

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
«ΑΝΑΠΤΥΞΗ ΕΦΑΡΜΟΓΗΣ ΑΝΑΓΝΩΡΙΣΗΣ
ΣΥΝΑΙΣΘΗΜΑΤΩΝ ΑΠΟ ΕΚΦΡΑΣΕΙΣ ΠΡΟΣΩΠΟΥ
ΚΑΙ ΠΡΟΣΑΡΜΟΓΗΣ ΜΟΥΣΙΚΟΥ
ΠΕΡΙΕΧΟΜΕΝΟΥ ΣΕ ΚΙΝΗΤΕΣ ΣΥΣΚΕΥΕΣ»



Του φοιτητή
Ευριπίδης Τσιαρδακλής
Αρ. Μητρώου: 2019216

Επιβλέπων
Αδαμίδης Παναγιώτης
Καθηγητής

Θεσσαλονίκη, Σεπτέμβριος 2026

Τίτλος Δ.Ε.: Ανάπτυξη εφαρμογής αναγνώρισης συναισθημάτων από εκφράσεις προσώπου και προσαρμογής μουσικού περιεχομένου σε κινητές συσκευές

Κωδικός Δ.Ε.: 26213

Ονοματεπώνυμο φοιτητή: Ευριπίδης Τσιαρδακλής

Ονοματεπώνυμο εισηγητή: Παναγιώτης Αδαμίδης

Ημερομηνία ανάληψης Δ.Ε.: 24 Μαρτίου 2026

Ημερομηνία περάτωσης Δ.Ε.: 6 Σεπτεμβρίου 2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Ευριπίδη Τσιαρδακλή που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (downloading), ανάρτηση (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραίτητα και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Πρόλογος

Η παρούσα διπλωματική εργασία εστιάζει στη διασύνδεση δύο δυναμικών επιστημονικών πεδίων που κατέχουν κεντρική θέση στη σύγχρονη τεχνολογία: της τεχνητής νοημοσύνης και της ανάπτυξης εφαρμογών για φορητές συσκευές. Βασικό κίνητρο για την εκπόνησή της αποτέλεσε η μελέτη της συναισθηματικής υπολογιστικής (affective computing) και η εφαρμογή της σε ένα σύστημα εξατομικευμένης σύστασης μουσικής, με βάση τη συναισθηματική κατάσταση του χρήστη.

Για την επίτευξη αυτού του στόχου, η εργασία συνδυάζει τη θεωρητική διερεύνηση με την πρακτική υλοποίηση. Ειδικότερα, εξετάστηκαν αρχιτεκτονικές βαθιάς μάθησης και μεθοδολογίες εκπαίδευσης νευρωνικών δικτύων, ενώ παράλληλα αναπτύχθηκε μια εφαρμογή Android με τη χρήση σύγχρονων βιβλιοθηκών, όπως το MediaPipe και το TensorFlow Lite. Επιπλέον, μέσω της ενσωμάτωσης του Spotify Web API, εφαρμόστηκαν στην πράξη οι αρχές λειτουργίας των RESTful APIs. Η προσέγγιση ολοκληρώνεται με τη μελέτη αρχών από τον χώρο της ψυχολογίας της μουσικής και την εφαρμογή μεθόδων αξιολόγησης της χρηστικότητας του συστήματος.

Περίληψη

Η παρούσα διπλωματική εργασία πραγματεύεται τον σχεδιασμό, την υλοποίηση και την αξιολόγηση μιας κινητής εφαρμογής Android, η οποία αναγνωρίζει αυτόματα τη συναισθηματική κατάσταση του χρήστη από εκφράσεις προσώπου σε πραγματικό χρόνο και παρέχει εξατομικευμένες μουσικές συστάσεις μέσω του Spotify API. Η εφαρμογή φέρει την ονομασία «Express Your Sound» και υλοποιείται για την πλατφόρμα Android με χρήση της γλώσσας προγραμματισμού Kotlin.

Για την αναγνώριση συναισθήματος αξιοποιείται το προεκπαιδευμένο συνελκτικό νευρωνικό δίκτυο MobileNetV2, το οποίο προσαρμόζεται στο πρόβλημα ταξινόμησης επτά κατηγοριών συναισθήματος μέσω διφασικής στρατηγικής λεπτομερούς προσαρμογής (fine-tuning) στο δημόσιο σύνολο δεδομένων FER-2013. Το εκπαιδευμένο μοντέλο εξάγεται σε μορφή TensorFlow Lite και εκτελείται αποκλειστικά τοπικά στη συσκευή (on-device), διαφυλάσσοντας την ιδιωτικότητα των βιομετρικών δεδομένων του χρήστη. Η ανίχνευση και προεπεξεργασία του προσώπου πραγματοποιείται μέσω της βιβλιοθήκης MediaPipe Face Landmarker, η οποία παράγει 478 γεωμετρικά σημεία αναφοράς και 52 συντελεστές έκφρασης σε κάθε καρέ της ροής βίντεο.

Η αντιστοίχιση του ανιχνευόμενου συναισθήματος σε μουσικές παραμέτρους βασίζεται στο κυκλικό μοντέλο των συναισθημάτων (Circumplex Model of Affect), μετατρέποντας κάθε ανιχνευόμενο συναίσθημα σε στοχευμένα ερωτήματα αναζήτησης βάσει μουσικού είδους και διάθεσης προς το Spotify Web API. Επιπλέον, το σύστημα καταγράφει το πρόσφατο ακροαματικό προφίλ του χρήστη μέσω του API, θέτοντας τη βάση για μελλοντική περαιτέρω εξατομίκευση των προτάσεων.

Η αξιολόγηση πραγματοποιείται στα επίπεδα της τεχνικής αξιολόγησης και της αξιολόγησης μέσω μελέτης χρηστών. Η τεχνική αξιολόγηση του μοντέλου πραγματοποιήθηκε στο σύνολο δοκιμής (test set) του FER-2013 με χρήση των μετρικών ακρίβειας (accuracy), F1-score ανά κατηγορία και πίνακα σύγχυσης (confusion matrix). Η αξιολόγηση χρηστικότητας μέσω μελέτης χρηστών (user study) με 5 συμμετέχοντες αξιοποιεί την κλίμακα ευχρηστίας συστήματος (System Usability Scale) καθώς και κλίμακες Likert για την αντιληπτή ποιότητα των μουσικών προτάσεων.

Τα αποτελέσματα παρέχουν ενδείξεις για την τεχνική εφικτότητα ενός ολοκληρωμένου συστήματος αναγνώρισης συναισθήματος και εξατομικευμένης μουσικής σύστασης σε κινητές συσκευές, αναδεικνύοντας τις δυνατότητες και τους περιορισμούς της εφαρμοσμένης συναισθηματικής υπολογιστικής (affective computing) στην καθημερινή ζωή.

Development of an Application for Emotion Recognition from Facial Expressions and Musical Content Adaptation on Mobile Devices

Evripidis Tsiardaklis

Abstract

The present thesis addresses the design, implementation, and evaluation of an Android mobile application that automatically recognizes the user's emotional state from facial expressions in real time and provides personalized music recommendations via the Spotify API. The application, named "Express Your Sound," is developed for the Android platform using the Kotlin programming language.

For emotion recognition, the pretrained convolutional neural network MobileNetV2 is employed, adapted to a seven-class emotion classification task through a two-phase fine-tuning strategy on the publicly available FER-2013 dataset. The trained model is exported in TensorFlow Lite format and executed entirely on-device, thereby preserving the privacy of the user's biometric data. Face detection and preprocessing are performed using the MediaPipe Face Landmarker library, which generates 478 geometric reference points and 52 expression coefficients (blendshapes) for each frame of the video stream.

The mapping of the detected emotion to musical parameters is based on the Circumplex Model of Affect, converting each detected emotion into targeted search queries based on genre and mood directed to the Spotify Web API. Additionally, the system records the user's recent listening profile via the API, laying the groundwork for future further personalization of recommendations.

The evaluation was conducted at two levels: technical evaluation and user-study-based evaluation. The technical evaluation of the CNN model was performed on the FER-2013 test set using accuracy, per-class F1-score, and a confusion matrix as metrics. The usability evaluation was conducted through a user study with 5 participants, employing the System Usability Scale (SUS) as well as Likert-type scales to assess the perceived quality of the music recommendations.

The results provide indications of the technical feasibility of an integrated system for emotion recognition and personalized music recommendation on mobile devices, highlighting both the potential and the limitations of applied affective computing in everyday life.

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε όλους εκείνους που συνέβαλαν στην επιτυχή ολοκλήρωση της παρούσας εργασίας. Πρωτίστως, ευχαριστώ πολύ τον καθηγητή μου τον κ. Αδαμίδα Παναγιώτη για την πολύτιμη βοήθεια, καθοδήγηση και υποστήριξή του καθ' όλη τη διάρκεια της εκπόνησης αυτής της μελέτης. Οι επιστημονικές του συμβουλές και η εμπειρία του αποτέλεσαν καθοριστικό παράγοντα για την ποιότητα και την επιτυχία του έργου.

Παράλληλα, θα ήθελα να εκφράσω την ευγνωμοσύνη μου προς όλους τους καθηγητές που είχα κατά τη διάρκεια των σπουδών μου. Οι γνώσεις και οι αρχές που μου μετέδωσαν, καθώς και η επιστημονική κατάρτιση που μου προσέφεραν, αποτέλεσαν το θεωρητικό υπόβαθρο επάνω στο οποίο στηρίχθηκε η παρούσα εργασία.

Ιδιαίτερες ευχαριστίες οφείλω στην οικογένεια και στους φίλους μου, καθώς και στην Ευδοξία Μακρίδου για τις επικοινωνιακές συμβουλές και την ανεκτίμητη βοήθειά της τόσο σε γραμματικά ζητήματα όσο και σε πρακτικά θέματα για την υλοποίηση της παρούσας εργασίας. Η υπομονή και η αμέριστη στήριξή τους υπήρξαν καθοριστικές για την ολοκλήρωση αυτού του εγχειρήματος. Η παρούσα εργασία δεν θα μπορούσε να ολοκληρωθεί χωρίς τη συλλογική προσπάθεια και συμβολή όλων των παραπάνω προσώπων.

Περιεχόμενα

Πρόλογος.....	iii
Περίληψη.....	iv
Abstract	v
Ευχαριστίες	vi
Περιεχόμενα	vii
Κατάλογος Πινάκων.....	x
Κατάλογος Σχημάτων	x
Συντομογραφίες.....	xi
Κεφάλαιο 1ο: Εισαγωγή	1
1.1 Αντικείμενο και κίνητρο της εργασίας.....	1
1.2 Σκοπός και στόχοι της εργασίας.....	1
1.3 Μεθοδολογία της εργασίας	2
1.4 Δομή της διπλωματικής εργασίας	3
Κεφάλαιο 2ο: Ανθρώπινα συναισθήματα και εκφράσεις προσώπου.....	5
2.1 Θεωρίες συναισθημάτων	5
2.1.1 Θεωρία βασικών συναισθημάτων: Paul Ekman.	5
2.1.2 Διαστατικά μοντέλα συναισθήματος.....	6
2.2 Ανατομία της συναισθηματικής έκφρασης προσώπου.....	7
2.3 Σύστημα κωδικοποίησης εκφράσεων προσώπου: FACS.....	9
2.4 Πολιτισμικές και ατομικές διαφορές στην έκφραση συναισθημάτων	11
Κεφάλαιο 3ο: Τεχνητή νοημοσύνη και αναγνώριση συναισθημάτων.....	12
3.1 Βασικές αρχές και κατηγορίες της μηχανικής μάθησης.....	12
3.1.1 Τεχνητά νευρωνικά δίκτυα και βαθιά μάθηση	13
3.1.2 Μεταφορά γνώσης και προσαρμογή του μοντέλου.....	15
3.2 Συνελκτικά νευρωνικά δίκτυα.....	15
3.2.1 Η αρχιτεκτονική MobileNetV2	17
3.2.2 CNNs στην αναγνώριση εκφράσεων.....	18
3.3 Ανίχνευση και εντοπισμός χαρακτηριστικών προσώπου με MediaPipe.....	19
3.3.1 Ρόλος του MediaPipe στη διαδικασία αναγνώρισης συναισθήματος.....	21
3.4 Αξιολόγηση μοντέλων ταξινόμησης	22
3.5 Συναισθηματική υπολογιστική: επισκόπηση πεδίου	23
Κεφάλαιο 4ο: Μουσική και συναισθηματική προσαρμογή.....	25

4.1	Μουσική και ανθρώπινο συναίσθημα	25
4.2	Μουσικοθεραπεία: Θεωρητικό υπόβαθρο και εφαρμογές	26
4.3	Αντιστοίχιση μουσικών χαρακτηριστικών σε συναισθηματικές καταστάσεις.....	27
4.4	Συστήματα προσαρμογής και σύστασης μουσικού περιεχομένου	29
Κεφάλαιο 5ο: Σχεδιασμός του συστήματος.....		30
5.1	Απαιτήσεις συστήματος	30
5.2	Αρχιτεκτονική του συστήματος	31
5.3	Αντιστοίχιση συναισθημάτων με μουσικά χαρακτηριστικά	32
5.4	Εξατομίκευση μέσω προφίλ χρήστη	33
5.5	Επιλογή τεχνολογιών υλοποίησης.....	35
Κεφάλαιο 6ο: Ανάπτυξη και εκπαίδευση του μοντέλου αναγνώρισης συναισθημάτων		37
6.1	Περιγραφή και χαρακτηριστικά του συνόλου δεδομένων FER-2013.....	37
6.2	Προεπεξεργασία δεδομένων.....	38
6.3	Αρχιτεκτονική νευρωνικού δικτύου.....	39
6.4	Διαδικασία εκπαίδευσης και Fine-Tuning.....	40
6.5	Μετατροπή και βελτιστοποίηση του μοντέλου σε TensorFlow Lite.....	42
Κεφάλαιο 7ο: Ανάπτυξη της κινητής εφαρμογής		43
7.1	Επισκόπηση υλοποίησης.....	43
7.2	Ανίχνευση προσώπου	44
7.3	Αναγνώριση συναισθήματος	45
7.4	Υλοποίηση της διεπαφής χρήστη	46
7.5	Διασύνδεση με το Spotify	47
7.6	Διαχείριση δικαιωμάτων πρόσβασης και ασφάλεια.....	48
7.7	Δοκιμή και αποσφαλμάτωση της εφαρμογής.....	49
Κεφάλαιο 8ο: Αξιολόγηση και συζήτηση αποτελεσμάτων		51
8.1	Μεθοδολογία αξιολόγησης.....	51
8.2	Αξιολόγηση του μοντέλου αναγνώρισης συναισθημάτων	51
8.3	Αξιολόγηση της εφαρμογής	55
8.3.1	Πρωτόκολλο δοκιμής	55
8.4	Αξιολόγηση χρηστικότητας εφαρμογής.....	56
8.5	Συζήτηση αποτελεσμάτων.....	59
8.6	Περιορισμοί της εργασίας	61
Κεφάλαιο 9ο: Συμπεράσματα και μελλοντική έρευνα.....		62
9.1	Συμπεράσματα.....	62
9.2	Συνεισφορά της εργασίας.....	63

9.3	Μελλοντική έρευνα	63
ΒΙΒΛΙΟΓΡΑΦΙΑ.....		65
ΠΑΡΑΡΤΗΜΑ Α : ΦΩΤΟΓΡΑΦΙΚΟ ΥΛΙΚΟ ΤΟΥ USER STUDY		68
ΠΑΡΑΡΤΗΜΑ Β : ΒΕΒΑΙΩΣΗ ΣΥΜΜΕΤΟΧΗΣ		70

Κατάλογος Πινάκων

Πίνακας 6-1: Πίνακας υπερπαραμέτρων.....	41
Πίνακας 8-1: Αναφορά ταξινόμησης - σύνολο δοκιμής FER-2013.....	52
Πίνακας 8-2: Σύγκριση ακρίβειας.....	55
Πίνακας 8-3: Ερωτηματολόγιο SUS	57
Πίνακας 8-4: Ερωτηματολόγιο αξιολόγησης προτάσεων	57
Πίνακας 8-5: Απαντήσεις χρηστών στα ερωτήματα SUS (κλίμακα Likert 1–5)	58
Πίνακας 8-6: Αποτελέσματα αξιολόγησης μουσικών προτάσεων	59

Κατάλογος Σχημάτων

Σχήμα 1-1: Αναγνώριση εκφράσεων προσώπου.....	3
Πηγή: Rodríguez-Azar, Mejía-Muñoz and Ochoa-Zezzatti (2022), Available: https://www.redalyc.org/journal/614/61472445002/html/	
Σχήμα 2-1: Άτλας συναισθημάτων του Paul Ekman.....	6
Πηγή: Ekman, P. and Ekman, E. (2016), The Ekmans' Atlas of Emotions. Available: https://atlasofemotions.org	
Σχήμα 2-2: Ο ηθοποιός Tim Roth παρουσιάζει εκφράσεις του προσώπου και την επεξήγησή τους..	10
Πηγή: Envision your Evolution (2019) 'How Paul Ekman's Theory of Emotion Can Help You Better Understand Others'. Envision your Evolution, 6 June. Available: https://www.envisionyourevolution.com/evolution-emotion/paul-ekman-theory-of-emotion/2149/	
Σχήμα 3-1: Ερευνητικές θεματικές περιοχές στο πεδίο της υπολογιστικής του συναισθήματος.....	12
Πηγή: Pei, G. et al. (2024) Five research themes in affective computing. EurekaAlert! Available: https://www.eurekaalert.org/multimedia/1016725	
Σχήμα 3-2: Πυκνό πλέγμα 478 landmark 3D σημείων αναφοράς.	20
Πηγή: Medium (2023) Mediapipe landmark face/hand/pose sequence number list view, Available: https://medium.com/@hotakoma/mediapipe-landmark-face-hand-pose-sequence-number-list-view-778364d6c414	
Σχήμα 4-1: Κυκλικό μοντέλο συναισθημάτων του Russell.	28
Πηγή: ResearchGate (2025) Toward music-based stress management, Available: https://www.researchgate.net/figure/Russells-1980-circumplex-model-of-affect_fig1_393923809	
Σχήμα 5-1: Αρχιτεκτονική MVVM.....	36
Πηγή: Google Developers (2024) Guide to app architecture: Model-View-ViewModel (MVVM). Available: https://developer.android.com/topic/architecture	
Σχήμα 6-1: Αρχιτεκτονική MobileNetV2 & Classification Head.	40
Πηγή: Sandler, M., Howard, A., Chow, M., Zhmoginov, A. and Chen, L.C. (2018) 'MobileNetV2: Inverted residuals and linear bottlenecks', Computer Vision and Pattern Recognition (CVPR). Available: https://arxiv.org/abs/1801.04381	
Σχήμα 8-1: Πίνακας σύγχυσης συνόλου δοκιμής.....	54
Πηγή: Ίδια επεξεργασία.	

Συντομογραφίες

Δ.Ε.	Διπλωματική Εργασία
ΔΙΠΙΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
CNN	Συνελικτικό Νευρωνικό Δίκτυο (Convolutional Neural Network)
FACS	Σύστημα Κωδικοποίησης Δράσης Προσώπου (Facial Action Coding System)
FER	Αναγνώριση Εκφράσεων Προσώπου (Facial Expression Recognition)
AU	Μονάδα Δράσης (Action Unit)
API	Διεπαφή Προγραμματισμού Εφαρμογών (Application Programming Interface)
GDPR	Γενικός Κανονισμός Προστασίας Δεδομένων (General Data Protection Regulation)
GPU	Μονάδα Επεξεργασίας Γραφικών (Graphics Processing Unit)
MVVM	Model-View-ViewModel
NNAPI	Neural Networks API (Android)
OAuth	Open Authorization
SDK	Πακέτο Ανάπτυξης Λογισμικού (Software Development Kit)
TFLite	TensorFlow Lite

Κεφάλαιο 1ο: Εισαγωγή

1.1 Αντικείμενο και κίνητρο της εργασίας

Η ταχεία εξέλιξη της τεχνητής νοημοσύνης και η ολοένα αυξανόμενη ενσωμάτωσή της στην καθημερινή ζωή έχουν δημιουργήσει νέες προοπτικές στον τομέα της αλληλεπίδρασης ανθρώπου-υπολογιστή. Μεταξύ των πλέον αξιοσημείωτων ερευνητικών κατευθύνσεων στο πεδίο αυτό συγκαταλέγεται η αυτόματη αναγνώριση συναισθημάτων από εκφράσεις προσώπου, η οποία αποτελεί κεντρικό αντικείμενο του ευρύτερου πεδίου της συναισθηματικής υπολογιστικής (affective computing) (Picard, 1997) [17]. Η δυνατότητα ενός υπολογιστικού συστήματος να αντιλαμβάνεται τη συναισθηματική κατάσταση του χρήστη σε πραγματικό χρόνο και να προσαρμόζει αναλόγως τη συμπεριφορά του αποτελεί σήμερα αντικείμενο εντατικής έρευνας, με εφαρμογές σε τομείς όπως η ψυχική υγεία, η εκπαίδευση, η ψυχαγωγία και η εξατομικευμένη αλληλεπίδραση (X. Lowe-Brown et al., 2024) [16].

Παράλληλα, η μουσική αναγνωρίζεται ευρέως ως ένα από τα ισχυρότερα μέσα ρύθμισης του ανθρώπινου συναισθήματος. Εμπειρικές μελέτες τεκμηριώνουν ότι η ακρόαση μουσικής επιδρά άμεσα στη συναισθηματική κατάσταση του ακροατή, επηρεάζοντας φυσιολογικές παραμέτρους όπως ο καρδιακός ρυθμός και τα επίπεδα κορτιζόλης, ενώ συμβάλλει ουσιαστικά στη μείωση του άγχους και στην ενίσχυση θετικών συναισθηματικών καταστάσεων (Thoma et al., 2013) [12]. Η άνοδος των πλατφορμών μουσικής streaming, και ιδιαίτερα του Spotify, έχει θέσει στη διάθεση της ερευνητικής κοινότητας πλούσια δεδομένα ακουστικών χαρακτηριστικών μουσικής που επιτρέπουν την αλγοριθμική αντιστοίχιση μουσικού περιεχομένου με συναισθηματικές καταστάσεις (Schedl et al., 2018) [29].

Η σύζευξη αυτόματης αναγνώρισης συναισθημάτων από εκφράσεις προσώπου με εξατομικευμένη σύσταση μουσικής αποτελεί τη θεματική περιοχή στην οποία εντάσσεται η παρούσα διπλωματική εργασία. Το κίνητρο για την επιλογή του συγκεκριμένου θέματος απορρέει από τη διαπίστωση ότι, παρά την ύπαρξη ερευνητικών εργασιών που αντιμετωπίζουν ξεχωριστά το πρόβλημα της αναγνώρισης συναισθήματος και το πρόβλημα της σύστασης μουσικής, ολοκληρωμένες κινητές εφαρμογές που συνδυάζουν τα δύο αυτά συστατικά σε πραγματικό χρόνο, on-device και με εξατομίκευση βάσει προφίλ χρήστη παραμένουν σχετικά σπάνιες στη βιβλιογραφία. Στο πλαίσιο αυτό, η εργασία διερευνά την ανάπτυξη μιας λειτουργικής Android εφαρμογής που αξιοποιεί τεχνικές βαθιάς μάθησης για αναγνώριση συναισθήματος και το Spotify Web API για εξατομικευμένη σύσταση μουσικής.

1.2 Σκοπός και στόχοι της εργασίας

Σκοπός της παρούσας διπλωματικής εργασίας είναι ο σχεδιασμός, η υλοποίηση και η αξιολόγηση κινητής εφαρμογής Android που αναγνωρίζει αυτόματα τη συναισθηματική κατάσταση του χρήστη από εκφράσεις προσώπου σε πραγματικό χρόνο και προτείνει εξατομικευμένες μουσικές συστάσεις μέσω του Spotify API, συνδυάζοντας τεχνικές βαθιάς μάθησης με θεωρητικά τεκμηριωμένη αντιστοίχιση συναισθήματος-μουσικής.

Για την επίτευξη του ανωτέρω σκοπού, τίθενται οι ακόλουθοι επιμέρους στόχοι:

Αρχικά, επιχειρείται η μελέτη και η κριτική ανασκόπηση της σχετικής βιβλιογραφίας στα πεδία της αυτόματης αναγνώρισης συναισθημάτων από εκφράσεις προσώπου, της ψυχολογίας της μουσικής

και των συστημάτων σύστασης μουσικής, με ιδιαίτερη έμφαση σε εργασίες που αξιοποιούν βαθιά μάθηση και συνελκτικά νευρωνικά δίκτυα.

Σε επόμενο στάδιο, η παρούσα διπλωματική εργασία επιδιώκει την εκπαίδευση μοντέλου αναγνώρισης συναισθήματος βάσει της αρχιτεκτονικής MobileNetV2 με στρατηγική μεταφοράς γνώσης (transfer learning) και λεπτομερούς προσαρμογής (fine-tuning) στο δημόσια διαθέσιμο σύνολο δεδομένων FER-2013, με σκοπό την επίτευξη ικανοποιητικής ακρίβειας ταξινόμησης σε επτά κατηγορίες συναισθήματος εντός των υπολογιστικών περιορισμών κινητής συσκευής.

Κατόπιν, επιδιώκεται η ανάπτυξη μίας λειτουργικής Android εφαρμογής που ενσωματώνει το εκπαιδευμένο μοντέλο σε μορφή TensorFlow Lite για εξαγωγή συμπερασμάτων επί της συσκευής (on-device inference), το MediaPipe Face Landmarker για ανίχνευση και προεπεξεργασία προσώπου, και το Spotify Web API για ανάκτηση εξατομικευμένων μουσικών προτάσεων βάσει του ανιχνευόμενου συναισθήματος και του προφίλ ακρόασης του χρήστη.

Η εργασία ολοκληρώνεται με την αξιολόγηση του συστήματος τόσο σε επίπεδο της τεχνικής απόδοσης του μοντέλου αναγνώρισης συναισθήματος στο σύνολο δοκιμής (test set) του FER-2013 με μετρικές accuracy, precision, recall, F1-score και τον πίνακα σύγχυσης, όσο και με την αξιολόγηση της χρηστικότητας της εφαρμογής με μελέτη χρηστών (user study) αξιοποιώντας το ερωτηματολόγιο System Usability Scale (SUS) (Brooke, 1996) [31].

1.3 Μεθοδολογία της εργασίας

Η παρούσα εργασία ακολουθεί μεθοδολογία τεσσάρων διαδοχικών φάσεων, οι οποίες αντιστοιχούν στα βασικά στάδια ανάπτυξης ενός συστήματος τεχνητής νοημοσύνης για κινητές εφαρμογές.

Η πρώτη φάση της μεθοδολογίας αφορά την επιλογή κατάλληλης αρχιτεκτονικής νευρωνικού δικτύου για την εργασία ταξινόμησης συναισθήματος από εικόνες προσώπου. Λαμβάνοντας υπόψη κριτήρια υπολογιστικής αποδοτικότητας για εξαγωγή συμπερασμάτων σε κινητές συσκευές Android, τον περιορισμένο αριθμό παραμέτρων, καθώς και την τεκμηριωμένη απόδοση στο σύνολο δεδομένων FER-2013, επιλέγεται η αρχιτεκτονική MobileNetV2 (Sandler et al., 2018) [15], σε συνδυασμό με στρατηγική μεταφοράς γνώσης από προεκπαιδευμένα βάρη ImageNet. Η επιλογή αυτή κρίνεται προτιμότερη έναντι της εκπαίδευσης από μηδενική βάση, τόσο ως προς την υπολογιστική αποδοτικότητα όσο και ως προς την επίτευξη υψηλότερης ακρίβειας με περιορισμένο σύνολο εκπαίδευσης (Chowdhury, Idris and Abas, 2024) [3].

Η δεύτερη φάση αφορά την εκπαίδευση του μοντέλου στο σύνολο δεδομένων FER-2013 (Goodfellow et al., 2013) [30], το οποίο περιλαμβάνει 35.887 εικόνες προσώπου κατανεμημένες σε επτά κατηγορίες συναισθήματος, όπως διακρίνεται στο σχήμα 1-1. Στο πλαίσιο αυτό, εφαρμόζεται διφασική στρατηγική λεπτομερούς προσαρμογής: κατά την πρώτη φάση εκπαιδεύεται αποκλειστικά μία νέα κεφαλή ταξινόμησης (classification head), με τον κορμό (backbone) να παραμένει παγωμένο, ενώ κατά τη δεύτερη φάση αποπαγώνουν τα τελευταία επίπεδα του κορμού, τα οποία και επανεκπαιδεύονται με χαμηλό ρυθμό μάθησης (learning rate). Προκειμένου να αντιμετωπιστεί η ανισορροπία κλάσεων που χαρακτηρίζει το FER-2013, εφαρμόζονται τεχνικές επαύξησης δεδομένων (data augmentation), καθώς και η συνάρτηση απώλειας διασταυρούμενης εντροπίας με στάθμιση κλάσεων (class-weighted cross-entropy loss) (Mahendran et al., 2024) [4]. Η εκπαίδευση πραγματοποιείται σε περιβάλλον Google Colab με χρήση GPU, ενώ το τελικό εκπαιδευμένο μοντέλο εξάγεται σε μορφή TensorFlow Lite, με εφαρμογή της ποσοτικοποίησης δυναμικού εύρους (dynamic range quantization), ώστε να καταστεί δυνατή η ενσωμάτωσή του στην κινητή εφαρμογή.



