ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΜΕΛΕΤΗ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΛΟΓΙΣΜΙΚΟΥ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΜΟΥΣΙΚΩΝ ΚΟΜΜΑΤΙΩΝ

ΑΧΙΛΛΕΑΣ ΠΑΠΑΣΤΑΜΑΤΙΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Ευάγγελος Σπύρου
Επίκουρος Καθηγητής

Λαμία, 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΜΕΛΕΤΗ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΛΟΓΙΣΜΙΚΟΥ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΜΟΥΣΙΚΩΝ ΚΟΜΜΑΤΙΩΝ

ΑΧΙΛΛΕΑΣ ΠΑΠΑΣΤΑΜΑΤΙΟΥ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Ευάγγελος Σπύρου
Επίκουρος Καθηγητής

Λαμία, 2022

UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

STUDY AND IMPLEMENTATION OF SOFTWARE FOR MUSIC TRACK ANALYSIS

ACHILLEAS PAPASTAMATIOU

FINAL THESIS

ADVISOR

Evaggelos Spyrou
Assistant Professor

Lamia, 2022

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [(1)], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

*1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω τον συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

*2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: 15/07/2022

Ο –Η Δηλ.

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση
του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων
σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

ΠΕΡΙΛΗΨΗ

Ο διαχωρισμός αρχείων ήχου μη απωλεστικής συμπίεσης σε γνήσια και διακωδικοποιημένα παραμένει ανοιχτό πρόβλημα στη βιομηχανία διανομής της μουσικής. Σκοπός αυτής της εργασίας είναι η εξερεύνηση μεθόδων αξιολόγησης της συμπίεσης ήχου υψηλής ποιότητας και η πρόταση μίας νέας προσέγγισης με χρήση τεχνητής νοημοσύνης και βαθιάς μάθησης. Το συνελικτικό νευρωνικό δίκτυο που προέκυψε από τα πειράματα παρουσιάζει ανταγωνίσιμη ακρίβεια και ταχύτητα και θα μπορούσε να χρησιμοποιηθεί από φίλους της μουσικής ή πλατφόρμες διανομής μουσικής ως εργαλείο εκτίμησης της γνησιότητας αρχείων flac και wav. Το προτεινόμενο μοντέλο μπορεί να δοκιμαστεί διαδικτυακά μέσω της εφαρμογής *penthy*.

ABSTRACT

The discrimination of lossless audio files into genuine and transcoded persists as an open-ended problem in the music distribution industry. The goal of this study is the exploration of assessment methods for the compression of high quality sound and the proposal of a novel approach using artificial intelligence and deep learning. The convolutional neural network that has been implemented through the experiments manifests competitive accuracy and speed and it could be used by audiophiles or music distribution platforms as an authenticity evaluation tool for flac and wav files. The suggested model can be tested online, through the *penthy* application.

Table of Contents

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή

1.1 Το πρόβλημα της γνησιότητας

1.1.α Βασικά χαρακτηριστικά

Η αποθήκευση του ήχου σε ψηφιακή μορφή περιλαμβάνει την εύρεση μίας ισορροπίας ανάμεσα σε ποιότητα ακρόασης, ακρίβεια αναπαραγωγής και κατανάλωση αποθηκευτικού χώρου. Ο ψηφιακός ήχος περιγράφεται από διάφορα χαρακτηριστικά, όπως η συχνότητα δειγματοληψίας (sample frequency), το βάθος bit (bit depth), ο αριθμός των καναλιών (channels) και ο ρυθμός μετάδοσης (bit rate).

- Η συχνότητα δειγματοληψίας, (μετριέται σε Hz) είναι το πλήθος των δειγμάτων ενός σήματος που καταγράφονται για την ψηφιοποίησή του ή την κωδικοποίησή του σε διάρκεια σήματος ενός δευτερολέπτου.
- Το βάθος bit είναι το μέγεθος κάθε δείγματος και αντιστοιχεί στο πλήθος πιθανών τιμών που μπορεί να πάρει ένα δείγμα.
- Ο αριθμός των καναλιών είναι το πλήθος των τμημάτων του σήματος που μπορούν να αναπαραχθούν σε ξεχωριστά σημεία (για παράδειγμα, σε διαφορετικά ηχεία) ή να επεξεργαστούν ξεχωριστά. Δύο κανάλια δημιουργούν ένα στερεοφωνικό σήμα.
- Ο ρυθμός μετάδοσης είναι ένας άλλος τρόπος μέτρησης της ανάλυσης του σήματος. Υπολογίζεται ως πλήθος δεδομένων ανά μονάδα χρόνου, συνήθως kbps (kilobits per second). (Faruq, 2022)

Τα παραπάνω μεγέθη μπορούν να χρησιμοποιηθούν για να χαρακτηρίσουν την ποσότητα πληροφορίας που υπάρχει σε ένα αρχείο ήχου και επηρεάζουν το μέγεθός του. Αναφορές στην ανάλυση του ήχου (audio resolution) παραπέμπουν σε αυτά τα μεγέθη.

1.1.β Συμπίεση ήχου

Όπως με κάθε ψηφιακό σήμα ή αρχείο, ο ήχος συμπιέζεται για την εξοικονόμηση χώρου. Η συμπίεση είναι η διαδικασία μετατροπής του τρόπου αποθήκευσης των δεδομένων σε μορφή που απαιτεί λιγότερο αποθηκευτικό χώρο. Υπάρχουν δύο βασικές κατηγορίες συμπίεσης, η μη απωλεστική (lossless compression) και η απωλεστική (lossy compression). Μη απωλεστική είναι η συμπίεση που διατηρεί τα δεδομένα στην αρχική τους μορφή όταν αποσυμπιέζονται, όπως η δημιουργία ενός αρχείου zip από έναν φάκελο αρχείων. Αντιθέτως, στην απωλεστική συμπίεση, ένα μικρό μέρος της αποθηκευμένης πληροφορίας χάνεται ή αλλοιώνεται. Το πλεονέκτημα της απωλεστικής συμπίεσης είναι η δυνατότητα εξοικονόμησης περισσότερου χώρου. Φυσικά, κάθε υπάρχουσα επιλογή στοχεύει σε συγκεκριμένη χρήση. (Cunningham and McGregor, 2019)

Υπάρχουν διάφοροι κωδικοποιητές ήχου που χρησιμοποιούν μη απωλεστική συμπίεση και διάφοροι που χρησιμοποιούν απωλεστική. Σε αυτήν την εργασία θα αναφερθούμε κυρίως σε αρχεία μη απωλεστικής συμπίεσης *flac* (Free Lossless Audio Codec), αρχεία απωλεστικής συμπίεσης *mp3* (MPEG Audio Layer III) και ασυμπίεστα αρχεία *wav* (Waveform Audio File Format). Με τον όρο ασυμπίεστα, εννοείται ότι η κωδικοποίηση των δεδομένων δε χρησιμοποιεί κάποια μέθοδο μείωσης του μεγέθους τους.

Ένα μουσικό κομμάτι αποθηκευμένο σε κωδικοποίηση flac καταναλώνει πολύ περισσότερο αποθηκευτικό χώρο σε σύγκριση με το ίδιο κομμάτι αποθηκευμένο ως mp3. Ωστόσο, τα αρχεία flac προτιμούνται από πολλούς ακροατές, φίλους της μουσικής ή και επαγγελματίες της βιομηχανίας της μουσικής για τον πιο πλούσιο ήχο τους και τη μικρότερη διαφορά τους από το πρωτότυπο κομμάτι. Στον χώρο διανομής της μουσικής, είναι φυσικό η ψηλότερη ανάλυση του ήχου ή η πιο γενναιόδωρη συμπίεση να κοστολογούνται ακριβότερα. Για παράδειγμα, η πλατφόρμα Spotify προσφέρει ψηλότερη ποιότητα ήχου στους συνδρομητές της premium έκδοσης, από ότι στους χρήστες της δωρεάν έκδοσης (Spotify, 2022). Ομοίως, η πλατφόρμα Tidal παρέχει δύο πλάνα επί πληρωμή, με το ακριβότερο να προσφέρει ψηλότερη ανάλυση (Tidal, 2022). Επίσης, η αγορά ενός άλμπουμ σε μορφή flac μπορεί να είναι ακριβότερη από ότι σε μορφή mp3.

1.1.γ Διακωδικοποίηση

Η αλλαγή της κωδικοποίησης ενός σήματος ονομάζεται διακωδικοποίηση (transcoding). Η μετατροπή ενός αρχείου flac σε mp3 αποτελεί παράδειγμα transcoding. Το αρχείο που παράγεται χρειάζεται σημαντικά λιγότερο αποθηκευτικό χώρο, αλλά εξασκημένοι ακροατές θα το κρίνουν ως κατώτερης ποιότητας από το αρχικό. (Corbett, 2012)

Η αντίστροφη μετατροπή, από mp3 σε flac, είναι δυνατή, αλλά παράγει ένα αρχείο που συνδυάζει τα μειονεκτήματα των δύο κωδικοποιήσεων: μέγεθος περίπου όσο το πρωτότυπο flac και ποιότητα ακρόασης όμοια με το mp3. Επομένως, ένα τέτοιο αρχείο δεν έχει κάποια πρακτική αξία και μπορεί να θεωρηθεί παραπλανητικό, καθώς δεν προσφέρει την αναμενόμενη εμπειρία. Στη συνέχεια, θα αναφερόμαστε σε τέτοιου είδους αρχεία ως *transcoded flac*. Ο όρος αυτός δεν έχει καμία σχέση με μετατροπή ενός αρχείου flac σε flac με διαφορετική ανάλυση ή με μετατροπή ενός wav σε flac.

Αν, στο ίδιο παράδειγμα, γνωρίζουμε ότι το αρχικό flac είχε ποιότητά πανομοιότυπη με το τελικό προϊόν που παρήγαγε το μουσικό στούντιο και δεν είχε υποστεί κάποια αλλοίωση κατά τη διάρκεια της κυκλοφορίας, τότε μπορούμε να το θεωρήσουμε γνήσιο. Ένα γνήσιο κομμάτι αναμένεται να είναι πραγματικά μη απωλεστικά συμπιεσμένο ή ασυμπίεστο. Δηλαδή, είτε είναι αποθηκευμένο σε flac, είτε σε wav είτε σε κάποια άλλη κωδικοποίηση μη απωλεστικής συμπίεσης, να μην έχει υποβληθεί σε transcoding προς καμία απωλεστική κωδικοποίηση. Στη συνέχεια, τέτοια αρχεία θα αναφέρονται ως *truly lossless* και εφόσον είναι αρχεία flac, *truly lossless flac*. Για παράδειγμα, ένα πρωτότυπο wav κομμάτι θεωρείται truly lossless. Η κωδικοποίησή του σε flac με την ίδια ανάλυση παράγει ένα truly lossless flac. Η μετατροπή αυτού του αρχείου σε flac με χαμηλότερη ανάλυση (transrating) παράγει ξανά ένα truly lossless flac. Όμως, η κωδικοποίηση του αρχικού wav σε mp3 και η διαδοχική μετατροπή του mp3 σε flac παράγει ένα transcoded flac.

Το πρόβλημα προκύπτει όταν χρειαζόμαστε κάποιας μορφής εγγύηση ότι ένα αρχείο είναι truly lossless. Με άλλα λόγια, πώς ξεχωρίζουμε ένα truly lossless flac από ένα transcoded flac; Ακόμη χειρότερα, πώς αναγνωρίζουμε αν ένα αρχείο είναι truly lossless flac ή transcoded flac αν έχουμε μόνο μία εκδοχή του και όχι το πρωτότυπο για να τα συγκρίνουμε; Αυτή η εργασία προσπαθεί να απαντήσει αυτό το ερώτημα με μία λύση που βασίζεται στην τεχνητή νοημοσύνη.

1.1.δ Άλλες πτυχές του προβλήματος

Η κατηγοριοποίηση flac ή wav αρχείων σε γνήσια και παραποιημένα είναι περίπλοκη, γιατί ο ήχος μπορεί να αλλοιωθεί με πολλούς διαφορετικούς τρόπους. Το transcoding είναι ένας από αυτούς. Δηλαδή ένα αρχείο που αποδεικνύεται ότι είναι truly lossless δεν είναι αυτομάτως γνήσιο. Μία απλή επεξεργασία του αρχείου όπως η αφαίρεση ενός τμήματός του και μείωση της διάρκειάς του καταστρέφει το έργο, χωρίς να επηρεάζει τα χαρακτηριστικά του. Επιπλέον, η ψευδής αύξηση της ανάλυσης (upsampling ή upscaling), παρομοίως με το transcoding, δημιουργεί ένα αρχείο που καταλαμβάνει περισσότερο χώρο από όσο χρειάζεται και εξαπατά τον χρήστη για την πραγματική ποσότητα πληροφορίας που περιέχει. Η αναγνώριση αρχείων με συχνότητα δειγματοληψίας ή βάθος bit μεγαλύτερο από το πραγματικό δεν είναι εύκολη. Ο προφανής έλεγχος των μεταδεδομένων (metadata) δεν επαρκεί, καθώς τα μεταδεδομένα περιγράφουν την τωρινή κωδικοποίηση του αρχείου και όχι τις προηγούμενες μορφές του.

Για τις περιπτώσεις transcoding από πηγές απωλεστικής συμπίεσης, η αναγνώριση αρχείων transcoded από mp3 είναι διαφορετική υπόθεση από την αναγνώριση transcoding από άλλες κωδικοποιήσεις. Για παράδειγμα, ένα transcoded flac που προήλθε από ένα αρχείο απωλεστικής συμπίεσης Vorbis, μπορεί να παρουσιάζει διαφορετικές αλλοιώσεις από ένα transcoded από mp3.

Καταλήγουμε έτσι στο συμπέρασμα, ότι η πιστοποίηση ενός αρχείου ως γνήσιο απαιτεί την επαλήθευση απουσίας κάθε πιθανής παραμόρφωσης. Εφόσον δεν υπάρχει πρόσβαση σε μία ήδη πιστοποιημένη εκδοχή του ίδιου κομματιού για να γίνει σύγκριση, κάθε σενάριο πιθανής παραμόρφωσης πρέπει να ελεγχθεί.

Σε αυτή την εργασία θα μελετηθεί μόνο το transcoding από πηγές mp3. Η προτεινόμενη προσέγγιση διαχωρίζει αποκλειστικά σε truly lossless ως προς mp3 transcoding και transcoded ως προς mp3 transcoding. Άλλοι πιθανοί τύποι lossy transcoding και άλλες πιθανές παραμορφώσεις δεν ελέγχονται.

Η αναγνώριση mp3 transcoding, όπως και η αναγνώριση οποιασδήποτε άλλης αλλοίωσης ενός αρχείου από το πρωτότυπο, μπορεί να είναι χρήσιμη για όλες τις πλευρές της βιομηχανίας διακίνησης της μουσικής. Οι χρήστες υπηρεσιών ροής (streaming) και οι αγοραστές ψηφιακών αντιτύπων δίσκων μπορούν να επιβεβαιώσουν ότι παρέλαβαν αυτό για το οποίο πλήρωσαν. Οι πλατφόρμες μεταπώλησης και streaming μπορούν να ελέγξουν το περιεχόμενο που προσθέτουν στο σύστημά τους και να εξασφαλίσουν αξιοπιστία για το πελατειακό τους κοινό. Ίσως να μπορούν ακόμα και να προωθήσουν τις υπηρεσίες τους με το επιχείρημα του επιπρόσθετου ποιοτικού ελέγχου. Οι καλλιτέχνες επωφελούνται γνωρίζοντας ότι περισσότεροι ακροατές θα ακούσουν το έργο τους στη μορφή που οι ίδιοι αποζήτησαν. Η ανάπτυξη τεχνολογιών αξιολόγησης ποιότητας ήχου ενδέχεται μέχρι και να επαυξήσει την ενημερότητα των ακροατών, με αποτέλεσμα διευρυμένη ζήτηση ήχου υψηλών προδιαγραφών. Τελικώς, τόσο η εύρεση τεχνικών σφαλμάτων όσο και η διερεύνηση ενδείξεων απάτης με τέτοιες τεχνολογίες θα μπορούσαν να ωφελήσουν τη μουσική παραγωγή και εμπορία.

1.2 Βιβλιογραφική Επισκόπηση

Η υπάρχουσα έρευνα που αφορά λύσεις του προβλήματος που περιγράφηκε είναι περιορισμένη. Παρ' όλα αυτά, έχουν εμφανιστεί κατά καιρούς διάφορες εφαρμογές για την αξιολόγηση της γνησιότητας flac αρχείων.

Χαρακτηριστικό παράδειγμα αποτελεί το πρόγραμμα Audio Checker (Opris, 2013), το οποίο όμως δε συντηρείται πλέον. Το Audio Checker εκτιμάει την προέλευση ενός μουσικού κομματιού και αναγνωρίζει πιθανές MPEG πηγές. Υποστηρίζει μόνο αρχεία περιορισμένης ανάλυσης.

Το πρόγραμμα Lossless Audio Checker είναι άλλη μία προσέγγιση του προβλήματος και καθιστά σαφές πως η γνησιότητα των αρχείων flac επηρεάζεται από πολλούς παράγοντες, όπως αναφέρθηκε στο 1.1.δ. Στοχεύει στην αναγνώριση ψευδούς ανάλυσης και transcoding από απωλεστική κωδικοποίηση AAC (Advanced Audio Coding). Η κωδικοποίηση AAC έχει πολλά κοινά με την κωδικοποίηση των αρχείων mp3. (Lacroix et al., 2015)
Ο αλγόριθμος που χρησιμοποιείται δε στηρίζεται σε μηχανική μάθηση και επιτυγχάνει εντυπωσιακή ακρίβεια για αρχεία της AAC κωδικοποίησης. Βασική μέθοδος είναι η εύρεση σφάλματος κβαντισμού (quantization error) που δημιουργείται κατά την μετατροπή του αρχείου. (Derrien, 2019)

Μία άλλη σχετική διαφοροποίηση αρχείων ήχου είναι με βάση την ανάλυση κομματιών απωλεστικής συμπίεσης με την ίδια κωδικοποίηση. Οι D'Alessandro και Shi δημιούργησαν ένα σύστημα αναγνώρισης του ρυθμού μετάδοσης αρχείων mp3. Με τη χρήση μίας Μηχανής Διανυσμάτων Υποστήριξης (Support Vector Machine – SVM), ανιχνεύεται η πραγματική ανάλυση του σήματος. (D'Alessandro and Shi, 2009)
Αν και αυτή η έρευνα ασχολείται με τις διαφορές μεταξύ σημάτων απωλεστικής συμπίεσης, η ποικιλομορφία των αρχείων mp3 ως προς τον ρυθμό μετάδοσης σχετίζεται άμεσα με την αναγνώριση mp3 transcoding. Ένα εργαλείο ανίχνευσης transcoded flac πρέπει να αναγνωρίζει όλες τις πιθανές εκδοχές μίας mp3 πηγής. Εναλλακτικά, η διαφορά μεταξύ truly lossless flac και transcoded flac πρέπει να είναι ορατή για transcoded flac χαμηλής αλλά και υψηλής ανάλυσης. Το προτεινόμενο μοντέλο που θα περιγραφεί παρακάτω ξεπερνάει αυτό το εμπόδιο.
Οι D'Alessandro και Shi χρησιμοποίησαν μόνο τις ψηλότερες συχνότητες των μουσικών κομματιών για την ταξινόμησή τους, καθώς σε εκείνο το φάσμα οι διαφορές είναι πιο προφανείς. Αυτή η παρατήρηση συνάδει με τον μηχανισμό του προτεινόμενου μοντέλου.

