκέψης και συλλογιστικής και άλλοι στη συμπεριφορά, ενώ ένας διαφορετικός διαχωρισμός των ορισμών αυτών γίνεται με βάση την ομοιότητα ενός υπολογιστικού συστήματος με τον είτε με τον άνθρωπο είτε με ένα ιδανικό ορθολογικό σύστημα[9]. Ενδεικτικά, αναφέρουμε τον ορισμό κατά Rich και Knight, σύμφωνα με τον οποίον, η Τεχνητή Νοημοσύνη είναι η μελέτη του πώς μπορούμε να κάνουμε τους υπολογιστές να κάνουν πράγματα στα οποία, προς το παρόν, οι άνρθωποι είναι καλύτεροι [10]. Ένας από τους πιο πλήρης ορισμούς της Τεχνητής Νοημοσύνης έχει προταθεί από τον Alan Turing, επομένως και ονομάζεται ‘‘Δοκιμασία Turing’’. Σύμφωνα με αυτό τον ορισμό, ένα υπολογιστικό σύστημα θεωρείται σύστημα Τεχνητής Νοημοσύνης αν ένας άνθρωπος εξεταστής, ο οποίος θέτει γραπτές ερωτήσεις, δεν μπορεί να αποφανθεί αν οι επίσης γραπτές απαντήσεις προέρχονται από άνθρωπο ή από υπολογιστή [11].

2.2 Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι ο επιμέρους κλάδος της Τεχνητής Νοημοσύνης ο οποίος περιλαμβάνει υπολογιστικές μεθόδους που αξιοποιούν την εμπειρία με στόχο να κάνουν ακριβείς προβλέψεις ή να βελτιώσουν τις επιδόσεις τους σε ένα συγκεκριμένο πεδίο [12]. Αυτό επί της ουσίας σημαίνει ότι, όπως και οι άνθρωποι, το υπολογιστικό σύστημα κατασκευάζει νοητικά πρότυπα, δηλαδή συσχετισμούς μεταξύ παρατηρήσεων και αποτελεσμάτων [13], και στη

συνέχεια εφαρμόζει αυτά τα νοητικά πρότυπα για να προβλέψει τα αποτελέσματα καινούριων παρατηρήσεων.

Η Μηχανική Μάθηση βρίσκει εφαρμογή στην επίλυση προβλημάτων που είναι δύσκολο να οριστούν με αυστηρούς μαθηματικούς όρους αλλά για τα οποία υπάρχει διαθέσιμη μία πληθώρα παραδειγμάτων. Τα παραδείγματα αυτά συνθέτουν το σύνολο εκπαίδευσης, δηλαδή την ''εμπειρία'' η οποία αξιοποιείται κατά τη φάση της εκπαίδευσης με σκοπό τη δημιουργία ενός μαθηματικού μοντέλου. Χαρακτηριστικές περιπτώσεις τέτοιων προβλημάτων είναι η αναγνώριση συγκεκριμένων αντικειμένων σε εικόνες, η αναγνώριση συναισθήματος σε ένα μουσικό κομμάτι, αλλά και η ταξινόμηση αρχείων ήχων ανάλογα με το μουσικό είδος, το οποίο είναι και το θέμα της παρούσας εργασίας.

Οι αλγόριθμοι Μηχανικής Μάθησης, ανάλογα με το είδος των δεδομένων και τον τρόπο με τον οποίο αυτά χρησιμοποιούνται κατά της φάση της εκπαίδευσης, μπορούν να ταξινομηθούν σε τρεις ξεχωριστές κατηγορίες:

- Επιβλεπόμενη Μάθηση (Supervised Learning)
- Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)
- Ενισχυτική Μάθηση (Reinforcement Learning)

2.2.1 Επιβλεπόμενη Μάθηση

Η Επιβλεπόμενη Μάθηση είναι η περίπτωση Μηχανικής Μάθησης κατά την οποία τα δεδομένα του συνόλου εκπαίδευσης (training set) είναι επισημασμένα [12]. Αυτό σημαίνει ότι για κάθε δείγμα εισόδου, γνωρίζουμε εκ των προτέρων την επιθυμητή έξοδο και ο στόχος είναι η δημιουργία ενός μοντέλου με τη δυνατότητα γενίκευσης. Με τον όρο γενίκευση εννοούμε τη δυνατότητα του μοντέλου να παράγει σωστές προβλέψεις για εισόδους που δεν έχει συναντήσει ξανά στη φάση της εκπαίδευσης [14]. Τα προβλήματα που αντιμετωπίζονται με μεθόδους Επιβλεπόμενης Μάθησης διαχωρίζονται σε δύο κατηγορίες με βάση τη μορφή της εξόδου που παράγουν [15]:

- Προβλήματα Ταξινόμησης (Classification)

  Στα προβλήματα αυτά, η έξοδος του μοντέλου λαμβάνει μια διακριτή τιμή μέσα από ένα προκαθορισμένο σύνολο. Επί της ουσίας, ο στόχος ενός αλγορίθμου ταξινόμησης είναι η εύρεση της κατηγορίας στην οποία ανήκει ένα συγκεκριμένο δείγμα, δηλαδή ένα ποιοτικό χαρακτηριστικό του εν λόγω δείγματος. Στο σημείο αυτό αξίζει να αναφερθεί ότι η μόνη περίπτωση που η έξοδος ενός μοντέλου ταξινόμησης είναι συνεχής αριθμός είναι όταν εκφράζει την πιθανότητα να ανήκει το δείγμα σε μία συγκεκριμένη κατηγορία.

- Προβλήματα Παλινδρόμησης (Regression)

  Στα προβλήματα αυτά, η έξοδος του μοντέλου λαμβάνει μια συνεχή τιμή. Η τιμή αυτή ανήκει σε ένα συνεχές σύνολο τιμών και αντιπροσωπεύει ένα ποσοτικό χαρακτηριστικό του δείγματος στο οποίο αναφέρεται.

2.2.2 Μη Επιβλεπόμενη Μάθηση

Η Μη Επιβλεπόμενη Μάθηση είναι η περίπτωση Μηχανικής Μάθησης κατά την οποία τα δεδομένα του συνόλου εκπαίδευσης δεν είναι επισημασμένα. Ο στόχος ενός αλγορίθμου Μη Επιβλεπόμενης Μάθησης είναι να εξάγει τα εγγενή χαρακτηριστικά των δεδομένων εισόδου και να εντοπίσει μοτίβα και κρυμμένα πρότυπα που μοντελοποιούν την κατανομή τους [16]. Τα κυριότερα προβλήματα τα οποία αντιμετωπίζονται με αυτό το είδος Μηχανικής Μάθησης είναι τα ακόλουθα:

- Προβλήματα Συσταδοποίησης (Clustering)

  Στα προβλήματα αυτά, ο στόχος είναι ο διαχωρισμός σε συστάδες (clusters) με βάσει την ομοιότητα των δειγμάτων μεταξύ τους. Πιο συγκεκριμένα, δείγματα παρόμοια μεταξύ τους ταξινομούνται στην ίδια συστάδα, ενώ δείγματα που δεν έχουν κοινά χαρακτηριστικά ταξινομούνται σε διαφορετικές συστάδες [17].

- Προβλήματα Μείωσης Διαστάσεων (Dimensionality Reduction)

  Στα προβλήματα αυτά, ο στόχος είναι, όπως είναι εύκολα κατανοητό και από το όνομα, η μείωση των διαστάσεων ενός συνόλου δεδομένων. Αυτό γίνεται γιατί αρκετές φορές οι επιμέρους διαστάσεις ενός συνόλου δεδομένων είναι συσχετισμένες μεταξύ τους με αποτέλεσμα να δημιουργείται πλεονασμός και να δυσχεραίνεται η οπτικοποίηση των δεδομένων αυτών αλλά και η εκτέλεση υπολογισμών με αυτά.

2.2.3 Ενισχυτική Μάθηση

Η Ενισχυτική Μάθηση είναι μία κατηγορία Μηχανικής Μάθησης διαφορετική από τις προηγούμενες. Τα δεδομένα, όπως και στην περίπτωση της Μη Επιβλεπόμενης Μάθησης είναι μη επισημασμένα, δηλαδή ο αλγόριθμος δεν γνωρίζει κατά την εκπαίδευση ποια είναι η επιθυμητή έξοδος για κάθε περίπτωση εισόδου. Αντιθέτως, αλληλεπιδρά με το περιβάλλον του, επηρεάζοντάς το μάλιστα σε ορισμένες περιπτώσεις, και ανάλογα με το αποτέλεσμα της κάθε αλληλεπίδρασης είτε ανταμείβεται είτε τιμωρείται. Ο στόχος του αλγορίθμου είναι να μεγιστοποιήσει την αθροιστική ανταμοιβή του, καλούμενος σε κάθε βήμα να αποφασίσει ανάμεσα στην εξερεύνηση (exploration), δηλαδή να δοκιμάσει άγνωστες ως τώρα ενέργειες, και στην εκμετάλλευση (exploitation), δηλαδή να παραμείνει πιστός σε ήδη δοκιμασμένες ενέργειες που είναι γνωστό ότι προσφέρουν μεγάλη επιβράβευση [12].

2.3 Νευρωνικά Δίκτυα

Στο υποκεφάλαιο αυτό θα ασχοληθούμε με την τεχνολογία των Νευρωνικών Δικτύων, εστιάζοντας αρχικά στα τεχνητά νευρωνικά δίκτυα, πώς αυτά δομούνται, καθώς και ορισμένους βασικούς αλγορίθμους που διέπουν τη λειτουργία τους.

2.3.1 Τεχνητά Νευρωνικά Δίκτυα

Στη βιβλιογραφία υπάρχει διαθέσιμος ένας μεγάλος αριθμός ορισμών για τα Τεχνητά Νευρωνικά Δίκτυα. Σύμφωνα με τον Simon Haykin, ένα Τεχνητό Νευρωνικό Δίκτυο είναι ένας

τεράστιος παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από τη φύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση. Μοιάζει με τον ανθρώπινο σε δύο σημεία:

1. Το δίκτυο προσλαμβάνει τη γνώση από το περιβάλλον του, μέσω μιας διαδικασίας μάθησης.

2. Η ισχύς των συνδέσεων μεταξύ των νευρώνων, που αποκαλείται συναπτικό βάρος, χρησιμοποιείται για την αποθήκευση της γνώσης που αποκτιέται [18].

Τα Νευρωνικά Δίκτυα έχουν αναπτυχθεί ιδιαίτερα τα τελευταία χρόνια λόγω της μεγαλύτερης διαθεσιμότητας δεδομένων και της ευκολότερης πρόσβασης σε αυτά, η οποία έχει επέλθει ως φυσικό επακόλουθο της ανάπτυξης του διαδικτύου και της γενικότερης διάδοσης της πληροφορίας. Επιπλέον, κομβικό ρόλο έχει επιτελέσει και η μεγάλη πρόοδος στον τομέα της παράλληλης επεξεργασίας των δεδομένων με την αλματώδη βελτίωση των Γραφικών Μονάδων Επεξεργασίας (GPU), η οποία, λόγω και της εγγενούς παραλληλίας των νευρωνικών δικτύων, καθιστά δυνατή την επεξεργασία τεράστιου όγκου δεδομένων σε μικρό χρονικό διάστημα.

Επιπλέον, τα νευρωνικά δίκτυα διαθέτουν κάποια σημαντικά πλεονεκτήματα, τα οποία συμβάλλουν ακόμα περισσότερο στην ανάπτυξή τους. Τα βασικότερα εξ αυτών είναι:

1. Μη γραμμικότητα

2. Αντιστοίχιση εισόδου - εξόδου

3. Προσαρμοστικότητα

4. Ενδεικτική απόκριση

5. Πληροφορία σχετική με το περιεχόμενο

6. Ανοχή σε βλάβες

7. Δυνατότητα υλοποίησης σε VLSI

8. Ομοιομορφία ανάλυσης και σχεδίασης

9. Αναλογία με νευροφυσιολογία του εγκεφάλου [18]

2.3.2 Ο Αλγόριθμος Perceptron

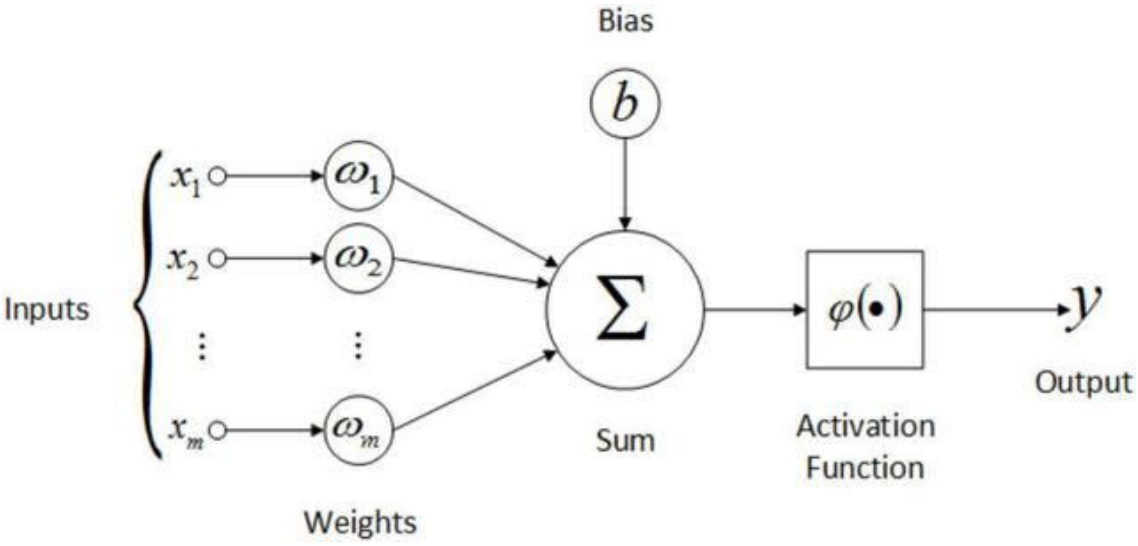
Το Perceptron είναι η δομική μονάδα ενός νευρωνικού δικτύου, η οποία δρα ως γραμμικός ταξινομητής. Η έννοια του Perceptron εισήχθη για πρώτη φορά το 1957 από τον Αμερικανό ψυχολόγο Frank Rosenblatt και βασίζεται στο μη γραμμικό μοντέλο νευρώνα του McCulloch-Pitts.

Όπως φαίνεται και στο Σχήμα 2.1, ο νευρώνας αποτελείται από τρία βασικά στοιχεία:

1. Ένα σύνολο συνάψεων, κάθε μία από τις οποίες δέχεται ένα σήμα εισόδου και το πολλαπλασιάζει επί το δικό της συναπτικό βάρος. Τα βάρη αυτά μπορούν να παίρνουν είτε θετικές είτε αρνητικές τιμές.

2. Έναν αθροιστή, ο ρόλος του οποίου είναι να αθροίζει τα (σταθμισμένα με τα συναπτικά βάρη) σήματα εισόδου.

3. Μία συνάρτηση ενεργοποίησης η οποία περιορίζει το πλάτος του σήματος εξόδου εντός πεπερασμένων ορίων, συνήθως [-1, 1] ή [0, 1] [18].



Σχήμα 2.1: *Η Δομή ενός Τεχνητού Νευρώνα*

Παρατηρώντας την αρχιτεκτονική του νευρώνα, μπορεί να γίνει αντιληπτό πως η έξοδός του y, δοσμένου ενός διανύσματος εισόδου $x = [x_1, x_2, x_3, ..., x_n] \in \mathbb{R}^n$, ενός διανύσματος συναπτικών βαρών $w = [w_1, w_2, w_3, ..., w_n] \in \mathbb{R}^n$ μίας τιμής πόλωσης b και μίας συνάρτησης ενεργοποίησης ϕ, εκφρασμένη με μαθηματικούς όρους, μπορεί να γραφτεί ως εξής:

$$y = \phi(x^T w + b).$$

Η παραπάνω αρχιτεκτονική μπορεί να εφαρμοστεί στην περίπτωση που ο στόχος είναι η ταξινόμηση δειγμάτων σε δύο γραμμικά διαχωρίσιμες κλάσεις. Για το πρόβλημα των n γραμμικά διαχωρίσιμων κλάσεων, μπορούν να χρησιμοποιηθούν n παράλληλα Perceptron ενός επιπέδου. Αντίθετα, στην περίπτωση μη γραμμικά διαχωρίσιμων κλάσεων, το Perceptron ενός επιπέδου δεν μπορεί να λειτουργήσει ως ταξινομητής και υπάρχει ανάγκη χρησιμοποίησης του Perceptron πολλών επιπέδων, το οποίο θα περιγραφεί αναλυτικά στην επόμενη ενότητα.

2.3.3 Πολυεπίπεδα Νευρωνικά Δίκτυα

Η ανάγκη για την εισαγωγή των Πολυεπίπεδων Νευρωνικών Δικτύων, ή αλλιώς Δικτύων Πρόσθιας Τροφοδότησης, προέκυψε από την αδυναμία των Νευρωνικών Δικτύων ενός επιπέδου να λειτουργούν ως ταξινομητές στην περίπτωση των μη γραμμικά διαχωρίσιμων κλάσεων. Σε αντίθεση με τα Νευρωνικά Δίκτυα ενός επιπέδου, τα Πολυεπίπεδα Νευρωνικά Δίκτυα έχουν τη δυνατότητα να προσεγγίσουν μη γραμμικές συναρτήσεις και κατά συνέπεια να χρησιμοποιηθούν για την επίλυση μη γραμμικών προβλημάτων.

Ένα Πολυεπίπεδο Νευρωνικό Δίκτυο, όπως μπορεί κανείς εύκολα να αντιληφθεί και από την ονομασία του, αποτελείται από πολλούς νευρώνες οργανωμένους σε διαδοχικά επίπεδα.

Συνολικά, μπορούμε να πούμε πως ένα τέτοιο δίκτυο χωρίζεται σε τρία βασικά μέρη, όπως μπορούμε να διακρίνουμε και στο Σχήμα 2.2.

- Ένα επίπεδο εισόδου

  Το επίπεδο αυτό είναι υπεύθυνο για την εισαγωγή της εξωτερικής πληροφορίας στο εσωτερικό του δικτύου. Αποτελείται από τόσους νευρώνες όσους και τα σήματα εισόδου και δεν επιτελεί κάποια επεξεργασία πάνω σε αυτά - απλά τα μεταβιβάζει στο επόμενο επίπεδο.

- Ένα σύνολο κρυφών επιπέδων

  Τα κρυφά επίπεδα είναι υπεύθυνα για την εξαγωγή χρήσιμης πληροφορίας από τα σήματα που εισάγονται στο επίπεδο εισόδου. Η έξοδος του κάθε κρυφού επιπέδου τροφοδοτείται σαν είσοδος στο επόμενο κρυφό επίπεδο, εκτός από την έξοδο του τελευταίου κρυφού επιπέδου που τροφοδοτείται στο επίπεδο εξόδου. Στην περίπτωση που το σύνολο κρυφών επιπέδων περιλαμβάνει περισσότερα από ένα επίπεδα, τότε γίνεται λόγος για ένα δίκτυο που εμπίπτει στην κατηγορία των βαθιών νευρωνικών δικτύων. Μάλιστα, στην περίπτωση αυτή, υπάρχει μια ιεραρχία στον τύπο των χαρακτηριστικών που εξάγονται σε κάθε επίπεδο. Πιο συγκεκριμένα, τα αρχικά επίπεδα εξάγουν πιο γενικά χαρακτηριστικά της εισόδου, ενώ καθώς κινούμαστε προς την έξοδο, τα χαρακτηριστικά αυτά γίνονται όλο και πιο συγκεκριμένα και σχετίζονται περισσότερο με το πρόβλημα που καλείται να επιλύσει το εν λόγω νευρωνικό.

- Ένα επίπεδο εξόδου

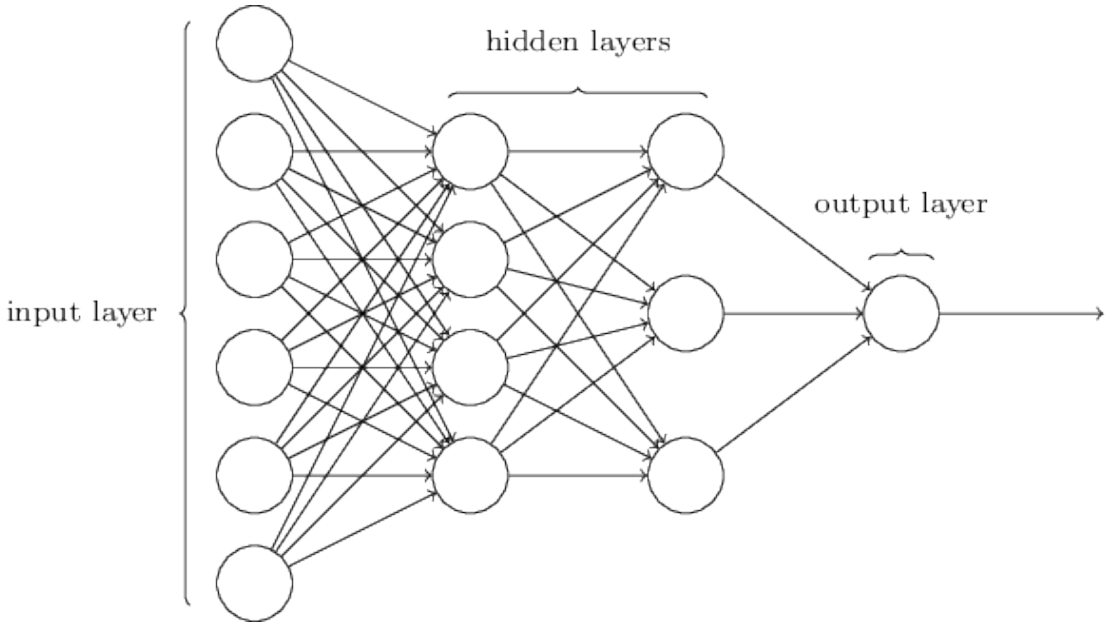
  Στο επίπεδο αυτό αποτελεί τη συνολική έξοδο του νευρωνικού δικτύου, όπου μετά από όλους τους υπολογισμούς που έχουν προηγηθεί στα προηγούμενα επίπεδα, λαμβάνεται η τελική απόφαση του δικτύου. Το επίπεδο εξόδου αποτελείται από τόσους νευρώνες όσες και οι συνιστώσες της επιθυμητής εξόδου.

2.4 Συνελικτικά Νευρωνικά Δίκτυα

Τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks) είναι ένα εξειδικευμένο είδος νευρωνικών δικτύων τα οποία χρησιμοποιούνται για την επεξεργασία δεδομένων με γνωστή, πλεγματοειδή τοπολογία, όπως για παράδειγμα εικόνες ή δεδομένα χρονοσειρών [19].

Η δομή και η λειτουργία αυτού του τύπου νευρωνικών δικτύων διαφέρει σε ορισμένα βασικά σημεία από τα Πολυεπίπεδα Νευρωνικά Δίκτυα που αναλύθηκαν στην προηγούμενη ενότητα:

- Η βασική διαφορά των δύο τύπων δικτύων έγκειται στο γεγονός ότι τα συνελικτικά νευρωνικά δίκτυα μπορούν να λαμβάνουν ως είσοδο τα δεδομένα στην πρωτογενή τους μορφή και να μαθαίνουν από αυτά. Αντίθετα, η είσοδος των πολυεπίπεδων νευρωνικών δικτύων πρέπει να είναι συγκεκριμένα χρήσιμα χαρακτηριστικά που έχουν εξαχθεί από τα πρωτογενή δεδομένα.



Σχήμα 2.2: *Πολυεπίπεδο Νευρωνικό Δίκτυο*

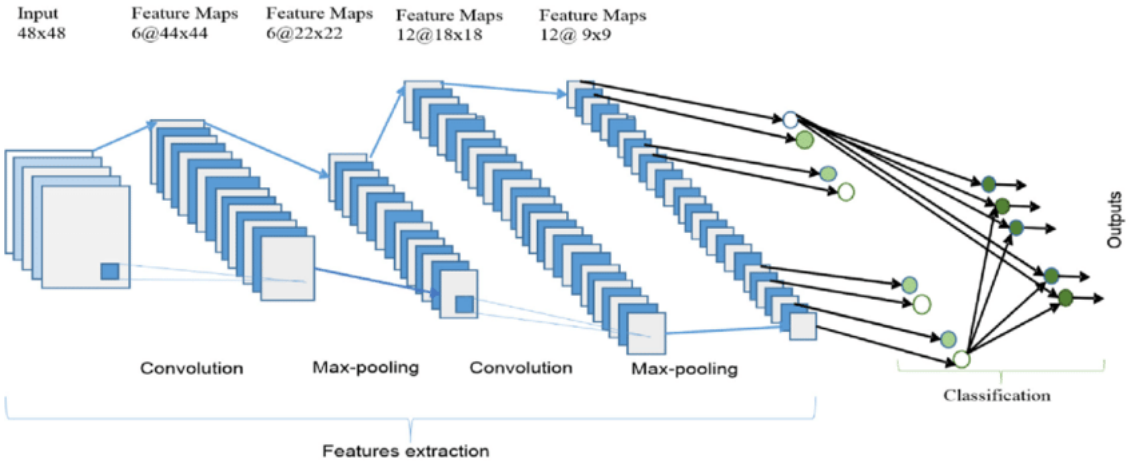
- Μία ακόμα σημαντική διαφορά παρατηρείται στον τρόπο με τον οποίο τα επίπεδα του κάθε τύπου δικτύου συνδέονται μεταξύ τους. Πιο συγκεκριμένα, στα πολυεπίπεδα νευρωνικά δίκτυα, η έξοδος κάθε νευρώνα μεταβιβάζεται σαν είσοδος κάθε νευρώνα του επόμενου επιπέδου. Στα συνελικτικά νευρωνικά δίκτυα, σε κάθε είσοδο ενός επιπέδου διαβιβάζεται μια τιμή η οποία προκύπτει από τη συνέλιξη και σχετίζεται με τη χωρική γειτονιά στην έξοδο του προηγούμενου επιπέδου.

- Διαφοροποίηση μπορεί να εντοπιστεί επίσης στο πλήθος των διαστάσεων της εισόδου. Τα πολυεπίπεδα νευρωνικά δίκτυα έχουν τη δυνατότητα να δέχονται μονοδιάστατη είσοδο, ενώ τα συνελικτικά νευρωνικά δίκτυα δέχονται δισδιάστατη ή τρισδιάστατη είσοδο, ανάλογα με το αν η εικόνες που τροφοδοτούνται στο δίκτυο είναι μονοκαναλικές ή πολυκαναλικές.

- Τέλος, αξίζει να αναφερθεί και η μέθοδος της δειγματοληψίας, η οποία εφαρμόζεται μεταξύ των επιπέδων στην πλειοψηφία των συνελικτικών νευρωνικών δικτύων. Χάρη σε αυτή περιορίζεται η ευαισθησία του δικτύου σε διαφοροποιήσεις της εισόδου που σχετίζονται με τη μετατόπιση [20].

2.4.1 Επίπεδα Επεξεργασίας

Όσον αφορά την αρχιτεκτονική ενός Συνελικτικού Νευρωνικού Δικτύου, μπορούμε να πούμε πως αποτελείται από τέσσερα διαδοχικά επίπεδα. Το κάθε ένα από αυτά επιτελεί μια διαφορετική λειτουργία, με το κυριότερο εξ αυτών να είναι το συνελικτικό επίπεδο. Συνολικά, η πλήρης αρχιτεκτονική ενός Συνελικτικού Νευρωνικού Δικτύου μπορεί να συνοψιστεί στο Σχήμα 2.3. Στη συνέχεια θα εξηγήσουμε αναλυτικά τον τρόπο λειτουργίας του καθενός από αυτά τα επίπεδα.



Σχήμα 2.3: *Συνελικτικό Νευρωνικό Δίκτυο*

Επίπεδο Εισόδου

Το επίπεδο αυτό, όπως είναι εύκολα κατανοητό, είναι υπεύθυνο για την τροφοδότηση των δεδομένων στο εσωτερικό του δικτύου. Οι διαστάσεις του, δηλαδή το μήκος, το πλάτος και το βάθος σε αριθμό νευρώνων, μπορούν να πάρουν διάφορες τιμές και καθορίζονται από τις αντίστοιχες διαστάσεις των δεδομένων.

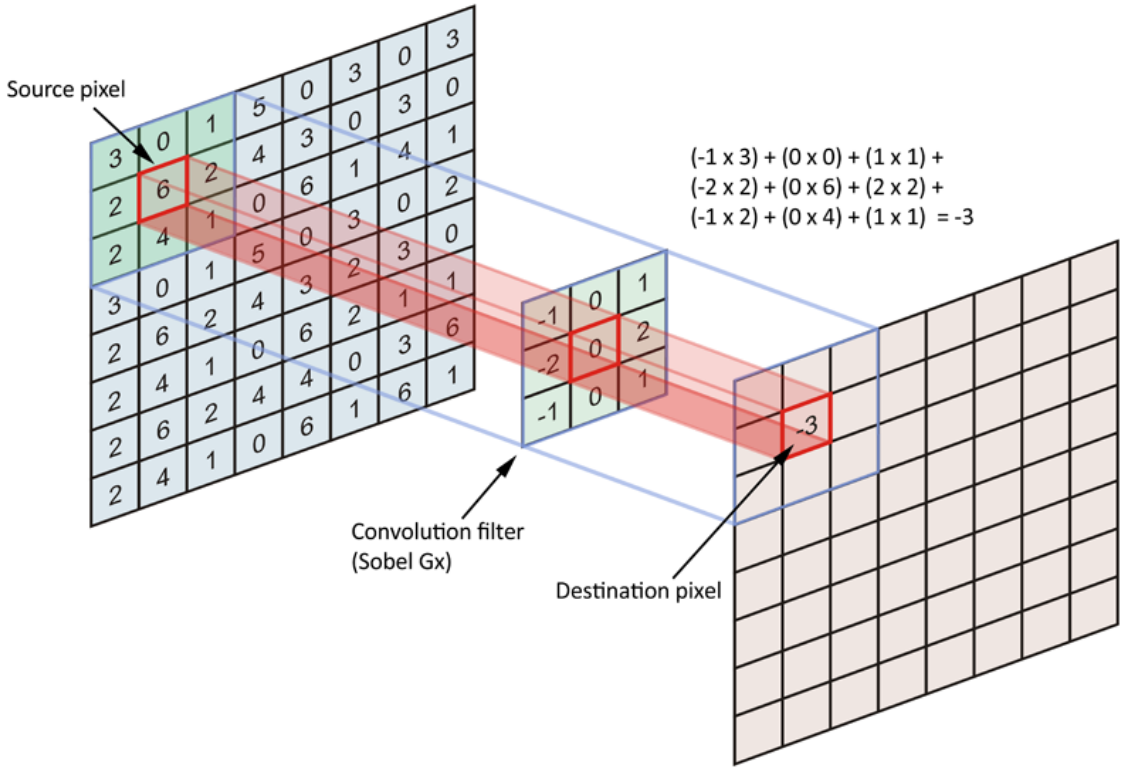
Συνελικτικό Επίπεδο

Το συνελικτικό επίπεδο αποτελεί την καρδιά του συνελικτικού νευρωνικού δικτύου, όπως είναι άλλωστε αναμενόμενο από την ονομασία του. Σε υψηλό επίπεδο, είναι υπεύθυνο για την εξαγωγή χρήσιμων χαρακτηριστικών από τα δεδομένα που δέχεται σαν είσοδο, μέσω της εφαρμογής διάφορων μετασχηματισμών στα δεδομένα αυτά.

Πιο συγκεκριμένα, πραγματοποιείται συνέλιξης μεταξύ ενός πυρήνα (kernel) ή αλλιώς φίλτρου (filter) και των δεδομένων εισόδου. Η συνέλιξη είναι μια μαθηματική πράξη, η οποία στην περίπτωση των πινάκων δύο διαστάσεων μπορεί να μεταφραστεί ως ένα προς ένα πολλαπλασιασμός του φίλτρου επί το τμήμα του δείγματος εισόδου που καλύπτει το φίλτρο, καθώς αυτό ολισθαίνει. Το ολισθαίνον φίλτρο περνά σταδιακά από όλη την επιφάνεια του δείγματος εισόδου σε κάθε βήμα της συνέλιξης, όπως μπορούμε να δούμε στο Σχήμα 2.4

Με αυτό τον τρόπο δημιουργείται ένας Χάρτης Χαρακτηριστικών (Feature Map) ο οποίος περιλαμβάνει το αποτέλεσμα της συνέλιξης του δείγματος εισόδου και του φίλτρου. Ανάλογα με τις υπερπαραμέτρους του δικτύου, υπάρχουν ένα ή περισσότερα φίλτρα, το καθένα εκ των οποίων είναι υπεύθυνο για την εξαγωγή ενός διαφορετικού χαρακτηριστικού και το οποίο εν τέλει δημιουργεί έναν διαφορετικό χάρτη χαρακτηριστικών.

Οι χάρτες χαρακτηριστικών που προκύπτουν στοιβάζονται και έτσι δημιουργείται ένας πίνακας χαρακτηριστικών τριών διαστάσεων, με βάθος ίσο με το πλήθος των διαφορετικών φίλτρων που χρησιμοποιήθηκαν στη συνέλιξη. Το μήκος και το πλάτος του τελικού πίνακα χαρακτηριστικών εξαρτώνται από τις διαστάσεις της εισόδου, τις διαστάσεις του φίλτρου, το βήμα ολίσθησης, καθώς και την ύπαρξη ή μη παραγεμίσματος (padding). Με τον όρο παραγέμισμα, εννοούμε την επέκταση των περιθωρίων του δείγματος εισόδου (συνήθως με



Σχήμα 2.4: *Συνελικτικό Επίπεδο*

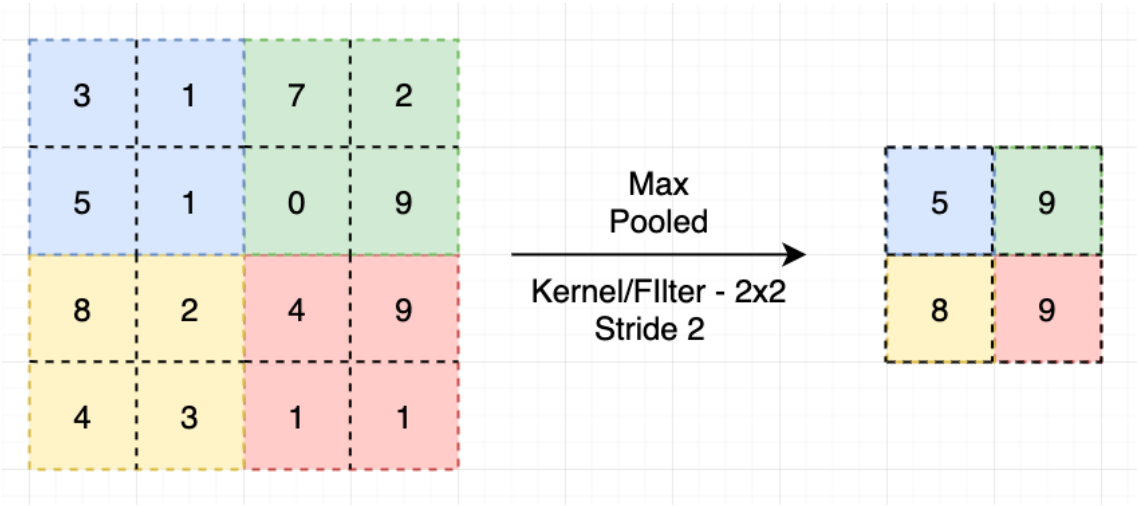
μηδενικές τιμές), ώστε να μην προκύπτει μείωση διαστάσεων λόγω αδυναμίας πλήρους κάλυψης της εισόδου από το φίλτρο όταν το βήμα ολίσθησης είναι μεγαλύτερο από ένα.