Σχήμα 1-1: Αναγνώριση εκφράσεων προσώπου

Η τρίτη φάση αφορά την ανάπτυξη της κινητής εφαρμογής σε γλώσσα προγραμματισμού Kotlin, αξιοποιώντας το αρχιτεκτονικό μοτίβο MVVM. Η εφαρμογή ενσωματώνει τρία βασικά τεχνολογικά συστατικά: το MediaPipe Face Landmarker, για την ανίχνευση και τη χαρτογράφηση χαρακτηριστικών προσώπου σε πραγματικό χρόνο· τον TFLite Interpreter, για την on-device εκτέλεση του μοντέλου ταξινόμησης συναισθήματος· και το Spotify Web API, με εξουσιοδότηση μέσω OAuth 2.0, για την ανάκτηση εξατομικευμένων μουσικών προτάσεων. Η αντιστοίχιση του ανιχνευόμενου συναισθήματος σε παραμέτρους αναζήτησης μουσικής βασίζεται στο Circumplex Model of Affect (Russell, 1980) [32], ενώ η εξατομίκευση των προτάσεων υλοποιείται μέσω αξιοποίησης του ιστορικού ακρόασης του χρήστη ως seeds στο αντίστοιχο recommendation endpoint.

Η τέταρτη και τελευταία φάση αφορά την αξιολόγηση του συστήματος σε δύο επίπεδα. Αρχικά, αξιολογείται η τεχνική απόδοση του μοντέλου CNN στο σύνολο δοκιμής του FER-2013, με χρήση μετρικών όπως accuracy, precision, recall και F1-score ανά κατηγορία, macro-averaged F1, καθώς και τον πίνακα σύγχυσης. Στη συνέχεια, αξιολογείται η χρηστικότητα και η αντιληπτή αποτελεσματικότητα της εφαρμογής μέσω μελέτης χρηστών με τη συμμετοχή 5 χρηστών, αξιοποιώντας το ερωτηματολόγιο System Usability Scale (Brooke, 1996) [31], σε συνδυασμό με συμπληρωματικό ερωτηματολόγιο τύπου Likert για την αξιολόγηση της ποιότητας των προτεινόμενων μουσικών κομματιών.

1.4 Δομή της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται σε εννέα κεφάλαια, η δομή των οποίων αντικατοπτρίζει τη μεθοδολογική πορεία που περιγράφηκε στην ενότητα 1.3. Αρχικά, παρουσιάζεται το αντικείμενο, το κίνητρο, ο σκοπός, οι στόχοι, η μεθοδολογία και η δομή της εργασίας, ενώ στη συνέχεια πραγματεύεται το θεωρητικό υπόβαθρο της αναγνώρισης συναισθημάτων, αναλύοντας τις βασικές αρχές μηχανικής και βαθιάς μάθησης, τα συνελκτικά νευρωνικά δίκτυα, την αρχιτεκτονική MobileNetV2, τη στρατηγική μεταφοράς γνώσης, η ροή επεξεργασίας (pipeline) ανίχνευσης προσώπου μέσω MediaPipe Face Landmarker, καθώς και τις μετρικές αξιολόγησης και μια εισαγωγή στη

συναισθηματική υπολογιστική. Ακολουθώντας, εξετάζεται εκτενέστερα το πεδίο της τεχνητής νοημοσύνης και της αναγνώρισης συναισθημάτων, όπου αναλύονται οι κατηγορίες τους, η δομή και εκπαίδευση των τεχνητών νευρωνικών δικτύων, οι τεχνικές μεταφοράς γνώσης και λεπτομερούς προσαρμογής, η λειτουργία των CNNs στην εξαγωγή χαρακτηριστικών, η εξειδικευμένη αρχιτεκτονική MobileNetV2 στην αναγνώριση εκφράσεων, καθώς και ο ρόλος του Face Landmarker του MediaPipe στον εντοπισμό χαρακτηριστικών προσώπου, με το θεωρητικό μέρος να ολοκληρώνεται με μια επισκόπηση της συναισθηματικής υπολογιστικής. Επιπλέον, καλύπτεται το θεωρητικό υπόβαθρο της μουσικής και του συναισθήματος, εξετάζοντας τη ρύθμιση συναισθημάτων, τη μουσικοθεραπεία, το Circumplex Model of Affect ως βάση αντιστοίχισης, καθώς και τα σύγχρονα συστήματα σύστασης μουσικής μέσω πλατφορμών streaming.

Βασισμένη σε αυτό το θεωρητικό πλαίσιο, η εργασία προχωρά στον σχεδιασμό του συστήματος, ορίζοντας τις λειτουργικές και μη λειτουργικές απαιτήσεις, την αρχιτεκτονική τεσσάρων επιπέδων της εφαρμογής, τον πίνακα αντιστοίχισης συναισθήματος-audio features βάσει Russell, τον μηχανισμό εξατομίκευσης μέσω top tracks seeds και την αιτιολόγηση του τεχνολογικού stack. Έπειτα, περιγράφεται αναλυτικά η διαδικασία εκπαίδευσης του μοντέλου αναγνώρισης συναισθήματος με τη χρήση του συνόλου δεδομένων FER-2013, τις τεχνικές προεπεξεργασίας και επαύξησης, τη διφασική στρατηγική λεπτομερούς προσαρμογής και τη μετατροπή του μοντέλου σε TFLite. Η πρακτική εφαρμογή αποτυπώνεται στην υλοποίηση της κινητής εφαρμογής Android, όπου αναλύεται η ενσωμάτωση του συστήματος αντίχενυσης προσώπου με CameraX και MediaPipe, του ταξινομητή συναισθήματος, της διεπαφής χρήστη, του Spotify OAuth 2.0 και του εξατομικευμένου μηχανισμού σύστασης. Τέλος, παρουσιάζονται η μεθοδολογία και τα αποτελέσματα της αξιολόγησης, στα οποία περιλαμβάνονται την πλήρης κεφαλή ταξινόμησης, ο πίνακας σύγχυσης (Confusion Matrix) του μοντέλου CNN και τα ευρήματα της πιλοτικής αξιολόγησης χρηστών με το SUS score και την ανάλυση των μουσικών προτάσεων, για να κλείσει η εργασία με τα τελικά συμπεράσματα, τη συνεισφορά, τους περιορισμούς και τις προτάσεις για μελλοντική έρευνα.

Κεφάλαιο 2ο: Ανθρώπινα συναισθήματα και εκφράσεις προσώπου

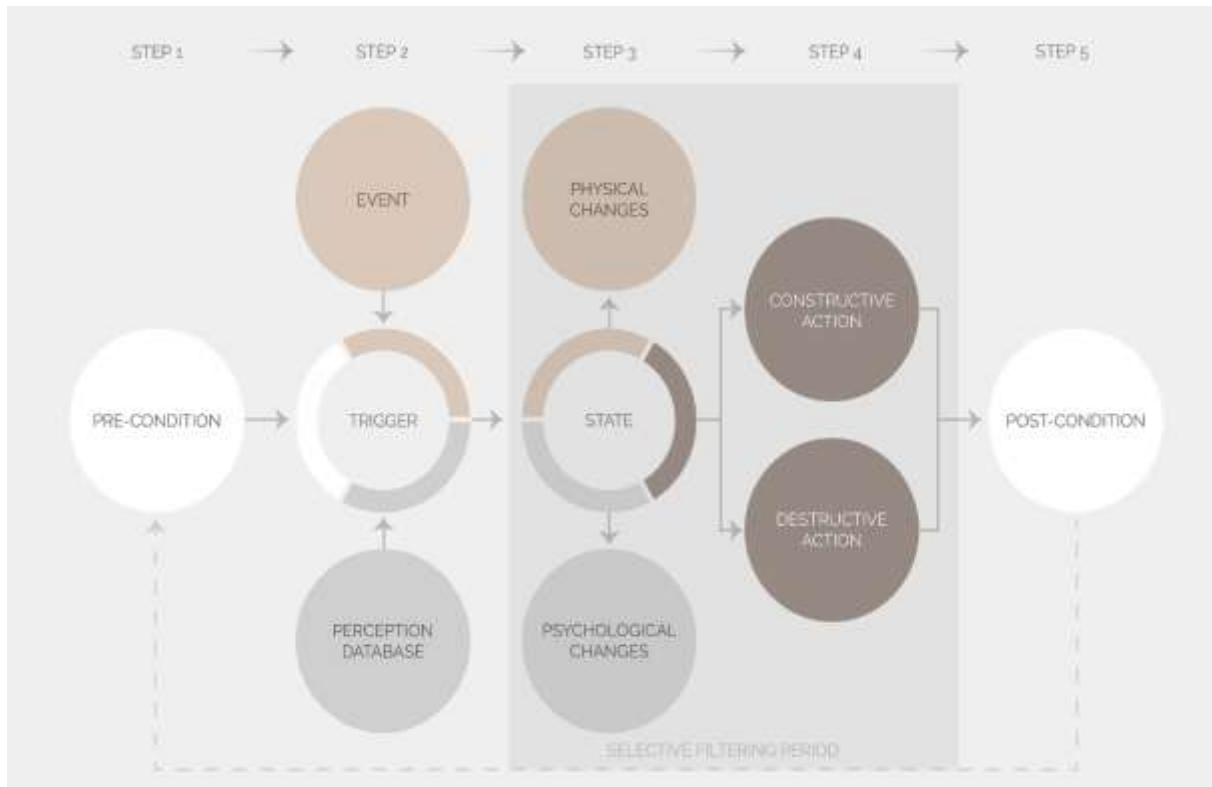
2.1 Θεωρίες συναισθημάτων

Η επιστημονική μελέτη των ανθρώπινων συναισθημάτων αποτελεί ένα ιδιαίτερος σύνθετο και πολυδιάστατο πεδίο έρευνας, το οποίο εκτείνεται σε μια ευρεία γκάμα επιστημονικών κλάδων, συμπεριλαμβανομένων της ψυχολογίας, της νευροεπιστήμης, της γνωσιακής επιστήμης, της ανθρωπολογίας και της φιλοσοφίας. Παρά την εκτεταμένη βιβλιογραφία και την πολυετή ερευνητική δραστηριότητα, δεν έχει επιτευχθεί μέχρι σήμερα μια καθολική επιστημονική συμφωνία ως προς τον ακριβή ορισμό, τη βιολογική προέλευση, τους μηχανισμούς εκδήλωσης και τη δομική οργάνωση των συναισθηματικών καταστάσεων. Η δυσκολία αυτή αποδίδεται πρωτίστως στην πολυεπίπεδη φύση του συναισθήματος, το οποίο συνδυάζει υποκειμενικές εσωτερικές εμπειρίες με αντικειμενικά παρατηρήσιμες φυσιολογικές μεταβολές, εκφράσεις του προσώπου και τάσεις συμπεριφοράς.

Στην προσπάθεια ποσοτικοποίησης και υπολογιστικής μοντελοποίησης αυτής της ανθρώπινης ιδιότητας, δύο κυρίαρχες θεωρητικές παραδόσεις έχουν διαμορφώσει το επιστημονικό πεδίο και εξακολουθούν να επηρεάζουν καθοριστικά τη σχεδίαση, την αρχιτεκτονική και τη λειτουργία των υπολογιστικών συστημάτων αυτόματης αναγνώρισης συναισθήματος (Affective Computing). Πρόκειται για τη θεωρία των διακριτών βασικών συναισθημάτων (Categorical/Basic Emotion Theory) και τα διαστατικά μοντέλα συναισθήματος (Dimensional Emotion Models), τα οποία προσεγγίζουν τη συναισθηματική αναπαράσταση από διαφορετικές αλλά συμπληρωματικές οπτικές γωνίες.

2.1.1 Θεωρία βασικών συναισθημάτων: Paul Ekman.

Η θεωρία των βασικών συναισθημάτων (Basic Emotions Theory), η οποία αναπτύχθηκε συστηματικά από τον Paul Ekman κατά τη δεκαετία του 1960, συνιστά τη θεμελιώδη θεωρητική βάση για την πλειονότητα των σύγχρονων συστημάτων αυτόματης αναγνώρισης συναισθήματος από μορφασμούς του προσώπου, συμπεριλαμβανομένου και του υπολογιστικού συνόλου δεδομένων FER-2013 που αξιοποιείται στην παρούσα εργασία. Σύμφωνα με τη θεμελιώδη παραδοχή του Ekman, υφίσταται ένα πεπερασμένο σύνολο έξι πρωτογενών ή βασικών συναισθημάτων: η χαρά (happiness), η λύπη (sadness), ο θυμός (anger), ο φόβος (fear), η αηδία (disgust) και η έκπληξη (surprise), όπως παρουσιάζονται στον άτλαντα του P. Ekman για την οργάνωση των ανθρώπινων συναισθημάτων στο σχήμα 2-1 παρακάτω. Σύμφωνα με τον Paul Ekman «Τα συναισθήματα αυτά είναι εξελικτικά προσαρμοστικά, βιολογικά καθορισμένα και εμφανίζουν παγκόσμιο, πολιτισμικά αναλλοίωτο χαρακτήρα.». Η θεμελιώδης αυτή θέση εδραιώθηκε μέσα από διαπολιτισμικές μελέτες του πεδίου που διεξήγαγε ο Ekman σε απομονωμένους, προβιομηχανικούς πληθυσμούς της Παπούας Νέας Γουινέας. Εκεί διαπιστώθηκε ότι τα μέλη ιθαγενών κοινοτήτων, χωρίς καμία προγενέστερη επαφή με τη δυτική κουλτούρα ή τα μέσα μαζικής ενημέρωσης, ήταν σε θέση όχι μόνο να αναγνωρίζουν με υψηλή ακρίβεια τις εκφράσεις προσώπου ατόμων από δυτικούς πληθυσμούς, αλλά και να παράγουν αντίστοιχες μορφαστικές εκφράσεις για τις ίδιες συναισθηματικές κατηγορίες, επιβεβαιώνοντας έτσι την κοινή και βιολογική ανθρωπολογική βάση των εκφράσεων αυτών (Ekman, 1992) [\[11\]](#).



Σχήμα 2-1: Άτλας συναισθημάτων του Paul Ekman

Κεντρικό θεωρητικό πυλώνα του οικοδομήματος του Ekman αποτελεί η υπόθεση των αυτόνομων νευρωνικών προγραμμάτων συναισθήματος (affect programs). Σύμφωνα με τον μηχανισμό αυτόν, η πυροδότηση ενός βασικού συναισθήματος ενεργοποιεί αυτομάτως ένα εγγενές, ταχύτατο και προβλέψιμο μοτίβο απόκρισης, το οποίο εκδηλώνεται σε πολλαπλά επίπεδα: περιλαμβάνει εξειδικευμένες συσπάσεις των μυών του προσώπου, αυτόνομες φυσιολογικές μεταβολές (όπως διακυμάνσεις στον καρδιακό ρυθμό, την αγωγιμότητα του δέρματος και την αναπνοή), καθώς και συγκεκριμένες προδιαθέσεις δράσης (action tendencies). Η μορφαστική έκφραση του προσώπου συνιστά την πλέον ορατή, άμεση και ποσοτικοποιήσιμη εκδήλωση αυτού του νευροφυσιολογικού μοτίβου, γεγονός που εξηγεί την κεντρική θέση που κατέχει στην υπολογιστική όραση και την αυτοματοποιημένη αναγνώριση συναισθημάτων (Ekman, 2003) [18].

Στο πλαίσιο μεταγενέστερων ερευνητικών επισκοπήσεων, η αρχική λίστα των έξι βασικών συναισθημάτων διευρύνθηκε με την εισαγωγή της κατηγορίας του ουδέτερου (neutral). Η κατηγορία αυτή αντιστοιχεί στην απουσία κάποιου έκδηλου συναισθηματικού μορφασμού ή σε μια κατάσταση ηρεμίας των μυών του προσώπου. Η εν λόγω προσθήκη αποδείχθηκε καθοριστική για τις υπολογιστικές εφαρμογές, καθώς ενσωματώθηκε στα σύγχρονα σύνολα δεδομένων εκπαίδευσης ως η έβδομη διακριτή κλάση. Έτσι, η επτακατηγορική ταξινόμηση είναι: θυμός, αηδία, φόβος, χαρά, λύπη, έκπληξη και ουδέτερο, αποτελεί το δομικό σχήμα κατηγοριοποίησης πάνω στο οποίο βασίστηκε η συγκρότηση του συνόλου δεδομένων FER-2013 (Goodfellow et al., 2013) [30].

2.1.2 Διαστατικά μοντέλα συναισθήματος

Παράλληλα με τη θεωρία βασικών συναισθημάτων, αναπτύχθηκαν και εδραιώθηκαν οι διαστατικές προσεγγίσεις στην ψυχολογία και τη γνωσιακή επιστήμη. Οι προσεγγίσεις αυτές

αντιμετωπίζουν τις συναισθηματικές εκδηλώσεις όχι ως απομονωμένες και περιχαρακωμένες κατηγορίες, αλλά ως συγκεκριμένες συντεταγμένες ή σημεία τοποθετημένα σε έναν συνεχή, πολυδιάστατο διανυσματικό χώρο. Το πλέον διαδεδομένο και ευρέως αποδεκτό διαστατικό υπόδειγμα είναι το κυκλικό μοντέλο συναισθήματος (Circumplex Model of Affect) του James Russell (1980), η θεωρητική θεμελίωση του οποίου αναλύεται διεξοδικά στο κεφάλαιο 4.3, στο πλαίσιο της λειτουργικής αντιστοίχισης με τα ηχητικά χαρακτηριστικά (audio features) της πλατφόρμας Spotify. Το μοντέλο του Russell οργανώνει τις συναισθηματικές καταστάσεις πάνω σε έναν δισδιάστατο καρτεσιανό χώρο, ο οποίος ορίζεται από δύο ορθογώνιους, ανεξάρτητους άξονες: τον άξονα της συναισθηματικής βαρύτητας (valence), που αποτυπώνει το συνεχές μεταξύ θετικής και αρνητικής συναισθηματικής ποιότητας, και τον άξονα της συναισθηματικής διέγερσης (arousal), που εκφράζει το επίπεδο νευροφυσιολογικής ενεργοποίησης από την κατάσταση ηρεμίας έως την κατάσταση υψηλής έντασης. Η συνεχή αυτή αναπαράσταση επιτρέπει την αποτύπωση λεπτόκοκκων (fine-grained) συναισθηματικών διακυμάνσεων, υπερβαίνοντας τους αυστηρούς ταξινομικούς περιορισμούς των επτά διακριτών κατηγοριών (Russell, 1980) [32].

Αξίζει να σημειωθεί ότι οι δύο προαναφερθείσες θεωρητικές παραδόσεις δεν λειτουργούν ως αμοιβαία αποκλειόμενα σχήματα, αλλά ως συμπληρωματικά επίπεδα ανάλυσης. Τα βασικά κατηγορικά συναισθήματα της ταξινόμησης του Ekman μπορούν να χαρτογραφηθούν στον δισδιάστατο χώρο του Russell καταλαμβάνοντας συγκεκριμένα τεταρτημόρια και περιοχές τιμών. Για παράδειγμα, η κατηγορία της χαράς τοποθετείται σε περιοχή υψηλού valence και μέτριας έως υψηλής διέγερσης (arousal), ο θυμός αντιστοιχεί σε χαμηλό valence και υψηλή διέγερση, ενώ η λύπη αποτυπώνεται σε περιοχή χαμηλού valence και χαμηλής διέγερσης. Η αντιστοίχιση αυτή αποτελεί ένα ισχυρό θεωρητικό υπόβαθρο, επιτρέποντας έτσι την μετάφραση μεταξύ της κατηγορικής και της διαστατικής αναπαράστασης του συναισθήματος (Russell, 1980) [32].

Η πρακτική εφαρμογή και η εξέλιξη των διαστατικών μοντέλων στον τομέα της υπολογιστικής όρασης και της αυτόματης αναγνώρισης συναισθημάτων ενισχύθηκε σημαντικά κατά τα τελευταία έτη με τη δημιουργία εκτεταμένων και σύνθετων βάσεων δεδομένων. Χαρακτηριστικό παράδειγμα αποτελεί το σύνολο δεδομένων AffectNet, το οποίο προσφέρει διπλή επισημάνση στα δείγματα εικόνων προσώπου, παρέχοντας ταυτόχρονα επισημειώσεις τόσο για τις διακριτές συναισθηματικές κατηγορίες όσο και για τις συνεχείς τιμές των διαστάσεων valence και arousal, διευκολύνοντας έτσι την εκπαίδευση προηγμένων μοντέλων βαθιάς μάθησης σε αμφότερα τα θεωρητικά πλαίσια (Mollahosseini et al., 2019) [14].

2.2 Ανατομία της συναισθηματικής έκφρασης προσώπου

Η ανατομική δομή του ανθρώπινου προσώπου συνιστά ένα ιδιαίτερα σύνθετο και ευέλικτο μυϊκό σύστημα, το οποίο αποτελείται από περισσότερους από 40 εξειδικευμένους μιμικούς μύες. Οι μύες αυτοί νευρώνονται από το 7ο κρανιακό νεύρο (προσωπικό νεύρο) και επιτρέπουν την πραγματοποίηση εξαιρετικά λεπτομερών, ταχέων και αλληλεπικαλυπτόμενων μορφαστικών μεταβολών στην επιφάνεια του προσώπου. Σε αντίθεση με τη συντριπτική πλειονότητα των σκελετικών μυών του σώματος, οι οποίοι συνδέονται σε οστά και παράγουν κίνηση στις αρθρώσεις, πολλοί από τους μιμικούς μύες φύονται ή καταφύονται απευθείας στους ιστούς του δέρματος. Αποτέλεσμα της ιδιομορφίας αυτής είναι ότι η συστολή τους δεν μετακινεί οστικά στοιχεία, αλλά παραμορφώνει άμεσα τη μορφολογία του δέρματος, προκαλώντας χαρακτηριστικές πτυχώσεις, οριζόντιες και κατακόρυφες ρυτίδες, καθώς και τοπικές ανυψώσεις ή κατασπάσεις σε συγκεκριμένες ανατομικές περιοχές (Ekman and Friesen, 1978) [19]. Η λεπτομερής κατανόηση αυτής της μυϊκής ανατομίας αποτελεί σημαντικό κομμάτι για την ερμηνεία των διανυσματικών εξόδων και των χωρικών σημείων αναφοράς (facial

landmarks) που εξάγει το εργαλείο MediaPipe Face Landmarker, καθώς και για την αντιστοίχισή τους με τις διακριτές κατηγορίες συναισθημάτων.

Στο ανώτερο τμήμα του προσώπου, η περιοχή των φρυδιών και του μετώπου συνιστά μία από τις πλέον εκφραστικές ζώνες, διαδραματίζοντας καθοριστικό ρόλο στη διαφοροποίηση των αρνητικών συναισθηματικών εκδηλώσεων. Ο συνοφρυωτικός μυς (corrugator supercilii) προκαλεί τη σύγκλιση των φρυδιών προς τη μέση γραμμή και την ταυτόχρονη κατάσπασή τους, δημιουργώντας χαρακτηριστικές κατακόρυφες ρυτίδες στην περιοχή του μεσόφρυου. Η κίνηση αυτή αποτελεί δεσπόζον μορφαστικό γνώρισμα στις εκφράσεις του θυμού, της λύπης και του φόβου. Αντίθετα, ο μετωπιαίος μυς (frontalis) ενεργεί ως ανελκτήρας των φρυδιών, προκαλώντας παράλληλες οριζόντιες ρυτίδες στην περιοχή του μετώπου, μια ενέργεια που παρατηρείται κυρίως κατά την εκδήλωση της έκπληξης και του φόβου. Επειδή ορισμένα συναισθήματα μοιράζονται κοινές μυϊκές ενεργοποιήσεις στην περιοχή των φρυδιών, η ακριβής διακριτική ταξινόμηση επιτυγχάνεται μέσω της συνδυαστικής ανάλυσης των μεταβολών στις υπόλοιπες ανατομικές περιοχές του προσώπου, και ιδιαίτερα στην περιμετρική ζώνη των οφθαλμών και του στόματος (Ekman and Friesen, 1978) [19].

Η περιμετρική περιοχή των οφθαλμών, η οποία περιλαμβάνει τα άνω και κάτω βλέφαρα, τις οφθαλμικές κόγχες και τους κανθούς, παρέχει εξαιρετικά πλούσια συναισθηματική πληροφορία για την αυθεντικότητα της έκφρασης. Ο σφικτήρας μυς των βλεφάρων (orbicularis oculi) ελέγχει τη σύγκλιση των βλεφάρων και την ανύσωση των παρειών προς τα πάνω κατά τη διάρκεια του χαμόγελου. Στη βιβλιογραφία γίνεται σαφής διαχωρισμός μεταξύ του αυθεντικού χαμόγελου βιωμένης χαράς (Duchenne smile) και του κοινωνικού ή προσποιητού χαμόγελου. Ένα γνήσιο χαμόγελο Duchenne απαιτεί τη συντονισμένη και ταυτόχρονη ενεργοποίηση τόσο του μείζονος ζυγωματικού μυός (zygomaticus major), ο οποίος ανασύρει τις γωνίες του στόματος προς τα πάνω και έξω, όσο και του σφικτήρα των βλεφάρων (orbicularis oculi), ο οποίος προκαλεί τις χαρακτηριστικές ακτινωτές ρυτίδες («πόδι της χήνας») στον έξω κανθό των ματιών. Αντίθετα, το τυπικό κοινωνικό ή συμβατικό χαμόγελο περιορίζεται στην εκούσια σύσπαση μόνο του μείζονος ζυγωματικού μυός (Ekman, 2003) [18]. Η ανατομική αυτή διάκριση προσφέρει ένα ισχυρό κριτήριο για την αυτόματη ανίχνευση της γνησιότητας των συναισθηματικών μορφασμών από υπολογιστικά μοντέλα.

Τέλος, το κάτω τμήμα του προσώπου, με επίκεντρο την περιοχή του στόματος, των χειλέων και της κάτω γνάθου, παρέχει κρίσιμα ανατομικά δεδομένα για τη διαμόρφωση της τελικής συναισθηματικής πρόβλεψης. Το χαρακτηριστικό άνοιγμα της κάτω γνάθου, που αντιστοιχεί στη χαλάρωση ή κατάσπαση του μασητήρα μυός, αποτελεί πρωτεύον γνώρισμα της έκπληξης και του έντονου φόβου. Η ανύσωση των γωνιών του στόματος προς τα πάνω σηματοδοτεί τη χαρά, ενώ η κατάσπαση των γωνιών του στόματος αποτελεί ένδειξη λύπης (Ekman, 2003) [18]. Παράλληλα, η ανύσωση του άνω χείλους και η αποκάλυψη των οδόντων συνιστά θεμελιώδη ανατομική εκδήλωση της αηδίας. Στο πλαίσιο της παρούσας υλοποίησης, οι συγκεκριμένες μυϊκές μεταβολές καταγράφονται ως συμπληρωματικά δεδομένα μορφοποίησης (blendshapes) που παρέχει το MediaPipe, όπως τα mouthSmileLeft/Right, mouthFrownLeft/Right και jawOpen. Η αντιστοίχιση αυτή παρέχει τη δυνατότητα εξαγωγής δεικτών έντασης μυϊκών συσπάσεων, δημιουργώντας μια υποδομή δεδομένων χωρίς ωστόσο να εμπλέκεται άμεσα στην τρέχουσα διαδικασία συναισθηματικής ταξινόμησης του CNN (Google AI Edge, 2023) [22].

2.3 Σύστημα κωδικοποίησης εκφράσεων προσώπου: FACS

Το σύστημα κωδικοποίησης εκφράσεων προσώπου (Facial Action Coding System, FACS), το οποίο αναπτύχθηκε από τους Ekman και Friesen (1978), σχεδιάστηκε με σκοπό την παροχή ενός πλήρους, αντικειμενικού και ανατομικά θεμελιωμένου πλαισίου για την κωδικοποίηση κάθε παρατηρήσιμης κίνησης του ανθρώπινου προσώπου. Μέχρι και σήμερα, το FACS παραμένει το πλέον διαδεδομένο και επιστημονικά αποδεκτό σύστημα ανάλυσης της εκφραστικότητας του προσώπου στην επιστημονική έρευνα, αποτελώντας τη βάση πάνω στην οποία οικοδομήθηκε η πλειονότητα των υπολογιστικών συστημάτων αυτόματης αναγνώρισης συναισθήματος.

Η θεμελιώδης δομική μονάδα του συστήματος είναι η μονάδα δράσης (Action Unit, AU), η οποία αντιστοιχεί στη μεμονωμένη ή συνδυαστική σύσπαση ανατομικά διακριτών μιμικών μυών. Κάθε AU αναγνωρίζεται μέσω ενός μοναδικού αριθμητικού κωδικού και συνοδεύεται από την αντίστοιχη περιγραφή της μορφολογικής μεταβολής που προκαλεί στο πρόσωπο. Ενδεικτικά παραδείγματα βασικών μονάδων περιλαμβάνουν: την AU1 (Inner Brow Raise, ανύψωση εσωτερικού τμήματος φρυδιού μέσω του μετωπιαίου μυός, *mus frontalis*), την AU4 (Brow Lowerer, κατέβασμα φρυδιών μέσω του συνοφρυωτικού μυός *corrugator supercilii*), την AU6 (Cheek Raiser, ανύψωση παρειών μέσω του σφιγκτήρα των βλεφάρων, *mus orbicularis oculi*), την AU12 (Lip Corner Puller, ανύψωση γωνιών χειλιών μέσω του μείζονος ζυγωματικού μυός, *mus zygomaticus major*) και την AU17 (Chin Raiser, ανύψωση πιγουνιού μέσω του γενειοκακού μυός, *mus mentalis*) (Ekman and Friesen, 1978) [\[19\]](#).

