2η Εργαστήριακή Άσκηση Αναγνώριση Προτύπων
Αναγνώριση Είδους και Εξαγωγή Συναισθήματος από Μουσική

Άρης Μαρκογιαννάκης 03120085

Contents

Προπαρασκευή

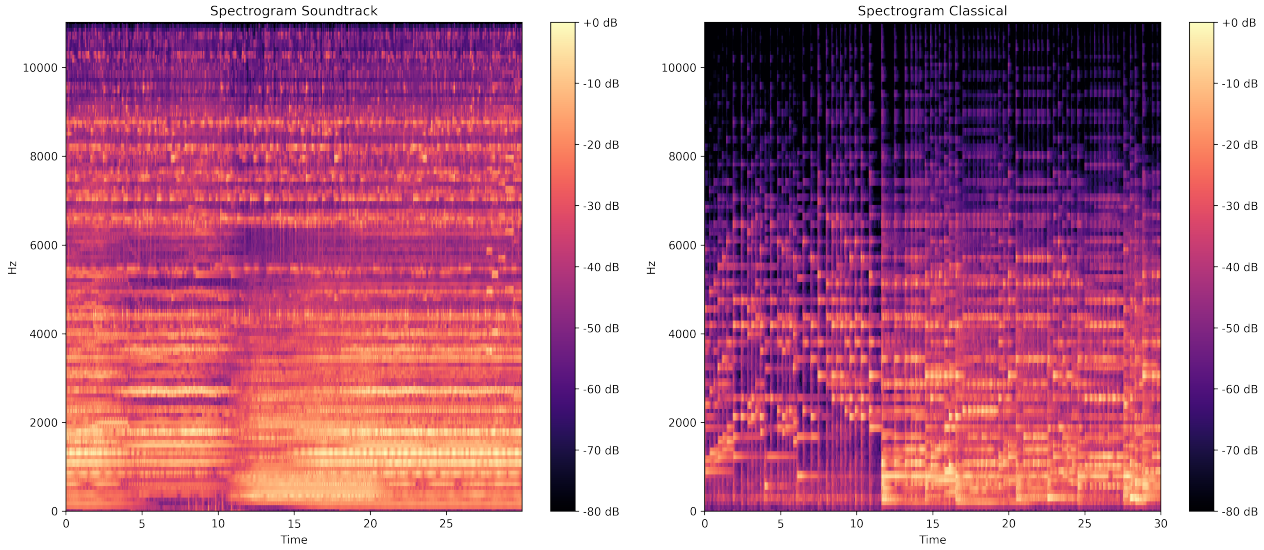
Αρχικά, επισκεπτόμαστε το Kaggle και χρησιμοποιώντας τον kernel που μας δίνεται ως οδηγό δημιουργούμε έναν δικό μας «κλώνο» τον οποίο χρησιμοποιούμε για την υλοποίηση της άσκησης.

Βήμα 1: Εξοικείωση με φασματογραφήματα στην κλίμακα mel

(α) Επιλέγουμε δύο τυχαίες γραμμές με διαφορετικά labels. Επιλέγουμε με τυχαίο τρόπο μία γραμμή με label "Soundtrack" και μία με label "Classical".

(β) Έπειτα φορτώνουμε από τον φάκελο fma_genre_spectrograms/train τα αντίστοιχα αρχεία για τα φασματογράφημα σε κλίμακα mel των αντίστοιχων γραμμών που επιλέξαμε προηγουμένως.

(γ) Έπειτα απεικονίζουμε τα φασματογραφήματα με χρήση της συνάρτησης librosa.display.specshow. Τα αποτελέσματα φαίνονται παρακάτω:



Soundtrack: Έχει ισχυρή παρουσία στις μεσαίες και υψηλές συχνότητες, ιδιαίτερα μεταξύ 2000-8000 Hz, που είναι χαρακτηριστικό της μουσικής με έντονα ρυθμικά ή μελωδικά στοιχεία. Οι περιοχές με έντονα χρώματα (πιο φωτεινές) δείχνουν ισχυρή ένταση, υποδηλώνοντας πλούσιες αρμονικές ή συχνούς παλμούς στον ήχο. Υπάρχει διαρκής παρουσία ενέργειας στις υψηλές συχνότητες, που συνδέεται με εφέ ή δυναμικά μουσικά στοιχεία (όπως τύμπανα ή ηλεκτρονικά όργανα).

Classical: Διακρίνουμε έντονη δραστηριότητα στις χαμηλές και μεσαίες συχνότητες (κάτω από 4000 Hz), γεγονός που συνδέεται με την παρουσία οργάνων όπως έγχορδα ή πνευστά. Υπάρχουν χρονικά κενά μεταξύ έντονων περιοχών, κάτι που αντανακλά τη χρήση παύσεων ή αλλαγών στην ένταση που είναι χαρακτηριστικές της κλασικής μουσικής. Οι στάθμες των συχνοτήτων είναι πιο ευδιάκριτες και συχνά διαχωρισμένες, που υποδηλώνει μεγαλύτερη έμφαση στη μελωδία και στην πολυφωνία.

(δ) Η κλίμακα Mel είναι μια κλίμακα συχνοτήτων που δημιουργήθηκε με βάση τον τρόπο που το ανθρώπινο αυτί αντιλαμβάνεται τον ήχο. Αναπτύχθηκε μέσω ψυχοακουστικών πειραμάτων, όπου οι άνθρωποι κλήθηκαν να κρίνουν αν δύο ήχοι έχουν ίσες "ακουστικές αποστάσεις". Η κλίμακα Mel συμπιέζει τις υψηλές συχνότητες και επεκτείνει τις χαμηλές, ώστε να προσεγγίσει τη μη γραμμική ευαισθησία της ανθρώπινης ακοής. Χρησιμοποιείται στην επεξεργασία μουσικών σημάτων, ειδικά στη δημιουργία Mel-spectrograms, διότι προσομοιώνει τη λειτουργία του αυτιού, βελτιώνοντας την ποιότητα χαρακτηριστικών που είναι πιο σημαντικά για την ανάλυση και αναγνώριση μουσικών σημάτων.

Βήμα 2: Συγχρονισμός φασματογραφημάτων στο ρυθμό της μουσικής (beat-synced spectrograms)

(α) Παρακάτω φαίνονται οι διαστάσεις των φασματογραφημάτων του βήματος 1:

```
Soundtrack shape of spectrogram: (128, 1291)
Classical  shape of spectrogram: (128, 1292)
```

Βλέπουμε ότι στο Soundtrack έχουμε 1291 χρονικά βήματα και στο Classical 1292 χρονικά βήματα.

Ένα LSTM μοντέλο είναι κατάλληλο για επεξεργασία ακολουθιακών δεδομένων. Αφού τα φασματογραφήματα μας είναι στην ουσία μία ακολουθία που μας δίνει τις εντάσεις των συχνοτήτων σε μία ακολουθία χρονικών στιγμών, το LSTM θα έχει μάλλον καλή επίδοση.

Τα 1291-1292 χρονικά βήματα είναι σχετικά μεγάλα σε αριθμό κι έτσι αν χρησιμοποιηθεί ένα LSTM χωρίς προεπεξεργασία (π.χ. υποδειγματοληψία ή μείωση διαστάσεων), ενδέχεται να απαιτηθούν σημαντικοί πόροι υπολογισμού (μνήμη, χρόνος). Ακόμη, όσο μεγαλύτερη η ακολουθία, τόσο πιο δύσκολο για το LSTM να διατηρήσει μακροχρόνιες εξαρτήσεις. Επίσης, μεγάλες είσοδοι και λίγα δεδομένα εκπαίδευσης ενδέχεται να οδηγήσουν το μοντέλο σε overfitting.

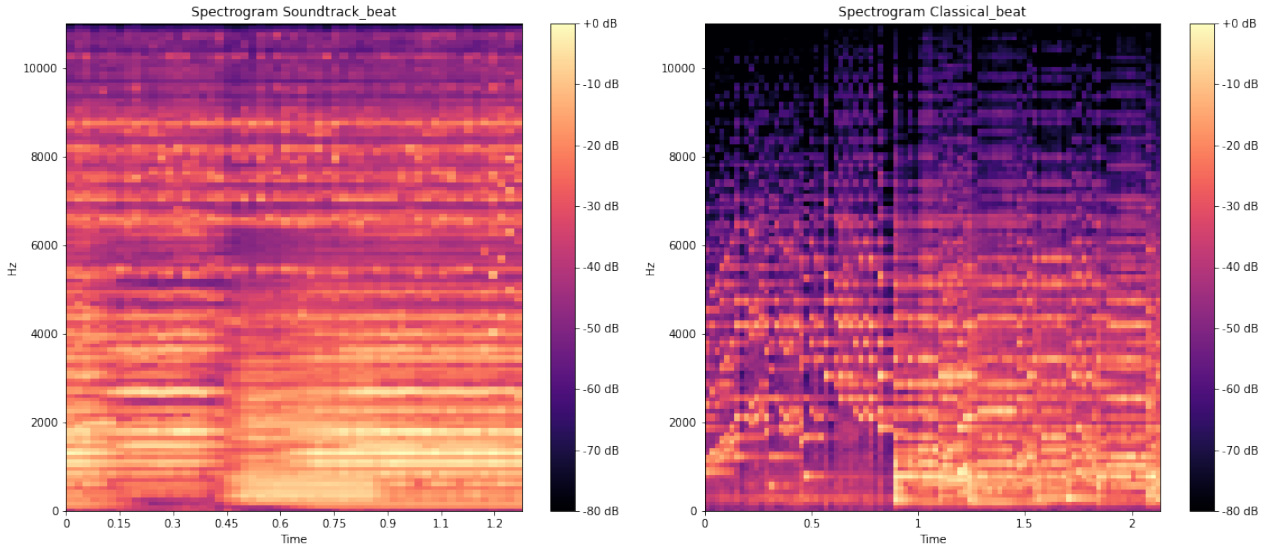
(β) Ένας τρόπος να μειώσουμε τα χρονικά βήματα είναι να συγχρονίσουμε τα φασματογραφήματα πάνω στο ρυθμό. Για αυτό το λόγο παίρνουμε τη διάμεσο (median) ανάμεσα στα σημεία που χτυπάει το beat της μουσικής.Παίρνουμε τα αρχεία που μας δίνονται δίνονται στο φάκελο /input/patreco3-multitaskaffectivemusic/data/fma_genre_spectrogram_beat.