Συγκεντρωτικό Επίπεδο

Στις περισσότερες περιπτώσεις, το συνελικτικό επίπεδο (convolutional layer) ακολουθείται από ένα συγκεντρωτικό επίπεδο (pooling layer). Ο ρόλος αυτού του επιπέδου είναι η μείωση των διαστάσεων του μήκους και του πλάτους του χάρτη χαρακτηριστικών που προέκυψε στο προηγούμενο επίπεδο, αφήνοντας ανέπαφη τη διάσταση του βάθους. Ο λόγος για τον οποίο γίνεται αυτή η διαδικασία είναι ώστε να μειωθεί το πλήθος των παραμέτρων του συστήματος και κατά συνέπεια ο χρόνος εκπαίδευσης, ενώ ταυτόχρονα αντιμετωπίζεται και το πρόβλημα της υπερπροσαρμογής.

Η μείωση των διαστάσεων γίνεται ως εξής: Αντίστοιχα με τη συνέλιξη, έχουμε και πάλι ένα ολισθαίνον παράθυρο, του οποίου η χρησιμότητα είναι να ομαδοποιεί τις γειτονικές τιμές του χάρτη χαρακτηριστικών. Στη συνέχεια, από αυτές τις γειτονικές τιμές συνήθως είτε επιλέγεται η μέγιστη (max pooling) όπως βλέπουμε και στο Σχήμα 2.5, είτε υπολογίζεται ο μέσος όρος αυτών (average pooling).

Το μέγεθος του παραθύρου, καθώς και το βήμα ολίσθησης ανήκουν στις υπερπαραμέτρους που ορίζονται από το χρήστη και καθορίζουν το μέγεθος του χάρτη χαρακτηριστικών στην έξοδο του επιπέδου.



Σχήμα 2.5: *Συγκεντρωτικό Επίπεδο*

Πλήρως Συνδεδεμένο Επίπεδο

Ύστερα από μία σειρά επαναλαμβανόμενων συνελικτικών και συγκεντρωτικών επιπέδων, ακολουθεί μία ομάδα από ένα ή περισσότερα πλήρως συνδεδεμένα επίπεδα. Ο ρόλος αυτών των επιπέδων είναι η ταξινόμηση σε μία ή περισσότερες κλάσεις. Ένα πλήρως συνδεδεμένο δίκτυο είναι επί της ουσίας ένα πολυεπίπεδο νευρωνικό δίκτυο, όπου κάθε νευρώνας ενός επιπέδου συνδέεται με όλους τους νευρώνες του επόμενου επιπέδου. Ένα τέτοιο δίκτυο μπορεί να διαχειριστεί μόνο μονοδιάστατη είσοδο, επομένως, ως ένα ενδιάμεσο βήμα μεταξύ του συγκεντρωτικού και του πλήρως συνδεδεμένου επιπέδου, η έξοδος του συγκεντρωτικού επιπέδου επιπεδώνεται (flattened) ώστε να είναι συμβατή με το επόμενο επίπεδο, ενώ ταυτόχρονα εξασφαλίζεται και ότι δεν υπάρχει απώλεια πληροφορίας.

2.5 Αναδρομικά Νευρωνικά Δίκτυα

Τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks) είναι και αυτά ένα εξειδικευμένο είδος νευρωνικών δικτύων τα οποία χρησιμοποιούνται για την επεξεργασία δεδομένων με ακολουθιακή μορφή, όπως είναι για παράδειγμα τα δεδομένα χρονοσειρών [19]. Η χαρακτηριστική τους ιδιότητα είναι η ‘‘μνήμη’’, καθώς, σε αντίθεση με τα είδη νευρωνικών δικτύων που έχουμε δει ως τώρα, η έξοδος των Αναδρομικών Νευρωνικών Δικτύων εξαρτάται και από τα προηγούμενες τιμές της ακολουθίας εισόδου και όχι μόνο από την τρέχουσα τιμή [21].

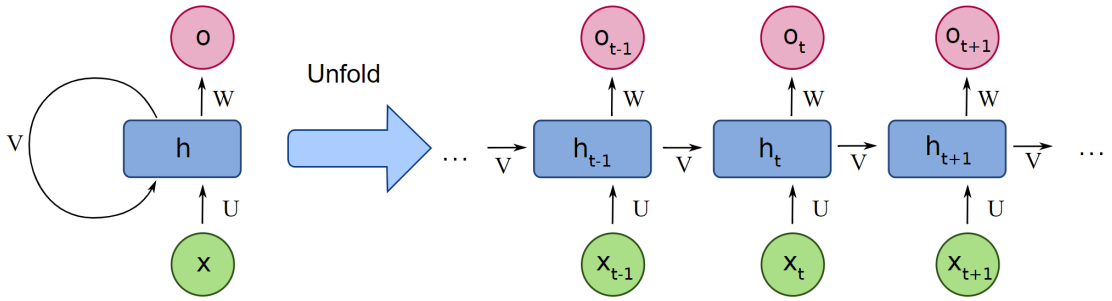
Η ιδιότητα της ‘‘μνήμης’’ οφείλεται στην δομή των Αναδρομικών Νευρωνικών Δικτύων και πιο συγκεκριμένα στο διάνυσμα κρυφής κατάστασης που διαθέτουν. Όπως μπορούμε να διακρίνουμε στο Σχήμα 2.6, κάθε επόμενη κρυφή κατάσταση h_{i+1} προκύπτει από την προηγούμενη κρυφή κατάσταση h_i και την τρέχουσα είσοδο x_{i+1}. Με αυτόν τον τρόπο το RNN μπορεί να θυμάται το περιεχόμενο της εισόδου που έχει δεχθεί ήδη κατά την διάρκεια της εκπαίδευσης.

Με μαθηματικούς όρους, αυτό μπορεί να εκφραστεί ως εξής:

$$h_t = f(Vh_{t-1} + Ux_t + b_h) \tag{2.1}$$

$$o_t = g(Wh_t + b_o) \tag{2.2}$$

Η σχέση 2.1 μας δίνει την κρυφή κατάσταση του δικτύου τη χρονική στιγμή t, h_t, με βάση την κρυφή κατάσταση του δικτύου τη χρονική στιγμή $t-1$, h_{t-1}, το διάνυσμα εισόδου τη χρονική στιγμή t, x_t, και την πόλωση b_h. Η σχέση 2.2 μας δίνει την έξοδο του δικτύου τη χρονική στιγμή t, o_t, με βάση την κρυφή κατάσταση την ίδια χρονική στιγμή, h_t και την πόλωση b_o. Υπάρχουν τρεις διαφορετικοί πίνακες βαρών: U για τα βάρη από την είσοδο στο κρυφό επίπεδο, V για τα βάρη από το ένα κρυφό επίπεδο στο άλλο και W για τα βάρη από το κρυφό επίπεδο προς την έξοδο. Επιπλέον, οι f και g είναι συναρτήσεις ενεργοποίησης για την κρυφή κατάσταση και για την έξοδο αντίστοιχα.



Σχήμα 2.6: *Αναδρομικό νευρωνικό δίκτυο σε συμπτυγμένη και αναπτυγμένη μορφή.*

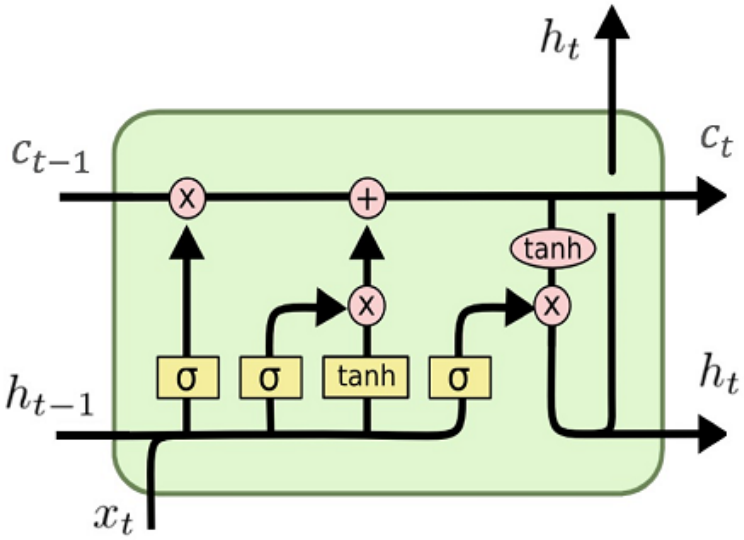
Παρά τα σημαντικά πλεονεκτήματα τον απλών αναδρομικών δικτύων στην παραγωγή προβλέψεων λαμβάνοντας υπόψιν τις χρονικές εξαρτήσεις μεταξύ των δειγμάτων της εισόδου, η απόδοσή τους δεν είναι ικανοποιητική όταν τα δείγματα αυτά απέχουν κατά μεγάλη χρονική απόσταση. Η συγκεκριμένη αδυναμία αποτελεί σημαντικό εμπόδιο στην επίλυση των προβλημάτων όπου η τρέχουσα είσοδος εξαρτάται και από εισόδους που δεν βρίσκονται πολύ κοντά χρονικά σε αυτή, όπως είναι για παράδειγμα η πρόβλεψη του μουσικού είδους ενός ηχητικού κομματιού, που καλούμαστε να αντιμετωπίσουμε στην παρούσα εργασία. Για το σκοπό αυτό, η αρχιτεκτονική των δικτύων αυτών έχει εξελιχθεί με σκοπό να καλυφθεί και αυτή η ανάγκη, όπως θα δούμε στις αμέσως επόμενες ενότητες.

2.5.1 Νευρωνικά Δίκτυα Long Short-Term Memory

Όπως ήδη αναφέρθηκε, τα απλά αναδρομικά νευρωνικά δίκτυα δεν έχουν τη δυνατότητα να λάβουν υπόψιν πληροφορία μεταξύ δειγμάτων εισόδου που απέχουν αρκετά μεταξύ τους. Αυτό συμβαίνει διότι σε αυτή την περίπτωση, κατά τη διάρκεια του αλγορίθμου οπίσθιας διάδοσης (backpropagation), οι κλίσεις (gradients) τείνουν προς το μηδέν (vanishing gradients) ή προς το άπειρο (exploding gradients).

Το πρόβλημα αυτό λύνεται από τα δίκτυα LSTM μέσω μηχανισμών που ονομάζονται πύλες και ελέγχουν την ροή της πληροφορίας. Πιο συγκεκριμένα, ένα κελί LSTM αποτελείται από την πύλη απώλειας (forget gate) η οποία ελέγχει ποιες πληροφορίες θα διατηρηθούν και ποιες θα διαγραφούν - ξεχαστούν, καθώς και από τις πύλες εισόδου και εξόδου, οι οποίες είναι υπεύθυνες για το ποιες πληροφορίες θα εισαχθούν στο εσωτερικό του κελιού ή θα

εξαχθούν στο εξωτερικό περιβάλλον, αντίστοιχα. Επιπλέον, ένα κελί LSTM περιλαμβάνει και την κατάσταση κελιού (cell state) πέρα από την κρυφή κατάσταση (hidden state).



Σχήμα 2.7: *Κελί LSTM*

Οι μαθηματικές εξισώσεις που διέπουν τη λειτουργία του κελιού LSTM (όπως περιγράφονται και από το Σχήμα 2.7 είναι οι εξής:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \tag{2.3}$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \tag{2.4}$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \tag{2.5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \tag{2.6}$$

$$h_t = o_t \odot \sigma_h(c_t) \tag{2.7}$$

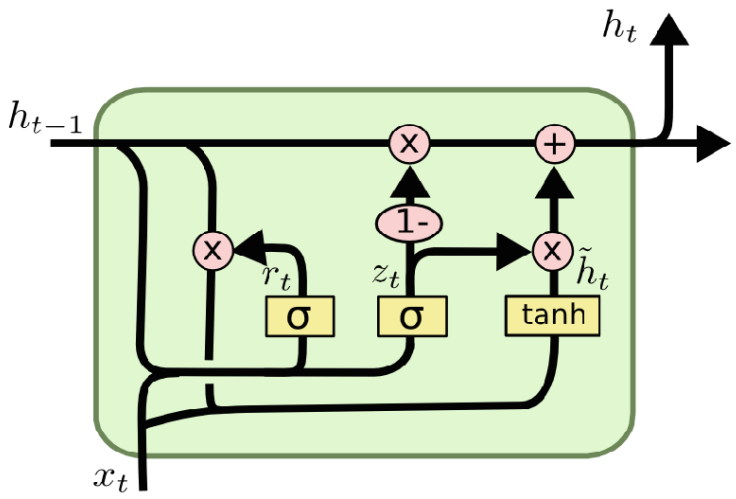
Όπου x_t το διάνυσμα εισόδου, f_t το διάνυσμα ενεργοποίησης της πύλης απώλειας, i_t το διάνυσμα ενεργοποίησης της πύλης εισόδου, o_t το διάνυσμα ενεργοποίησης της πύλης εξόδου, h_t το διάνυσμα κρυφής κατάστασης, c_t το διάνυσμα κατάστασης κελιού. Επιπλέον, τα W, U είναι οι πίνακες βαρών και τα b είναι τα διανύσματα πόλωσης. Τέλος, σ είναι η συναρτήσεις ενεργοποίησης και $\odot$ είναι ο τελεστής Hadamard που εκφράζει τον πολλαπλασιασμό πινάκων ένα-προς-ένα.

2.5.2 Νευρωνικά Δίκτυα Gated Recurrent Unit

Όπως μπορεί κανείς εύκολα να υποψιαστεί από το πλήθος και την περιπλοκότητα των μαθηματικών σχέσεων που διέπουν τη λειτουργία του κελιού LSTM, η επεξεργασία δεδομένων

στις εν λόγω μονάδες είναι ιδιαίτερα ακριβή υπολογιστικά και χρονοβόρα. Το πρόβλημα αυτό έρχονται να λύσουν οι μονάδες GRU, οι οποίες διατηρούν την αποτελεσματικότητα των μονάδων LSTM μειώνοντας ταυτόχρονα σε μεγάλο βαθμό την περιπλοκότητά τους.

Πιο συγκεκριμένα, η απλοποίηση προκύπτει από τον τρόπο που ορίζονται οι πύλες του κελιού GRU. Σε αντίθεση με τις τρεις πύλες του κελιού LSTM, το κελί GRU αποτελείται από δύο πύλες: την πύλη ανανέωσης (update gate) η οποία είναι υπεύθυνη για το ποιες πληροφορίες θα εισαχθούν στο εσωτερικό του κελιού, όπως ακριβώς η πύλη εισόδου στο κελί LSTM, και την πύλη επαναφοράς (reset gate) η οποία αντιστοιχεί στο συνδυασμό των πυλών εξόδου και απώλειας του κελιού LSTM και είναι υπεύθυνη για τον όγκο πληροφορίας που η μονάδα πρέπει να διαγράψει.



Σχήμα 2.8: *Κελί GRU*

Οι μαθηματικές εξισώσεις που διέπουν τη λειτουργία του κελιού GRU (όπως περιγράφονται και από το Σχήμα 2.8 είναι οι εξής:

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z) \tag{2.8}$$

$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r) \tag{2.9}$$

$$\hat{h}_t = \phi_h(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{2.10}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \hat{h}_t \tag{2.11}$$

Όπου x_t το διάνυσμα εισόδου, h_t το διάνυσμα εξόδου, $\hat{h}_t$ το υποψήφιο διάνυσμα ενεργοποίησης, z_t το διάνυσμα της πύλης ανανέωσης, r_t το διάνυσμα της πύλης επαναφοράς. Επιπλέον, τα W, U είναι οι πίνακες βαρών και τα b είναι τα διανύσματα πόλωσης. Τέλος, σ και ϕ είναι η συναρτήσεις ενεργοποίησης και $\odot$ είναι ο τελεστής Hadamard που εκφράζει τον πολλαπλασιασμό πινάκων ένα-προς-ένα.

2.6 Εκπαίδευση Νευρωνικών Δικτύων

2.6.1 Συνάρτηση Κόστους

Η συνάρτηση κόστους ορίζεται ως ένας μετρήσιμος τρόπος αξιολόγησης των επιδόσεων ενός αλγορίθμου. Αυτό είναι αναγκαίο ώστε να καθοριστεί η απόσταση μεταξύ της τρέχουσας εξόδου του αλγορίθμου και της επιθυμητής εξόδου. Η μέτρηση αυτή χρησιμοποιείται σαν ένας μηχανισμός ανάδρασης ώστε να επαναπροσαρμοστεί ο τρόπος λειτουργίας του αλγορίθμου. Αυτή ακριβώς η επαναπροσαρμογή της λειτουργίας του αλγορίθμου ονομάζεται μάθηση [22]. Με άλλα λόγια, η συνάρτηση κόστους εκφράζει το σφάλμα του αλγορίθμου κατά την εκτέλεση της λειτουργίας για την οποία έχει δημιουργηθεί. Συνεπώς, όπως είναι αναμενόμενο, η τιμή του σφάλματος λαμβάνει πραγματικές τιμές και πρέπει να είναι κάτω φραγμένη ώστε να μπορεί να ελαχιστοποιηθεί. Ωστόσο, υπάρχουν και ορισμένες περιπτώσεις όπου η συνάρτηση κόστους είναι άνω φραγμένη και στόχος είναι η μεγιστοποίησή της.

Υπάρχουν αρκετές διαφορετικές συναρτήσεις κόστους, ανάλογα με το πρόβλημα που καλείται να επιλύσει ο αντίστοιχος αλγόριθμος. Με αυτή τη λογική, θα μπορούσαμε να διαχωρίσουμε τις συναρτήσεις κόστους σε δύο βασικές κατηγορίες: σε αυτές που χρησιμοποιούνται σε προβλήματα ταξινόμησης και σε αυτές που χρησιμοποιούνται σε προβλήματα παλινδρόμησης. Δεδομένου ότι το θέμα της παρούσας εργασίας είναι ένα πρόβλημα ταξινόμησης σε πολλές κλάσεις, αξίζει να γίνει μια λεπτομερέστερη αναφορά σε μία από τις πιο ευρέως χρησιμοποιούμενες συναρτήσεις κόστους αυτής της κατηγορίας: στην συνάρτηση κατηγορικής διασταυρούμενης εντροπίας (categorical cross-entropy).

Η συνάρτηση αυτή χρησιμοποιείται για να υπολογίσει την απόσταση ανάμεσα σε δύο διαφορετικές διακριτές κατανομές πιθανότητας. Με μαθηματικούς όρους ορίζεται ως εξής:

$$Loss = -\sum_{i=1}^{\text{output size}} y_i \cdot \log \hat{y}_i \tag{2.12}$$

Όπου, $\hat{y}_i$ είναι η i-οστή έξοδος του αλγορίθμου, y_i η αντίστοιχη επιθυμητή έξοδος και output size είναι το πλήθος των διακριτών τιμών στην έξοδο του αλγορίθμου.

2.6.2 Αλγόριθμος Οπισθοδιάδοσης (Backpropagation)

Ο αλγόριθμος οπισθοδιάδοσης (Backpropagation) είναι το πιο ουσιαστικό μέρος της διαδικασίας εκπαίδευσης ενός νευρωνικού δικτύου. Είναι η διαδικασία προσαρμογής των παραμέτρων του δικτύου, βάσει της συνάρτησης κόστους που αναλύθηκε στην προηγούμενη ενότητα. Η κατάλληλη προσαρμογή των παραμέτρων είναι υψίστης σημασίας γιατι εξασφαλίζει την αξιοπιστία του αλγορίθμου, αυξάνοντας την ικανότητά του να γενικεύει.

Ο αλγόριθμος αυτός εκτελείται μετά από τη διαδικασία της εμπρόσθιας διάδοσης (forward propagation), κατά την οποία μία είσοδος τροφοδοτείται στο δίκτυο και μετά από μία σειρά υπολογισμών προκύπτει η αντίστοιχη έξοδος. Έπειτα, εφόσον γνωρίζουμε την επιθυμητή έξοδο, υπολογίζεται το σφάλμα L, με τη βοήθεια της συνάρτησης κόστους. Κατά την εκτέλεση του αλγορίθμου οπισθοδιάδοσης, υπολογίζονται οι μερικές παράγωγοι του σφάλματος ως προς κάθε παράμετρο του δικτύου, ξεκινώντας από την έξοδο του δικτύου και προχωρώντας προς

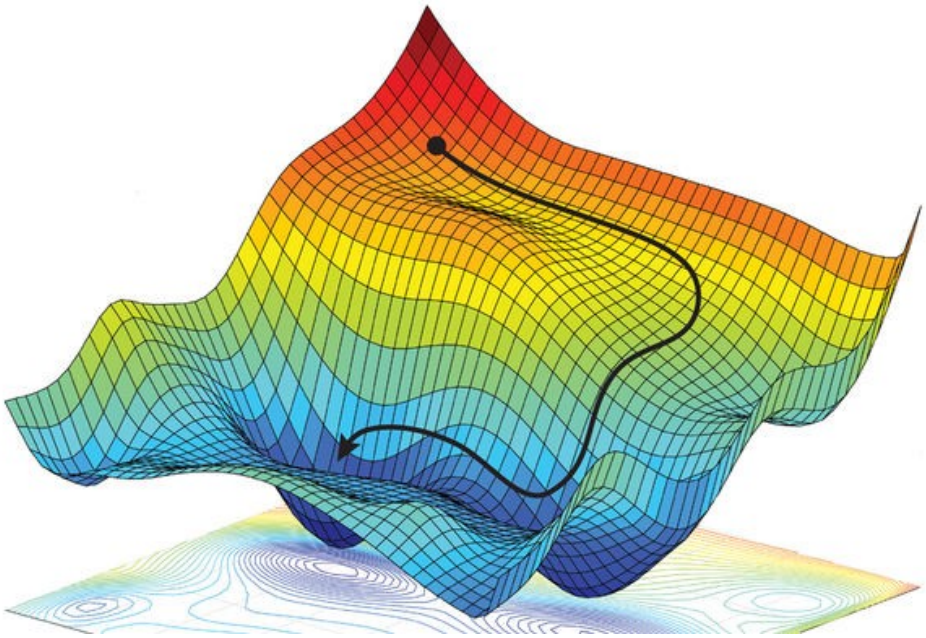
την είσοδο, με τη χρήση του κανόνα της αλυσίδας. Στη συνέχεια, οι μερική παράγωγος $\frac{\partial L(\Theta)}{\partial \Theta}$ ως προς την κάθε παράμετρο Θ θα αξιοποιηθεί για την κατάλληλη προσαρμογή της αντίστοιχης παραμέτρου, με βάση τον αλγόριθμο κατάβασης κλίσης, ο οποίος παρουσιάζεται στην επόμενη ενότητα.

2.6.3 Αλγόριθμος Κατάβασης Κλίσης (Gradient Descent)

Ο αλγόριθμος κατάβασης κλίσης, όπως φαίνεται γραφικά στην εικόνα 2.9 είναι μια επαναληπτική μέθοδος βελτιστοποίησης που χρησιμοποιείται για την εύρεση των παραμέτρων μίας συνάρτησης f η οποία ελαχιστοποιεί μία συνάρτηση κόστους $L(\Theta)$. Σύμφωνα με την απλή εκδοχή του αλγορίθμου, μία παράμετρος Θ ενημερώνεται με βάση την παρακάτω μαθηματική σχέση:

$$\Theta_{t+1} = \Theta_t - \rho \cdot \frac{\partial L(\Theta)}{\partial \Theta} \tag{2.13}$$

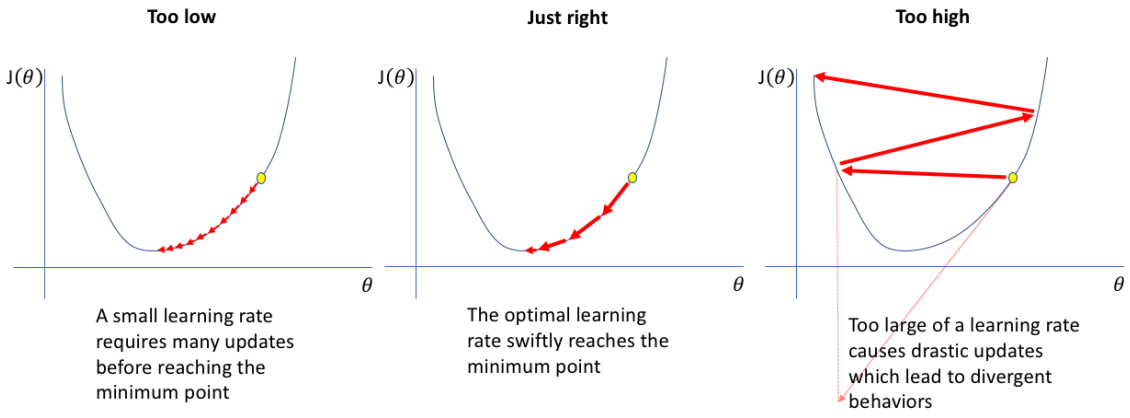
Όπου Θ_{t+1} η ενημερωμένη τιμή της παραμέτρου Θ, Θ_t η τιμή της παραμέτρου Θ στο τρέχον βήμα της επαναληπτικής διαδικασίας, $\frac{\partial L(\Theta)}{\partial \Theta}$ η μερική παράγωγος της συνάρτησης κόστους L ως προς την παράμετρο Θ όπως αυτή υπολογίστηκε από τον αλγόριθμο οπισθοδιάδοσης και ρ είναι ο ρυθμός μάθησης (learning rate) της διαδικασίας εκπαίδευσης, για τον οποίο ισχύει $0 \leq \rho \leq 1$.



Σχήμα 2.9: *Αλγόριθμος κατάβασης κλίσης.*

Όπως μπορούμε να διακρίνουμε και στην εικόνα 2.10, η τιμή του ρυθμού μάθησης παίζει σημαντικό ρόλο στο αν και πόσο γρήγορα τελικά ο αλγόριθμος θα καταφέρει να βρει την ελάχιστη τιμή της συνάρτησης. Ένας πολύ μικρός ρυθμός μάθησης μπορεί να καθυστερήσει σημαντικά τη σύγκλιση του αλγορίθμου, ενώ ένας μεγάλος ρυθμός μάθησης μπορεί να οδηγήσει στην πλήρη αποτυχία εύρεσης του ελαχίστου.

Στην πράξη, υπάρχουν διαφορετικές εκδοχές του αλγορίθμου κατάβασης κλίσης όσον αφο-



Σχήμα 2.10: *Επίδραση του ρυθμού μάθησης στον αλγόριθμο κατάβασης κλίσης.*

ρά τη χρονική στιγμή της ενημέρωσης των παραμέτρων σε κάθε επανάληψη. Οι σημαντικότερες εξ αυτών είναι οι εξής:

- Batch Gradient Descent

  Στην περίπτωση αυτή, η ενημέρωση γίνεται όταν το σύνολο των δεδομένων εκπαίδευσης περάσει από τη φάση της πρόσθιας και οπίσθιας διάδοσης, οπότε η συνάρτηση κόστους και συνεπώς η μερικές παράγωγοί της υπολογίζονται για το σύνολο των δεδομένων εκπαίδευσης. Η αδυναμία της συγκεκριμένης μεθόδου έγκειται στην εξαιρετικά χρονοβόρα διαδικασία εκπαίδευσης, αλλά και σε επιπλέον πρακτικά προβλήματα καθώς σε αυτή την περίπτωση το σύνολο των δεδομένων εκπαίδευσης πρέπει να είναι ταυτόχρονα στη μνήμη - πράγμα πολλές φορές αδύνατο λόγω του όγκου των δεδομένων.

- Stochastic Gradient Descent

  Στην περίπτωση αυτή, η ενημέρωση γίνεται κάθε φορά που ένα δείγμα του συνόλου δεδομένων εκπαίδευσης περάσει από τη φάση της πρόσθιας και οπίσθιας διάδοσης, οπότε η συνάρτηση κόστους και συνεπώς η μερικές παράγωγοί της υπολογίζονται μόνο για το συγκεκριμένο δείγμα εκπαίδευσης. Η μέθοδος αυτή μπορεί να βοηθήσει τον αλγόριθμο να βρει νέα τοπικά ελάχιστα λόγω των συνεχών ενημερώσεων, αλλά αυτό ταυτόχρονα μπορεί να καθυστερήσει σημαντικά τη διαδικασία σύγκλισης.

- Mini Batch Gradient Descent

  Στην περίπτωση αυτή, η οποία είναι η ενδιάμεση των δύο προαναφερθέντων, το σύνολο των δεδομένων εκπαίδευσης χωρίζεται σε επιμέρους τμήματα - παρτίδες. Η ενημέρωση γίνεται όταν μία παρτίδα περάσει από τη φάση της πρόσθιας και οπίσθιας διάδοσης, οπότε η συνάρτηση κόστους και συνεπώς η μερικές παράγωγοί της υπολογίζονται για το συγκεκριμένο υποσύνολο του συνόλου εκπαίδευσης.

2.6.4 Αλγόριθμοι Βελτιστοποίησης

Οι αλγόριθμοι βελτιστοποίησης είναι μέθοδοι οι οποίες χρησιμεύουν στο να καθορίσουν τον τρόπο με τον οποίο θα αλλάξουν οι παράμετροι μιας συνάρτησης με στόχο την ελαχιστοποίηση

μιας συνάρτησης κόστους. Ο αλγόριθμος κατάβασης κλίσης που παρουσιάστηκε αναλυτικά στην προηγούμενη ενότητα είναι μια περίπτωση αλγορίθμου βελτιστοποίησης και συγκεκριμένα είναι ένας αλγόριθμος βελτιστοποίησης πρώτης τάξης.

Ο διαχωρισμός των αλγορίθμων βελτιστοποίησης σε Πρώτης και Δεύτερης Τάξης γίνεται αναλόγως με την τάξη της μερικής παραγώγου της συνάρτησης κόστους ως προς μία παράμετρο που χρησιμοποιείται για τον υπολογισμό της νέας τιμής της εν λόγω παραμέτρου. Πιο συγκεκριμένα, οι αλγόριθμοι βελτιστοποίησης πρώτης τάξης χρησιμοποιούν την τιμή της πρώτης παραγώγου, ενώ αντίστοιχα οι αλγόριθμοι βελτιστοποίησης δεύτερης τάξης χρησιμοποιούν την τιμή της δεύτερης παραγώγου. Οι τελευταίοι αν και περισσότερο αποτελεσματικοί (καθώς λαμβάνουν υπόψιν και την κυρτότητα πέρα από την κλίση) είναι πολύ ακριβοί υπολογιστικά και έτσι δεν χρησιμοποιούνται πολύ συχνά στην πράξη.

Ορμή (Momentum)

Ένα από τα προβλήματα που αντιμετωπίζουν συχνά οι αλγόριθμοι βελτιστοποίησης είναι το φαινόμενο των ταλαντώσεων γύρω από τοπικά ελάχιστα. Η έννοια της ορμής έρχεται να παρακάμψει αυτό το πρόβλημα, μειώνοντας ταυτόχρονα τον χρόνο που απαιτείται για τη σύγκλιση του αλγορίθμου βελτιστοποίησης. Με μαθηματικούς όρους, ο αλγόριθμος κατάβασης κλίσης με ορμή γράφεται ως εξής:

$$\Theta = \Theta - V(t) \tag{2.14}$$

$$V(t) = \gamma V(t-1) + r\nabla L(\Theta) \tag{2.15}$$

Εκτός από τον αλγόριθμο κατάβασης κλίσης με ορμή, υπάρχουν και άλλοι αλγόριθμοι βελτιστοποίησης που ενσωματώνουν την έννοια της ορμής. Οι βασικότεροι εξ αυτών είναι οι εξής:

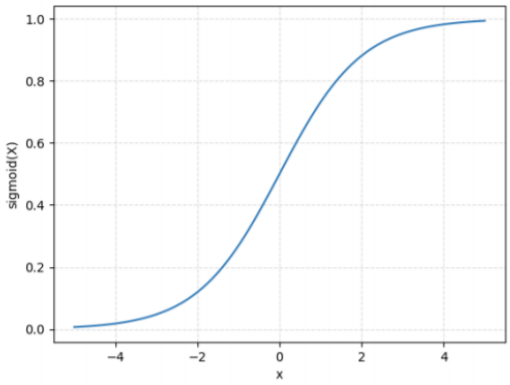
- Adagrad [23]
- Adadelta [24]
- Adam [25]

2.6.5 Συνάρτηση Ενεργοποίησης

Στις προηγούμενες ενότητες έχει αναφερθεί αρκετές φορές η έννοια της συνάρτησης ενεργοποίησης, ως βήμα κατά την επεξεργασία που λαμβάνει χώρα στο εσωτερικό ενός νευρώνα. Ο ρόλος της συνάρτησης ενεργοποίησης είναι να φράσει την έξοδο του νευρώνα μεταξύ συγκεκριμένων ορίων και να εισάγει τη μη γραμμικότητα στο σύστημα. Η ονομασία της εν λόγω συνάρτησης δεν είναι τυχαία, καθώς, δεδομένου του περιορισμού της τιμής εξόδου του νευρώνα εντός ενός συγκεκριμένου διαστήματος, καθορίζεται αν ο εν λογω νευρώνας θα μεταδώσει μια σημαντική τιμή στον επόμενο (δηλαδή θα ενεργοποιηθεί) ή μια αμελητέα τιμή (δηλαδή θα απενεργοποιηθεί). Μερικές από τις πιο συνηθισμένες συναρτήσεις ενεργοποίησης, μαζί με τους μαθηματικούς τους τύπους και τις αντίστοιχες γραφικές αναπαραστάσεις είναι οι ακόλουθες:

Σιγμοειδής συνάρτηση (Sigmoid)

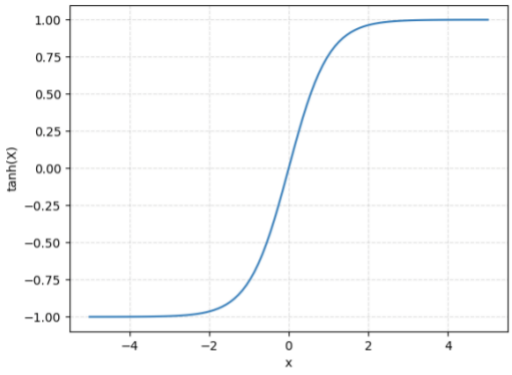
$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.16}$$



Σχήμα 2.11: *Σιγμοειδής συνάρτηση.*

Συνάρτηση υπερβολικής εφαπτομένης (Hyperbolic Tangent)

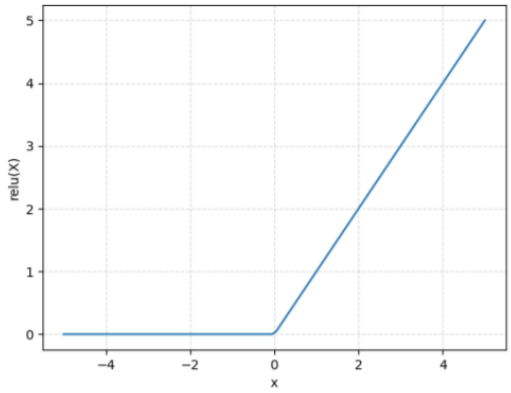
$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.17}$$



Σχήμα 2.12: *Συνάρτηση υπερβολικής εφαπτομένης.*

Συνάρτηση Rectified Linear Unit - ReLU

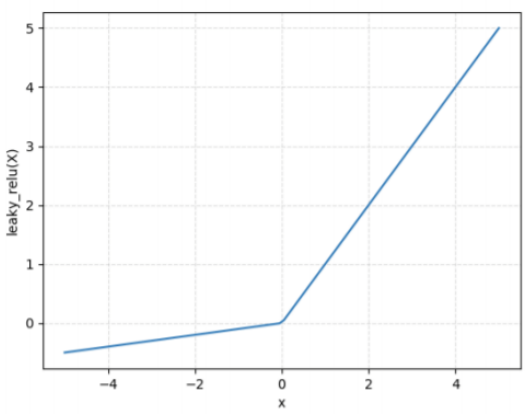
$$f(x) = max(0, x) \tag{2.18}$$



Σχήμα 2.13: *Συνάρτηση Rectified Linear Unit - ReLU.*

Συνάρτηση Leaky Rectified Linear Unit - Leaky ReLU

$$f(x) = \begin{cases} x & x > 0 \\ ax & x \leq 0 \end{cases} \tag{2.19}$$



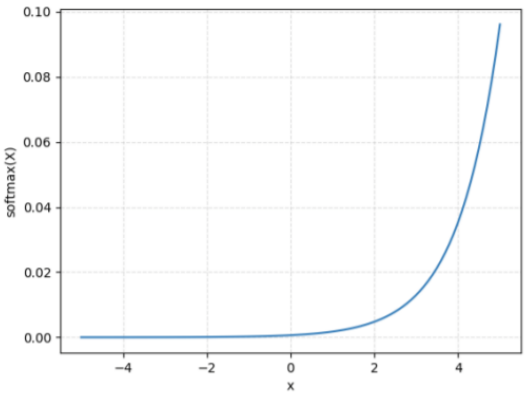
Σχήμα 2.14: *Συνάρτηση Rectified Linear Unit - ReLU.*

Συνάρτηση Softmax

$$\sum_{k=1}^{K} y_k = \sum_{k=1}^{K} Softmax(z)_k = 1 \tag{2.20}$$

2.7 Αξιολόγηση Επιδόσεων

Αφού η διαδικασία της εκπαίδευσης φτάσει στο τέλος της, ακολουθεί η αξιολόγηση των επιδόσεων του συστήματος που δημιουργήθηκε. Αυτή η διαδικασία είναι αναγκαία ώστε να μπορέσουν να εντοπιστούν τα δυνατά σημεία αλλά και οι αδυναμίες του μοντέλου. Επιπλέον, δεδομένου ότι με την διαδικασία αξιολόγησης προκύπτουν τόσο ποιοτικά όσο και ποσοτικά



Σχήμα 2.15: *Συνάρτηση Softmax.*

αποτελέσματα, καθίσταται δυνατή η σύγκριση διαφορετικών προσεγγίσεων του προβλήματος που καλούμαστε να λύσουμε πάνω σε μια κοινή βάση. Ωστόσο για να γίνει αυτό, είναι απαραίτητο η αξιολόγηση να λαμβάνει χώρα μέσα σε ένα σαφώς ορισμένο κλειστό πλαίσιο. Αυτό σημαίνει ότι τόσο οι μετρικές αξιολόγησης όσο και το σύνολο των δεδομένων αξιολόγησης πάνω στο οποίο αυτές υπολογίζονται πρέπει να είναι κοινά για κάθε κύκλο αξιολόγησης που πραγματοποιείται.