Στο πλαίσιο της κατηγορικής ταξινόμησης, κάθε βασικό συναίσθημα αντιστοιχεί σε έναν χαρακτηριστικό και στερεοτυπικό συνδυασμό Μονάδων Δράσης. Ειδικότερα, η συναισθηματική έκφραση της χαράς κωδικοποιείται πρωτίστως μέσω των AU6 και AU12, ενώ η εκδήλωση της λύπης χαρακτηρίζεται από τη σύγχρονη ενεργοποίηση των AU1, AU4, AU15 και AU17. Ο θυμός αποτυπώνεται ανατομικά μέσω του συνθετικού μοτίβου των AU4, AU5, AU7, AU23 και AU24, ο φόβος εκφράζεται από το σύνθετο πλέγμα των AU1, AU2, AU4, AU5, AU7, AU20 και AU26, η αηδία κωδικοποιείται από τις AU9, AU15 και AU16, ενώ η έκπληξη αναγνωρίζεται από την ταυτόχρονη παρουσία των AU1, AU2, AU5, AU26 και AU27 (σχήμα 2-2) (Ekman and Friesen, 1978) [\[19\]](#).



Σχήμα 2-2: Ο ηθοποιός Tim Roth παρουσιάζει εκφράσεις του προσώπου και την επεξήγησή τους

Αξιοσημείωτο είναι το γεγονός ότι ορισμένες συναισθηματικές κατηγορίες μοιράζονται κοινές Μονάδες Δράσης. Για παράδειγμα, οι εκφράσεις του φόβου και της έκπληξης παρουσιάζουν σημαντική ανατομική επικάλυψη, καθώς και οι δύο εμπεριέχουν τις μονάδες AU1, AU2 και AU5. Αυτή η κοινή μυϊκή δραστηριότητα εξηγεί σε θεωρητικό και ανατομικό επίπεδο τη συχνή υπολογιστική σύγχυση μεταξύ των εν λόγω συναισθημάτων κατά την αυτόματη ταξινόμηση, όπως αναδεικνύεται συστηματικά στους πίνακες σύγχυσης (confusion matrices) των σχετικών ερευνητικών εργασιών (Khan, 2022) [1].

Η εννοιολογική διασύνδεση του FACS με σύγχρονες υπολογιστικές βιβλιοθήκες, όπως το MediaPipe Face Landmarker, είναι άμεση, καθώς τα 52 ψηφιακά blendshapes που παράγει το εργαλείο αντιστοιχούν λειτουργικά στις μονάδες δράσης του FACS, παρέχοντας ένα ποσοτικό αποτύπωμα των κινήσεων του προσώπου. Παρά τη δυνατότητα άμεσης κανονιστικής ταξινόμησης βάσει των κανόνων του FACS, στην παρούσα διπλωματική εργασία υιοθετείται η προσέγγιση της end-to-end εκπαίδευσης ενός Συνελκτικού Νευρωνικού Δικτύου (CNN) στο σύνολο δεδομένων FER-2013. Η επιλογή αυτή αναθέτει στο ίδιο το δίκτυο την αυτόματη εκμάθηση της βέλτιστης αντιστοίχισης εικόνας και συναισθήματος, παρακάμπτοντας την ανάγκη για ρητή κωδικοποίηση FACS (Chowdary, Nguyen and Hemanth, 2023) [6].

2.4 Πολιτισμικές και ατομικές διαφορές στην έκφραση συναισθημάτων

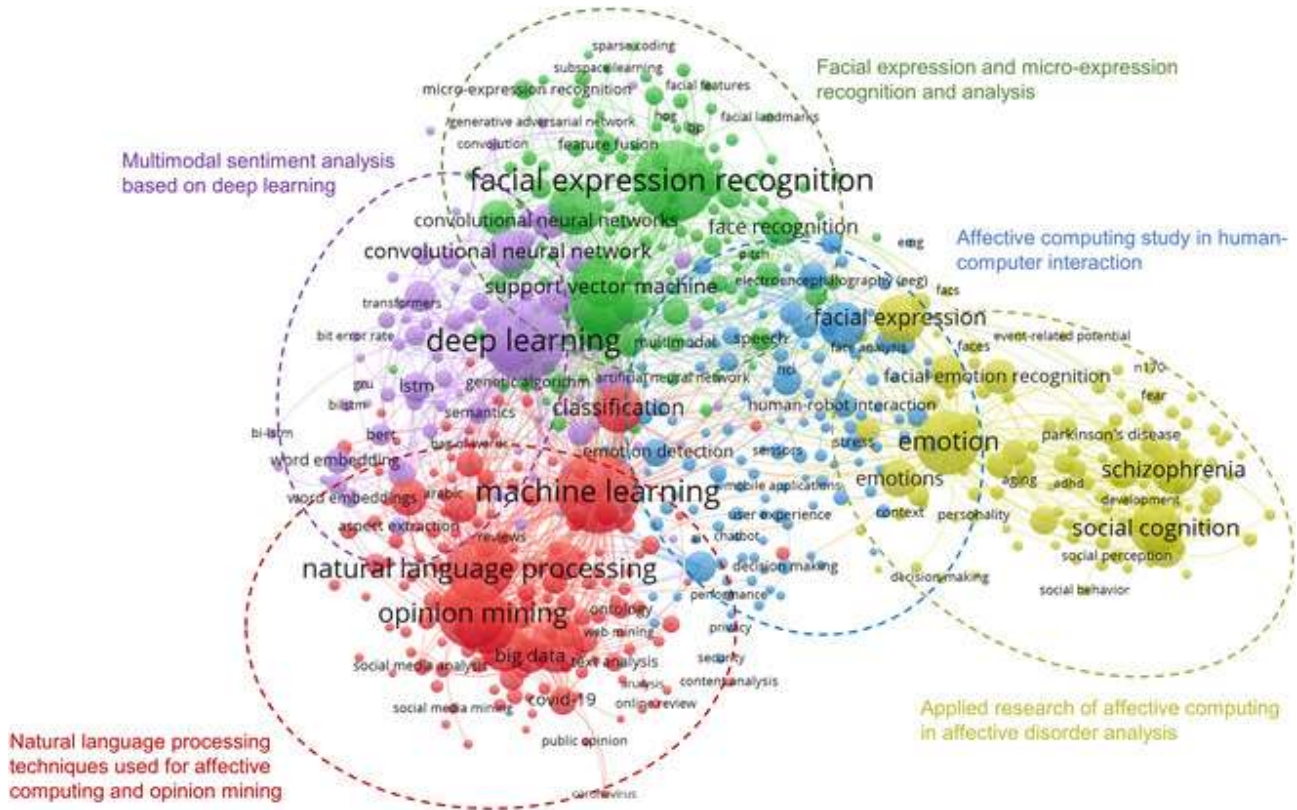
Παρά τη διατυπωμένη θέση του Ekman σχετικά με την παγκοσμιότητα των βασικών συναισθηματικών εκφράσεων, η μεταγενέστερη και σύγχρονη βιβλιογραφία αναδεικνύει σημαντικές πολιτισμικές διαφοροποιήσεις, οι οποίες επηρεάζουν καθοριστικά τόσο τον τρόπο εξωτερίκευσης όσο και τη διαδικασία αντίληψης και αναγνώρισης των συναισθημάτων μεταξύ ατόμων με διαφορετικό πολιτισμικό υπόβαθρο. Σε αυτό το πλαίσιο, η έννοια των κανόνων εμφάνισης (display rules) αναφέρεται στις κοινωνικά και πολιτισμικά διαμορφωμένες νόρμες που ρυθμίζουν το πότε, το πώς, αλλά και με ποια ένταση επιτρέπεται ή επιβάλλεται να εκφράζονται τα συναισθήματα στο δημόσιο πεδίο (Ekman, 2003) [18]. Για παράδειγμα, οι συλλογικιστικές κουλτούρες τείνουν συστηματικά να αποθαρρύνουν τη δημόσια εκδήλωση αρνητικών συναισθημάτων για λόγους διατήρησης της κοινωνικής αρμονίας. Ως εκ τούτου, τα αυτόματα υπολογιστικά μοντέλα που έχουν εκπαιδευτεί σχεδόν αποκλειστικά σε σύνολα δεδομένων δυτικού πληθυσμού εμφανίζουν αισθητά μειωμένη ακρίβεια ταξινόμησης όταν καλούνται να επεξεργαστούν εκφράσεις χρηστών από διαφορετικές πολιτισμικές περιοχές (Li and Deng, 2022) [7].

Επιπροσθέτως, η αυτόματη αναγνώριση επηρεάζεται σημαντικά από τις ατομικές διαφορές στην εκφραστικότητα (expressivity), δηλαδή από την έμφυτη ή επίκτητη τάση κάθε ατόμου να εξωτερικεύει τα συναισθήματά του μέσω ορατών μορφασμών του προσώπου. Τα άτομα με χαμηλή εκφραστικότητα εκδηλώνουν συναισθήματα χαμηλότερης έντασης και υψηλότερης λεπτότητας (subtle expressions), τα οποία καθίστανται ιδιαιτέρως δυσχερή στην ανίχνευση από αλγόριθμους που εκπαιδεύτηκαν σε τυπικές, τεχνητά παραγόμενες εκφράσεις εργαστηριακού περιβάλλοντος (posed expressions) (Ekman and Friesen, 1978) [19]. Το σύνολο δεδομένων FER-2013, αποτελούμενο από εικόνες που συλλέχθηκαν απευθείας από το διαδίκτυο, περιλαμβάνει τόσο προσποητές όσο και αυθόρμητες, φυσικές εκφράσεις. Αυτό το χαρακτηριστικό το καθιστά σημαντικά πιο αντιπροσωπευτικό των πραγματικών συνθηκών χρήσης σε σύγκριση με τα ελεγχόμενα εργαστηριακά σύνολα δεδομένων (datasets), αλλά ταυτόχρονα αυξάνει αισθητά τον βαθμό δυσκολίας της αυτόματης ταξινόμησης (Goodfellow et al., 2013) [30].

Οι παραπάνω θεωρητικοί και πρακτικοί περιορισμοί καθίστανται ιδιαίτερα κρίσιμοι για την ερμηνεία των αποτελεσμάτων αξιολόγησης που παρουσιάζονται αναλυτικά στο κεφάλαιο 8 της παρούσας εργασίας. Συγκεκριμένα, αναδεικνύουν ότι η υψηλή ακρίβεια ενός μοντέλου FER σε ένα αυστηρά προσαρμοσμένο ή εργαστηριακό περιβάλλον δοκιμών δεν συνεπάγεται απαραίτητα ανάλογη απόδοση σε πραγματικές συνθήκες αλληλεπίδρασης, όπου η ετερογένεια της πολιτισμικής προέλευσης και η διακύμανση της ατομικής εκφραστικότητας των χρηστών είναι ουσιαστικά μεγαλύτερες.

Κεφάλαιο 3ο: Τεχνητή νοημοσύνη και αναγνώριση συναισθημάτων

3.1 Βασικές αρχές και κατηγορίες της μηχανικής μάθησης



Σχήμα 3-1: Ερευνητικές θεματικές περιοχές στο πεδίο της υπολογιστικής του συναισθήματος (Affective Computing)

Η μηχανική μάθηση (Machine Learning, ML) αποτελεί κομβικό κλάδο της τεχνητής νοημοσύνης, ο οποίος εστιάζει στην ανάπτυξη αλγοριθμικών συστημάτων με την ικανότητα να εκπαιδεύονται από δεδομένα και να βελτιώνουν αυτόνομα την απόδοσή τους βάσει εμπειρίας, καταργώντας την ανάγκη για ρητό προγραμματισμό κάθε επιμέρους λειτουργίας (Khan, 2022) [1]. Η θεμελιώδης διαφορά της μηχανικής μάθησης από τον παραδοσιακό υπολογιστικό προγραμματισμό έγκειται στον μηχανισμό παραγωγής της γνώσης. Ενώ στα κλασικά συστήματα ο προγραμματιστής κωδικοποιεί εκ των προτέρων αυστηρούς κανόνες και λογικές αποφάσεων, στη μηχανική μάθηση το σύστημα εξάγει αυτόματα στατιστικές αναπαραστάσεις και σύνθετα μοτίβα από εκτεταμένα σύνολα παραδειγμάτων, τα οποία στη συνέχεια αξιοποιεί για τη γενίκευσή τους σε νέα δεδομένα (Li and Deng, 2022) [7]. Η ικανότητα αυτή της γενίκευσης, δηλαδή η αποτελεσματική εφαρμογή της αποκτηθείσας γνώσης σε αόρατες εισόδους εκτός του συνόλου εκπαίδευσης, συνιστά το κεντρικό ζητούμενο κάθε μοντέλου μηχανικής μάθησης και το βασικό κριτήριο αξιολόγησης της ποιότητάς του.

Ανάλογα με τη μορφή της διαθέσιμης πληροφορίας και τον τρόπο παροχής της κατά τη διαδικασία της εκπαίδευσης, τα μοντέλα μηχανικής μάθησης διακρίνονται σε τρεις θεμελιώδεις κατηγορίες, καθεμία από τις οποίες απευθύνεται σε διαφορετικού τύπου υπολογιστικά προβλήματα.

Στο πλαίσιο της επιβλεπόμενης μάθησης (supervised learning), το μοντέλο εκπαιδεύεται επί ενός συνόλου δεδομένων όπου κάθε παράδειγμα εισόδου συνοδεύεται από την αντίστοιχη επιθυμητή ετικέτα εξόδου, με στόχο την εκμάθηση μιας συνάρτησης αντιστοίχισης ικανής να παράγει ορθές προβλέψεις για νέες εισόδους. Η κατηγορία αυτή περιλαμβάνει προβλήματα ταξινόμησης (classification), στα οποία η έξοδος ανήκει σε ένα προκαθορισμένο και πεπερασμένο σύνολο διακριτών κατηγοριών, καθώς και προβλήματα παλινδρόμησης (regression), όπου η έξοδος λαμβάνει συνεχή αριθμητική τιμή. Η αυτόματη αναγνώριση συναισθηματικών εκφράσεων προσώπου, η οποία συνιστά το αντικείμενο της παρούσας διπλωματικής εργασίας, εντάσσεται στην κατηγορία της επιβλεπόμενης ταξινόμησης, καθώς κάθε εικόνα του συνόλου εκπαίδευσης φέρει σημασιολογική ετικέτα που αντιστοιχεί σε μία από τις επτά βασικές συναισθηματικές κατηγορίες (Goodfellow et al., 2013) [30].

Στη μη επιβλεπόμενη μάθηση (unsupervised learning), το μοντέλο τροφοδοτείται με δεδομένα που δεν φέρουν εκ των προτέρων ετικέτες και καλείται να ανακαλύψει αυτόνομα εσωτερικές δομές, ομαδοποιήσεις ή συμπίεσμένες αναπαραστάσεις. Αντιπροσωπευτικές εφαρμογές της κατηγορίας αυτής αποτελούν η ομαδοποίηση (clustering) μέσω αλγορίθμων όπως ο k-means, η μείωση διαστατικότητας μέσω της Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis, PCA), καθώς και η εκμάθηση αναπαραστάσεων μέσω αυτοκωδικοποιητών (autoencoders). Η προσέγγιση αυτή αξιοποιείται κυρίως σε σενάρια όπου η χειροκίνητη επισήμανση δεδομένων είναι υπολογιστικά ή οικονομικά απαγορευτική.

Τέλος, η ενισχυτική μάθηση (reinforcement learning) αφορά σενάρια όπου ένας αυτόνομος πράκτορας (agent) αλληλεπιδρά δυναμικά με ένα περιβάλλον, λαμβάνοντας επιβραβεύσεις ή ποινές ανάλογα με τις ενέργειες που επιλέγει, με απώτερο σκοπό την εκμάθηση μιας βέλτιστης πολιτικής (policy) που μεγιστοποιεί τη συσσωρευμένη ανταμοιβή σε μακροπρόθεσμο ορίζοντα. Η ενισχυτική μάθηση βρίσκει εφαρμογή και σε προσαρμοστικά συστήματα σύστασης μουσικής βάσει συναισθήματος, επιτρέποντας τη δυναμική αναπροσαρμογή των προτάσεων ανάλογα με τη συνεχή ανατροφοδότηση του χρήστη (Deldjoo et al., 2022) [13].

3.1.1 Τεχνητά νευρωνικά δίκτυα και βαθιά μάθηση

Η βαθιά μάθηση (Deep Learning) συνιστά μια εξειδικευμένη υποκατηγορία της μηχανικής μάθησης, η οποία βασίζεται στη χρήση τεχνητών νευρωνικών δικτύων (Artificial Neural Networks, ANNs) μεγάλου βάθους, δηλαδή υπολογιστικών αρχιτεκτονικών με πολλαπλά διαδοχικά επίπεδα επεξεργασίας. Η θεμελιώδης δομική μονάδα ενός τεχνητού νευρωνικού δικτύου είναι ο τεχνητός νευρώνας (perceptron), ο οποίος λαμβάνει ένα διάνυσμα σταθμισμένων εισόδων, υπολογίζει το σταθμισμένο άθροισμά τους προσθέτοντας έναν σταθερό όρο προκατάληψης (bias), και στη συνέχεια διοχετεύει το αποτέλεσμα σε μια μη γραμμική συνάρτηση ενεργοποίησης (activation function) για την παραγωγή της τελικής εξόδου του. Η εισαγωγή της μη γραμμικότητας αποτελεί καθοριστικό παράγοντα, καθώς, χωρίς αυτήν, οσοδήποτε μεγάλος αριθμός διαδοχικών γραμμικών μετασχηματισμών θα ισοδυναμούσε υπολογιστικά με έναν και μοναδικό γραμμικό μετασχηματισμό, περιορίζοντας δραστικά την εκφραστική ικανότητα του δικτύου και την απόδοσή του σε σύνθετα μη γραμμικά προβλήματα (Chowdary, Nguyen and Hemanth, 2023) [6].

Στην πράξη, μεταξύ των πλέον διαδεδομένων συναρτήσεων ενεργοποίησης συγκαταλέγεται η ReLU (Rectified Linear Unit), η οποία ορίζεται μαθηματικά ως $f(x) = \max(0, x)$ και προτιμάται συστηματικά στα κρυφά επίπεδα λόγω της υπολογιστικής της απλότητας αλλά και της αποτελεσματικής αντιμετώπισης του προβλήματος εξαφάνισης της κλίσης (vanishing gradient problem). Παράλληλα, στο επίπεδο εξόδου προβλημάτων πολυκατηγορικής ταξινόμησης εφαρμόζεται η συνάρτηση Softmax, η

οποία μετασχηματίζει τις ασταθμητές εξόδους (logits) σε μια ομαλοποιημένη κατανομή πιθανοτήτων επί των διαθέσιμων κλάσεων.

Δομικά, η αρχιτεκτονική ενός τυπικού βαθούς νευρωνικού δικτύου απαρτίζεται από τρία βασικά είδη επιπέδων: το επίπεδο εισόδου (input layer), το οποίο υποδέχεται τα ακατέργαστα δεδομένα, τα κρυφά επίπεδα (hidden layers), τα οποία εκτελούν διαδοχικούς μετασχηματισμούς εξάγοντας ιεραρχικά χαρακτηριστικά, και το επίπεδο εξόδου (output layer), το οποίο παράγει την τελική πρόβλεψη. Καθώς το βάθος του δικτύου αυξάνεται, τα παραγόμενα χαρακτηριστικά αποκτούν σταδιακά υψηλότερο βαθμό αφαιρετικότητας και σημασιολογικού πλούτου. Ειδικότερα, τα πρώτα επίπεδα εξάγουν τοπικά χαρακτηριστικά χαμηλού επιπέδου, όπως ακμές, γωνίες και χρωματικές διαβαθμίσεις, ενώ τα βαθύτερα επίπεδα συνθέτουν σύνθετες, υψηλού επιπέδου εννοιολογικές αναπαραστάσεις, όπως η μορφολογία ενός χαμογελαστού στόματος ή το σφίξιμο των φρυδιών (Li and Deng, 2022) [7].

Η διαδικασία εκπαίδευσης ενός βαθούς νευρωνικού δικτύου στηρίζεται στην ελαχιστοποίηση μιας κατάλληλα ορισμένης συνάρτησης απώλειας (loss function), η οποία ποσοτικοποιεί την απόκλιση μεταξύ των προβλεπόμενων εξόδων του μοντέλου και των πραγματικών ετικετών του συνόλου εκπαίδευσης. Για προβλήματα πολυκατηγορικής ταξινόμησης, όπως η αναγνώριση συναισθηματικών εκφράσεων, η πλέον ενδεδειγμένη συνάρτηση απώλειας είναι η διασταυρούμενη εντροπία (cross-entropy loss), η οποία επιβάλλει αυστηρότερες ποινές στα σφάλματα ταξινόμησης που συνοδεύονται από υψηλή εμπιστοσύνη του μοντέλου. Σε σενάρια όπου παρατηρείται ανισορροπία κλάσεων στο σύνολο δεδομένων, όπου ορισμένες συναισθηματικές κατηγορίες εκπροσωπούνται από σημαντικά μεγαλύτερο αριθμό δειγμάτων σε σχέση με άλλες, εφαρμόζεται η σταθμισμένη παραλλαγή της διασταυρούμενης εντροπίας (class-weighted cross-entropy loss). Στην περίπτωση αυτή, κάθε κατηγορία συνεισφέρει στη συνολική απώλεια με συντελεστή βαρύτητας ανάλογο της αντίστροφης συχνότητας εμφάνισής της, εξασφαλίζοντας έτσι την ισόρροπη εκπαίδευση του δικτύου και την αποφυγή μεροληψίας υπέρ των επικρατέστερων κλάσεων (Mahendran et al., 2024) [4].

Η βελτιστοποίηση και ελαχιστοποίηση της συνάρτησης απώλειας επιτυγχάνεται μέσω του αλγορίθμου οπισθοδιάδοσης σφάλματος (backpropagation), ο οποίος αξιοποιεί τον κανόνα της αλυσίδας του διαφορικού λογισμού για τον υπολογισμό των μερικών παραγώγων (κλίσεων) της απώλειας ως προς κάθε επιμέρους παράμετρο (βάρος και bias) του δικτύου. Στη συνέχεια, οι υπολογισθείσες κλίσεις (gradients) τροφοδοτούνται σε έναν αλγόριθμο βελτιστοποίησης, ο οποίος ενημερώνει τις παραμέτρους του δικτύου κατά την κατεύθυνση που ελαχιστοποιεί την απώλεια. Ο πλέον διαδεδομένος αλγόριθμος βελτιστοποίησης στη βαθιά μάθηση είναι ο Adam (Adaptive Moment Estimation), ο οποίος διατηρεί προσαρμοστικούς ρυθμούς μάθησης για κάθε παράμετρο ξεχωριστά, υπολογίζοντας εκτιμήσεις της πρώτης (μέση τιμή) και της δεύτερης ροπής (διασπορά) των κλίσεων. Η προσέγγιση αυτή προσφέρει ταχύτερη, σταθερότερη και αποτελεσματικότερη σύγκλιση σε σχέση με τον κλασικό αλγόριθμο στοχαστικής καταβίβασης κλίσης (Stochastic Gradient Descent, SGD) (Li and Deng, 2022) [7].

Επιπροσθέτως, στο πλαίσιο της προετοιμασίας και υλοποίησης της εκπαιδευτικής διαδικασίας, το διαθέσιμο σύνολο δεδομένων χωρίζεται αυστηρά σε τρία επιμέρους υποσύνολα: το σύνολο εκπαίδευσης (training set), το οποίο προορίζεται για τη συνεχή προσαρμογή των παραμέτρων του δικτύου, το σύνολο επικύρωσης (validation set), το οποίο αξιολογείται παράλληλα για την παρακολούθηση της ικανότητας γενίκευσης σε άορατα δεδομένα και την εφαρμογή τεχνικών πρώιμου τερματισμού (early stopping) προς αποφυγή της υπερπροσαρμογής (overfitting), καθώς και το σύνολο δοκιμής (test set), το οποίο χρησιμοποιείται για την τελική και αντικειμενική αξιολόγηση της απόδοσης του μοντέλου μετά το πέρας της εκπαίδευσης. Στην παρούσα διπλωματική εργασία, για την εκπαίδευση

και αξιολόγηση αξιοποιείται το σύνολο δεδομένων FER-2013, στο οποίο ο διαχωρισμός αυτός παρέχεται εκ των προτέρων από τους δημιουργούς του συνόλου των δεδομένων, περιλαμβάνοντας 28.709 εικόνες στο σύνολο εκπαίδευσης και 7.178 εικόνες στο σύνολο δοκιμής (Goodfellow et al., 2013) [30].

3.1.2 Μεταφορά γνώσης και προσαρμογή του μοντέλου

Ιδιαίτερης σημασίας για τις εφαρμογές ανάλυσης εικόνας σε περιβάλλοντα που χαρακτηρίζονται από περιορισμένη διαθεσιμότητα δεδομένων εκπαίδευσης και υπολογιστικών πόρων συνιστά η τεχνική της μεταφοράς γνώσης. Η προσέγγιση αυτή θεμελιώνεται στη διαπίστωση ότι τα κατώτερα και ενδιάμεσα επίπεδα ενός βαθέος νευρωνικού δικτύου, το οποίο έχει προεκπαιδευτεί σε σύνολα δεδομένων εξαιρετικά μεγάλης κλίμακας, όπως το ImageNet, που περιλαμβάνει περισσότερες από 14 εκατομμύρια εικόνες καταναμημένες σε 1.000 γενικές κατηγορίες, αποκτούν ισχυρές και γενικεύσιμες αναπαραστάσεις οπτικών χαρακτηριστικών. Τα χαρακτηριστικά αυτά (π.χ. ακμές, γωνίες, υφές, γεωμετρικά σχήματα και χρωματικές διαβαθμίσεις) φέρουν καθολικό χαρακτήρα και καθίστανται άμεσα μεταφέρσιμα και αξιοποιήσιμα σε ένα ευρύ φάσμα εργασιών όρασης υπολογιστών, ανεξάρτητα από το συγκεκριμένο αντικείμενο αναγνώρισης του τελικού προβλήματος (Sandler et al., 2018) [15].

Στην υπολογιστική πράξη, η μεταφορά γνώσης υλοποιείται κυρίως μέσω της τεχνικής της λεπτομερούς προσαρμογής (fine-tuning). Κατά τη διαδικασία αυτή, το προεκπαιδευμένο δίκτυο αξιοποιείται ως αρχικό σημείο εκκίνησης, αντικαθίσταται το τελικό επίπεδο ταξινόμησης ώστε να ανταποκρίνεται στον ακριβή αριθμό των κατηγοριών της νέας εργασίας, και στη συνέχεια ακολουθεί επανεκπαίδευση η οποία δομείται σε δύο διαδοχικές φάσεις. Στην πρώτη φάση της αρχικής εκπαίδευσης του ταξινομητή (Frozen Backbone), τα βάρη του κύριου προεκπαιδευμένου τμήματος του δικτύου (backbone) παραμένουν πλήρως παγωμένα (frozen). Εκπαιδεύεται αποκλειστικά το νέο προστιθέμενο επίπεδο ταξινόμησης με σχετικά υψηλό ρυθμό μάθησης, επιτρέποντας την ταχεία προσαρμογή των νέων παραμέτρων στις ιδιαιτερότητες του νέου προβλήματος χωρίς να διαταράσσονται οι προϋπάρχουσες αναπαραστάσεις. Στη δεύτερη φάση, της λεπτής προσαρμογής (Unfrozen Fine-Tuning), αποπαγώνουν τα τελευταία βαθύτερα επίπεδα του κορμού και επανεκπαιδεύονται παράλληλα με τον ταξινομητή, χρησιμοποιώντας ωστόσο ιδιαίτερα χαμηλό ρυθμό μάθησης (της τάξης του 10^{-4}). Η επιλογή αυτού του χαμηλού ρυθμού επιτρέπει την ήπια ευθυγράμμιση των εσωτερικών αναπαραστάσεων στις λεπτομέρειες των νέων δεδομένων, αποτρέποντας το φαινόμενο της καταστροφικής λήθης (catastrophic forgetting), δηλαδή την απώλεια της προαποκτηθείσας γενικής γνώσης του μοντέλου (Chowdhury, Idris and Abas, 2024) [3]. Η αποτελεσματικότητα και η μεθοδολογική υπεροχή της στρατηγικής λεπτομερούς προσαρμογής έναντι της απλής εξαγωγής χαρακτηριστικών (feature extraction), στην οποία τα βάρη του βασικού δικτύου παραμένουν αμετάβλητα καθ' όλη τη διάρκεια της διαδικασίας, έχει τεκμηριωθεί εκτενώς στη διεθνή βιβλιογραφία, παρουσιάζοντας βελτίωση της ακρίβειας ταξινόμησης κατά 1-10%. Η διαπίστωση αυτή καθιστά τη μεταφορά γνώσης με λεπτομερή προσαρμογή ως την καταλληλότερη μεθοδολογική πρακτική για την ανάπτυξη αποδοτικών μοντέλων αναγνώρισης συναισθημάτων στο πλαίσιο της παρούσας διπλωματικής εργασίας (Mahendran et al., 2024) [4].