Σε αυτό το σημείο, αξίζει να αναφερθεί ότι υπάρχουν μελέτες για την ανίχνευση transcoding και άλλων αλλοιώσεων σε σήματα βίντεο. Παρόμοια με τον ήχο, το βίντεο μπορεί να τροποποιηθεί με τρόπο που χάνει την αξία του. Η επαλήθευση της γνησιότητας αρχείων βίντεο είναι χρήσιμη στην εγκληματολογία, αφού επιβεβαιώνει αν ένα στοιχείο παρουσιάζει πραγματικά γεγονότα ή έχει υποστεί ψηφιακή επεξεργασία. (Singh and Aggarwal, 2015. Xu et al., 2012)

1.3 Προτεινόμενη Λύση

Ανακεφαλαιώνοντας, η αξιόπιστη διαφοροποίηση γνήσιων αρχείων ήχου μη απωλεστικής συμπίεσης και παραποιημένων αρχείων κωδικοποιημένων με την ίδια μορφή είναι ένα ευρύ και δύσκολο πρόβλημα δυαδικής κατηγοριοποίησης (binary classification). Μέρος του προβλήματος είναι η αναγνώριση αρχείων flac κωδικοποιημένων προηγουμένως ως mp3 (transcoded flac, βλέπε 1.1.γ). Το προτεινόμενο σύστημα είναι ένα μοντέλο βαθιάς μάθησης (deep learning) και συγκεκριμένα ένα Συνελικτικό Νευρωνικό Δίκτυο (Convolutional Neural Network – CNN) που αναγνωρίζει τις απαραίτητες αλλοιώσεις στο φασματογράφημα (spectrogram) ενός κομματιού με αξιοπρεπή ακρίβεια και ταχύτητα. Η είσοδος του συστήματος είναι ένα μουσικό κομμάτι, το οποίο διαιρείται σε μικρά χρονικά διαστήματα και ένα φασματογράφημα δημιουργείται για κάθε τμήμα. Τα φασματογραφήματα αποτελούν την είσοδο του CNN, το οποίο τα αναλύει και επιστρέφει έναν πραγματικό αριθμό. Αυτός ο αριθμός ερμηνεύεται κατάλληλα και επομένως η έξοδος του συστήματος είναι ο ισχυρισμός ότι το συγκεκριμένο αρχείο είτε είναι ή δεν είναι transcoded. Στα επόμενα κεφάλαια περιγράφεται πιο αναλυτικά η δομή του μοντέλου.

- Στο κεφάλαιο 2 δίνονται περισσότερες πληροφορίες σχετικά με τις διαθέσιμες επιλογές κωδικοποίησης ήχου με απωλεστική ή μη απωλεστική συμπίεση, πως λειτουργεί το πρότυπο mp3 και πως συγκρίνεται με το flac (2.1). Επίσης, παρουσιάζεται η αρχή παραγωγής και η χρήση των φασματογραφημάτων ήχου (2.2). Τέλος, εξηγούνται βασικοί όροι τεχνητής νοημοσύνης και η διαδικασία που χρησιμοποιεί ένα CNN για να λειτουργήσει (2.3).
- Στο κεφάλαιο 3 ξεκινάει η προβολή του προτεινόμενου μοντέλου. Περιγράφεται το σύνολο δεδομένων που χρησιμοποιήθηκε και η προετοιμασία του, αλλά και η πραγματική αρχιτεκτονική του νευρωνικού δικτύου.
- Στο κεφάλαιο 4 παρουσιάζονται τα αποτελέσματα των πειραμάτων και η ακρίβεια του τελικού μοντέλου στο σύνολο δεδομένων επαλήθευσης.
- Τελικώς, στο κεφάλαιο 5 συνοψίζονται τα συμπεράσματα της εργασίας.

ΚΕΦΑΛΑΙΟ 2 Θεωρητικό Υπόβαθρο

2.1 Κωδικοποίηση ήχου

2.1.α Το πρότυπο mp3

Η πρώτη εκδοχή του προτύπου mp3 (MPEG Audio Layer III) παρουσιάστηκε το 1991. Για να επιτευχθεί αυτό, μία ομάδα από εμπειρογνώμονες, το Moving Picture Experts Group (MPEG), επιχειρούσε στην ανάπτυξη του προτύπου από την ίδρυσή του το 1988. Η συνεργασία αυτή δημιουργήθηκε από τον Διεθνή Οργανισμό Τυποποίησης (International Organization for Standardization – ISO) και με συμβολή της Διεθνούς Ηλεκτροτεχνικής Επιτροπής (International Electrotechnical Commission – IEC). Η Εταιρεία Φράουνχοφερ (Fraunhofer Society | Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung) δημιούργησε σημαντικές εφαρμογές λογισμικού, αλλά και υλικού για το mp3 και επηρέασε την εμπορική του διάδοση. Η σύντομη εκδοχή του ονόματος του προτύπου ως *mp3* ήταν μία προσπάθεια προώθησης στο κοινό. (Musmann, 2006)

Η διάδοση του τύπου κωδικοποίησης mp3 με απωλεστική συμπίεση άλλαξε τα δεδομένα στον κόσμο αποθήκευσης και διαμοιρασμού της μουσικής. Τα νέα αρχεία ήχου έγιναν εύχρηστα και προσιτά. Σημαντικό επίτευγμα ήταν η μείωση του χώρου που επιτυγχάνεται με αυτή την κωδικοποίηση. Αυτή η βελτίωση βασίζεται κυρίως στα ψυχοακουστικά φαινόμενα που συνδέονται με την ακρόαση από τους ανθρώπους. Πρόκειται για την ανακάλυψη ότι κάποιοι ήχοι μπορούν να παραλειφθούν ή να παραποιηθούν με τέτοιο τρόπο που η αποθήκευση των δεδομένων γίνεται ευκολότερη χωρίς να βεβηλώνεται αισθητά η ποιότητα του κομματιού. Συνοπτικά, οι 3 πιθανές περιπτώσεις ψυχοακουστικής που εκμεταλλεύεται το πρότυπο είναι:

- Ο άνθρωπος δεν ακούει κάποιες συχνότητες.
- Ο άνθρωπος δεν ακούει όλες τις συχνότητες με την ίδια ένταση.
- Όταν δύο συχνότητες ακούγονται ταυτόχρονα, ο άνθρωπος κατανοεί κυρίως την ισχυρότερης έντασης συχνότητα.

(Jayant et al., 1993)

Την έλλειψη υψηλών συχνοτήτων που προκύπτει από τη χρήση ψυχοακουστικής, χρησιμοποιεί και το προτεινόμενο μοντέλο, για την αναγνώριση mp3 κωδικοποίησης.

2.1.β Το πρότυπο flac

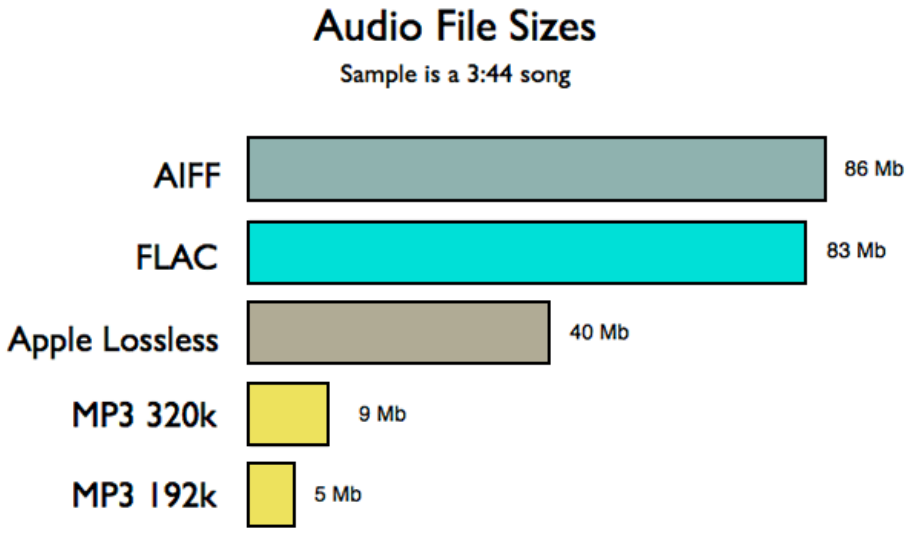
Από την απαρχή του προτύπου το 2000, το flac (Free Lossless Audio Codec) δημοσιεύτηκε επίσημα το 2001 από τον Josh Coalson. Το 2003 το ίδρυμα Xiph (Xiph.Org Foundation) ανέλαβε την ανάπτυξη και συντήρηση του νέου προτύπου. (Xiph.org, 2022)

Με διαφορετικό σκοπούμενο κοινό, ή μάλλον διαφορετική σκοπούμενη χρήση, η κωδικοποίηση flac με μη απωλεστική συμπίεση στόχευε στην περιορισμένη εξοικονόμηση χώρου με πρώτη προτεραιότητα την διατήρηση των δεδομένων στην αρχική τους μορφή. Ακριβότερος υλικός εξοπλισμός μπορεί να αξιοποιήσει την ποιοτική διαφορά του ήχου που αναπαράγεται από αυτή την κωδικοποίηση, αλλά ο απαιτούμενος χώρος είναι διακριτά μεγαλύτερος από τις κωδικοποιήσεις απωλεστικής συμπίεσης. Έτσι, παρεμποδίζεται η εξαπλωμένη χρήση του flac σε φορητές συσκευές όπως κινητά τηλέφωνα και συσκευές

αναπαραγωγής πολυμέσων. Φυσικά, η ποιοτική σύγκριση με το mp3 έχει νόημα μόνο αν ορίσουμε την ανάλυση του κάθε αρχείου, καθώς η κωδικοποίηση σε mp3 με πολύ υψηλό ρυθμό μετάδοσης (για παράδειγμα. 320 kbps) καλύπτει όλες τις ανάγκες των περισσότερων ακροατών.

2.1.γ Σύγκριση απόδοσης mp3 και flac

Όπως προαναφέρθηκε, η κωδικοποίηση mp3 εξοικονομεί πολύ περισσότερο χώρο από την κωδικοποίηση flac. Δεν είναι υπερβολή να ισχυριστούμε ότι σε γενικά πλαίσια το ίδιο μουσικό κομμάτι ως flac καταναλώνει περίπου 10 φορές το μέγεθος του ως mp3. Το επόμενο παράδειγμα περιλαμβάνει διαφορετικές αναλύσεις mp3 και άλλες κωδικοποιήσεις ως μέτρο σύγκρισης, για ένα κομμάτι διάρκειας 3 λεπτών και 44 δευτερολέπτων.



Εικόνα 1 – (Callum, 2018)

Αξίζει ακόμη να μελετήσουμε τον χώρο που εξοικονομείται από τη συμπίεση flac, σε σύγκριση με ασυμπίεστο ήχο, όπως θα αποθηκευόταν δηλαδή σε ένα αρχείο wav. Κατά κανόνα, ένα αρχείο flac χρειάζεται περίπου 50% - 70% λιγότερο χώρο από την ασυμπίεστη μορφή του (Pigeon, 2022). Επομένως, η μη απωλεστική συμπίεση μπορεί να επιφέρει οφέλη και επίσης υπό αυτό το πρίσμα παρουσιάζεται η εντυπωσιακή μείωση μεγέθους της απωλεστικής συμπίεσης.

Για τη σύγκριση του mp3 με το αρχικό υλικό προ συμπίεσης, παρατίθενται δύο πίνακες με τα εκτιμώμενα μεγέθη αρχείων, υπολογισμένα με βάση τον ρυθμό μετάδοσης, για στερεοφωνικό ήχο.

Settings	Bitrate	File size per second	File size per minute	File size per hour
16 bit, 44.1 KHz	1,411.2 Kbps	176.4 KB	10.584 MB	635.04 MB
16 bit, 48 KHz	1,536 Kbps	192 KB	11.520 MB	691.2 MB
24 bit, 48KHz	2,304 Kbps	288 KB	17.28 MB	1.036 GB
24 bit, 96KHz	4,608 Kbps	576 KB	34.56 MB	2.0736 GB

Εικόνα 2 – Μη συμπιεσμένα αρχεία - Εκτίμηση μεγέθους

Bitrate	File size per second	File size per minute	File size per hour
8 Kbps	1 KB	60 KB	3.6 MB
16 Kbps	2 KB	120 KB	7.2 MB
32 Kbps	4 KB	240 KB	14.4 MB
40 Kbps	5 KB	300 KB	18.0 MB
48 Kbps	6 KB	360 KB	21.6 MB
56 Kbps	7 KB	420 KB	25.2 MB
64 Kbps	8 KB	480 KB	28.8 MB
80 Kbps	10 KB	600 KB	36.0 MB
96 Kbps	12 KB	720 KB	43.2 MB
112 Kbps	14 KB	840 KB	50.4 MB
128 Kbps	16 KB	960 KB	57.6 MB
160 Kbps	20 KB	1.20 MB	72.0 MB
192 Kbps	24 KB	1.44 MB	86.4 MB
224 Kbps	28 KB	1.68 MB	100.8 MB
256 Kbps	32 KB	1.92 MB	115.2 MB
320 Kbps	40 KB	2.40 MB	144.0 MB

Εικόνα 3 – Αρχεία MP3 - Εκτίμηση μεγέθους

(AudioMountain, 2022)

2.1.δ Άλλες κωδικοποιήσεις

Πέρα από τις κωδικοποιήσεις που μας αφορούν άμεσα για αυτή την εργασία, αξίζει να αναφερθούν άλλες επιλογές απωλεστικής και μη απωλεστικής συμπίεσης που έχουν αναπτυχθεί για διάφορους σκοπούς. Με τη διαθέσιμη ποικιλία προτύπων, γίνεται φανερό πόσοι πιθανοί συνδυασμοί transcoding μπορούν να υπάρξουν και επομένως πόσο πολύπλευρο είναι το πρόβλημα της γνησιότητας. Ακολουθούν επιγραμματικά επιπρόσθετες κωδικοποιήσεις ήχου.

Απωλεστικής συμπίεσης:

- *Advanced Audio Coding (AAC)*, δημιουργήθηκε από μία μεγάλη συνεργασία εταιρειών με σκοπό να διαδεχθεί το πρότυπο mp3 ως βελτιωμένη εκδοχή του. Το πρότυπό του ενσωματώθηκε στο σώμα MPEG.
- *Dolby AC-3* (*Audio Compression format 3*, συνδεδεμένο με το *Dolby Digital*), σχεδιάστηκε από την Dolby για τον κινηματογράφο και αργότερα χρησιμοποιήθηκε και σε άλλα μέσα.
- *aptX (audio processing technology X)*, δημιουργήθηκε από την Qualcomm για ασύρματες επικοινωνίες.
- *Enhanced Voice Services (EVS)*, αναπτύχθηκε από μία συνεργία εταιρειών για μετάδοση ομιλίας.
- *Vorbis/Ogg*, αναπτύχθηκε από το ίδρυμα Xiph.Org. Η κωδικοποίηση Vorbis χρησιμοποιείται σε αρχεία Ogg, όπως και οι κωδικοποιήσεις *Opus* και *Constrained Energy Lapped Transform (CELT)* που συντηρούνται από το ίδιο ίδρυμα. Μπορούν να θεωρηθούν ελεύθερες εναλλακτικές του προτύπου mp3 και προβάλλουν ανταγωνιστικά χαρακτηριστικά.
- *Windows Media Audio (WMA)*, αναπτύχθηκε από τη Microsoft ως μέρος μίας οικογένειας κωδικοποιήσεων. Θεωρήθηκε αντίστοιχο του mp3.

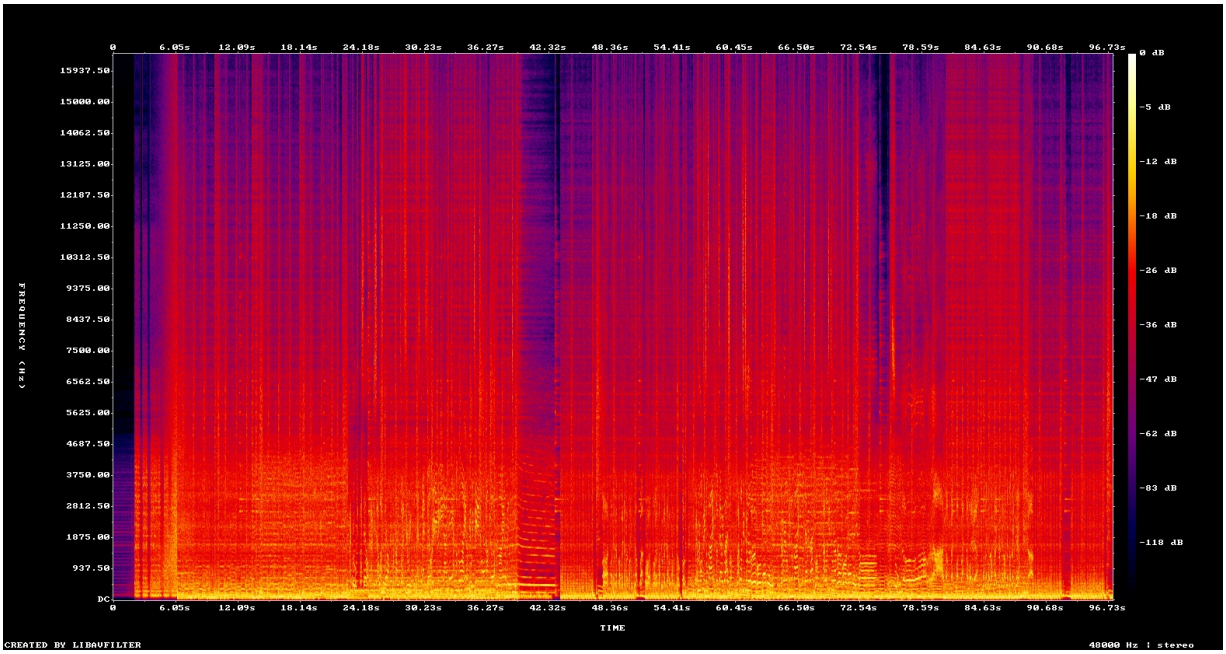
Μη απωλεστικής συμπίεσης:

- *Apple Lossless Audio Codec (ALAC)*, αναπτύχθηκε από την Apple για μουσική.
- *Monkey's Audio (ape)*, έργο του Matthew T. Ashland.
- *OptimFROG*, αναπτύχθηκε από τον Florin Ghido, βελτιστοποιημένο για υψηλή συμπίεση. Επιτυγχάνει μικρό μέγεθος αρχείων, αυξάνοντας τον χρόνο κωδικοποίησης και αποκωδικοποίησης.
- *True Audio (TTA)*, χρησιμοποιείται για ήχο πολλών καναλιών. Δημιουργήθηκε από τον Aleksander Djourik.
- *Windows Media Audio Lossless (WMA Lossless)*, πρόκειται για τη μη απωλεστική κωδικοποίηση της οικογένειας WMA της Microsoft.