2.7.1 Μετρικές Αξιολόγησης

Ανάλογα με τον τύπο του προβλήματος που καλούμαστε να επιλύσουμε, γίνεται αντίστοιχη επιλογή των μετρικών αξιολόγησης που χρησιμοποιούνται. Εν προκειμένω, εφόσον ο στόχος της παρούσας διπλωματικής εργασίας είναι ένα πρόβλημα ταξινόμησης, θα παρουσιαστούν οι μετρικές αξιολόγησης που βρίσκουν εφαρμογή στο συγκεκριμένο ζήτημα. Πρώτα όμως θα ορίσουμε μερικές έννοιες που θα χρησιμοποιηθούν με τη σειρά τους στον ορισμό των μετρικών αξιολόγησης:

- TP - True Positives

  Είναι το πλήθος των δειγμάτων που προβλέφθηκαν ως δείγματα της θετικής κλάσης και είναι όντως δείγματα της θετικής κλάσης.

- FP - False Positives

  Είναι το πλήθος των δειγμάτων που προβλέφθηκαν ως δείγματα της θετικής κλάσης ενώ στην πραγματικότητα δεν ήταν δείγματα της θετικής κλάσης.

- TN - True Negatives

  Είναι το πλήθος των δειγμάτων που προβλέφθηκαν ως δείγματα της αρνητικής κλάσης και είναι όντως δείγματα της αρνητικής κλάσης.

- FN - False Negatives

  Είναι το πλήθος των δειγμάτων που προβλέφθηκαν ως δείγματα της αρνητικής κλάσης ενώ στην πραγματικότητα δεν ήταν δείγματα της αρνητικής κλάσης.

Προχωρώντας θα ορίσουμε τις πιο κοινές μετρικές αξιολόγησης για προβλήματα ταξινόμησης σε 2 κλάσεις:

Ακρίβεια (Precision)

Η Ακρίβεια (Precision) μας δείχνει τι ποσοστό των δείγματων που κατηγοριοποιήθηκαν ως δείγματα της θετικής κλάσης είναι όντως δείγματα της θετικής κλάσης. Με απλά λόγια, υψηλή Ακρίβεια σημαίνει πως μπορούμε να εμπιστευόμαστε το μοντέλο σε ό,τι κατηγοριοποιεί ως θετική κλάση.

$$Precision = \frac{TP}{TP + FP} \tag{2.21}$$

Ανάκληση (Recall)

Η Ανάκληση (Recall) μας δείχνει τι ποσοστό των δειγμάτων που ήταν όντως δείγματα της θετικής κλάσης κατηγοριοποιήθηκαν ως δείγματα της θετικής κλάσης. Με απλά λόγια, υψηλή Ανάκληση σημαίνει οτι μπορούμε να εμπιστευόμαστε το μοντέλο σε ό,τι *δεν* κατηγοριοποιεί ως θετική κλάση.

$$Recall = \frac{TP}{TP + FN} \tag{2.22}$$

F1-Score

Το F1-Score είναι ο αρμονικός μέσος της Ακρίβειας και της Ανάκλησης.

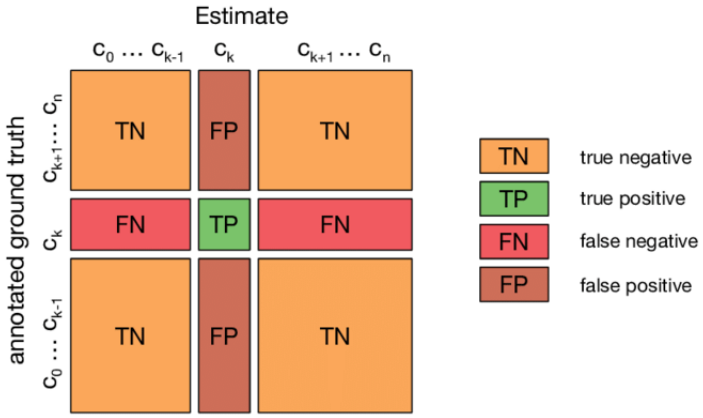
$$Recall = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.23}$$

Οι παραπάνω μετρικές μπορούν να χρησιμοποιηθούν και σε προβλήματα ταξινόμησης σε πολλές κλάσεις, όμως σε αυτή την περίπτωση πρέπει να υπολογιστούν ξεχωριστά για κάθε δυάδα πιθανών κλάσεων. Με τον τρόπο αυτό, όπως είναι αναμενόμενο, προκύπτει ένας μεγάλος όγκος πληροφορίας που είναι δύσκολο να καταναλωθεί. Στην περίπτωση αυτή είναι χρήσιμη η οπτικοποίηση των δεδομένων η οποία γίνεται με τη βοήθεια του πίνακα σύγχυσης (confusion matrix).

Πίνακας Σύγχυσης (Confusion Matrix)

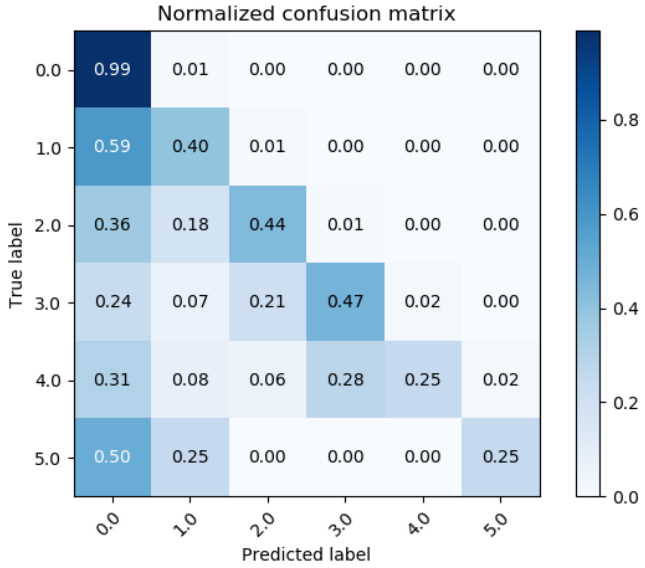
Όπως μπορούμε να παρατηρήσουμε και στο Σχήμα 2.16, ο πίνακας σύγχυσης αποτελεί μία γραφική αναπαράσταση των TP, FP, TN, FN. Η μια από τις δύο διαστάσεις του πίνακα εκφράζει τις πραγματικές κλάσεις και η άλλη τις προβλέψεις. Ο στόχος είναι να έχουμε υψηλές τιμές στην κύρια διαγώνιο, δηλαδή υψηλές τιμές TP.

Πολλές φορές ο Πίνακας Σύγχυσης είναι κανονικοποιημένος, οπότε και περιέχει ποσοστά αντί για απόλυτες τιμές, όπως φαίνεται στο Σχήμα 2.17. Επιπλέον, σε αυτή την περίπτωση ορίζεται ένας χάρτης χρώματος (color map) ώστε τα χρώματα που χρησιμοποιούνται να έχουν επεξηγηματικό ρόλο. Η οπτικοποίηση του Πίνακα Σύγχυσης συμβάλλει στην ποιοτική ερμηνεία των αποτελεσμάτων. Για παράδειγμα, η τιμή 0.99 στο πρώτο κελί του πίνακα του Σχήματος 2.17 δείχνει ότι το εν λόγω μοντέλο είναι πολύ αποτελεσματικό στο να ταξινομεί σωστά τα δείγματα της κλάσης 0. Αντίθετα, η τιμή 0.25 στο κάτω δεξιά κελί του πίνακα



Σχήμα 2.16: *Πίνακας Σύγχυσης (Confusion Matrix) - Ορισμός.*

δείχνει την αδυναμία του μοντέλου να ταξινομεί σωστά δείγματα της κλάσης 5. Μάλιστα, η τιμή 0.50 στο πρώτο κελί της ίδιας γραμμής σημαίνει πως το μοντέλο μας ¨μπερδεύεται' και με συχνότητα 50% προβλέπει τα δείγματα της κλάσης 5 ως δείγματα της κλάσης 0.



Σχήμα 2.17: *Πίνακας Σύγχυσης (Confusion Matrix) - Παράδειγμα.*

Τέλος, για τα προβλήματα ταξινόμησης σε πολλές κλάσεις όπου υπάρχει μια σχετική ισορροπία μεταξύ του πλήθους των δειγμάτων ανά κλάση, ιδιαίτερα χρήσιμη είναι και η μετρική της Ορθότητας (Accuracy).

Ορθότητα (Accuracy)

Η Ορθότητα (Accuracy) στην περίπτωση της ταξινόμησης σε πολλές κλάσεις ορίζεται ως το πλήθος των ορθώς ταξινομημένων δειγμάτων TP_c για κάθε κλάση c του συνόλου των επιτρεπτών κλάσεων C, προς το συνολικό πλήθος των δειγμάτων N του συνόλου δεδομένων.

$$Accuracy = \frac{\sum_{c \in C} TP_c}{N} \tag{2.24}$$

Μέρος II

Πρακτικό Μέρος Α΄ - Ταξινόμηση σε μουσικό είδος

Κεφάλαιο 3

Δεδομένα και Προεπεξεργασία

Το κεφάλαιο αυτό περιλαμβάνει πληροφορίες για τα δεδομένα που χρησιμοποιήθηκαν στο πρώτο μέρος της διπλωματικής εργασίας. Αρχικά, θα παρουσιαστεί το σύνολο δεδομένων μαζί με τα ποιοτικά και ποσοτικά του χαρακτηριστικά. Στη συνέχεια θα αναλυθούν οι τρόποι με τους οποίους έγινε η εξαγωγή χρήσιμων χαρακτηριστικών από τα δεδομένα αυτά, η τεχνικές προεπεξεργασίας που εφαρμόστηκαν, καθώς και οι μεθοδολογίες που χρησιμοποιήθηκαν από το πεδίο της Ψηφιακής Επεξεργασίας Σήματος.

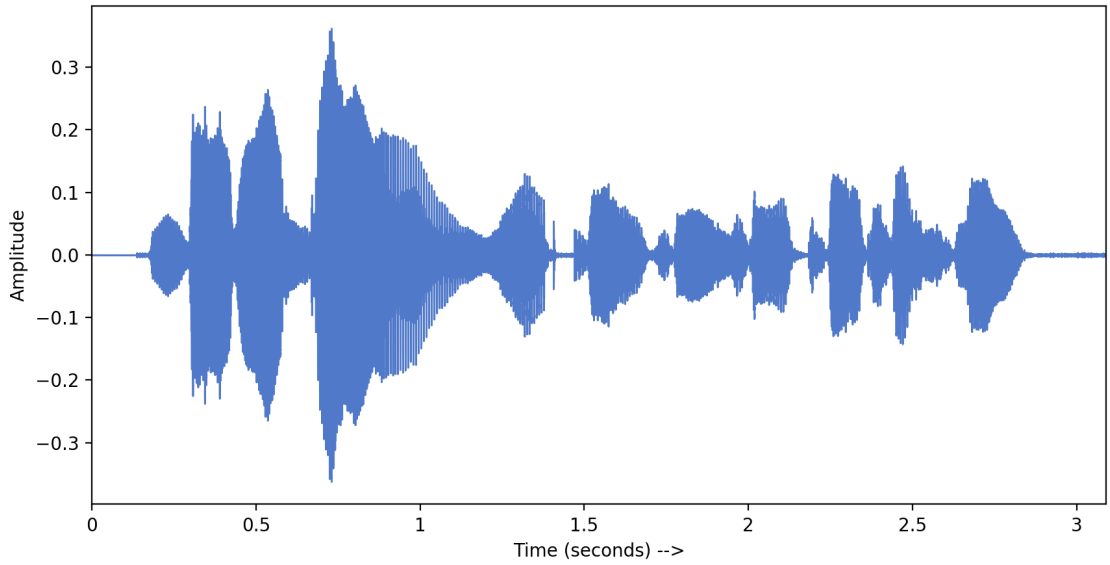
3.1 Σύνολο Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στο πρώτο μέρος της εργασίας είναι το σύνολο GTZAN το οποίο αναφέρθηκε για πρώτη φορά στη βιβλιογραφία το 2002 από τους G. Tzanetakis και P. Cook [1]. Πρόκειται για ένα σύνολο αποτελούμενο από 1000 ηχητικά αρχεία wav των 30 δευτερολέπτων το καθένα. Το κάθε αρχείο είναι επισημασμένο με ένα από τα δέκα βασικά μουσικά είδη: hip hop, country, disco, metal, reggae, blues, rock, classical, jazz, pop. Με αυτόν τον τρόπο λοιπόν γίνεται ένας διαχωρισμός των μουσικών κομματιών σε δέκα κατηγορίες. Η ταξινόμηση σε αυτές τις κατηγορίες είναι ισορροπημένη, το οποίο σημαίνει πως ο αριθμός δειγμάτων ανά κατηγορία είναι σταθερός και επομένως έχουμε 100 μουσικά κομμάτια ανά είδος.

Το συγκεκριμένο σύνολο δεδομένων είναι σχετικά μικρού μεγέθους και πλήρως διαθέσιμο στο ευρύ κοινό. Αυτά του τα χαρακτηριστικά το έχουν καταστήσει ένα από τα πιο ευρέως διαδεδομένα σύνολα δεδομένων, με αναφορά σε τουλάχιστον 100 δημοσιεύσεις στη βιβλιογραφία, ενώ θεωρείται το πιο πολυχρησιμοποιημένο σύνολο δεδομένων αξιολόγησης για τα προβλήματα ταξινόμησης σε μουσικό είδος. Παρόλα αυτά, από έρευνες που διενεργήθηκαν μεταγενέστερα, διαπιστώθηκε ότι το GTZAN περιλαμβάνει ορισμένα λάθη, όπως επαναλήψεις κομματιών ή λάθος επισημάνσεις. Το πλήθος των λαθών όμως δεν το καθιστά ακατάλληλο για τη συγκεκριμένη χρήση, ωστόσο είναι σημαντικό για την όποια αξιολόγηση πραγματοποιείται πάνω στο συγκεκριμένο σύνολο δεδομένων να λαμβάνονται υπόψιν και τα ελαττώματά του [26].

3.2 Προεπεξεργασία και Εξαγωγή Χρήσιμων Χαρακτηριστικών

Όπως φαίνεται και στην εικόνα 3.1, ένα ηχητικό σήμα στην ακατέργαστη μορφή του είναι μια κυματομορφή, δηλαδή μία χρονοσειρά που δείχνει πώς μεταβάλλεται το πλάτος (amplitude) του σήματος σε συνάρτηση με το χρόνο. Για να αναπαρασταθεί η πληροφορία αυτή με ψηφιακό τρόπο, το πλάτος του λαμβάνει τιμές σε τακτές, διακριτές χρονικές στιγμές, οι οποίες καθορίζονται από τη συχνότητα δειγματοληψίας (sampling rate). Επιπλέον, οι τιμές του πλάτους υφίστανται κβάντιση, δηλαδή λαμβάνουν τιμές από ένα προκαθορισμένο σύνολο δυνατών τιμών. Το πλήθος αυτών των τιμών και συνεπώς η ακρίβεια του πλάτους εξαρτώνται από το bit depth δηλαδή το πλήθος των δυαδικών ψηφίων που χρησιμοποιούνται για την κωδικοποίησή του. Τελικά, ένα ψηφιακό ηχητικό σήμα αναπαρίσταται από μία ακολουθία κβαντισμένων δειγμάτων.



Σχήμα 3.1: *Κυματομορφή ηχητικού σήματος*

Ένα ηχητικό σήμα σε αυτή τη μορφή συνήθως δεν είναι αρκετό για την αυτόματη διάκριση σε κλάσεις από ένα νευρωνικό δίκτυο. Για να μπορέσουν τα δεδομένα του συνόλου να τροφοδοτηθούν σε ένα νευρωνικό δίκτυο και να αξιοποιηθούν από αυτό θα πρέπει να μετατραπούν στην κατάλληλη μορφή. Στην ενότητα αυτή θα περιγραφούν τα βήματα αυτής της προεπεξεργασίας, ενώ επίσης θα επεξηγηθούν οι τρόποι με τους οποίους εξάγονται τα κατάλληλα χαρακτηριστικά από τα ηχητικά δεδομένα εισόδου.

3.2.1 Διαίρεση Κομματιού

Το πρώτο βήμα που εφαρμόστηκε στην διαδικασία προεπεξεργασίας του συνόλου δεδομένων είναι η διαίρεση του κάθε κομματιού σε επιμέρους τμήματα. Αυτό γίνεται διότι στην πλειοψηφία των περιπτώσεων, το νευρωνικό δίκτυο μπορεί να αναγνωρίσει την χρήσιμη πληροφορία ακόμα και σε ένα σχετικά μικρό χρονικό παράθυρο. Επομένως, χωρίζοντας κάθε αρχείο ήχου σε μικρότερα αυξάνουμε σημαντικά το πλήθος των δειγμάτων εισόδου, χωρίς να χάνουμε

χρήσιμη πληροφορία. Επιπλέον, ο διαχωρισμός αυτός έχει μια επιπλέον χρησιμότητα: εφόσον το μοντέλο είναι σε θέση να αναγνωρίσει το μουσικό είδος από ένα σύντομο χρονικά ηχητικό σήμα, μπορεί να παράξει πολλές, ανεξάρτητες μεταξύ τους, προβλέψεις για ένα μεγαλύτερο σε χρονική διάρκεια αρχείο ήχου. Αυτό σημαίνει πως η συνολική πρόβλεψη για το ενιαίο αρχείο ήχου μπορεί να προκύψει ως συνδυασμός των επιμέρους προβλέψεων (π.χ. μέσω ψηφοφορίας - voting), οπότε και το τελικό αποτέλεσμα να είναι περισσότερο αξιόπιστο.

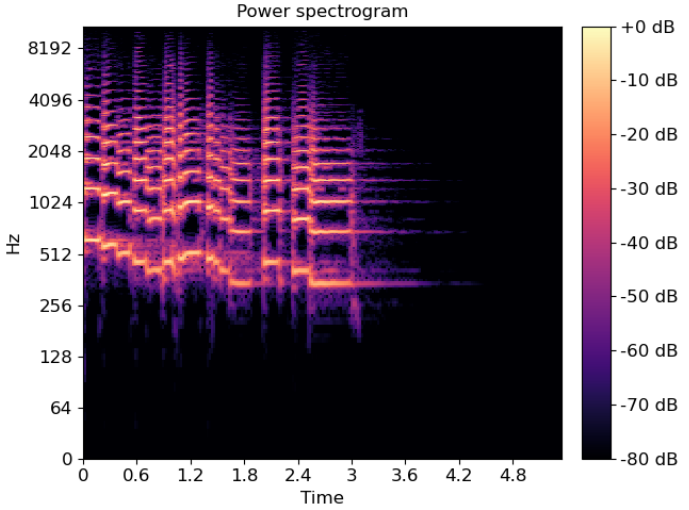
Για τον σκοπό της παρούσας εργασίας, εφαρμόσαμε δύο διαφορετικούς διαχωρισμούς ανάλογα με την προτεινόμενη αρχιτεκτονική του νευρωνικού και συγκρίναμε τα αποτελέσματα. Πιο συγκεκριμένα, το κάθε αρχείο ήχου διάρκειας 30 δευτερολέπτων διαχωρίστηκε σε τμήματα διάρκειας 3 δευτερολέπτων (πρώτη προσέγγιση) και 1.5 δευτερολέπτων (δεύτερη προσέγγιση) και 50% επικάλυψη μεταξύ των διαδοχικών τμημάτων. Έτσι λοιπόν, από κάθε αρχείο ήχου διάρκειας 30 δευτερολέπτων, προέκυψαν 19 ή 39 επιμέρους τμήματα, αντίστοιχα, αυξάνοντας σημαντικά των όγκο των δειγμάτων του συνόλου εκπαίδευσης.

3.2.2 Μετασχηματισμοί Ηχητικού Σήματος

Μετασχηματισμός Short Time Fourier Transform - STFT

Ο μετασχηματισμός Short-time Fourier transform (STFT) χρησιμοποιείται για να ορίσει την ημιτονοειδή συχνότητα και το φασικό περιεχόμενο μέσα σε μικρά, χρονικά περιορισμένα τμήματα ενός σήματος. Στην πράξη, για τον υπολογισμό του μετασχηματισμού STFT το σήμα χωρίζεται σε μικρότερα επικαλυπτόμενα τμήματα ίσης διάρκειας και στη συνέχεια υπολογίζεται ο διακριτός μετασχηματισμός Fourier για καθένα από αυτά.

Το αποτέλεσμα αυτής της διαδικασίας αναπαρίσταται (εν μέρει) οπτικά με τη μορφή φασματογραφήματος (spectrogram) όπως φαίνεται και στο σχήμα 3.2, δηλαδή μίας δισδιάστατης αναπαράστασης όπου ο οριζόντιος άξονας δείχνει τα επιμέρους χρονικά παράθυρα και ο κάθετος δείχνει το *μέτρο* του φασματικού περιεχομένου του αντίστοιχου παραθύρου για τα διαφορετικά εύρη συχνοτήτων.



Σχήμα 3.2: *Φασματογράφημα μετασχηματισμού STFT*

Με μαθηματικούς όρους, ο μετασχηματισμός STFT για ένα σήμα διακριτού χρόνου $x[n]$

δίνεται από την εξής σχέση:

$$\textbf{STFT}\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{3.1}$$

όπου, $w[n]$ είναι μια συνάρτηση παράθυρο (συνήθως τύπου Hann ή Γκαουσιανό).

Στην πράξη, ο μετασχηματισμός STFT του κάθε δείγματος του συνόλου δεδομένων υπολογίστηκε με τη βοήθεια της βιβλιοθήκης librosa [27], χρησιμοποιώντας τις παραμέτρους n_fft=1024, hop_length=512 στην περίπτωση του διαχωρισμού σε τμήματα των 3 δευτερολέπτων και n_fft=1024, hop_length=256 στην περίπτωση του διαχωρισμού σε τμήματα των 1.5 δευτερολέπτων. Και στις δύο περιπτώσεις, οδηγούμαστε σε ένα πίνακα διαστάσεων 513×130, όπου η πρώτη διάσταση είναι τα εύρη συχνοτήτων και η δεύτερη τα χρονικά παράθυρα.

Μετασχηματισμός Chroma Short Time Fourier Transform - Chroma STFT

Τα χαρακτηριστικά chroma ή αλλιώς κλάσεις τονικού ύψους προκύπτουν από την παρατήρηση ότι οι άνθρωποι αντιλαμβάνονται δύο μουσικούς τόνους ως παρόμοιους εάν αυτοί απέχουν κατά μία οκτάβα. Έτσι λοιπόν, ένας μουσικός τόνος μπορεί να αναλυθεί σε δύο συνιστώσες: το τονικό ύψος και το χρώμα chroma. Με τη λογική αυτή, μπορούμε να πούμε ότι οι κλάσεις τονικού ύψους μπορούν να αντιστοιχιστούν με τις τιμές της ισοβαθμισμένης τονικής κλίμακας της δυτικής μουσικής:

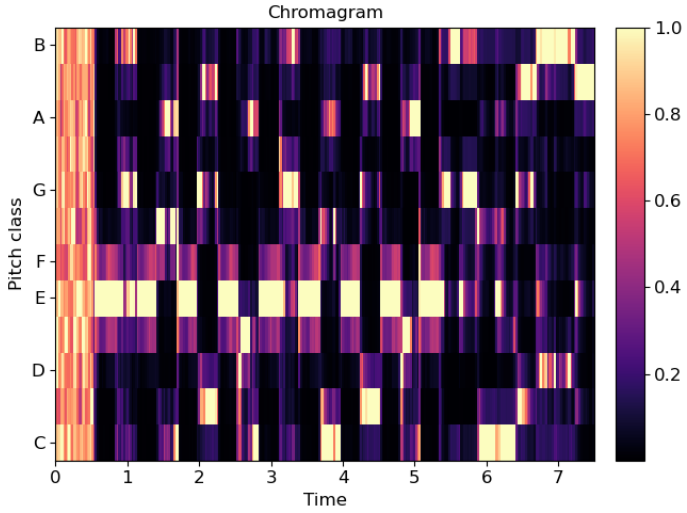
$$C, C\sharp, D, D\sharp, E, F, F\sharp, G, G\sharp, A, A\sharp, B. \tag{3.2}$$

Στη πράξη, για τον υπολογισμό του μετασχηματισμού Chroma STFT, υπολογίζεται ο μετασχηματισμός STFT όπως αναλύθηκε στην προηγούμενη παράγραφο, εξάγεται το μέτρο αυτού και στη συνέχεια για κάθε χρονικό παράθυρο υπολογίζονται τα χαρακτηριστικά chroma. Αυτό σημαίνει ότι το φασματικό περιεχόμενο που αντιστοιχεί σε κάθε κλάση τονικού ύψους συγκεντρώνεται σε ένα συντελεστή ανά κλάση [28]. Η οπτικοποίηση το συγκεκριμένου μετασχηματισμού μπορεί να γίνει και πάλι με τη χρήση του φασματογραφήματος όπως φαίνεται και στην εικόνα 3.3.

Τέλος, αξίζει να αναφερθεί ότι η εξαγωγή του μετασχηματισμού από τα δείγματα του συνόλου δεδομένων έγινε και πάλι με τη χρήση της βιβλιοθήκης librosa [27], χρησιμοποιώντας τις ίδιες παραμέτρους που χρησιμοποιήθηκαν και για την εξαγωγή του μετασχηματισμού STFT. Τελικά, για κάθε δείγμα του συνόλου δεδομένων προέκυψε ένας πίνακας διαστάσεων 12×130.

Μετασχηματισμός Mel Spectrogram

Η κλίμακα mel είναι ένας μη γραμμικός μετασχηματισμός που εφαρμόζεται πάνω σε τιμές συχνοτήτων. Ο στόχος του μετασχηματισμού είναι να επανατοποθετήσει τις συχνότητες πάνω σε μία κλίμακα όχι με βάση την πραγματική τους απόσταση αλλά με βάση το πώς γίνεται αντιληπτή αυτή η απόσταση από το ανθρώπινο αυτί. Η μετατροπή σε αυτή την κλίμακα έχει αξία, διότι μια σταθερή διαφορά συχνοτήτων γίνεται όλο και λιγότερο αντιληπτή από τον ακροατή καθώς οι συχνότητες αυτές αυξάνονται [29].



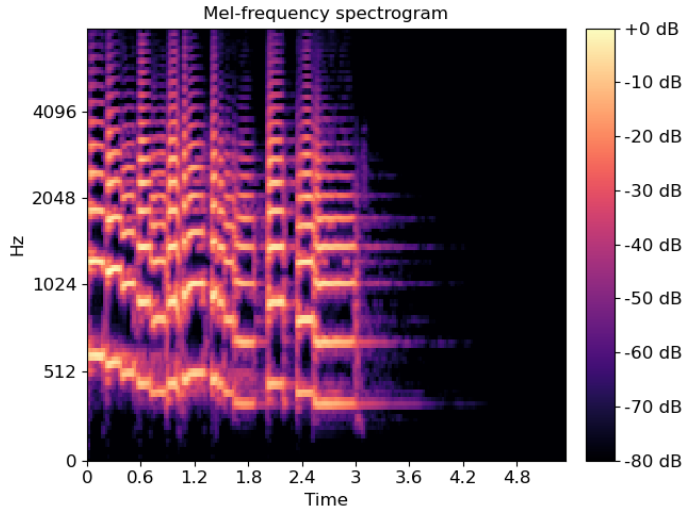
Σχήμα 3.3: *Φασματογράφημα μετασχηματισμού Chroma STFT*

Με μαθηματικούς όρους, ο μετασχηματισμός αυτός εκφράζεται ως εξής:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.3}$$

όπου f η συχνότητα σε Hz και m η συχνότητα σε mel.

Ουσιαστικά, ο μετασχηματισμός Mel Spectrogram προκύπτει από το μετασχηματισμό STFT όπου οι συχνότητες κανονικοποιούνται στην κλίμακα mel. Ο μετασχηματισμός αυτός μπορεί να αναπαρασταθεί γραφικά με τη μορφή φασματογραφήματος όπως φαίνεται και στην εικόνα 3.4, όπου ο κάθετος άξονας αναπαριστά το μέτρο του φασματικού περιεχομένου στην κλίμακα mel χωρισμένο σε εύρη συχνοτήτων ενώ ο οριζόντιος τα χρονικά παράθυρα.



Σχήμα 3.4: *Φασματογράφημα μετασχηματισμού Mel Spectrogram*

Στην πράξη, η εξαγωγή του μετασχηματισμού Mel Spectrogram γίνεται και πάλι με τη χρήση της βιβλιοθήκης librosa [27], χρησιμοποιώντας τις παραμέτρους που χρησιμοποιήθηκαν για την εξαγωγή του μετασχηματισμού STFT και επιπλέον n_mels=128. Έτσι για κάθε δείγμα του συνόλου δεδομένων προκύπτει ένας πίνακας διαστάσεων 128×130 όπου η πρώτη

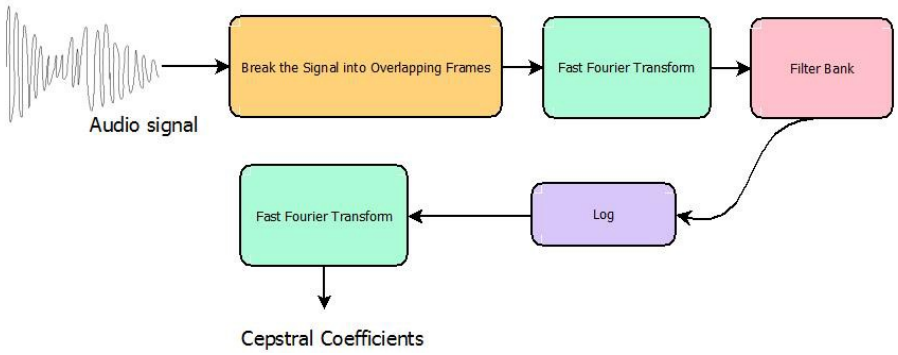
διάσταση είναι τα εύρη συχνοτήτων στην κλίμακα mel και η δεύτερη τα χρονικά παράθυρα.

Μετασχηματισμός Mel-Frequency Cepstrum Coefficients - MFCC

Ο μετασχηματισμός MFCC προκύπτει με βάση τον παρακάτω αλγόριθμο, οποίος περιλαμβάνει ως αρχικά βήματα την εξαγωγή μετασχηματισμών που έχουμε ήδη αναλύσει στις προηγούμενες υποενότητες:

1. Υπολογισμός του μετασχηματισμού STFT.
2. Εξαγωγή της ισχύς του φασματικού περιεχομένου, μέσω του τετραγώνου του μέτρου του μετασχηματισμού STFT - εξαγωγή power spectrogram.
3. Μετατροπή του power spectrogram στην κλίμακα mel.
4. Μετατροπή του mel power spectrogram σε λογαριθμική κλίμακα.
5. Εφαρμογή του Διακριτού Μετασχηματισμού Συνημιτόνου (DCT) στο log mel power spectrogram.
6. Οι συντελεστές MFCC είναι τα πλάτη του τελευταίου μετασχηματισμού.

Η διαδικασία αυτή συνοψίζεται και στο σχήμα 3.5.



Σχήμα 3.5: *Διαδικασία Εξαγωγής μετασχηματισμού MFCC*

Η βιβλιοθήκη librosa [27] χρησιμοποιείται και πάλι σε αυτή την περίπτωση. Για κάθε χρονικό παράθυρο εξάγονται 20 συντελεστές MFCC, επομένως για κάθε δείγμα του συνόλου δεδομένων προκύπτει ένας πίνακας διαστάσεων 20×130. Επιπλέον, υπολογίζουμε την πρώτη και την δεύτερη παράγωγο αυτών των συντελεστών και τις συνενενώνουμε σε έναν ενιαίο πίνακα χαρακτηριστικών μαζί με τους αρχικούς συντελεστές, με αποτέλεσμα τελικά τη δημιουργία ενός πίνακα διαστάσεων 60×130 για κάθε δείγμα του συνόλου δεδομένων.

Η αξία του μετασχηματισμού MFCC έγκειται το ότι εφαρμόζει την κλίμακα mel και επομένως κωδικοποιεί το ηχητικό σήμα σε μία μορφή που ενσωματώνει τον τρόπο με τον οποίο αυτό γίνεται αντιληπτό από τον άνθρωπο.