3.2 Συνελκτικά νευρωνικά δίκτυα

Τα συνελκτικά νευρωνικά δίκτυα (Convolutional Neural Networks, CNNs) αποτελούν μια εξειδικευμένη και ιδιαίτερα ισχυρή κατηγορία βαθέων νευρωνικών δικτύων, η οποία σχεδιάστηκε συγκεκριμένα για την επεξεργασία και ανάλυση δεδομένων που διαθέτουν εγγενή χωρική δομή πλέγματος, όπως οι δισδιάστατες εικόνες και οι αλληλουχίες βίντεο. Σε αντίθεση με τα παραδοσιακά,

πλήρως συνδεδεμένα τεχνητά νευρωνικά δίκτυα (fully connected networks), στα οποία κάθε νευρώνας ενός επιπέδου συνδέεται με κάθε στοιχείο εισόδου του προηγούμενου επιπέδου, τα CNNs θεμελιώνονται σε τρεις κεντρικές αρχιτεκτονικές αρχές: την τοπική συνδεσιμότητα (local connectivity), την κοινή χρήση βαρών (weight/parameter sharing) και την ιεραρχική δομή αναπαράστασης χαρακτηριστικών. Οι ιδιότητες αυτές περιορίζουν δραστικά τον συνολικό αριθμό των προς εκπαίδευση παραμέτρων, ενώ παράλληλα προσδίδουν στο σύστημα υψηλή ανεκτικότητα σε χωρικές μετατοπίσεις, περιστροφές και τοπικές παραμορφώσεις της εισόδου (Li and Deng, 2022) [7]. Τα χαρακτηριστικά αυτά καθιστούν τα συνελκτικά δίκτυα την πλέον ενδεδειγμένη τεχνολογία για εργασίες υπολογιστικής όρασης, όπως η αυτόματη αναγνώριση συναισθηματικών εκφράσεων προσώπου, όπου τα κρίσιμα διαγνωστικά μοτίβα (όπως η κλίση των φρυδιών, η γωνία των χειλιών ή η πτύχωση του μετώπου) εντοπίζονται σε τοπικές περιοχές της εικόνας ανεξαρτήτως της ακριβούς χωρικής τους θέσης (Sarvakar and Rana, 2025) [9].

Δομικά, η αρχιτεκτονική ενός τυπικού συνελκτικού νευρωνικού δικτύου συγκροτείται από τη συστηματική διαδοχή και σύνθεση επιπέδων τριών βασικών κατηγοριών: τα συνελκτικά επίπεδα (convolutional layers), τα επίπεδα συγκέντρωσης (pooling layers) και τα πλήρως συνδεδεμένα επίπεδα (fully connected layers), τα οποία συνδυάζονται σε ειδικές αρχιτεκτονικές διατάξεις ανάλογα με τη φύση του προβλήματος.

Όσον αφορά τα συνελκτικά επίπεδα, το κεντρικό υπολογιστικό στοιχείο ενός CNN είναι η πράξη της συνέλιξης (convolution). Σε κάθε συνελκτικό επίπεδο ορίζεται ένα σύνολο από φίλτρα (filters ή kernels) διαθέσιμα για εκπαίδευση, συνήθως διαστάσεων 3×3 ή 5×5 pixels, τα οποία ολισθαίνουν συστηματικά πάνω στην εικόνα εισόδου. Αντίστοιχα, σε κάθε θέση υπολογίζεται το εσωτερικό γινόμενο μεταξύ του φίλτρου και της αντίστοιχης τοπικής περιοχής της εισόδου, παράγοντας ένα χάρτη χαρακτηριστικών (feature map) που αντικατοπτρίζει την παρουσία του συγκεκριμένου μοτίβου σε κάθε σημείο της εικόνας. Τα βάρη κάθε φίλτρου είναι σταθερά και κοινά για όλες τις θέσεις της εικόνας (parameter sharing), κάτι που εξασφαλίζει την ανίχνευση του ίδιου χαρακτηριστικού ανεξαρτήτως θέσης στην εικόνα (translation invariance) και μειώνει σημαντικά τον αριθμό εκπαιδευσιμων παραμέτρων σε σύγκριση με πλήρως συνδεδεμένα επίπεδα (Chowdary, Nguyen and Hemanth, 2023) [6]. Μετά τη συνέλιξη, εφαρμόζεται συνάρτηση ενεργοποίησης, συνήθως η ReLU (Rectified Linear Unit), για την εισαγωγή μη γραμμικότητας στο υπολογιστικό σύστημα.

Αναφορικά με τα επίπεδα συγκέντρωσης, μετά από ένα ή περισσότερα συνελκτικά επίπεδα, παρεμβάλλεται συνήθως ένα επίπεδο συγκέντρωσης (pooling layer) με σκοπό την σταδιακή μείωση των χωρικών διαστάσεων, των παραγόμενων χαρτών χαρακτηριστικών και επομένως του υπολογιστικού φόρτου των επόμενων επιπέδων, ενώ ταυτόχρονα ενισχύεται η ανεκτικότητα του δικτύου σε μικρές μεταφορές και στρεβλώσεις της εισόδου. Η πιο διαδεδομένη μέθοδος είναι η μέγιστη συγκέντρωση (max pooling), κατά την οποία κάθε περιοχή του χάρτη χαρακτηριστικών αντικαθίσταται από τη μέγιστη τιμή της, διατηρώντας τα πιο έντονα ενεργοποιημένα χαρακτηριστικά.

Προσθέτοντας, τα πλήρως συνδεδεμένα επίπεδα στο τελικό τμήμα της αρχιτεκτονικής, οι χάρτες χαρακτηριστικών ισοπεδώνονται (flatten) σε ένα μονοδιάστατο διάνυσμα και τροφοδοτούνται σε ένα ή περισσότερα πλήρως συνδεδεμένα επίπεδα, τα οποία συνδυάζουν τα εξαχθέντα χαρακτηριστικά για την παραγωγή της τελικής κατηγορικής απόφασης. Το τελικό πλήρως συνδεδεμένο επίπεδο διαθέτει τόσους νευρώνες όσες και οι κατηγορίες ταξινόμησης, ενώ η εφαρμογή συνάρτησης softmax μετατρέπει τις εξόδους σε κατανομή πιθανοτήτων επί των κατηγοριών (Li and Deng, 2022) [7].

Ένα από τα πλέον θεμελιώδη πλεονεκτήματα των συνελκτικών δικτύων έγκειται στην ικανότητά τους να εκτελούν ιεραρχική εξαγωγή χαρακτηριστικών αυξανόμενης σημασιολογικής πολυπλοκότητας όσο αυξάνεται το βάθος του δικτύου. Στα αρχικά συνελκτικά επίπεδα, τα φίλτρα ενεργοποιούνται από βασικά οπτικά στοιχεία χαμηλού επιπέδου, όπως οριζόντιες, κατακόρυφες και διαγώνιες ακμές, τοπικές υφές και χρωματικές κλίσεις. Στα ενδιάμεσα επίπεδα, οι πληροφορίες αυτές συνδυάζονται σε πιο σύνθετα μοτίβα, όπως τόξα, γωνίες και βασικά γεωμετρικά σχήματα, ενώ στα βαθύτερα επίπεδα οι αναπαραστάσεις αποκτούν υψηλό σημασιολογικό πλούτο. Ειδικότερα στο πρόβλημα της ανάλυσης εκφράσεων προσώπου, τα βαθύτερα επίπεδα αναπαριστούν εξειδικευμένες δομές, όπως η γωνία σχηματισμού των χειλιών, το ανασήκωμα ή η σύγκλιση των φρυδιών, καθώς και η διαστολή των ρουθουνιών, στοιχεία που συνδέονται άμεσα με συγκεκριμένες συναισθηματικές καταστάσεις (Khan, 2022) [11].

Η αυτοματοποιημένη αυτή ιεραρχική μάθηση αναπαραστάσεων (end-to-end feature learning) συνιστά το βασικό σημείο υπεροχής των CNNs έναντι των παραδοσιακών τεχνικών υπολογιστικής όρασης, οι οποίες βασίζονταν σε χειροκίνητα σχεδιασμένα χαρακτηριστικά (handcrafted features), όπως οι αλγόριθμοι Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG) ή τα φίλτρα Gabor. Σε αντίθεση με τις παραδοσιακές μεθόδους που παρήγαγαν δύσκαμπτους μηχανισμούς εξαγωγής, τα συνελκτικά νευρωνικά δίκτυα προσαρμόζουν αυτόματα τα φίλτρα τους στα δεδομένα εκπαίδευσης, βελτιστοποιώντας δυναμικά τη διαδικασία αναγνώρισης (Li and Deng, 2022) [7].

3.2.1 Η αρχιτεκτονική MobileNetV2

Παρά τις εντυπωσιακές επιδόσεις που καταγράφουν οι σύνθετες αρχιτεκτονικές των βαθέων συνελκτικών νευρωνικών δικτύων, η πρακτική τους ενσωμάτωση και εκτέλεση σε φορητές και κινητές συσκευές (mobile and edge devices) προσκρούει σε σημαντικούς τεχνικούς περιορισμούς. Οι περιορισμένοι διαθέσιμοι υπολογιστικοί πόροι, η περιορισμένη χωρητικότητα μνήμης, οι αυστηρές απαιτήσεις χαμηλού χρόνου απόκρισης κατά την εξαγωγή συμπερασμάτων (low inference latency) και η ανάγκη για δραστική εξοικονόμηση ενεργειακής καταναλώσεως καθιστούν την εφαρμογή παραδοσιακών, βαρέων μοντέλων ιδιαίτερα δυσχερή ή ακόμη και απαγορευτική σε περιβάλλοντα πραγματικού χρόνου. Προς την κατεύθυνση της αντιμετώπισης αυτών των προκλήσεων, αναπτύχθηκε η εξειδικευμένη αρχιτεκτονική MobileNetV2 (Sandler et al., 2018) [15], η οποία αποτελεί εξελιγμένη μορφή της οικογένειας υπολογιστικών μοντέλων MobileNet. Ο αρχιτεκτονικός σχεδιασμός του MobileNetV2 στοχεύει στην επίτευξη εξαιρετικά αποδοτικής εκτέλεσης υπολογιστικών συμπερασμάτων απευθείας επί της συσκευής, διατηρώντας στο ακέραιο την εκφραστική ικανότητα του δικτύου και επιτυγχάνοντας ελάχιστη δυνατή θυσία στην τελική ακρίβεια ταξινόμησης.

Η αρχιτεκτονική MobileNetV2 βασίζεται σε δύο καινοτόμες δομικές αρχές. Η πρώτη συνίσταται στην υιοθέτηση των διαχωρίσιμων συνελίξεων κατά βάθος (depthwise separable convolutions), οι οποίες αντικαθιστούν την κλασική, υπολογιστικά δαπανηρή πράξη της συνέλιξης με δύο διαδοχικούς υπολογιστικούς μηχανισμούς χαμηλότερης πολυπλοκότητας. Αρχικά, εφαρμόζεται μια συνέλιξη κατά βάθος (depthwise convolution), η οποία εκτελεί χωρική σάρωση ανεξάρτητα και μεμονωμένα σε κάθε κανάλι της εισόδου. Εν συνεχεία, ακολουθεί μια σημειακή συνέλιξη (pointwise convolution) διαστάσεων 1×1 , η οποία αναλαμβάνει να συνδυάσει γραμμικά τα αποτελέσματα των επιμέρους καναλιών για τη δημιουργία των τελικών αναπαραστάσεων. Ο δομικός αυτός διαχωρισμός επιτυγχάνει δραστική μείωση του συνολικού πλήθους των απαιτούμενων πολλαπλασιαστικών πράξεων κατά 8 έως 9 φορές σε σύγκριση με τις συμβατικές συνελίξεις, επιφέροντας αμελητέα μόνο έκπτωση στην εκφραστική δυναμική του μοντέλου (Sandler et al., 2018) [15].

Η δεύτερη θεμελιώδης καινοτομία αφορά την εισαγωγή των αναστραμμένων υπολειμματικών δομικών μονάδων με γραμμικούς στενωπούς (inverted residual blocks with linear bottlenecks). Στη συγκεκριμένη διάταξη, η αναπαράσταση χαμηλής διαστατικότητας επεκτείνεται αρχικά σε έναν χώρο υψηλότερης διαστατικότητας μέσω ενός επιπέδου επέκτασης (expansion layer) με σημειακή συνέλιξη 1×1 . Ακολούθως, εφαρμόζεται η συνέλιξη κατά βάθος για την αποτελεσματική εξαγωγή των χωρικών χαρακτηριστικών, και στη συνέχεια η πληροφορία συμπίεζεται εκ νέου σε έναν χώρο χαμηλής διαστατικότητας. Για την αποφυγή απώλειας κρίσιμων πληροφοριών, στο τελικό επίπεδο συμπίεσης δεν εφαρμόζεται μη γραμμική συνάρτηση ενεργοποίησης (linear bottleneck), καθώς η επιβολή μη γραμμικότητας σε χώρους χαμηλής διαστατικότητας έχει αποδειχθεί ότι καταστρέφει πολύτιμη σημασιολογική πληροφορία. Παράλληλα, η ενσωμάτωση συνδέσεων παράκαμψης (residual connections) απευθείας μεταξύ των στενωπών χαμηλής διαστατικότητας της εισόδου και της εξόδου κάθε δομικής μονάδας διευκολύνει την απρόσκοπτη ροή των κλίσεων κατά τη διαδικασία της οπισθοδιάδοσης, επιτρέποντας την εκπαίδευση βαθύτερων δικτύων χωρίς τον κίνδυνο εμφάνισης του φαινομένου εξαφάνισης της κλίσης (Sandler et al., 2018) [15].

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, η αρχιτεκτονική MobileNetV2, έχοντας προεκπαιδευτεί στο εκτενές σύνολο δεδομένων ImageNet, αξιοποιείται ως βασικός κορμός εξαγωγής χαρακτηριστικών για την εφαρμογή τεχνικών λεπτομερούς προσαρμογής στο σύνολο δεδομένων FER-2013. Σύμφωνα με τη σχετική ερευνητική βιβλιογραφία (Sharma, Nagpal and Raja Praveen, 2025) [2], η συγκεκριμένη μεθοδολογική προσέγγιση επιτυγχάνει υψηλά επίπεδα απόδοσης, καταγράφοντας ακρίβεια της τάξης του 73.82% και εμβαδόν κάτω από την καμπύλη ROC (Area Under the Curve, AUC) ίσο με 95.16%. Παράλληλα, ο απαιτούμενος χρόνος εκτέλεσης συμπερασμού διαμορφώνεται σε μόλις 43ms ανά εικόνα, επιβεβαιώνοντας την εξαιρετική υπολογιστική της επάρκεια. Τα ανωτέρω πειραματικά ευρήματα τεκμηριώνουν την καταλληλότητα του MobileNetV2 για την υλοποίηση εφαρμογών αναγνώρισης συναισθημάτων σε πραγματικό χρόνο (real-time execution) στο περιβάλλον φορητών συσκευών Android (Sharma, Nagpal and Raja Praveen, 2025) [2].

3.2.2 CNNs στην αναγνώριση εκφράσεων

Η εφαρμογή των CNNs στο πρόβλημα της αυτόματης αναγνώρισης εκφράσεων προσώπου (Facial Expression Recognition, FER) αποτελεί ενεργό πεδίο έρευνας τουλάχιστον από τα μέσα της δεκαετίας του 2010. Η κύρια πρόκληση του προβλήματος έγκειται στη μεγάλη ενδοκατηγορική μεταβλητότητα, δηλαδή η ίδια συναισθηματική έκφραση μπορεί να εκδηλώνεται με σημαντικά διαφορετικό τρόπο μεταξύ ατόμων διαφορετικής ηλικίας, φύλου και πολιτισμικού υπόβαθρου, καθώς και στη μικρή διακατηγορική μεταβλητότητα μεταξύ συγγενών εκφράσεων, όπως φόβος και έκπληξη ή θυμός και αηδία (Li and Deng, 2022) [7].

Το σύνολο δεδομένων FER-2013, που δημιουργήθηκε από τους Goodfellow et al. (2013) για τον διαγωνισμό αναγνώρισης συναισθημάτων ICML 2013, αποτελεί το πιο ευρέως χρησιμοποιούμενο benchmark για την αξιολόγηση μοντέλων FER. Περιλαμβάνει 35.887 εικόνες προσώπου σε κλίμακα του γκρι (grayscale) διαστάσεων 48×48 pixels, καταναμημένες σε επτά κατηγορίες συναισθήματος: θυμός (anger), αηδία (disgust), φόβος (fear), χαρά (happiness), λύπη (sadness), έκπληξη (surprise) και ουδέτερο (neutral). Αξιοσημείωτο χαρακτηριστικό του συνόλου δεδομένων είναι η έντονη ανισορροπία κλάσεων: η κατηγορία «χαρά» αριθμεί περίπου 8.989 εικόνες, ενώ η κατηγορία «αηδία» μόλις 547, δημιουργώντας πρόκληση για τα μοντέλα ταξινόμησης ως προς τη δίκαιη αναγνώριση σπάνιων εκφράσεων (Goodfellow et al., 2013) [30]. Η ανθρώπινη ακρίβεια στο FER-2013 εκτιμάται στο 65–70%, γεγονός που αναδεικνύει την εγγενή δυσκολία του συνόλου δεδομένων ακόμα και για την ανθρώπινη αντίληψη (Khan, 2022) [1]. Σύγχρονες αρχιτεκτονικές βαθιάς μάθησης και λεπτομερούς

προσαρμογής επιτυγχάνουν ακρίβεια στο εύρος 62–75% στο σύνολο δοκιμής, με τα καλύτερα αποτελέσματα να προέρχονται από σύνθετες αρχιτεκτονικές και τεχνικές επαύξησης δεδομένων (Chowdary, Nguyen and Hemanth, 2023) [6].

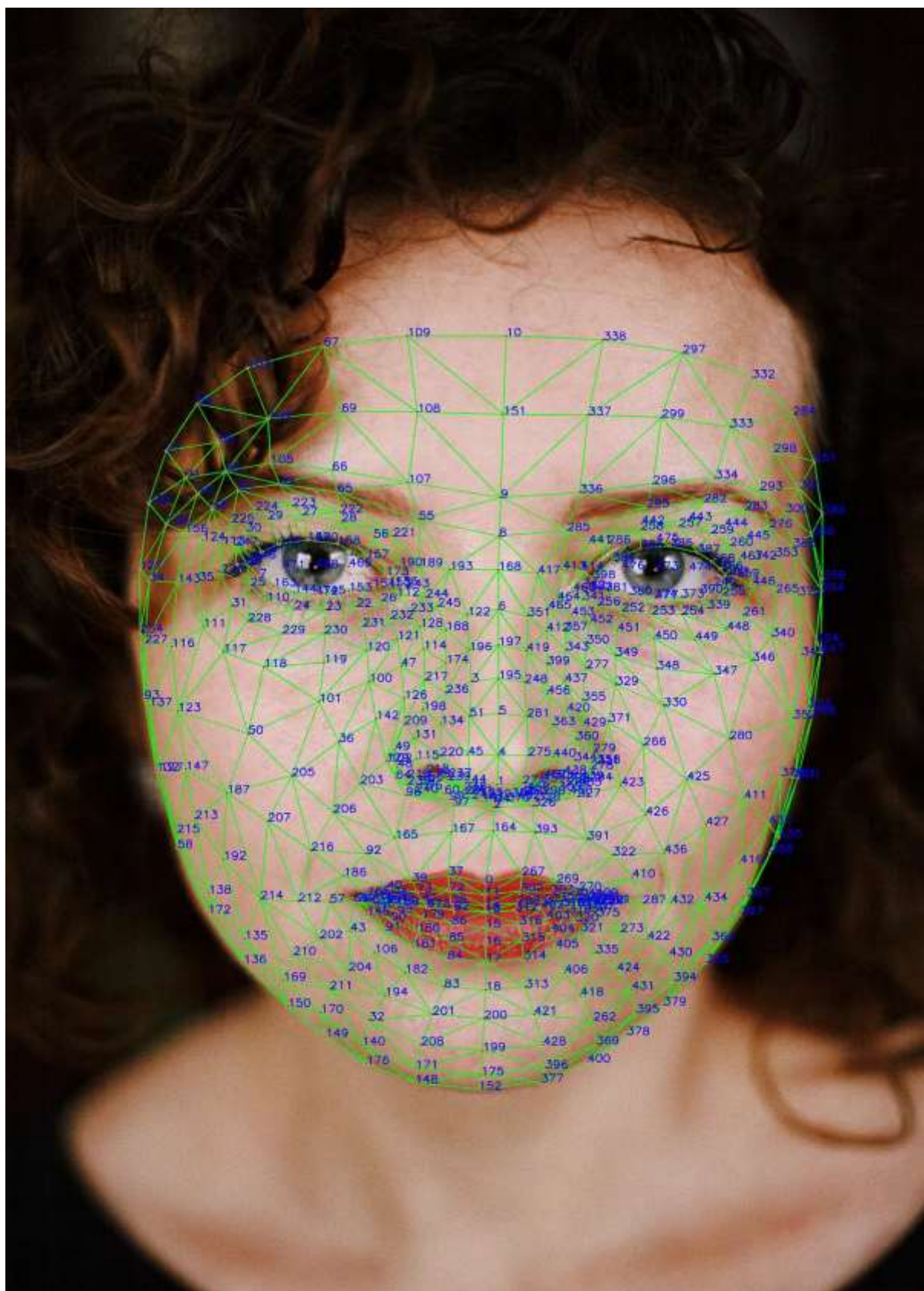
Η υιοθέτηση συνελκτικών νευρωνικών δικτύων (CNNs) αποτελεί την κυρίαρχη κατεύθυνση στη σύγχρονη βιβλιογραφία για την αναγνώριση συναισθημάτων προσώπου, όπως τεκμηριώνεται σε εκτενείς επισκοπήσεις του πεδίου (Raikar et al., 2023) [25].

3.3 Ανίχνευση και εντοπισμός χαρακτηριστικών προσώπου με MediaPipe

Προαπαιτούμενο για την αποτελεσματική εφαρμογή οποιουδήποτε υπολογιστικού μοντέλου αναγνώρισης συναισθήματος σε περιβάλλον πραγματικού χρόνου αποτελεί η αξιόπιστη, ακριβής και υπολογιστικά αποδοτική ανίχνευση και απομόνωση της περιοχής του προσώπου εντός του οπτικού πεδίου της κάμερας. Ιστορικά, η ανίχνευση προσώπου βασιζόταν σε κλασικές αλγοριθμικές προσεγγίσεις, με κυριότερη αυτή των Viola και Jones, η οποία αξιοποιεί τεχνικές εξαγωγής χαρακτηριστικών τύπου Haar και ταξινομητές κλιμακωτής διάταξης (cascade classifiers) για την ταχεία επεξεργασία εικόνων. Ωστόσο, οι παραδοσιακές αυτές μέθοδοι παρουσιάζουν σημαντικές λειτουργικές αδυναμίες όταν καλούνται να λειτουργήσουν σε ανεξέλεγκτα περιβάλλοντα, επιδεικνύοντας υψηλή ευαισθησία σε έντονες διακυμάνσεις του περιβάλλοντος φωτισμού, σε ακραίες γωνίες περιστροφής της κεφαλής, καθώς και σε περιπτώσεις μερικής απόκρυψης (occlusion) των χαρακτηριστικών του προσώπου (Li and Deng, 2022) [7].

Η αλματώδης ανάπτυξη των βαθέων συνελκτικών νευρωνικών δικτύων επέτρεψε τον σχεδιασμό ουσιαδώς ανώτερων συστημάτων ανίχνευσης. Στο πλαίσιο αυτό, η τεχνολογία MediaPipe Face Landmarker της Google ξεχωρίζει ως μία από τις πλέον διαδεδομένες και βελτιστοποιημένες λύσεις για εφαρμογές πραγματικού χρόνου σε φορητές συσκευές περιορισμένων υπολογιστικών πόρων (Google AI Edge, 2023) [22].

Η αρχιτεκτονική του MediaPipe Face Landmarker συνιστά ένα ολοκληρωμένο υπολογιστικό σύστημα ανίχνευσης και χαρτογράφησης προσώπου (όπως φαίνεται στο παρακάτω σχήμα 3-2), η λειτουργία του οποίου βασίζεται σε μια διαδοχική διοχέτευση ροής επεξεργασίας τριών εξειδικευμένων νευρωνικών μοντέλων, καθένα εκ των οποίων επιτελεί έναν διακριτό υπολογιστικό ρόλο (Google AI Edge, 2023) [22].



Σχήμα 3-2: Πυκνό πλέγμα 478 landmark 3D σημείων αναφοράς

Το πρώτο μοντέλο, γνωστό ως BlazeFace, είναι ένας ελαφρύς ανιχνευτής προσώπου βελτιστοποιημένος για εκτέλεση σε μονάδες επεξεργασίας γραφικών (GPU) κινητών συσκευών. Το BlazeFace λαμβάνει ως είσοδο το πλήρες καρέ βίντεο της κάμερας και παράγει ως έξοδο ορθογώνια πλαίσια οριοθέτησης (bounding boxes) που οριοθετούν κάθε ανιχνευόμενο πρόσωπο, μαζί με έξι βασικά σημεία αναφοράς (anchor landmarks) για την εκτίμηση της στάσης του κεφαλιού. Η αρχιτεκτονική του χαρακτηρίζεται από ιδιαίτερα μικρό αριθμό παραμέτρων και εκτελείται σε λιγότερο από 10ms σε σύγχρονα smartphones, επιτρέποντας real-time επεξεργασία βίντεο σε υψηλούς ρυθμούς ανανέωσης καρέ (fps) (Google AI Edge, 2023) [23]. Αξιοσημείωτο χαρακτηριστικό είναι ότι ο ανιχνευτής (tracker) καλείται μόνο όταν είναι απαραίτητος: μετά την αρχική ανίχνευση, το σύστημα χρησιμοποιεί τα landmarks του προηγούμενου καρέ για την αυτόματη περικοπή του επόμενου, επικαλούμενο εκ νέου τον ανιχνευτή μόνο σε περίπτωση απώλειας παρακολούθησης.

Παράλληλα, το δεύτερο μοντέλο είναι το πλέγμα προσώπου (Face Mesh), ένα νευρωνικό δίκτυο βαθιάς παλινδρόμησης που δέχεται ως είσοδο την περικομμένη περιοχή προσώπου που παρήγαγε το BlazeFace και παράγει 478 τρισδιάστατα γεωμετρικά σημεία αναφοράς (landmarks) καταναμημένα σε ολόκληρη την επιφάνεια του προσώπου, συμπεριλαμβανομένων των φρυδιών, των οφθαλμών, της μύτης, των χειλιών, του σαγονιού και του περιγράμματος. Κάθε landmark αναπαρίσταται ως τριάδα κανονικοποιημένων συντεταγμένων (x, y, z), όπου x και y αντιστοιχούν στη θέση εντός της εικόνας και z αντιπροσωπεύει το εκτιμώμενο βάθος σε σχέση με το κέντρο του προσώπου (Google AI Edge, 2023) [23]. Η πυκνή αυτή χωρική χαρτογράφηση επιτρέπει την ακριβή αναπαράσταση της τρισδιάστατης γεωμετρίας του προσώπου κάνοντας χρήση μίας και μόνο συμβατικής κάμερας RGB, χωρίς την ανάγκη ειδικών αισθητήρων βάθους.

Τέλος, το τρίτο μοντέλο είναι ο προβλεπτής μορφών ανάμειξης (blendshape predictor), ο οποίος λαμβάνει ως είσοδο την έξοδο του Face Mesh και παράγει 52 βαθμωτές τιμές (blendshape scores) στο διάστημα $[0, 1]$, καθεμία εκ των οποίων αντιπροσωπεύει το βαθμό ενεργοποίησης μιας συγκεκριμένης κίνησης μυϊκής ομάδας του προσώπου. Παραδείγματα blendshapes περιλαμβάνουν τα browDownLeft, browDownRight (κατέβασμα φρυδιών), mouthSmileLeft, mouthSmileRight (χαμόγελο), eyeWideLeft, eyeWideRight (άνοιγμα ματιών) και jawOpen (άνοιγμα στόματος). Τα blendshapes αυτά αντιστοιχούν εννοιολογικά στις μονάδες δράσης (Action Units) του συστήματος κωδικοποίησης δράσης προσώπου (Facial Action Coding System - FACS) που περιγράφεται στο κεφάλαιο 2, παρέχοντας έτσι ένα ποσοτικό φορτίο κινήσεων προσώπου υψηλής εκφραστικής αξίας (Google AI Edge, 2023) [22].

3.3.1 Ρόλος του MediaPipe στη διαδικασία αναγνώρισης συναισθήματος

Η σαφής διάκριση του ρόλου του MediaPipe Face Landmarker κρίνεται απαραίτητη εντός της συνολικής αρχιτεκτονικής (pipeline) αναγνώρισης συναισθήματος που υλοποιείται στην παρούσα εργασία. Το MediaPipe εκτελεί αποκλειστικά τη φάση της ανίχνευσης, της χωρικής απομόνωσης και της προεπεξεργασίας του προσώπου: εντοπίζει το πρόσωπο στο καρέ βίντεο, παράγει τα 478 landmarks και τα 52 blendshapes, και παραδίδει την απομονωμένη και κανονικοποιημένη εικόνα της περιοχής προσώπου ως είσοδο στο επόμενο στάδιο. Η ταξινόμηση του συναισθήματος σε μία από τις επτά διακριτές κατηγορίες δεν αποτελεί λειτουργία του MediaPipe, αλλά εκτελείται αποκλειστικά από το συνελκτικό νευρωνικό δίκτυο (MobileNetV2 CNN), το οποίο έχει υποστεί εξειδικευμένη εκπαίδευση, μέσω της στρατηγικής της λεπτομερούς προσαρμογής, στο σύνολο δεδομένων FER-2013 το οποίο περιγράφεται στο κεφάλαιο 6.1 (Google AI Edge, 2023) [23].