Οι διαστάσεις των φασματογραφημάτων τώρα είναι:

```
Soundtrack shape of spectrogram: (140, 55)
Classical  shape of spectrogram: (128, 92)
```

Βλέπουμε ότι στο Sountrack έχουμε μείνει με το 4% περίπου των αρχικών χρονικών βημάτων ενώ στο classical περίπου με το 7% των αρχικών χρονικών βημάτων.

Τα φασματογραφήματα έπειτα από την μείωση των αρχικών χρονικών βημάτων:



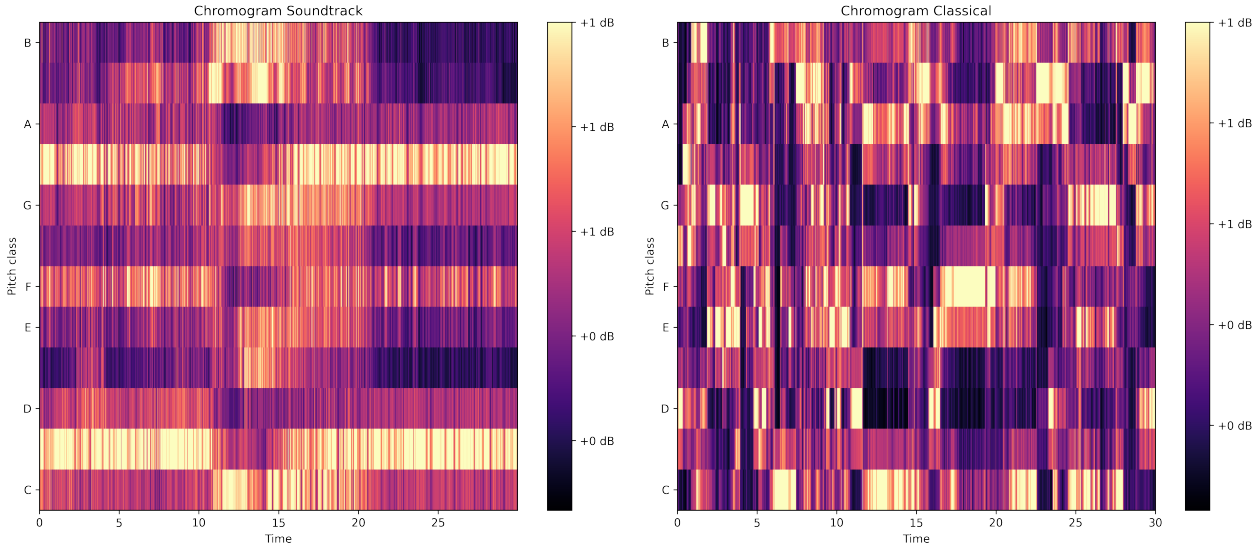
Παρατηρούμε ότι τα δύο φασματογαφήματα παρά την σημαντική μείωση των χρονικών βημάτων, μοιάζουν αρκετά με τα αρχικά.

Βήμα 3: Εξοικείωση με χρωμογραφήματα

Παρακάτω φαίνονται οι διαστάσεις των αρχικών χρωμογραφημάτων:

```
Soundtrack shape of chromogram: (12, 1291)
Classical  shape of chromogram: (12, 1292)
```

Απεικονίζουμε τα χρωμογραφήματα για τα διαφορετικά με χρήση της συνάρτησης librosa.display.specshow. Τα αποτελέσματα φαίνονται παρακάτω:



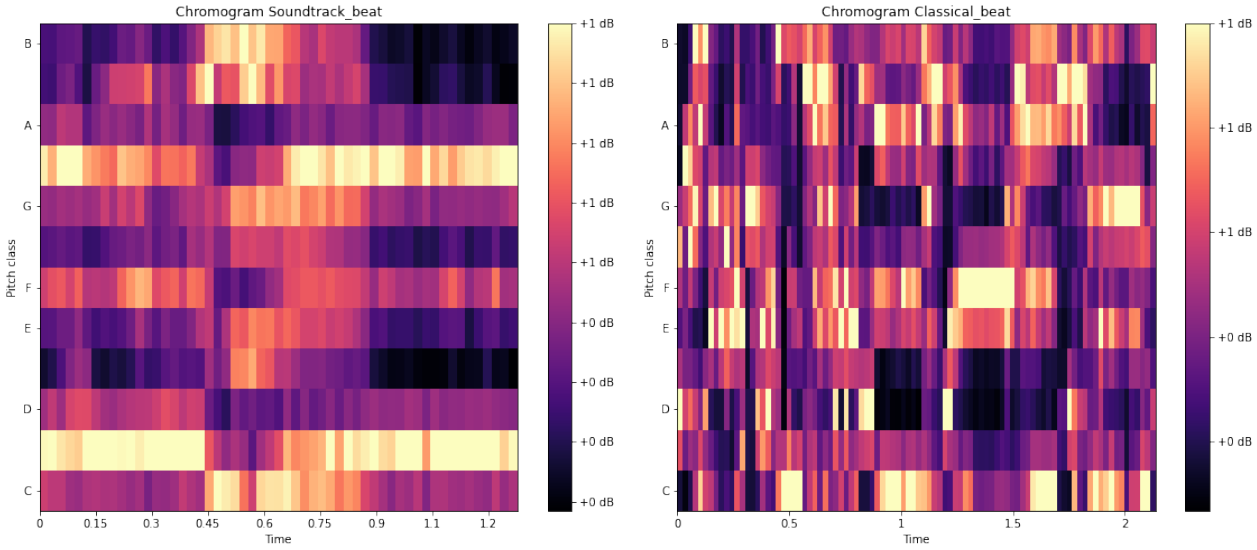
Sountrack: Οι φωτεινές περιοχές δείχνουν συχνότητες όπου οι τόνοι είναι έντονοι. Στο soundtrack υπάρχει γενικά μεγάλη "πληρότητα" τόνων, δηλαδή πολλοί διαφορετικοί τόνοι συμμετέχουν με παρόμοια ένταση σε πολλά χρονικά σημεία. Παρατηρείται γενικά ομαλή κατανομή και λιγότερα κενά στον χρόνο, κάτι που υποδεικνύει ότι το soundtrack έχει πιο σταθερό αρμονικό υπόβαθρο, ίσως λόγω της χρήσης συγχορδιών ή drones που διατηρούνται για μεγάλα χρονικά διαστήματα. Το chromogram δείχνει ότι το soundtrack βασίζεται σε μεγάλη ποικιλία τονικών κατηγοριών ταυτόχρονα, πιθανώς για να δημιουργήσει πολυδιάστατα ηχητικά περιβάλλοντα.

Classical: Το chromogram του classical δείχνει πιο διάσπαρτη χρήση των τόνων με πολλές εναλλαγές έντασης ανάμεσα στις κατηγορίες (pitch classes). Υπάρχουν σαφώς καθορισμένα χρονικά κενά και εναλλαγές στις φωτεινές περιοχές. Αυτό δείχνει ότι το classical χρησιμοποιεί περισσότερες παύσεις ή διαλείμματα, καθώς και μεταβάσεις μεταξύ διαφορετικών αρμονιών. Υπάρχουν στιγμές όπου κυριαρχούν συγκεκριμένοι τόνοι, γεγονός που υποδεικνύει ότι η κλασική μουσική είναι πιο μελωδική και πιθανόν ακολουθεί κλίμακες με βάση την τονικότητα της σύνθεσης.
Οι διαστάσεις των χρωμογραφημάτων μετά την μείωση των αρχικών χρονικών βημάτων είναι:

```
Soundtrack_beat shape of chromogram: (12, 55)
Classical_beat  shape of chromogram: (12, 92)
```

Βλέπουμε ότι στο Sountrack έχουμε μείνει με το 4% περίπου των αρχικών χρονικών βημάτων ενώ στο classical περίπου με το 7% των αρχικών χρονικών βημάτων.

Τα χρωμογραφήματα έπειτα από την μείωση των αρχικών χρονικών βημάτων:



Παρατηρούμε ότι τα δύο χρωμογραφήματα παρά την σημαντική μείωση των χρονικών βημάτων, μοιάζουν αρκετά με τα αρχικά.

Βήμα 4: Φόρτωση και ανάλυση δεδομένων

(α) Ο κώδικας υλοποιεί μια κλάση `SpectrogramDataset` που είναι υπεύθυνη για την ανάγνωση και επεξεργασία φασματογραφημάτων (spectrograms) από δεδομένα που βρίσκονται σε μορφή `.npy`. Οι κύριες λειτουργίες της κλάσης είναι:

1. **Εκκίνηση του `Dataset`:**
   - Η κλάση διαβάζει δεδομένα από τους φακέλους `train` ή `test`, επιλέγοντας το κατάλληλο σύνολο εκπαίδευσης ή αξιολόγησης χρησιμοποιώντας την συνάρτηση `get_files_labels`.
   - Υποστηρίζει τη συγχώνευση ή αφαίρεση παρόμοιων κατηγοριών (labels) μέσω του `CLASS_MAPPING`.
2. **Ανάγνωση Φασματογραμμάτων:**
   - Τα φασματογραφήματα φορτώνονται μέσω της συνάρτησης `read_spectrogram`.
   - Υποστηρίζονται δύο τύποι χαρακτηριστικών:
     - `mel`: Επιστρέφονται τα 128 πρώτα κανάλια του φασματογραφήματος.
     - `chroma`: Επιστρέφονται τα κανάλια από το index 128 και μετά.
3. **Επεξεργασία Ετικετών:**
   - Χρησιμοποιείται ο μετασχηματιστής `LabelTransformer` για τη μετατροπή των κατηγοριών (π.χ. `Rock`) σε αριθμητικές τιμές.
   - Το `CLASS_MAPPING` εφαρμόζεται για τη συγχώνευση ή αφαίρεση κατηγοριών, όπως απαιτείται.
4. **Μήκος και Padding:**
   - Δεδομένου ότι το μήκος κάθε φασματογραφήματος μπορεί να διαφέρει, εφαρμόζεται padding μέσω της κλάσης `PaddingTransform`, ώστε όλα τα δείγματα να έχουν το ίδιο μήκος (`max_length`).
   - Το padding συμπληρώνει με μηδενικά τα φασματογραφήματα που έχουν μικρότερο μήκος από το μέγιστο.
5. **Επιστροφή Δειγμάτων:**
   - Κάθε δείγμα επιστρέφει:
     (a) Το φασματογράφημα, κανονικοποιημένο και με padding.