Μετασχηματισμός Constant Q Transform - CQT

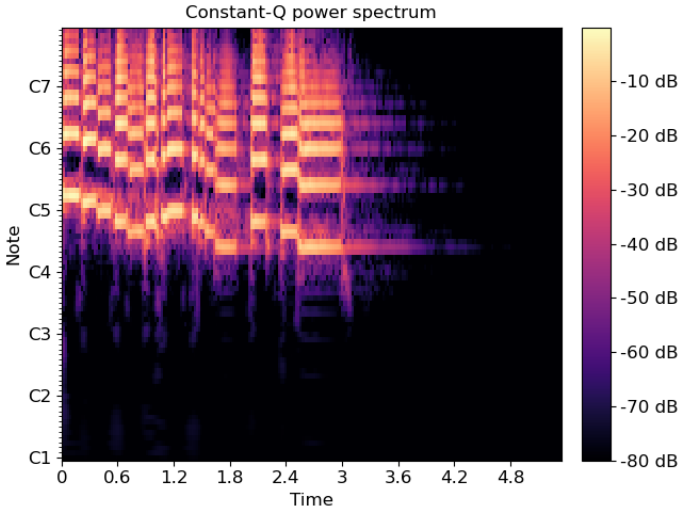
Ο μετασχηματισμός Constant Q Transform - CQT είναι ένας ακόμα μετασχηματισμός ο οποίος μεταφέρει ένα σύνολο δεδομένων από το πεδίο του χρόνου στο πεδίο της συχνότητας.

Μπορεί να θεωρηθεί ως μια σειρά φίλτρων f_k, τοποθετημένα σε λογαριθμικές αποστάσεις στο πεδίο της συχνότητας, με το φασματικό πλάτος δf_k του k-οστού φίλτρου να είναι ακέραιο πολλαπλάσιο του φασματικού πλάτους του προηγούμενου φίλτρου. Με μαθηματικούς όρους, αυτό μπορεί να εκφραστεί ως εξής:

$$\delta f_k = 2^{1/n} \cdot \delta f_{k-1} = \left(2^{1/n}\right)^k \cdot \delta f_{min}, \delta f_k = 2^{1/n} \cdot \delta f_{k-1} = \left(2^{1/n}\right)^k \cdot \delta f_{min}, \tag{3.4}$$

όπου δf_k είναι το πλάτους τουk-οστού φίλτρου, f_{min} είναι η κεντρική συχνότητα του χαμηλότερου φίλτρου και n είναι το πλήθος φίλτρων ανά οκτάβα.

Ο μετασχηματισμός αυτός μπορεί όπως επίσης να αναπαρασταθεί εποπτικά με τη βοήθεια φασματογραφήματος, όπως φαίνεται και στην εικόνα 3.6.



Σχήμα 3.6: *Φασματογράφημα μετασχηματισμού CQT*

Ο υπολογισμός του μετασχηματισμού έγινε με την βοήθεια της βιβλιοθήκης librosa [27], διατηρώντας τις παραμέτρους που χρησιμοποιήθηκαν και στους υπόλοιπους μετασχηματισμούς και επιπλέον $n_bins = 85$. Εν τέλει, προκύπτει ένας πίνακας διαστάσεων 85×130 για κάθε δείγμα του συνόλου δεδομένων.

3.2.3 Διαχωρισμός Συνόλου Δεδομένων

Αφότου έχει ολοκληρωθεί η διαδικασία της προεπεξεργασίας όπως αναφέρθηκε στις προηγούμενες ενότητες, έχουμε καταλήξει με ένα σύνολο δεδομένων αποτελούμενο από 19.000 ή 39.000 διαφορετικά δείγματα, ανάλογα με τον τρόπο με τον οποίο έγινε η διαίρεση του κομματιού. Το σύνολο δεδομένων παραμένει ισορροπημένο όπως είναι αναμενόμενο, με την καθεμία από τις 10 κλάσεις να περιλαμβάνει το 10% επί του συνολικού πλήθους των δειγμάτων. Το τελευταίο βήμα πριν προχωρήσουμε στη διαδικασία της εκπαίδευσης, είναι ο διαχωρισμός των δειγμάτων αυτών σε 3 επιμέρους υποσύνολα: το σύνολο εκπαίδευσης (training set), το σύνολο επικύρωσης (validation set) και το σύνολο ελέγχου (test set). Η αναλογία διαχωρισμού είναι 80%-10%-10% αντίστοιχα, ενώ φροντίζουμε ώστε να παραμένουν ισορροπημένα όλα τα επιμέρους υποσύνολα. Τέλος, αξίζει να αναφερθεί πως τα δείγματα που αναφέρονται στο ίδιο

αρχικό κομμάτι, παρέμειναν στο ίδιο υποσύνολο δεδομένων, ώστε να μπορέσει να εφαρμοστεί η διαδικασία ψηφοφορίας (voting) για την παραγωγή της τελικής πρόβλεψης.

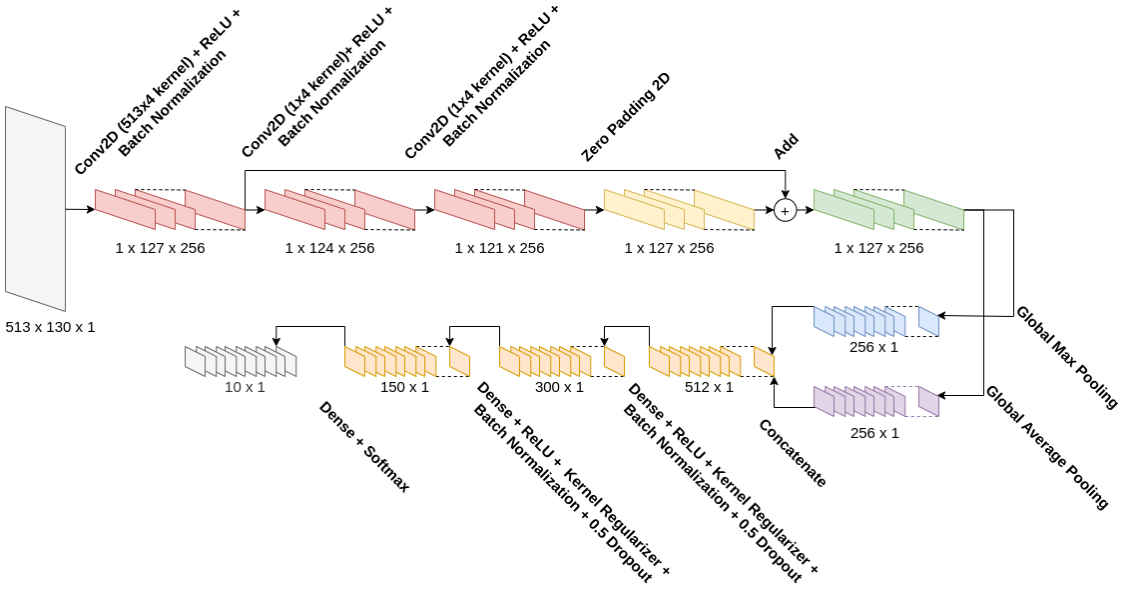
Κεφάλαιο 4

Αρχιτεκτονικές Νευρωνικών Δικτύων

Στο κεφάλαιο αυτό παρουσιάζονται οι διαφορετικές προσεγγίσεις που ακολουθήθηκαν όσον αφορά τις αρχιτεκτονικές των νευρωνικών δικτύων. Πραγματοποιήθηκαν αρκετοί διαφορετικοί πειραματισμοί, ωστόσο εδώ θα εστιάσουμε στις 2 αρχιτεκτονικές από τις οποίες προέκυψαν τα καλύτερα και πιο ενδιαφέροντα αποτελέσματα.

4.1 Αρχιτεκτονική Νευρωνικού A1

Η αρχιτεκτονική, ο τρόπος εκπαίδευσης, καθώς και ο τρόπος προεπεξεργασίας των δεδομένων για την περίπτωση νευρωνικού A1 είναι κατά κύριο λόγο εμπνευσμένες από τη δουλειά που παρουσιάζεται στο [30].



Σχήμα 4.1: *Αρχιτεκτονική Νευρωνικού A1*

Το δίκτυο, το οποίο φαίνεται και αναλυτικά στο Σχήμα 4.1 είναι επί της ουσίας ένα συνελικτικό νευρωνικό δίκτυο. Εποπτικά, το δίκτυο αποτελείται από 3 συνελικτικά μπλοκ, το καθένα αποτελούμενο από ένα συνελικτικό επίπεδο με συνάρτηση ενεργοποίησης ReLU και ένα επίπεδο κανονικοποίησης δέσμης (batch normalization). Στη συνέχεια η έξοδος του πρώτου και του τρίτου συνελικτικού μπλοκ αθροίζονται, αφού στην τελευταία εφαρμοστεί

Zero Padding ώστε να επανέλθει στις διαστάσεις της εξόδου του πρώτου μπλοκ. Με τον όρο Zero Padding αναφερόμαστε στην τεχνική κατά την οποία αυξάνουμε το μέγεθος ενός πίνακα προσθέτοντας ένα περίγραμμα στην περίμετρό του και συμπληρώνοντας με μηδενικά τις επιπλέον θέσεις. Έτσι λοιπόν προκύπτει ένας πίνακας χαρακτηριστικών διαστάσεων 127×256. Ακολούθως, ο δισδιάστατος πίνακας χαρακτηριστικών μετατρέπεται σε μονοδιάστατο διάνυσμα ως εξής: τροφοδοτείται παράλληλα σε ένα επίπεδο Global Max Pooling και σε ένα επίπεδο Global Average Pooling και οι έξοδοί τους συνενώνονται σε ένα διάνυσμα μήκους 512. Τα επίπεδα Global Max Pooling και Global Average Pooling μετατρέπουν ένα δισδιάστατο πίνακα σε μονοδιάστατο διάνυσμα επιστρέφοντας το μέγιστο (ή το μέσο όρο αντίστοιχα) από κάθε πίνακα χαρακτηριστικών. Συνεχίζοντας, το διάνυσμα αυτό τροφοδοτείται σε 3 πλήρως συνδεδεμένα επίπεδα, μεταξύ των οποίων υπάρχει ένα στρώμα κανονικοποίησης δέσμης (batch normalization) και ένα στρώμα εγκατάλειψης (dropout). Τα 2 πρώτα πλήρως συνδεδεμένα επίπεδα διαθέτουν ομαλοποίηση πυρήνα (kernel regularizer) και συνάρτηση ενεργοποίησης ReLU, ενώ το τελευταίο περιλαμβάνει συνάρτηση ενεργοποίησης Softmax και επιτελεί την κατηγοριοποίηση σε μία από τις δέκα πιθανές κλάσεις.

Η ιδιαιτερότητα του δικτύου έγκειται σε δύο βασικά χαρακτηριστικά:

- Σχήμα και διαστάσεις των πυρήνων των συνελίξεων.
- Σύνδεση Συντόμευσης (Shortcut Connection - Residual Block).

Επιπλέον, μπορούμε να παρατηρήσουμε ορισμένους όρους που δεν έχουν παρουσιαστεί ως τώρα:

- Κανονικοποίηση Δέσμης (Batch Normalization).
- Στρώματα Εγκατάλειψης (Dropout Layers).
- Ομαλοποίηση Πυρήνα (Kernel Regularizer).

4.1.1 Σχήμα και Διαστάσεις των Πυρήνων των Συνελίξεων

Όπως μπορούμε να παρατηρήσουμε και στο σχήμα, ο πυρήνας του πρώτου συνελικτικού μπλοκ έχει διαστάσεις 513×4. Καταλαμβάνει δηλαδή όλη τη διάσταση των συχνοτήτων και εκτείνεται κατά 4 χρονικά παράθυρα στο πεδίο του χρόνου. Αυτό έχει ως αποτέλεσμα να επιτευχθεί ένας γραμμικός συνδυασμός των συχνοτικών χαρακτηριστικών σε ένα μονοδιάστατος διάνυσμα. Τα επόμενα συνελικτικά μπλοκ έχουν πυρήνες διαστάσεων 1×4 έχουν ως στόχο να μοντελοποιήσουν τις χρονικές εξαρτήσεις, καθώς συνδυάζουν χαρακτηριστικά που βρίσκονται κοντά μεταξύ τους στο πεδίο του χρόνου.

4.1.2 Σύνδεση Συντόμευσης (Shortcut Connection - Residual Block)

Όπως φαίνεται και στο σχήμα, υπάρχει μία σύνδεση συντόμευσης (shortcut) ανάμεσα στο πρώτο και στο τρίτο συνελικτικό μπλοκ. Αυτό σημαίνει ότι η έξοδος του πρώτου μπλοκ τροφοδοτείται στην είσοδο του δεύτερου αλλά ταυτόχρονα, προσπερνά το δεύτερο και το τρίτο μπλοκ και συνδυάζεται με την έξοδο του τρίτου μπλοκ.

Η αξία της χρήσης συνδέσεων συντόμευσης έγκειται στο ότι συμβάλλει στην επίλυση του προβλήματος της υποβάθμισης (degradation), κατά το οποίο καθώς αυξάνεται το βάθος ενός νευρωνικού δικτύου, παρατηρείται ένας κορεσμός στην ακρίβεια (accuracy saturation), ο οποίος ακολουθείται από μία απότομη πτώση. Αυτή η υποβάθμιση, δεν οφείλεται στο φαινόμενο της υπερεκπαίδευσης και μάλιστα συνδέεται με την περαιτέρω αύξηση του σφάλματος εκπαίδευσης, καθώς το δίκτυο γίνεται βαθύτερο [4].

4.1.3 Κανονικοποίηση Δέσμης (Batch Normalization)

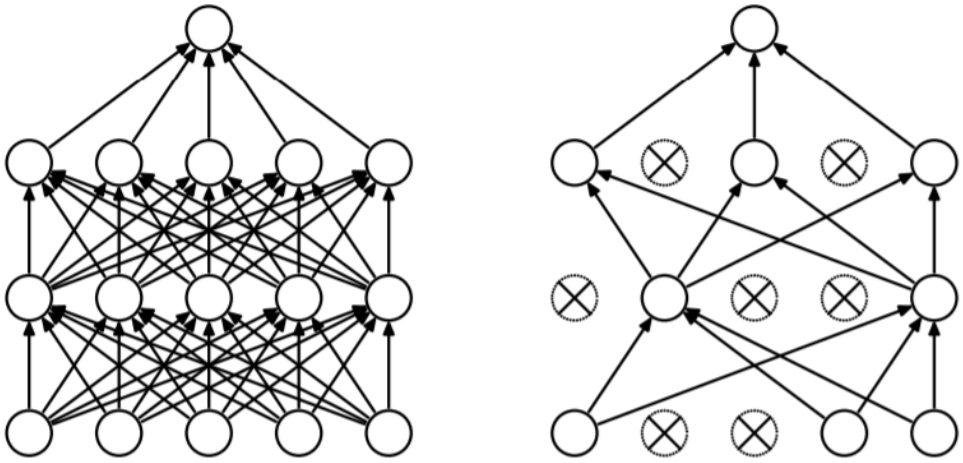
Η κανονικοποίηση δέσμης είναι μια τεχνική που στοχεύει στην αντιμετώπιση του φαινομένου internal covariate shift. Κατά το φαινόμενο αυτό, η κατανομή των εισόδων του κάθε επιπέδου αλλάζει κατά τη διάρκεια της εκπαίδευσης, καθώς οι παράμετροι των προηγούμενων επιπέδων μεταβάλλονται. Αυτό έχει ως αποτέλεσμα την επιβράδυνση της διαδικασίας της εκπαίδευσης, λόγω της ανάγκης για μικρούς ρυθμούς εκπαίδευσης και κατάλληλη αρχικοποίηση παραμέτρων. Η μέθοδος κανονικοποίησης δέσμης επιχειρεί να επιλύσει αυτό το πρόβλημα ενσωματώνοντας στην αρχιτεκτονική του δικτύου (ως ένα επιπλέον επίπεδο) την κανονικοποίηση των εισόδων των επιπέδων, οποία εφαρμόζεται μετά από το πέρασμα κάθε δέσμης - τμήματος από το δίκτυο. Αυτό έχει ως αποτέλεσμα την δυνατότητα χρήσης μεγαλύτερων ρυθμών εκπαίδευσης και κατ΄ επέκταση την συνολική επιτάχυνση της διαδικασίας της εκπαίδευσης. Επιπλέον, οι μεγαλύτεροι ρυθμοί μάθησης έχουν ως ένα επιπλέον πλεονέκτημα την αποφυγή του εγκλωβισμού του αλγορίθμου σε τοπικά ελάχιστα και άρα μπορούν δυνητικά να βελτιώσουν τις συνολικές επιδόσεις του μοντέλου [31].

4.1.4 Στρώματα Εγκατάλειψης (Dropout Layers)

Τα στρώματα εγκατάλειψης έχουν ως βασικό στόχο την αποφυγή του φαινομένου της υπερεκπαίδευσης, το οποίο είναι ιδιαίτερα συχνό στις περιπτώσεις βαθιών νευρωνικών δικτύων, λόγω του μεγάλου αριθμού παραμέτρων που περιλαμβάνουν. Η βασική τους λειτουργικότητα είναι η τυχαία επιλογή ενός αριθμού νευρώνων καθώς και των συνδέσεών τους και η απενεργοποίησή τους κατά τη διάρκεια της εκπαίδευσης. Με τον όρο απενεργοποίηση, εννοούμε πως οι νευρώνες αυτοί δεν λαμβάνονται υπόψη κατά τη διάρκεια μιας συγκεκριμένης διάσχισης του νευρωνικού είτε προς τα εμπρός ή προς τα πίσω, όπως φαίνεται και στην εικόνα 4.2. Αυτό έχει ως αποτέλεσμα τη χρήση διαφορετικών νευρώνων σε κάθε εποχή, οπότε αποφεύγεται η προσαρμογή του συνόλου των βαρών στα δεδομένα εκπαίδευσης. Επί της ουσίας, το συνολικό δίκτυο χωρίζεται σε επιμέρους αραιότερα δίκτυα, τα οποία εκπαιδεύονται ταυτόχρονα [32].

4.1.5 Ομαλοποίηση Πυρήνα (Kernel Regularizer)

Οι ομαλοποιητές (regularizers) δίνουν τη δυνατότητα της εφαρμογής ‘‘ποινών’’ στις παραμέτρους ή στην ενεργοποίηση ενός επιπέδου κατά τη διαδικασία της βελτιστοποίησης. Οι ποινές αυτές συμπεριλαμβάνονται στον υπολογισμό της συνάρτησης κόστους του αλγορίθμου. Συγκεκριμένα, η ομαλοποίηση πυρήνα επιβάλλει ποινές στον πυρήνα (kernel) ενός επιπέδου.



(a) Standard Neural Net (b) After applying dropout.

Σχήμα 4.2: *Εφαρμογή τεχνικής dropout*

4.2 Αρχιτεκτονική Νευρωνικού A2

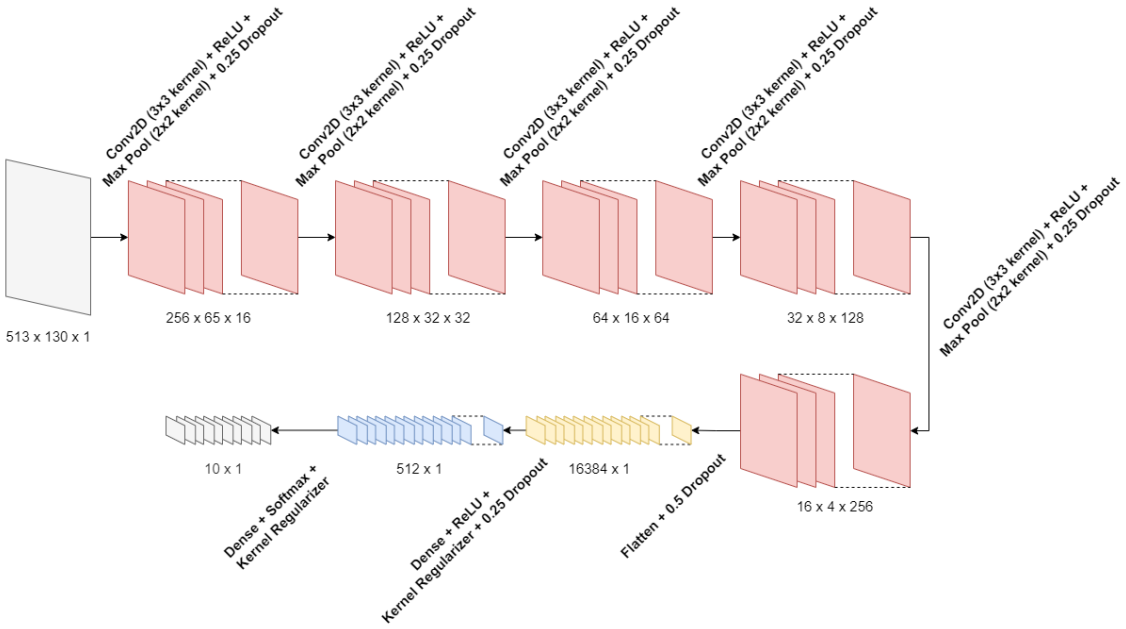
Η αρχιτεκτονική του Νευρωνικού Δικτύου A2 είναι κατά βάση εμπνευσμένη από τη μεθοδολογία που παρουσιάζεται στο [33]. Για την αρχιτεκτονική αυτή πραγματοποιήθηκαν αρκετοί πειραματισμοί όσον αφορά τον τύπο των χαρακτηριστικών που τροφοδοτούνται στο μοντέλο. Καθώς όμως διαφορετικοί τύποι χαρακτηριστικών συνεπάγονται και διαφορετικές διαστάσεις εισόδου, σε κάποιες περιπτώσεις κρίθηκε απαραίτητο να εφαρμοστούν ορισμένες τροποποιήσεις στην αρχιτεκτονική του δικτύου οι οποίες θα παρουσιαστούν στη συνέχεια. Πιο συγκεκριμένα, το συγκεκριμένο νευρωνικό (είτε στην αρχική του είτε στην τροποποιημένη του μορφή) δοκιμάστηκε με τις εξής εισόδους:

- STFT μετασχηματισμός - διαστάσεις 513×130
- Chroma STFT μετασχηματισμός - διαστάσεις 12×130
- Mel Spectrogram μετασχηματισμός - διαστάσεις 128×130
- MFCC μετασχηματισμός - διαστάσεις 30×130
- MFCC μετασχηματισμός συνενωμένος με την πρώτη και τη δεύτερη παράγωγο - διαστάσεις 60×130
- CQT μετασχηματισμός - διαστάσεις 85×130

4.2.1 Είσοδος STFT

Το νευρωνικό που παρουσιάζεται στην παρούσα ενότητα είναι επί της ουσίας είναι ένα απλό συνελικτικό δίκτυο, αποτελούμενο από 5 συνελικτικά μπλοκ. Το κάθε συνελικτικό μπλοκ συντίθεται με τη σειρά του από ένα συνελικτικό επίπεδο, με συνάρτηση ενεργοποίησης ReLU, ένα επίπεδο Max Pooling το οποίο μειώνει τις διαστάσεις του πίνακα χαρακτηριστικών κατά 50% και ένα επίπεδο εγκατάλειψης με πιθανότητα 0.25. Όπως φαίνεται και στο σχήμα 4.3,

από το κάθε συνελικτικό μπλοκ εξάγονται σταδιακά περισσότεροι πίνακες χαρακτηριστικών, φτάνοντας στους 256 πίνακες στην έξοδο του τελευταίου μπλοκ. Στη συνέχεια οι πίνακες αυτοί επιπεδοποιούνται (flattened) ώστε να μετατραπούν σε ένα μονοδιάστατο διάνυσμα, το οποίο ακολούθως τροφοδοτείται σε δύο πλήρως συνδεδεμένα επίπεδα. Εν τέλει, η έξοδος του τελευταίου πλήρως συνδεδεμένου επιπέδου είναι ένα διάνυσμα 10 θέσεων, στην καθεμία από τις οποίες αντιστοιχεί η πιθανότητα του να ανήκει το δείγμα στο αντίστοιχο μουσικό είδος. Δεδομένου ότι η συνάρτηση ενεργοποίησης του τελευταίου επιπέδου είναι η Softmax, οι εν λόγω πιθανότητες αθροίζουν υποχρεωτικά στη μονάδα. Έτσι, τελική πρόβλεψη για το μουσικό είδος προκύπτει ως το είδος που αντιστοιχεί στη θέση με τη μεγαλύτερη πιθανότητα.



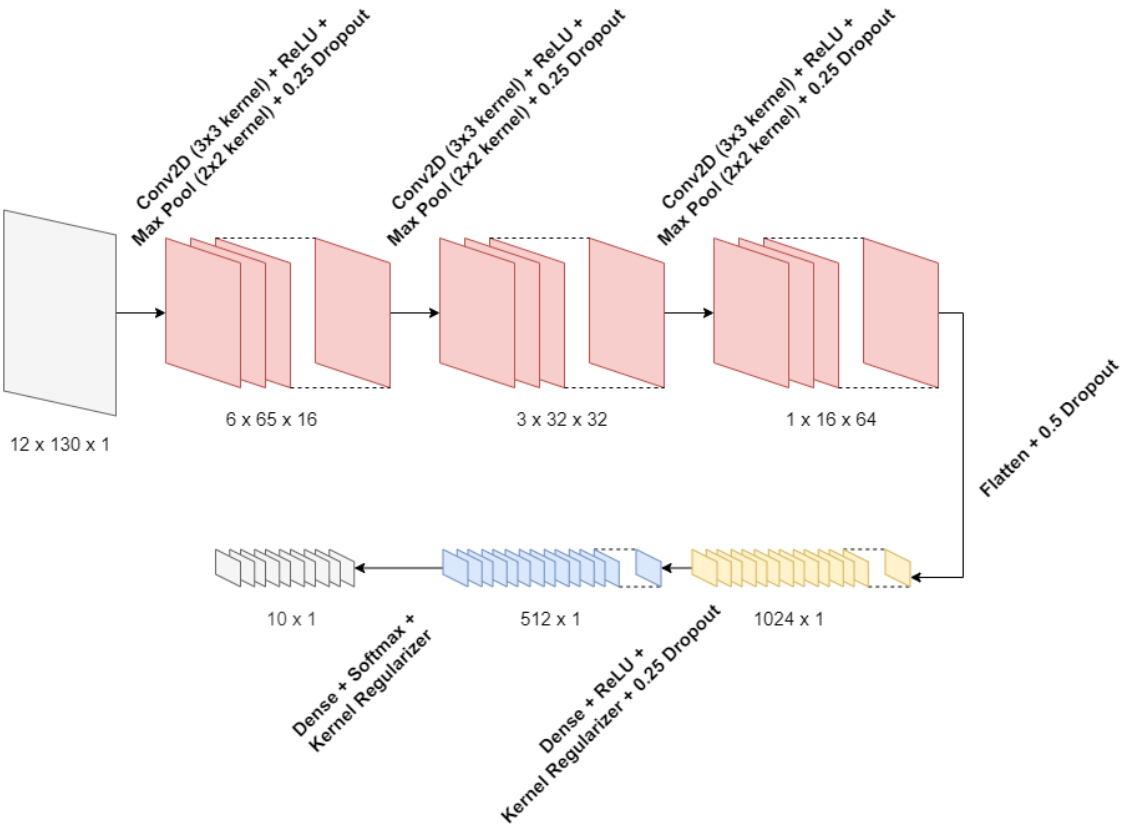
Σχήμα 4.3: *Αρχιτεκτονική Νευρωνικού Α2 - Είσοδος STFT*

4.2.2 Είσοδος Chroma STFT

Στο σχήμα 4.4 μπορούμε να παρατηρήσουμε την αρχιτεκτονική του νευρωνικού δικτύου όπως αυτή τροποποιήθηκε για να μπορεί να δέχεται ως είσοδο τον μετασχηματισμό Chroma STFT, ο οποίος έχει διαστάσεις 12×130 για κάθε δείγμα. Λόγω της πολύ μικρής σε μέγεθος διάστασης των συχνοτήτων, αφαιρέθηκαν τα 2 τελευταία συνελικτικά μπλοκ, καθώς το επίπεδο Max pooling που αυτά περιλαμβάνουν δεν μπορούσε να μειώσει περεταίρω τις διαστάσεις. Έτσι λοιπόν, το τελικό δίκτυο αποτελείται από 3 συνελικτικά μπλοκ, ακολουθούμενα όπως και προηγουμένως από 2 πλήρως συνδεδεμένα επίπεδα.

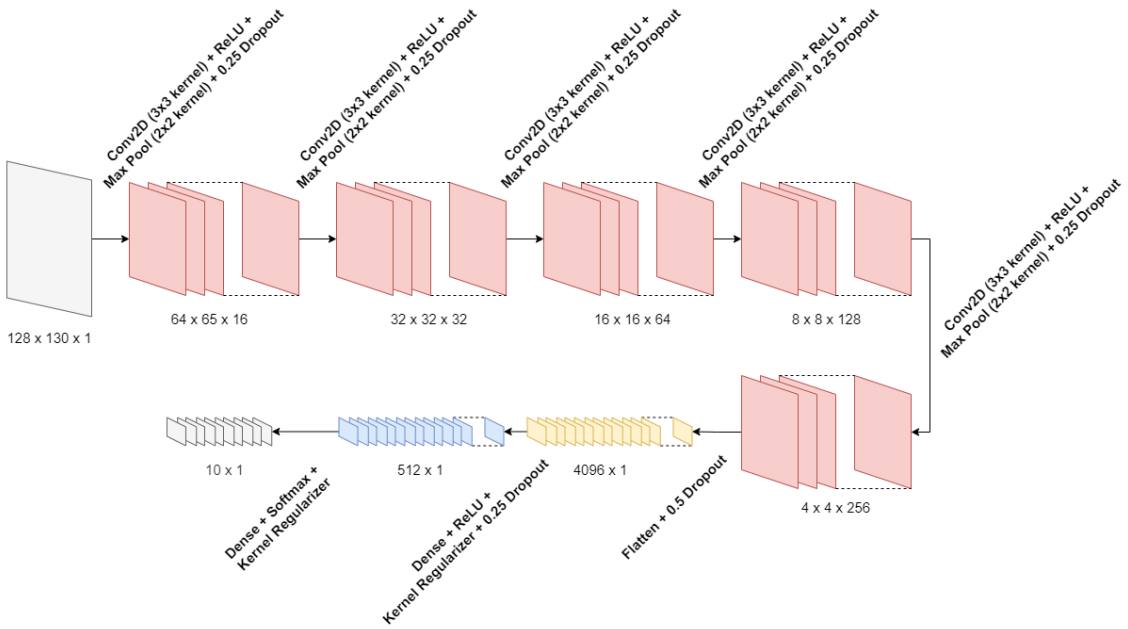
4.2.3 Είσοδος Mel Spectrogram

Στο σχήμα 4.5 παρουσιάζεται η αρχιτεκτονική του δικτύου για την επεξεργασία του μετασχηματισμού Mel Spectrogram. Στην προκειμένη περίπτωση, οι διαστάσεις της εισόδου (12×130) δεν δημιουργούν πρόβλημα αρνητικών διαστάσεων όπως στην περίπτωση του μετασχηματισμού Chroma STFT. Έτσι το δίκτυο παραμένει ως έχει και η μόνη διαφοροποίη-



Σχήμα 4.4: *Αρχιτεκτονική Νευρωνικού A2 - Είσοδος Chroma STFT*

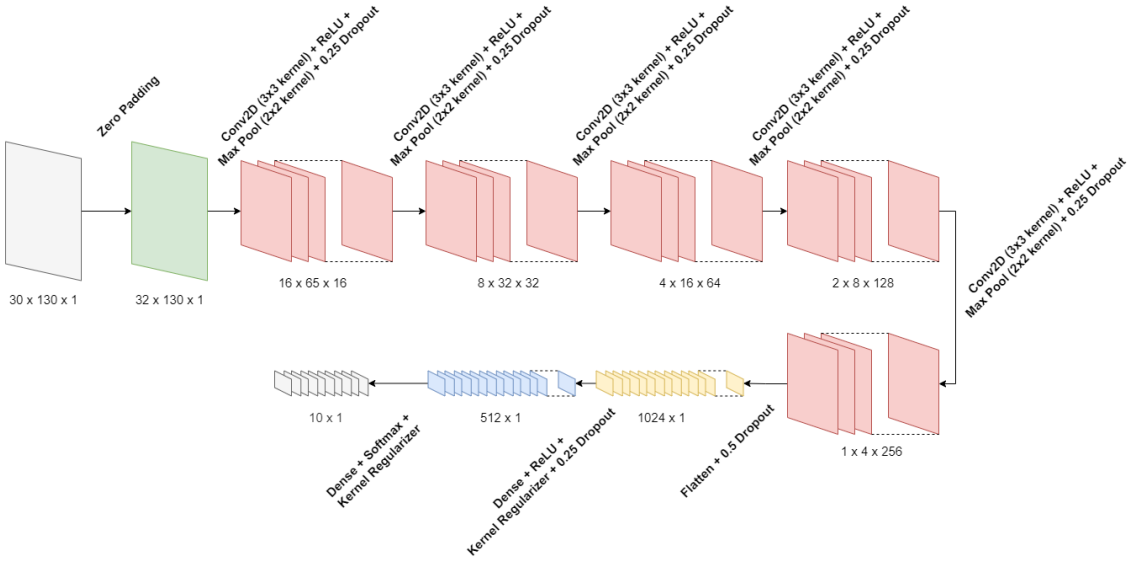
ση παρατηρείται στις διαστάσεις των ενδιάμεσων πινάκων χαρακτηριστικών και συνεπώς στο μήκος του διανύσματος που προκύπτει μετά την επιπεδοποίηση.



Σχήμα 4.5: *Αρχιτεκτονική Νευρωνικού A2 - Είσοδος Mel Spectrogram*

4.2.4 Είσοδος MFCC

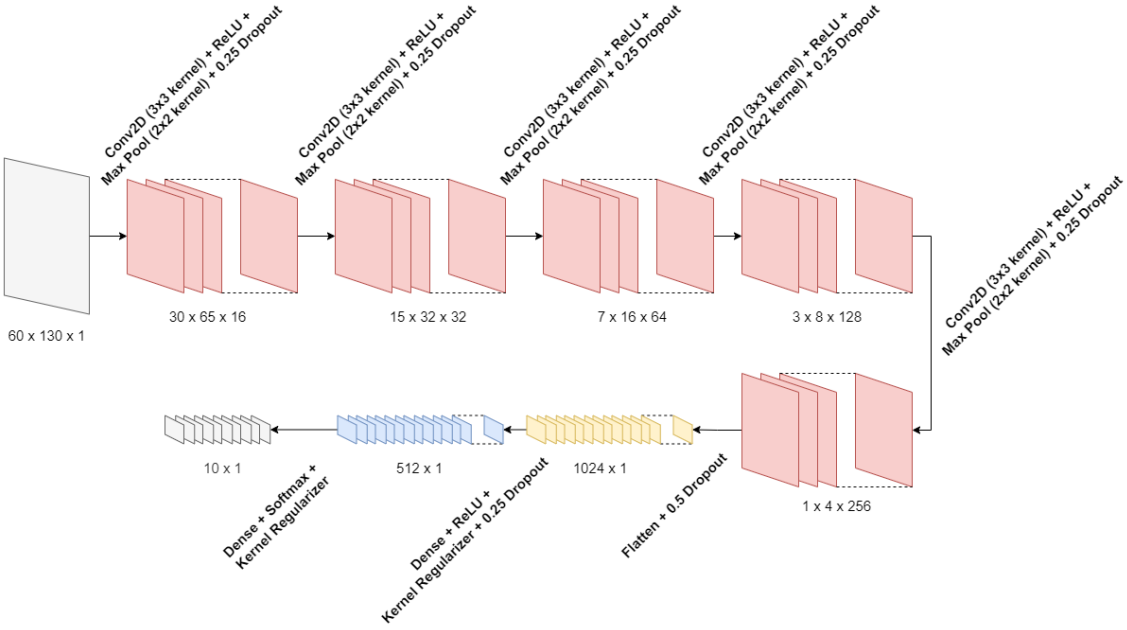
Προχωρώντας στο σχήμα 4.6 φαίνεται το νευρωνικό δίκτυο τροποποιημένο για να δέχεται στην είσοδο το μετασχηματισμό MFCC, ο οποίος έχει διαστάσεις 30×130. Λόγω της μικρής διάστασης των συχνοτήτων, προέκυψε και πάλι το πρόβλημα της αρνητικής διάστασης λόγω της μείωσης διαστάσεων που έχουν ως αποτέλεσμα τα επίπεδα Max pooling. Δεδομένου όμως ότι η διάσταση των συχνοτήτων διαφέρει μόνο κατά 2 από το ελάχιστο μέγεθος που μπορεί να δεχτεί το δίκτυο, προτιμήθηκε η λύση του Zero Padding έναντι της αφαίρεσης συνελικτικού μπλοκ.