(Wikipedia, 2022)

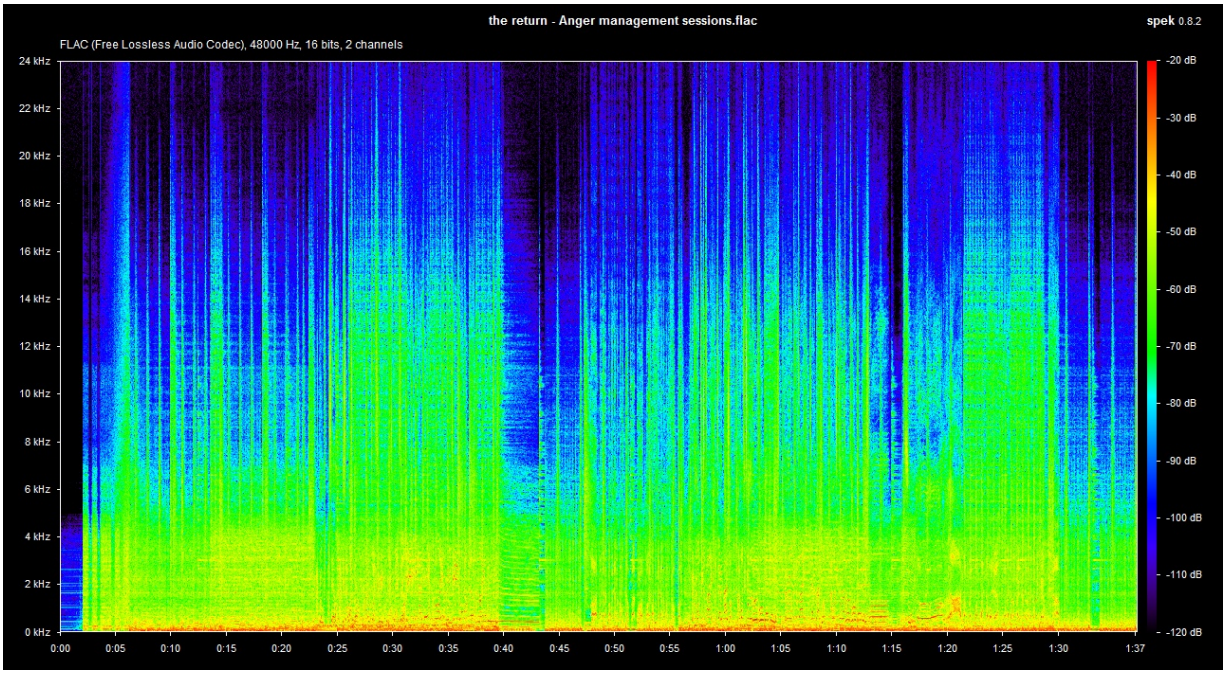
2.2 Φασματογράφημα

Το φασματογράφημα (spectrogram) είναι ένας τύπος διαγράμματος που αποτυπώνει πληροφορία για κάποια φάσμα. Στον ήχο, απεικονίζει την ένταση για το φάσμα της συχνότητας. Πιο συγκεκριμένα, είναι ένα γράφημα με 3 άξονες. Ο οριζόντιος άξονας αντιστοιχεί στον χρόνο, ο κατακόρυφος στη συχνότητα του ήχου και με τον χρωματισμό αντιπροσωπεύεται ο τρίτος άξονας, η ένταση του ήχου. Το spectrogram ενός κομματιού είναι μία μέθοδος να σχηματιστεί η ηχητική πληροφορία σε μορφή εικόνας. Με αυτόν τον τρόπο μπορούμε να εντοπίσουμε μοτίβα που θα ήταν δύσκολο να προσέξουμε σε μία ακρόαση και προγραμματιστικά υπάρχει η δυνατότητα της αποθήκευσης της εικόνας σε μορφή πίνακα για ευκολότερη επεξεργασία ή ανάλυση. Ακολουθούν παραδείγματα spectrograms του αρχείου *Anger management sessions* από τους *the return* (εικόνες 4, 5 και 6).

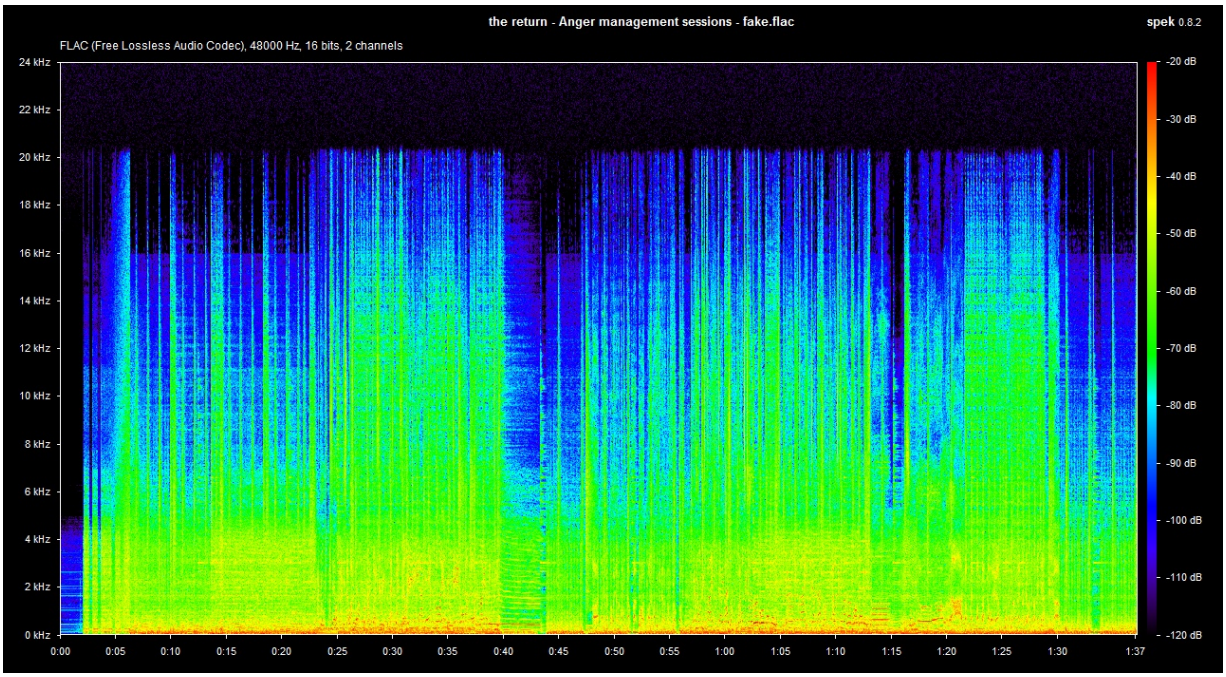


Εικόνα 4 – Φασματογράφημα του αρχείου *Anger management sessions* από τους *the return* Παράχθηκε με εργαλείο της βιβλιοθήκης FFmpeg.

Στη συνέχεια ας συγκρίνουμε τα spectrograms του ίδιου κομματιού, αποθηκευμένου ως flac για ένα αρχείο truly lossless (εικόνα 5) και ένα transcoded από mp3 (εικόνα 6). Έμφαση πρέπει να δοθεί στις ψηλότερες συχνότητες, κυρίως στο διάστημα 20 – 24 kHz, δηλαδή στο ανώτερο τμήμα των εικόνων. Όσο πιο σκουρόχρωμη η περιοχή, τόσο λιγότερη πληροφορία υπάρχει διαθέσιμη για αναπαραγωγή. Το αν η πληροφορία που λείπει από το transcoded flac σε αυτές τις συχνότητες είναι άμεσα ανιχνεύσιμη από την ανθρώπινη ακοή δε μας απασχολεί. Στόχος είναι η εύρεση ενός χαρακτηριστικού, εύκολα ανιχνεύσιμου σε μορφή εικόνας, που μαρτυράει την ύπαρξη transcoding.



Εικόνα 5 – Truly lossless flac για το μουσικό κομμάτι *Anger management sessions* από τους *the return*



Εικόνα 6 – Transcoded flac για το μουσικό κομμάτι *Anger management sessions* από τους *the return*

Φασματογραφήματα που παράχθηκαν με το πρόγραμμα spek.

Παρατηρούμε μεγάλη απώλεια πληροφορίας πάνω από τα 20 kHz, σε συμφωνία με ό,τι έχει αναφερθεί σε προηγούμενα κεφάλαια. Αυτή είναι επαρκής διαφορά για να εκπαιδευτεί το νευρωνικό δίκτυο.

2.3 Τεχνητή Νοημοσύνη

2.3.α Βασικοί όροι

Η Τεχνητή Νοημοσύνη (Artificial Intelligence – AI) είναι ένας ευρύς όρος που χρησιμοποιείται για πολλές διαφορετικές τεχνολογίες. Τα τελευταία χρόνια έχει σημειώσει τεράστια ανάπτυξη, καταλήγοντας να αποτελεί μία ξεχωριστή ερευνητική περιοχή της Πληροφορικής. Θα μπορούσαμε να θεωρήσουμε ως AI όλες τις προγραμματιστικές μεθόδους που αυτοματοποιούν εργασίες που προηγουμένως θα γινόντουσαν από ανθρώπους λόγω πολυπλοκότητας. Με τεχνολογίες AI ανοίγονται δρόμοι για εφαρμογές που θα ήταν πολύ δύσκολο να υλοποιηθούν με παραδοσιακό, ντετερμινιστικό προγραμματισμό. (Chollet, 2021)

Η Μηχανική Μάθηση (Machine Learning) αποτελεί μέρος της Τεχνητής Νοημοσύνης. Με τον όρο Machine Learning αναφερόμαστε σε αυτές τις υποκατηγορίες AI που επιτρέπουν στο πρόγραμμα να «μάθει» και να εξελιχθεί με κάποιο βαθμό αυτονομίας, ξεφεύγοντας από το κλασσικό μοντέλο αυστηρών κανόνων του παραδοσιακού προγραμματισμού. Όταν υπάρχει ανάγκη για μία μαθηματική συνάρτηση που θα περιγράφει ένα σύνολο δεδομένων, αλλά οι συσχετίσεις των δεδομένων είναι πολύ περίπλοκες για να παρατηρηθούν από έναν άνθρωπο, η Μηχανική Μάθηση παρέχει μία εναλλακτική λύση. Αντί να βρεθεί η συνάρτηση και να χρησιμοποιηθεί σε άλλα δεδομένα, η μηχανή «μαθαίνει» από τα δεδομένα τις διάφορες σχέσεις και μπορεί να χρησιμοποιηθεί ως «μαύρο κουτί» στη θέση της άγνωστης συνάρτησης. Μερικές φορές, αυτό σημαίνει μειωμένη ακρίβεια σε σχέση με μία ρητή συνάρτηση, αλλά υπάρχουν εφαρμογές που αυτή είναι αποδεκτή θυσία. Υπάρχουν ακόμη εφαρμογές που ούτε μία μαθηματική συνάρτηση δε θα περιέγραφε με ακρίβεια όλα τα δεδομένα, ούτως ή άλλως. Πίσω από το μαύρο κουτί, δε βρίσκεται τίποτα άλλο παρά μία δομή δεδομένων που βασίζεται σε μαθηματικές πράξεις, για να δημιουργήσει ή να προβλέψει την επιθυμητή πληροφορία. (Chollet, 2021)

Ένα από τα βασικά εργαλεία της Μηχανικής Μάθησης είναι το Νευρωνικό Δίκτυο (Neural Network). Πρόκειται για μία δομή δεδομένων που αποθηκεύει πληροφορία σε ένα επίπεδο, την επεξεργάζεται μαθηματικά και την περνάει στο επόμενο επίπεδό της, επαναληπτικά. Υπάρχει πληθώρα νευρωνικών δικτύων, με κάθε κατηγορία τους να μπορεί να πάρει ποικιλία μορφών. Η είσοδος ενός νευρωνικού δικτύου μπορεί να είναι ένας αριθμός ή μία εικόνα αποθηκευμένη ως πίνακας αριθμών ή κάποια άλλη δομή δεδομένων. Κάθε επίπεδο (layer) από το οποίο περνάνε τα δεδομένα μπορεί να απεικονιστεί ως μία στρώση που αποτελείται από νευρώνες (neurons), κόμβους από αποθηκεύουν πληροφορία σαν μεταβλητές. Συνήθως, η πράξη που γίνεται στη σύνδεση ενός νευρώνα με τον νευρώνα της επόμενης στρώσης είναι πολλαπλασιασμός. Κάθε νευρώνας, λοιπόν, μπορεί να θεωρηθεί το αποτέλεσμα μίας συνάρτησης γινομένου της τιμής ενός προηγούμενου νευρώνα μία μία τιμή βάρους (weight), που επίσης αποθηκεύεται στο δίκτυο και μεταβάλλεται. Οι νευρώνες μπορούν να είναι συνδεδεμένοι αραιά ή πυκνά μεταξύ τους. Όταν ένας νευρώνας είναι συνδεδεμένος με πολλούς από το προηγούμενο επίπεδο, όπως συχνά γίνεται, η συνάρτηση της τιμής του είναι το άθροισμα όλων των επί μέρους γινομένων. Ωστόσο, γίνεται να εφαρμοστεί και μία επιπρόσθετη πράξη, με μία συνάρτηση ενεργοποίησης (activation function) ανάμεσα στα επίπεδα. Κάθε νευρώνας μπορεί να επικοινωνήσει άμεσα μόνο με νευρώνες των δύο γειτονικών επιπέδων. Το τελευταίο επίπεδο παράγει την έξοδο του δικτύου, έναν αριθμό ή ένα άλλο σύνολο δεδομένων που περνώντας από μία συνάρτηση ενεργοποίησης παίρνει πιο εύχρηστη μορφή. (Keim, 2019)

Το πρώτο επίπεδο του δικτύου έχει έναν νευρώνα για κάθε είσοδο και αντίστοιχα το τελευταίο επίπεδο έχει πλήθος νευρώνων ίσο με το πλήθος εξόδων. Τα ενδιάμεσα επίπεδα ονομάζονται κρυφά (hidden layers). Όταν ένα νευρωνικό δίκτυο έχει πάνω από ένα hidden layer καλείται βαθύ (deep neural network). Ο όρος Βαθιά Μάθηση (Deep Learning) αφορά το υποσύνολο της Μηχανικής Μάθησης που παράγει αρκετά περίπλοκα μοντέλα για την επίλυση περίπλοκων προβλημάτων. Όσο περισσότερα επίπεδα έχει ένα νευρωνικό δίκτυο, τόσο πιο βαθύ είναι. Με την αύξηση του βάθους, αυξάνεται και η απαιτούμενη μνήμη και επεξεργαστική ισχύς για τη διάσχιση του δικτύου. Η επαναληπτική διάσχιση του νευρωνικού δικτύου και μεταβολή των τιμών του λέγεται εκπαίδευση (training). Με το κατάλληλο σύνολο δεδομένων και επαρκή και κατάλληλη εκπαίδευση, το δίκτυο ίσως μπορεί να παράξει την επιθυμητή έξοδο για οποιαδήποτε είσοδο με κάποιο βαθμό ακρίβειας. Ο βασικότερος αλγόριθμος εκπαίδευσης είναι η Οπισθοδιάδοση (Backpropagation). (Rojas, 1996)

2.3.β Είδη νευρωνικών δικτύων

Οι βασικότερες κατηγορίες νευρωνικών δικτύων μπορούν να δώσουν μία αίσθηση της λειτουργικότητας της Μηχανικής Μάθησης.

- Το *αντίληπτρο (perceptron)* είναι η πιο απλή μορφή νευρωνικού δικτύου, αφού αποτελείται από έναν μόνο νευρώνα.
- Το *εμπροσθοτροφοδοτούμενο (feedforward)* είναι η βάση της Μηχανικής Όρασης. Ονομάζεται και αντίληπτρο πολλαπλών επιπέδων (multi-layer perceptron – MLP), αν και οι νευρώνες του δεν είναι perceptrons. Μπορεί να έχει ένα ή περισσότερα hidden layers.
- Το *αναδρομικό νευρωνικό δίκτυο (recurrent neural network – RNN)* διαφέρει από το feedforward λόγω της ύπαρξης βρόχων στους νευρώνες.
- Το *συνελικτικό νευρωνικό δίκτυο (convolutional neural network – CNN)* είναι μία πιο πολύπλοκη, βαθιά μορφή δικτύου που ειδικεύεται στο να ανιχνεύει μοτίβα σε εικόνες.

(IBM, 2020)

2.3.γ Συνελικτικά νευρωνικά δίκτυα

Στα μαθηματικά, *συνέλιξη* είναι η πράξη μεταξύ δύο συναρτήσεων που ορίζεται από την εξίσωση (1).

$$h(x) = \int_{-\infty}^{+\infty} f(x-y)g(y)\,dy = \int_{-\infty}^{+\infty} f(y)g(x-y)\,dy;$$

Εξίσωση (1)

όπου η συνάρτηση h είναι το αποτέλεσμα της συνέλιξης των συναρτήσεων f και g. Η συνέλιξη συμβολίζεται ως $f*g$. (Encyclopedia of Mathematics, 2022)

Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks ή CNNs) έχουν δει ευρεία χρήση στη Μηχανική Όραση τα τελευταία χρόνια, όντας συνηθισμένη επιλογή για οποιουδήποτε είδους ανάλυση εικόνας με AI. Μπορούν να λειτουργήσουν αποτελεσματικά τόσο με φωτογραφίες όσο και συνθετικές εικόνες. Η δομή τους διαφέρει από τα είδη νευρωνικών δικτύων που έχουν προαναφερθεί, καθώς περιλαμβάνουν παραπάνω από ένα

είδος επιπέδων. Επίσης, το πρώτο μέρος τους περιλαμβάνει πρωτότυπα επίπεδα για την ανίχνευση συγκεκριμένων χαρακτηριστικών, αλλά στη συνέχεια η ροή της πληροφορίας γίνεται σε απλούστερα επίπεδα, όμοια με των κλασσικών νευρωνικών δικτύων (σαν το feedforward). Στη συνέχεια, θα αναλυθεί ο τρόπος λειτουργίας των CNNs, καθώς ένα CNN είναι και ο πυρήνας του προτεινόμενου μοντέλου.
Ένα CNN δέχεται ως είσοδο μία εικόνα και αφού εφαρμόσει διάφορα *φίλτρα (filters)* πάνω της, εξάγει πληροφορία η οποία στη συνέχεια χρησιμοποιείται εσωτερικά με παρόμοιο τρόπο με ένα απλούστερο νευρωνικό δίκτυο. Η έξοδος του CNN μπορεί να είναι ένας ή περισσότεροι αριθμοί ή ακόμη και μία καινούργια εικόνα. (Wang et al., 2020)

Τα πρωτότυπα είδη επιπέδων τους είναι:

- Το *επίπεδο συνέλιξης (convolutional layer)*, όπου συμβαίνει η πράξη της συνέλιξης μεταξύ εικόνας και φίλτρου.
- Το *επίπεδο συγκέντρωσης* ή *υποδειγματοληψίας (pooling layer)*, όπου το προϊόν της συνέλιξης αποκτά μειωμένες διαστάσεις.