Η αρχιτεκτονική επιλογή χρήσης του MediaPipe για την πρώτη φάση της ροής επεξεργασίας προσφέρει τρία πρακτικά πλεονεκτήματα. Πρώτον, απαλλάσσει το κύριο μοντέλο ταξινόμησης (CNN)

από την ανάγκη εντοπισμού θέσης του προσώπου, επιτρέποντάς του να εστιάσει αποκλειστικά στην αναγνώριση των συναισθηματικών μοτίβων, γεγονός που οδηγεί σε υψηλότερη ακρίβεια ταξινόμησης. Δεύτερον, εξασφαλίζει ότι η είσοδος που τροφοδοτείται στο CNN είναι σταθερά περικομμένη και επικεντρωμένη στο πρόσωπο, εξαλείφοντας τις αρνητικές επιδράσεις του περιβάλλοντος φόντου και των εξωτερικών οπτικών παρεμβολών. Τρίτον, το MediaPipe εκτελείται εξ ολοκλήρου τοπικά στη συσκευή (on-device execution) χωρίς την ανάγκη αποστολής οπτικών ή βιομετρικών δεδομένων σε εξωτερικούς διακομιστές. Η τοπική αυτή επεξεργασία διαφυλάσσει την ιδιωτικότητα των ευαίσθητων προσωπικών δεδομένων του χρήστη, ικανοποιώντας πλήρως τις αυστηρές απαιτήσεις του Γενικού Κανονισμού Προστασίας Δεδομένων (GDPR).

3.4 Αξιολόγηση μοντέλων ταξινόμησης

Η αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης αποτελεί αναπόσπαστο τμήμα κάθε ερευνητικής εργασίας στο πεδίο της αναγνώρισης συναισθημάτων, καθώς παρέχει ποσοτικά μέτρα που επιτρέπουν τη σύγκριση μεταξύ διαφορετικών προσεγγίσεων και τον εντοπισμό αδυναμιών του μοντέλου ανά κατηγορία συναισθήματος. Στο παρόν πλαίσιο, η αξιολόγηση πραγματοποιείται στο ανεξάρτητο σύνολο δοκιμής του FER-2013, το οποίο αποτελείται από 7.178 εικόνες (ενοποίηση PublicTest και PrivateTest). Το σύνολο αυτό δεν έχει χρησιμοποιηθεί σε κανένα στάδιο της διαδικασίας εκπαίδευσης ή επικύρωσης, εξασφαλίζοντας έτσι μια πλήρως αντικειμενική εκτίμηση της ικανότητας γενίκευσης του μοντέλου.

Για τη συνολική και επιμέρους αποτίμηση της συμπεριφοράς του ταξινομητή αξιοποιούνται οι βασικές μετρικές αξιολόγησης της μηχανικής μάθησης, οι οποίες προσφέρουν συμπληρωματικές πληροφορίες σχετικά με την ακρίβεια των προβλέψεων.

Η πρώτη βασική μετρική αξιολόγησης είναι η ακρίβεια (Accuracy), η οποία ορίζεται ως το πηλίκο των ορθών προβλέψεων προς το συνολικό πλήθος παραδειγμάτων, όπως παρακάτω που αναφέρεται στη σχέση (3.1):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3.1)$$

όπου: TP (True Positives) είναι τα ορθώς ταξινομημένα θετικά παραδείγματα, TN (True Negatives) τα ορθώς ταξινομημένα αρνητικά, FP (False Positives) τα λανθασμένα θετικά και FN (False Negatives) τα λανθασμένα αρνητικά.

Παρά την απλότητα και την ευρεία χρήση της, η συνολική ακρίβεια μπορεί να παρέχει παραπλανητικές ενδείξεις σε προβλήματα με έντονη ανισορροπία κλάσεων (class imbalance). Για παράδειγμα, στο σύνολο FER-2013, ένα απλοϊκό μοντέλο που θα προέβλεπε συστηματικά την κλάση «χαρά» θα επιτύγχανε ακρίβεια της τάξης του ~25%, χωρίς ωστόσο να διαθέτει την ικανότητα αναγνώρισης κανενός άλλου συναισθήματος (Chowdary, Nguyen and Hemanth, 2023) [6].

Η δεύτερη βασική μετρική αξιολόγηση ονομάζεται Precision, Recall και F1-score, κατά την οποία υπολογίζονται για κάθε κατηγορία c ξεχωριστά οι μετρικές precision (ακρίβεια πρόβλεψης), recall (ευαισθησία ανίχνευσης) και F1-score (αρμονικός μέσος precision και recall) όπως φαίνονται στις σχέσεις (3.2) και (3.3) παρακάτω:

$$Precision_c = \frac{TP_c}{(TP_c + FP_c)} \quad (3.2)$$

$$Recall_c = \frac{TP_c}{(TP_c + FN_c)} \quad (3.3)$$

$$F1_c = \frac{2 \times (Precision_c \times Recall_c)}{(Precision_c + Recall_c)} \quad (3.4)$$

Για τη συνολική αξιολόγηση σε προβλήματα με ανισορροπία κλάσεων, προσδιορίζεται το macro-averaged F1-score, το οποίο υπολογίζει τον αριθμητικό μέσο των F1-scores όλων των κατηγοριών χωρίς στάθμιση ανάλογα με την κατανομή τους, παρέχοντας έτσι ίση βαρύτητα σε κάθε κατηγορία ανεξαρτήτως πλήθους παραδειγμάτων (Mahendran et al., 2024) [4]. Η μετρική αυτή, είναι ιδιαίτερα κατάλληλη για το FER-2013, στο οποίο η υπο-εκπροσωπούμενη κατηγορία «αηδία» πρέπει να αξιολογείται ισότιμα με την πλούσια σε δείγματα κατηγορία «χαρά», όπως φαίνεται στην σχέση (3.4).

Τελευταία βασική μετρική αξιολόγηση, αποτελεί η Confusion Matrix. Ειδικότερα, ο πίνακας σύγχυσης (confusion matrix) αποτελεί εργαλείο οπτικής ανάλυσης που αναπαριστά τη συχνότητα όλων των συνδυασμών πραγματικής (ground truth) και προβλεπόμενης κατηγορίας σε μορφή πίνακα $C \times C$, όπου C ο αριθμός κατηγοριών υπό εξέταση. Η κύρια διαγώνιος αντιστοιχεί στις ορθές ταξινομήσεις (True Positives), ενώ τα εκτός διαγωνίου στοιχεία αποκαλύπτουν τα συστηματικά σφάλματα του μοντέλου. Στο πλαίσιο της αναγνώρισης συναισθημάτων, ο πίνακας σύγχυσης είναι ιδιαίτερα πληροφοριακός καθώς αναδεικνύει τα ζεύγη εκφράσεων που το μοντέλο τείνει να συγχέει, π.χ. φόβος/έκπληξη ή θυμός/αηδία. Η λεπτομερής αυτή χαρτογράφηση των σφαλμάτων παρέχει τις απαραίτητες κατευθύνσεις για τη στοχευμένη βελτίωση του συστήματος, είτε μέσω της ποιοτικής ενίσχυσης του συνόλου δεδομένων εκπαίδευσης είτε μέσω της βελτιστοποίησης της εξαγωγής χαρακτηριστικών (Khan, 2022) [1].

3.5 Συναισθηματική υπολογιστική: επισκόπηση πεδίου

Η συναισθηματική υπολογιστική (Affective Computing) ορίζεται ως ο κλάδος της επιστήμης υπολογιστών και της τεχνητής νοημοσύνης που αντικείμενο έχει τη μελέτη, τη σχεδίαση και την ανάπτυξη συστημάτων τα οποία διαθέτουν την ικανότητα να αναγνωρίζουν, να ερμηνεύουν, να επεξεργάζονται και να προσομοιώνουν την ανθρώπινη συναισθηματική συμπεριφορά και εκδήλωση (Picard, 1997) [17]. Η ιδέα αναπτύχθηκε από την Rosalind Picard στο MIT Media Lab, η οποία υποστήριξε ότι η ικανότητα των υπολογιστών να «κατανοούν» συναισθήματα, είναι θεμελιώδης για την ανάπτυξη φυσικής και αποτελεσματικής αλληλεπίδρασης ανθρώπου-υπολογιστή. Το πεδίο διασταυρώνεται με την ψυχολογία, τη γνωσιακή επιστήμη (Cognitive Science), την τεχνητή νοημοσύνη και την αποδοτική αλληλεπίδραση ανθρώπου με υπολογιστή (Human-Computer Interaction - HCI), αντλώντας γνώση και μεθόδους από όλους αυτούς τους κλάδους.

Τα συστήματα αναγνώρισης συναισθήματος διακρίνονται σε μονότροπα (unimodal), τα οποία αξιοποιούν μία μεμονωμένη πηγή/κανάλι πληροφορίας, όπως τις οπτικές εκφράσεις του προσώπου, την ακουστική ανάλυση της φωνής, το γραπτό κείμενο ή τα φυσιολογικά βιοσήματα (π.χ. ηλεκτροκαρδιογράφημα, γαλβανική αντίδραση δέρματος). Καθώς και σε πολυτροπικά (multimodal), τα οποία συνδυάζουν ταυτόχρονα πολλαπλές ετερογενείς πηγές πληροφορίας, στοχεύοντας στην επίτευξη υψηλότερης ανθεκτικότητας (robustness) και ακρίβειας αναγνώρισης (Zhang et al., 2023) [10]. Η παρούσα διπλωματική εργασία εντάσσεται στην κατηγορία των μονότροπων συστημάτων βάσει οπτικής ανάλυσης εικόνων προσώπου. Η συγκεκριμένη αρχιτεκτονική επιλογή υπαγορεύεται από δύο βασικούς λειτουργικούς παράγοντες όπου η πρώτη, η μη παρεμβατική του φύση (Non-intrusiveness), δεν απαιτεί την εφαρμογή φυσικών αισθητήρων στο σώμα του χρήστη ούτε την επιβολή φωνητικών

αλληλεπιδράσεων, διατηρώντας τη φυσικότητα της αλληλεπίδρασης. Ενώ η δεύτερη, η προσβασιμότητα σε hardware υλικού (Hardware Availability), αξιοποιεί απευθείας τις ενσωματωμένες κάμερες που διαθέτει το σύνολο των σύγχρονων έξυπνων κινητών συσκευών (smartphones), καθιστώντας την εφαρμογή άμεσα προσβάσιμη και λειτουργική σε πραγματικά περιβάλλοντα καθημερινής χρήσης.

Παρά τη ραγδαία εξέλιξη των αλγορίθμων μηχανικής μάθησης, αυτό το πεδίο αντιμετωπίζει σημαντικές επιστημολογικές και λειτουργικές προκλήσεις. Κυριότερη αυτών αποτελεί ο διαχωρισμός μεταξύ του εσωτερικού συναισθήματος (felt emotion) και του εξωτερικά εκφραζόμενου συναισθήματος (expressed emotion). Οι άνθρωποι συχνά καταστέλλουν, τροποποιούν ή συγκαλύπτουν τις εκφράσεις τους υπό την επίδραση κοινωνικών κανόνων συμπεριφοράς (social display rules), με αποτέλεσμα η οπτικά ανιχνευόμενη έκφραση του προσώπου να μην αντικατοπτρίζει πάντοτε με ακρίβεια την υποκείμενη συναισθηματική κατάσταση. Επιπλέον, οι πολιτισμικές διαφοροποιήσεις στον τρόπο έκφρασης και αντίληψης των συναισθημάτων επηρεάζουν τη δυνατότητα γενίκευσης (generalizability) των μοντέλων όταν αυτά καλούνται να λειτουργήσουν σε πληθυσμούς διαφορετικού υπόβαθρου από αυτό των δεδομένων εκπαίδευσης (Li and Deng, 2022) [7].

Παρά τις εγγενείς αυτές δυσκολίες, η ενσωμάτωση τεχνικών συναισθηματικής υπολογιστικής σε πεδία όπως η εξατομικευμένη σύσταση περιεχομένου και μουσικής, η υποστηρικτική τεχνολογία στην ψυχική υγεία, η προσαρμοστική ηλεκτρονική μάθηση (e-learning) και η ψηφιακή ψυχαγωγία παρουσιάζει αλματώδη ανάπτυξη. Η εξέλιξη αυτή καθιστά τη συναισθηματική υπολογιστική έναν από τους πλέον υποσχόμενους και δυναμικούς τομείς αιχμής της σύγχρονης Επιστήμης Υπολογιστών (Picard, 1997) [17].

Κεφάλαιο 4ο: Μουσική και συναισθηματική προσαρμογή

4.1 Μουσική και ανθρώπινο συναίσθημα

Η σχέση μεταξύ μουσικής και συναισθήματος αποτελεί αντικείμενο συστηματικής επιστημονικής διερεύνησης εδώ και δεκαετίες, με ευρήματα που τεκμηριώνουν σταθερά την ικανότητα της μουσικής να επηρεάζει, να ρυθμίζει και να μετασχηματίζει τη συναισθηματική κατάσταση του ακροατή. Η μουσική δεν αποτελεί απλώς μέσο αισθητικής απόλαυσης, αλλά λειτουργεί ως ένα από τα ισχυρότερα εργαλεία συναισθηματικής ρύθμισης (emotion regulation) που διαθέτει ο άνθρωπος στην καθημερινή του ζωή (Juslin and Sloboda, 2010) [20].

Ως συναισθηματική ρύθμιση ορίζεται το δυναμικό σύνολο των γνωσιακών, συμπεριφορικών και νευροφυσιολογικών στρατηγικών που επιστρατεύει το άτομο με σκοπό την τροποποίηση του είδους, της έντασης, της χρονικής διάρκειας και της έκφρασης των βιωματικών του συναισθημάτων, είτε προς την κατεύθυνση της ενίσχυσης των θετικών συναισθηματικών καταστάσεων είτε προς την άμβλυνση των αρνητικών συναισθηματικών φορτίων.

Έρευνες στο πεδίο της ψυχολογίας της μουσικής έχουν αναδείξει ότι οι άνθρωποι χρησιμοποιούν ενεργητικά τη μουσική για διαφορετικούς ρυθμιστικούς σκοπούς ανάλογα με τη συναισθηματική τους κατάσταση και το πλαίσιο ακρόασης. Συγκεκριμένα, η βιβλιογραφία διακρίνει δύο βασικές στρατηγικές μουσικής ακρόασης: την ισομορφική (Iso Principle), κατά την οποία ο ακροατής επιλέγει μουσικά ερεθίσματα που αντιστοιχούν πλήρως στην τρέχουσα συναισθηματική του διάθεση (π.χ. επιλογή αργού ρυθμού και μελαγχολικής αρμονικής δομής σε καταστάσεις θλίψης). Η προσέγγιση αυτή προσφέρει συναισθηματική επικύρωση (emotional validation), διευκολύνοντας την επεξεργασία και την αποδοχή της υφιστάμενης βιωματικής κατάστασης. Καθώς και τη μετασχηματιστική (transformative), κατά την οποία ο ακροατής επιλέγει μουσική της οποίας τα υφολογικά και ακουστικά χαρακτηριστικά διαφοροποιούνται από την τρέχουσα διάθεσή του, στοχεύοντας στην ενεργή τροποποίηση και βελτίωσή της (π.χ. ακρόαση μουσικής υψηλής ενεργητικότητας για την αντιμετώπιση της κοπώσεως ή της πλήξης). (Juslin and Sloboda, 2010) [20]. Η σχεδίαση και υλοποίηση ενός ευφυούς συστήματος σύστασης μουσικής βάσει συναισθήματος (Emotion-Based Music Recommendation System) οφείλει να ενσωματώνει τη διπολική αυτή στρατηγική. Η παροχή της δυνατότητας στον χρήστη να επιλέξει είτε την αντικατοπτρική (iso-principle) είτε τη μετασχηματιστική (transformative) λειτουργία σύστασης συνιστά κρίσιμη παράμετρο για την προσαρμοστικότητα και την αποτελεσματικότητα της εφαρμογής.

Παράλληλα, ποσοτικά ευρήματα από μεγάλης κλίμακας μελέτες επιβεβαιώνουν τη διαδεδομένη χρήση της μουσικής ως στρατηγικής ρύθμισης συναισθήματος. Ειδικότερα, έρευνα που εξέτασε τις στρατηγικές αντιμετώπισης άγχους κατά τη διάρκεια της πανδημίας COVID-19 ανέδειξε ότι ακουστικά χαρακτηριστικά Spotify όπως υψηλό valence δηλαδή θετικότητα, energy ως ενεργητικότητα, danceability για την καταλληλότητα της κίνησης και του χωρού, ακόμα και tempo στην οποία αντιστοιχεί η ταχύτητα εκτέλεσης σε κτυπήματα ανά λεπτό (BPM). Αυτά τα χαρακτηριστικά συσχετίζονται σημαντικά με τη μουσική που επιλέγουν οι χρήστες για συναισθηματική αντιμετώπιση δύσκολων καταστάσεων (Chong et al., 2024) [24]. Επιπλέον, πειραματικές μελέτες έχουν τεκμηριώσει ότι η ακρόαση μουσικής προκαλεί μετρήσιμες νευροφυσιολογικές αντιδράσεις, συμπεριλαμβανομένης της απελευθέρωσης ντοπαμίνης στο ραβδωτό σώμα (striatum), κατά την ακρόαση μουσικής υψηλής συναισθηματικής αξίας, επιβεβαιώνοντας τη νευροβιολογική βάση της συναισθηματικής επίδρασης της μουσικής (Thoma et al., 2013) [12].

Ιδιαίτερα εκτεταμένη είναι η επιστημονική τεκμηρίωση σχετικά με την ευεργετική επίδραση της μουσικής στη μείωση του άγχους (stress reduction) σε κλινικά και μη κλινικά περιβάλλοντα. Συστηματικές αναλύσεις επιβεβαιώνουν ότι η στοχευμένη ακρόαση μουσικής επιφέρει σημαντική μείωση των επιπέδων της κορτιζόλης στο σάλιο, το οποίο αποτελεί τον πρωτεύοντα βιολογικό δείκτη ορμονικού στρες, ενώ παράλληλα προκαλεί καταστολή της δραστηριότητας του συμπαθητικού νευρικού συστήματος, επιφέροντας μείωση του καρδιακού ρυθμού και της αρτηριακής πίεσης (Thoma et al., 2013) [12].

Τα αποτελέσματα αυτά έχουν άμεση πρακτική εφαρμογή στην αρχιτεκτονική του προτεινόμενου συστήματος: η αυτόματη ανίχνευση συναισθηματικών καταστάσεων υψηλής διέγερσης και αρνητικού σθένους (όπως ο φόβος ή ο θυμός) ενεργοποιεί αλγορίθμους σύστασης που προκρίνουν μουσικά κομμάτια χαμηλής ενεργητικότητας (low energy), χαμηλού τέμπο (low tempo) και υψηλής ακουστικότητας (high acousticness), στοχεύοντας στη διευκόλυνση της συναισθηματικής αποφόρτισης και χαλάρωσης του χρήστη κατά την καθημερινή χρήση, χωρίς αυτό να συνεπάγεται κλινική ή θεραπευτική αποτελεσματικότητα.

4.2 Μουσικοθεραπεία: Θεωρητικό υπόβαθρο και εφαρμογές

Η μουσικοθεραπεία (Music Therapy) αποτελεί κλινική εφαρμογή της επιστήμης της ψυχολογίας της μουσικής, κατά την οποία εκπαιδευμένοι θεραπευτές αξιοποιούν μουσικές παρεμβάσεις για την επίτευξη θεραπευτικών στόχων σε πλαίσιο ιατρικής, ψυχολογικής ή εκπαιδευτικής φροντίδας (Juslin and Sloboda, 2010) [20]. Ως εφαρμοσμένος κλάδος της ψυχολογίας της μουσικής, η μουσικοθεραπεία δεν αντιμετωπίζει τη μουσική ως απλή ψυχαγωγική δραστηριότητα, αλλά ως ένα δομημένο, πολυδιάστατο ερέθισμα ικανό να προκαλέσει στοχευμένες νευροφυσιολογικές, γνωσιακές και συναισθηματικές μεταβολές.

Το θεωρητικό υπόβαθρο του πεδίου διακρίνεται για τον εγγενώς διεπιστημονικό του χαρακτήρα, αντλώντας αρχές και ευρήματα από τη γνωσιακή νευροεπιστήμη, την κλινική ψυχολογία, την ψυχοακουστική και τη φυσιολογία.

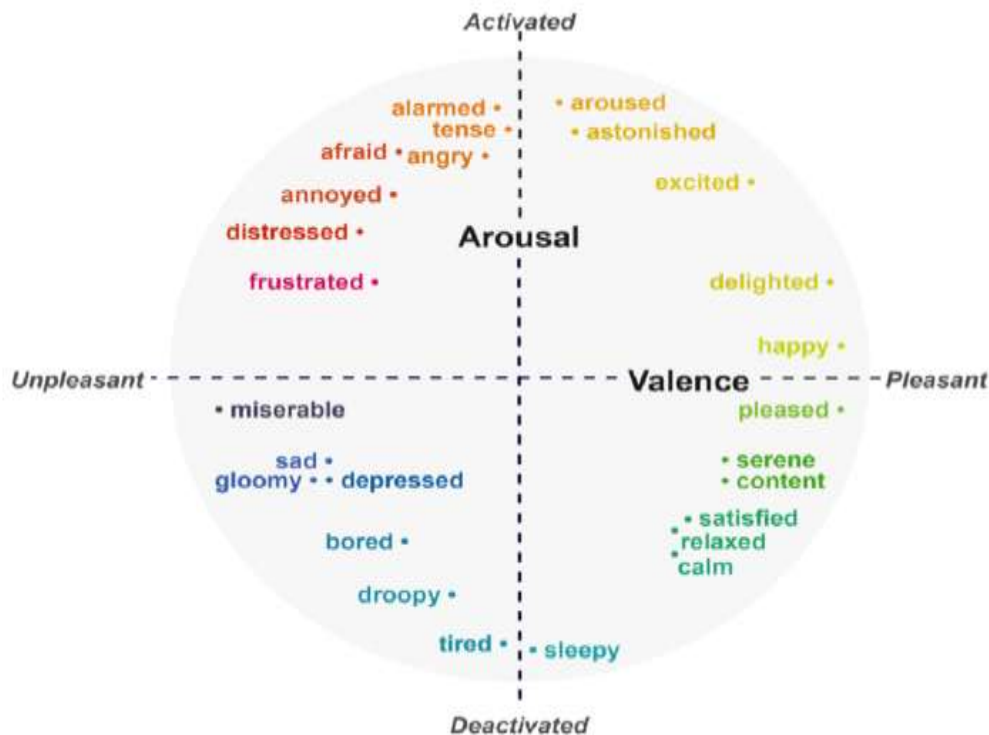
Θεωρητικά, η επίδραση και διαμόρφωση που προκαλεί η μουσική στη συναισθηματική κατάσταση ερμηνεύεται μέσω πολλαπλών μηχανισμών. Για παράδειγμα, ο μηχανισμός της μουσικής προσδοκίας (musical expectation) αναφέρεται στη δυνατότητα της μουσικής να δημιουργεί και να παραβιάζει προσδοκίες του ακροατή σχετικά με τη δομή της μουσικής (αρμονική διαδοχή, ρυθμικό μοτίβο, μελωδική εξέλιξη). Η προσωρινή αναστολή ή η αιφνίδια ανατροπή μιας αναμενόμενης μουσικής εξέλιξης προκαλεί συναισθηματική διέγερση (arousal) και ένταση, ενώ η επακόλουθη επίλυση (resolution) επιφέρει αίσθημα συναισθηματικής εκτόνωσης και ανακούφισης. Ενώ, ο μηχανισμός της συναισθηματικής μόλυνσης (emotional contagion) αναφέρεται στην αυτόματη, ασυνείδητη τάση του ακροατή να «αντικατοπτρίζει» εσωτερικά το συναισθηματικό φορτίο που εκφράζεται μέσω της μουσικής. Ο μηχανισμός αυτός ερμηνεύεται νευροφυσιολογικά μέσω της ενεργοποίησης του συστήματος νευρώνων-καθρεπτών (mirror neuron system), το οποίο αναπαράγει εσωτερικά τα μορφολογικά και εκφραστικά χαρακτηριστικά του ακουστικού ερεθίσματος, προκαλώντας ταυτόσημη συναισθηματική απόκριση στον δέκτη. Τέλος, ο μηχανισμός της μνημονικής επίκλησης (episodic memory) λειτουργεί ως ισχυρός ακουστικός δείκτης (auditory cue) που ενεργοποιεί αυτόματα επεισοδιακές μνήμες από το παρελθόν του ατόμου. Η ανάκληση αυτών των αυτοβιογραφικών συμβάντων συνοδεύεται από την αναβίωση των συνδεδεμένων συναισθηματικών καταστάσεων, εξηγώντας τη βαθιά υποκειμενική και εξατομικευμένη συναισθηματική αξία που αποκτούν συγκεκριμένα μουσικά έργα για κάθε χρήστη. (Juslin and Sloboda, 2010) [20].

Στην κλινική πράξη, η αποτελεσματικότητα της μουσικοθεραπείας τεκμηριώνεται από πλήθος ποσοτικών και ποιοτικών μελετών σε ποικίλα θεραπευτικά πεδία. Σε προεγχειρητικά περιβάλλοντα, για παράδειγμα, η στοχευμένη ακρόαση μουσικής επιφέρει σημαντική μείωση των επιπέδων στρες, περιορίζοντας την ανάγκη χορήγησης ηρεμιστικών σκευασμάτων. Στη διαχείριση του πόνου σε χρόνιες παθήσεις, η μουσική συμβάλλει στη μείωση της υποκειμενικής αντίληψής του πόνου μέσω της γνωσιακής απόσπασης της προσοχής (cognitive distraction) και της παράλληλης διέγερσης έκκρισης ενδογενών οπιοειδών (ενδορφινών). Σε άτομα στο φάσμα του αυτισμού, ενισχύει τη μη λεκτική επικοινωνία, τη συναισθηματική έκφραση και την κοινωνική αλληλεπίδραση, βελτιώνοντας τη συνολική κοινωνική λειτουργικότητα, ενώ αποδεικνύεται αποτελεσματική και στην ανακούφιση συμπτωμάτων κατάθλιψης, καθώς διευκολύνει τη συναισθηματική επεξεργασία και ενισχύει το θετικό συναισθηματικό σθένος (Thoma et al., 2013) [12]. Η επιστημονική τεκμηρίωση αυτών των αποτελεσμάτων παρέχει πολύτιμες θεωρητικές κατευθύνσεις ως προς τον τρόπο με τον οποίο η μουσική επηρεάζει τη διάθεση. Ωστόσο, υπογραμμίζεται ότι η αναπτυχθείσα εφαρμογή δεν αποτελεί σύστημα μουσικοθεραπείας ούτε φέρει κλινικό ή θεραπευτικό χαρακτήρα, αλλά λειτουργεί αποκλειστικά ως ένα ψηφιακό εργαλείο καθημερινής συναισθηματικής προσαρμογής και ψυχαγωγίας.

4.3 Αντιστοίχιση μουσικών χαρακτηριστικών σε συναισθηματικές καταστάσεις

Η αντιστοίχιση μουσικών ιδιοτήτων σε συναισθηματικές καταστάσεις αποτελεί θεμελιώδες επιστημονικό ζήτημα στη διασταύρωση της ψυχολογίας της μουσικής και της ανάκτησης μουσικής πληροφορίας (Music Information Retrieval – MIR). Η κυρίαρχη θεωρητική πλαίσιοση του προβλήματος βασίζεται κατά κύριο λόγο στο κυκλικό μοντέλο συναισθηματικής διάθεσης (Circumplex Model of Affect) του James Russell (1980), σύμφωνα με το οποίο το σύνολο των ανθρώπινων συναισθημάτων χαρτογραφείται σε έναν συνεχή δισδιάστατο ευκλείδειο χώρο. Ο χώρος αυτός ορίζεται από δύο ορθογώνιους άξονες: τον άξονα του συναισθηματικού σθένους (valence), που αποτυπώνει την κλίμακα θετικότητας ή αρνητικότητας από το έντονα δυσάρεστο έως το ιδιαίτερα ευχάριστο, και τον άξονα της συναισθηματικής διέγερσης (arousal), που εκφράζει το επίπεδο νευροφυσιολογικής ενεργοποίησης από την κατάσταση ηρεμίας έως την κατάσταση υψηλής εγρήγορσης. Η συνδυαστική τοποθέτηση σε αυτό το σύστημα συντεταγμένων καθορίζει τη φύση του συναισθήματος, όπου για παράδειγμα η χαρά αντιστοιχεί σε περιοχή υψηλού σθένους και υψηλής διέγερσης, η θλίψη καταλαμβάνει την περιοχή χαμηλού σθένους και χαμηλής διέγερσης, ενώ ο θυμός και ο φόβος χαρακτηρίζονται από χαμηλό σθένος αλλά ιδιαίτερα αυξημένη διέγερση. (Russell, 1980) [32].