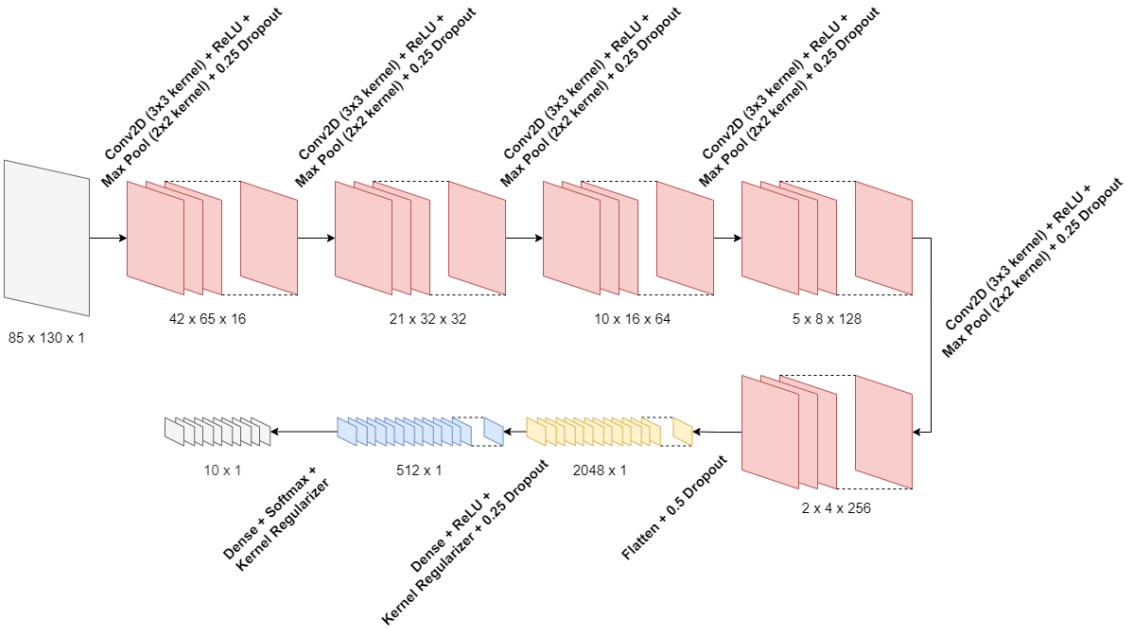
Σχήμα 4.6: *Αρχιτεκτονική Νευρωνικού A2 - Είσοδος MFCC*

4.2.5 Είσοδος MFCC & delta και CQT

Τέλος, στα σχήματα 4.7 και 4.8 μπορούμε να παρατηρήσουμε την αρχιτεκτονική του δικτύου για την επεξεργασία του μετασχηματισμού MFCC συνενωμένου με τις παραγώγους του και του μετασχηματισμού CQT αντίστοιχα. Καθώς οι διαστάσεις των μετασχηματισμών αυτών είναι 60×130 για τον πρώτο και 85×130 για τον δεύτερο, δεν ήταν απαραίτητο να γίνει κάποια αλλαγή στην αρχιτεκτονική του νευρωνικού. Επομένως, πέρα από τις αλλαγές στις διαστάσεις των ενδιάμεσων πινάκων χαρακτηριστικών, η αρχιτεκτονική παραμένει η ίδια με αυτή του δικτύου που δέχεται ως είσοδο τον μετασχηματισμό STFT.



Σχήμα 4.7: *Αρχιτεκτονική Νευρωνικού A2 - Είσοδος MFCC & delta*



Σχήμα 4.8: *Αρχιτεκτονική Νευρωνικού A2 - Είσοδος CQT*

Κεφάλαιο 5

Αποτελέσματα και Αξιολόγηση

Στο παρόν κεφάλαιο παρουσιάζονται αναλυτικά τα πειράματα που πραγματοποιήθηκαν στα πλαίσια της διαδικασίας εκπαίδευσης ενός μοντέλου για την αναγνώριση του μουσικού είδους. Θα γίνει αναφορά στους διαφορετικούς συνδυασμούς χαρακτηριστικών και αρχιτεκτονικών που χρησιμοποιήθηκαν καθώς και στις υπερπαραμέτρους της εκπαίδευσης. Έπειτα, για το κάθε πείραμα που διεξήχθη θα παρουσιαστούν οι μετρικές αξιολόγησης που προέκυψαν.

Σε όλα τα πειράματα που θα περιγραφούν, τόσο η υλοποίηση των μοντέλων όσο και η εκπαίδευση και η αξιολόγηση αυτών υλοποιήθηκαν με τη χρήση της βιβλιοθήκης Tensorflow [34] και της διεπαφής Keras [35]. Η Tensorflow είναι μια βιβλιοθήκη που χρησιμοποιείται ευρέως για την ανάπτυξη και την εκπαίδευση μοντέλων Μηχανικής Μάθησης, οποία υλοποιήθηκε από την Google και έκτοτε χρησιμοποιείται στα πλαίσια έρευνας και ανάπτυξης αλλά και σε περιβάλλον παραγωγής από την εταιρία. Το Keras είναι μια βιβλιοθήκη λογισμικού ανοιχτού κώδικα που λειτουργεί ως διεπαφή για τη βιβλιοθήκη Tensorflow στη γλώσσα προγραμματισμού Python.

Η εξαγωγή των χαρακτηριστικών, η εκπαίδευση των νευρωνικών και η αξιολόγηση των τελικών μοντέλων εγιναν στην πλατφόρμα υπολογιστικού νέφους (cloud computing platform) AWS - Amazon Web Services [36]. Συγκεκριμένα χρησιμοποιήθηκε η υπηρεσία Amazon SageMaker, ο ρόλος της οποίας είναι να δίνει τη δυνατότητα σε επιστήμονες δεδομένων και προγραμματιστες να υλοποιούν, να εκπαιδεύουν και να αναπτύσσουν μοντέλα μηχανικής μάθησης [37]. Από την υπηρεσία αυτή αξιοποιήθηκαν Estimators και Processors οι οποίοι ενσωματώνουν τη διαδικασία εκπαίδευσης μοντέλων και προεπεξεργασίας δεδομένων αντίστοιχα.

5.1 Πειράματα Αρχιτεκτονικής Νευρωνικού Α1

Στην παρούσα ενότητα θα παρουσιαστούν τα πειράματα που διεξήχθησαν χρησιμοποιώντας ως αρχιτεκτονική του δικτύου την Αρχιτεκτονική Α1. Η είσοδος του μοντέλου σε όλα τα πειράματα που θα περιγραφούν είναι ο μετασχηματισμός STFT. Για την εκπαίδευση της εν λόγω αρχιτεκτονικής, το σύνολο δεδομένων εκπαίδευσης χωρίζεται σε επιμέρους τμήματα - παρτίδες με μέγεθος (batch size) 20 δείγματα η καθεμία, ενώ στο τέλος της κάθε εποχής το σύνολο αυτό ‘‘ανακατεύεται’’ (shuffle) και έπειτα χωρίζεται εκ νέου σε παρτίδες, με στόχο να αποφευχθεί το φαινόμενο της υπερεκπαίδευσης. Το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας ως συνάρτηση κόστους την κατηγορική διασταυρούμενη εντροπία (categorical crossentropy loss), για συνολικά 100 εποχές, χρησιμοποιώντας την τεχνική Early Stopping [38]. Ο ρόλος αυτής της τεχνικής είναι ο έλεγχος της υπερεκπαίδευσης. Πιο συγκεκριμένα, επιλέγεται μια

μετρική η οποία τίθεται υπό παρακολούθηση (στην περίπτωσή μας η τιμή της συνάρτησης κόστους για το σύνολο επαλήθευσης) και ελέγχεται αν έχει την επιθυμητή συμπεριφορά μετά το τέλος κάθε εποχής (στην περίπτωσή μας η αναμενόμενη συμπεριφορά του validation loss είναι η μείωση). Αν το validation loss δεν ελαττωθεί για παραπάνω από 3 εποχές (διαδοχικές ή μη) τότε η διαδικασία της εκπαίδευσης διακόπτεται, διότι θεωρείται ότι το φαινόμενο της υπερεκπαίδευσης αρχίζει να εμφανίζεται.

Όσον αφορά τον αλγόριθμο βελτιστοποίησης και το ρυθμό μάθησης, το άρθρο [30] δεν ανέφερε καμία σχετική πληροφορία, επομένως κρίθηκε χρήσιμο να πραγματοποιηθούν ορισμένα πειράματα με διαφορετικούς αλγορίθμους και ρυθμούς μάθησης. Συγκεκριμένα, δοκιμάστηκαν οι αλγόριθμοι Adam [25], Adagrad [23] και Adadelta [24] με 4 διαφορετικούς ρυθμούς μάθησης ο κάθε ένας. Τα αποτελέσματα όσον αφορά την ορθότητα και την τιμή της συνάρτησης κόστους στο σύνολο αξιολόγησης παρουσιάζονται στον πίνακα 5.1.

Πίνακας 5.1: *Αξιολόγηση Πειραμάτων Αρχιτεκτονικής A1*

Αλγόριθμος Βελτιστοποίησης	**Ρυθμός Μάθησης**	**Test Loss**	**Test Accuracy**	**Εποχές**
Adam	0.01	4.9548	25%	11
Adam	0.005	3.0251	35%	8
Adam	0.001	2.1138	54%	6
Adam	0.0001	2.8507	55%	16
Adagrad	0.01	2.4003	56%	17
Adagrad	0.001	4.9602	54%	35
Adagrad	0.0005	5.0084	56%	100
Adagrad	0.0001	7.4401	36%	11
Adadelta	0.01	3.8538	56%	59
Adadelta	0.001	7.6903	31%	7
Adadelta	0.0005	7.8133	27%	18
Adadelta	0.0001	8.0502	25%	44

Τα παραπάνω αποτελέσματα δεν είναι ιδιαίτερα ικανοποιητικά, καθώς σύμφωνα με το [30] η ορθότητα θα έπρεπε να προσεγγίζει το 88%. Αυτή η μεγάλη απόκλιση κατά πάσα πιθανότητα οφείλεται στην παράλειψη του άρθρου όσον αφορά στον αλγόριθμο βελτιστοποίησης και στον αντίστοιχο ρυθμό μάθησης. Η παράλειψη αυτή μας οδήγησε σε αυθαίρετη επιλογή των παραμέτρων αυτών η οποία βάσει των αποτελεσμάτων δεν είναι η βέλτιστη. Επιπλέον, ενδεχομένως να υπάρχει και κάποια επιπλέον τεχνική ή παράμετρος η οποία ναι μεν χρησιμοποιήθηκε για την παραγωγή των αποτελεσμάτων του άρθρου, αλλά δεν συμπεριλαμβάνεται στην περιγραφή της μεθοδολογίας που ακολουθήθηκε. Για το λόγο αυτό θεωρήθηκε προτιμότερος ο πειραματισμός με μια διαφορετική προσέγγιση, όπως αυτή θα αναλυθεί στην επόμενη ενότητα.

5.2 Πειράματα Αρχιτεκτονικής Νευρωνικού A2

Στην παρούσα ενότητα θα παρουσιαστούν τα πειράματα που διεξήχθησαν χρησιμοποιώντας ως αρχιτεκτονική του δικτύου την Αρχιτεκτονική A2 (με τις κατάλληλες τροποποιήσεις ανάλο-

για τον τύπο της εισόδου, όπως αυτές αναλύθηκαν στον προηγούμενο κεφάλαιο). Όσον αφορά την μέθοδο εκπαίδευσης καθώς και τις τιμές των υπερπαραμέτρων της εκπαίδευσης, αυτές διατηρήθηκαν σταθερές για όλα τα πειράματα που διεξάχθηκαν. Οι διαφοροποιήσεις μεταξύ των πειραμάτων αφορούν στα χαρακτηριστικά που τροφοδοτήθηκαν στην είσοδο. Επιπλέον, εκτός από τα διαφορετικά χαρακτηριστικά στην είσοδο του νευρωνικου, πραγματοποιήθηκαν και ορισμένα πειράματα συνδυασμών των ξεχωριστών μοντέλων που προέκυψαν κατά την εκπαίδευση μέσω της μεθόδου της ψηφοφορίας (voting).

Το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας ως συνάρτηση κόστους την κατηγορική διασταυρούμενη εντροπία (categorical crossentropy loss), για συνολικά 200 εποχές. Το σύνολο δεδομένων εκπαίδευσης χωρίζεται σε επιμέρους τμήματα - παρτίδες με μέγεθος (batch size) 128 δείγματα η καθεμία, ενώ στο τέλος της κάθε εποχής το σύνολο αυτό ‘‘ανακατεύεται’’ (shuffle) και έπειτα χωρίζεται εκ νέου σε παρτίδες, με στόχο να αποφευχθεί το φαινόμενο της υπερεκπαίδευσης. Όσον αφορά τον αλγόριθμο βελτιστοποίησης, επιλέχθηκε ο Adam [25], με αρχική τιμή ρυθμού μάθησης 0.001, ενώ εφαρμόστηκε και η τεχνική της μείωσης του ρυθμού μάθησης όταν η τιμή της συνάρτησης κόστους για το σύνολο επικύρωσης (validation loss) σταματά να φθίνει (Reduce Learning Rate on Plateau) [39]. Πιο συγκεκριμένα, κάθε φορά που η τιμή της συνάρτησης κόστους για το σύνολο επικύρωσης δεν ελαττωνόταν πέραν της τρέχουσας ελάχιστης τιμής της για παραπάνω από 3 εποχές (διαδοχικές ή μη), τότε ο ρυθμός μάθησης μειωνόταν κατά 5%. Μετά από κάθε τέτοια μείωση δινόταν ένα περιθώριο 2 εποχών προτού το validation loss επανέλθει υπό παρακολούθηση, ενώ επιπλέον είχε τεθεί ένα κάτω όριο στις επιτρεπτές τιμές του ρυθμού μάθησης στο 0.000001, πέραν του οποίου δεν γινόταν περεταίρω μείωση.

Κατά την αξιολόγηση του μοντέλου, αφότου έγινε η πρόβλεψη για κάθε δείγμα του συνόλου δεδομένων αξιολόγησης (test set), οι προβλέψεις για δείγματα τα οποία πριν την τεχνική διαίρεσης κομματιού άνηκαν στο ίδιο κομμάτι συνδυάστηκαν μεταξύ τους. Στο σημείο αυτό, αξίζει να υπενθυμίσουμε πως το κάθε αρχικό κομμάτι διάρκειας 30 δευτερολέπτων διαχωρίστηκε σε επιμέρους τμήματα διάρκειας 1.5 δευτερολέπτων με 50% επικάλυψη των διαδοχικών τμημάτων. Επομένως, για κάθε κομμάτι παράχθηκαν 39 διαφορετικές προβλέψεις για το μουσικό είδος. Οι προβλέψεις αυτές συνδυάστηκαν μεταξύ τους με την τεχνική της ψηφοφορίας πλειοψηφίας (majority voting). Αυτό σημαίνει πως η τελική πρόβλεψη ενός κομματιού προκύπτει ως η πρόβλεψη που έχει παραχθεί για τα περισσότερα επιμέρους τμήματα.

5.2.1 Αποτελέσματα Πειραμάτων για Διαφορετικές Εισόδους

Ο πίνακας 5.2 περιλαμβάνει τα αποτελέσματα ως προς την ορθότητα (accuracy) που σημειώθηκε στο σύνολο δεδομένων αξιολόγησης (test set) από κάθε μοντέλο εκπαιδευμένο στον αντίστοιχο τύπο χαρακτηριστικών.

Από τον πίνακα αυτόν προκύπτουν 3 ενδιαφέρουσες παρατηρήσεις:

- **Μέγιστη ορθότητα στο μοντέλο που εκπαιδεύτηκε στον Μετασχηματισμό STFT**

  Το αποτέλεσμα αυτό μπορεί να θεωρηθεί αναμενόμενο, δεδομένων των μεγάλων διαστάσεων του μετασχηματισμού STFT (513×130). Οι αυξημένες διαστάσεις υποδηλώνουν πως το μεγαλύτερο μέρος της πληροφορίας τόσο στη διάσταση του χρόνου

Πίνακας 5.2: *Αξιολόγηση Πειραμάτων Αρχιτεκτονικής A2*

Τύπος Εισόδου	**Test Accuracy - no Voting**	**Test Loss - no Voting**	**Test Accuracy - Voting**
STFT	81%	0.4178	90%
Chroma STFT	10%	2.3026	10%
Mel Spectrogram	76%	0.9045	82%
MFCC	70%	0.8799	80%
MFCC και παράγωγοι	73%	1.2033	84%
CQT	76%	0.8033	86%

όσο και σε αυτή των συχνοτήτων έχει διατηρηθεί και μάλιστα είναι αρκετά χρήσιμο για το μοντέλο ώστε, βάσει αυτού, να μπορέσει να εξάγει συμπεράσματα αναφορικά με το μουσικό είδος.

- **Αξιοσημείωτα χαμηλή ορθότητα στο μοντέλο που εκπαιδεύτηκε στον Μετασχηματισμό Chroma STFT**

  Η συμπεριφορά αυτή είναι και πάλι αναμενόμενη και μπορεί να ερμηνευθεί με βάσει το περιεχόμενο του μετασχηματισμού Chroma STFT. Όπως περιγράφηκε αναλυτικά στο προηγούμενο κεφάλαιο, ο μετασχηματισμός Chroma STFT αποτελεί μία αναπαράσταση του χρώματος των μουσικών τόνων ανά σταθερά προκαθορισμένα χρονικά παράθυρα. Διαισθητικά, μπορεί κανείς να καταλάβει πως το μουσικό είδος δεν μπορεί να συσχετιστεί αποκλειστικά με τις νότες οι οποίες περιλαμβάνονται σε ένα μουσικό κομμάτι. Αυτό επιβεβαιώνεται και από τις επιδόσεις του αλγορίθμου οι οποίες είναι σημαντικά χαμηλότερες στην περίπτωση αυτή, σε σχέση με όλες τις υπόλοιπες περιπτώσεις χαρακτηριστικών εισόδου.

- **Σημαντική βελτίωση της ορθότητας με την εφαρμογή της τεχνικής της ψηφοφορίας (voting)**

  Όπως μπορούμε να δούμε και στον πίνακα, στις περισσότερες περιπτώσεις χαρακτηριστικών εισόδου, παρατηρείται μια αύξηση περίπου 10% στην ορθότητα με την εφαρμογή της τεχνικής majority voting. Αυτό συμβαίνει διότι έχοντας ένα σχετικά μεγάλο πλήθος ‘‘ψηφοφόρων’’, δίνεται η δυνατότητα τυχόν λάθη που έχουν προκύψει να διορθωθούν μέσω των προβλέψεων για γειτονικά μουσικά τμήματα. Τα λάθη αυτά μπορεί να προκύψουν είτε λόγω της μικρής διάρκειας των τμημάτων η οποία ενδέχεται να δημιουργήσει σύγχυση στο μοντέλο, είτε λόγω των διαφορετικών διακυμάνσεων του μουσικού είδους μέσα στο ίδιο κομμάτι. Ένα παράδειγμα αυτού θα μπορούσε να είναι ένα κατά βάση ροκ κομμάτι που περιέχει υποπεριοχές που προσομοιάζουν περισσότερο σε μέταλ ή ποπ.

Κλείνοντας την αξιολόγηση των πειραμάτων διαφορετικών εισόδων, κρίθηκε χρήσιμος ο υπολογισμός των πινάκων σύγχυσης, οι οποίοι φαίνονται συγκεντρωτικά στην εικόνα 5.1. Από αυτούς τους πίνακες μπορούμε να συμπεράνουμε για το κάθε μοντέλο ποιες είναι οι κλάσεις στις οποίες έχει καλύτερες επιδόσεις και ποιες είναι αυτές που προκαλούν τη μεγαλύτερη σύγχυση.

5.2.2 Αποτελέσματα Πειραμάτων Συνδυασμού Εισόδων

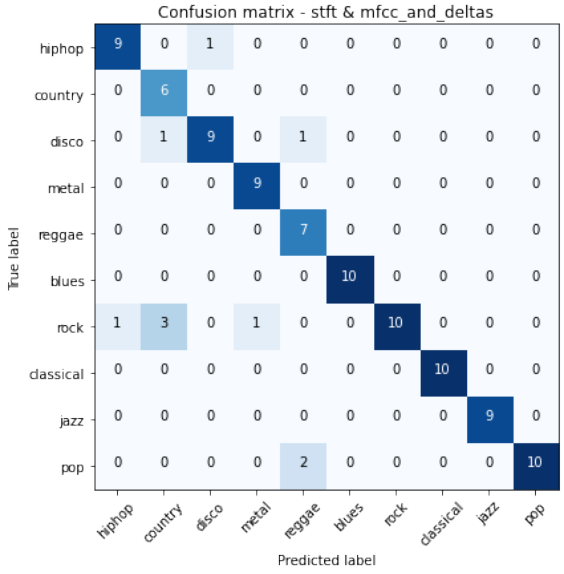
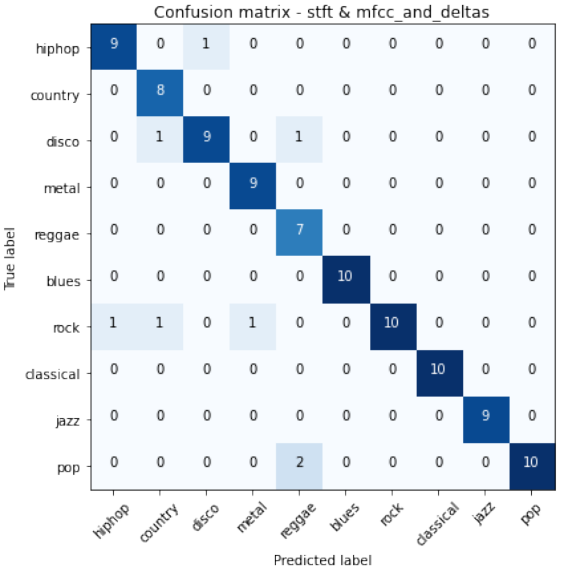
Παρατηρώντας τους πίνακες σύγχυσης του σχήματος 5.1, όπου φαίνονται οι επιδόσεις ανά κλάση για τα μοντέλα εκπαιδευμένα σε διαφορετικά χαρακτηριστικά, είναι εμφανές πως υπάρχουν ορισμένα μοντέλα τα οποία αποδίδουν με συμπληρωματικό τρόπο. Αυτό σημαίνει πως το ένα μοντέλο είναι καλύτερο στις κλάσεις που το δεύτερο μοντέλο εμφανίζει αδυναμίες και το αντίστροφο. Η παρατήρηση αυτή υπήρξε το κίνητρο για την διεξαγωγή πειραμάτων για τον έλεγχο της δυνατότητας του ενός μοντέλου να διορθώνει το άλλο (και το αντίστροφο) με στόχο την δημιουργία ενός υβριδικού μοντέλου με συνολικά καλύτερες επιδόσεις.

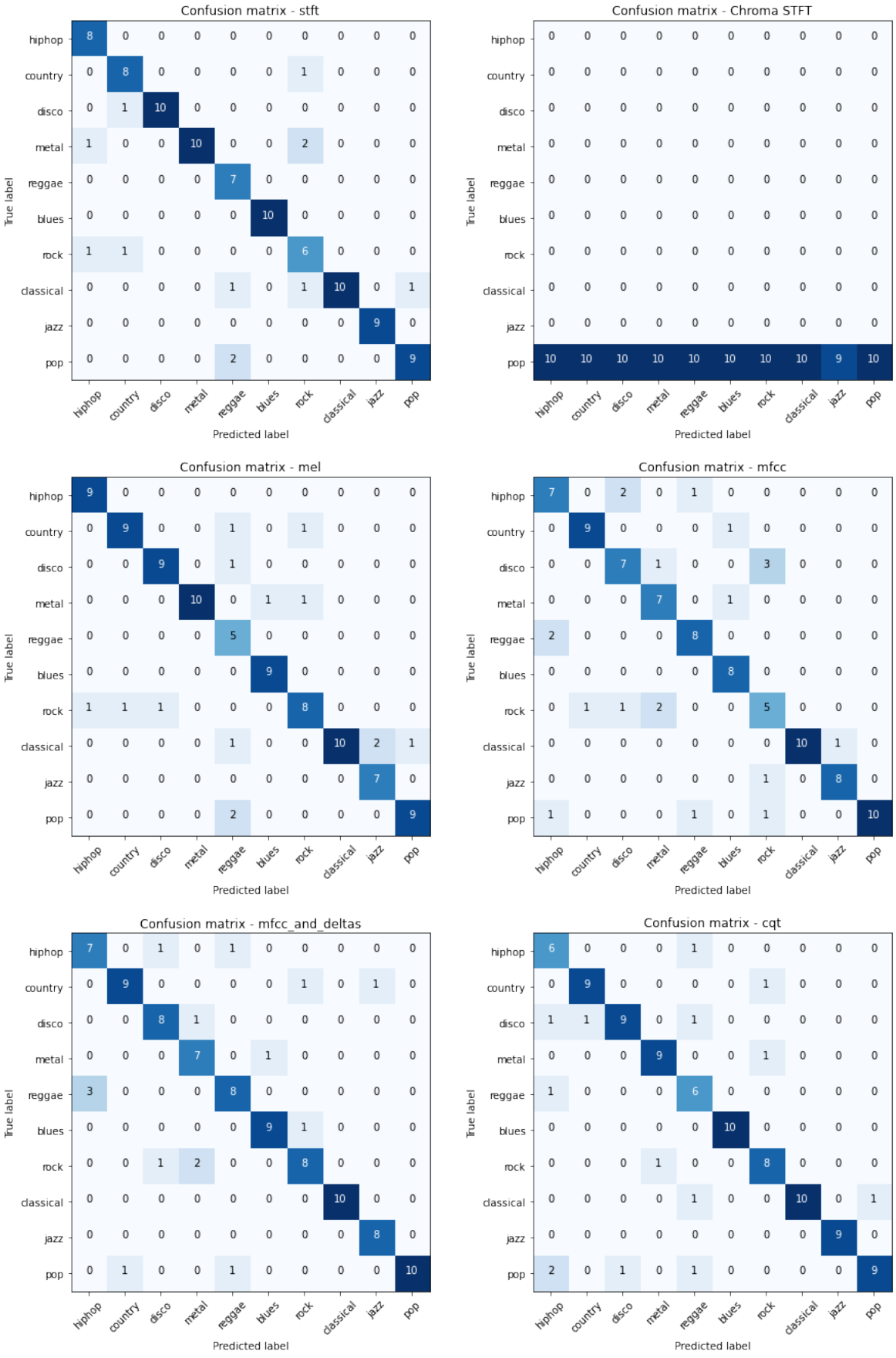
Στη λογική αυτή δοκιμάστηκαν δύο διαφορετικοί τρόπου συνδυασμού των επιμέρους μοντέλων. Η βασική λογική πίσω από αυτούς τους συνδυασμούς ήταν η διενέργεια ‘‘ψηφοφορίας’’ μεταξύ των προβλέψεων των μοντέλων. Πιο συγκεκριμένα, στην πρώτη περίπτωση, εφαρμόστηκε η τεχνική του Hard Voting όπου, για το κάθε μουσικό κομμάτι γινόταν η απόλυτη πρόβλεψη για το κάθε επιμέρους τμήμα του από το κάθε μοντέλο και η τελική απόφαση προέκυπτε από την πρόβλεψη με τη μεγαλύτερη συχνότητα εμφάνισης. Στη δεύτερη περίπτωση, εφαρμόστηκε η τεχνική του Soft Voting κατά την οποία λαμβάνεται υπόψιν όχι μόνο η απόλυτη πρόβλεψη του κάθε μοντέλου αλλά και η ‘‘σιγουριά’’ (confidence) με την οποία λαμβάνεται αυτή η πρόβλεψη. Αυτό μεταφράζεται ως ο υπολογισμός του μέσου όρου της πιθανότητας να ανήκει το επιμέρους τμήμα σε κάθε κλάση, για όλα τα τμήματα και για όλα τα μοντέλα. Εν τέλει, η πρόβλεψη προκύπτει ως η κλάση στην οποία αντιστοιχεί η μεγαλύτερη πιθανότητα.

Όσον αφορά τα επιμέρους μοντέλα που συνδυάστηκαν, επιλέχθηκε το μοντέλο που εκπαιδεύτηκε πάνω στο μετασχηματισμός STFT και το μοντέλο που εκπαιδεύτηκε πάνω στο μετασχηματισμό MFCC μαζί με τις παραγώγους του. Η επιλογή αυτού του συνδυασμού έγινε μετά από την παρατήρηση των αντίστοιχων πινάκων σύγχυσης. Πιο συγκεκριμένα, είναι εμφανές ότι το πρώτο μοντέλο παρουσιάζει αδυναμία στις κλάσεις reggae και rock, στις οποίες το δεύτερο μοντέλο έχει καλύτερες επιδόσεις. Επιπλέον, το δεύτερο μοντέλο με τη σειρά του φαίνεται να δυσκολεύεται στις κλάσεις hiphop και metal, ενώ ταυτόχρονα το πρώτο μοντέλο αποδίδει καλύτερα σε αυτές.

Τα αποτελέσματα ανά μέθοδο συνδυασμού φαίνονται στον πίνακα 5.3. Παρατηρούμε πως η μέθοδος Soft Voting έφερε τα καλύτερα αποτελέσματα, με τελικό ποσοστό ορθότητας 92%. Επίσης, είναι εμφανής και η επιτυχία της μεθόδου αυτής στη διόρθωση των λαθών του ενός μοντέλου από το άλλο. Συγκεκριμένα, βλέπουμε πως από τις 4 κλάσεις (hiphop, metal, reggae, rock) όπου τα δύο μοντέλα εμφάνιζαν συμπληρωματικά λάθη, οι 3 έχουν πλέον μεγαλύτερα ποσοστά ορθής αναγνώρισης, ενώ στη 1 το ποσοστό παρέμεινε ίδιο.

Πίνακας 5.3: *Αξιολόγηση Μεθόδων Συνδυασμού Μοντέλων*

Hard Voting	**Soft Voting**
	
Accuracy: 90%	Accuracy: 92%



Σχήμα 5.1: *Πίνακες Σύγχυσης Αρχιτεκτονικής Α2 - Διαφορετικές Είσοδοι*

Μέρος III

Πρακτικό Μέρος Β΄ - Ταξινόμηση σε μουσικό υποείδος

Κεφάλαιο 6

Δεδομένα και Προεπεξεργασία

Στο κεφάλαιο αυτό γίνεται αναφορά στα δεδομένα που χρησιμοποιήθηκαν στο δεύτερο μέρος της διπλωματικής εργασίας. Η πρώτη ενότητα του κεφαλαίου περιλαμβάνει πληροφορίες σχετικά με το σύνολο δεδομένων και τα χαρακτηριστικά του. Η δεύτερη ενότητα εστιάζει στην προεπεξεργασία που εφαρμόστηκε στο σύνολο δεδομένων αυτό για τη μετατροπή του στην κατάλληλη μορφή ώστε να μπορέσει να τροφοδοτηθεί σε ένα νευρωνικό δίκτυο.

6.1 Σύνολο Δεδομένων

6.1.1 Σύνολο FMA

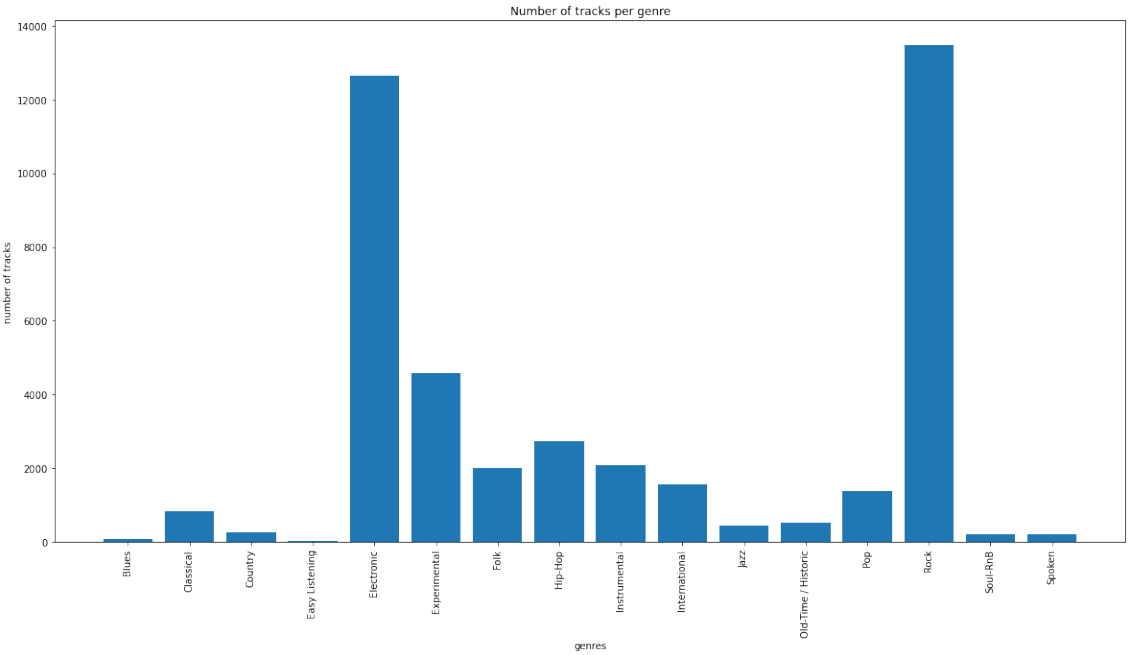
Τα δεδομένα που χρησιμοποιήθηκαν στο δεύτερο μέρος της εργασίας αποτελούν μέρος του συνόλου FMA. Πρόκειται για ένα σύνολο δεδομένων το οποίο περιλαμβάνει αρχεία ήχου αδειοδοτημένα μέσω Creative Commons συνολικού όγκου 917 GiB (343 μέρες σε διάρκεια). Το ηχητικό υλικό συνίσταται από 106,574 κομμάτια 16,341 καλλιτεχνών και 14,854 άλμπουμ, οργανωμένα σε μια ιεραρχία 161 μουσικών ειδών. Από αυτά τα 161 μουσικά είδη, τα 16 αποτελούν βασικά είδη ενώ τα υπόλοιπα είναι υποείδη αυτών σε διάφορα επίπεδα (βάσει συγκεκριμένης ιεραρχίας). Το μεγαλύτερο μέρος από τα 106,574 κομμάτια είναι κωδικοποιημένα σε μορφή mp3, με ρυθμό δειγματοληψίας 44,100 Hz και bitrate 320 kbit/s.