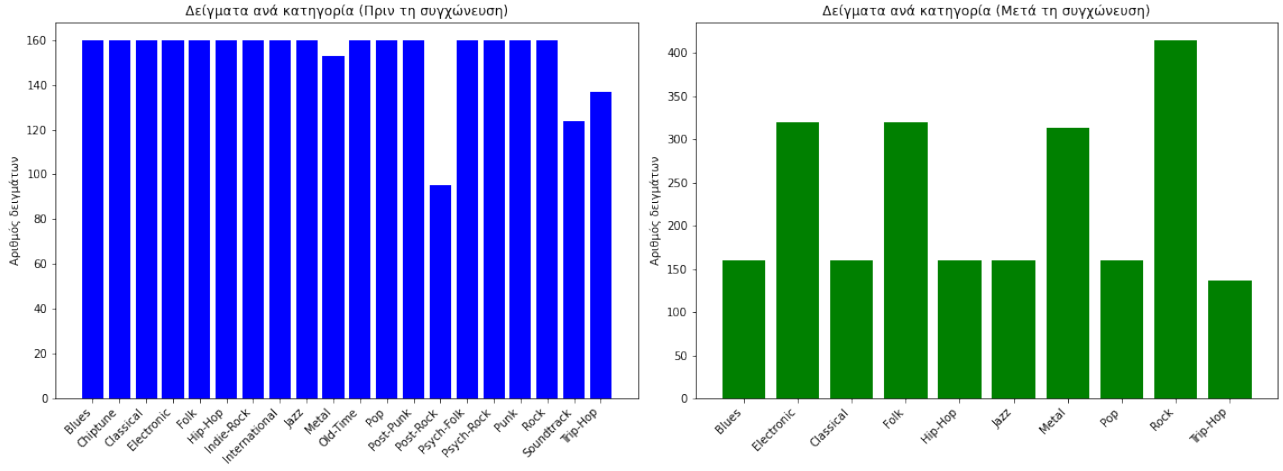
(b) Την αντίστοιχη κατηγορία (label) ως αριθμητική τιμή.
(c) Το αρχικό του μήκος (χρήσιμο για επεξεργασία από RNNs).

(β) Η διαδικασία συγχώνευσης/αφαίρεσης πραγματοποιείται μέσω του `CLASS_MAPPING`:

- **Συγχώνευση Κατηγοριών:** Παρόμοιες κατηγορίες (π.χ. `Rock` και `Psych-Rock`) συγχωνεύονται στην κατηγορία `Rock`.
- **Αφαίρεση Κατηγοριών:** Κατηγορίες που αντιστοιχούν σε λίγα δείγματα ή δεν είναι αντιπροσωπευτικές (π.χ. `Indie-Rock`, `Old-Time`) αφαιρούνται.

Η συγχώνευση και αφαίρεση γίνεται για την εξισορρόπηση των δεδομένων και την αποφυγή προβλημάτων *overfitting* σε κατηγορίες με λίγα δείγματα.

(γ) Τα δύο ιστογράμματα που θα δείχνουν πόσα δείγματα αντιστοιχούν σε κάθε κλάση, ένα πριν από τη διαδικασία του βήματος 4β και ένα μετά:



Από την παραπάνω ανάλυση διακρίνουμε τόσο τα πλεονεκτήματα όσο και τα μειονεκτήματα της διαδικασίας συγχώνευσης. Αρχικά, παρατηρούμε ότι ορισμένες κατηγορίες πλέον διαθέτουν πολύ περισσότερα δεδομένα για εκπαίδευση σε σύγκριση με άλλες. Παρόλα αυτά, ο διαχωρισμός τους είναι ευκολότερος σε σχέση με τις πιο εξειδικευμένες κατηγορίες που είχαμε προηγουμένως, καθώς αρκετές από αυτές παρουσιάζουν πολύ παρόμοια χαρακτηριστικά. Επομένως, είναι πιο αποτελεσματικό να τις ενοποιήσουμε σε μια πιο γενική κατηγορία.

Βήμα 5: Αναγνώριση μουσικού είδους με LSTM

(α) Για την υλοποίηση του LSTM χρησιμοποιούμε των κώδικα που φτιάξαμε στην προηγούμενη εργαστηριακή άσκηση. Ενεργοποιούμε επίσης την GPU θέτοντας:

```
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
```

Για την εκπαίδευση των μοντέλων χρησιμοποιούμε το αρχείο train.py από τον βοηθητικό κώδικα αυτής της εργαστηρικακής άσκησης.

(β) Συμπληρώνουμε την συνάρτηση overfit_with_a_couple_of_batches() του αρχείου train.py σύμφωνα με τις οδηγίες της εκφώνησης.

Εκπαιδεύουμε τα μοντέλα παρακάτω σύμφωνα με τις οδηγίες της εκφώνησης και στο Βήμα 6 παρουσιάζουμε τα αναλυτικά αποτελέμσματα για το test set.

(γ) Validation Accuracy: 0.3232

(δ) Validation Accuracy: 0.3492

(ε) Validation Accuracy: 0.2343

(ζ) Validation Accuracy: 0.3319

Βήμα 6: Αξιολόγηση των μοντέλων

Τι δείχνει το accuracy / precision / recall / F1 score;

- **Accuracy:** Δείχνει το ποσοστό των σωστά ταξινομημένων δειγμάτων σε σχέση με το συνολικό πλήθος των δειγμάτων. Χρησιμοποιείται ως γενικός δείκτης απόδοσης, αλλά δεν είναι ιδανικός σε προβλήματα με μη ισορροπημένες κλάσεις.
- **Precision:** Υπολογίζει το ποσοστό των σωστά ταξινομημένων δειγμάτων σε σχέση με τα δείγματα που προβλέφθηκαν στη συγκεκριμένη κλάση. Είναι χρήσιμο σε προβλήματα όπου θέλουμε να ελαχιστοποιήσουμε τα false positives.
- **Recall:** Υπολογίζει το ποσοστό των σωστά ταξινομημένων δειγμάτων σε σχέση με τα συνολικά δείγματα που ανήκουν στη συγκεκριμένη κλάση. Είναι σημαντικό σε προβλήματα όπου θέλουμε να ελαχιστοποιήσουμε τα false negatives.
- **F1 score:** Είναι ο αρμονικός μέσος του precision και του recall. Χρησιμοποιείται όταν θέλουμε μια ισορροπία μεταξύ αυτών των δύο μετρικών.

Τι δείχνει το micro / macro averaged precision / recall / F1 score;

- **Micro-averaged:** Υπολογίζει τις μετρικές λαμβάνοντας υπόψη το συνολικό πλήθος των true positives, false positives, και false negatives για όλες τις κλάσεις. Δίνει μεγαλύτερη βαρύτητα στις συχνότερες κλάσεις.
- **Macro-averaged:** Υπολογίζει τις μετρικές ξεχωριστά για κάθε κλάση και στη συνέχεια υπολογίζει τον μέσο όρο. Δίνει ίση βαρύτητα σε όλες τις κλάσεις, ανεξάρτητα από το μέγεθός τους.

Πότε προκύπτει μεγάλη απόκλιση ανάμεσα στο accuracy / F1 score και τι σημαίνει αυτό;

Η απόκλιση ανάμεσα στο accuracy και το F1 score προκύπτει συνήθως σε προβλήματα με μη ισορροπημένες κλάσεις. Για παράδειγμα:

- **Υψηλό accuracy, χαμηλό F1 score:** Σημαίνει ότι το μοντέλο ταξινομεί σωστά τις συχνότερες κλάσεις αλλά αποτυγχάνει να αναγνωρίσει τις σπάνιες.
- **Χρήση F1 score:** Σε αυτή την περίπτωση, το F1 score αποτελεί καλύτερη μετρική καθώς λαμβάνει υπόψη τόσο τα false positives όσο και τα false negatives.

Πότε προκύπτει μεγάλη απόκλιση ανάμεσα στο micro/macro F1 score και τι σημαίνει αυτό;

Η απόκλιση ανάμεσα στο micro και macro F1 score προκύπτει σε προβλήματα με μη ισορροπημένες κλάσεις.

- **Μεγάλο micro F1, μικρό macro F1:** Σημαίνει ότι το μοντέλο αποδίδει καλά στις συχνότερες κλάσεις αλλά αποτυγχάνει να ταξινομήσει σωστά τις σπανιότερες.
- **Μεγάλο macro F1:** Απαιτείται όταν η απόδοση στις σπάνιες κλάσεις είναι εξίσου σημαντική με τις συχνότερες.