Η εφαρμογή του θεωρητικού μοντέλου του Russell στην ανάλυση μουσικής επιτρέπει την αντιστοίχιση ακουστικών χαρακτηριστικών (audio features) τραγουδιών σε θέσεις του valence–arousal χώρου, ανοίγοντας τον δρόμο για αλγοριθμική σύσταση μουσικής βάσει στοχευόμενης συναισθηματικής κατάστασης. Στο πλαίσιο αυτό, το Spotify Web API παρέχει για κάθε τραγούδι της βάσης δεδομένων του ένα σύνολο ακουστικών χαρακτηριστικών που αντιστοιχούν θεωρητικά αλλά όχι απόλυτα, στις διαστάσεις του Circumplex Model. Ειδικότερα, το χαρακτηριστικό valence ορίζεται από το Spotify ως μέτρο της μουσικής θετικότητας που μεταδίδει ένα κομμάτι σε κλίμακα 0.0–1.0: τιμές πλησίον του 1.0 αντιστοιχούν σε χαρούμενα, ευφορικά κομμάτια, ενώ τιμές πλησίον του 0.0 σε θλιμμένα ή οργισμένα. Ενώ το χαρακτηριστικό energy αντικατοπτρίζει την αντιληπτή ενεργητικότητα και ένταση του κομματιού και αντιστοιχεί υποκειμενικά ως ο υπολογιστικός ισοδύναμος δείκτης στον άξονα arousal του μοντέλου Russell (σχήμα 4-1).



Σχήμα 4-1: Κυκλικό μοντέλο συναισθημάτων του Russell

Εμπειρική επικύρωση της αντιστοίχισης αυτής έχει παρασχεθεί από ερευνητική μελέτη με 244 συμμετέχοντες, η οποία συνέκρινε τις αυτοματοποιημένες τιμές valence και energy του Spotify με υποκειμενικές αξιολογήσεις ακροατών. Τα ευρήματα κατέδειξαν ισχυρή συσχέτιση μεταξύ του Spotify energy και των ανθρώπινων αξιολογήσεων arousal, καθώς και μέτρια συσχέτιση μεταξύ Spotify valence και ανθρώπινων αξιολογήσεων valence, επιβεβαιώνοντας τη χρησιμότητα των αυτοματοποιημένων μετρήσεων για εφαρμογές αναγνώρισης συναισθήματος μουσικής εφόσον χρησιμοποιούνται με κατάλληλη επιφυλακτικότητα.

Πέραν των valence και energy, σύγχρονες έρευνες στο πεδίο της αναγνώρισης συναισθήματος μουσικής (Music Emotion Recognition, MER) έχουν αναδείξει ότι τα χαρακτηριστικά acousticism (ακουστικός χαρακτήρας), instrumentality (απουσία φωνής) και tempo (ρυθμός σε BPM) συμβάλλουν επίσης σημαντικά στη διάκριση συναισθηματικών κατηγοριών (Schedl et al., 2018) [29]. Για τον σκοπό αυτό, το μοντέλο της παρούσας εργασίας διεκδικείται με τη συνεκτίμηση του ακουστικού χαρακτήρα (acousticism), της πιθανότητας απουσίας φωνητικών (instrumentality), του ρυθμικού βηματισμού σε κτυπήματα ανά λεπτό (tempo) και της χορευτικότητας (danceability). Συγκεκριμένα, υψηλή acousticism και χαμηλή energy τείνουν να χαρακτηρίζουν μουσική κατάλληλη για καταστάσεις λύπης ή ηρεμίας, ενώ υψηλό tempo και energy χαρακτηρίζουν μουσική κατάλληλη για θυμό ή ευφορία. Βάσει αυτών των παραδοχών, στην παρούσα εργασία ορίζεται ένας σχεδιαστικός κανόνας αντιστοίχισης μεταξύ συναισθηματικών κατηγοριών και εύρους τιμών audio features, ο οποίος λειτουργεί και αναλύεται ως ευρετικό πλαίσιο (heuristic framework) σύστασης στο Κεφάλαιο 5.

4.4 Συστήματα προσαρμογής και σύστασης μουσικού περιεχομένου

Η άνοδος των υπηρεσιών ψηφιακών πλατφορμών μουσικής (streaming) έχει μετασηματίσει ριζικά τον τρόπο με τον οποίο οι χρήστες ανακαλύπτουν και καταναλώνουν μουσικό περιεχόμενο, παρέχοντας ταυτόχρονα στην ερευνητική κοινότητα πρωτοφανή πρόσβαση σε δεδομένα συμπεριφοράς ακρόασης και ακουστικών χαρακτηριστικών μουσικής σε κλίμακα εκατομμυρίων κομματιών. Μεταξύ των πλατφορμών αυτών, το Spotify έχει καθιερωθεί ως η κυρίαρχη υπηρεσία παγκοσμίως, με άνω των 600 εκατομμυρίων ενεργών χρηστών και βάση δεδομένων που υπερβαίνει τα 100 εκατομμύρια κομμάτια, καθιστώντας το Spotify API το πιο διαδεδομένο εργαλείο για ερευνητικές και εφαρμοσμένες εργασίες σύστασης μουσικής (Schedl et al., 2018) [29].

Τα συστήματα σύστασης μουσικής (Music Recommender Systems) αποτελούν εξειδικευμένη κατηγορία συστημάτων σύστασης που εφαρμόζουν τεχνικές συνεργατικού φιλτραρίσματος (collaborative filtering), φιλτραρίσματος βάσει περιεχομένου (content-based filtering) ή υβριδικές προσεγγίσεις για την πρόβλεψη μουσικών προτιμήσεων χρηστών. Στο συνεργατικό φιλτράρισμα, οι συστάσεις βασίζονται στις ομοιότητες μεταξύ χρηστών ή μεταξύ αντικειμένων: το σύστημα συνιστά κομμάτια που άρεσαν σε χρήστες με παρόμοιο προφίλ ακρόασης. Στο φιλτράρισμα βάσει περιεχομένου, οι συστάσεις βασίζονται στα ακουστικά χαρακτηριστικά των κομματιών: το σύστημα συνιστά κομμάτια με παρόμοια χαρακτηριστικά με αυτά που ο χρήστης έχει ήδη ακούσει. Η παρούσα εργασία υλοποιεί μια υβριδική προσέγγιση που συνδυάζει στοιχεία και των δύο: τα audio features targets προκύπτουν από το ανιχνευμένο συναίσθημα (content-based), ενώ τα seed tracks αντλούνται από το ιστορικό ακρόασης του χρήστη (collaborative-based) μέσω κλήσης στο endpoint GET /me/top/tracks του Spotify API (Schedl et al., 2018) [29].

Σημαντική ερευνητική κατεύθυνση στο πεδίο της μουσικής σύστασης αποτελεί η ενσωμάτωση συναισθηματικού πλαισίου (affective context) στη διαδικασία σύστασης, με στόχο τη μετάβαση από στατικά προφίλ χρήστη σε δυναμικές, πλαίσιοεναίσθητες προτάσεις που ανταποκρίνονται στην εκάστοτε συναισθηματική κατάσταση του χρήστη. Εργασίες στο πεδίο αυτό έχουν αναδείξει ότι η ενσωμάτωση συναισθηματικής πληροφορίας βελτιώνει σημαντικά την αντιληπτή σχετικότητα και ικανοποίηση από τις μουσικές προτάσεις σε σύγκριση με συστήματα που βασίζονται αποκλειστικά σε ιστορικό ακρόασης (Deldjoo et al., 2022) [13]. Η παρούσα εργασία εντάσσεται σε αυτή την ερευνητική κατεύθυνση, προτείνοντας ένα σύστημα που αξιοποιεί αναγνώριση συναισθήματος σε πραγματικό χρόνο μέσω εκφράσεων προσώπου ως πηγή πλαίσιακής πληροφορίας για τη διαμόρφωση εξατομικευμένων μουσικών προτάσεων.

Η βιβλιογραφία επιβεβαιώνει την αποτελεσματικότητα ανάλογων υλοποιήσεων, όπου υπολογιστικές αρχιτεκτονικές συνελκτικών νευρωνικών δικτύων (CNN), όπως το μοντέλο VGG16, επιστρατεύονται για την εξαγωγή συναισθηματικών δεικτών από το πρόσωπο του χρήστη με σκοπό την αυτόματη επιλογή του κατάλληλου μουσικού περιεχομένου (Kiran et al., 2024) [26]. Παράλληλα, ευρύτερες μελέτες στο πεδίο επιβεβαιώνουν την αποτελεσματικότητα της αντιστοίχισης συναισθηματικών καταστάσεων και μεταβλητών διάθεσης (mood-based processing) με δυναμικά προσαρμοσμένες μουσικές λίστες (Kumar, 2024) [27], (Tidke et al., 2022) [28].

Το προτεινόμενο σύστημα επεκτείνει τις αρχές αυτές, συνδυάζοντας την υπολογιστική όραση χαμηλών απαιτήσεων με τους αλγόριθμους σύστασης του Spotify για την παροχή μιας άμεσης, εξατομικευμένης και συναισθηματικά προσαρμοσμένης ακροαματικής εμπειρίας.

Κεφάλαιο 5ο: Σχεδιασμός του συστήματος

5.1 Απαιτήσεις συστήματος

Η φάση ανάλυσης απαιτήσεων αποτελεί το θεμέλιο κάθε διαδικασίας ανάπτυξης λογισμικού, καθώς καθορίζει με σαφήνεια τόσο τις λειτουργικές προδιαγραφές, δηλαδή το τι πρέπει να επιτελεί το σύστημα, όσο και τις μη λειτουργικές απαιτήσεις, οι οποίες ορίζουν τις συνθήκες και τους περιορισμούς υπό τους οποίους το σύστημα οφείλει να λειτουργεί. Στο πλαίσιο της παρούσας εργασίας, οι απαιτήσεις ορίστηκαν βάσει τριών κεντρικών αρχών: της λειτουργίας σε πραγματικό χρόνο εντός κινητής συσκευής, της προστασίας των βιομετρικών δεδομένων του χρήστη και της εξατομίκευσης των μουσικών προτάσεων.

Οι λειτουργικές απαιτήσεις (ΛΑ) (functional requirements) προσδιορίζουν τις συγκεκριμένες διεργασίες και λειτουργίες που οφείλει να εκτελεί το σύστημα, εξεταζόμενες από την οπτική γωνία του τελικού χρήστη. Αναφορικά με τη συγκεκριμένη εφαρμογή, η πρώτη θεμελιώδης απαίτηση αφορά την ανίχνευση προσώπου σε πραγματικό χρόνο (ΛΑ-01), όπου το σύστημα εντοπίζει αυτόματα το πρόσωπο του χρήστη μέσω της κάμερας της κινητής συσκευής και εξάγει τα γεωμετρικά χαρακτηριστικά του από κάθε καρέ της ροής βίντεο. Στη συνέχεια πραγματοποιείται η αναγνώριση συναισθήματος (ΛΑ-02), κατά την οποία το ενσωματωμένο μοντέλο ταξινομεί την ανιχνευόμενη έκφραση σε μία από επτά καθορισμένες κατηγορίες συναισθήματος, οι οποίες περιλαμβάνουν τον θυμό, την αηδία, τον φόβο, τη χαρά, τη λύπη, την έκπληξη καθώς και την ουδέτερη κατάσταση, επιτυγχάνοντας ακρίβεια επαρκή για πρακτική εφαρμογή.

Η πληροφορία αυτή τροφοδοτεί ακολούθως τον μηχανισμό σύστασης μουσικής βάσει συναισθήματος (ΛΑ-03), ο οποίος αντιστοιχίζει το αναγνωρισμένο συναίσθημα σε κατάλληλες παραμέτρους αναζήτησης, ώστε να ανακτήσει μια δυναμική λίστα προτεινόμενων κομματιών μέσω του Spotify API. Για την περαιτέρω βελτιστοποίηση των αποτελεσμάτων, υλοποιείται η ενσωμάτωση του προφίλ του χρήστη (ΛΑ-04), επιτρέποντας στο σύστημα να αξιοποιεί το ιστορικό ακρόασης (top tracks) ως αρχικά δεδομένα (seeds) για την εξατομίκευση των προτάσεων, ενώ παράλληλα προβλέπεται και η εναλλακτική χρήση κριτηρίων βασισμένων σε μουσικά είδη (genre-based seeds) σε περιπτώσεις απουσίας ιστορικού.

Σε επίπεδο αλληλεπίδρασης, η εφαρμογή εξασφαλίζει την αναπαραγωγή μουσικής (ΛΑ-05), παρέχοντας τη δυνατότητα άμεσης εκτέλεσης των προτεινόμενων μουσικών κομματιών, καθώς και τη δυνατότητα αποθήκευσης των επιλεγμένων προτάσεων απευθείας στις προσωπικές λίστες αναπαραγωγής του χρήστη, αξιοποιώντας εκ νέου τις σχετικές διεπαφές του Spotify API. Τέλος, η εύρυθμη και ασφαλής εκτέλεση των ανωτέρω λειτουργιών βασίζεται στην εξουσιοδότηση του χρήστη (ΛΑ-06), η οποία επιτυγχάνεται μέσω της υλοποίησης της ροής OAuth 2.0 για την προστασία των δεδομένων του λογαριασμού και την αυτόματη διαχείριση της ανανέωσης των διακριτικών πρόσβασης (tokens).

Πέραν των λειτουργικών προδιαγραφών, η ορθή και αποδοτική εκτέλεση των διεργασιών διασφαλίζεται μέσω ενός συνόλου αυστηρών μη λειτουργικών απαιτήσεων (ΜΛΑ) (non-functional requirements) που ορίζουν τα ποιοτικά χαρακτηριστικά και τους περιορισμούς λειτουργίας του συστήματος, προσδιορίζοντας τις προδιαγραφές που διασφαλίζουν την ορθή και αποδοτική εκτέλεση των διεργασιών του. Στο πλαίσιο αυτό, η πρώτη κρίσιμη απαίτηση αφορά την απόδοση σε πραγματικό χρόνο (ΜΛΑ-01), σύμφωνα με την οποία η συνολική καθυστέρηση από τη λήψη του καρέ βίντεο έως την παραγωγή της κατηγορίας συναισθήματος οφείλει να μην υπερβαίνει τα 100ms σε τυπικές κινητές

συσκευές Android, εξασφαλίζοντας με τον τρόπο αυτό μια αδιάλειπτη εμπειρία χρήστη χωρίς αντιληπτή καθυστέρηση. Παράλληλα, καθορίζεται η ανάγκη για on-device επεξεργασία (ΜΛΑ-02), η οποία επιβάλλει η ανίχνευση προσώπου και η ταξινόμηση συναισθήματος να εκτελούνται αποκλειστικά τοπικά στη συσκευή, χωρίς την αποστολή εικόνων ή βιομετρικών δεδομένων σε εξωτερικούς διακομιστές, εναρμονιζόμενη πλήρως με τις αρχές ελαχιστοποίησης δεδομένων που ορίζει ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR).

Επιπλέον, ιδιαίτερη έμφαση δίνεται στους περιορισμούς που αφορούν τη μνήμη και τον αποθηκευτικό χώρο (ΜΛΑ-03), όπου το ενσωματωμένο μοντέλο TensorFlow Lite οφείλει να έχει μέγεθος μικρότερο των 15MB προκειμένου να μην επιβαρύνει υπερβολικά τον αποθηκευτικό χώρο της συσκευής, ενώ ταυτόχρονα η κατανάλωση μνήμης RAM κατά την εκτέλεση της εφαρμογής δεν πρέπει να ξεπερνά τα 150MB. Όσον αφορά τη συμβατότητα του συστήματος (ΜΛΑ-04), η εφαρμογή σχεδιάζεται ώστε να είναι συμβατή με συσκευές Android που εκτελούν την έκδοση Android 8.0 (API level 26) ή νεότερη, καλύπτοντας έτσι την πλειονότητα των εν χρήσει συσκευών σε παγκόσμιο επίπεδο. Τέλος, η απαίτηση για τη χρηστικότητα της εφαρμογής (ΜΛΑ-05) ορίζει ότι η δομή του συστήματος επιτυγχάνει βαθμολογία μεγαλύτερη ή ίση του 68 στην κλίμακα System Usability Scale (SUS), γεγονός που αντιστοιχεί σε ένα αποδεκτό επίπεδο χρηστικότητας σύμφωνα με τα καθιερωμένα διεθνή πρότυπα αξιολόγησης.

5.2 Αρχιτεκτονική του συστήματος

Η αρχιτεκτονική της εφαρμογής σχεδιάστηκε με γνώμονα τη θεμελιώδη αρχή του διαχωρισμού των ευθυνών (Separation of Concerns), διασφαλίζοντας ότι κάθε επιμέρους υποσύστημα αναλαμβάνει μια αυστηρά ορισμένη, διακριτή και ανεξάρτητη λειτουργία. Το συνολικό λογισμικό οργανώνεται σε μια αλληλουχία τεσσάρων διαδοχικών επιπέδων επεξεργασίας, τα οποία διοχετεύουν τη ροή των δεδομένων από την αρχική λήψη του οπτικού ερεθίσματος έως την τελική παρουσίαση των εξατομικευμένων μουσικών προτάσεων στον χρήστη.

Η αφετηρία των επιπέδων της επεξεργαστικής αλυσίδας εντοπίζεται στον μηχανισμό λήψης της ροής βίντεο σε πραγματικό χρόνο από την εμπρόσθια κάμερα της κινητής συσκευής, ο οποίος αναλαμβάνει τη διαχείριση της κάμερας, τον έλεγχο των δικαιωμάτων πρόσβασης και τη συνεχή λήψη των ανανεωμένων καρέ. Τα εισερχόμενα καρέ μετατρέπονται κατάλληλα σε δομή πίνακα εικόνας, προκειμένου να τροφοδοτήσουν το επόμενο στάδιο επεξεργασίας.

Εκεί, στο δεύτερο επίπεδο το υποσύστημα αναλαμβάνει τον εντοπισμό του προσώπου και την εξαγωγή των χωρικών του χαρακτηριστικών, μέσω της εκτέλεσης της ροής επεξεργασίας MediaPipe Face Landmarker, όπως περιγράφεται αναλυτικά στην ενότητα 3.3. Η έξοδος του επιπέδου αυτού είναι η απομονωμένη και κανονικοποιημένη εικόνα της περιοχής προσώπου, διαστάσεων 224×224 pixels σε μορφή RGB, καθώς και τους 52 δείκτες μορφολογικής έκφρασης (blendshape scores) για πιθανή μελλοντική αξιοποίηση. Σε περίπτωση μη ανίχνευσης προσώπου στο καρέ, π.χ. λόγω κακής φωτεινότητας ή απουσίας χρήστη, το επίπεδο επιστρέφει κατάλληλο σήμα αποτυχίας και το σύστημα αναμένει το επόμενο καρέ χωρίς να ενεργοποιεί άσκοπα το μοντέλο ταξινόμησης.

Αμέσως μετά την επιτυχή απομόνωση της περιοχής ενδιαφέροντος, η κανονικοποιημένη εικόνα προωθείται στο στάδιο της συναισθηματικής ταξινόμησης. Για τον σκοπό αυτό, στο τρίτο επίπεδο επιστρατεύεται ένα εξειδικευμένο μοντέλο MobileNetV2 (fine-tuned) σε συμπιεσμένη μορφή TensorFlow Lite (.tflite), το οποίο εκτελείται τοπικά στη συσκευή με σκοπό την ταχεία επεξεργασία και την προστασία των προσωπικών δεδομένων. Για τη βελτιστοποίηση της υπολογιστικής απόδοσης και την επίτευξη χαμηλής καθυστέρησης, το σύστημα αξιοποιεί τον Neural Networks API (NNAPI) του

Android, αναθέτοντας το βάρος των υπολογισμών στον επεξεργαστή νευρωνικών δικτύων της συσκευής όπου αυτός διατίθεται (hardware acceleration). Η έξοδος του μοντέλου συνίσταται σε ένα διάνυσμα επτά τιμών πιθανότητας μέσω της συνάρτησης Softmax, οι οποίες αντιστοιχούν στις ισάριθμες βασικές κατηγορίες συναισθήματος, με την επικρατέστερη τιμή να επιλέγεται ως η τρέχουσα εκτίμηση. Ωστόσο, προκειμένου να εξαιρεθούν οι ανεπιθύμητες διακυμάνσεις και η αστάθεια λόγω φυσιολογικού θορύβου μεταξύ διαδοχικών καρτέ, ενσωματώνεται τεχνική χρονικής εξομάλυνσης (temporal smoothing). Ο μηχανισμός αυτός υπολογίζει τον κυλιόμενο μέσο όρο των προβλέψεων σε ένα παράθυρο των τελευταίων πέντε καρτέ, διασφαλίζοντας μια σταθερή και αξιόπιστη εκτίμηση της συναισθηματικής κατάστασης.

Στο τελικό επίπεδο της αρχιτεκτονικής, η σταθεροποιημένη συναισθηματική εκτίμηση μεταφράζεται σε λειτουργικές παραμέτρους αναζήτησης μουσικού περιεχομένου. Η λογική αντιστοίχισης βασίζεται στον προσδιορισμό συγκεκριμένων τιμών για τα ακουστικά χαρακτηριστικά (audio features), όπως ορίζονται στον πίνακα που περιγράφεται στην ενότητα 5.3. Στη συνέχεια, ο μηχανισμός επικοινωνεί με το Spotify Web API αξιοποιώντας τη βιβλιοθήκη HTTP client Retrofit για τη διαχείριση των δικτυακών αιτημάτων και το Spotify Android Auth SDK για την ασφαλή εξουσιοδότηση του χρήστη. Κατά τη διαδικασία παραγωγής των συστάσεων, το σύστημα συνδυάζει τις συναισθηματικές παραμέτρους με το ακροαματικό προφίλ του χρήστη, ενώ διαθέτει και έναν ενσωματωμένο μηχανισμό ασφαλείας (fallback logic) για την αντιμετώπιση περιπτώσεων απουσίας ιστορικού ακρόασης, όπως αναλύεται στην ενότητα 5.4, εξασφαλίζοντας έτσι την αδιάλειπτη παροχή συναφών μουσικών προτάσεων.

5.3 Αντιστοίχιση συναισθημάτων με μουσικά χαρακτηριστικά

Η αντιστοίχιση ανιχνευόμενου συναισθήματος σε παραμέτρους αναζήτησης μουσικής αποτελεί κρίσιμο σχεδιαστικό στοιχείο της εφαρμογής και βασίζεται στη θεωρητική θεμελίωση του Circumplex Model of Affect (Russell, 1980) [32], σύμφωνα με το οποίο κάθε συναισθηματική κατάσταση χαρτογραφείται σε έναν δισδιάστατο χώρο που ορίζεται από τους άξονες του συναισθηματικού σθένους (valence) και της νευροφυσιολογικής διέγερσης (arousal), όπως αναλύθηκε στο κεφάλαιο 4.3. Στο πλαίσιο του λογισμικού, η αντιστοίχιση υλοποιείται ως μια στατική, βελτιστοποιημένη δομή δεδομένων στον πηγαίο κώδικα της Kotlin (μέσω εξειδικευμένων αντικειμένων τύπου data class), η οποία εμπεριέχει για κάθε μία από τις επτά αναγνωρίσιμες κατηγορίες συναισθήματος του μοντέλου FER-2013 τις βέλτιστες τιμές (target values), καθώς και τα αντίστοιχα κάτω και άνω όρια (min/max bounds) για τα ακουστικά χαρακτηριστικά (audio features) του Spotify. Η επιλογή των εύρων τιμών αντί αυστηρά σταθερών σημείων κρίνεται απαραίτητη, προκειμένου να εξασφαλίζεται η απαιτούμενη αλγοριθμική ευελιξία κατά την ανάκτηση των δεδομένων από το API, διατηρώντας παράλληλα τη σύσταση εντός των ορίων της ισομορφικής στρατηγικής (iso principle), δηλαδή της παροχής μουσικής που αντικατοπτρίζει και επικυρώνει την τρέχουσα συναισθηματική διάθεση του χρήστη.

Ειδικότερα, για τις συναισθηματικές καταστάσεις θετικού σθένους και υψηλής διέγερσης, το σύστημα ορίζει παραμέτρους που προάγουν τον ευφορικό και δυναμικό χαρακτήρα της ακρόασης. Στην κατηγορία της χαράς (happiness), η οποία εντοπίζεται στο τεταρτημόριο υψηλού valence και υψηλού arousal του μοντέλου Russell, ορίζονται τιμές *target_valence* 0.7–1.0, *target_energy* 0.6–1.0 και *target_danceability* 0.6–1.0, αντικατοπτρίζοντας μουσική υψηλής θετικότητας, ενεργητικότητας και χορευτικού χαρακτήρα, η οποία, σύμφωνα με τη στρατηγική της ομοιότητας, ενισχύει την παραγωγή ντοπαμίνης και διατηρεί τη θετική συναισθηματική διέγερση. Παράλληλα, για την κατηγορία της έκπληξης (surprise), η οποία χαρακτηρίζεται επίσης από αυξημένα επίπεδα εγρήγορσης αλλά με πιο ευμετάβλητο συναισθηματικό σθένος, ορίζονται τιμές *target_valence* στο διάστημα 0.5–0.8,

target_energy στο 0.6–0.9 και αυξημένο *target_tempo*, ευνοώντας μουσικές συνθέσεις με δυναμικές εναλλαγές και έντονο το στοιχείο της αρμονικής ή ρυθμικής ανατροπής.

Στον αντίποδα, για τις συναισθηματικές καταστάσεις που διέπονται από αρνητικό σθένος και αυξημένη νευροφυσιολογική ένταση, οι αλγοριθμικές παράμετροι προσαρμόζονται ώστε να διοχετεύσουν ή να εκτονώσουν την υποκείμενη συναισθηματική πίεση. Στην περίπτωση του θυμού (*anger*), όπου το χαμηλό σθένος συνδυάζεται με ιδιαίτερα υψηλή διέγερση, το σύστημα επιβάλλει χαμηλό *target_valence* στο 0–0.35, υψηλό *target_energy* στο 0.7–1 και ελάχιστο ρυθμό *min_tempo* 130 BPM, προκρίνοντας μουσικά είδη με έντονη κρουστική δομή και υψηλή ηχητική πυκνότητα που προσφέρουν συναισθηματική εκτόνωση. Όσον αφορά την κατηγορία του φόβου και του άγχους (*fear*), η οποία καταλαμβάνει την περιοχή χαμηλού σθένους και μέτριας έως υψηλής διέγερσης, ορίζονται τιμές *target_valence* 0–0.4, *target_energy* 0.5–0.8 και *target_instrumentalness* μεταξύ 0.5–1. Η έμφαση στην οργανική μουσική χωρίς φωνητικά στίχων (*instrumental*) βασίζεται στη βιβλιογραφία της μουσικοθεραπείας, καθώς η απουσία λεκτικών ερεθισμάτων μειώνει το γνωσιακό φορτίο του εγκεφάλου και συμβάλλει στη σταδιακή ομαλοποίηση των δεικτών του αυτόνομου νευρικού συστήματος. Επιπρόσθετα, για την κατηγορία της αηδίας (*disgust*), η οποία εκφράζει μια κατάσταση έντονης αποστροφής με μέτρια διέγερση, ορίζονται χαμηλές τιμές *target_valence* από 0–0.3 και *target_energy* στο 0.3–0.6 σε συνδυασμό με χαμηλή ακουστικότητα, αποδίδοντας πιο σκοτεινά και βαρέα ακουστικά προφίλ.

Συμπληρωματικά, για τις καταστάσεις χαμηλής διέγερσης καθώς και για την ουδέτερη συναισθηματική ζώνη, ο αλγόριθμος υιοθετεί παραμέτρους που προάγουν την εσωτερίκευση ή τη συναισθηματική ισορροπία. Στην κατηγορία της λύπης (*sadness*), η οποία εντοπίζεται στο τεταρτημόριο χαμηλού σθένους και χαμηλής διέγερσης, ορίζονται τιμές *target_valence* 0–0.35, *target_energy* 0–0.4 και υψηλό *target_acousticness* στο 0.6–1. Η προτίμηση σε ακουστικές συνθέσεις χαμηλής ενέργειας προσφέρει στον ακροατή αίσθημα συναισθηματικής επικύρωσης και παρηγορίας. Στην ουδέτερη (*neutral*) κατάσταση, η οποία λειτουργεί ως σημείο αναφοράς, ορίζονται κεντρικές τιμές ισορροπίας με *target_valence* 0.4–0.6, *target_energy* στο 0.4–0.6 και *target_acousticness* από 0.3–0.7. Το σύνολο αυτών των τιμών παραμένει ο θεωρητικός πυρήνας της αντιστοίχισης και αξιοποιείται πλέον έμμεσα, αντί να αποστέλλεται ως αριθμητική παράμετρος σε αίτημα συστάσεων. Κατευθύνει τη διαμόρφωση λεκτικών ερωτημάτων αναζήτησης (*Search Queries*) με βάση το είδος και τη διάθεση προς το σημείο πρόσβασης *GET /search* του Spotify Web API, όπως αναλύεται διεξοδικά στην ενότητα 7.5. Η επιλογή αυτή είναι αναγκαία κατόπιν της απόσυρσης του σημείου πρόσβασης *GET /v1/recommendations* από τη Spotify για εφαρμογές που βρίσκονται σε στάδιο ανάπτυξης (*Development Mode*) μετά το 2024. (Schedl et al., 2018) [29].