Το FMA είναι ένα σύνολο δεδομένων ιδανικό για το πρόβλημα της Αναγνώρισης Μουσικού Υποείδους, καθώς περιλαμβάνει πολύ λεπτομερείς πληροφορίες σχετικά με το είδος (όπως για παράδειγμα πολλαπλά μουσικά είδη ανά κομμάτι) και είναι οργανωμένο βάσει μιας εσωτερικής ιεραρχίας αναφορικά με τα μουσικά είδη. Επιπλέον, τα είδη αυτά έχουν αποδοθεί σε κάθε κομμάτι από τους ίδιους του καλλιτέχνες, γεγονός ενδεικτικό της ποιότητας της επισήμανσης. Είναι διαθέσιμο σε 3 διαφορετικές μορφές/υποσύνολα:

- FMA small
- FMA medium
- FMA full

Για τους σκοπούς της παρούσας εργασίας θα χρησιμοποιηθεί μέρος του FMA medium. Το σύνολο αυτό συνίσταται από τα κομμάτια του ευρύτερου συνόλου τα οποία διαθέτουν μία και μοναδική μοναδική επισήμανση από τα 16 βασικά είδη (και μία ή περισσότερες από τα αντίστοιχα

μουσικά υποείδη). Συνολικά αποτελείται από 25,000 κομμάτια, διάρκειας 30 δευτερολέπτων το καθένα. Είναι μη ισορροπημένο, καθώς οι 16 επιμέρους κλάσεις περιλαμβάνουν από 21 έως 7,103 δείγματα η καθεμία, όπως φαίνεται και στην εικόνα 6.1 [40].

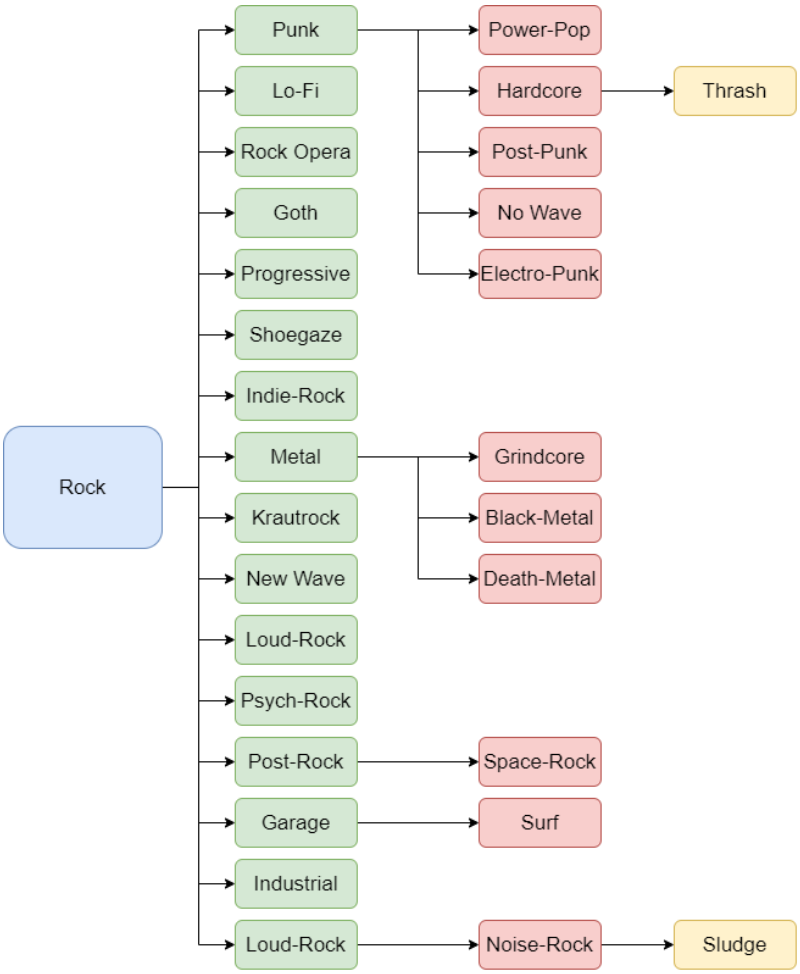


Σχήμα 6.1: *Κατανομή συνόλου FMA medium*

6.1.2 Υποσύνολο του FMA για την Ταξινόμηση σε Μουσικό Υποείδος

Καθώς το δεύτερο μέρος της διπλωματικής εργασίας εστιάζει στη δημιουργία ενός μοντέλου για την αναγνώριση μουσικού υποείδους, θα συλλεχθεί ένα υποσύνολο του FMA medium ώστε να καλυφθούν οι ανάγκες αυτού του προβλήματος. Στο σχήμα 6.2 παρουσιάζεται η ιεραρχία των υποειδών της Rock μουσικής, όπως αυτή οργανώνεται στο σύνολο FMA. Όπως μπορούμε να παρατηρήσουμε, η rock μουσική πρόκειται για ένα μουσικό είδος το οποίο διαθέτει πληθώρα υποειδών, επομένως συνιστά ιδανική περίπτωση για πρόβλημα αναγνώρισης μουσικού υποείδους.

Ένα σημείο που αξίζει αναφοράς είναι πως ενώ η χρήση του συνόλου FMA medium μας εξασφαλίζει πως τα κομμάτια που το αποτελούν διαθέτουν μοναδική επισήμανση είδους, δεν είναι βέβαιο πως η επισήμανση υποείδους είναι επίσης μοναδική. Έτσι λοιπόν, δεδομένου ότι επιχειρούμε να επιλύσουμε το πρόβλημα ταξινόμησης ενός κομματιού σε ένα και μοναδικό μουσικό υποείδος, είναι αναγκαίο να φροντίσουμε και τα δείγματα του συνόλου δεδομένων που θα χρησιμοποιηθούν στην εκπαίδευση να πληρούν αυτή την προϋπόθεση. Έτσι λοιπόν, επιλέχθηκαν τα κομμάτια τα οποία διαθέτουν μία και μοναδική επισήμανση από το πρώτο επίπεδο υποειδών. Κάποιες κλάσεις από αυτές περιλάμβαναν ένα πολύ μικρό αριθμό δειγμάτων, επομένως κρίθηκε ότι δεν ήταν επαρκώς αντιπροσωπούμενες και απορρίφθηκαν. Τελικά καταλήξαμε με ένα σύνολο 3,663 κομματιών 30 δευτερολέπτων κατανεμημένα σε 10 κλάσεις όπως φαίνεται στο σχήμα 6.3.



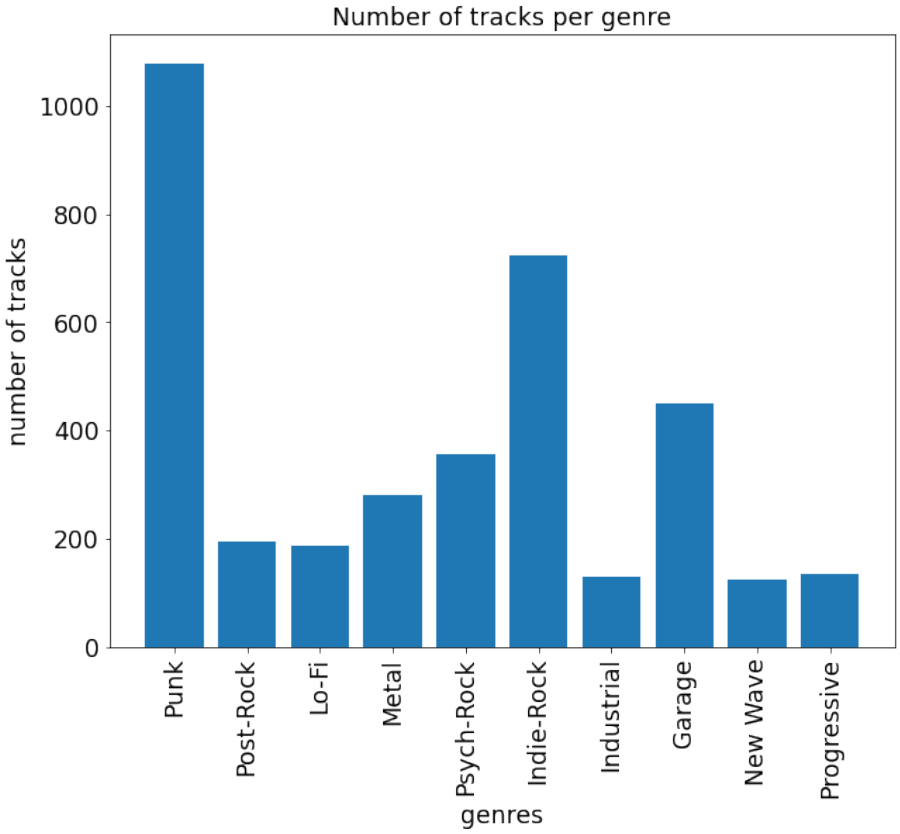
Σχήμα 6.2: *Ιεραρχία Υποειδών Rock*

6.2 Προεπεξεργασία και Εξαγωγή Χρήσιμων Χαρακτηριστικών

Όπως περιγράφηκε αναλυτικά στο Κεφάλαιο 3, ένα ηχητικό σήμα στην ακατέργαστη μορφή του, δηλαδή στη μορφή κυματομορφής, δεν μπορεί να χρησιμοποιηθεί για την εκπαίδευση ενός νευρωνικού δικτύου για την αναγνώριση μουσικού είδους (και συνεπώς μουσικού υποείδους). Στην ενότητα αυτή θα αναλυθούν οι τρόποι με τους οποίους το αρχικό ηχητικό σήμα μετατρέπεται σε μια μορφή χρήσιμη για να τροφοδοτηθεί ως σύνολο εκπαίδευσης στο νευρωνικό δίκτυο. Ορισμένα βήματα της προεπεξεργασίας είναι κοινά με τα βήματα που εφαρμόστηκαν στην προεπεξεργασία των δεδομένων στο Πρακτικό Μέρος Α΄, επομένως δεν θα γίνει εκτενής περιγραφή σε αυτά και θα υπάρχει παραπομπή στην αντίστοιχη ενότητα.

6.2.1 Διαίρεση Κομματιού

Το πρώτο βήμα στη διαδικασία προεπεξεργασίας που εφαρμόστηκε είναι κοινό με το πρώτο βήμα της διαδικασίας προεπεξεργασίας του συνόλου δεδομένων για την αναγνώριση μουσικού είδους. Όπως αναλύθηκε στην ενότητα 3.2.1, η τεχνική αυτή μπορεί να βοηθήσει σημαντικά στην αύξηση του πλήθους των δειγμάτων εισόδου, δίχως την απώλεια σημαντικής πληροφο-



Σχήμα 6.3: *Κατανομή στα υποείδη της Rock*

ρίας. Όσον αφορά τον ακριβή τρόπο διαίρεσης, εφαρμόστηκε η ίδια διαδικασία που χρησιμοποιήθηκε για τα πειράματα της Αρχιτεκτονικής Α2 (ενότητα 5.2), δηλαδή κάθε αρχείο ήχου διάρκειας 30 δευτερολέπτων διαχωρίστηκε σε τμήματα διάρκειας 1.5 δευτερολέπτων με 50% επικάλυψη μεταξύ των διαδοχικών τμημάτων. Ετσι λοιπόν, από κάθε αρχείο ήχου διάρκειας 30 δευτερολέπτων, προέκυψαν 39 επιμέρους τμήματα.

6.2.2 Εξαγωγή Μετασχηματισμού STFT και Κανονικοποίηση

Το επόμενο βήμα στη διαδικασία προεπεξεργασίας ήταν η εφαρμογή ενός μετασχηματισμού ώστε το κάθε δείγμα να μετατραπεί από το πεδίο του χρόνου, στο πεδίο χρόνου-συχνότητας (time-frequency domain). Για το σκοπό αυτό εφαρμόστηκε ο μετασχηματισμός STFT ο οποίος, με βάση τα πειράματα που διεξάχθηκαν στο Πρακτικό Μέρος Α΄, φαίνεται να ενσωματώνει την πληροφορία με το βέλτιστο δυνατό τρόπο ώστε να γίνει αντιληπτή από ένα νευρωνικό δίκτυο για την αναγνώριση του μουσικού είδους. Η εξαγωγή του μετασχηματισμού έγινε και πάλι με τη βοήθεια της βιβλιοθήκης librosa [27], χρησιμοποιώντας τις παραμέτρους $n\_fft = 1024$, $hop\_length = 256$. Το αποτέλεσμα είναι ένας πίνακας διαστάσεων 513×130 για κάθε δείγμα του συνόλου δεδομένων, όπου η πρώτη διάσταση είναι τα εύρη συχνοτήτων και η δεύτερη τα χρονικά παράθυρα.

Επιπλέον, εφαρμόστηκε η τεχνική της κανονικοποίησης στους πίνακες που προέκυψαν. Πιο συγκεκριμένα, ο στόχος ήταν το σύνολο δεδομένων να ακολουθεί την τυποποιημένη κανονική κατανομή ώστε η μέση τιμή να είναι μηδενική και η τυπική απόκλιση ίση με τη μονάδα. Για

το σκοπό αυτό υπολογίστηκε η μέση τιμή μ και η τυπική απόκλιση σ πάνω στο σύνολο των μετασχηματισμένων δεδομένων και η κανονικοποίηση έγινε ως εξής:

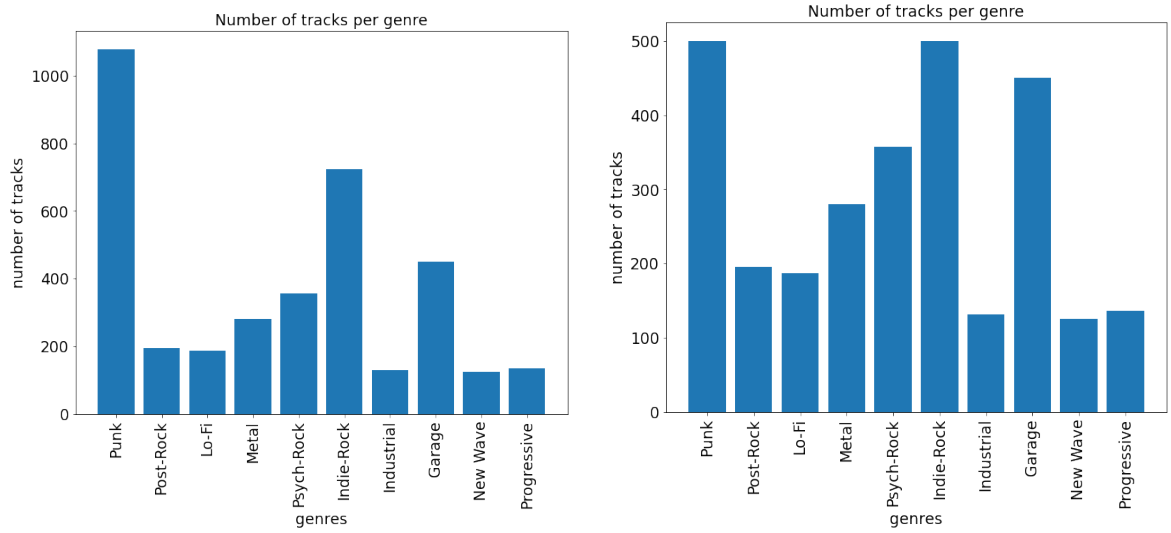
$$X_{norm} = \frac{X - \mu}{\sigma} \tag{6.1}$$

όπου X_{norm} είναι το κανονικοποιημένο δείγμα και X το αρχικό.

6.2.3 Υποδειγματοληψία

Όπως μπορεί κανείς να παρατηρήσει στην εικόνα 6.3 υφίσταται μια μεγάλη ανισορροπία μεταξύ των κλάσεων του συνόλου δεδομένων. Αυτή η ανισορροπία μπορεί να επηρεάσει σημαντικά τις επιδόσεις του μοντέλου, καθώς οι λιγότερο αντιπροσωπούμενες κλάσεις ενδέχεται εν τέλει να αγνοηθούν. Ένας τρόπος να αντιμετωπιστεί αυτό το πρόβλημα είναι η τεχνική της Υποδειγματοληψίας (Undersampling), κατά την οποία ορισμένα δείγματα των κλάσεων με τη μεγαλύτερη εκπροσώπηση διαγράφονται [41]. Υπάρχουν αρκετοί διαφορετικοί τρόποι με τους οποίους μπορεί να γίνει η επιλογή των δειγμάτων που θα διαγραφούν. Ο πιο απλοϊκός τρόπος είναι η τυχαία επιλογή (Random Undersampling) τον οποίο και θα χρησιμοποιήσουμε στην περίπτωσή μας.

Δεδομένου του πολύ μικρού πλήθους δεδομένων στη λιγότερο αντιπροσωπούμενη κλάση (126 δείγματα στην κλάση New Wave), η τυχαία υποδειγματοληψία με στόχο την δημιουργία ενός πλήρως ισορροπημένου συνόλου δεδομένων, με 126 δείγματα ανά κλάση δεν ήταν μια ρεαλιστική επιλογή. Αυτό διότι με αυτό τον τρόπο το σύνολο δεδομένων θα μειωνόταν δραματικά και έτσι δεν θα μπορούσε να εξαχθεί πληροφορία κατά την εκπαίδευση του νευρωνικού. Για το λόγο αυτό, θεωρήθηκε προτιμότερο να εφαρμοστεί η τεχνική της δειγματοληψίας μόνο στις 2 κλάσεις με το μεγαλύτερο πλήθος δειγμάτων (Punk και Industrial). Στις κλάσεις αυτές επιλέχθηκαν τυχαία δείγματα προς διαγραφή ώστε το τελικό πλήθος των δειγμάτων τους να είναι 500. Τελικά, η κατανομή των δειγμάτων στις κλάσεις φαίνεται στην εικόνα 6.4.



Σχήμα 6.4: *Κατανομή στα υποείδη της Rock πριν και μετά την Υποδειγματοληψία*

Όπως μπορεί κανείς να παρατηρήσει, το σύνολο δεδομένων δεν είναι και πάλι ισορροπημένο,

όμως τα δείγματα είναι πιο ομαλά κατανεμημένα ανάμεσα στις κλάσεις. Επίσης, εξαλείφθηκαν οι ακραίες διαφορές μεταξύ των κλάσεων και επομένως αποφεύγεται το φαινόμενο της προκατάληψης υπέρ των κλάσεων με τη μεγαλύτερη εκπροσώπηση εκ μέρους του μοντέλου.

6.2.4 Επαύξηση Δεδομένων (Data Augmentation)

Η επαύξηση δεδομένων είναι μία τεχνική η οποία χρησιμοποιείται με σκοπό την αύξηση του πλήθους των δεδομένων, μέσω της προσθήκης ελαφρώς τροποποιημένων αντιγράφων των ήδη υπαρχόντων δειγμάτων ή και συνθετικών δεδομένων. Η τεχνική είναι ένας από τους τρόπους αντιμετώπισης του φαινομένου της υπερεκπαίδευσης [42].

Δεδομένου ότι στην περίπτωσή μας τα δεδομένα του συνόλου εκπαίδευσης είναι στη μορφή πινάκων 2 διαστάσεων που αναπαριστούν τον STFT μετασχηματισμό, είναι αναγκαίο οι τροποποιήσεις που θα εφαρμοστούν στα αντίγραφα των δειγμάτων να γίνονται με τέτοιο τρόπο ώστε το προκύπτον δείγμα να διατηρεί τα χαρακτηριστικά ενός STFT μετασχηματισμού. Για το σκοπό αυτό ακολουθήθηκε η μεθοδολογία που παρουσιάζεται στο [43], όπου περιγράφεται αναλυτικά ένα πλαίσιο επαύξησης δεδομένων εφαρμόσιμο σε φασματογραφήματα.

Πιο συγκεκριμένα, εφαρμόστηκαν δύο μέθοδοι επαύξησης δεδομένων:

- **Εφαρμογή μάσκας στον άξονα των συχνοτήτων (Frequency Masking)**

  Κατά τη μέθοδο αυτή επιλέγεται με τυχαίο τρόπο (βάσει ομοιόμορφης κατανομής) μία ζώνη που εκτείνεται σε όλο τον άξονα του χρόνου και σε ένα τμήμα του άξονα των συχνοτήτων που αντιστοιχεί σε ένα ποσοστό από 0% έως 20% του τελευταίου. Η τιμές του μετασχηματισμού STFT σε αυτή τη ζώνη τίθενται ίσες με το μηδέν. Δεδομένου μάλιστα ότι το σύνολο των δεδομένων είναι κανονικοποιημένο ώστε να ακολουθεί την τυποποιημένη κανονική κατανομή, η τιμή 0 αντιστοιχεί στη μέση τιμή του συνόλου.

- **Εφαρμογή μάσκας στον άξονα του χρόνου (Time Masking)**

  Κατ΄ αναλογία, σύμφωνα με τη μέθοδο αυτή επιλέγεται μία ζώνη που εκτείνεται σε όλο τον άξονα των συχνοτήτων και σε ένα τμήμα του άξονα του χρόνου που αντιστοιχεί σε ένα ποσοστό από 0% έως 20% του τελευταίου. Η τιμές του μετασχηματισμού STFT σε αυτή τη ζώνη τίθενται ίσες με το μηδέν, δηλαδή ίσες με τη μέση τιμή του συνόλου δεδομένων, αν αυτό ακολουθεί τυποποιημένη κανονική κατανομή.

Οι δύο παραπάνω μέθοδοι εφαρμόζονται στα δεδομένα με τυχαίο τρόπο κατά τη διάρκεια της εκπαίδευσης.

6.2.5 Διαχωρισμός Συνόλου Δεδομένων

Το τελευταίο βήμα πριν προχωρήσουμε στη διαδικασία της εκπαίδευσης είναι ο διαχωρισμός των δειγμάτων σε 3 επιμέρους υποσύνολα: το σύνολο εκπαίδευσης (training set), το σύνολο επικύρωσης (validation set) και το σύνολο ελέγχου (test set). Η αναλογία διαχωρισμού είναι 80%-10%-10% αντίστοιχα, δηλαδή ίδια με αυτή που ορίστηκε στην ενότητα 3.2.3 του Πρακτικού Μέρους Α΄. Αξίζει στο σημείο αυτό να αναφερθεί πως η αναλογία αυτή τηρήθηκε για κάθε επιμέρους κλάση, ώστε η κατανομή στις κλάσεις να είναι κοινή και για τα 3 υποσύνολα.

Κεφάλαιο 7

Αρχιτεκτονικές Νευρωνικών Δικτύων

Το κεφάλαιο αυτό περιλαμβάνει πληροφορίες αναφορικά με τις αρχιτεκτονικές που χρησιμοποιήθηκαν στο Πρακτικό Μέρος Β΄ της διπλωματικής εργασίας, δηλαδή κατά τη δημιουργία ενός συστήματος αναγνώρισης μουσικού υποείδους. Οι αρχιτεκτονικές αυτές είναι στην πλειοψηφία τους εμπνευσμένες από τη δουλειά που παροουσιάζεται στο [7]. Επιπλέον, στον πίνακα 7.1 φαίνεται το πλήθος παραμέτρων (συνολικών και εκπαιδεύσιμων) για την κάθε αρχιτεκτονική που χρησιμοποιήθηκε.

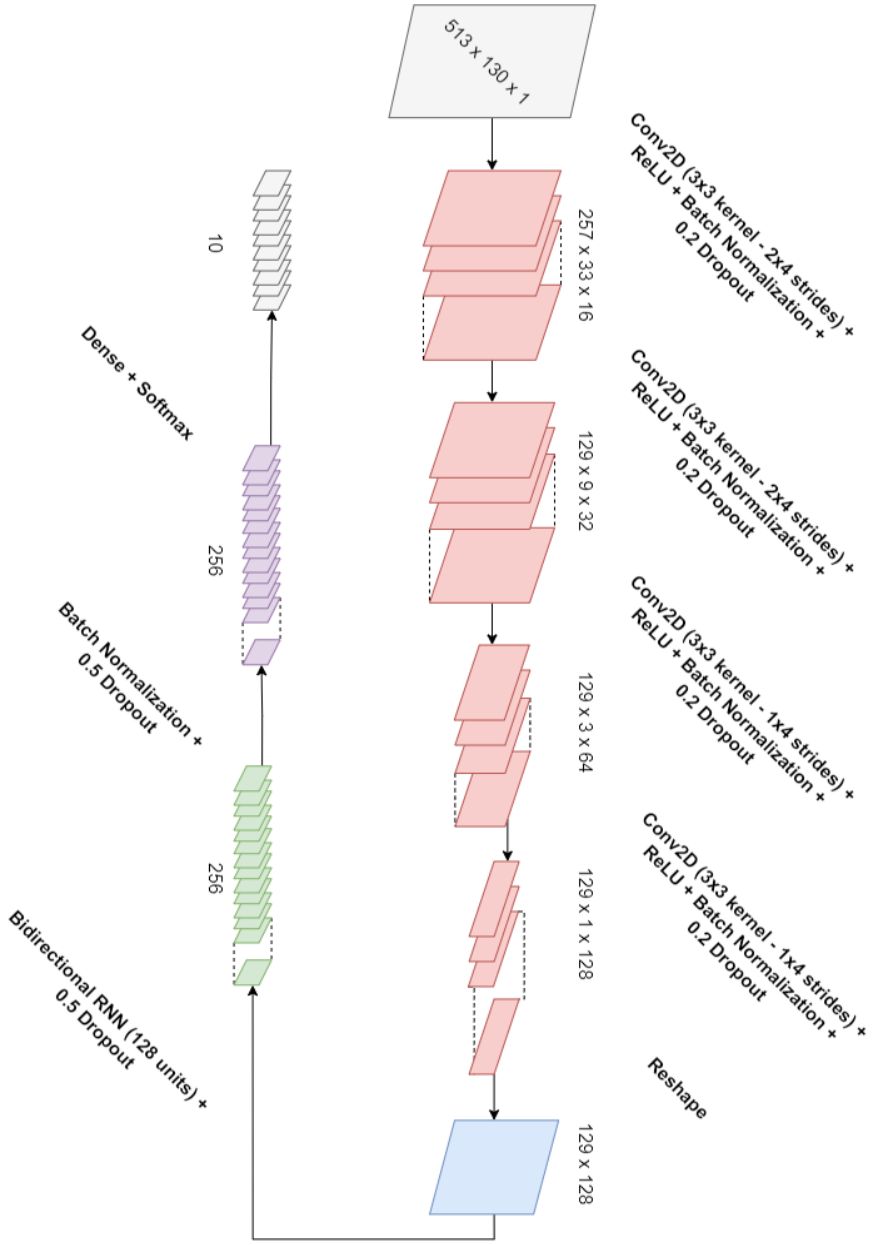
Πίνακας 7.1: *Πλήθος παραμέτρων Αρχιτεκτονικών Μέρους Β΄*

Αρχιτεκτονική	**Πλήθος παραμέτρων**	**Πλήθος εκπαιδεύσιμων παραμέτρων**
Αρχιτεκτονική B1	300,618	299,626
Αρχιτεκτονική B2	475,338	473,738
Αρχιτεκτονική B3	474,666	473,562
Αρχιτεκτονική B4	846,310	845,442
Αρχιτεκτονική B5	376,970	376,362

7.1 Αρχιτεκτονική Νευρωνικού B1

Όπως μπορούμε να παρατηρήσουμε στο Σχήμα 7.1, το Νευρωνικό Δίκτυο B1 είναι ένα δίκτυο CRNN. Με τον όρο αυτό εννοούμε ένα δίκτυο που συνδυάζει CNN επίπεδα των οποίων ο ρόλος είναι η εξαγωγή χαρακτηριστικών και RNN που συνδυάζουν αυτά τα χαρακτηριστικά, λαμβάνοντας υπόψιν τις χρονικές εξαρτήσεις μεταξύ τους. Πιο συγκεκριμένα, το CNN τμήμα αποτελείται από 4 διαδοχικά μπλοκ, το καθένα από τα οποία συνίσταται από ένα συνελικτικό επίπεδο με συνάρτηση ενεργοποίησης ReLU, ένα επίπεδο κανονικοποίησης δέσμης (Batch Normalization) και ένα επίπεδο εγκατάλειψης (Dropout Layer) με πιθανότητα 0.2. Όσον αφορά τα συνελικτικά επίπεδα, αυτά διαθέτουν πυρήνα διαστάσεων 3×3 και βήμα 2×4 ή 1×4. Ο ρόλος του μη μοναδιαίου βήματος στα συνελικτικά επίπεδα είναι η μείωση των διαστάσεων, καθώς όπως βλέπουμε δεν υπάρχουν επίπεδα Max Pooling. Εν τέλει, μετά και από ένα επίπεδο Reshape, προκύπτει ένας πίνακας χαρακτηριστικών μεγέθους 129×128. Αυτός ο πίνακας τροφοδοτείται σε ένα αναδρομικό επίπεδο διπλής κατεύθυνσης (Bidirectional RNN), με 128 μονάδες (units) και εγκατάλειψη με πιθανότητα 0.5, στην έξοδο του οποίου προκύπτει ένα διάνυσμα μήκους 256 θέσεων. Ακολούθως, το διάνυσμα αυτό αφού περάσει

από ένα επίπεδο κανονικοποίησης δέσμης, διέρχεται από ένα πλήρως συνδεδεμένο επίπεδο με συνάρτηση ενεργοποίησης softmax, η έξοδος του οποίου είναι ένα διάνυσμα 10 θέσεων το οποίο αντιστοιχεί στην πιθανότητα το δείγμα εισόδου να ανήκει σε κάθε μία από τις 10 πιθανές κλάσεις. Εφόσον η συνάρτησης ενεργοποίησης του τελευταίου επιπέδου είναι softmax, οι πιθανότητες αυτές αθροίζουν στη μονάδα.



Σχήμα 7.1: *Αρχιτεκτονική Νευρωνικού B1*

7.2 Αρχιτεκτονική Νευρωνικού B2

Στο Σχήμα 7.2 φαίνεται η αρχιτεκτονική του Νευρωνικού Δικτύου B2. Όπως μπορεί κανείς να παρατηρήσει, πρόκειται για ένα δίκτυο με παράλληλη αρχιτεκτονική, αποτελούμενο από δύο μέρη. Το πρώτο μέρος είναι αποτελείται στο σύνολό του από 5 συνελικτικά μπλοκ.

Το κάθε μπλοκ συνίσταται από ένα συνελικτικό επίπεδο με πυρήνα διαστάσεων 3×3 και συνάρτηση ενεργοποίησης ReLU, ένα επίπεδο κανονικοποίησης δέσμης (Batch Normalization), ένα επίπεδο Max Pooling το οποίο μειώνει τις διαστάσεις στο 50% ή στο 25% και ένα επίπεδο εγκατάλειψης (Dropout Layer) με πιθανότητα 0.2. Η έξοδος του τελευταίου μπλοκ επιπεδοποιείται (Flattened) σε ένα διάνυσμα 256 θέσεων. Το δεύτερο παράλληλο μέρος είναι επί της ουσίας ίδιο με την Αρχιτεκτονική B1 που περιγράφηκε στην προηγούμενη ενότητα, και πιο συγκεκριμένα μέχρι και το αναδρομικό επίπεδο διπλής κατεύθυνσης (Bidirectional RNN). Η έξοδος του αναδρομικού επιπέδου είναι ένα διάνυσμα 256 θέσεων, το οποίο συνενώνεται (Concatenate) με το αντίστοιχο διάνυσμα που προέκυψε από το πρώτο παράλληλο μέρος, με αποτέλεσμα τη δημιουργία ενός διανύσματος 512 θέσεων. Ακολούθως, το διάνυσμα αυτό διέρχεται από ένα επίπεδο εγκατάλειψης με πιθανότητα 0.5 και έπειτα τροφοδοτείται σε ένα πλήρως συνδεδεμένο επίπεδο με συνάρτηση ενεργοποίησης softmax. Η έξοδος του τελευταίου είναι η τελική έξοδος του δικτύου, δηλαδή ένα διάνυσμα 10 πιθανοτήτων.

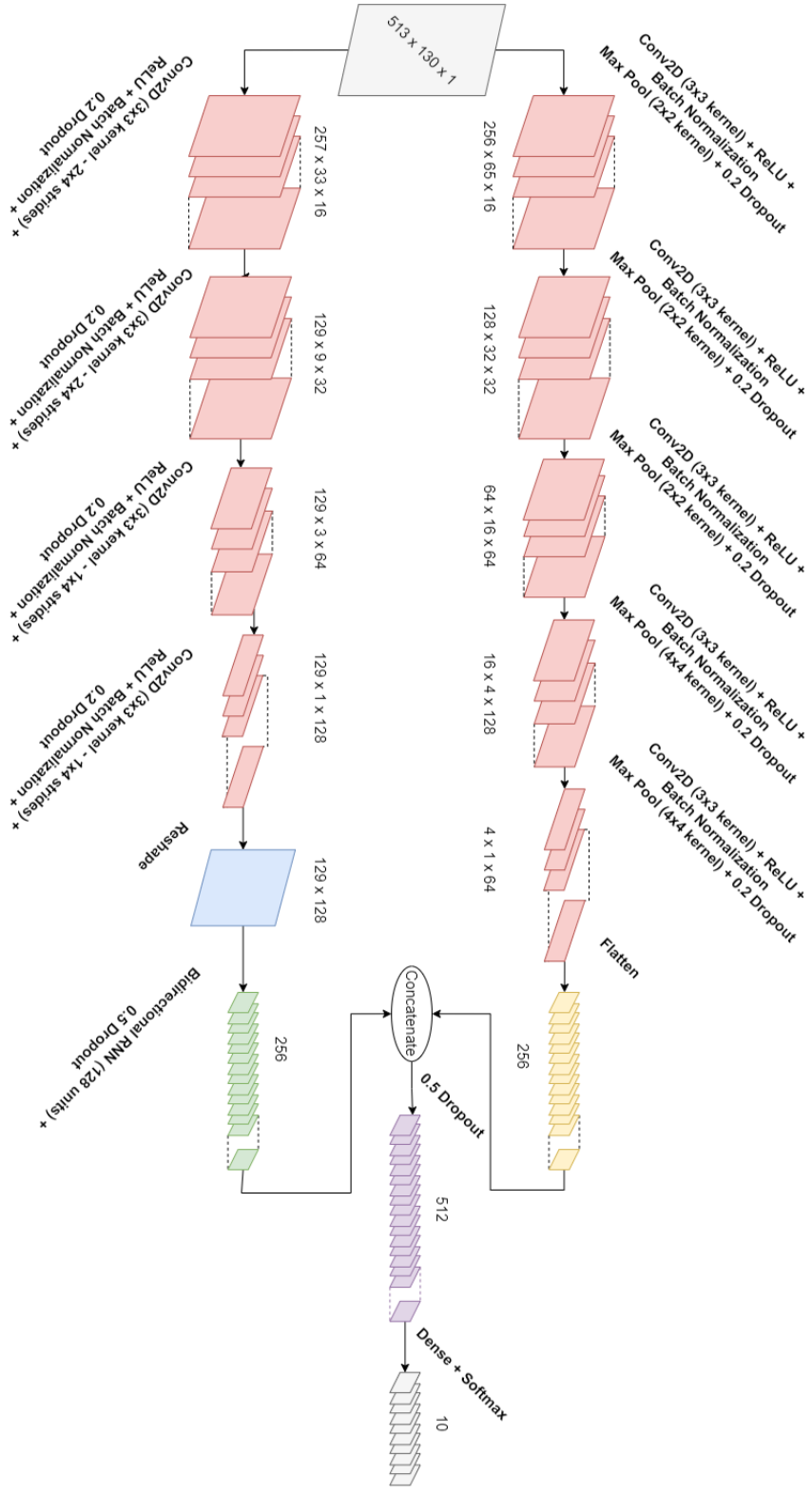
7.3 Αρχιτεκτονική Νευρωνικού B3

Το Σχήμα 7.3 περιλαμβάνει την αρχιτεκτονική του Νευρωνικού Δικτύου B3. Σε γενικές γραμμές μοιάζει σημαντικά με την αρχιτεκτονική B2. Η διαφορά μεταξύ τους έγκειται στο δεύτερο παράλληλο μέρος (δηλαδή αυτό που καταλήγει στο αναδρομικό επίπεδο διπλής κατεύθυνσης. Στα συνελικτικά μπλοκ αυτού του μέρους, η μείωση των διαστάσεων δεν γίνεται μέσω του βήματος του συνελικτικού επιπέδου (το οποίο τώρα είναι μοναδιαίο), αλλά μέσω ενός επιπέδου Max Pooling σε κάθε μπλοκ, το οποίο βρίσκεται αμέσως μετά το συνελικτικό επίπεδο.