Αυτά τα επίπεδα μπορούν να συνδυαστούν και να επαναληφθούν με αρκετή ελευθερία, ανάλογα με την εφαρμογή. Πολλαπλά convolutional layers μπορούν να προστεθούν στη δομή πριν το πρώτο pooling. Στη συνέχεια το δίκτυο επεκτείνεται με ένα ή περισσότερα *πυκνά* ή *πλήρως συνδεδεμένα επίπεδα* (*dense* ή *fully connected layers*).

Στο convolutional layer δημιουργείται ένα ή περισσότερα φίλτρα τα οποία διασχίζουν την εικόνα και δημιουργούν έναν καινούργιο πίνακα με το αποτέλεσμα της συνέλιξης. Η εικόνα μπορεί να είναι δισδιάστατη ή τρισδιάστατη, ανάλογα με το χρώμα. Οι έγχρωμες εικόνες αποθηκεύονται ως τρισδιάστατοι πίνακες με την τρίτη διάσταση να αντιστοιχεί στα κανάλια χρώματος (για παράδειγμα RGB). Για μία εικόνα I με διαστάσεις n_H (ύψος), n_W (πλάτος) και n_C (βάθος) και ένα τετράγωνο φίλτρο K με μήκος f, η συνέλιξη ορίζεται όπως στην εξίσωση (2).

$$\text{conv}\,(I,\ K)_{x,y} = \sum_{i=1}^{n_H} \sum_{j=1}^{n_W} \sum_{k=1}^{n_C} K_{i,j,k} I_{x+i-1,y+j-1,k}$$

Εξίσωση (2)

Μπορεί να οριστεί πιο συγκεκριμένα ο τρόπος που το φίλτρο διασχίζει την εικόνα. Επειδή τα εικονοστοιχεία (pixels) στις άκρες τις εικόνας δε θα υπολογιστούν με τον επιθυμητό τρόπο, προαιρετικά προστίθεται περιθώριο (padding) μερικών pixel στην εικόνα. Συνήθως το padding γεμίζει με μηδενικές τιμές. Επιπλέον, το μέγεθος του παραχθέντος πίνακα εξαρτάται από το μέγεθος του βήματος (stride) που γίνεται κατά τη διάσχιση. Συμβολίζοντας το padding με p και το stride με s, οι διαστάσεις του πίνακα της συνέλιξης δίνονται από τη σχέση (3).

$$\dim(\text{conv}(I, K)) = \left(\left\lfloor \frac{n_H + 2p - f}{s} + 1 \right\rfloor, \left\lfloor \frac{n_W + 2p - f}{s} + 1 \right\rfloor\right); \; s > 0$$
$$= (n_H + 2p - f, \; n_W + 2p - f); \; s = 0$$

Εξίσωση (3)

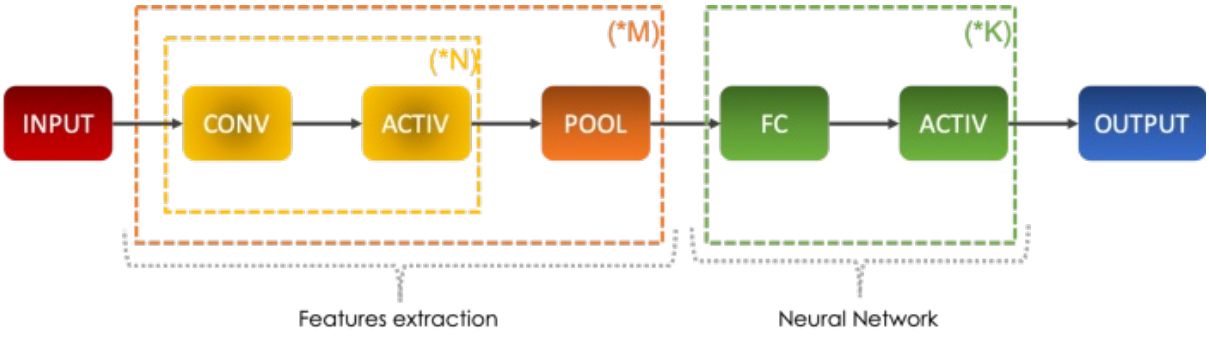
Αυτός ο πίνακας είναι και η έξοδος του επιπέδου, που προωθείται στο επόμενο επίπεδο. (Mebsout, 2020)

Μετά από ένα ή περισσότερα convolutional layers, ακολουθεί το pooling layer. Το pooling μειώνει την κλίμακα της εικόνας, αλλά δεν επηρεάζει το βάθος της, δηλαδή το πλήθος των καναλιών μένει ίδιο για τρισδιάστατες εικόνες. Η μείωση του ύψους και πλάτους γίνεται με ένα τετράγωνο φίλτρο με μήκος f. Συνηθίζεται το f να είναι ίσο με 2. Τα στοιχεία του πίνακα που βρίσκονται μέσα στο φίλτρο ανά επανάληψη συνθέτουν ένα νέο στοιχείο για το προϊόν του pooling, που είναι ένας μικρότερος πίνακας. Ο υπολογισμός του νέου στοιχείου μπορεί να γίνει με διάφορους τρόπους, όπως με μέσο όρο των στοιχείων ή με εύρεση της μέγιστης τιμής. Αν χρησιμοποιηθεί η μέγιστη τιμή, το επίπεδο μπορεί να ονομαστεί και *max pooling*. Οι διαστάσεις του νέου πίνακα υπολογίζονται όπως στην εξίσωση (4).

$$\dim(\text{pooling}(\text{image})) = \left(\left\lfloor \frac{n_H + 2p - f}{s} + 1 \right\rfloor, \left\lfloor \frac{n_W + 2p - f}{s} + 1 \right\rfloor, n_C\right); \; s > 0$$
$$= (n_H + 2p - f, n_W + 2p - f, n_C); \; s = 0$$

Εξίσωση (4)

Το ζεύγος convolution και pooling μπορεί να επαναληφθεί αρκετές φορές για την εξαγωγή χαρακτηριστικών από την αρχική εικόνα. Στα πρώτα στάδια αναγνωρίζονται απλά μοτίβα και μετέπειτα πιο σύνθετα. Τελικώς, τα χαρακτηριστικά θα χρησιμοποιηθούν από ένα ή περισσότερα dense / fully connected layers ως είσοδος. Σε αυτό το μέρος του δικτύου αναμένεται συμπεριφορά παρόμοια με των απλούστερων κατηγοριών νευρωνικών δικτύων. Τα βάρη των επιπέδων μεταβάλλονται κατά την εκπαίδευση, με σκοπό να παράγουν την κατάλληλη έξοδο ή εξόδους του συνολικού νευρωνικού δικτύου. Εκτός από τα βάρη, συχνά υπάρχει και άλλο ένα μέγεθος που προστίθεται στο γινόμενο που υπολογίζεται για κάθε νευρώνα (βλέπε 2.3.α). Αυτό το μέγεθος είναι η *πόλωση (bias)* και συνεισφέρει στη σωστή ενεργοποίηση των νευρώνων. Ολόκληρη η δομή του CNN, μαζί με τις συναρτήσεις ενεργοποίησης που προσαρμόζουν την έξοδο του ενός επιπέδου πριν περάσει στο επόμενο, σχηματίζεται όπως φαίνεται στην Εικόνα 7.



Εικόνα 7 – Δομή CNN

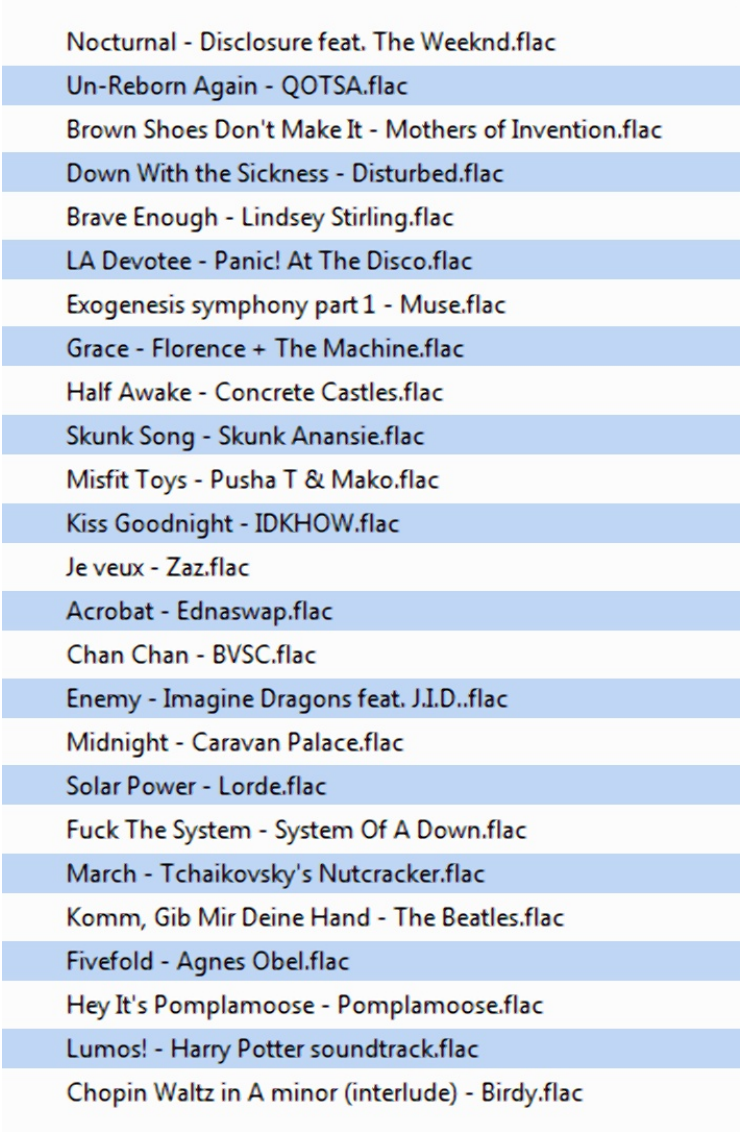
Οι μεταβλητές N, M και K συμβολίζουν το πλήθος κάθε τμήματος στο δίκτυο. (Mebsout, 2020)

ΚΕΦΑΛΑΙΟ 3 Σύνολο δεδομένων και Αρχιτεκτονική

3.1 Σύνολο δεδομένων

Όπως συμβαίνει συχνά στην Τεχνητή Νοημοσύνη, η μορφή του συνόλου δεδομένων (dataset) επηρεάζει πολύ την ακρίβεια του νευρωνικού δικτύου και την έκταση του προβλήματος που θα είναι ικανό να αντιμετωπίσει. Το μέγεθος του dataset, η ποικιλομορφία των δειγμάτων και η αναλογία truly lossless και transcoded flac έχουν σημασία για την εκπαίδευση και επαλήθευση του μοντέλου. Η σωστή προεπεξεργασία των συνόλου μπορεί να βελτιώσει τα αποτελέσματα σε κάποιες περιπτώσεις. Ωστόσο, τεχνικοί και πρακτικοί περιορισμοί μπορούν να αναγκάσουν μειωμένους πειραματισμούς. Ένα τέτοιο παράδειγμα είναι η ανάλυση των εικόνων που παράγονται από τα μουσικά κομμάτια, αφού μεγαλύτερη ανάλυση θα απαιτούσε περισσότερη επεξεργαστική ισχύ και μνήμη. Επίσης, η παροχή truly lossless αρχείων από διαφορετικά μουσικά είδη δεν είναι εγγυημένη. Παρά τους περιορισμούς που πρέπει να ληφθούν υπ' όψη, το dataset που χρησιμοποιήθηκε παρουσίασε ικανοποιητικά αποτελέσματα.

Το dataset της εκπαίδευσης βασίζεται σε 65 τραγούδια από διάφορους καλλιτέχνες που ανήκουν σε ποικίλα είδη. Σε αυτά συμπεριλαμβάνονται ροκ, ποπ, μέταλ, ηλεκτρονική μουσική, κλασική μουσική και ραπ. Κάθε spectrogram αυτών των τραγουδιών μελετήθηκε για την επαλήθευση της γνησιότητάς του και συμπληρωματικά το πρόγραμμα Audiochecker χρησιμοποιήθηκε (Opris, 2013). Το συνολικό μέγεθός τους σε ανάλυση 44100 Hz και 16 bit depth είναι 1,52 GB. Η συνολική διάρκειά τους είναι 3:59:08 (3 ώρες, 59 λεπτά και 8 δευτερόλεπτα). Θεωρώντας την αρχική μορφή αυτών των αρχείων ως truly lossless flac, παράχθηκαν 3 διαφορετικές εκδοχές transcoded flac για κάθε τραγούδι. Οι δύο δημιουργήθηκαν με τη βιβλιοθήκη FFmpeg, μία εκδοχή από transcoding σε mp3 με bitrate 128 kbps και μία για bitrate 320 kbps. Η τρίτη εκδοχή transcoded flac ήταν και αυτή από mp3 με bitrate 320 kbps, αλλά παράχθηκε μέσω του προγράμματος Audacity. Η ανάγκη για διαφορετικά bitrate έγκειται στο ποιες συχνότητες αποκόπτονται, καθώς σε ψηλές αναλύσεις οι διαφορές από το γνήσιο αρχείο είναι μικρότερες. Η ύπαρξη δύο διαφορετικών εκδοχών για 320 kbps επεκτείνει τη λειτουργικότητα του μοντέλου για διαφορετικά αποτελέσματα transcoding. Παρατηρήθηκε ότι τα transcoded flac που παράχθηκαν μέσω του Audacity έχουν περισσότερη πληροφορία στις υψηλές συχνότητες και πιθανό θόρυβο (dither) στο φάσμα της αποκοπής.

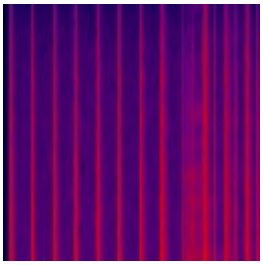


Εικόνα 8 – Τμήμα του dataset εκπαίδευσης (ονόματα αρχείων)

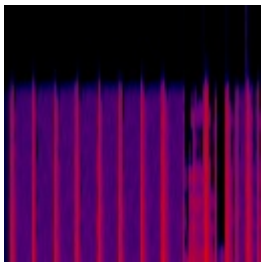
Το νευρωνικό δίκτυο δεν αποδίδει ικανοποιητικά για datasets με πολύ μεγαλύτερο πλήθος transcoded flac δειγμάτων από ότι truly lossless flac. Εφόσον τα transcoded flac αρχεία που αναφέρθηκαν είναι τα τριπλάσια σε πλήθος από τα αρχικά, μόνο κάποια από αυτά χρησιμοποιήθηκαν. Όλες οι 3 κατηγορίες transcoded flac αντιπροσωπεύτηκαν στο dataset της εκπαίδευσης. Συνολικά, χρησιμοποιήθηκαν 65 αρχεία ως truly lossless και 135 αρχεία ως transcoded.

Κάθε τραγούδι που δίνεται στο σύστημα περνάει από προεπεξεργασία πριν χρησιμοποιηθεί ως είσοδος του νευρωνικού δικτύου. Το ίδιο ισχύει και για τα τραγούδια που χρησιμοποιήθηκαν στην εκπαίδευση. Κάθε αρχείο διαιρείται σε τμήματα των 8 δευτερολέπτων και μία τετράγωνη εικόνα του spectrogram παράγεται από κάθε τμήμα. Το spectrogram περιέχει μόνο το φάσμα συχνοτήτων από 16200 Hz μέχρι 22000 Hz για κάθε τμήμα. Ακόμα και για κομμάτια μεγαλύτερης ανάλυσης στα οποία υπάρχει πληροφορία πάνω από τα 22000 Hz, ισχύει η ίδια επιλογή. Οι εικόνες είναι έγχρωμες σε μορφή RGB (Red, Green, Blue) και έχουν διαστάσεις 128 × 128 pixels. Αποθηκεύονται προσωρινά σε πίνακες με διαστάσεις 128 × 128 × 3. Το πλήθος γραφημάτων ανά μουσικό κομμάτι

εξαρτάται από τη διάρκειά του. Τα τμήματα έχουν επικάλυψη (overlap) 50% με τα χρονικά γειτονικά τους. Για παράδειγμα, το πρώτο δείγμα διαρκεί από την αρχή ενός αρχείου μέχρι το 8ο δευτερόλεπτο, το δεύτερο δείγμα διαρκεί από το 4ο μέχρι το 12ο δευτερόλεπτο κ.ο.κ. Έτσι επιτυγχάνεται μεγαλύτερο πλήθος δειγμάτων. Το τελικό dataset που χρησιμοποιήθηκε στην εκπαίδευση του CNN, δημιουργημένο από τα αρχεία που προαναφέρθηκαν, αποτελούνταν από 10.729 δείγματα. Κάθε δείγμα αντιστοιχεί σε μία εικόνα spectrogram. Από αυτά τα δείγματα, τα 10.086 (περίπου το 94%) εκπαίδευσαν το δίκτυο και τα υπόλοιπα 643 χρησίμευσαν για την άμεση επαλήθευση ανά δείγμα (spectrogram) κατά τη διάρκεια της διαδικασίας. Τα δείγματα ανακατεύτηκαν με τυχαία σειρά πριν την εκπαίδευση και δεν αποθηκεύτηκε κάποια πληροφορία που να συσχετίζει κάθε εικόνα με το αρχείο ήχου από το οποίο παράχθηκε. Δεν υπήρξε καμία κανονικοποίηση των αριθμητικών τιμών. Για την εκπαίδευση, τα δείγματα συγκεντρώθηκαν σε *δεσμίδες* ή *παρτίδες (batches)* των 16 δειγμάτων.