Υπάρχουν προβλήματα όπου το precision μας ενδιαφέρει περισσότερο από το recall και αντίστροφα;

Ναι, υπάρχουν τέτοια προβλήματα:

- **Precision σημαντικότερο από Recall:** Σε προβλήματα όπου τα false positives είναι πιο κοστοβόρα, π.χ. σε recommendation systems.
- **Recall σημαντικότερο από Precision:** Σε προβλήματα όπου τα false negatives είναι πιο κρίσιμα, όπως στην ανίχνευση ασθενειών ή σε συστήματα ασφάλειας.

Είναι μια καλή τιμή accuracy / F1 αρκετή σε αυτές τις περιπτώσεις για να επιλέξω ένα μοντέλο;

Όχι απαραίτητα. Ανάλογα με το πρόβλημα, ίσως χρειάζεται να δώσουμε περισσότερη έμφαση στο precision ή το recall, και το accuracy ή το F1 score μόνο του δεν είναι αρκετό για την επιλογή του μοντέλου. Χρειάζεται πάντα να εξετάζουμε τις απαιτήσεις της εφαρμογής.

Παρακάτω φαίνονται τα αποτελέσματα που πήραμε στο test set για κάθε μοντέλο:

Table 1: Classification Report - Spectrogrmas

Class	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.51	0.47	0.49	40
2	0.33	0.57	0.42	80
3	0.33	0.35	0.34	80
4	0.25	0.28	0.26	40
5	0.00	0.00	0.00	40
6	0.33	0.50	0.40	78
7	0.00	0.00	0.00	40
8	0.28	0.39	0.33	103
9	0.00	0.00	0.00	34
Accuracy			0.32	
Macro Avg	0.20	0.26	0.22	575
Weighted Avg	0.24	0.32	0.27	575

Table 2: Classification Report - Beat-Synced Spectrograms

Class	Precision	Recall	F1-Score	Support
0	0.14	0.03	0.04	40
1	0.34	0.55	0.42	40
2	0.37	0.53	0.43	80
3	0.33	0.66	0.44	80
4	0.21	0.12	0.16	40
5	0.50	0.03	0.05	40
6	0.45	0.59	0.51	78
7	0.00	0.00	0.00	40
8	0.35	0.28	0.31	103
9	0.00	0.00	0.00	34
Accuracy			0.34	
Macro Avg	0.26	0.27	0.23	575
Weighted Avg	0.30	0.34	0.29	575

Table 3: Classification Report - Chromograms

Class	Precision	Recall	F1-Score	Support
0	0.20	0.03	0.04	40
1	0.12	0.07	0.09	40
2	0.17	0.25	0.20	80
3	0.23	0.41	0.30	80
4	0.43	0.15	0.22	40
5	0.27	0.07	0.12	40
6	0.32	0.38	0.35	78
7	0.00	0.00	0.00	40
8	0.23	0.31	0.26	103
9	0.07	0.06	0.06	34
Accuracy			0.23	
Macro Avg	0.21	0.17	0.17	575
Weighted Avg	0.22	0.23	0.20	575

Table 4: Classification Report - Spectrograms+Chromograms

Class	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	40
1	0.44	0.45	0.44	40
2	0.27	0.61	0.38	80
3	0.33	0.59	0.42	80
4	0.00	0.00	0.00	40
5	0.12	0.03	0.04	40
6	0.40	0.68	0.50	78
7	0.00	0.00	0.00	40
8	0.26	0.17	0.21	103
9	0.00	0.00	0.00	34
Accuracy			0.32	
Macro Avg	0.18	0.25	0.20	575
Weighted Avg	0.22	0.32	0.25	575

Από τα αποτελέσματα στους παραπάνω πίνακες, προκύπτουν οι εξής παρατηρήσεις:

- Το **μοντέλο που χρησιμοποιεί μόνο τα chromograms** παρουσιάζει τις χαμηλότερες επιδόσεις. Η ακρίβεια (accuracy) είναι μόλις 23%, ενώ και οι τιμές F1-Score για τις περισσότερες κλάσεις είναι πολύ χαμηλές. Αυτό πιθανώς οφείλεται στο γεγονός ότι τα chromograms δεν διατηρούν επαρκώς πληροφορίες για τη χρονική δυναμική του σήματος, που είναι σημαντική για πολλές κατηγορίες.

- Το **μοντέλο που χρησιμοποιεί τα beat-synced spectrograms** έχει τις καλύτερες συνολικές επιδόσεις με ακρίβεια 34% και σχετικά υψηλότερα F1-Scores για αρκετές κλάσεις. Αυτό πιθανώς συμβαίνει επειδή τα beat-synced spectrograms διατηρούν πληροφορίες χρονικού συγχρονισμού με το ρυθμό, ενώ ταυτόχρονα μειώνουν τον θόρυβο που δεν ευθυγραμμίζεται με τον ρυθμό.

- Το **μοντέλο με συνδυασμό spectrograms και chromograms** έχει ακρίβεια 32%. Παρόλο που θεωρητικά θα έπρεπε να αξιοποιεί πληροφορίες τόσο χρονικές όσο και αρμονικές, τα αποτελέσματα δείχνουν ότι ο συνδυασμός αυτός δεν βελτιώνει σημαντικά την απόδοση. Ίσως απαιτείται περαιτέρω επεξεργασία ή βελτιστοποίηση χαρακτηριστικών.

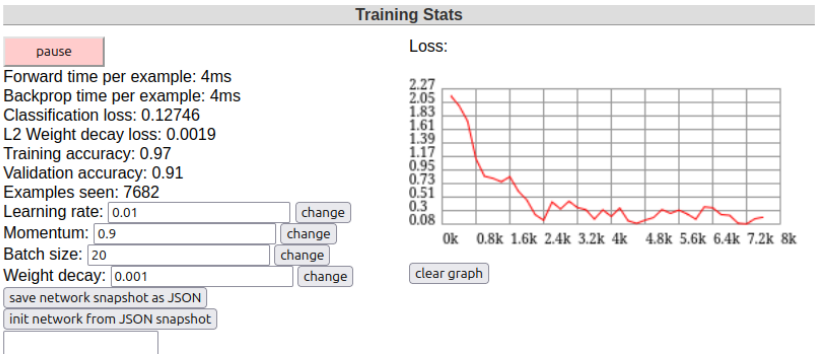
- Σχετικά με τις επιμέρους κλάσεις, οι συχνότερες κλάσεις στο dataset, όπως η κλάση 6, παρουσιάζουν τις καλύτερες επιδόσεις (π.χ., F1-Score 0.50 για τα beat-synced spectrograms). Αντίθετα, οι υποεκπροσωπούμενες κλάσεις, όπως οι κλάσεις 5 και 7, έχουν πολύ χαμηλά F1-Scores, ανεξάρτητα από τη μέθοδο εισόδου.

- Οι διαφορές μεταξύ **Macro Avg** και **Weighted Avg** δεν είναι ιδιαίτερα μεγάλες, γεγονός που υποδηλώνει ότι οι προβλέψεις του μοντέλου είναι σχετικά ισορροπημένες, λαμβάνοντας υπόψη την ανισοκατανομή του dataset.
- Τέλος, παρατηρούμε ότι τα spectrograms από μόνα τους δίνουν παρόμοια αποτελέσματα με τα beat-synced spectrograms, αλλά τα τελευταία υπερέχουν λόγω της ενσωμάτωσης πληροφοριών χρονικού συγχρονισμού, που φαίνεται να είναι κρίσιμες για την κατηγοριοποίηση.

Κυρίως Μέρος

Βήμα 7.1: 2D CNN

α) Παρακάτω βλέπουπε τις υπερπαραμέτρους του δικτύου και πως συμπεριφέρεται το training loss με την πάροδο του χρόνου. Παρατηρούμε ότι μετά από λίγες εποχές το training loss μειώνεται σημαντικά αφού το δίκτυο έχει μάθει πια να ταξινομεί τα ψηφία.



Παρακάτω βλέπουμε την αρχιτεκτονική του δικτύου:

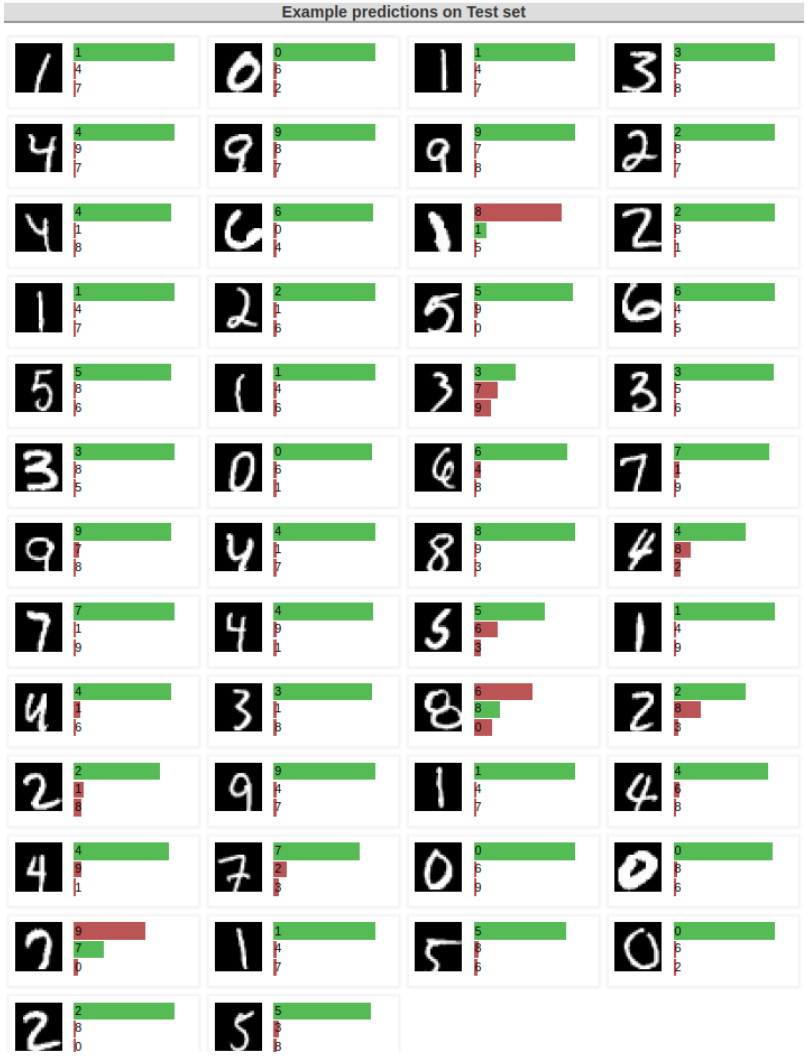
```
layer_defs = [];
layer_defs.push({type:'input', out_sx:24, out_sy:24, out_depth:1});
layer_defs.push({type:'conv', sx:5, filters:8, stride:1, pad:2, activation:'relu'});
layer_defs.push({type:'pool', sx:2, stride:2});
layer_defs.push({type:'conv', sx:5, filters:16, stride:1, pad:2, activation:'relu'});
layer_defs.push({type:'pool', sx:3, stride:3});
layer_defs.push({type:'softmax', num_classes:10});

net = new convnetjs.Net();
net.makeLayers(layer_defs);

trainer = new convnetjs.SGDTrainer(net, {method:'adadelta', batch_size:20, l2_decay:0.001});
```