5.4 Εξατομίκευση μέσω προφίλ χρήστη

Όπως αναλύθηκε στο κεφάλαιο 4.4, η υιοθέτηση μιας αμιγώς στατικής αντιστοίχισης μεταξύ συναισθηματικών καταστάσεων και μουσικών παραμέτρων παρουσιάζει εγγενείς περιορισμούς, καθώς παραβλέπει τις ιδιαίτερες αισθητικές και υφολογικές προτιμήσεις του εκάστοτε ακροατή. Για την επίτευξη υψηλής ποιότητας εμπειρίας χρήστη, η αρχιτεκτονική του συστήματος ενσωματώνει έναν διπλό μηχανισμό εξατομίκευσης, ο οποίος συνδυάζει την τρέχουσα συναισθηματική κατάσταση με το πραγματικό ακροαματικό ιστορικό του χρήστη. Η προσέγγιση αυτή διασφαλίζει ότι οι παραγόμενες μουσικές προτάσεις δεν είναι μόνο συναισθηματικά συμβατές (*emotionally aligned*), αλλά εναρμονίζονται παράλληλα με τις υποκειμενικές μουσικές συνήθειες του ατόμου, επιτυγχάνοντας μια πολυπαραγοντική προσαρμογή του περιεχομένου.

Ειδικότερα, ο πρωτεύων μηχανισμός εξατομίκευσης βασίζεται στη χρήση αρχικών δειγμάτων προτιμώμενων κομματιών (Top Tracks Seeds). Κατά την έναρξη της διαδικασίας σύστασης, εκτελείται κλήση στο σημείο απόληξης (endpoint) και κάνοντας χρήση των ανάλογων παραμέτρων, το σύστημα αιτείται από το Spotify API τα πέντε κομμάτια που ο χρήστης έχει ακούσει περισσότερο κατά την πρόσφατη περίοδο (τελευταίες 4 εβδομάδες), αποτυπώνοντας έτσι τις τρέχουσες μουσικές τάσεις του χρήστη. Για την επιτυχή εκτέλεση της κλήσης, απαιτείται η προηγούμενη εξουσιοδότηση μέσω του Spotify Android Auth SDK με το κατάλληλο πεδίο δικαιωμάτων πρόσβασης (OAuth scope user-top-read). Στον αρχικό σχεδιασμό, τα μοναδικά αναγνωριστικά (track URIs/IDs) των πέντε επικρατέστερων συνθέσεων εκχυλίζονται και εισάγονται ως παράμετροι seed_tracks στο κεντρικό αίτημα παραγωγής συστάσεων. Στο ίδιο αίτημα διοχετεύονται ταυτόχρονα και οι τιμές των ακουστικών χαρακτηριστικών (*target_valence*, *target_energy*, κ.ά.) που προέκυψαν από τη διαδικασία συναισθηματικής ταξινόμησης (Schedl et al., 2018) [29]. Στην τελική υλοποίηση, όμως, κατόπιν της απόσυρσης του σημείου πρόσβασης GET /v1/recommendations από τη Spotify για εφαρμογές σε στάδιο ανάπτυξης, ο μηχανισμός seed_tracks και η άμεση εισαγωγή διανυσμάτων ακουστικών χαρακτηριστικών καταργήθηκαν. Στην τελική αρχιτεκτονική της εφαρμογής, η κλήση στο GET /me/top/tracks διατηρείται ως μηχανισμός καταγραφής και ανάλυσης του ακροαματικού προφίλ του χρήστη, χωρίς όμως να τροφοδοτεί πλέον άμεσα την παραγωγή των μουσικών προτάσεων. Η ουσιαστική εξατομίκευση προσαρμόστηκε αναγκαστικά και επιτυγχάνεται πλέον έμμεσα, μέσω της αντιστοίχισης της αναγνωρισμένης συναισθηματικής κατάστασης σε λεκτικά ερωτήματα κατηγορίας και ύφους (genre/mood queries), τα οποία διαβιβάζονται στο endpoint αναζήτησης GET /search, όπως αναλύεται διεξοδικά στην ενότητα 7.5.

Ωστόσο, στα συστήματα σύστασης ανακύπτει συχνά το φαινόμενο της ψυχρής εκκίνησης (cold start problem), το οποίο στην παρούσα εφαρμογή εμφανίζεται όταν ο χρήστης δεν διαθέτει ιστορικό ακρόασης στο Spotify, λόγω του ότι ο χρήστης είναι νέος στην υπηρεσία, δεν διαθέτει επαρκές ιστορικό ακρόασης κατά την πρόσφατη περίοδο ή αρνείται τη χορήγηση των σχετικών αδειών πρόσβασης στο προφίλ του. Το ζήτημα αυτό αντιμετωπίζεται πλέον εγγενώς από τη δομή του μηχανισμού αναζήτησης. Αντί για την εκτέλεση μίας μεμονωμένης κλήσης, ο αλγόριθμος ορίζει για κάθε ανιχνευόμενο συναίσθημα μια διαβαθμισμένη ακολουθία έως και τριών εναλλακτικών συστάσεων θεματικών μουσικών ειδών, η οποία εκτελείται διαδοχικά μέχρι την ανάκτηση μη κενών αποτελεσμάτων (για παράδειγμα, στην περίπτωση της λύπης εκτελούνται διαδοχικά οι όροι αναζήτησης genre:acoustic, genre:acoustic sad και acoustic). Παράλληλα, για τη διασφάλιση της ποιότητας των προτάσεων, στα ανακτηθέντα αποτελέσματα εφαρμόζεται ένα επιπλέον φίλτρο απόρριψης κομματιών που προέρχονται από μουσικές συλλογές (αποκλείοντας εγγραφές με καλλιτέχνη «Various Artists» ή τίτλους άλμπουμ που περιέχουν όρους όπως «compilation», «greatest hits» και «playlist»). Σε περίπτωση που το φιλτράρισμα αυτό αφήσει κενή λίστα, το σύστημα ενεργοποιεί αυτόματα έναν μηχανισμό ασφαλείας (fallback) επαναφέροντας τα αρχικά μη φιλτραρισμένα αποτελέσματα. Αυτή η πολυεπίπεδη αρχιτεκτονική προστατεύει την εφαρμογή από τον κίνδυνο λειτουργικής αποτυχίας και διασφαλίζει την παροχή των κατάλληλων μουσικών συστάσεων, ανεξάρτητα από την ύπαρξη ή μη ακροαματικού ιστορικού.

5.5 Επιλογή τεχνολογιών υλοποίησης

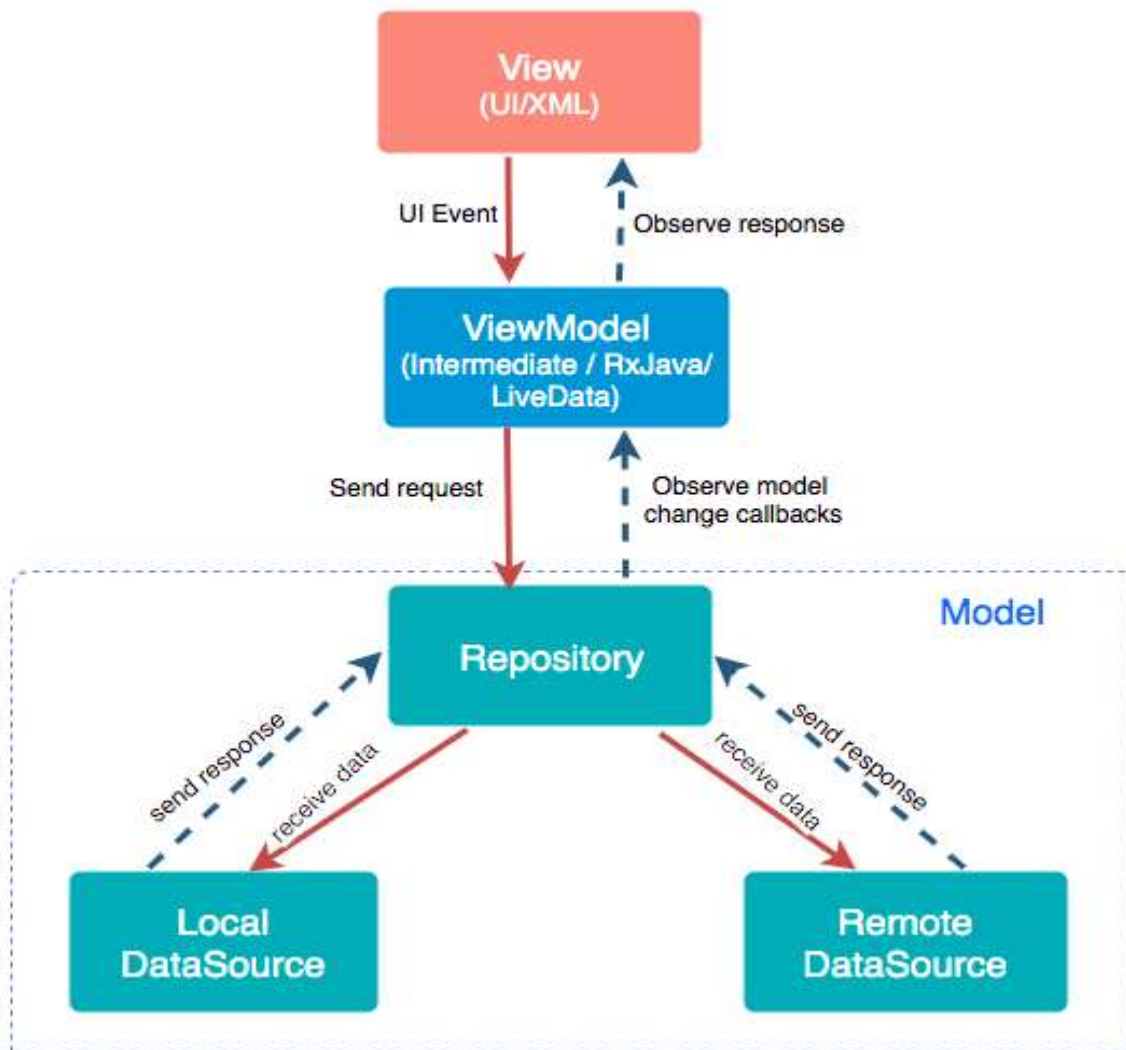
Η επιλογή της τεχνολογικής στοίβας (tech stack) για την ανάπτυξη της εφαρμογής βασίστηκε σε τρία κεντρικά κριτήρια: την ύπαρξη ώριμων και υψηλής απόδοσης βιβλιοθηκών για την εκτέλεση μηχανικής μάθησης απευθείας στη συσκευή (on-device machine learning), την πλήρη υποστήριξη του πλαισίου MediaPipe Tasks Vision για τον εντοπισμό των χαρακτηριστικών του προσώπου, καθώς και την ευκολία ενσωμάτωσης του Spotify API. Η απαίτηση για τοπική εκτέλεση των μοντέλων τεχνητής νοημοσύνης υπαγορεύτηκε τόσο από την ανάγκη επίτευξης ελάχιστης καθυστέρησης (low latency) κατά την επεξεργασία της ροής βίντεο σε πραγματικό χρόνο, όσο και από τη διασφάλιση της ιδιωτικότητας των βιομετρικών δεδομένων του χρήστη, καθώς η οπτική πληροφορία δεν μεταφέρεται σε εξωτερικούς εξυπηρετητές.

Η εφαρμογή αναπτύχθηκε στο Android Studio σε συνδυασμό με το επίσημο Android SDK, ενώ ως γλώσσα προγραμματισμού επιλέχθηκε η Kotlin. Η επιλογή της Kotlin έναντι της Java ευθυγραμμίζεται πλήρως με την επίσημη στρατηγική της Google (Kotlin-first development), η οποία καθιέρωσε τη γλώσσα ως την προτιμώμενη υποδομή για την ανάπτυξη εφαρμογών Android από το 2019. Πέραν της επίσημης υποστήριξης, η γλωσσική αρχιτεκτονική της Kotlin προσφέρει κατασκευές που ενισχύουν σημαντικά την ποιότητα και την αξιοπιστία του πηγαίου κώδικα. Η εγγενής ασφάλεια έναντι κενών αναφορών (null safety) εξαλείφει σε μεγάλο βαθμό τα σφάλματα τύπου NullPointerException, ενώ οι συναρτήσεις επέκτασης (extension functions) επιτρέπουν την καθαρή επέκταση της λειτουργικότητας υφιστάμενων κλάσεων του Android SDK. Επιπλέον, ο μηχανισμός των συρουτινών (coroutines) και των ροών δεδομένων (StateFlow/SharedFlow) παρέχει ένα εξαιρετικά αποδοτικό πλαίσιο για τη διαχείριση ασύγχρονων εργασιών, επιτρέποντας την εκτέλεση απαιτητικών υπολογισμών, όπως η επεξεργασία καρέ και οι δικτυακές κλήσεις, σε νήματα υποβάθρου (background threads), διατηρώντας το κύριο νήμα της διεπαφής χρήστη (UI thread) απόλυτα ανταποκρίσιμο.

Όσον αφορά την ενσωμάτωση των μοντέλων μηχανικής μάθησης, πραγματοποιείται μέσω του TensorFlow Lite, το οποίο παρέχει ένα βελτιστοποιημένο περιβάλλον εκτέλεσης νευρωνικών δικτύων της Google για περιβάλλοντα χαμηλής κατανάλωσης πόρων όπως οι εφαρμογές σε κινητές συσκευές. Ο TensorFlow Lite Interpreter αναλαμβάνει την εκτέλεση του συμπιεσμένου μοντέλου ταξινόμησης συναισθήματος, αξιοποιώντας μηχανισμούς επιτάχυνσης υλικού μέσω του Neural Networks API (NNAPI) του Android ή της μονάδας επεξεργασίας γραφικών (GPU Delegate) όπου αυτοί είναι διαθέσιμοι στη συσκευή. Παράλληλα, για τον αρχικό εντοπισμό και την εξαγωγή και χαρτογράφηση των γεωμετρικών σημείων του προσώπου χρησιμοποιείται το MediaPipe Tasks Vision SDK, το οποίο ενσωματώνεται στην εφαρμογή ως μια από τις βιβλιοθήκες του Maven για το Android. Στο επίπεδο της δικτυακής επικοινωνίας, η αλληλεπίδραση με το Spotify Web API υλοποιείται μέσω του συνδυασμού των βιβλιοθηκών Retrofit και OkHttp. Η Retrofit μετατρέπει τις RESTful κλήσεις HTTP σε διεπαφές Kotlin ασφαλούς τύπου (type-safe interfaces), απλοποιώντας τη σειριοποίηση και αποσειριοποίηση των JSON δεδομένων, ενώ η OkHttp διαχειρίζεται το επίπεδο μεταφοράς, την προσωρινή μνήμη (caching) και τους αναχαιτιστές αιτημάτων (interceptors) για τη διαχείριση των κεφαλίδων. Για τη διαδικασία αυθεντικοποίησης και εξουσιοδότησης βάσει του πρωτοκόλλου OAuth 2.0, χρησιμοποιείται το Spotify Android Auth SDK, το οποίο διεκπεραιώνει με ασφάλεια τη ροή δικαιωμάτων πρόσβασης.

Τέλος, η αρχιτεκτονική της εφαρμογής ακολουθεί το πρότυπο MVVM (Model-View-ViewModel, το οποίο προτείνεται από τις επίσημες αρχιτεκτονικές κατευθυντήριες γραμμές (Android Architecture Guidelines) της Google για εφαρμογές Android. Η εφαρμογή του MVVM διασφαλίζει τον πλήρη διαχωρισμό της λογικής της εφαρμογής (business logic) από το επίπεδο παρουσίασης (presentation layer). Το επίπεδο του Model εσωκλείει τις πηγές δεδομένων (τη διασύνδεση με την

κάμερα, το μοντέλο TFLite και τα δικτυακά endpoints), το ViewModel διαχειρίζεται την κατάσταση της διεπαφής (UI state) και την προετοιμασία των δεδομένων διατηρώντας έτσι τα δεδομένα κατά τις αλλαγές προσανατολισμού της συσκευής, ενώ το View αναλαμβάνει αποκλειστικά την οπτική αποτύπωση των στοιχείων, όπως φαίνεται στο παρακάτω σχήμα (5-1). Ο συνδυασμός του MVVM με το πρότυπο Repository διασφαλίζει την δυνατότητα δοκιμής (testability), την ευκολία συντήρησης και την επεκτασιμότητα του λογισμικού, καθιστώντας την αρχιτεκτονική αρθρωτή (modular) και εύκολα διαχειρίσιμη.



Σχήμα 5-1: Αρχιτεκτονική MVVM

Κεφάλαιο 6ο: Ανάπτυξη και εκπαίδευση του μοντέλου αναγνώρισης συναισθημάτων

6.1 Περιγραφή και χαρακτηριστικά του συνόλου δεδομένων FER-2013

Η επιλογή του κατάλληλου συνόλου δεδομένων εκπαίδευσης συνιστά καθοριστικό παράγοντα για την επιτυχή ανάπτυξη οποιουδήποτε υπολογιστικού μοντέλου αναγνώρισης συναισθήματος, καθώς η ποιότητα, η ποικιλομορφία και η κατανομή των δεδομένων ορίζουν άμεσα τα ανώτατα όρια απόδοσης και τη δυνατότητα γενίκευσης του συστήματος σε πραγματικές συνθήκες λειτουργίας (in-the-wild conditions). Στο πλαίσιο της παρούσας διπλωματικής εργασίας, επιλέχθηκε το σύνολο δεδομένων FER-2013 (Facial Expression Recognition 2013), το οποίο αποτελεί το πλέον αναγνωρισμένο σημείο αναφοράς (benchmark) στη διεθνή βιβλιογραφία για την αξιολόγηση αλγορίθμων συναισθηματικής ταξινόμησης από εικόνες προσώπου. Η επιλογή αυτή αιτιολογείται από το γεγονός ότι το FER-2013 παρέχει έναν επαρκή όγκο δεδομένων για τη διαδικασία της λεπτομερούς προσαρμογής προεκπαιδευμένων βαθιών νευρωνικών δικτύων, προσφέροντας παράλληλα ρεαλιστική ποικιλία οπτικών ερεθισμάτων εντός των υπολογιστικών περιορισμών μιας ακαδημαϊκής διπλωματικής εργασίας. (Huang et al., 2023) [8].

Ως προς την δομή και τα χαρακτηριστικά του, το FER-2013 δημιουργήθηκε από τους Goodfellow et al. (2013) για τον διαγωνισμό αναγνώρισης συναισθημάτων που διοργανώθηκε στο πλαίσιο του ICML 2013 (Workshop on Challenges in Representation Learning). Το σύνολο δεδομένων (dataset) συγκεντρώθηκε αυτόματα μέσω μηχανής αναζήτησης Google Images, αξιοποιώντας λέξεις-κλειδιά που συνδύαζαν συναισθηματικές κατηγορίες με δημογραφικά χαρακτηριστικά, και στη συνέχεια επισημάνθηκε από ανθρώπινους κριτές μέσω πλατφόρμας κατανεμημένης εργασίας (crowdsourcing). Συγκεκριμένα, αποτελείται συνολικά από 35.887 εικόνες προσώπου σε κλίμακα γκρι (grayscale), τυποποιημένων διαστάσεων 48×48 pixels, κατανεμημένων σε επτά κατηγορίες συναισθήματος: θυμός (4.953 εικόνες), αηδία (547 εικόνες), φόβος (5.121 εικόνες), χαρά (8.989 εικόνες), λύπη (6.077 εικόνες), έκπληξη (4.002 εικόνες) και ουδέτερο (6.198 εικόνες).

Στη διεθνή βιβλιογραφία, το σύνολο δεδομένων FER-2013 περιγράφεται συχνά ως προδιαγεγραμμένο σε τρία διακριτά υποσύνολα (Training, PublicTest, PrivateTest), με σκοπό την σύγκριση των αποτελεσμάτων μεταξύ μελετών. Στην παρούσα εργασία αξιοποιείται η ευρέως διαθέσιμη μορφή του συνόλου δεδομένων FER-2013, η οποία περιλαμβάνει το σύνολο εκπαίδευσης (28.709 εικόνες) και ένα ενοποιημένο σύνολο δοκιμής 7.178 εικόνων (αντιστοιχεί στην ένωση PublicTest και PrivateTest). Το σύνολο δοκιμής παρέμεινε αυστηρά απομονωμένο και δεν χρησιμοποιήθηκε σε κανένα στάδιο της εκπαιδευτικής διαδικασίας· η επικύρωση που απαιτείται για την εφαρμογή του μηχανισμού πρώιμης διακοπής (early stopping) κατά τη δεύτερη φάση εκπαίδευσης πραγματοποιήθηκε αποκλειστικά σε ανεξάρτητο υποσύνολο 2.868 εικόνων (10% του συνόλου εκπαίδευσης) που προέκυψε από τυχαίο διαχωρισμό (validation split), διασφαλίζοντας ότι το σύνολο παρέμεινε πλήρως αόρατο (unseen) καθ' όλη τη διάρκεια της εκπαίδευσης (Goodfellow et al., 2013) [30].

Ένα θεμελιώδες τεχνικό χαρακτηριστικό του FER-2013, το οποίο καθορίζει τις μεταγενέστερες επιλογές σχεδιασμού του μοντέλου, είναι η έντονη ανισορροπία των κλάσεων (class imbalance). Συγκεκριμένα, προκύπτει από τα αριθμητικά δεδομένα ότι η κατηγορία «χαρά» υπερεκπροσωπείται με 8.989 εικόνες, πάνω από 16 φορές περισσότερες από την υπο-εκπροσωπούμενη κατηγορία «αηδία» (547 εικόνες). Η ανισορροπία αυτή, αν δεν αντιμετωπιστεί κατάλληλα, οδηγεί σε μοντέλα που

εμφανίζουν πλασματική υψηλή συνολική ακρίβεια (accuracy), λόγω σωστής ταξινόμησης συχνών κατηγοριών, αλλά ιδιαίτερα χαμηλή ανάκληση (recall) για σπάνιες κατηγορίες. Επιπροσθέτως, η φύση των εικόνων, οι οποίες ελήφθησαν από το διαδίκτυο χωρίς ελεγχόμενες εργαστηριακές συνθήκες φωτισμού, γωνίας λήψης ή απόκρυψης προσώπου (occlusions), σε συνδυασμό με τον θόρυβο επισημείωσης (label noise) που προέκυψε από τον πληθοπορισμό (crowdsourcing), περιορίζει την ανθρώπινη ακρίβεια ταξινόμησης στο συγκεκριμένο σύνολο στο 65–70%. Το ποσοστό αυτό θέτει ένα ρεαλιστικό, θεωρητικό ανώτατο όριο στην αναμενόμενη επίδοση οποιουδήποτε αλγορίθμου μηχανικής μάθησης, καθιστώντας το FER-2013 ένα ιδιαίτερα απαιτητικό αλλά και αντιπροσωπευτικό πλαίσιο δοκιμασίας για εφαρμογές σε πραγματικό χρόνο (Goodfellow et al., 2013) [30].

6.2 Προεπεξεργασία δεδομένων

Πριν από τη διοχέτευση των πρωτογενών οπτικών δεδομένων στο δίκτυο MobileNetV2, το σύνολο δεδομένων FER-2013 υποβάλλεται σε μια σαφή και καθορισμένη σειρά μετασχηματισμών προεπεξεργασίας. Στόχος των μετασχηματισμών αυτών είναι η προσαρμογή των δεδομένων στις διαστατικές και στατιστικές απαιτήσεις εισόδου του προεκπαιδευμένου μοντέλου, η εξάλειψη της υπολογιστικής ασυμβατότητας, καθώς και η ενίσχυση της γενικευσιμότητας του συστήματος.

Σχετικά με την κανονικοποίηση τους, επειδή οι εικόνες του FER-2013 είναι διαμορφωμένες σε κλίμακα γκρι (grayscale) μονού καναλιού με τιμές εικονοστοιχείων στο εύρος $[0, 255]$. Αρχικά, για τη συμβατότητα με το MobileNetV2, το οποίο έχει εκπαιδευτεί σε έγχρωμες εικόνες τριών καναλιών (RGB), εφαρμόζεται μετατροπή σε έγχρωμη μορφή τριών καναλιών (grayscale $\rightarrow$ RGB) μέσω αντιγραφής του μοναδικού καναλιού σε τρία πανομοιότυπα κανάλια. Η προσαρμογή αυτή κρίνεται απαραίτητη προκειμένου το εισερχόμενο σήμα να καταστεί συμβατό με τα βάρη του MobileNetV2, το οποίο έχει προεκπαιδευτεί στο σύνολο ImageNet. Στη συνέχεια, οι εικόνες αλλάζουν μέγεθος (resize) από 48×48 σε 224×224 pixels μέσω διγραμμικής παρεμβολής (bilinear interpolation), ώστε να ταιριάζουν με την απαιτούμενη είσοδο του MobileNetV2. Τέλος, οι τιμές εικονοστοιχείων κανονικοποιούνται στο εύρος $[-1, 1]$ μέσω της σχέσης (6.1):

$$pixel_normalized = \left(\frac{pixel}{127.5} \right) - 1 \quad (6.1)$$

Όπου *pixel* συμβολίζει την αρχική τιμή του εικονοστοιχείου. Σύμφωνα με τη στρατηγική κανονικοποίησης που χρησιμοποιήθηκε κατά την αρχική εκπαίδευση του MobileNetV2 στο ImageNet (Sandler et al., 2018) [15].

Παρομοίως, για την αποτελεσματική αντιμετώπιση του υπερπροσαρμογής (overfitting) και την ενίσχυση της ανθεκτικότητας του μοντέλου απέναντι σε φυσιολογικές μεταβολές των συνθηκών λήψης, εφαρμόζεται μια συστηματική αλυσίδα τεχνικών επαύξησης δεδομένων (data augmentation) αποκλειστικά κατά τη διάρκεια της εκπαίδευσης, όχι κατά την αξιολόγηση. Οι τεχνικές αυτές δημιουργούν τεχνητές παραλλαγές των εικόνων εκπαίδευσης, αυξάνοντας αποτελεσματικά το μέγεθος και τη διαφορετικότητα του συνόλου εκπαίδευσης χωρίς συλλογή νέων δεδομένων.

Συγκεκριμένα, εφαρμόζονται οι ακόλουθοι μετασχηματισμοί: η οριζόντια αντιστροφή (horizontal flip) με πιθανότητα 50%, η οποία είναι εννοιολογικά ουδέτερη για αναγνώριση συναισθήματος καθώς τα συναισθήματα εκφράζονται συμμετρικά, η τυχαία στροφή (random rotation) έως ± 10 μοίρες, προσομοιώνοντας ελαφρά κλίση κεφαλής, η τυχαία αλλαγή φωτεινότητας και αντίθεσης (color jitter) με παράγοντα $\pm 20\%$, προσομοιώνοντας διαφορετικές συνθήκες φωτισμού και η τυχαία περικοπή και επαναφορά στο αρχικό μέγεθος (random resized crop) με αναλογία περικοπής 80%

έως 100% της αρχικής επιφάνειας της εικόνας, προσομοιώνοντας μικρές μεταφορές του προσώπου εντός του φωτογραφικού πλαισίου (Mahendran et al., 2024) [4].