Εικόνα 9 – Truly lossless flac



Εικόνα 10 – Transcoded flac (320 kbps)

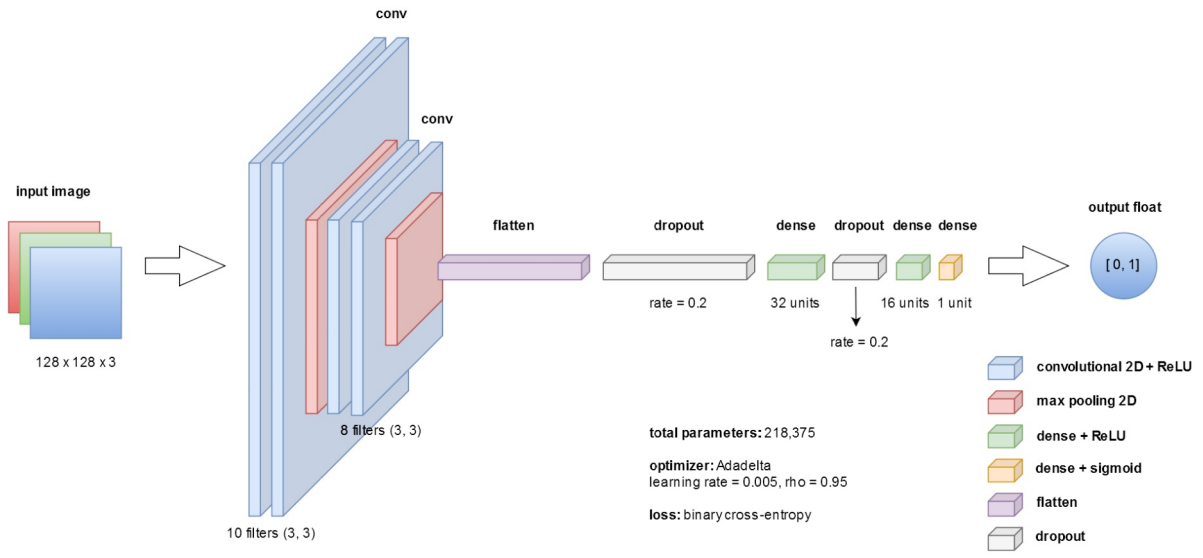
Δύο δείγματα του dataset εκπαίδευσης (spectrograms)
Απεικονίζονται τα πρώτα 8 δευτερόλεπτα του τραγουδιού *Back in Black* των *AC/DC* σε truly lossless και transcoded μορφές για το εύρος από 16200 Hz μέχρι 22000 Hz.

Για τη χειροκίνητη επαλήθευση του μοντέλου μετά το τέλος της εκπαίδευσης χρησιμοποιήθηκαν τραγούδια που δεν υπήρχαν στο αρχικό dataset. Ακολούθησε η ίδια διαδικασία ελέγχου της γνησιότητάς τους και παραγωγής διαφορετικών ποιοτήτων transcoded αρχείων. Η ποικιλομορφία των μουσικών ειδών παραμένει σημαντική και για το στάδιο της επαλήθευσης. Αρχικά, 10 τραγούδια και οι παραποιημένες εκδοχές τους (40 αρχεία συνολικά) χρησιμοποιήθηκαν για τα διάφορα πειράματα, δηλαδή ως μέτρο σύγκρισης της ακρίβειας του μοντέλου μεταξύ πειραματικών εκπαιδεύσεων. Για την τελική αξιολόγηση της επικρατούσας εκπαίδευσης χρησιμοποιήθηκε διαφορετικό dataset επαλήθευσης, με σκοπό τη διασφάλιση της γενικευμένης επίλυσης του προβλήματος. Αυτό το dataset επαλήθευσης αποτελείται από 23 γνήσια αρχεία και τις παραποιημένες εκδοχές τους σε διάφορες ποιότητες, ακολουθώντας τους ίδιους κανόνες με προηγουμένως. Κανένα από αυτά δεν υπήρχε στο dataset εκπαίδευσης ή στα ενδιάμεσα dataset επαλήθευσης. Από τα 92 αρχεία που προκύπτουν, χρησιμοποιήθηκαν τα 46 για τον υπολογισμό της ακρίβειας, 23 truly lossless flac και 23 transcoded flac διάφορων ποιοτήτων, ώστε η ακρίβεια να αντικατοπτρίζεται σε 50% truly lossless και 50% transcoded αρχεία. Αυτό το τελικό στάδιο επαλήθευσης έγινε ανά αρχείο και όχι ανά δείγμα.

3.2 Αρχιτεκτονική νευρωνικού δικτύου

Το προτεινόμενο σύστημα αποτελείται από δύο μέρη. Στο πρώτο μέρος, δημιουργούνται τα κατάλληλα spectrograms από αρχεία ήχου. Το δεύτερο μέρος είναι το CNN που αναλύει τα spectrograms. Επομένως, η είσοδος του συστήματος είναι αρχεία flac ή wav, τα οποία το σύστημα διαχειρίζεται για την παραγωγή των εικόνων. Η είσοδος του νευρωνικού δικτύου είναι οι εικόνες που περιγράφηκαν παραπάνω, σε μορφή πινάκων.

Το προτεινόμενο νευρωνικό δίκτυο είναι ένα συμβατικό CNN, επαρκώς βαθύ για την ανάλυση τετράγωνων έγχρωμων εικόνων των 128 pixels. Υπάρχουν δύο φάσεις συνέλιξης για την εξαγωγή χαρακτηριστικών, με δύο convolutional layers στην καθεμία. Στο τέλος κάθε φάσης γίνεται max pooling για σμίκρυνση του πίνακα εξόδου. Μετά από τη συνέλιξη ακολουθεί μία φάση απλούστερης δομής νευρωνικού δικτύου με πυκνά συνδεδεμένα layers. Το μέγεθος των layers φθίνει όπως η ροή των δεδομένων κατευθύνεται προς την έξοδο, η οποία είναι μόνο ένας πραγματικός αριθμός. Ενδιάμεσα στα dense layers χρησιμοποιούνται και μετασχηματισμοί dropout για τη μείωση του φαινόμενου υπερπροσαρμογής (overfitting). Το dropout μετατρέπει κάποια στοιχεία σε 0, επιλέγοντάς τα τυχαία, εδώ με ρυθμό (rate) 0,2. Με αυτόν τον τρόπο επιτυγχάνεται μικρή διαφορά ακρίβειας ανάμεσα στα σύνολα εκπαίδευσης και επαλήθευσης. Ακολουθεί διάγραμμα με την ακριβή δομή του CNN.



Εικόνα 11 – Το προτεινόμενο CNN

Το επίπεδο flatten απλώνει τη δισδιάστατη έξοδο του τελευταίου pooling σε μονοδιάστατη, ώστε να μπορεί να χρησιμοποιηθεί στα dense layers.

Η στοχαστική τεχνική βελτιστοποίησης Adadelta χρησιμοποιείται για την εκπαίδευση, με ρυθμό μάθησης (learning rate) ίσο με 0,005 και ρυθμό μείωσης της μάθησης (rho) ίσο με 0,95. Συνολικά για όλο το δίκτυο, δημιουργούνται 218.375 παράμετροι που μπορούν να μεταβληθούν με την εκπαίδευση. Η εκπαίδευση του μοντέλου εκτελείται για 50 εποχές. Σε κάθε εποχή χρησιμοποιείται ολόκληρο το dataset εκπαίδευσης, διαιρεμένο σε batches των 16 δειγμάτων.

Στην πρώτη φάση συνέλιξης, τα δύο convolutional layers εφαρμόζουν από 10 φίλτρα το καθένα. Στη δεύτερη φάση, τα φίλτρα είναι 8 για κάθε convolutional layer. Και για τα τέσσερα επίπεδα, όλα τα φίλτρα έχουν διαστάσεις 3 × 3 pixels. Για τα pooling layers, το παράθυρο σμίκρυνσης έχει διαστάσεις 2 × 2 pixels.

Στο στάδιο κλασικού νευρωνικού δικτύου που ακολουθεί χρησιμοποιούνται τα χαρακτηριστικά που έχουν εξαχθεί και ανιχνεύονται οι συσχετίσεις τους με τις πιθανές εξόδους του δικτύου. Υπάρχουν 3 dense layers. Το πρώτο layer έχει 32 νευρώνες, το δεύτερο έχει 16 νευρώνες και το τελευταίο έχει 1 νευρώνα, για να ταιριάζει με το πλήθος των εξόδων του δικτύου.

Κάθε layer, είτε convolutional, είτε dense, περιέχει bias καθώς και activation function πριν την έξοδο. Χρησιμοποιούνται δύο activation functions, η *Συνάρτηση Ανορθωμένης Γραμμικής Μονάδας (Rectified Linear Unit – ReLU)* και η *Σιγμοειδής Συνάρτηση (Sigmoid Function)*.

- Η ReLU είναι μία συνάρτηση που μηδενίζει τις αρνητικές τιμές εισόδου της, χωρίς να επηρεάζει τις θετικές τιμές (Becker, 2018).
- Η sigmoid μεταφράζει την είσοδό της σε ένα συγκεκριμένο μικρό διάστημα, ως μηχανισμός κανονικοποίησης ή μετατροπής σε μορφή πιθανότητας (Saeed, 2021).

Όλα τα convolutional layers εφαρμόζουν ReLU στο αποτέλεσμά τους πριν το μεταφέρουν στο επόμενο επίπεδο, όπως και τα πρώτα δύο dense layers. Με αυτόν τον τρόπο, πιθανές αρνητικές τιμές που θα υπονόμευαν την ακρίβεια του δικτύου παύουν να αποτελούν πρόβλημα. Το τρίτο dense layer που παράγει την έξοδο του δικτύου εφαρμόζει συνάρτηση sigmoid, ώστε η έξοδος να βρίσκεται πάντα στο κλειστό διάστημα [0, 1].

Παρακάτω αναλύεται το μέγεθος εξόδου κάθε στρώσης, από την είσοδο του δικτύου προς την έξοδο.

- Πρώτο convolutional layer: 126 × 126 × 10
- Δεύτερο convolutional layer: 124 × 124 × 10
- Πρώτο pooling layer: 62 × 62 × 10
- Τρίτο convolutional layer: 60 × 60 × 8
- Τέταρτο convolutional layer: 58 × 58 × 8
- Δεύτερο pooling layer: 29 × 29 × 8
- Flatten layer: 6.728
- Πρώτο dropout layer: 6.728
- Πρώτο dense layer: 32
- Δεύτερο dropout layer: 32
- Δεύτερο dense layer: 16
- Τρίτο dense layer: 1

Η τελευταία αριθμητική τιμή των γινομένων των convolutional και pooling layers αντιστοιχεί στο πλήθος φίλτρων, αφού κάθε φίλτρο παράγει έναν καινούργιο πίνακα. Οι πίνακες είναι δισδιάστατοι αρχικά και γίνονται μονοδιάστατοι με την εφαρμογή του flatten.

3.3 Χειρισμός εξόδου

Η έξοδος του νευρωνικού δικτύου για κάθε δείγμα είναι ένας πραγματικός αριθμός από 0 μέχρι 1. Όταν είναι ίση με 0, το δείγμα κρίνεται ως mp3 transcoded και όταν είναι ίση με 1 κρίνεται ως truly lossless. Οι ενδιάμεσες τιμές χρειάζονται έναν κανόνα για να ερμηνευτούν. Για αυτόν τον λόγο, χρησιμοποιείται μία τιμή κατωφλιού (threshold) για τον δυαδικό διαχωρισμό του διαστήματος [0, 1]. Το κατώφλι είναι ίσο με 0,6. Όταν η έξοδος είναι μεγαλύτερη από 0,6 ερμηνεύεται ως truly lossless και όταν είναι μικρότερη ή ίση με 0,6 ερμηνεύεται ως transcoded.

Προηγουμένως αναφέρθηκε επαλήθευση ανά δείγμα και ανά αρχείο. Για την αξιολόγηση ενός αρχείου με το ήδη εκπαιδευμένο μοντέλο μπορούμε να χρησιμοποιήσουμε πολλαπλά δείγματα, αφού κάθε δείγμα αφορά μόνο 8 δευτερόλεπτα του κομματιού. Το νευρωνικό δίκτυο εκτελεί την αξιολόγηση για κάθε τέτοιο δείγμα ανεξάρτητα από το τραγούδι στο οποίο ανήκει. Αντί να εκτελεστεί η αξιολόγηση ενός δείγματος για να κριθεί ολόκληρο το αρχείο, αξιολογούνται όλα τα δείγματα και η τελική απάντηση δίνεται από τον συνδυασμό των πολλαπλών αξιολογήσεων. Έτσι, προκύπτει το πρόβλημα της εύρεσης βέλτιστου μηχανισμού για τον συνδυασμό των αξιολογήσεων όλων των δειγμάτων ενός αρχείου. Ακόμη και απλές προσεγγίσεις αυξάνουν την ακρίβεια του μοντέλου σε σύγκριση με την ακρίβεια για ένα μεμονωμένο δείγμα. Ούτως ή άλλως, η ακρίβεια ανά αρχείο περιγράφει την πραγματική χρήση του συστήματος.

Μία μέθοδος είναι ο υπολογισμός μέσου όρου των τιμών εξόδου για όλα τα δείγματα. Με παρόμοια ακρίβεια στο dataset επαλήθευσης, μπορεί να χρησιμοποιηθεί και ένα σύστημα ψήφων. Στο σύστημα ψήφων, η έξοδος κάθε δείγματος έχει μία ψήφο για την αξιολόγηση του αρχείου. Η ψήφος είναι είτε υπέρ truly lossless, είτε υπέρ transcoded και όποια κατηγορία έχει περισσότερες ψήφους καθορίζει την ποιότητα του κομματιού. Εναλλακτικά, θα μπορούσε να υπάρχει μία τιμή κατωφλιού για πλήθος ψήφων truly lossless και να απαιτείται συντριπτική πλειοψηφία για να οριστεί το αρχείο ως truly lossless.

ΚΕΦΑΛΑΙΟ 4 Αποτελέσματα

4.1 Μετρήσεις απόδοσης

Για την αντικειμενική αξιολόγηση του προτεινόμενου μοντέλου, θα χρησιμοποιηθούν διάφορες μετρικές για την απόδοση. Υπενθυμίζεται ότι το πρόβλημα της ταξινόμησης αρχείων σε truly lossless και transcoded από mp3 είναι δυαδικό, όπως και η έξοδος του μοντέλου. Για την αξιολόγηση χρησιμοποιείται το dataset επαλήθευσης 46 αρχείων flac που περιγράφηκε στο κεφάλαιο 3.1 και αποτελείται από 23 truly lossless αρχεία και 23 transcoded flac που παράχθηκαν από αυτά, σε διάφορες ποιότητες mp3 κωδικοποίησης. Δεν υπάρχουν κοινά δείγματα με το dataset εκπαίδευσης.

Αρχικά, τα αποτελέσματα των προβλέψεων παρουσιάζονται σε μορφή *πίνακα σύγχυσης (confusion matrix)*, ώστε να παρατίθεται το πλήθος των λανθασμένων απαντήσεων. Σε έναν πίνακα σύγχυσης, η μία διάσταση αντιστοιχεί στις πραγματικές τιμές των στοιχείων και η άλλη στις τιμές των προβλέψεων. Για δυαδικά προβλήματα, ο πίνακας έχει τέσσερις καταχωρήσεις προβλέψεων:

- Αληθώς Θετικές (True Positives)
- Ψευδώς Θετικές (False Positives)
- Αληθώς Αρνητικές (True Negatives)
- Ψευδώς Αρνητικές (False Negatives)

Θετική ορίζεται η έξοδος για truly lossless flac και αρνητική για transcoded flac.
(Fawcett, 2006)

Σύνολο: 46		Τιμές προβλέψεων	
		Θετικές (17)	**Αρνητικές** (29)
Πραγματικές τιμές	**Θετικές** (23)	17	6
	Αρνητικές (23)	0	23

Confusion matrix

Παρατηρούμε ότι υπάρχουν 6 αρχεία στην κατηγορία False Negative, δηλαδή truly lossless αρχεία που ψευδώς κρίθηκαν ως transcoded. Θα ερευνηθεί παρακάτω γιατί το μοντέλο είναι τόσο αυστηρό με κάποια αρχεία. Ωστόσο, δε βρέθηκε κανένα False Positive στοιχείο. Επομένως, όταν ένα αρχείο αξιολογηθεί ως truly lossless, το αποτέλεσμα είναι πιθανότατα σωστό.

Γνωρίζοντας τα πλήθη σωστών και λανθασμένων προβλέψεων, μπορούμε να υπολογίσουμε τα μέτρα απόδοσης. Διαφορετικά μέτρα προσφέρουν διαφορετικές οπτικές γωνίες για την επαλήθευση του μοντέλου.

- Η *ορθότητα (accuracy)* εκφράζεται με τον λόγο των σωστών προβλέψεων προς το σύνολο των προβλέψεων.
- Η *ακρίβεια (precision)* εκφράζεται με τον λόγο των αληθώς θετικών προβλέψεων προς το σύνολο των θετικών προβλέψεων.
- Η *ανάκληση (recall)* εκφράζεται με τον λόγο των αληθώς θετικών προβλέψεων προς το σύνολο των πραγματικών θετικών στοιχείων.
- Η *μετρική F_1 (F_1 score)* ισούται με το διπλάσιο γινόμενο της ακρίβειας επί την ανάκληση διαιρεμένο με το άθροισμα της ακρίβειας και της ανάκλησης (βλέπε εξίσωση (5)).

(Fawcett, 2006)

$$F_1\ score\ =\ 2 * \frac{precision * recall}{precision + recall}$$

Εξίσωση (5)

Με βάση τους προηγούμενους ορισμούς, η ορθότητα του μοντέλου υπολογίζεται ως εξής:

40 / 46 = 0,8696

Η ακρίβεια του μοντέλου θα είναι:

17 / 17 = 1

Παρομοίως, η ανάκληση θα είναι:

17 / 23 = 0,7391

Η ακρίβεια δείχνει πόσες από τις θετικές προβλέψεις είναι σωστές. Συμπληρωματικά, η ανάκληση δείχνει για πόσες από τις πραγματικά θετικές τιμές έγινε σωστή πρόβλεψη. Με αυτές τις δύο μετρικές υπολογίζεται το F_1:

2 * 1 * 0,7391 / (1 + 0,7391) = 0,85

Η μετρική F_1 συνδυάζει δύο σημαντικά μέτρα απόδοσης και περιγράφει το νευρωνικό δίκτυο πιο ολοκληρωμένα.

Συνοψίζοντας, η απόδοση του προτεινόμενου μοντέλου φαίνεται στον επόμενο πίνακα.