Παρακάτω βλέπουμε τη διαδικασία επεξεργασίας εικόνας ενός ψηφίου MNIST από CNN. Περιλαμβάνει επίπεδα όπως convolutional (για εξαγωγή χαρακτηριστικών), ReLU (για μη γραμμικότητα), pooling (για μείωση διαστάσεων) και fully connected (για κατηγοριοποίηση). Κάθε επίπεδο εμφανίζει ενεργοποιήσεις, βαθμίδες και βάρη, επιτρέποντας την ανάλυση της επίδρασης κάθε στρώσης. Βλέπουμε σε κάθε layer την επίδραση που έχουν τα φίλτρα και πως μετασχηματίζεται η εικόνα μετά την εφαρμογή αυτών. Το softmax στο τέλος προβλέπει την πιθανότητα της σωστής κατηγορίας, και η ενεργοποίηση υποδεικνύει σωστή αναγνώριση.

Παρακάτω βλέπουμε τις προβλέψεις του μοντέλου για διάφορες εισόδους. Παρατηρούμε ότι κατά κύριο λόγο το μοντέλο ταξινομεί σωστά σε μεγάλο ποσοστό τα δείγματα με εξαίρεση 3 ψηφία μόνο τα οποία ταξινομεί λάθος.



β) Υλοποιούμε το CNN σύμφωνα με τις οδηγίες της εκφώνησης. Οι παράμετροι που επιλέξαμε στο CNN είναι:

```
input_dims = torch.Size([150, 128])
in_channels = 1
filters = [32, 64, 128, 256]
feature_size = 1000
```

Επίσης στα 2 πρώτα επίπεδα τα φίλτρα έχουν διαστάσεις 5×5 με stride=2 και στα υπόλοιπα 2 επίπεδα τα φίλτρα έχουν διαστάσεις 3×3 με stride=2

γ)

- **Convolutions:** Οι convolutional layers αναγνωρίζουν τοπικά μοτίβα σε εικόνες ή άλλα δεδομένα εισόδου, εφαρμόζοντας φίλτρα σε μικρά τμήματα της εισόδου. Αυτές οι λειτουργίες βοηθούν το μοντέλο να ανιχνεύσει χαρακτηριστικά όπως άκρα, υφές και σχήματα σε διάφορες κλίμακες.
- **Batch Normalization:** Το Batch Normalization είναι μια τεχνική που κανονικοποιεί τα δεδομένα εισόδου σε κάθε επίπεδο του νευρωνικού δικτύου, με σκοπό την επιτάχυνση της εκπαίδευσης και τη βελτίωση της σταθερότητας. Συγκεκριμένα, η κανονικοποίηση πραγματοποιείται για κάθε batch δεδομένων, υπολογίζοντας τον μέσο όρο και την τυπική απόκλιση των ενεργοποιήσεων για κάθε χαρακτηριστικό. Οι ενεργοποιήσεις κανονικοποιούνται αφαιρώντας τον μέσο όρο και διαιρώντας με την τυπική απόκλιση. Στη συνέχεια, εφαρμόζονται δύο παραμέτρους, την κλίμακα (γ) και τη μετατόπιση (β), οι οποίες επιτρέπουν στο μοντέλο να ανακτήσει την αρχική διανομή των δεδομένων,

αν χρειάζεται. Η κανονικοποίηση αυτή μειώνει την ενδοδικτυακή εξάρτηση από την αρχικοποίηση των παραμέτρων και επιτρέπει τη χρήση μεγαλύτερων ρυθμών εκπαίδευσης, ελαχιστοποιώντας τα προβλήματα vanishing και exploding gradients.

- **ReLU (Rectified Linear Unit):** Η ReLU είναι μια μη γραμμική συνάρτηση ενεργοποίησης που δίνει θετική έξοδο για θετικές τιμές εισόδου και μηδενίζει τις αρνητικές τιμές. Αυτό επιτρέπει στο μοντέλο να μάθει πιο σύνθετα και μη γραμμικά μοτίβα, ενώ ταυτόχρονα αποτρέπει το φαινόμενο vanishing gradients.

- **Max Pooling:** Η μέγιστη υποδειγματοληψία (Max Pooling) είναι μια τεχνική που μειώνει τη διάσταση των χαρακτηριστικών, κρατώντας το μέγιστο στοιχείο από κάθε τμήμα του χαρακτηριστικού χάρτη. Αυτό βοηθά στη μείωση της υπολογιστικής πολυπλοκότητας και στην εξάλειψη περιττών λεπτομερειών, ενισχύοντας τη γενίκευση του μοντέλου.

δ) Πραγρατοποιούμε την διαδικασία overfit_batch, όπως αναφέρεται και στην εκφώνηση
ε) Validation Accuracy: 0.4208

Τα αποτελέσματα στο test set:

Table 5: CNN for spectrograms

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.26	0.25	0.26	40
1	0.39	0.60	0.48	40
2	0.59	0.42	0.49	80
3	0.40	0.50	0.44	80
4	0.41	0.60	0.48	40
5	0.12	0.05	0.07	40
6	0.43	0.72	0.54	78
7	0.09	0.03	0.04	40
8	0.36	0.30	0.33	103
9	0.36	0.15	0.21	34
Accuracy		0.39		575
Macro Avg	0.34	0.36	0.33	575
Weighted Avg	0.37	0.39	0.37	575

Παρατηρούμε τα εξής:

- Η συνολική **ακρίβεια** (accuracy) είναι 39%, κάτι που αποτελεί βελτίωση σε σχέση με προηγούμενα μοντέλα. Ωστόσο, παραμένει χαμηλή για ένα πρόβλημα πολυκλασικής ταξινόμησης, δείχνοντας περιθώρια βελτίωσης.

- Οι κλάσεις με μεγαλύτερη εκπροσώπηση (support), όπως η κλάση 6, έχουν την καλύτερη απόδοση. Συγκεκριμένα, η κλάση 6 έχει F1-Score 0.54 και υψηλό Recall (0.72), γεγονός που υποδεικνύει ότι το μοντέλο αναγνωρίζει αρκετά καλά τις περισσότερες περιπτώσεις αυτής της κλάσης.

- Οι κλάσεις με χαμηλότερη εκπροσώπηση, όπως οι κλάσεις 5 και 7, παρουσιάζουν πολύ χαμηλές επιδόσεις, με F1-Scores μόλις 0.07 και 0.04 αντίστοιχα. Αυτό πιθανώς οφείλεται στην περιορισμένη εκπροσώπησή τους στο dataset.

- Οι διαφορές μεταξύ των **Macro Avg** (Precision: 0.34, Recall: 0.36, F1-Score: 0.33) και **Weighted Avg** (Precision: 0.37, Recall: 0.39, F1-Score: 0.37) είναι μικρές, γεγονός που υποδηλώνει ότι το μοντέλο αντιμετωπίζει τις κλάσεις με σχετική ισορροπία, παρά την ανισοκατανομή.

- Οι κλάσεις 1, 3, και 4 έχουν μέτρια απόδοση, με F1-Scores κοντά στο 0.44-0.48. Αυτό υποδεικνύει ότι το μοντέλο μπορεί να τις διαχειριστεί σχετικά καλά, αλλά χρειάζονται περαιτέρω βελτιώσεις.

Βήμα 7.2: AST - Audio Spectrogram Transformer

α) Οι αρχιτεκτονικές Transformer, όπως το "Audio Spectrogram Transformer" (AST), έχουν προσαρμοστεί για την επεξεργασία ηχητικών σημάτων εκμεταλλευόμενες την ικανότητά τους να συλλαμβάνουν μακροχρόνιες εξαρτήσεις και να αναλύουν πολύπλοκα μοτίβα. Το AST αποτελεί μια ιδιαίτερη υλοποίηση που βασίζεται στο Vision Transformer (ViT), αλλά προσαρμόζεται για την ανάλυση ακουστικών δεδομένων.
Το αρχικό ηχητικό σήμα μετατρέπεται πρώτα σε spectrogram. Το spectrogram αναπαριστά τη χρονική εξέλιξη του σήματος στις διαφορετικές συχνότητες, παρέχοντας μια δισδιάστατη απεικόνιση που μοιάζει με εικόνα. Το spectrogram χωρίζεται σε μικρότερα μη επικαλυπτόμενα τμήματα (patches), όπως γίνεται και στις εικόνες στο ViT. Αυτά τα patches μετατρέπονται σε διανύσματα (embeddings) μέσω γραμμικών μετασχηματισμών. Δεδομένου ότι τα spectrogram patches δεν περιέχουν πληροφορία θέσης, προστίθενται *positional encodings* στα embeddings, ώστε το μοντέλο να αναγνωρίζει τη σειρά και τη διάταξη των patches. Το επεξεργασμένο spectrogram περνά μέσα από πολλαπλά επίπεδα ενός *Transformer Encoder*. Κάθε επίπεδο αποτελείται από:

- **Multi-Head Self-Attention**: Συλλαμβάνει τις συσχετίσεις ανάμεσα στα patches, ακόμα και αν βρίσκονται μακριά το ένα από το άλλο.
- **Feed-Forward Layers**: Μετασχηματίζουν τις πληροφορίες των patches σε υψηλότερης διάστασης χαρακτηριστικά.
- **Residual Connections και Layer Normalization**: Βοηθούν στη σταθερότητα της εκπαίδευσης.