Επί της ουσίας, η βασική διαφορά μεταξύ των δύο μεθόδων μείωσης διαστάσεων είναι ότι η πρώτη είναι μία εκπαιδεύσιμη μέθοδος, γεγονός το οποίο σημαίνει ότι μπορεί να ‘‘μάθει’’ συγκεκριμένες ιδιότητες. Τις ιδιότητες αυτές το επίπεδο Pooling δεν έχει τη δυνατότητα να λάβει υπόψιν, δεδομένου του ότι επιτελεί μία σταθερή πράξη, ανεξάρτητη από άλλες παραμέτρους. Από την άλλη, το επίπεδο Pooling είναι υπολογιστικά φθηνότερο σε σχέση με ένα συνελικτικό επίπεδο, τόσο από την άποψη του απαιτούμενου υπολογιστικού χρόνου, όσο και από πλευράς πλήθους παραμέτρων.

7.4 Αρχιτεκτονική Νευρωνικού B4

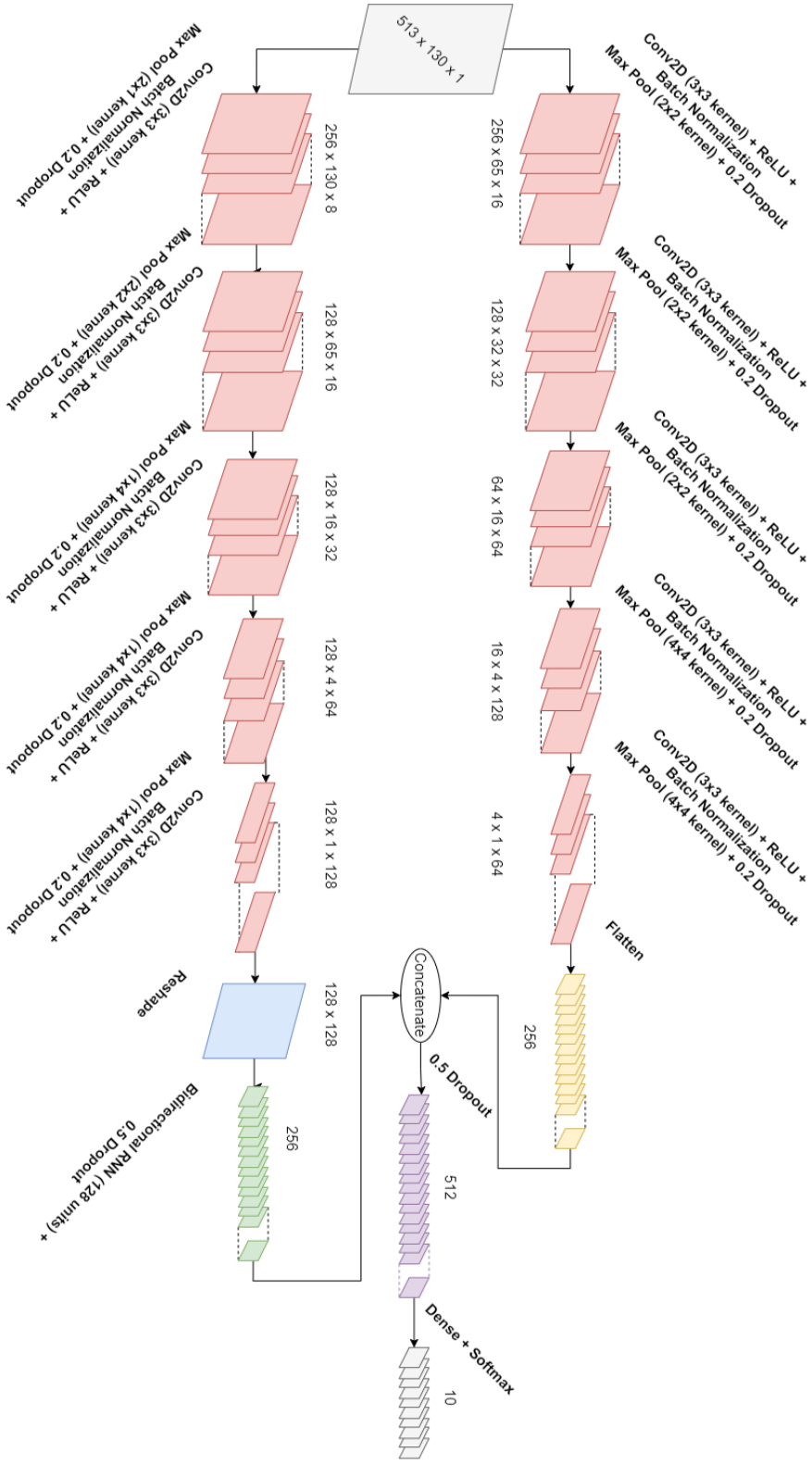
Στο Σχήμα 7.4 παρουσιάζεται η αρχιτεκτονική του Νευρωνικού Δικτύου B4. Το συνελικτικό μέρος παραμένει ίδιο με το αντίστοιχο μέρος των προηγούμενων ενοτήτων, ενώ στο δεύτερο μέρος παρατηρείται μια διαφοροποίηση. Η διαφοροποίηση αυτή έγκειται στην ύπαρξη 2 Max Pooling επιπέδων πριν το αναδρομικό επίπεδο, τα οποία έχουν ως στόχο τη μείωση των διαστάσεων της εισόδου του αναδρομικού επιπέδου. Αυτό γίνεται με τρόπο σταθερό, χωρίς τη χρήση συνελίξεων, σε αντίθεση με τις Αρχιτεκτονικές B2 και B3. Η συγχώνευση των δύο παράλληλων σκελών και ο υπολογισμός του τελικού διανύσματος εξόδου γίνονται με ακριβώς τον ίδιο τρόπο με όλες τις προαναφερθείσες παράλληλες αρχιτεκτονικές.



Σχήμα 7.2: *Αρχιτεκτονική Νευρωνικού B2*

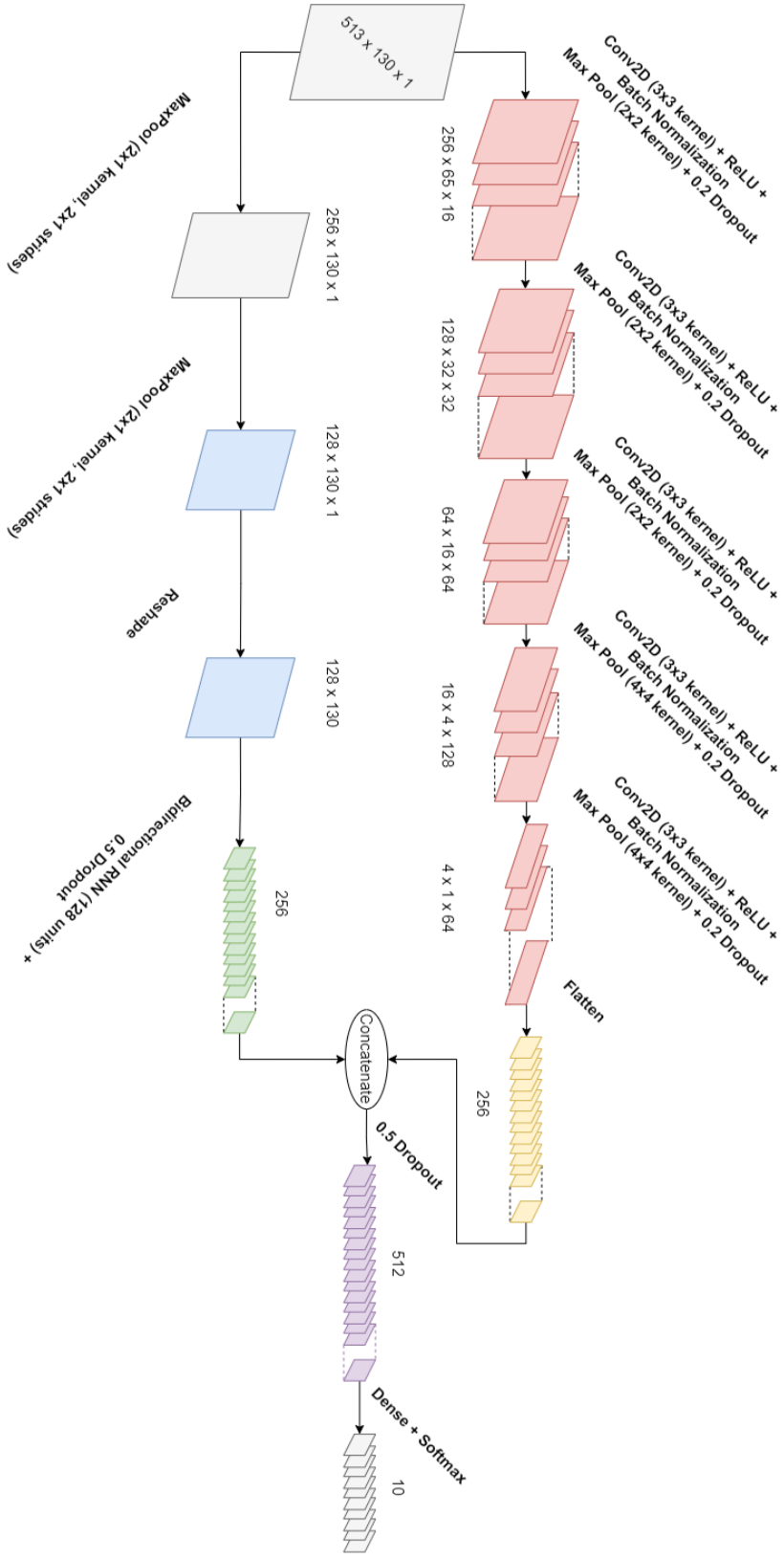
7.5 Αρχιτεκτονική Νευρωνικού B5

Τέλος, στο Σχήμα 7.5 φαίνεται η αρχιτεκτονική του Νευρωνικού Δικτύου B5. Πρόκειται και πάλι για μια παράλληλη αρχιτεκτονική. Το πρώτο παράλληλο μέρος της αρχιτεκτονικής



Σχήμα 7.3: *Αρχιτεκτονική Νευρωνικού B3*

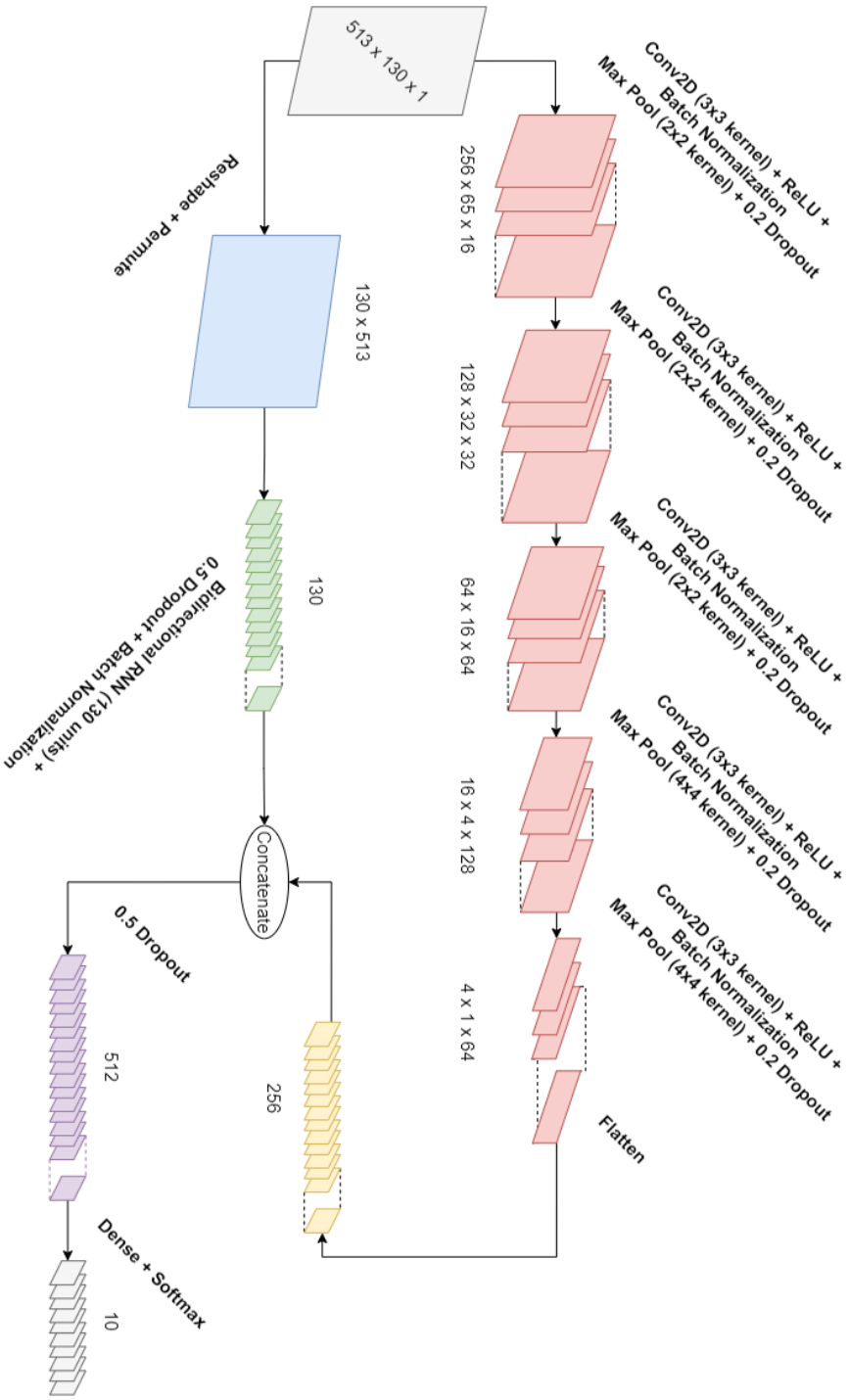
είναι αμιγώς συνελικτικό και είναι πανομοιότυπο με το αντίστοιχο συνελικτικό παράλληλο μέρος των αρχιτεκτονικών που παρουσιάστηκαν στις προηγούμενες ενότητες. Το δεύτερο παράλληλο μέρος είναι σημαντικά απλούστερο από το αντίστοιχο των προηγούμενων ενοτήτων, καθώς αποτελείται από δύο επίπεδα για την αλλαγή των διαστάσεων (Reshape και Permute)



Σχήμα 7.4: *Αρχιτεκτονική Νευρωνικού B4*

και ένα αναδρομικό επίπεδο διπλής κατεύθυνσης (Bidirectional RNN) ακολουθούμενο από ένα επίπεδο κανονικοποίησης δέσμης (Batch Normalization). Τα πρώτα επίπεδα δεν διαθέτουν εκπαιδεύσιμες παραμέτρους και δεν επιτελούν κάποια σημαντική αλλαγή. Το αναδρομικό επίπεδο

διπλής κατεύθυνσης (Bidirectional RNN) διαθέτει 130 μονάδες και εγκατάλειψη (Dropout) με πιθανότητα 0.5. Οι έξοδοι των δύο παράλληλων μερών συνενώνονται κατά τα γνωστά και τροφοδοτούνται σε ένα πλήρως συνδεδεμένο επίπεδο, από το οποίο προκύπτει ως έξοδος ένα διάνυσμα μεγέθους 10 θέσεων, όπως ακριβώς και στις αρχιτεκτονικές που περιγράφηκαν στις προηγούμενες ενότητες.



Σχήμα 7.5: *Αρχιτεκτονική Νευρωνικού B5*

Κεφάλαιο 8

Αποτελέσματα και Αξιολόγηση

Στο κεφάλαιο αυτό θα παρουσιαστούν αναλυτικά τα πειράματα που διεξήχθησαν με σκοπό τη δημιουργία ενός συστήματος ταξινόμησης της pop μουσικής στα 10 επιμέρους υποείδη της.

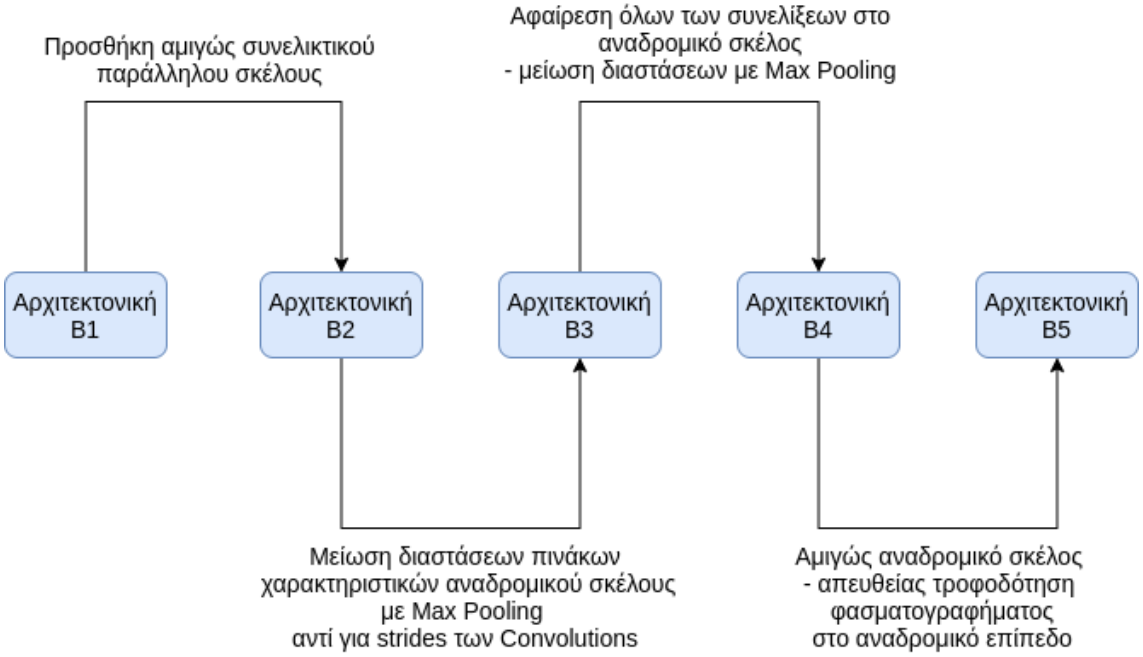
Το σύνολο εκπαίδευσης είναι κοινό σε όλα τα πειράματα. Στο σύνολο αυτό έχει εφαρμοστεί η προεπεξεργασία που περιγράφηκε αναλυτικά στο κεφάλαιο 6. Σε όλα τα πειράματα που θα περιγραφούν, το σύνολο εκπαίδευσης χωρίστηκε σε επιμέρους τμήματα - παρτίδες με μέγεθος (batch size) 8 δείγματα η καθεμία, ενώ στο τέλος της κάθε εποχής το σύνολο αυτό ''ανακατεύεται'' και έπειτα χωρίζεται εκ νέου σε παρτίδες με στόχο να αποφευχθεί το φαινόμενο της υπερεκπαίδευσης. Το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας ως συνάρτηση κόστους την κατηγορική διασταυρούμενη εντροπία (categorical crossentropy loss), για συνολικά 100 εποχές. Οσον αφορά τον αλγόριθμο βελτιστοποίησης, επιλέχθηκε ο Adam [25], με αρχική τιμή ρυθμού μάθησης 0.001, ενώ εφαρμόστηκε και η τεχνική της μείωσης του ρυθμού μάθησης όταν η τιμή της συνάρτησης κόστους για το σύνολο επικύρωσης (validation loss) σταματά να φθίνει (Reduce Learning Rate on Plateau) [39].

Όπως ακριβώς και στο Πρακτικό Μέρος Α΄, όλα τα πειράματα υλοποιήθηκαν με τη χρήση της βιβλιοθήκης Tensorflow [34] και της διεπαφής Keras [35]. Η εξαγωγή των χαρακτηριστικών, η εκπαίδευση των νευρωνικών και η αξιολόγηση των τελικών μοντέλων έγιναν χρησιμοποιώντας την υπηρεσία AmazonSageMaker [37] της πλατφόρμας υπολογιστικού νέφους (cloud computing platform) AWS - Amazon Web Services [36].

8.1 Πειράματα με Διαφορετικές Αρχιτεκτονικές

Η ενότητα αυτή περιλαμβάνει τα αποτελέσματα των πειραμάτων που διεξήχθησαν με σκοπό να εξεταστεί η επίδραση των διαφορετικών αρχιτεκτονικών που παρουσιάστηκαν στην προηγούμενη ενότητα στην απόδοση του μοντέλου. Αυτό σημαίνει πως στα πειράματα αυτά, το σύνολο δεδομένων, η προεπεξεργασία του, καθώς και οι υπερπαράμετροι της εκπαίδευσης διατηρήθηκαν αμετάβλητα, και η μόνη διαφοροποίηση μεταξύ των πειραμάτων ήταν η αρχιτεκτονική του νευρωνικού δικτύου. Για το λόγο αυτό, κρίνεται χρήσιμο να γίνει μια συγκεντρωτική αναφορά στις μεταβολές που λαμβάνουν χώρα από τον ένα τύπο αρχιτεκτονικής στον επόμενο. Οι μεταβολές αυτές φαίνονται συνοπτικά και στο διάγραμμα της εικόνας 8.1.

Ξεκινώντας από την Αρχιτεκτονική B1, το δίκτυο αποτελείται από ένα μόνο σκέλος, το οποίο, δεχόμενο ως είσοδο το φασματογράφημα, εξάγει χαρακτηριστικά μέσω μίας σειράς συ-



Σχήμα 8.1: *Διαφοροποιήσεις Αρχιτεκτονικών*

νελικτικών μπλοκ και στη συνέχεια τα τροφοδοτεί σε ένα αναδρομικό επίπεδο διπλής κατεύθυνσης. Προχωρώντας στην Αρχιτεκτονική B2, δεδομένων των ενθαρρυντικών αποτελεσμάτων των πειραμάτων του Πρακτικού Μέρους Α΄ (όπου είχαμε ένα αμιγώς συνελικτικό δίκτυο), δοκιμάστηκε να διατηρηθεί στο ακέραιο η αρχιτεκτονική του B1 και να συνδυαστεί με ένα παράλληλο, αμιγώς συνελικτικό σκέλος. Ο συνδυασμός των 2 παράλληλων σκελών έγινε με τη μέθοδο της συνένωσης (concatenation). Ακολούθως, η Αρχιτεκτονική B3 διαφέρει από την αρχιτεκτονική B2 όσον αφορά τον τρόπο μείωσης των διαστάσεων στους πίνακες χαρακτηριστικών που τροφοδοτούνται στο αναδρομικό επίπεδο. Πιο συγκεκριμένα, στην Αρχιτεκτονική B2 η μείωση γίνεται μέσω του βήματος των συνελίξεων, ενώ στη B3 μέσω επιπέδων Max Pooling. Στη συνέχεια, η Αρχιτεκτονική B4 διαφέρει από την B3 στο ένα παράλληλο σκέλος, όπου πλέον γίνεται αμιγώς αναδρομικό, καθώς δεν υπάρχουν πλέον συνελικτικά μπλοκ για την εξαγωγή χαρακτηριστικών, παραμόνο 2 Max Pooling επίπεδα που χρησιμεύουν στη μείωση των διαστάσεων των φασματογραφημάτων πριν αυτά τροφοδοτηθούν στο αναδρομικό επίπεδο. Τέλος, η Αρχιτεκτονική B5 είναι μια απλούστερη εκδοχή της Αρχιτεκτονικής B4, καθώς εδώ τα φασματογραφήματα τροφοδοτούνται κατευθείαν στο αναδρομικό επίπεδο, χωρίς μείωση διαστάσεων.

Στο σημείο αυτό, αξίζει να αναφερθεί πως για λόγους συγκρισιμότητας, το αναδρομικό επίπεδο σε όλες τις αρχιτεκτονικές που περιγράφηκαν ήταν ένα Bidirectional LSTM επίπεδο, με μέθοδο σύμπτυξης (merge mode) τη συνένωση (concatenation). Περισσότερες λεπτομέρειες σχετικά με αυτό θα αναφερθούν σε επόμενη ενότητα.

Ο πίνακας 8.1 περιλαμβάνει τις μετρικές αξιολόγησης και πιο συγκεκριμένα το ποσοστό ορθότητας στο σύνολο ελέγχου (Test Accuracy) και την τιμή της συνάρτησης κόστους για το σύνολο ελέγχου (Test Loss) για το κάθε πείραμα που πραγματοποιήθηκε.

Παρατηρώντας τον πίνακα, είναι εμφανές πως τόσο από πλευράς Accuracy όσο και από πλευράς Loss, η Αρχιτεκτονική B5 οδήγησε στα βέλτιστα αποτελέσματα. Αυτό μας οδηγεί

Πίνακας 8.1: *Αποτελέσματα πειραμάτων διαφορετικών Αρχιτεκτονικών*

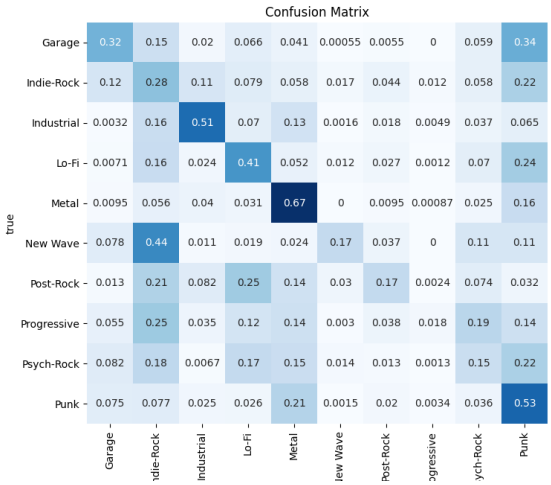
Αρχιτεκτονική	**Test Accuracy**	**Test Loss**
Αρχιτεκτονική B1	34%	1.8765
Αρχιτεκτονική B2	52%	1.4664
Αρχιτεκτονική B3	52%	1.6259
Αρχιτεκτονική B4	50%	1.5070
Αρχιτεκτονική B5	56%	1.4005

στο συμπέρασμα ότι η ξεκάθαρη διάκριση των λειτουργικοτήτων μεταξύ των δύο παράλληλων σκελών της αρχιτεκτονικής συνέβαλλε στην καλύτερη αντιμετώπιση του προβλήματος. Πιο συγκεκριμένα, φαίνεται πως το αμιγώς συνελικτικό σκέλος χρησίμευσε στην εξαγωγή ''χωρικής'' πληροφορίας (δεδομένου ότι το φασματογράφημα αντιμετωπίζεται ως μια εικόνα), ενώ το αμιγώς αναδρομικό επίπεδο λειτούργησε εξάγοντας χρονική πληροφορία, όπως είναι άλλωστε αναμενόμενο από τον ρόλο του. Σε αντιδιαστολή, παρατηρούμε πως από την Αρχιτεκτονική B1 προέκυψαν τα χειρότερα αποτελέσματα. Βάσει αυτού, συμπεραίνουμε πως ένα απλό νευρωνικό δίκτυο (όπως φαίνεται και από το πλήθος των παραμέτρων στον πίνακα 7.1) δεν επαρκεί για την αντιμετώπιση ενός τόσο απαιτητικού προβλήματος. Αντίθετα, η εισαγωγή μιας παράλληλης αρχιτεκτονικής, η οποία συνδυάζει με κάποιον τρόπο χρονικά και συχνοτικά χαρακτηριστικά της εισόδου, πράγματι λειτουργεί ενισχυτικά στις επιδόσεις του συστήματος. Από τη σύγκριση των αποτελεσμάτων των Αρχιτεκτονικών B2 και B3, τα οποία είναι σχετικά παρόμοια, προκύπτει ότι ο τρόπος μείωσης των διαστάσεων των πινάκων χαρακτηριστικών του αναδρομικού σκέλους δεν παίζει ιδιαίτερο ρόλο στο αποτέλεσμα για το εν λόγω πρόβλημα. Τέλος, αντιπαραβάλλοντας τα αποτελέσματα των Αρχιτεκτονικών B4 και B5, μπορούμε να αντιληφθούμε πως η αυθαίρετη μείωση των διαστάσεων του φασματογραφήματος μέσω επιπέδων Max Pooling επέφερε την απώλεια χρήσιμης πληροφορίας. Αντίθετα, όταν το φασματογράφημα τροφοδοτήθηκε ακέραιο στο αναδρομικό επίπεδο, η πληροφορία αυτή έγινε αντιληπτή από το δίκτυο, με αποτέλεσμα σημαντικά καλύτερες επιδόσεις.

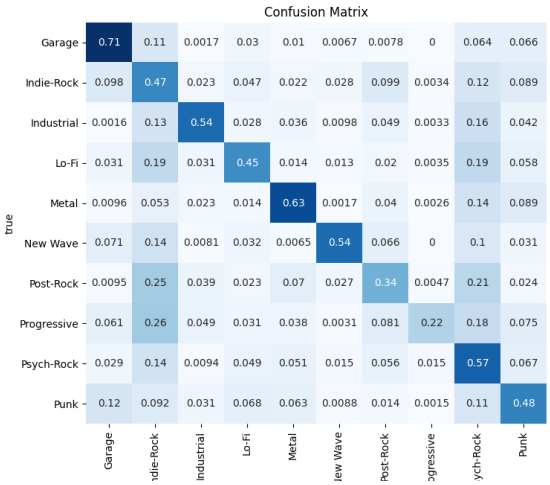
Για καλύτερη εποπτεία των αποτελεσμάτων των παραπάνω πειραμάτων, παρατίθενται και οι αντίστοιχοι πίνακες σύγχυσης στο σχήμα 8.2. Οι πίνακες αυτοί επιβεβαιώνουν και τα πιο συνοπτικά αποτελέσματα του πίνακα 8.1, προσφέροντας βαθύτερη κατανόηση στις επιδόσεις του κάθε μοντέλου σε κάθε κλάση του συνόλου δεδομένων.

8.2 Πειράματα με Διαφορετικά Αναδρομικά Επίπεδα

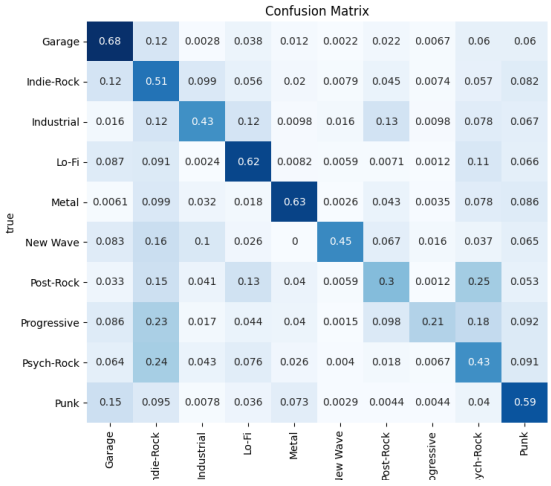
Η ενότητα αυτή περιλαμβάνει τα αποτελέσματα των πειραμάτων που διεξήχθησαν προκειμένου να εξεταστεί η επίδραση διαφορετικών αναδρομικών επιπέδων στις επιδόσεις του μοντέλου. Για το σκοπό των πειραμάτων αυτής της ενότητας, επιλέχθηκε το βέλτιστο μοντέλο που προέκυψε από τα πειράματα της προηγούμενης ενότητας, δηλαδή η Αρχιτεκτονική B5. Πάνω στη αρχιτεκτονική αυτή εφαρμόστηκαν ορισμένες τροποποιήσεις στο αναδρομικό επίπεδο. Συγκεκριμένα, πραγματοποιήθηκαν 3 πειράματα. Στο πρώτο εξ αυτών, το επίπεδο GRU [44] αντικαταστάθηκε από ένα επίπεδο Simple RNN [45], στο δεύτερο από ένα επίπεδο LSTM [46] και στο τρίτο από δύο επίπεδα LSTM. Ο πίνακας 8.2 περιλαμβάνει τις μετρι-



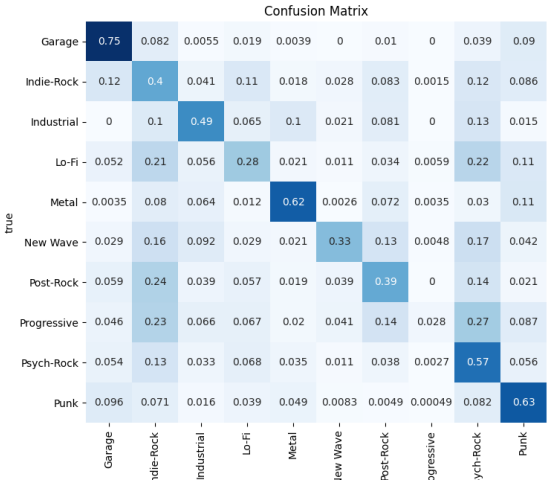
Αρχιτεκτονική B1



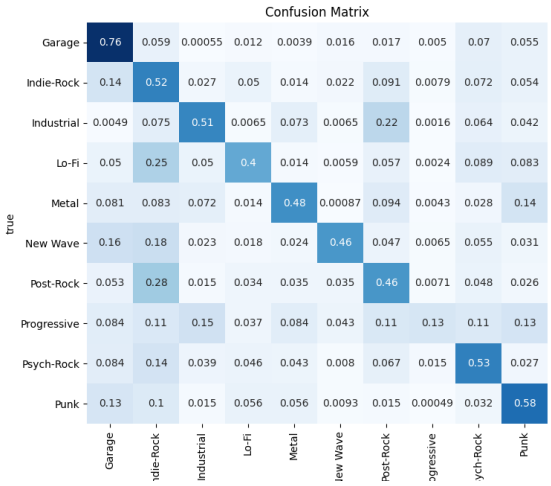
Αρχιτεκτονική B2



Αρχιτεκτονική B3



Αρχιτεκτονική B4



Αρχιτεκτονική B5

Σχήμα 8.2: *Διαφορετικές Αρχιτεκτονικές - Πίνακες Σύγχυσης*

κές αξιολόγησης και πιο συγκεκριμένα το ποσοστό ορθότητας στο σύνολο ελέγχου (Test Accuracy) και την τιμή της συνάρτησης κόστους για το σύνολο ελέγχου (Test Loss) για το κάθε πείραμα που πραγματοποιήθηκε. Επιπλέον, στο σχήμα 8.3 παρουσιάζονται οι αντίστοιχοι πίνακες σύγχυσης που προέκυψαν από την αξιολόγηση του κάθε πειράματος.