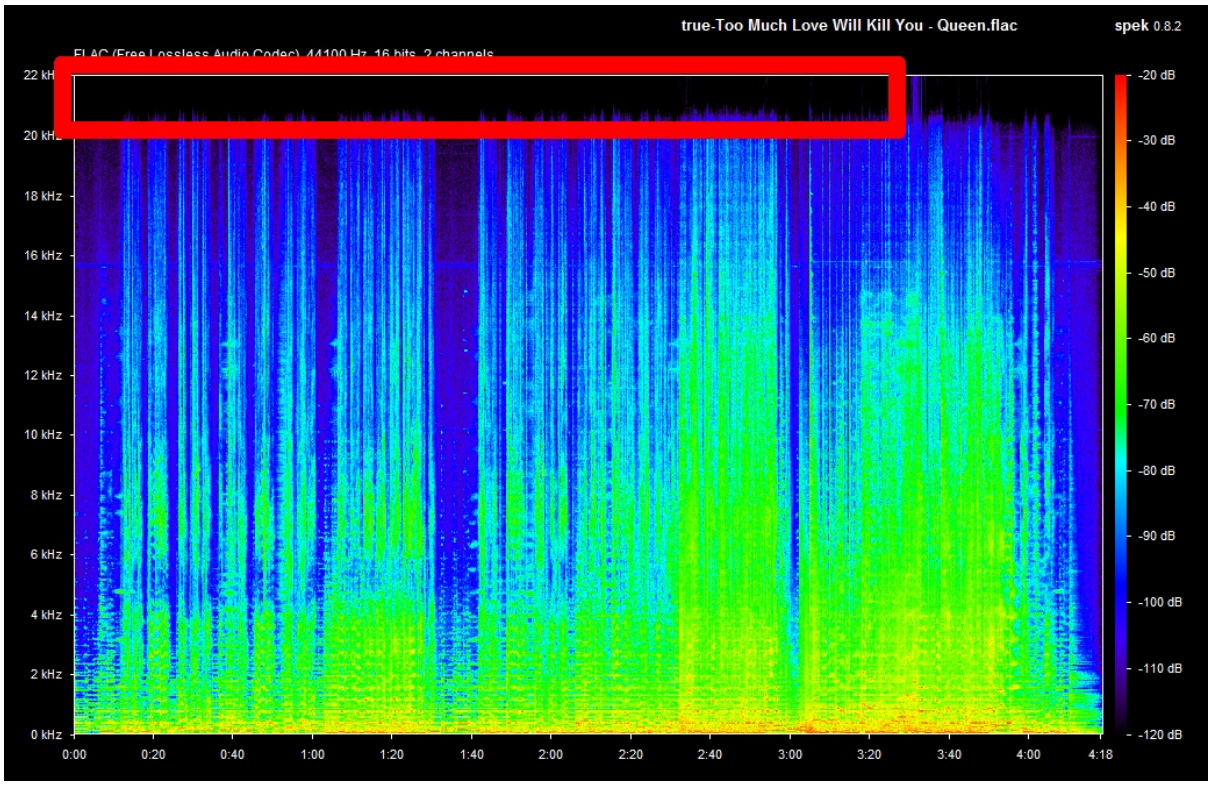
Accuracy	**Precision**	**Recall**	**F_1 Score**
86,96%	**100%**	**73,91%**	**85%**

Αποτελέσματα προτεινόμενου μοντέλου

4.2 Μελέτη λανθασμένων προβλέψεων

Η κατανόηση των λόγων που δημιούργησαν τις λανθασμένες προβλέψεις είναι σημαντική για να υπάρχει πλήρης αντίληψη της απόδοσης και πιθανή μελλοντική βελτίωση. Το CNN βασίζεται στην ύπαρξη πληροφορίας στις υψηλότερες συχνότητες του φάσματος, για να απορρίψει κωδικοποίηση mp3. Τραγούδια με μειωμένη ή καθόλου αποθηκευμένη πληροφορία στο εύρος συχνοτήτων που αναπαριστάται στα δείγματα spectrograms έχουν μειονέκτημα στην ακριβή αξιολόγησή τους. Σε αυτό το σενάριο εντάσσονται όλα τα 6 False Negatives αρχεία του dataset επαλήθευσης. Δεν παρατηρήθηκε κάποιο άλλο κοινό στοιχείο μεταξύ των αρχείων αυτών, καθώς ανήκουν σε διαφορετικά μουσικά είδη με εμφανή ποικιλομορφία μουσικών οργάνων, φωνητικών και ηχητικού σχεδιασμού.

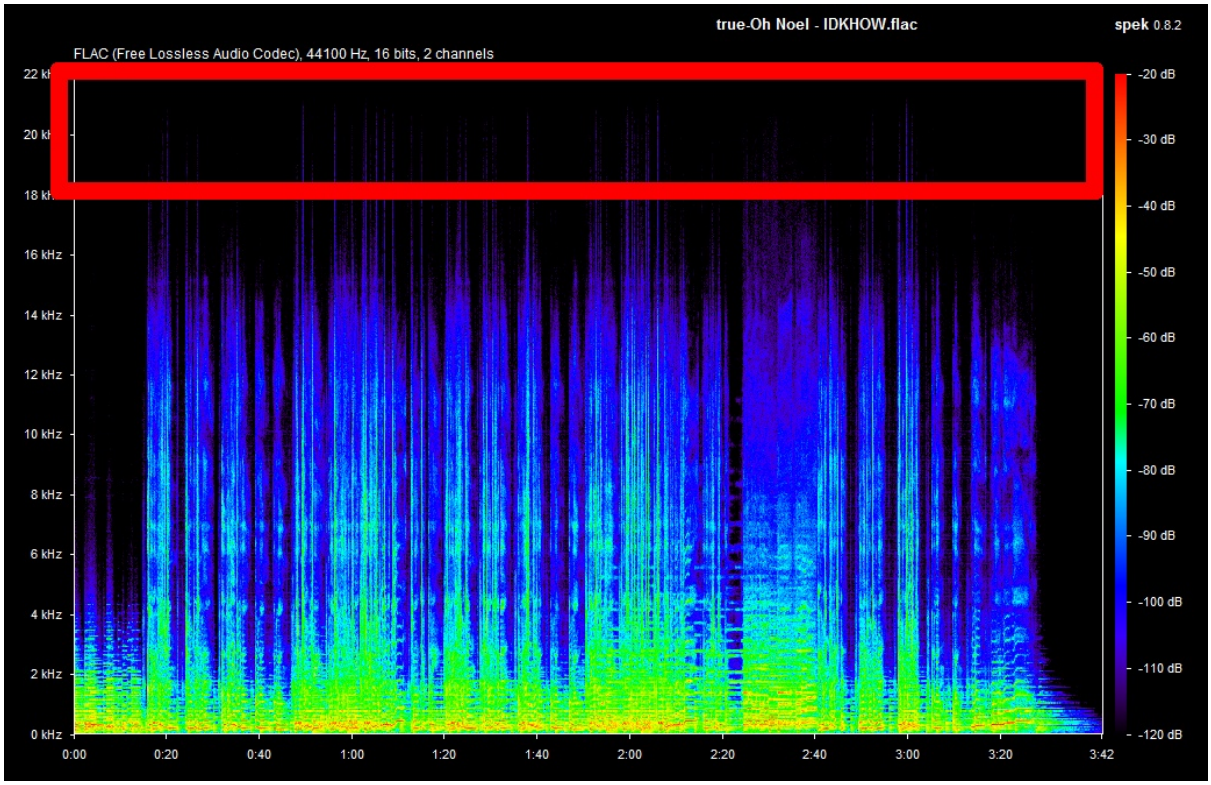
Στο παράδειγμα της εικόνας 12 απεικονίζεται ολόκληρο το spectrogram του τραγουδιού *Too much love will kill you* των *Queen* σε μορφή truly lossless flac με συχνότητα δειγματοληψίας 44100 Hz και βάθος bit 16. Αυτό το αρχείο αποτελεί ένα από τα 6 False Negatives. Με κόκκινο χρώμα υπογραμμίζεται το φάσμα συχνοτήτων όπου δεν υπάρχει ήχος με επαρκή ένταση για να αναγνωριστεί η κωδικοποίηση ως truly lossless. Παρά τη γνησιότητα του αρχείου, από τη φύση του τραγουδιού δεν υπάρχει η απαιτούμενη πληροφορία στο spectrogram για τον διαχωρισμό από μία πιθανή transcoded εκδοχή. Η υπογραμμισμένη περιοχή κυμαίνεται από πάνω από τα 20 kHz μέχρι τα 22 kHz.



Εικόνα 12 – Spectrogram του *Too much love will kill you* των *Queen* (truly lossless) Παράχθηκε με το πρόγραμμα spek.

Άλλο ένα παράδειγμα είναι το τραγούδι *Oh Noel* των *I Dont Know How But They Found Me* που ανήκει επίσης στα False Negative αρχεία. Στην εικόνα 13 παρουσιάζεται το πλήρες

spectrogram της truly lossless μορφής του με συχνότητα δειγματοληψίας 44100 Hz και βάθος bit 16. Η περιοχή που είναι υπογραμμισμένη με κόκκινο χρώμα δεν περιέχει ήχο μεγάλης έντασης και ως συνέπεια παραπλανά το CNN.



Εικόνα 13 – Spectrogram του *Oh Noel* των *IDKHOW* (truly lossless)
Παράχθηκε με το πρόγραμμα spek.

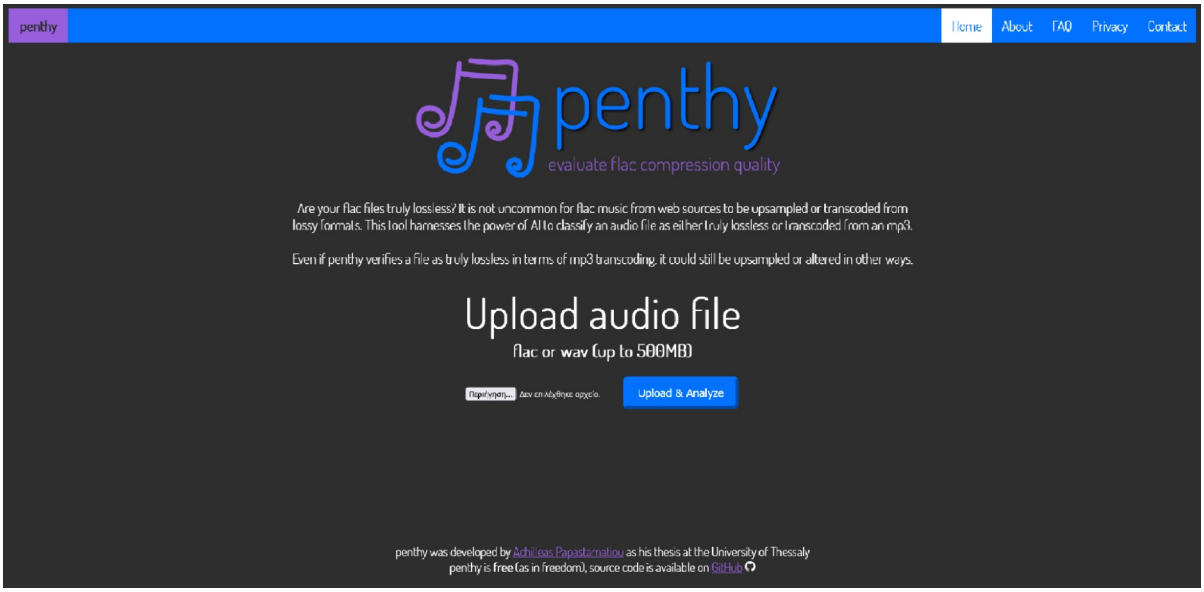
Τα spectrograms των υπόλοιπων τεσσάρων αρχείων που ορίστηκαν ως False Negative παρουσιάζουν την ίδια ιδιότητα. Συμπεραίνουμε ότι η αδυναμία του μοντέλου είναι γνήσια αρχεία που δεν περιέχουν πολύ υψηλές συχνότητες για σημαντικό κομμάτι της διάρκειάς τους. Η συμπεριφορά αυτή οφείλεται στα χαρακτηριστικά που εξάγονται στο στάδιο συνέλιξης του CNN.

ΚΕΦΑΛΑΙΟ 5 Συμπεράσματα

5.1 Η εφαρμογή *penthy*

5.1.α Λειτουργία της ιστοσελίδας

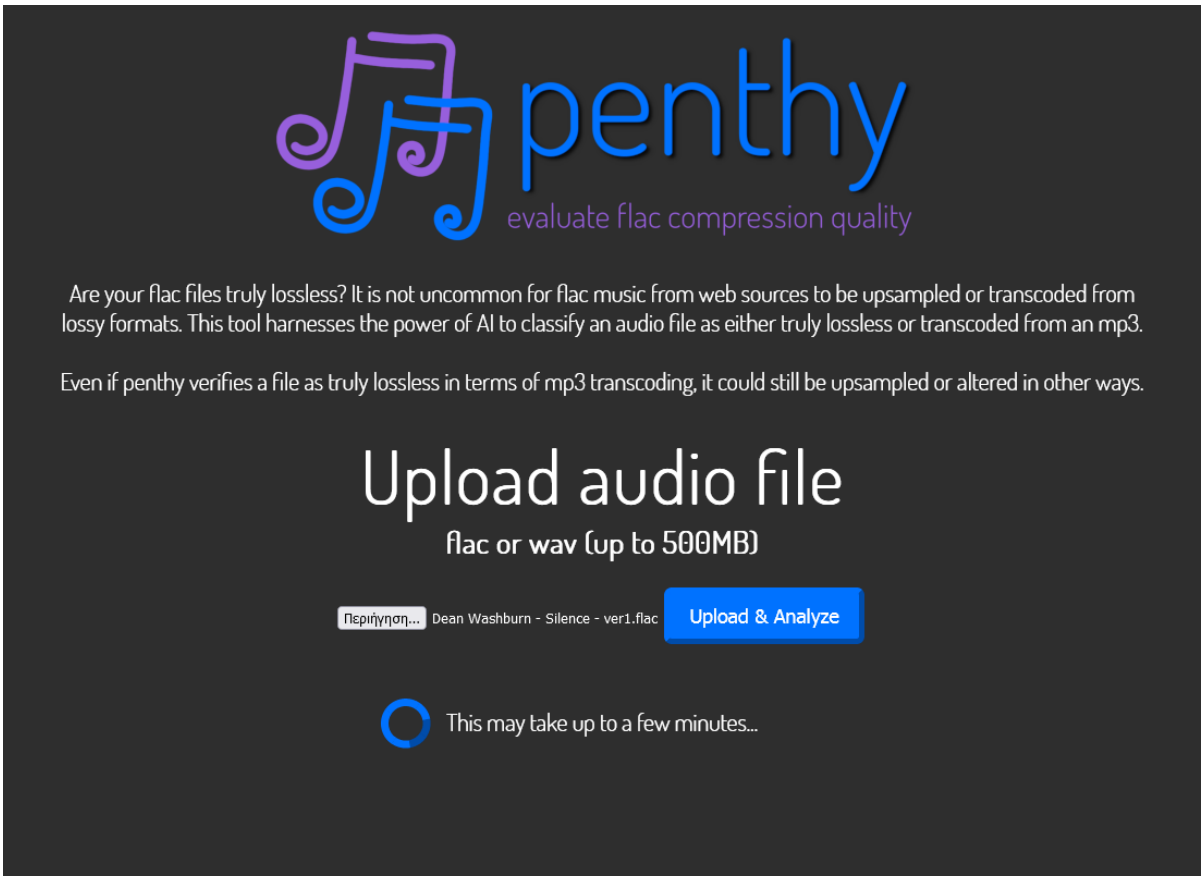
Το προτεινόμενο μοντέλο μπορεί να χρησιμοποιηθεί και μέσω μίας ιστοσελίδας για σκοπούς επίδειξης. Η ιστοσελίδα είναι πλήρως λειτουργική και μπορεί να βρεθεί στη διεύθυνση **penthy.eu** όπου φιλοξενείται μέσω εικονικής μηχανής (Virtual machine – VM). Σε αυτή την εκδοχή του μοντέλου γίνεται εφικτή η δοκιμή του σε οποιοδήποτε αρχείο flac ή wav του χρήστη, για οποιαδήποτε ανάλυση με μέγεθος μέχρι 500 MB. Κάθε αρχείο ανεβαίνει στον εξυπηρετητή (server) μεμονωμένα, δημιουργούνται τα spectrograms και τροφοδοτούνται στο νευρωνικό δίκτυο, το οποίο και παρέχει μία δυαδική απάντηση, η οποία προβάλλεται στον χρήστη. Η διαδικασία υποβολής ενός αρχείου στον server είναι αισθητά πιο αργή από την τοπική εκτέλεση του μοντέλου, αλλά δείχνει διαδραστικά τις δυνατότητες του μοντέλου χωρίς να χρειάζεται εγκατάστασή του. Ακολουθούν ενδεικτικές εικόνες από την ιστοσελίδα.



Εικόνα 14 – Αρχική σελίδα

Για λόγους εύκολης απομνημόνευσης, δόθηκε στο μοντέλο το εμπορικό όνομα penthy, το οποίο είναι η σύντομη μορφή του ονόματος Penthesilea. Η Πενθεσίλεια ήταν βασίλισσα των Αμαζόνων στον Τρωικό πόλεμο, σύμφωνα με την αρχαία ελληνική μυθολογία.

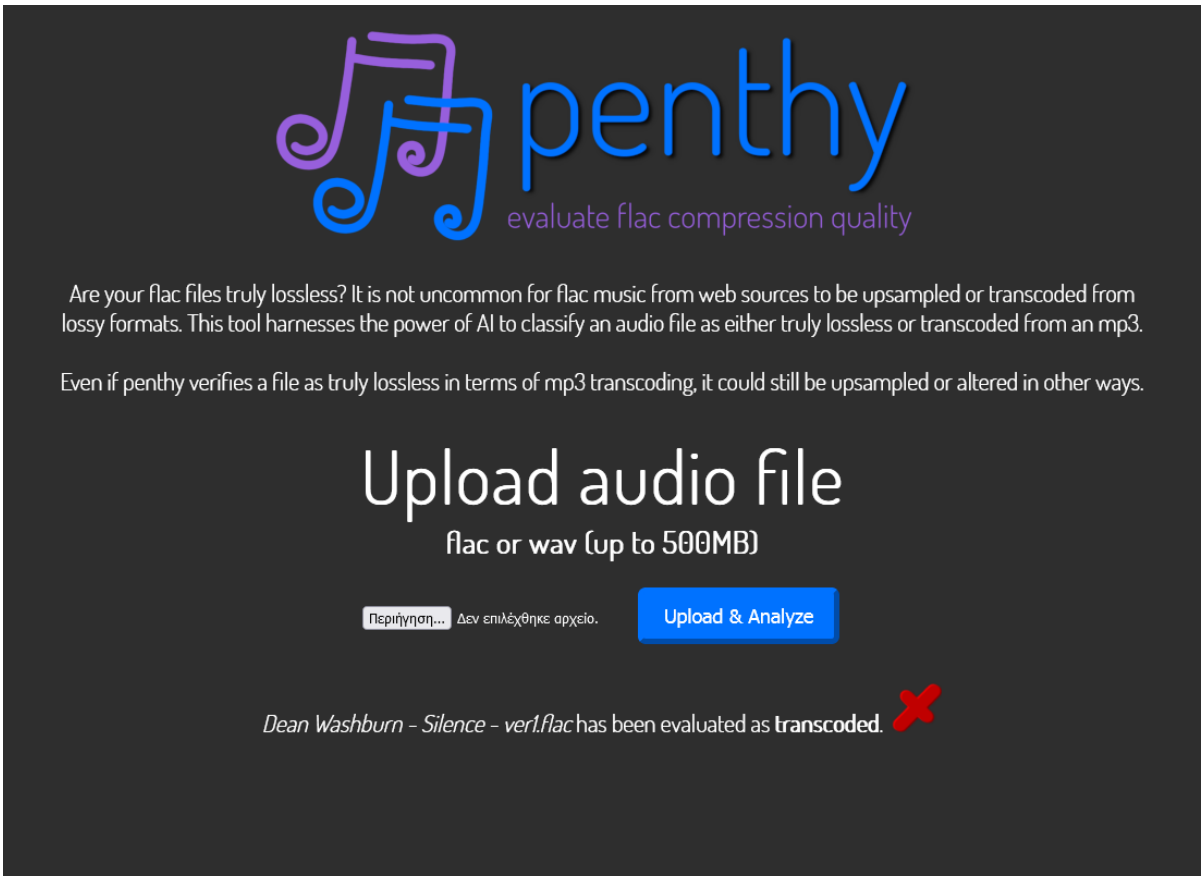
Με την επιλογή ενός αρχείου από τον υπολογιστή του χρήστη και το πάτημα του κουμπιού, ξεκινάει το ανέβασμα και η ανάλυση.