β) Θέτουμε τις παραμέτρους:

```
input_fdim = 128
input_tdim = MAX_LEN
feature_size = 1000
```

Validation Accuracy: 0.4664

Τα αποτελέσματα στο test set:

Table 6: AST for spectrograms

Class	Precision	Recall	F1-Score	Support
0	0.31	0.25	0.28	40
1	0.49	0.55	0.52	40
2	0.57	0.59	0.58	80
3	0.43	0.45	0.44	80
4	0.40	0.45	0.42	40
5	0.15	0.15	0.15	40
6	0.60	0.56	0.58	78
7	0.24	0.15	0.18	40
8	0.41	0.45	0.43	103
9	0.26	0.29	0.28	34
Accuracy		0.43		575
Macro Avg	0.39	0.39	0.39	575
Weighted Avg	0.42	0.43	0.43	575

Παρατηρούμε τα εξής:

- Η συνολική **ακρίβεια** (accuracy) ανέρχεται στο 43%, το οποίο, αν και δεν είναι ιδιαίτερα υψηλό, υποδεικνύει βελτίωση σε σχέση με προηγούμενα αποτελέσματα.
- Οι κλάσεις με μεγαλύτερη εκπροσώπηση (support), όπως οι κλάσεις 2, 3, 6 και 8, έχουν καλύτερη απόδοση σε σχέση με άλλες κλάσεις, με τις τιμές F1-Score να κυμαίνονται από 0.43 έως 0.58. Αυτό είναι αναμενόμενο, δεδομένου ότι οι περισσότερες παρατηρήσεις βρίσκονται σε αυτές τις κλάσεις.

- Οι κλάσεις με λιγότερη εκπροσώπηση, όπως οι κλάσεις 0, 5, 7, και 9, παρουσιάζουν σημαντικά χαμηλότερες επιδόσεις. Οι χαμηλές τιμές F1-Score για αυτές τις κλάσεις (π.χ. 0.15 για την κλάση 5) πιθανώς οφείλονται στην ανισοκατανομή του dataset.
- Η μικρή διαφορά μεταξύ των **Macro Avg** και **Weighted Avg** (και τα δύο περίπου στο 0.39-0.43) υποδεικνύει ότι το μοντέλο αντιμετωπίζει σχετικά ισορροπημένα όλες τις κλάσεις, παρόλο που υπάρχει ασυμμετρία στην κατανομή του dataset.

Βήμα 8: Εκτίμηση συναισθήματος - συμπεριφοράς με παλινδρόμηση

Σε αυτό το βήμα χρησομοποιούμε τα RNNBackbone(), CNNBackbone() και ASTBackbone() όμως αντί τον Classifier() χρησιμοποιούμε τον Regressor(). Χρησιμοποιούμε ως συνάρτηση κόστους τώρα την nn.MSELoss().

Τα αποτελέσματα στο test set:

Model	Task	MAE	MSE	Spearman Correlation
CNN	Danceability	0.1458	0.0329	0.1141
RNN	Danceability	0.1412	0.0311	0.1118
AST	Danceability	0.1353	0.0284	0.1400
CNN	Valence	0.2309	0.0745	0.0960
RNN	Valence	0.2247	0.0683	0.0520
AST	Valence	0.2241	0.0674	0.1138
CNN	Energy	0.2308	0.0764	0.1328
RNN	Energy	0.2240	0.0713	0.1416
AST	Energy	0.2204	0.0692	0.2533

Table 7: Test Metrics for CNN, RNN, and AST Models across Danceability, Valence, and Energy tasks.

Danceability

Το **AST μοντέλο** ξεπερνά σε επίδοση τα CNN και RNN, πετυχαίνοντας το χαμηλότερο MSE Loss (0.0284) και τον υψηλότερο συντελεστή Spearman Correlation (0.1400). Αυτό υποδεικνύει ότι το AST είναι το πιο αποτελεσματικό για την πρόβλεψη της "danceability".
Το **RNN μοντέλο** και το **CNN μοντελο** δείχνουν να έχουν παρόμοια απόδοση, δηλαδή Spearman Correlation (0.1118 και 0.1141 αντίστοιχα), με χαμηλότερο MSE Loss (0.311 και 0.0329 αντίστοιχα).

Valence

Το **AST μοντέλο** αποδίδει καλύτερα συνολικά, με το χαμηλότερο MSE Loss (0.0674) και τον υψηλότερο Spearman Correlation (0.1138). Αυτό δείχνει ότι το AST αποδίδει καλύτερα στα χαρακτηριστικά "valence".
Το **RNN μοντέλο** έχει παρόμοιο Loss με το RNN (0.0683), αλλά ο Spearman Correlation (0.0520) είναι χαμηλότερος, υποδεικνύοντας χαμηλότερη συνάφεια.
Το **CNN μοντέλο** έχει το υψηλότερο Loss (0.0745) και την δεύτερη καλύτερη επίδοση, δηλαδή Spearman Correlation (0.0960).

Energy

Το **AST μοντέλο** είναι ο ξεκάθαρος νικητής, με το χαμηλότερο Loss (0.0692) και τον υψηλότερο Spearman Correlation (0.2533), που είναι σημαντικά ανώτερο από τα υπόλοιπα μοντέλα.
Το **RNN μοντέλο** παρουσιάζει μέτρια απόδοση, με MSE Loss 0.713 και Spearman Correlation 0.1416. Αποδίδει καλύτερα από το CNN αλλά χειρότερα από το AST.

Το **CNN μοντέλο** αποδίδει χειρότερα από τα άλλα μοντέλα στην πρόβλεψη "energy", με το υψηλότερο MSE Loss (0.0764) και τον μικρότερο Spearman Correlation (0.1328).

Γενικές Παρατηρήσεις

Το **AST μοντέλο** παρουσιάζει σταθερά την καλύτερη απόδοση σε όλες τις εργασίες, με τους υψηλότερους συντελεστές Spearman Correlation, επιδεικνύοντας την ισχυρή προβλεπτική του ικανότητα.
Το **μοντέλο RNN** και το **μοντέλο CNN** ακολουθούν σε επιδόσεις, με το CNN να υπερέχει ελαφρώς, κυρίως λόγω της καλύτερης απόδοσής του στο task "valence".

Βήμα 9: Μεταφορά γνώσης (Transfer Learning)

α) Το paper αυτό διερευνά το transferability των χαρακτηριστικών σε deep CNN και καταλήγει ότι το specialization των ανώτερων επιπέδων μειώνει το transferability, ενώ οι δυσκολίες βελτιστοποίησης λόγω εξαρτημένων νευρώνων μπορούν επίσης να επιδεινώσουν την απόδοση. Παρά τις δυσκολίες, η χρήση transferable χαρακτηριστικών, ακόμη και από distant tasks, συνήθως υπερτερεί έναντι τυχαίων αρχικοποιήσεων, βελτιώνοντας τη γενίκευση.

β) Διαλέγουμε το AST από το βήμα 7.2, καθώς υπερτερούσε των υπόλοιπων μοντέλων σε όλες τις μετρικές (accuracu, precision, recall, f1 score, macro Avg, micro Avg). Επίσης και στο Βήμα 8, στα regression tasks πάλι το AST είχε την καλύτερ επίδοση, όποτε είναι λογικό να θεωρήσουμε ότι το AST θα είναι και το καλύτερο μοντέλο όταν θα κάνουμε transfer learning.

γ) Χρησιμοποιούμε το καλύτερο μοντέλο που εκπαιδεύσαμε στο βήμα 7.2, εφαρμόζοντάς το στο dataset fma_genre_spectrograms. Παρόλο που το accuracy δεν αποτελεί την ιδανική μετρική για ένα μη ισορροπημένο dataset, δεδομένου ότι οι τιμές του macro και weighted average score δεν παρουσιάζουν σημαντικές αποκλίσεις, μπορούμε να βασιστούμε σε αυτήν τη μετρική.

δ) Φορτώνουμε το μοντέλο και αντικαθιστούμε το Classifier() με το Regressor(). Επιλέγουμε να εκπαιδεύσουμε όλα τα βάρη σε όλα τα layers του μοντέλου. Ουσιαστικά με αυτόν τον τρόπο πετυχαίνουμε μία καλύτερη αρχικοποίηση των βαρών του ASTBackbone().

ε) Τα αποτελέσματα που πήραμε στο test set:

Model	Task	MAE	MSE	Spearman Correlation
AST	Valence	0.2241	0.0674	0.1138
AST with transfer Learning	Valence	0.2232	0.0670	0.3727

Παρατηρούμε ότι η χρήση του transfer learning βελτιώνει σημαντικά την επίδοση του μοντέλου, όπως φαίνεται από την αύξηση του Spearman Correlation στο task "Valence", από 0.1138 γίνεται 0.3727. Αυτή η βελτίωση υποδεικνύει ότι η προσθήκη γνώσης από προϋπάρχουσες εκπαιδευμένες παραμέτρους βοηθά το μοντέλο να γενικεύσει καλύτερα και να κατανοήσει καλύτερα τις σχέσεις ανάμεσα στις κλάσεις του συγκεκριμένου task και ότι τα representaions των spectrograms για το classification task (Βήμα 5) είναι μάλλον καλύτερα από αυτά του regression task (Βήμα 8).