Πίνακας 8.2: *Αποτελέσματα πειραμάτων διαφορετικών Αναδρομικών Επιπέδων*

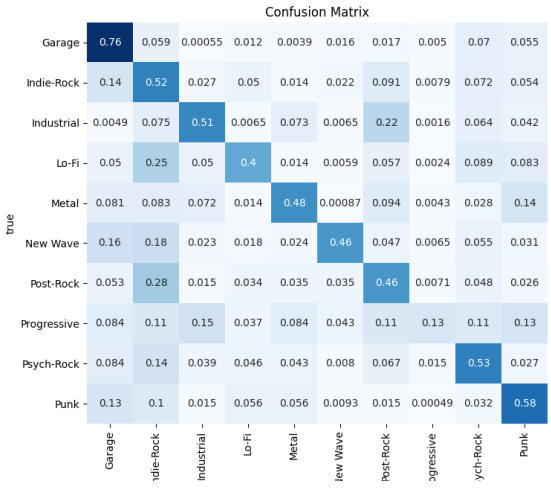
Αναδρομικό Επίπεδο	**Test Accuracy**	**Test Loss**
GRU	56%	1.4005
Simple RNN	53%	1.5906
LSTM	62%	1.3763
2 Stacked LSTMs	51%	1.6676

Παρατηρώντας τον πίνακα, το πρώτο σημαντικό συμπέρασμα που μπορούμε να συνάγουμε είναι πως καθώς η πολυπλοκότητα του αναδρομικού επιπέδου αυξάνεται, ξεκινώντας από το απλούστερο Simple RNN επίπεδο, προχωρώντας στο περισσότερο σύνθετο επίπεδο GRU και καταλήγοντας στο πιο περίπλοκο LSTM επίπεδο, τόσο βελτιώνονται και οι επιδόσεις του συστήματος. Αυτό μπορεί να επεξηγηθεί καθώς ένα πολύπλοκο πρόβλημα όπως αυτό που καλούμαστε να επιλύσουμε, απαιτεί και μια ανάλογη πολυπλοκότητα στην αρχιτεκτονική του, ώστε να μπορέσουν να μοντελοποιηθούν τα χαρακτηριστικά και οι προβλέψεις με τον κατάλληλο τρόπο [47]. Παρόλα αυτά, το παράδοξο είναι πως όταν επιχειρήθηκε επιπλέον πολυπλοκότητα στην αρχιτεκτονική, μέσω της εισαγωγής ενός επιπλεόν επιπέδου LSTM τότε οι επιδόσεις του μοντέλου μειώθηκαν αισθητά. Αυτό βέβαια μπορεί να ερμηνευθεί, καθώς οι επιδόσεις ενός συστήματος για την επίλυση ενός πολύπλοκου προβλήματος δεν αυξάνονται γραμμικά με την αύξηση της πολυπλοκότητας της αρχιτεκτονικής, ενώ επίσης και ο τρόπος με τον οποίο αυξάνεται η πολυπλοκότητα έχει ιδιαίτερη σημασία.

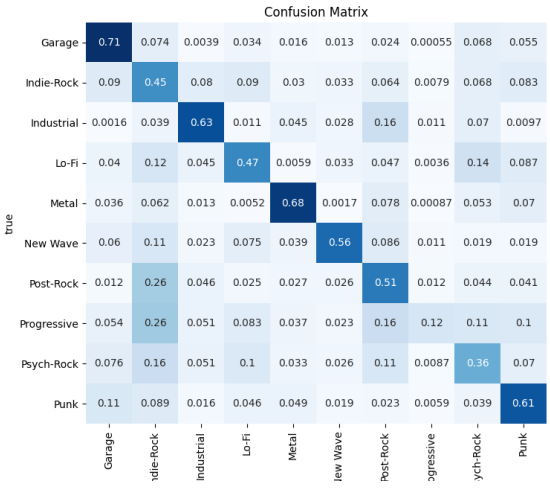
8.3 Πειράματα με Διαφορετικές Μεθόδους Σύμπτυξης

Η ενότητα αυτή περιλαμβάνει τα αποτελέσματα των πειραμάτων που διεξήχθησαν προκειμένου να εξεταστεί η επίδραση διαφορετικών μεθόδων σύμπτυξης του αναδρομικού επιπέδου διπλής κατεύθυνσης στις επιδόσεις του μοντέλου. Για το σκοπό των πειραμάτων αυτής της ενότητας, επιλέχθηκε το βέλτιστο μοντέλο που προέκυψε από τα πειράματα των προηγούμενων ενοτήτων, δηλαδή η Αρχιτεκτονική B5 με Bidirectional LSTM. Πάνω στη αρχιτεκτονική αυτή εφαρμόστηκαν ορισμένες τροποποιήσεις στη μέθοδο σύμπτυξης.

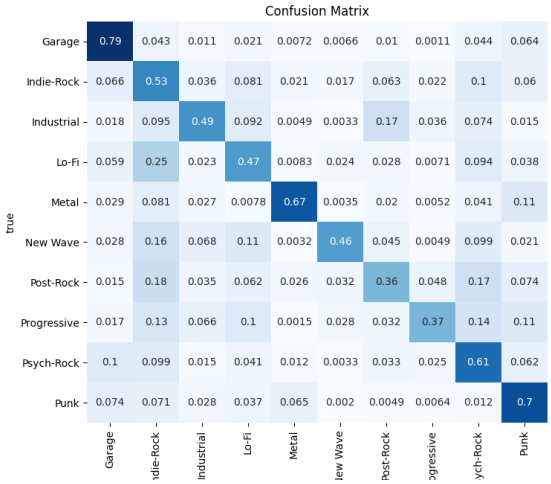
Στο σημείο αυτό, είναι σημαντικό να αναφερθούμε στην λειτουργία των αναδρομικών επιπέδων διπλής κατεύθυνσης και στον τρόπο με τον οποίο λειτουργούν οι διαφορετικές μέθοδοι σύμπτυξης merge modes. Τα αναδρομικά δίκτυα διπλής κατεύθυνσης παρουσιάστηκαν για πρώτη φορά το 1997α από τους Mike Schuster και Kuldip K. Paliwal [48]. Η χρησιμότητά τους έγκειται στο ότι συνδέουν δύο κρυφά επίπεδα που λειτουργούν σε αντίθετες κατευθύνσεις σε μία κοινή έξοδο, με αποτέλεσμα να μπορούν να επεξεργάζονται ταυτόχρονα πληροφορίες τόσο από παρελθοντικές όσο και από μελλοντικές καταστάσεις. Αυτό είναι ιδιαίτερα χρήσιμο στις περιπτώσεις όπου η είσοδος μία δεδομένη χρονική στιγμή εξαρτάται όχι μόνο από προηγούμενες εισόδους αλλά και από μελλοντικές. Στην πράξη τα επίπεδα αυτά υλοποιούνται



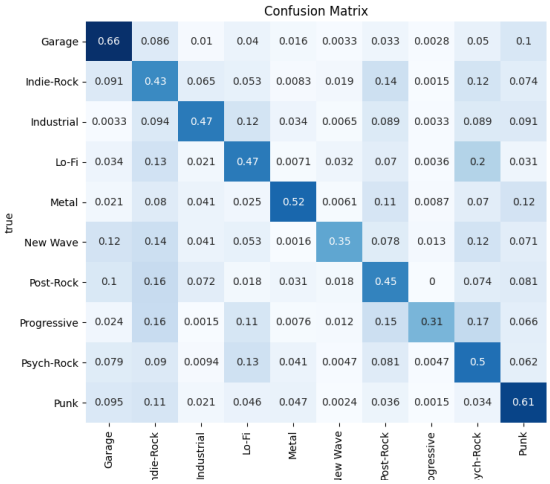
Αναδρομικό Επίπεδο: GRU



Αναδρομικό Επίπεδο: Simple RNN



Αναδρομικό Επίπεδο: LSTM



Αναδρομικό Επίπεδο: 2 Stacked LSTMs

Σχήμα 8.3: *Διαφορετικά Αναδρομικά Επίπεδα - Πίνακες Σύγχυσης*

μέσω ενός wrapper της διεπαφής Keras, ο οποίος μετατρέπει ένα απλό αναδρομικό επίπεδο σε αναδρομικό επίπεδο διπλής κατεύθυνσης [49].

Επί της ουσίας, σε ένα LSTM διπλής κατεύθυνσης, αντί να εκπαιδεύεται ένα επίπεδο, εκπαιδεύονται δύο. Το πρώτο επίπεδο μαθαίνει την ακολουθία της εισόδου όπως αυτή είναι, ενώ το δεύτερο μαθαίνει την ίδια ακολουθία αντεστραμμένη. Καθώς έχουμε λοιπόν δύο εκπαιδευμένα επίπεδα, είναι αναγκαίος ένας μηχανισμός ο οποίος συνδυάζει τις προβλέψεις τους. Ο μηχανισμός αυτός είναι η Μέθοδος Σύμπτυξης (Merge Mode) και μπορεί να υλοποιηθεί με τους ακόλουθους τρόπους [50]:

- Συνένωση (Concatenation)
- Άθροισμα (Sum)
- Πολλαπλασιασμός (Multiplication)
- Μέσος Όρος (Averaging)

Ο πίνακας 8.3 περιλαμβάνει τις μετρικές αξιολόγησης και πιο συγκεκριμένα το ποσοστό ορθότητας στο σύνολο ελέγχου (Test Accuracy) και την τιμή της συνάρτησης κόστους για το σύνολο ελέγχου (Test Loss) για το κάθε πείραμα με διαφορετική μέθοδο σύμπτυξης που πραγματοποιήθηκε. Επιπλέον, στο σχήμα 8.4 παρουσιάζονται οι αντίστοιχοι πίνακες σύγχυσης που προέκυψαν από την αξιολόγηση του κάθε πειράματος. Όπως μπορούμε να παρατηρήσουμε, από τη μέθοδο συνένωσης προκύπτουν τα βέλτιστα αποτελέσματα. Αξίζει να αναφερθεί πως η μέθοδος συνένωσης είναι η προκαθορισμένη μέθοδος και έτσι χρησιμοποιήθηκε σε όλα τα πειράματα των προηγούμενων ενοτήτων του κεφαλαίου.

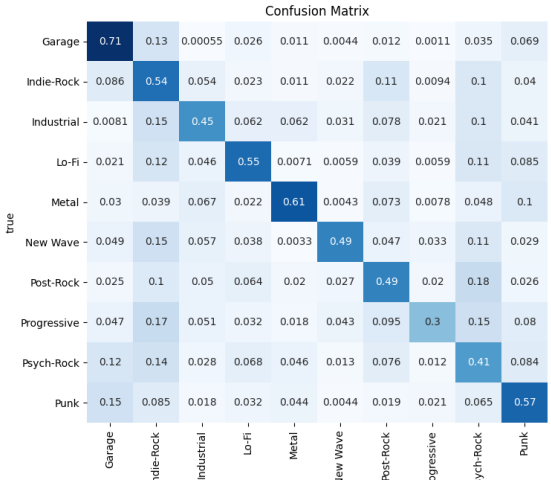
Πίνακας 8.3: *Αποτελέσματα πειραμάτων διαφορετικών Μεθόδων Σύμπτυξης*

Αναδρομικό Επίπεδο	**Test Accuracy**	**Test Loss**
LSTM - merge mode: concat	62%	1.3763
LSTM - merge mode: sum	60%	1.3934
LSTM - merge mode: mul	55%	1.7357
LSTM - merge mode: ave	52%	1.6833

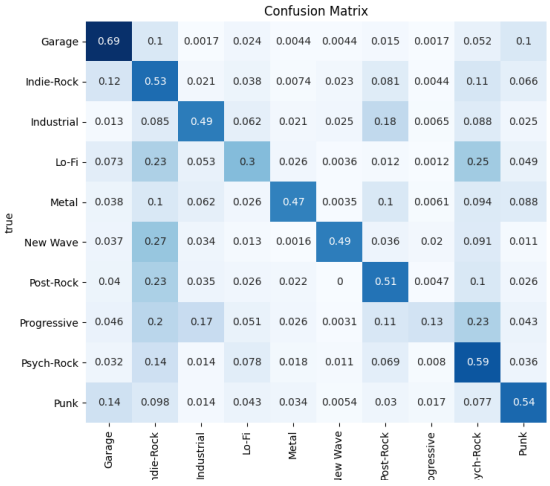
8.4 Πειράματα με Διαφορετικές Μεθόδους Συνδυασμού Παράλληλων Σκελών

Κλείνοντας τα πειράματα του Πρακτικού Μέρους Β΄, στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα των πειραμάτων που διεξήχθησαν προκειμένου να εξεταστεί η επίδραση διαφορετικών μεθόδων συνδυασμού των 2 παράλληλων σκελών της αρχιτεκτονικής του νευρωνικού. Για το σκοπό των πειραμάτων αυτής της ενότητας, επιλέχθηκε το βέλτιστο μοντέλο που προέκυψε από τα πειράματα των προηγούμενων ενοτήτων, δηλαδή η Αρχιτεκτονική B5 με Bidirectional LSTM και μέθοδο σύμπτυξης concat. Πάνω σε αυτή την αρχιτεκτονική δοκιμάστηκαν δύο διαφορετικοί τρόποι συνδυασμού των εξόδων των δύο παράλληλων σκελών, δηλαδή του συνελικτικού και του αναδρομικού σκέλους. Πιο συγκεκριμένα, δοκιμάστηκε η μέθοδος Concatenation όπου οι έξοδοι των δύο σκελών απλά συνενώνονται, αφού παρατεθούν η μία δίπλα στην άλλη, και η μέθοδος Add όπου πραγματοποιείται άθροισμα μεταξύ των αντίστοιχων θέσεων των επιμέρους εξόδων. Αυτό βέβαια φυσικά προϋποθέτει κοινό μήκος εξόδων από το κάθε παράλληλο σκέλος.

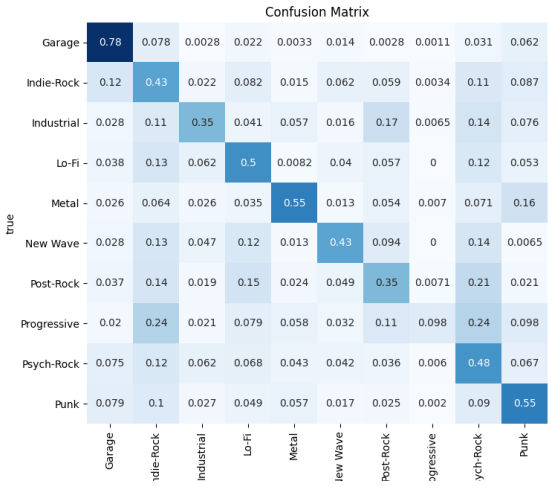
Ο πίνακας 8.4 περιλαμβάνει τις μετρικές αξιολόγησης και πιο συγκεκριμένα το ποσοστό ορθότητας στο σύνολο ελέγχου (Test Accuracy) και την τιμή της συνάρτησης κόστους για το σύνολο ελέγχου (Test Loss) για το κάθε πείραμα που πραγματοποιήθηκε. Επιπλέον, στο σχήμα 8.5 παρουσιάζονται οι αντίστοιχοι πίνακες σύγχυσης που προέκυψαν από την αξιολόγηση του κάθε πειράματος. Όπως μπορούμε να παρατηρήσουμε, η μέθοδος Concatenation δίνει τα βέλτιστα αποτελέσματα.



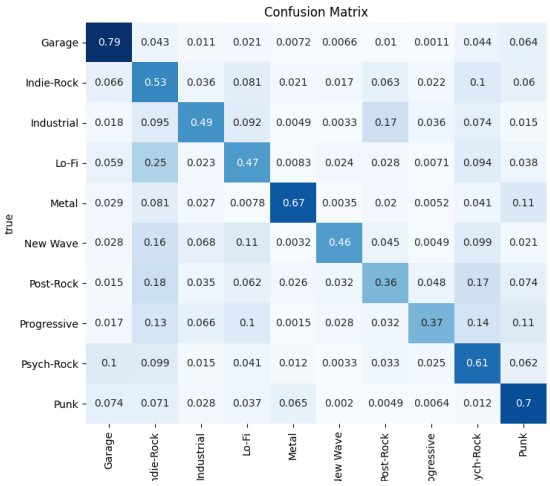
LSTM - merge mode: sum



LSTM - merge mode: mul



LSTM - merge mode: ave

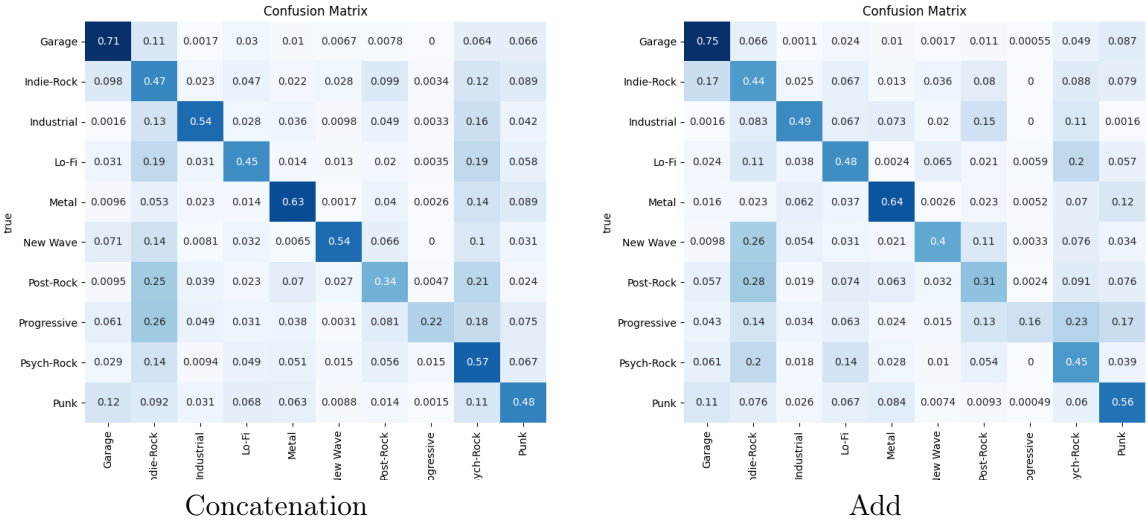


LSTM - merge mode: concat

Σχήμα 8.4: *Διαφορετικές Μέθοδοι Σύμπτυξης - Πίνακες Σύγχυσης*

Πίνακας 8.4: *Αποτελέσματα πειραμάτων διαφορετικών Μεθόδων Συνδυασμού Παράλληλων Σκελών*

Μέθοδος Συνδυασμού	**Test Accuracy**	**Test Loss**
Concatenation	62%	1.3763
Add	53%	1.6934



Σχήμα 8.5: *Διαφορετικές Μέθοδοι Συνδυασμού Παράλληλων Σκελών - Πίνακες Σύγχυσης*

Μέρος IV

Επίλογος

Κεφάλαιο 9

Σύνοψη και Μελλοντικές Επεκτάσεις

Στο κεφάλαιο αυτό συνοψίζεται η δουλειά που έλαβε χώρα στα πλαίσια της παρούσας διπλωματικής εργασίας και παρουσιάζονται επιγραμματικά τα συμπεράσματα που μπορούν να εξαχθούν μετά από το σύνολο των πειραματισμών που επιτελέστηκαν. Επιπλέον, περιγράφονται ορισμένες πιθανές κατευθύνσεις που θα μπορούσαν να ακολουθηθούν με σκοπό την περαιτέρω έρευνα και εμβάθυνση σχετικά με το αντικείμενο της εργασίας.

9.1 Σύνοψη

Η εργασία αυτή είχε ως στόχο να δημιουργήσει ένα σύστημα ταξινόμησης μουσικής, ξεκινώντας από το γενικότερο πρόβλημα της αναγνώρισης μουσικού είδους και προχωρώντας στο ειδικότερο πρόβλημα της αναγνώρισης μουσικού υποείδους.

Στο πρώτο σκέλος της εργασίας, αυτό της αναγνώρισης μουσικού είδους, χρησιμοποιήθηκε ένα ευρέως διαδεδομένο σύνολο δεδομένων, με πληθώρα βιβλιογραφικών αναφορών, το GTZAN. Δοκιμάστηκαν δύο διαφορετικές αρχιτεκτονικές νευρωνικών δικτύων, και οι δύο αμιγώς συνελικτικές. Η πρώτη εκπαιδεύτηκε χρησιμοποιώντας ως είσοδο τα mel spectrograms των διαιρεμένων ηχητικών κομματιών και έγινε προσπάθεια βελτιστοποίησης του συστήματος όσον αφορά τον αλγόριθμο βελτιστοποίησης και το ρυθμό μάθησης. Ωστόσο, τα αποτελέσματα δεν ήταν ικανοποιητικά, οπότε δεν επιχειρήθηκαν άλλες βελτιστοποιήσεις και προχωρήσαμε στη δεύτερη αρχιτεκτονική. Τα πειράματα που πραγματοποιήθηκαν πάνω στην αρχιτεκτονική αυτή είχαν ως σκοπό τον έλεγχο των επιδόσεών της για διαφορετικούς μετασχηματισμούς (εξαγωγή χαρακτηριστικών) του συνόλου εκπαίδευσης. Επομένως, όλες οι υπόλοιπες παράμετροι και υπερπαράμετροι διατηρήθηκαν σταθερές και η μόνη διαφοροποίηση ήταν ο τύπος των χαρακτηριστικών που τροφοδοτήθηκαν στο δίκτυο κατά την εκπαίδευση. Στη συνέχεια, δοκιμάστηκαν ορισμένοι συνδυασμοί με σκοπό τη δημιουργία ενός υβριδικού μοντέλου που δέχεται πολλαπλά χαρακτηριστικά στην είσοδό του.

Στο δεύτερο σκέλος της εργασίας επιχειρήθηκε η υλοποίηση ενός συστήματος αναγνώρισης μουσικού υποείδους. Το σύνολο δεδομένων που χρησιμοποιήθηκε είναι ένα υποσύνολο του επίσης ευρέως διαδεδομένου συνόλου FMA. Για την προεπεξεργασία των δεδομένων χρησιμοποιήθηκαν τα αποτελέσματα των πειραμάτων του πρώτου σκέλους, από όπου προέκυψε ότι ο μετασχηματισμός STFT οδηγεί στην καλύτερη ενσωμάτωση πληροφορίας και συνεπώς στις καλύτερες δυνατές επιδόσεις. Τα πειράματα που διεχήθησαν είχαν ως στοχο την εύρεση μιας βέλτιστης αρχιτεκτονικής νευρωνικού. Για το σκοπό αυτό, δοκιμάστηκαν σε πρώτη φάση

5 διαφορετικές αρχιτεκτονικές οι οποίες στο σύνολό τους περιείχαν αναδρομικά επίπεδα, ενώ οι 4 εξ αυτών αποτελούνταν από δύο παράλληλα μέρη που συνδυάζονταν στην έξοδο. Τα πειράματα αυτά οδήγησαν στην επιλογή της καλύτερης αρχιτεκτονικής, η οποία στη συνέχεια βελτιστοποιήθηκε περαιτέρω μέσω επιπλέον πειραματισμών. Αυτοί αφορούσαν το είδος του αναδρομικού επιπέδου, τη μέθοδο σύμπτυξης αυτού και την τεχνική συνδυασμού των παράλληλων μερών.

Αποτέλεσμα των παραπάνω ήταν η επιλογή δύο συστημάτων, ένα για το κάθε σκέλος. Στην περίπτωση του προβλήματος της αναγνώρισης μουσικού είδους καταλήξαμε σε ένα υβριδικό συνελικτικό δίκτυο, το οποίο προκύπτει από το συνδυασμό ενός μοντέλου εκπαιδευμένο στο μετασχηματισμό STFT και ενός μοντέλου εκπαιδευμένου στο μετασχηματισμό MFCC συνενωμένο με τις παραγώγους του. Ο συνδυασμός αυτός γίνεται με τη μέθοδο του Soft Voting και οδηγεί σε ποσοστό ορθότητας **92%**. Στην περίπτωση του προβλήματος της αναγνώρισης μουσικού υποείδους καταλήξαμε σε μια παράλληλη αρχιτεκτονική νευρωνικού δικτύου, αποτελούμενη από ένα αμιγώς συνελικτικό και ένα αμιγώς αναδρομικό μέρος που συνδυάζονται μέσω συνένωσης (Concatenation). Το αναδρομικό μέρος συνίσταται από ένα LSTM επίπεδο διπλής κατεύθυνσης με μέθοδο σύμπτυξης concat. Το ποσοστό ορθότητας του εν λόγω δικτύου ανέρχεται στο **62%**.

9.2 Μελλοντικές Επεκτάσεις

Έχοντας ως αφετηρία την παρούσα διπλωματική εργασία και τα ευρήματα που προέκυψαν από αυτή, ξεδιπλώνεται ένα ευρύ φάσμα διαφορετικών προεκτάσεων και κατευθύνσεων προς τις οποίες θα μπορούσε να πραγματοποιηθεί επιπλέον έρευνα.

Μια ενδιαφέρουσα και ιδιαίτερα χρήσιμη πρόταση είναι η ανάπτυξη εφαρμογής η οποία θα προσφέρει τη λειτουργικότητα της αναγνώρισης μουσικού είδους και υποείδους μέσω μίας εύχρηστης διεπαφής χρήστη (User Interface). Μέσω της εφαρμογής αυτής, ο χρήστης θα μπορεί να επιλέγει ένα συγκεκριμένο μουσικό κομμάτι και να ενημερώνεται για το μουσικό είδος στο οποίο ανήκει. Επιπλέον, ενδιαφέρουσα λειτουργικότητα μιας τέτοιας εφαρμογής είναι και η αναγνώριση μουσικού είδους ενός συνόλου μουσικών κομματιών, με σκοπό την οργάνωσή τους σε επιμέρους λίστες με κοινό μουσικό είδος και συνεπώς παρόμοιο ύφος. Αυτή μάλιστα θα μπορούσε να είναι και η βάση της δημιουργίας ενός συστήματος συστάσεων (Recommendation System).

Επιπλέον, περαιτέρω έρευνα θα μπορούσε να πραγματοποιηθεί με σκοπό την βελτιστοποίηση των μοντέλων που προέκυψαν. Στην κατεύθυνση αυτή, είναι ιδιαίτερα χρήσιμη η εκτενέστερη μελέτη των διαφορετικών υπερπαραμέτρων που χρησιμοποιούνται κατά την εκπαίδευση και ο πειραματισμός με πλήθος διαφορετικών συνδυασμών με σκοπό την εύρεση αυτού που μεγιστοποιεί την απόδοση του συστήματος. Ακόμα, η επανεκπαίδευση των μοντέλων με ένα μεγαλύτερο και καλύτερης ποιότητας σύνολο δεδομένων μπορεί να βοηθήσει σημαντικά τις επιδόσεις του αλγορίθμου, καθώς αυξάνεται η δυνατότητα γενίκευσης. Δεδομένης όμως της δυσκολίας εύρεσης κατάλληλων συνόλων δεδομένων για τόσο εξειδικευμένα προβλήματα όπως η ταξινόμηση σε μουσικό υποείδος, ενδιαφέρον θα παρουσίαζε η τεχνική της Μεταφοράς Μάθησης (Transfer Learning). Στην περίπτωση αυτή, ένα δίκτυο θα μπορούσε να εκπαιδευτεί σε ένα μεγάλο σύνολο δεδομένων με σκοπό τη δημιουργία συστήματος αναγνώρι-

σης μουσικού είδους, με σκοπό να μάθει καλά τα βασικά χαρακτηριστικά του γενικότερου προβλήματος. Στη συνέχεια το ήδη εκπαιδευμένο δίκτυο θα γίνει fine tune πάνω σε ένα μικρότερο σύνολο δεδομένων επισημασμένα με το μουσικό υποείδος. Έτσι, χρησιμοποιώντας την ήδη υπάρχουσα γενική γνώση θα μπορέσει να μοντελοποιήσει τα χαρακτηριστικά ενός ειδικότερου προβλήματος, ακόμα και χωρίς την ύπαρξη ενός ικανοποιητικά μεγάλου συνόλου δεδομένων.

Όσον αφορά τα δεδομένα, θα μπορούσαν να εμπλουτιστούν ώστε πέρα από το ηχητικό σήμα να υπάρχουν επιπλέον πληροφορίες διαθέσιμες, όπως για παράδειγμα τα μεταδεδομένα του κάθε κομματιού (π.χ. τίτλος, καλλιτέχνες κλπ) αλλά και η κειμενική πληροφορία των στίχων του κάθε τραγουδιού. Αυτό φυσικά μας οδηγεί σε ένα νέο κύκλο έρευνας, καθώς υποδηλώνει την ανάπτυξη ενός τελείως διαφορετικού μοντέλου ή συνδυασμού μοντέλων που να έχουν τη δυνατότητα να εκπαιδευτούν πάνω στα αντίστοιχα σύνολα δεδομένων.

Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει και η χρήση των ήδη υπαρχόντων μοντέλων (ενδεχομένως μετά από επανεκπαίδευση) για παρεμφερή προβλήματα, όπως η αναγνώριση υποείδους ενός είδους διαφορετικού από τη ροκ, η αναγνώριση της διάθεσης του κομματιού, η αναγνώριση της γλώσσας στην οποία τραγουδά ο καλλιτέχνης ή ακόμα και η αναγνώριση του μουσικού συνθέτη ενός κλασικού κομματιού.

Βιβλιογραφία - Αναφορές

Βιβλιογραφία

[1] G. Tzanetakis και P. Cook. *Musical genre classification of audio signals. IEEE transactions on speech and audio processing: a publication of the IEEE Signal Processing Society*, 10(5):293–302, 2002.

[2] Sander Dieleman και Benjamin Schrauwen. *End-to-end learning for music audio. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 6964–6968, 2014.

[3] Weibin Zhang, Wenkang Lei, Xiangmin Xu και Xiaofeng Xing. *Improved music genre classification with convolutional neural networks. Interspeech 2016*. ISCA, 2016.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren και Jian Sun. *Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition*, σελίδες 770–778, 2016.

[5] Jordi Pons, Thomas Lidy και Xavier Serra. *Experimenting with musically motivated convolutional neural networks. 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2016.

[6] Keunwoo Choi, György Fazekas, Mark Sandler και Kyunghyun Cho. *Convolutional recurrent neural networks for music classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, σελίδες 2392–2396, 2017.

[7] Rui Yang, Lin Feng, Huibing Wang, Jianing Yao και Sen Luo. *Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices. IEEE access: practical innovations, open solutions*, 8:19629–19637, 2020.

[8] John McCarthy. *WHAT IS ARTIFICIAL INTELLIGENCE?* 2004.

[9] Peter Norvig Stuart Russel. *Τεχνητή Νοημοσύνη - Μια σύγχρονη προσέγγιση*. Κλειδάριθμος, Αθήνα, 2005.

[10] E. Rich και K. Knight. *Artificial Intelligence*. Artificial Intelligence Series. McGraw-Hill, 1991.

[11] A. M. TURING. *I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind*, LIX(236):433–460, 1950.

[12] M. Mohri, A. Rostamizadeh και A. Talwalkar. *Foundations of Machine Learning, second edition*. Adaptive Computation and Machine Learning series. MIT Press, 2018.

[13] Γεωργούλη, Α. *Μηχανική Μάθηση*. 2015.

[14] *What is Generalization in Machine Learning?* https://deepai.space/what-is-generalization-in-machine-learning/. Ημερομηνία πρόσβασης: 15-5-2021.

[15] *Difference Between Classification and Regression in Machine Learning*. https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/. Ημερομηνία πρόσβασης: 15-5-2021.

[16] G.E. Hinton, T.J. Sejnowski και H.H.M.I.C.N.L.T.J. Sejnowski. *Unsupervised Learning: Foundations of Neural Computation*. A Bradford Book. Bradford University Press, 1999.

[17] Βερύκιος, Β., Καγκλής, Β., Σταυρόπουλος, Η. *Συσταδοποίηση*. 2015.

[18] Simon Haykin. *Νευρωνικά Δίκτυα και Μηχανική Μάθηση*. Εκδόσεις Παπασωτηρίου, Αθήνα, 2010.

[19] Ian Goodfellow, Yoshua Bengio και Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[20] Rafael C. Gonzalez και Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., USA, 2006.

[21] *Recurrent Neural Networks - IBM*. https://www.ibm.com/cloud/learn/recurrent-neural-networks. Ημερομηνία πρόσβασης: 29-5-2021.

[22] Nikhil Ketkar. *Deep learning with python: A hands-on introduction*. APRESS, 1η έκδοση, 2017.

[23] Agnes Lydia και Sagayaraj Francis. *Adagrad—An optimizer for stochastic gradient descent. Int. J. Inf. Comput. Sci.*, 6(5), 2019.

[24] Matthew D. Zeiler. *ADADELTA: An Adaptive Learning Rate Method. arXiv [cs.LG]*, 2012.

[25] Diederik P. Kingma και Jimmy Ba. *Adam: A method for stochastic optimization. arXiv [cs.LG]*, 2014.

[26] Bob L. Sturm. *The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. arXiv [cs.SD]*, 2013.

[27] *Librosa library*. https://librosa.org/doc/main/index.html. Ημερομηνία πρόσβασης: 8-6-2021.

[28] Meinard Muller. *Short-Time Fourier Transform and Chroma Features*. https://www.audiolabs-erlangen.de/content/05-fau/professor/00-mueller/02-teaching/2016s_apl/LabCourse_STFT.pdf.

[29] *Getting to Know the Mel Spectrogram*. https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0. Ημερομηνία πρόσβασης: 8-6-2021.

[30] Christine Senac, Thomas Pellegrini, Florian Mouret και Julien Pinquier. *Music feature maps with convolutional neural networks for music genre classification. Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*. ACM, 2017.

[31] Sergey Ioffe και Christian Szegedy. *Batch Normalization: Accelerating deep network training by reducing internal covariate shift. arXiv [cs.LG]*, 2015.

[32] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever και Ruslan Salakhutdinov. *Dropout: A simple way to prevent neural networks from overfitting*.

[33] *Music Genre classification using Convolutional Neural Networks*. https://github.com/Hguimaraes/gtzan.keras. Ημερομηνία πρόσβασης: 8-6-2021.

[34] *Tensorflow library*. https://www.tensorflow.org/. Ημερομηνία πρόσβασης: 20-6-2021.

[35] *Keras API*. https://keras.io/. Ημερομηνία πρόσβασης: 20-6-2021.

[36] *Amazon Web Services*. https://aws.amazon.com/?nc2=h_lg. Ημερομηνία πρόσβασης: 20-6-2021.

[37] *Amazon SageMaker*. https://aws.amazon.com/sagemaker/. Ημερομηνία πρόσβασης: 20-6-2021.

[38] *Early Stopping*. https://keras.io/api/callbacks/early_stopping/. Ημερομηνία πρόσβασης: 20-6-2021.

[39] *Reduce Learning Rate on Plateau*. https://keras.io/api/callbacks/reduce_lr_on_plateau/. Ημερομηνία πρόσβασης: 18-6-2021.

[40] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst και Xavier Bresson. *FMA: A Dataset For Music Analysis. arXiv [cs.SD]*, 2016.

[41] *Oversampling and Undersampling: A technique for Imbalanced Classification*. https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf. Ημερομηνία πρόσβασης: 21-6-2021.

[42] Connor Shorten και Taghi M. Khoshgoftaar. *A survey on image data augmentation for deep learning. Journal of big data*, 6(1), 2019.

[43] Daniel S. Park, William Chan, Yu Zhang, Chung Cheng Chiu, Barret Zoph, Ekin D. Cubuk και Quoc V. Le. *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. Interspeech 2019*. ISCA, 2019.

[44] *GRU layer*. https://keras.io/api/layers/recurrent_layers/gru/. Ημερομηνία πρόσβασης: 28-6-2021.

[45] *Simple RNN layer*. https://keras.io/api/layers/recurrent_layers/simple_rnn/. Ημερομηνία πρόσβασης: 28-6-2021.

[46] *LSTM layer*. https://keras.io/api/layers/recurrent_layers/lstm/. Ημερομηνία πρόσβασης: 28-6-2021.

[47] *Comparison of RNN layers*. https://medium.com/@saurabh.rathor092/simple-rnn-vs-gru-vs-lstm-difference-lies-in-more-flexible-control-5f33e07b1e57. Ημερομηνία πρόσβασης: 28-6-2021.

[48] M. Schuster και K.K. Paliwal. *Bidirectional recurrent neural networks*. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[49] *Bidirectional wrapper for RNNs in Keras*. https://keras.io/api/layers/recurrent_layers/bidirectional/. Ημερομηνία πρόσβασης: 28-6-2021.

[50] *Merge Modes for Bidirectional LSTMs*. https://towardsdatascience.com/lstm-and-bidirectional-lstm-for-regression-4fddf910c655. Ημερομηνία πρόσβασης: 28-6-2021.