Εικόνα 15 – Φόρτωση μετά το πάτημα του κουμπιού

Η διάρκεια αναμονής εξαρτάται από το μέγεθος του αρχείου, την ταχύτητα σύνδεσης του χρήστη και τη διάρκεια του κομματιού.

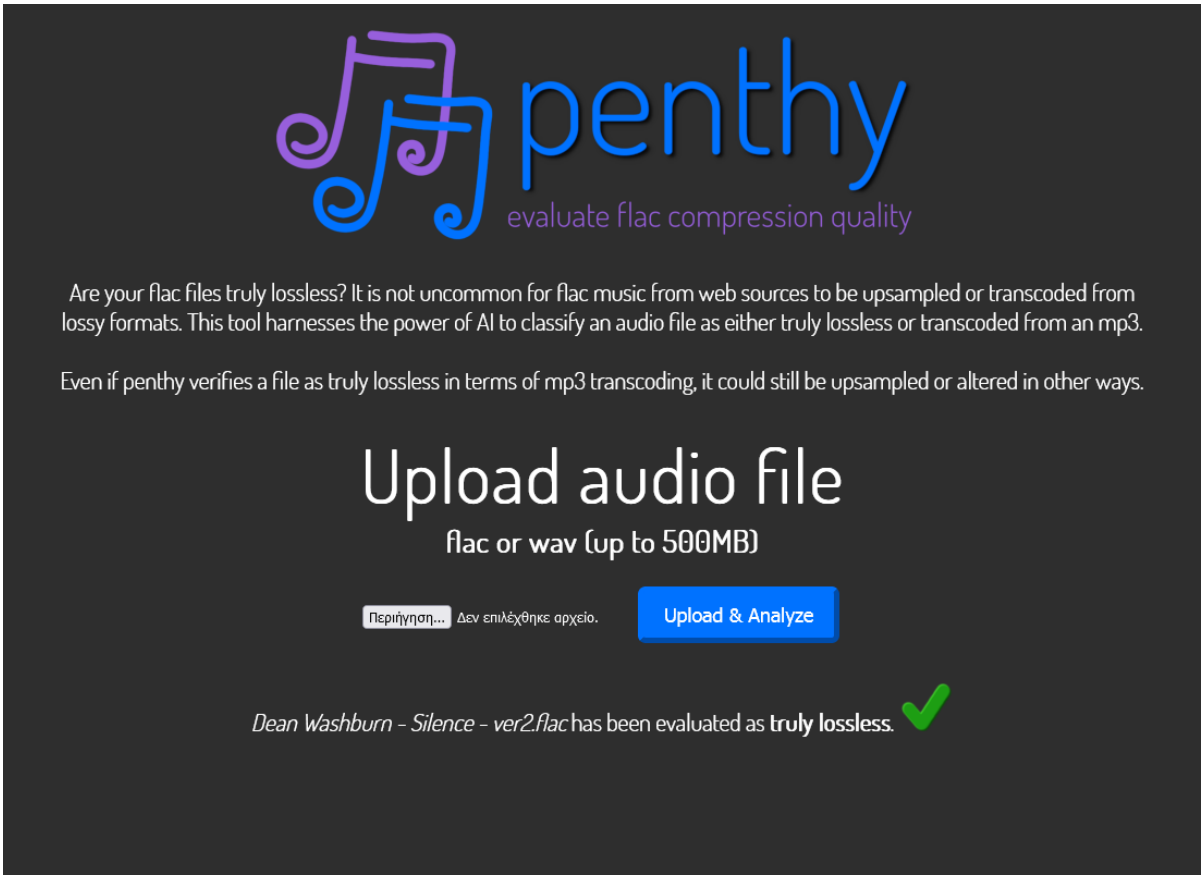
Όταν τελειώσει η ανάλυση του κομματιού, η σελίδα αναγράφει το αποτέλεσμα. Είτε ανιχνεύτηκε mp3 transcoding, είτε το αρχείο πέρασε τη δοκιμασία.



Εικόνα 16 – Αρνητικό αποτέλεσμα

(Οι εικόνες παρουσιάζονται μεγεθυσμένες στην περιοχή ενδιαφέροντος.)

Η απάντηση συνοδεύεται από χρωματιστό εικονίδιο για εύκολη αναγνώριση.

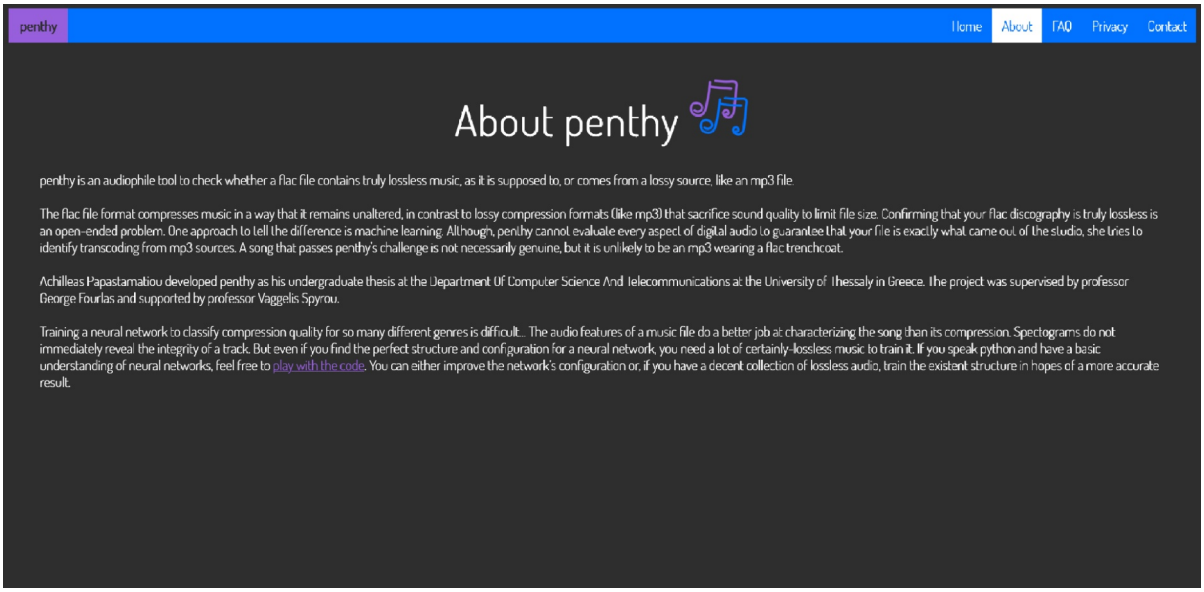


Εικόνα 17 – Θετικό αποτέλεσμα

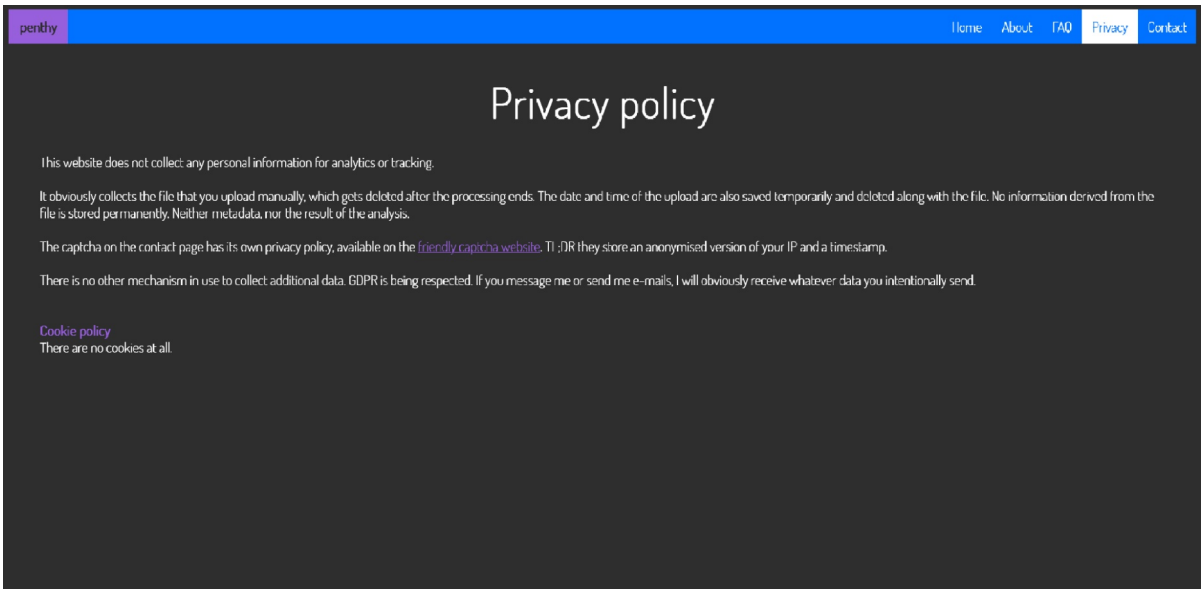
Το αρχείο διαγράφεται αυτόματα από τον server σε αυτό το σημείο.

5.1.β Επιπρόσθετες πληροφορίες

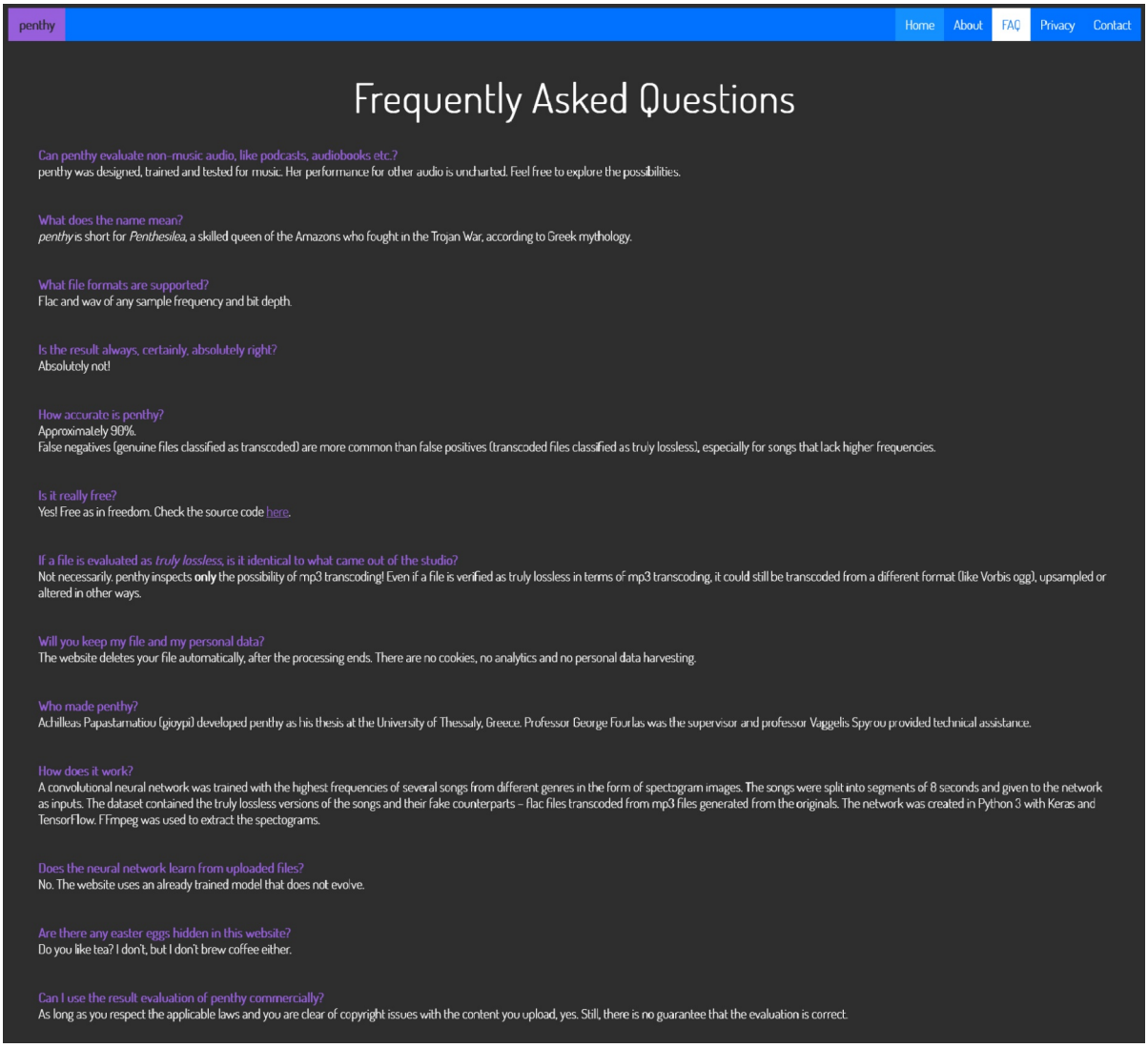
Ο ιστότοπος περιλαμβάνει όλες τις απαραίτητες πληροφορίες για την επεξήγηση του συστήματος. Ακολουθούν εικόνες από τις υπόλοιπες σελίδες.



Εικόνα 18 – Σελίδα «Σχετικά»

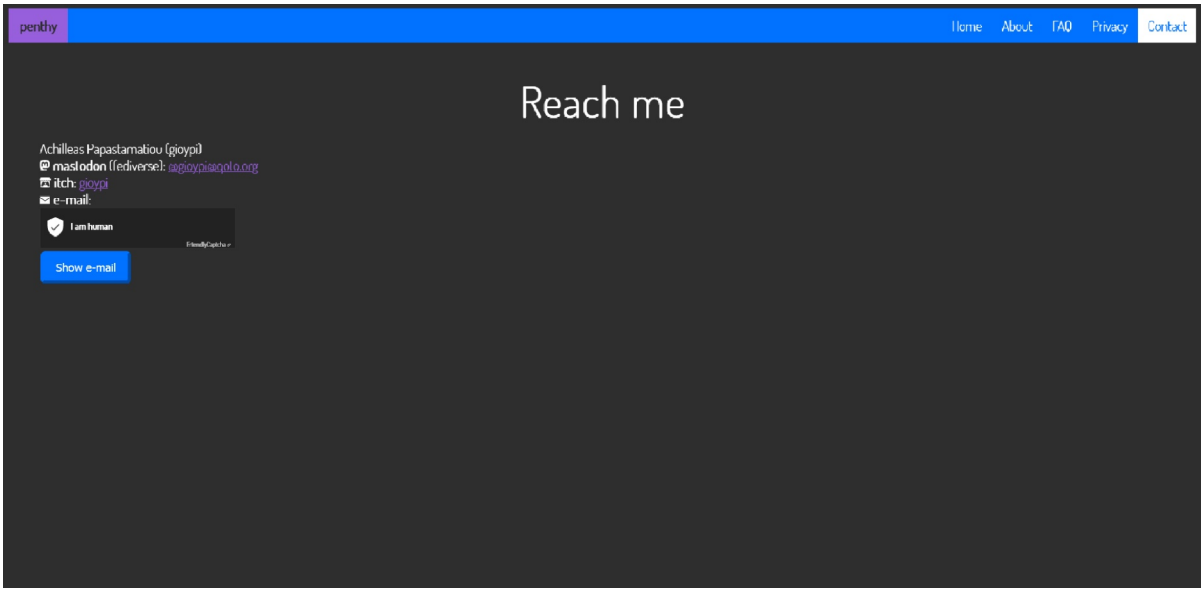


Εικόνα 19 – Σελίδα Πολιτικής απορρήτου



Εικόνα 20 – Σελίδα Συνηθισμένων ερωτήσεων

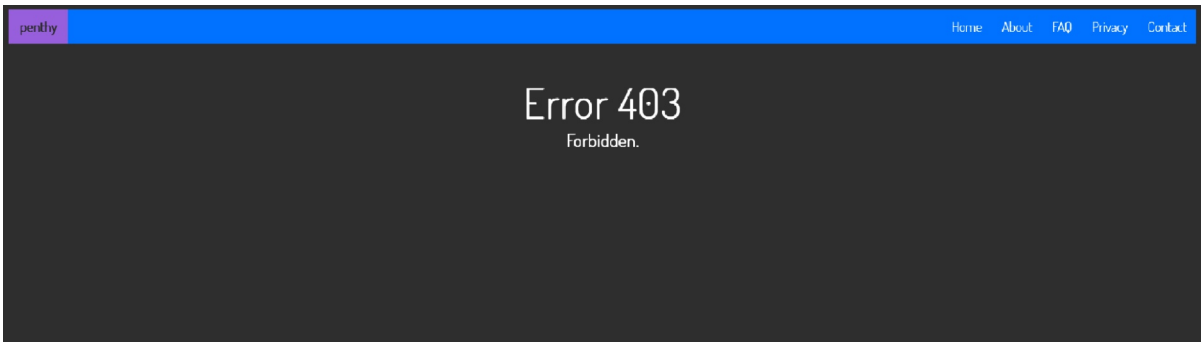
Η σελίδα επικοινωνίας ενσωματώνει έλεγχο CAPTCHA για την προστασία της διεύθυνσης email. Ο έλεγχος γίνεται με τεχνολογία αλυσίδας μπλοκ (blockchain), ώστε να μην απαιτείται από τον χρήστη να συμπληρώσει κάποια φόρμα χειροκίνητα.