Βήμα 10: Εκπαίδευση σε πολλαπλά προβλήματα (Multitask Learning)

α) Το paper "One Model To Learn Them All" αναδεικνύει την ικανότητα ενός νευρωνικού δικτύου να πετυχαίνει αρκετά καλή επίδοση σε μία ποικιλία από tasks τα οποία ανήκουν και σε διαφορετικά domains όπως vision, language και speech, χρησιμοποιώντας πολλά διαφορετικά είδη building blocks όπως convolutional layers, attention mechanisms και sparsely-gated layers. Φαίνεται επίσης ότι ακόμη και αν ένα block δεν είναι κρίσιμο για κάποιο task αυτό δεν επηρεάζει αρνητικά σχεδόν καθόλου την επίδοση στο task.

β), γ) Πραγματοποιούμε την εκπαίδευση σύμφωνα με τον τρόπο που αναφέρεται στην εκφώνηση. Θέτουμε το βάρος για το loss κάθε task ως trainable parameter στο Multitask Model που κατασκευάσαμε.

δ) Τα αποτελέσματα που πήραμε στο test set:

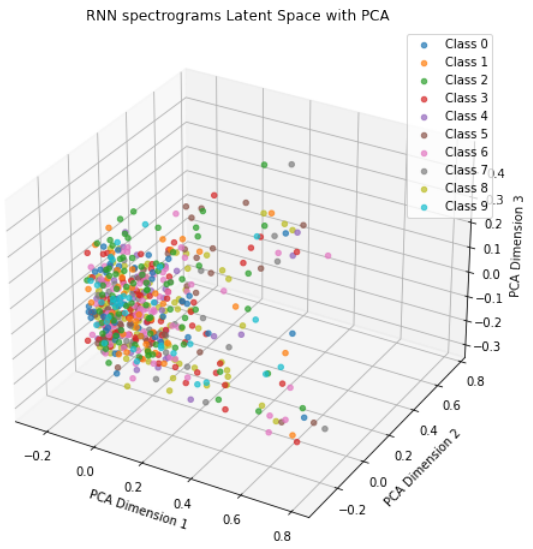
Model	**Task**	**MAE**	**MSE**	**Spearman Correlation**
AST	Valence	0.2241	0.0674	0.1138
AST	Energy	0.2204	0.0692	0.2533
AST	Danceability	0.1353	0.0284	0.1400
AST multitasking	Valence	0.2232	0.0673	0.1425
AST multitasking	Energy	0.2225	0.0681	0.3960
AST multitasking	Danceability	0.1352	0.0285	0.4598

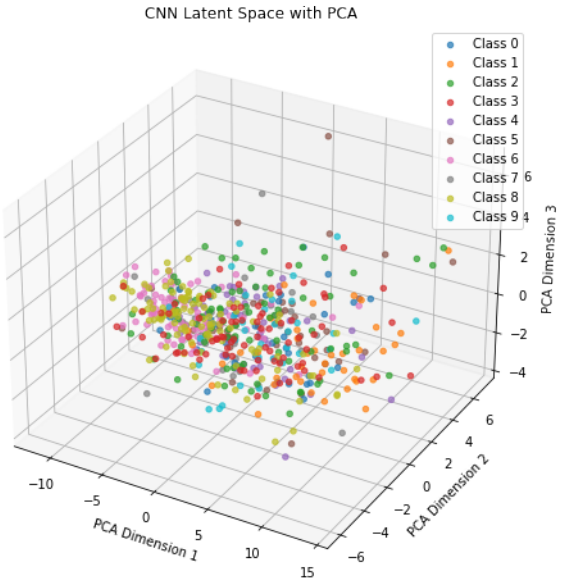
Τα αποτελέσματα δείχνουν ότι το μοντέλο AST multitasking υπερτερεί του απλού AST σε όλες σχεδόν τις μετρικές, ειδικά όσον αφορά τον Spearman Correlation. Συγκεκριμένα, στο task "Energy" και "Danceability", η πολυδιάστατη προσέγγιση του multitasking AST βελτιώνει σημαντικά τη συσχέτιση, με τιμές 0.3960 και 0.4598 αντίστοιχα, έναντι 0.2533 και 0.1400 για το μοντέλο AST χωρίς multitasking. Η βελτίωση που παρατηρείται στις μετρικές του multitasking AST, ειδικά στο "Valence", δείχνει ότι οι κοινές αναπαραστάσεις δεν θυσιάζουν την απόδοση σε μεμονωμένα tasks, ενώ παράλληλα ενισχύουν τη γενικότερη απόδοση του μοντέλου σε επίπεδο συσχέτισης. Συνολικά, τα αποτελέσματα υποστηρίζουν τη χρήση multitasking προσεγγίσεων για σύνθετα προβλήματα που εμπλέκουν συσχετισμένα tasks.

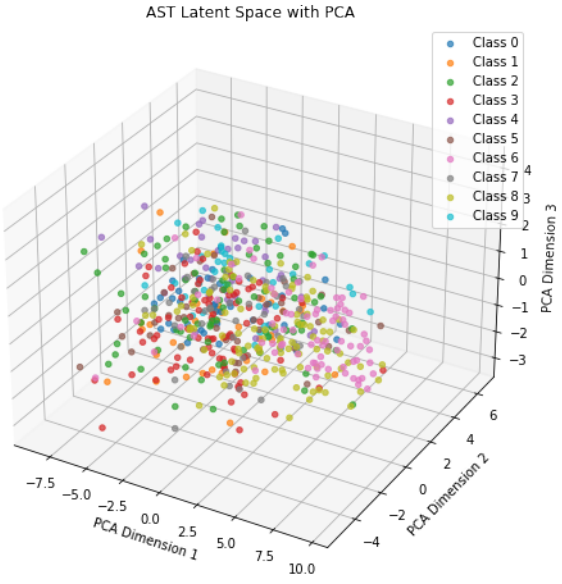
Βήμα 11 (Προαιρετικό): Οπτικοποίηση κρυφών αναπαραστάσεων

α) Για κάθε μοντέλο εξάγουμε το latent representation του κάθε τραγουδιού του test set (όπως αυτό εξάγεται από το τελευταίο layer του νευρωνικού πριν το output_layer) και από το αντίστοιχο label.

β) Παρακάτω παίρνουμε τα αποτελέσματα για τις οπτικοποιήσεις των latent representations για τα μοντέλα που παίρνουν ως είσοδο spectrograms:



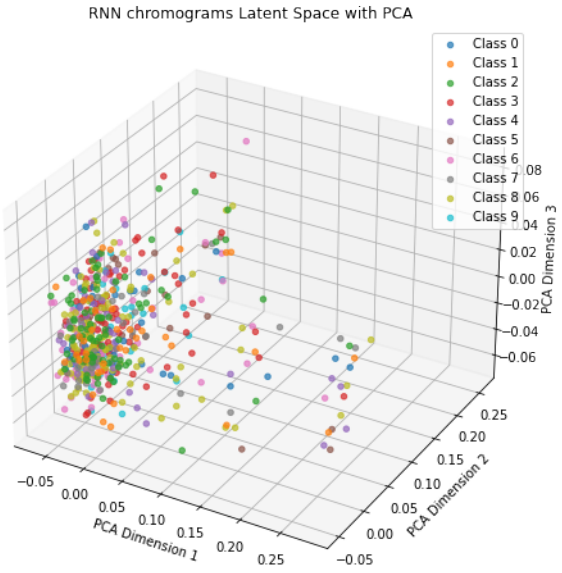




Παρατηρούμε ότι στον latent space του AST οι κλάσεις εμφανίζονται πιο σαφώς διαχωρίσιμες, υποδεικνύοντας καλύτερη απόδοση στη μάθηση χαρακτηριστικών. Αντίθετα, ο latent space του RNN παρουσιάζει τη χειρότερη επίδοση, με λιγότερο ευδιάκριτο διαχωρισμό μεταξύ των κλάσεων. Ο latent space του CNN είναι κάπως βελτιωμένος σε σύγκριση με αυτόν του RNN, αλλά εξακολουθεί να υπολείπεται σε ποιότητα σε σχέση με τον AST.

γ) Παρακάτω φαίνονται τα αποτελέσματα για το χειρότερο μοντέλο του ερωτήματος 5, δηλαδή αυτό που

παίρνει ως είσοδο μόνο τα chromograms.

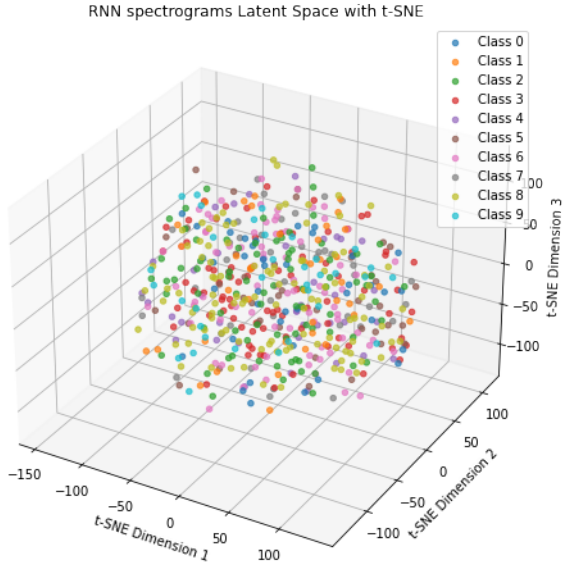


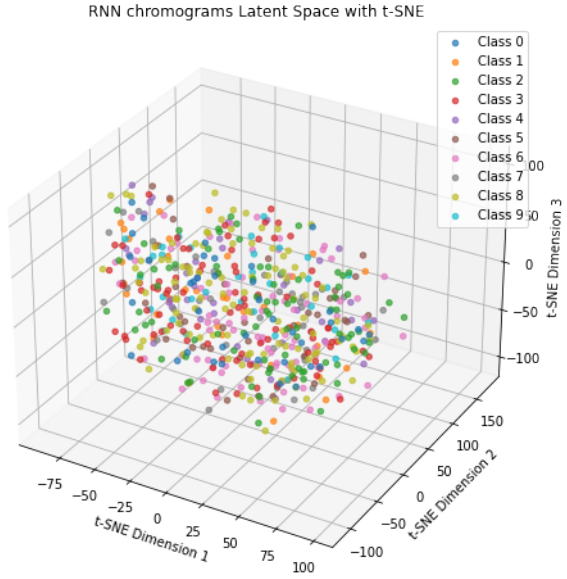
Σε σύγκριση με τα προηγούμενα μοντέλα φαίνεται ότι οι κλάσεις είναι πιο δυσδιάκριτες στο χώρο. Βλέπουμε επίσης ότι αυτός ο latent space μοιάζει με τον latent space του RNN του προηγούμενου ερωτήματος.

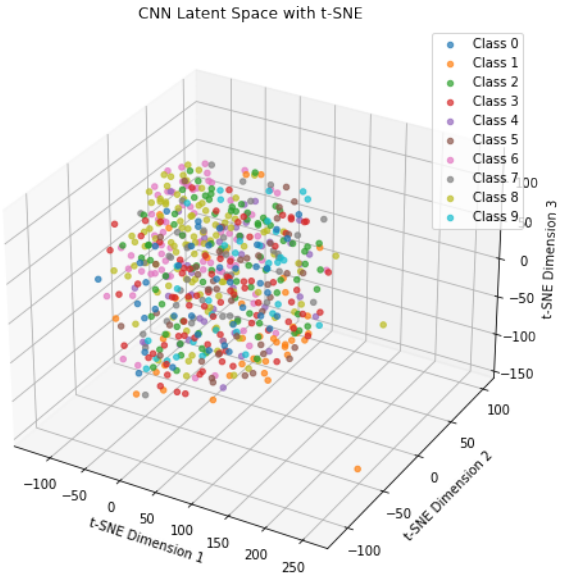
δ) Η μείωση διαστάσεων είναι κρίσιμη για την κατανόηση των χωρών χαρακτηριστικών υψηλών διαστάσεων, όπως τα latent representations από νευρωνικά δίκτυα. Η PCA προσφέρει μια βασική γραμμική μέθοδο που αναδεικνύει την κατανομή της διασποράς, αλλά δυσκολεύεται να αποκαλύψει μη γραμμικές σχέσεις. Η t-SNE και η UMAP είναι πιο εξελιγμένοι αλγόριθμοι που μπορούν να διατηρήσουν τοπικές σχέσεις ή να αναδείξουν clusters. Η t-SNE ειδικεύεται στη δημιουργία οπτικών διαχωρισμών τοπικών συστάδων αλλά είναι υπολογιστικά απαιτητική, ενώ η UMAP προσφέρει ταχύτερη και πιο ολοκληρωμένη οπτικοποίηση, διατηρώντας τόσο την τοπική όσο και τη γενική δομή δεδομένων.

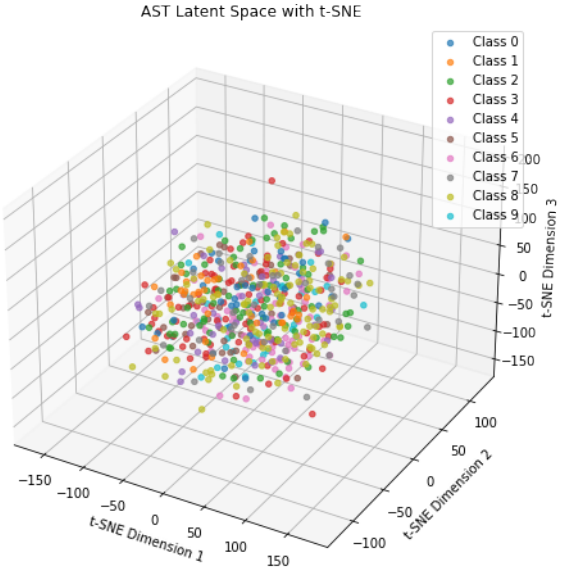
Υποψιαζόμαστε ότι οι UMAP και t-SNE θα δώσουν καλύτερες οπτικοποιήσιες καθώς όπως είδαμε και στα προηγούμενα ερωτήματα τα latent represenation των genres σε κανένα μοντέλο δεν είναι γραμμικά διαχωρίσιμα κι έτσι είναι λογκό να καταφύγουμε σε πιο εξελιγμένες μεθόδους.

Χρησιμοποιηώντας τον αλγόριθμο t-SNE έχουμε:

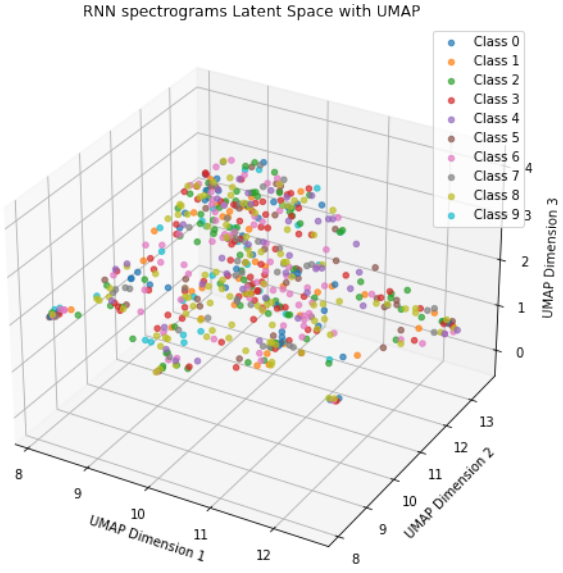

RNN spectrograms Latent Space with t-SNE
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
t-SNE Dimension 1
t-SNE Dimension 2
t-SNE Dimension 3

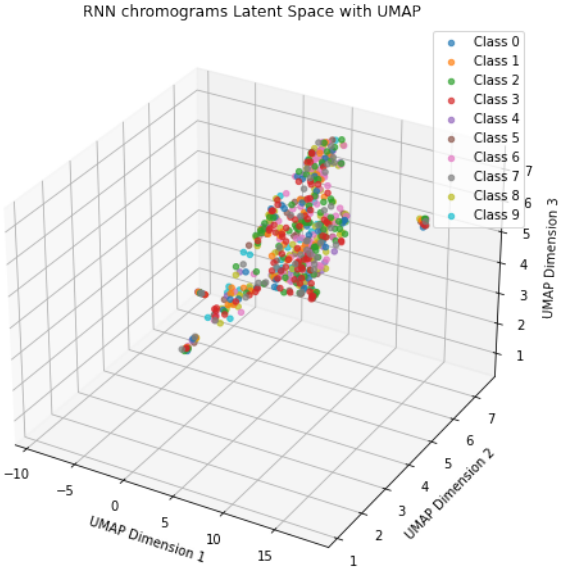

RNN chromograms Latent Space with t-SNE
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
t-SNE Dimension 1
t-SNE Dimension 2
t-SNE Dimension 3

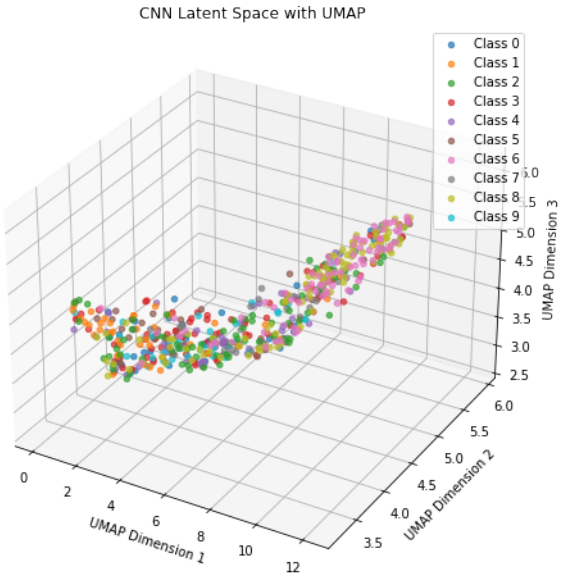


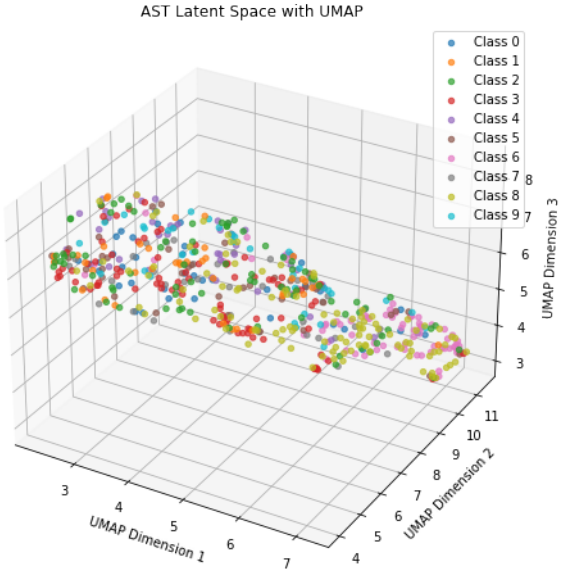


Χρησιμοποιώντας τον αλγόριθμο UMAP έχουμε:


RNN spectrograms Latent Space with UMAP
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
UMAP Dimension 1
UMAP Dimension 2
UMAP Dimension 3


RNN chromograms Latent Space with UMAP
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
Class 9
UMAP Dimension 1
UMAP Dimension 2
UMAP Dimension 3





Παρατηρούμε ότι οι αλγόριθμοι t-SNE και UMAP βελτιώνουν σημαντικά την οπτικοποίηση του latent space των μοντέλων, με τον UMAP να ξεχωρίζει, προσφέροντας τις πιο ποιοτικές και καθαρές απεικονίσεις.