Εικόνα 21 – Σελίδα Επικοινωνίας

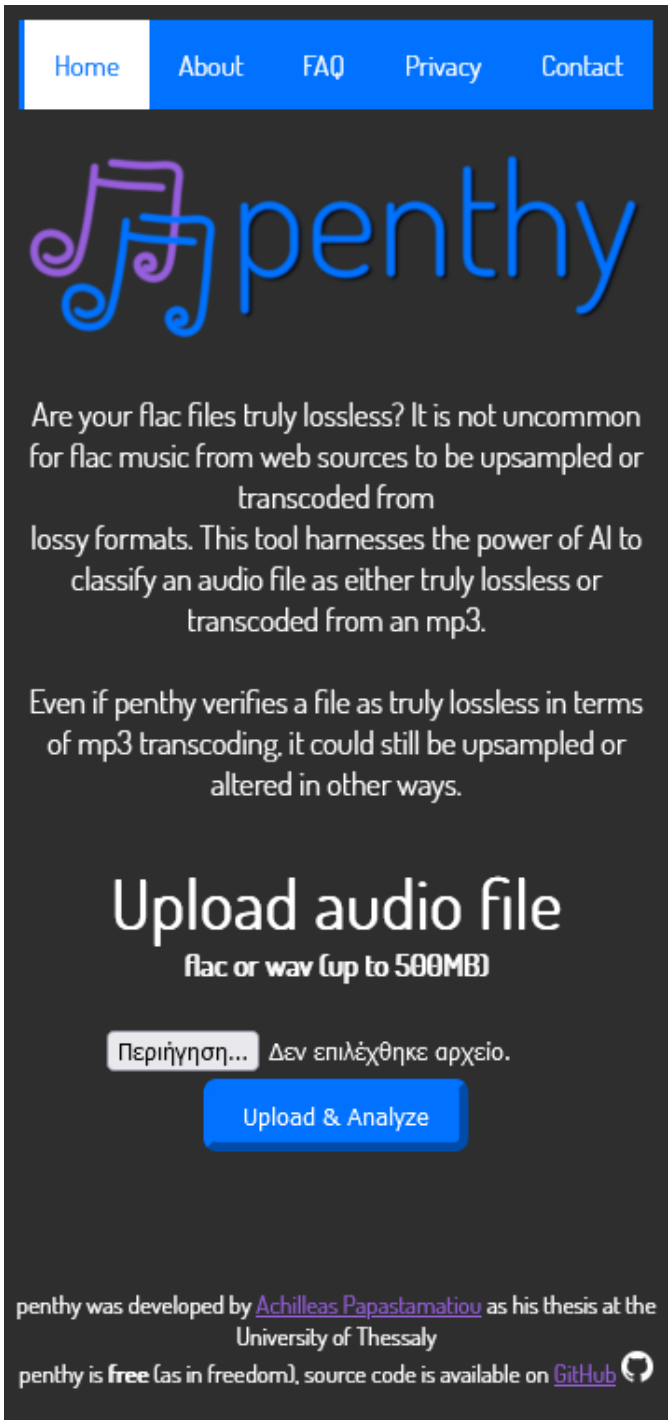
Η ιστοσελίδα είναι σχεδιασμένη με σκοπό την ασφάλεια των χρηστών και την προστασία της ιδιωτικότητάς τους. Δεν υπάρχει κανένας μηχανισμός καταγραφής δεδομένων των χρηστών.

Ακόμη, έχουν υλοποιηθεί εξατομικευμένες σελίδες σφαλμάτων.



Εικόνα 22 – Η σελίδα Σφάλματος 403 εμφανίζεται αν ο χρήστης προσπαθήσει να αποκτήσει πρόσβαση στον φάκελο των ανεβασμένων αρχείων.

Ο ιστότοπος υποστηρίζει διαφορετικές αναλύσεις οθονών για κάθετο και οριζόντιο προσανατολισμό, επομένως είναι εύχρηστος και από κινητά τηλέφωνα (responsive design).



Εικόνα 23 – Αρχική σελίδα σε προβολή κινητού τηλεφώνου

5.1.γ Ασφάλεια και απόδοση

Ο server φιλοξενείται από ένα VM που βρίσκεται στην Ολλανδία. Το εικονικό υλικό του είναι 1 πυρήνας CPU και 1 GB RAM και ο αποθηκευτικός χώρος είναι SSD (Solid State Drive). Το λειτουργικό σύστημα είναι Ubuntu Linux και το πρόγραμμα Apache χρησιμοποιείται για τη φιλοξενία της ιστοσελίδας. Το νευρωνικό δίκτυο εκτελείται μέσα στο VM και επικοινωνεί με την ιστοσελίδα.

Ο server είναι ρυθμισμένος για να στέλνει όλες τις απαραίτητες κεφαλίδες (headers) που υλοποιούν πολιτικές ασφάλειας, για την πρόληψη κοινών επιθέσεων και την κάλυψη γνωστών ευαίσθητων σημείων των πρωτοκόλλων του διαδικτύου.

Επίσης, έχει εκδοθεί πιστοποιητικό TLS (Transport Layer Security, διάδοχος του Secure Sockets Layer – SSL) και ο server έχει ρυθμιστεί για να κάνει τις κατάλληλες ανακατευθύνσεις από HTTP (Πρωτόκολλο Μεταφοράς Υπερκειμένου – Hypertext Transfer Protocol) σε HTTPS (HTTP Secure). Έτσι, επιτυγχάνεται η κρυπτογραφημένη σύνδεση των χρηστών με τον server.

Ανακατευθύνσεις γίνονται και για την κατάλληλη χρήση του προθέματος *www*, της IP του server και των ονομάτων των υποσελίδων. Για παράδειγμα, η διεύθυνση *http://www.penthy.eu* ανακατευθύνεται στη διεύθυνση *https://penthy.eu*. Πρόσθετες βελτιστοποιήσεις καταλήγουν στην ομαλή χρήση του ιστότοπου με ελάχιστες απαιτήσεις δικτύου.

Στον έλεγχο ασφάλειας του Mozilla Observatory, η ιστοσελίδα αξιολογείται ως A+ με βαθμό 110/100 (Mozilla, 2022). Επιπλέον, στον έλεγχο ταχύτητας του Pagespeed Insights της Google η απόδοση της ιστοσελίδας βαθμολογείται με 100% για υπολογιστές και 88% για κινητά τηλέφωνα (Google, 2022).

5.2 Προηγούμενες προσεγγίσεις

Τα πειράματα ξεκίνησαν με απλούστερες αρχιτεκτονικές νευρωνικών δικτύων, πριν σχεδιαστεί το προτεινόμενο μοντέλο που υλοποιεί ένα CNN. Αρχικά δοκιμάστηκαν αλγοριθμικά χαρακτηριστικά του ήχου αντί για spectrograms ως είσοδος σε ένα feedforward ή recurrent neural network. Χαρακτηριστικά του φάσματος του ήχου χρησιμοποιούνται συχνά ως είσοδος σε νευρωνικά δίκτυα που αναλύουν μουσική, για παράδειγμα για την κατηγοριοποίηση τραγουδιών ανά μουσικό είδος. Συγκεκριμένα, το *Φασματικό Κεντροειδές (Spectral Centroid)* και η *Διασπορά Φάσματος (Spread Spectrum)* παράχθηκαν από το dataset εκπαίδευσης και δοκιμάστηκαν ως δεδομένα–κλειδιά για τον διαχωρισμό truly lossless και transcoded flac.

Αυτό το μέρος των πειραμάτων δεν είχε επιτυχία. Τα μοντέλα αυτά είχαν ορθότητα περίπου ίση με 50%, που σε ένα δυαδικό πρόβλημα αντιπροσωπεύει τυχαίο αποτέλεσμα. Για τον λόγο αυτό, δεν είναι ουσιώδης η σύγκρισή τους με το προτεινόμενο μοντέλο. Γίνεται κατανοητό ότι τέτοια αλγοριθμικά χαρακτηριστικά είναι χρήσιμα όταν η περιγραφή του τραγουδιού είναι το επιθυμητό αποτέλεσμα. Στην προκειμένη περίπτωση, χρειαζόταν ένα χαρακτηριστικό που να περιγράφει τη συμπίεση του αρχείου και να παραμένει αξιόπιστο για κάθε τραγούδι, κάθε μουσικού είδους. Οι ανώτερες συχνότητες των spectrograms, που χρησιμοποιήθηκαν τελικώς, ταιριάζουν σε αυτό το σενάριο και παρουσιάζουν μεγαλύτερη διαφοροποίηση μεταξύ transcoded και μη transcoded εκδοχών του ίδιου τραγουδιού.

5.3 Ευρήματα της εργασίας

Η επαλήθευση της γνησιότητας αρχείων ήχου μη απωλεστικής συμπίεσης παραμένει ανοιχτό πρόβλημα. Ο διαχωρισμός μεταξύ truly lossless και transcoded flac, όπως αναλύθηκε, εξαρτάται από πολλούς παράγοντες, καθώς ένα αρχείο flac θα μπορούσε να υποστεί αλλοίωση με διάφορους τρόπους. Συγκεκριμένα, η ανίχνευση transcoding από πηγές mp3 κωδικοποίησης είναι εφικτή με το προτεινόμενο μοντέλο, με αξιοπρεπή απόδοση. Το μοντέλο που παρουσιάστηκε χρησιμοποιεί τις υψηλότερες συχνότητες μουσικών κομματιών σε μορφή εικόνων spectrograms και εκπαιδεύει ένα νευρωνικό δίκτυο. Το δίκτυο είναι συνελικτικό (CNN), αποτελείται από αρκετά επίπεδα και αναγνωρίζει τη διαφορά μεταξύ ενός πραγματικού flac αρχείου και ενός αρχείου mp3 κωδικοποιημένου ως flac, ανεξάρτητα από την ανάλυση της κωδικοποίησης. Με το πέρας των πειραμάτων και την πλήρη υλοποίηση του μοντέλου, καταλήγουμε στο συμπέρασμα ότι αυτός ο τομέας του προαναφερθέντος προβλήματος έχει λύση. Ωστόσο, είναι απαραίτητο να χρησιμοποιηθεί η κατάλληλη αρχιτεκτονική του νευρωνικού δικτύου, αλλά και να υπάρχει ένα dataset επαρκούς μεγέθους και ορθά προετοιμασμένο για εκπαίδευση.

Με την εφαρμογή αυτού του μοντέλου, όπως και άλλων λύσεων για τις λοιπές πιθανές παραμορφώσεις της μη απωλεστικής μουσικής, δίνεται η ευκαιρία αξιολόγησης της ποιότητας ήχου στη βιομηχανία διακίνησης της μουσικής. Τόσο οι εμπορικές πλατφόρμες μουσικής υψηλής ανάλυσης, όσο και οι χρήστες τους, ωφελούνται με την τήρηση ελέγξιμων προδιαγραφών. Οι ακροατές που γνωρίζουν ότι το προϊόν που αγοράζουν έχει ελεγχθεί για τη γνησιότητά του, έχουν περισσότερες πιθανότητες να μείνουν ευχαριστημένοι με την επένδυσή τους. Οι εταιρείες μεταπώλησης μουσικής υψηλής ανάλυσης μπορούν να δικαιολογήσουν στους πελάτες τους το μεγαλύτερο κόστος σε σχέση με αρχεία ή ροές απωλεστικής συμπίεσης, υποδεικνύοντας τη διαφορά της ποιότητας. Ακόμη, ίσως και να είναι εφικτό να χρησιμοποιήσουν το αποτέλεσμα τέτοιων εργαλείων προωθητικά, ως ένδειξη τήρησης υψηλών προδιαγραφών. Τέλος, οι καλλιτέχνες και οι μουσικοί παραγωγοί θα μπορούν να γνωρίζουν ότι στο τέλος της εμπορικής αλυσίδας οι απαιτητικοί ακροατές τους απολαμβάνουν τον ήχο ακριβώς όπως αυτοί επιθυμούσαν και όχι αλλοιωμένο. Επομένως, ηλεκτρονικές λύσεις για την επαλήθευση της συμπίεσης ήχου μπορούν να διεκδικήσουν μία θέση στην αγορά της μουσικής.

5.4 Μελλοντικές εργασίες

Αν αναλογιστούμε τρόπους βελτίωσης του προτεινόμενου μοντέλου, υπάρχουν κάποιες ευκαιρίες για μελλοντικές εργασίες. Το υπάρχον νευρωνικό δίκτυο θα μπορούσε να επανεκπαιδευτεί με άλλα datasets για καλύτερη απόδοση. Επίσης, μοντέλα με παρόμοιες αρχιτεκτονικές θα μπορούσαν να ερευνηθούν για ανίχνευση transcoding από άλλες κωδικοποιήσεις. Η βελτίωση του μοντέλου και ο συνδυασμός του με παρεμφερή εργαλεία σε ένα μεγαλύτερο σύστημα αξιολόγησης ποιότητας είναι πιθανό να έχει ευρύτερη πρακτική χρησιμότητα στην αγορά της μουσικής.

Μία πιθανή βελτίωση της απόδοσης θα ήταν η εκπαίδευση του νευρωνικού με την τωρινή αρχιτεκτονική, διευρύνοντας σημαντικά το dataset εκπαίδευσης. Ίσως μεγαλύτερο dataset να προσφέρει βελτιωμένη ακρίβεια, ειδικά αν περιλαμβάνει αρκετά δείγματα με έλλειψη υψηλών συχνοτήτων, σαν αυτά που εντάχθηκαν στα False Negatives. Επιπλέον, η αξιοπιστία του μοντέλου θα μπορούσε να αλλάξει θετικά αν υπήρχε πρόσβαση σε βάσεις δεδομένων μουσικών στούντιο, ώστε τα truly lossless δείγματα να προέρχονται από πιστοποιημένα γνήσιες πηγές.

Μία εναλλακτική πρόταση περιλαμβάνει την ανάπτυξη νέων μοντέλων, με καινούργιες αρχιτεκτονικές, για την αντιμετώπιση των υπόλοιπων πτυχών του προβλήματος της γνησιότητας. Για παράδειγμα, μικρές ρυθμίσεις του υπάρχοντος μοντέλου και νέα εκπαίδευση θα μπορούσε να παράξει ένα εργαλείο ανίχνευσης transcoding από πηγές AAC, καθώς η κωδικοποίηση AAC δε διαφέρει πολύ από την κωδικοποίηση mp3. Ομοίως, χρήσιμο θα ήταν και ένα νευρωνικό δίκτυο για ανίχνευση transcoding από Vorbis πηγές (ogg).

Ιδανικά, η βιομηχανία διακίνησης της μουσικής, θα μπορούσε να ωφεληθεί από λογισμικό που θα επαλήθευε πλήρως τη γνησιότητα των αρχείων flac, συμπεριλαμβανομένου του transcoding, του upsampling και άλλων αλλοιώσεων. Τέτοιο λογισμικό θα μπορούσε να αναπτυχθεί ίσως ως ένωση πολλαπλών μοντέλων για ανίχνευση κάθε συχνής παραποίησης. Αυτή η εργασία θέλει να δείξει ότι η αναγνώριση mp3 transcoding, ως μέρος της λύσης, είναι εφικτή.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Faruq P., ανακτήθηκε 2022. Sample Rate, Bit-Depth & Bitrate. Exclusivemusicplus. https://exclusivemusicplus.com/audio-processing/sample-rate-bit-depth-bitrate

Cunningham, S. and McGregor, I., 2019. Subjective Evaluation of Music Compressed with the ACER Codec Compared to AAC, MP3, and Uncompressed PCM. International Journal of Digital Multimedia Broadcasting, σελ.1-16.

Spotify, ανακτήθηκε 2022. Premium Plans.
https://support.spotify.com/us/article/premium-plans/

Tidal, ανακτήθηκε 2022. Pricing.
https://tidal.com/pricing

Corbett I., 2012. What Data Compression Does To Your Music. Soundonsound.
https://www.soundonsound.com/techniques/what-data-compression-does-your-music

Opris E., 2013. Lightweight, portable and simple-to-use app that checks the origin of APE, FLAC, SHN, WAV and LPAC files to verify if the format is lossless. Softpedia.
https://www.softpedia.com/get/Multimedia/Audio/Other-AUDIO-Tools/Audiochecker.shtml

Lacroix J., Prime Y., Remy A., and Derrien O., 2015. Lossless Audio Checker: A Software for the Detection of Upscaling, Upsampling, and Transcoding in Lossless Musical Tracks. 139th International Audio Engineering Society Convention (AES).

Derrien O., 2019. Detection of Genuine Lossless Audio Files: Application to the MPEG-AAC Codec. Journal of the Audio Engineering Society (JAES), Volume 67 Issue 3, σελ. 116-123.

D'Alessandro B. and Shi Y., 2009. Mp3 bit rate quality detection through frequency spectrum analysis. Proceedings of the 11th ACM workshop on Multimedia and security, σελ. 57-62.

Singh R. D. and Aggarwal N., 2015. Detection of re-compression, transcoding and frame-deletion for digital video authentication. 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS), σελ. 1-6.

Xu J., Su Y. and You X., 2012. Detection of video transcoding for digital forensics. International Conference on Audio, Language and Image Processing, σελ. 160-164.

Musmann H. G., 2006. Genesis of the MP3 audio coding standard. IEEE Transactions on Consumer Electronics, vol. 52, no. 3, σελ. 1043-1049.

Jayant N., Johnston J. and Safranek R., 1993. Signal compression based on models of human perception. Proceedings of the IEEE, vol. 81, no. 10, σελ. 1385-1422.

Xiph.org Foundation, ανακτήθηκε 2022. Flac news.
https://xiph.org/flac/news.html

Callum, 2018. Beginner's Guide To… Codecs and Compression. Richone.
https://blog.richtonemusic.co.uk/posts/audio-formats-codecs-compression/

Pigeon S., ανακτήθηκε 2022. Looking At Flac Compression Ratios. Harder, Better, Faster, Stronger.
https://hbfs.wordpress.com/2012/02/07/looking-at-flac-compression-ratios/

AudioMountain, ανακτήθηκε 2022. Audio File Size Calculations.
https://www.audiomountain.com/tech/audio-file-size.html

Wikipedia, ανακτήθηκε 2022. Άρθρα για τις διάφορες κωδικοποιήσεις ήχου. Ενδεικτικά: Comparison of audio coding formats.
https://en.wikipedia.org/wiki/Comparison_of_audio_coding_formats

Chollet F., 2021. Deep Learning with Python. Manning Publications Co., 2nd edition, σελ. 2-18.

Keim R., 2019. How to Perform Classification Using a Neural Network: What Is the Perceptron? All about circuits.

Rojas R., 1996. Neural Networks. Springer-Verlag, κεφάλαιο 8.

IBM, 2020. Neural Networks. IBM Cloud Education.
https://www.ibm.com/cloud/learn/neural-networks

Encyclopedia of Mathematics, ανακτήθηκε 2022. Convolution of functions.
https://encyclopediaofmath.org/wiki/Convolution_of_functions

Wang S. -Y., Wang O., Zhang R., Owens A. and Efros A. A., 2020. CNN-Generated Images Are Surprisingly Easy to Spot… for Now. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), σελ. 8692-8701.

Mebsout I., 2020. Convolutional Neural Networks' mathematics. Towards Data Science, Medium.
https://towardsdatascience.com/convolutional-neural-networks-mathematics-1beb3e6447c0

Becker D., 2018. Rectified Linear Units (ReLU) in Deep Learning. Kaggle.
https://www.kaggle.com/code/dansbecker/rectified-linear-units-relu-in-deep-learning/notebook

Saeed M., 2021. A Gentle Introduction To Sigmoid Function. Machine Learning Mastery.
https://machinelearningmastery.com/a-gentle-introduction-to-sigmoid-function/

Fawcett T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, Volume 27, Issue 8, σελ. 861-874.

Mozilla, ανακτήθηκε 2022. HTTP Observatory.
https://observatory.mozilla.org/analyze/penthy.eu

Google, ανακτήθηκε 2022. PageSpeed Insights. Google Developers. https://pagespeed.web.dev/report?url=https%3A%2F%2Fpenthy.eu%2F