Για την αντιμετώπιση της αρνητικής επίδρασης της έντονης ανισορροπίας των κλάσεων του FER-2013, η διαδικασία βελτιστοποίησης ενσωματώνει τη συνάρτηση απώλειας διασταυρούμενης εντροπίας με στάθμιση των κλάσεων. Στο πλαίσιο αυτό, εισάγεται ένας συντελεστής βαρύτητας w_c για κάθε συναισθηματική κατηγορία c , ο οποίος ορίζεται ως ο αντίστροφος λόγος της σχετικής της συχνότητας στο σύνολο εκπαίδευσης, όπως διακρίνεται στην σχέση (6.2):

$$w_c = \frac{N}{(C \times n_c)} \quad (6.2)$$

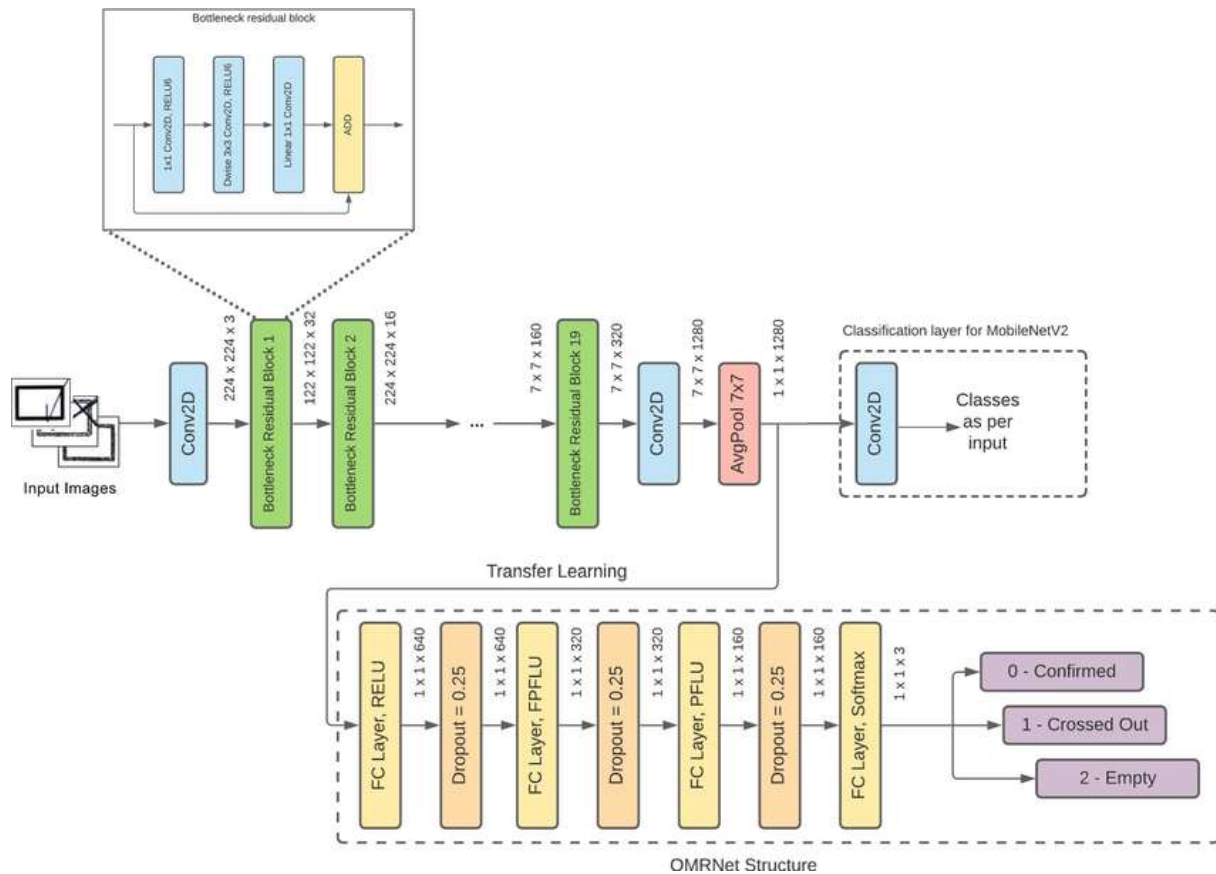
Όπου N το συνολικό πλήθος εικόνων των εικόνων εκπαίδευσης, C ο συνολικός αριθμός των κατηγοριών ($C = 7$) και n_c το πλήθος των δειγμάτων που ανήκουν στην κλάση c . Μέσω αυτής της μαθηματικής διαμόρφωσης, η υποεκπροσωπούμενη κατηγορία της «αηδίας» λαμβάνει συντελεστή βαρύτητας περίπου δεκαέξι φορές μεγαλύτερο από εκείνον της υπερεκπροσωπούμενης κατηγορίας της χαράς. Η συγκεκριμένη στάθμιση εξισορροπεί την επίδραση των λανθασμένων ταξινομήσεων στη συνολική συνάρτηση απώλειας, εξαναγκάζοντας τον αλγόριθμο να αποδίδει ισότιμη προσοχή στις σπάνιες συναισθηματικές εκφράσεις (Mahendran et al., 2024) [4].

6.3 Αρχιτεκτονική νευρωνικού δικτύου

Η αρχιτεκτονική του προτεινόμενου υπολογιστικού μοντέλου βασίζεται στη μεθοδολογία της μεταφοράς γνώσης, συνδέοντας ένα προεκπαιδευμένο δίκτυο κορμού (backbone network) με μια προσαρμοσμένη κεφαλή ταξινόμησης (classification head). Ως δίκτυο κορμού επιλέγεται η αρχιτεκτονική MobileNetV2 (Sandler et al., 2018) [15], η οποία σχεδιάστηκε ειδικά για την εκτέλεση απαιτητικών εργασιών υπολογιστικής όρασης σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους όπως αναλύθηκαν εκτενώς στην ενότητα 3.2.1. Συνοπτικά, το MobileNetV2 διακρίνεται για την υψηλή υπολογιστική του αποδοτικότητα, καθώς αποτελείται από περίπου 3,4 εκατομμύρια παραμέτρους και απαιτεί μόλις 300 εκατομμύρια πράξεις πολλαπλασιασμού-πρόσθεσης κινητής υποδιαστολής (Multiply-Accumulate - MAdd) ανά πρόβλεψη. Οι ιδιότητες αυτές οδηγούν στην παραγωγή ενός συμπαγούς αρχείου μοντέλου μεγέθους περίπου 14MB σε μεταγλωττισμένη μορφή TensorFlow Lite (TFLite), γεγονός που ικανοποιεί πλήρως τις αυστηρές μη λειτουργικές απαιτήσεις που τέθηκαν για την εκτέλεση της εφαρμογής στην κινητή συσκευή με χαμηλή κατανάλωση μνήμης και ενέργειας. Η αρχικοποίηση των βαρών του δικτύου κορμού πραγματοποιείται με τη χρήση των προεκπαιδευμένων παραμέτρων του συνόλου ImageNet, οι οποίες ενσωματώνουν πλούσια, γενικευμένη γνώση εξαγωγής χαμηλού και μέσου επιπέδου οπτικών χαρακτηριστικών (low- and mid-level feature representation) από εκατομμύρια εικόνες χιλίων διαφορετικών κατηγοριών.

Για την προσαρμογή του μοντέλου στις ειδικές απαιτήσεις της αναγνώρισης συναισθήματος από εικόνες προσώπου, το αρχικό επίπεδο εξόδου του MobileNetV2 (το οποίο αποτελείται από 1.000 νευρώνες) αφαιρείται και αντικαθίσταται από μια νέα, εξειδικευμένη κεφαλή ταξινόμησης. Η νέα αυτή δομή ξεκινά με ένα επίπεδο καθολικής συγκέντρωσης μέσου όρου (Global Average Pooling - GAP), το οποίο συμπεκνώνει τους χωρικούς χάρτες χαρακτηριστικών σε ένα μονοδιάστατο διάνυσμα. Η χρήση του GAP πλεονεκτεί έναντι των παραδοσιακών επιπέδων επιπέδωσης (Flatten), καθώς μειώνει δραστικά τον συνολικό αριθμό των εκπαιδευσιμων παραμέτρων, περιορίζοντας παράλληλα την τάση του μοντέλου για υπερπροσαρμογή. Στη συνέχεια, τοποθετείται ένα επίπεδο τυχαίας αποσύνδεσης (Dropout) με ποσοστό 0,3 για τη στατιστική ομαλοποίηση (regularization) των βαρών, το οποίο απενεργοποιεί τυχαία το 30% των συνδέσεων κατά τη διάρκεια της εκπαίδευσης, εξαναγκάζοντας το

δίκτυο να αναπτύξει πιο ανθεκτικές αναπαραστάσεις. Ακολουθεί ένα ενδιάμεσο πλήρως συνδεδεμένο επίπεδο (Fully Connected / Dense Layer) 128 νευρώνων με μη γραμμική συνάρτηση ενεργοποίησης ReLU (Rectified Linear Unit), το οποίο αναλαμβάνει τη σύνθεση των υψηλότερου επιπέδου συναισθηματικών μοτίβων. Η αρχιτεκτονική ολοκληρώνεται με ένα τελικό επίπεδο εξόδου 7 νευρώνων, ισάριθμων με τις συναισθηματικές κατηγορίες του FER-2013, το οποίο αξιοποιεί τη συνάρτηση ενεργοποίησης Softmax για την παραγωγή της τελικής διακριτής κατανομής πιθανότητας επί των επτά συναισθημάτων, παρουσιάζεται αναλυτικά στο σχήμα 6-1 παρακάτω (Chowdhury, Idris and Abas, 2024) [3].



Σχήμα 6-1: Αρχιτεκτονική MobileNetV2 & Classification Head

6.4 Διαδικασία εκπαίδευσης και Fine-Tuning

Η εκπαίδευση υλοποιείται σε δύο διαδοχικές φάσεις, σύμφωνα με τη διαφασική στρατηγική λεπτομερούς προσαρμογής (two-stage fine-tuning) που αναλύθηκε στην ενότητα 3.1.2.

Κατά την πρώτη φάση, τα βάρη ολόκληρου του δικτύου κορμού MobileNetV2 παραμένουν «παγωμένα» (frozen), εμποδίζοντας οποιαδήποτε ενημέρωση ή μεταβολή τους κατά τη διαδικασία της οπισθοδιάδοσης σφάλματος (backpropagation). Σε αυτό το στάδιο, η εκπαίδευση περιορίζεται αποκλειστικά στις παραμέτρους της νέας, προσαρμοσμένης κεφαλής ταξινόμησης. Η συγκεκριμένη φάση εκτελείται για 10 εποχές (epochs) με αρχικό ρυθμό μάθησης $\eta = 10^{-3}$ και με τη χρήση του βελτιστοποιητή Adam. Η προσέγγιση αυτή εξασφαλίζει τη γρήγορη εκμάθηση της βέλτιστης αντιστοίχισης μεταξύ των ήδη εξαγόμενων χαρακτηριστικών του ImageNet και των νέων κατηγοριών

συναισθήματος του FER-2013, μειώνοντας τον κίνδυνο καταστροφικής αλλοίωσης της προϋπάρχουσας γνώσης του δικτύου κορμού.

Κατά τη δεύτερη φάση, τα τελευταία 30 επίπεδα (layers) του δικτύου κορμού αποπαγώνονται και υπόκεινται σε επανεκπαίδευση, επιτρέποντας τη λεπτομερή προσαρμογή των χαρακτηριστικών υψηλού επιπέδου στις ιδιαιτερότητες των συσπάσεων και εκφράσεων του προσώπου. Η φάση αυτή εκτελείται για 20 εποχές με σημαντικά χαμηλότερο ρυθμό μάθησης, ίσο με $\eta = 10^{-4}$ (μια τάξη μεγέθους μικρότερη σε σχέση με την πρώτη φάση), προκειμένου να αποφευχθεί το φαινόμενο της καταστροφικής λήθης (catastrophic forgetting) της γενικευμένης γνώσης που αποκτήθηκε από το ImageNet. Για τη βελτιστοποίηση της διαδικασίας, χρησιμοποιείται η τεχνική της συνημιτονικής ανόπτησης (cosine annealing) ως χρονοδιάγραμμα μεταβολής του ρυθμού μάθησης (learning rate schedule), η οποία μειώνει ομαλά την τιμή από 10^{-4} έως 10^{-6} κατά τη διάρκεια των 20 εποχών, διευκολύνοντας την ακριβέστερη σύγκλιση του μοντέλου στα τελικά στάδια. Παράλληλα, εφαρμόζεται ο μηχανισμός της πρώιμης διακοπής (early stopping) με κριτήριο ανοχής (patience) 5 εποχών, ο οποίος τερματίζει αυτόματα την εκπαίδευση εάν η ακρίβεια στο σύνολο επικύρωσης δεν παρουσιάσει βελτίωση για πέντε συνεχόμενες εποχές (Mahendran et al., 2024) [4].

Ο πίνακας που ακολουθεί συνοψίζει τις κύριες υπερπαραμέτρους που χρησιμοποιούνται και στις δύο φάσεις εκπαίδευσης:

Πίνακας 6-1: Πίνακας υπερπαραμέτρων

Υπερπαραμέτρος	Φάση 1	Φάση 2
Εποχές	10	20
Ρυθμός μάθησης	1e-3	1e-4
Βελτιστοποιητής	Adam	Adam
Batch size	32	32
LR Schedule	Σταθερός	Cosine Annealing
Παγωμένα επίπεδα	Όλα (backbone)	Πρώτα 100 επίπεδα
Early Stopping	Όχι	Ναι (patience=5)

Μετά την επιτυχή ολοκλήρωση της διαδικασίας εκπαίδευσης, το μοντέλο μετατρέπεται σε μορφή TensorFlow Lite (.tflite), προκειμένου να καταστεί δυνατή η ενσωμάτωσή του στην Android εφαρμογή. Η μετατροπή αυτή πραγματοποιείται μέσω του εργαλείου TFLiteConverter και περιλαμβάνει τη βελτιστοποίηση του δικτύου μέσω της τεχνικής της κβάντισης δυναμικού εύρους (dynamic range quantization). Κατά τη διαδικασία αυτή, τα βάρη του μοντέλου μετατρέπονται από αναπαράσταση κινητής υποδιαστολής 32-bit (floating point) σε ακέραιες τιμές 8-bit (integer). Στην παρούσα υλοποίηση, η προσέγγιση αυτή πέτυχε μείωση του τελικού μεγέθους του αρχείου κατά περίπου 75% (από $\approx 14\text{MB}$ σε $\approx 3.5\text{MB}$) επιταχύνοντας σημαντικά την εκτέλεση της εξαγωγής συμπερασμάτων στη συσκευή, επιφέροντας παράλληλα ελάχιστη μείωση στην ακρίβεια του μοντέλου (η οποία βιβλιογραφικά κυμαίνεται κάτω του 1%). Το τελικό βελτιστοποιημένο αρχείο .tflite τοποθετείται στον κατάλογο πόρων (assets) της Android εφαρμογής και φορτώνεται κατά την εκκίνηση της δραστηριότητας μέσω του TFLite Interpreter.

6.5 Μετατροπή και βελτιστοποίηση του μοντέλου σε TensorFlow Lite

Μετά την ολοκλήρωση της εκπαίδευσης, το τελικό μοντέλο MobileNetV2 μετατρέπεται από τη μορφή Keras/TensorFlow σε μορφή TensorFlow Lite (.tflite), προκειμένου να καταστεί κατάλληλο για τοπική εκτέλεση (on-device) σε περιβάλλον Android με περιορισμένους υπολογιστικούς πόρους.

Η προσαρμογή και η συμπίεση του μοντέλου σε μορφή TensorFlow Lite (.tflite) επιτρέπει την άμεση εγκατάσταση και λειτουργία του σε κινητά Android. Μέσω της τεχνικής κβάντισης δυναμικού εύρους (Dynamic Range Quantization) μέσω του εργαλείου TFLiteConverter. Έτσι, οι παράμετροι του δικτύου συμπιέζονται, μειώνοντας δραστικά τις υπολογιστικές απαιτήσεις κατά τη διαδικασία εξαγωγής συμπερασμάτων, με τα αναλυτικά ποσοτικά αποτελέσματα και τη συγκριτική αξιολόγηση των μετρικών να παρατίθενται αναλυτικά στο Κεφάλαιο 8, καθώς συνδέονται άμεσα με τη συνολική αποτίμηση της εφαρμογής στο πειραματικό πλαίσιο. Ως σημείο αναφοράς (baseline) για την αξιολόγηση των επιδόσεων, η σύγχρονη βιβλιογραφία καταγράφει ακρίβεια της τάξης του 73,82% για αρχιτεκτονικές MobileNetV2 που έχουν υποστεί εξειδικευμένη εκπαίδευση (fine-tuning) στο FER-2013 (Sharma, Nagpal and Raja Praveen, 2025) [2]. Στην παρούσα εργασία, η προσέγγιση ή και η υπέρβαση του συγκεκριμένου ορίου επιδιώκεται μέσω της ενσωμάτωσης μιας συνάρτησης απώλειας με στάθμιση κλάσεων (class-weighted loss) και της εφαρμογής της βελτιστοποιημένης στρατηγικής δύο φάσεων όπως αναλύθηκαν προηγουμένως.

Κεφάλαιο 7ο: Ανάπτυξη της κινητής εφαρμογής

7.1 Επισκόπηση υλοποίησης

Η ανάπτυξη της κινητής εφαρμογής «Express Your Sound» πραγματοποιήθηκε στο περιβάλλον Android Studio με χρήση της γλώσσας προγραμματισμού Kotlin, αξιοποιώντας το Android SDK και μια σειρά εξειδικευμένων βιβλιοθηκών ανοιχτού κώδικα (open source) για την υλοποίηση των επιμέρους λειτουργικών ενοτήτων. Η αρχιτεκτονική της εφαρμογής ακολουθεί το πρότυπο σχεδίασης MVVM (Model-View-ViewModel) σε συνδυασμό με το πρότυπο Repository, διασφαλίζοντας τον διαχωρισμό μεταξύ των αρμοδιοτήτων (separation of concerns) και την αρθρωτή δομή (modularity) του κώδικα. Στο πλαίσιο αυτό, το επίπεδο του Model εσωκλείει την επιχειρησιακή λογική, τη διαχείριση των τοπικών πηγών δεδομένων (όπως ο μηχανισμός εκτέλεσης του TensorFlow Lite) και των απομακρυσμένων διεπαφών (Spotify Web API). Το επίπεδο της παρουσίασης δεδομένων (ViewModel) και το επίπεδο διεπαφής χρήστη (View) διαχωρίζονται πλήρως, διευκολύνοντας τη δυνατότητα μελλοντικής συντήρησης και τη δυνατότητα δοκιμής του κώδικα. Ο πηγαίος κώδικας της εφαρμογής «Express Your Sound» είναι διαθέσιμος σε δημόσιο αποθετήριο GitHub, στη διεύθυνση: <https://github.com/evrip-tsia/ExpressYourSound/releases/tag/v1.0-thesis> (έκδοση v1.0-thesis, η οποία αντιστοιχεί στην έκδοση της εφαρμογής κατά τη συγγραφή της παρούσας εργασίας). Η επικοινωνία μεταξύ ViewModel και View υλοποιείται μέσω LiveData, το οποίο εξασφαλίζει ότι οι ενημερώσεις διεπαφής πραγματοποιούνται αποκλειστικά στο κύριο νήμα (main thread) και με σεβασμό στον κύκλο ζωής των Android components. Η επιλογή αυτή εξασφαλίζει ότι οι ενημερώσεις της διεπαφής χρήστη πραγματοποιούνται αποκλειστικά όταν τα αντίστοιχα στοιχεία (Activities ή Fragments) βρίσκονται σε ενεργή κατάσταση (STARTED ή RESUMED), αποτρέποντας αποτελεσματικά την εμφάνιση διαρροών μνήμης (memory leaks) ή ανεπιθύμητων εξαιρέσεων κατά την αλλαγή κατάστασης της εφαρμογής. Παράλληλα, οι απαιτητικές υπολογιστικές διεργασίες, όπως η επεξεργασία των καρέ της κάμερας, η εκτέλεση του μοντέλου ταξινόμησης και οι ασύγχρονες κλήσεις HTTP προς το Spotify API, εκτελούνται σε νήματα υποβάθρου (background threads) με τη χρήση συρουτινών Kotlin (Kotlin Coroutines), ενώ η τελική ενημέρωση του LiveData δρομολογείται αυτόματα στο κύριο νήμα (main thread) για τη διασφάλιση της ομαλής απόδοσης του UI.

Η εφαρμογή οργανώνεται σε τρεις κύριες οθόνες (*Activities*): την οθόνη σύνδεσης (*LoginActivity*) η οποία διαχειρίζεται την αυθεντικοποίηση και τις εξουσιοδοτήσεις του χρήστη μέσω του Spotify Android Auth SDK, την κύρια οθόνη αναγνώρισης συναισθήματος μέσω κάμερας (*MainActivity*) με τη βοήθεια του MediaPipe Tasks Vision SDK και την εξαγωγή συναισθήματος χρησιμοποιώντας το TFLite Interpreter, καθώς και την οθόνη εμφάνισης μουσικών προτάσεων (*RecommendationsActivity*) η οποία αναλαμβάνει την άμεση αναπαραγωγή των μουσικών κομματιών και την αλληλεπίδραση με το Spotify Web API. Τελειώνοντας, κάθε Activity συνοδεύεται από αντίστοιχο ViewModel που διαχειρίζεται την κατάσταση (state) της οθόνης και εκτελεί τις απαραίτητες κλήσεις προς τα υποκείμενα Repository επίπεδα επιτυγχάνοντας υψηλή ικανότητα για δοκιμές (testability) και διατήρηση του πηγαίου κώδικα.

7.2 Ανίχνευση προσώπου

Η διαδικασία λήψης της εικόνας από την κάμερα της κινητής συσκευής υλοποιείται μέσω του Android CameraX API, το οποίο αποτελεί μια σύγχρονη βιβλιοθήκη πρόσβασης που λαμβάνει υπόψη τον κύκλο ζωής των συστατικών της εφαρμογής (lifecycle-aware). Το συγκεκριμένο API προτείνεται επίσημα από την Google για την ανάπτυξη νέων εφαρμογών στο περιβάλλον Android. Η ιδιότητα αυτή απλοποιεί τη διαχείριση των πόρων της κάμερας, αποτρέποντας διαρροές μνήμης και σφάλματα δέσμευσης υλικού κατά τις εναλλαγές κατάστασης των *Activities*. Το CameraX παρέχει τρεις βασικές περιπτώσεις χρήσης (use cases): την προεπισκόπηση (*Preview*) για τη ζωντανή προβολή της εικόνας στο γραφικό συστατικό *PreviewView*, τη λήψη στατικών φωτογραφιών (*ImageCapture*), καθώς και την ανάλυση εικόνας (*ImageAnalysis*) για την απευθείας προσπέλαση και επεξεργασία των επιμέρους καρέ της ροής βίντεο. Στην παρούσα εφαρμογή ενεργοποιούνται παράλληλα οι λειτουργίες *Preview* και *Image Analysis*, διασφαλίζοντας ταυτόχρονη ομαλή οπτική ανατροφοδότηση στον χρήστη και υπολογιστική επεξεργασία της οπτικής πληροφορίας (Google AI Edge, 2023) [22].

Η λειτουργία ανάλυσης εικόνας (*ImageAnalysis*) διαμορφώνεται με τη στρατηγική διατήρησης μόνο του τελευταίου καρέ (*STRATEGY_KEEP_ONLY_LATEST*). Μέσω αυτής της προσέγγισης απορρίπτονται αυτόματα τα παλαιότερα καρέ που δεν έχουν προλάβει να υποστούν επεξεργασία, εξασφαλίζοντας ότι το σύστημα αναλύει πάντοτε το πιο πρόσφατο διαθέσιμο οπτικό στιγμιότυπο. Για την αποφυγή του αποκλεισμού του κύριου νήματος εκτέλεσης (UI thread) και την αποτροπή του σφάλματος μη απόκρισης της εφαρμογής (ANR - Application Not Responding), η επεξεργασία κάθε καρέ εκτελείται σε ξεχωριστό νήμα υποβάθρου (background thread) μέσω της υπηρεσίας *ExecutorService*. Ως προεπιλογή ορίζεται η εμπρόσθια κάμερα της συσκευής (*CameraSelector.DEFAULT_FRONT_CAMERA*), καθώς η φύση της εφαρμογής προϋποθέτει ότι ο χρήστης έχει στραμμένη την οθόνη προς το πρόσωπό του.

Κάθε καρέ που παράγεται από τη ροή του CameraX τροφοδοτείται άμεσα στον αγωγό επεξεργασίας MediaPipe Tasks Vision SDK, και ειδικότερα στην ροή επεξεργασίας του MediaPipe Face Landmarker, ο οποίος εκτελεί την ανίχνευση και τη χαρτογράφηση των χαρακτηριστικών του προσώπου. Ο συγκεκριμένος αγωγός ρυθμίζεται με την παράμετρο *setNumFaces(1)*, καθώς η εφαρμογή επικεντρώνεται αποκλειστικά στον κύριο χρήστη, ενώ το ελάχιστο όριο εμπιστοσύνης για την ανίχνευση προσώπου (*minFaceDetectionConfidence*) ορίζεται στην τιμή 0.5, η οποία αποτελεί την προεπιλεγμένη τιμή του MediaPipe Face Landmarker. Παράλληλα, η επιλογή *setOutputFaceBlendshapes(true)* ενεργοποιεί την παραγωγή 52 δεικτών μορφοποίησης (blendshape scores), οι οποίοι αποθηκεύονται στη δομή δεδομένων της εφαρμογής, παρέχοντας έτοιμη υποδομή για πιθανή μελλοντική αξιοποίηση σε αναλύσεις λεπτομερών μυϊκών συσπάσεων.

Η οριοθέτηση της περιοχής ενδιαφέροντος (Region of Interest - ROI) που περιλαμβάνει το πρόσωπο πραγματοποιείται μέσω του υπολογισμού ενός πλαισίου οριοθέτησης (bounding box), αξιοποιώντας τα 478 αναγνωρισμένα χωρικά σημεία (landmarks) που παρέχει το MediaPipe σε κανονικοποιημένες συντεταγμένες $(x, y) \in [0,1] \times [0,1]$. Συγκεκριμένα, εντοπίζονται οι ακραίες τιμές των αξόνων $x_{\min}, x_{\max}$ και $y_{\min}, y_{\max}$ σε κανονικοποιημένες συντεταγμένες, οι οποίες στη συνέχεια μετατρέπονται σε εικονοστοιχεία (pixels) $X_{\min}, X_{\max}, Y_{\min}, Y_{\max}$ βάσει του πλάτους W και του ύψους H του αρχικού καρέ, όπως αποτυπώνεται στην παρακάτω σχέση (7.1):

$$\begin{aligned} X_{\min} &= x_{\min} \times W, & X_{\max} &= x_{\max} \times W \\ Y_{\min} &= y_{\min} \times H, & Y_{\max} &= y_{\max} \times H \end{aligned} \quad (7.1)$$

Για τη διασφάλιση της ακεραιότητας της περιμετρικής πληροφορίας του προσώπου και την αποφυγή αποκοπής κρίσιμων χαρακτηριστικών (όπως το σαγόνι ή το μέτωπο), εφαρμόζεται ένα επιπλέον περιθώριο ασφαλείας της τάξης του 10% περιμετρικά του υπολογισθέντος πλαισίου. Το προκύπτον πλαίσιο αποκόπτεται (cropped) από το αρχικό καρέ, υποβάλλεται σε αλλαγή κλίμακας (resizing) στις διαστάσεις 224x224 pixels και αποτελεί τη βελτιστοποιημένη είσοδο που τροφοδοτείται στο συνελκτικό νευρωνικό δίκτυο (CNN) MobileNetV2 για την τελική ταξινόμηση του συναισθήματος (Google AI Edge, 2023) [23].

7.3 Αναγνώριση συναισθήματος

Το εκπαιδευμένο μοντέλο MobileNetV2 ενσωματώνεται στην αρχιτεκτονική της εφαρμογής σε συμπίεσμένη μορφή TensorFlow Lite (.tflite) και τοποθετείται στον ειδικό κατάλογο πόρων (assets) του συστήματος αρχείων, από όπου φορτώνεται κατά την εκκίνηση της εφαρμογής μέσω του TFLite Interpreter. Για τη βελτιστοποίηση της διαχείρισης των υπολογιστικών πόρων της κινητής συσκευής, ο διερμηνέας παραμετροποιείται κατάλληλα ώστε να αξιοποιεί το Android Neural Networks API (NNAPI) εφόσον αυτό υποστηρίζεται από το λειτουργικό σύστημα, επιτρέποντας την υπολογιστική επιτάχυνση της εξαγωγής συμπερασμάτων μέσω εξειδικευμένων μονάδων επεξεργασίας νευρωνικών δικτύων (NPU) ή επεξεργαστών γραφικών (GPU).

Η προεπεξεργασία κάθε απομονωμένης εικόνας εισόδου πραγματοποιείται μέσω της κλάσης ImageProcessor της βιβλιοθήκης TFLite Support, η οποία αναλαμβάνει τον μετασχηματισμό των δεδομένων πριν αυτά τροφοδοτηθούν στο δίκτυο. Κατά τη διαδικασία αυτή, εφαρμόζεται χωρική προσαρμογή μεγέθους (resize) στα 224x224 εικονοστοιχεία (pixels) πριν από την τροφοδοσία τους στο δίκτυο. Στο ίδιο πλαίσιο, εφαρμόζεται κανονικοποίηση των τιμών των εικονοστοιχείων από το αρχικό εύρος [0, 255] στο διάστημα [-1, 1], εναρμονιζόμενη πλήρως με τη στρατηγική εκπαίδευσης που ακολουθεί η αρχιτεκτονική του MobileNetV2 αναπαρίσταται από την παρακάτω σχέση (7.2)

(Sandler et al., 2018) [15].

$$x_{\text{norm}} = \frac{x - 127.5}{127.5} \quad (7.2)$$

Η έξοδος του δικτύου συνίσταται σε ένα διάνυσμα επτά τιμών της συνάρτησης Softmax (σχέση 7.3):

$$p = [p_1, p_2, \dots, p_7] \in [0,1]^7 \text{ με } \sum_{i=1}^7 p_i = 1 \quad (7.3)$$

το οποίο αντιπροσωπεύει την κατανομή πιθανότητας πάνω στις επτά συναισθηματικές κατηγορίες του FER-2013 (Οργή, Αηδία, Φόβος, Χαρά, Λύπη, Έκπληξη, Ουδέτερο). Η κλάση που συγκεντρώνει τη μέγιστη τιμή πιθανότητας ορίζεται αρχικά ως το στιγμιαίο ανιχνευόμενο συναίσθημα $\hat{c} = \arg \max_i p_i$.

Για την αντιμετώπιση των ακαριαίων διακυμάνσεων της ταξινόμησης από καρέ σε καρέ, οι οποίες οφείλονται σε θόρυβο ή στιγμιαίες αλλαγές του φωτισμού και θα προκαλούσαν ανεπιθύμητες, ταχείες εναλλαγές στην προβαλλόμενη κατηγορία, εφαρμόζεται τεχνική χρονικής εξομάλυνσης (temporal smoothing) μέσω κυλόμενου παραθύρου (sliding window) πέντε διαδοχικών καρέ ($N = 5$). Σε κάθε νέο καρέ t , η εκτιμώμενη κατηγορία $\hat{c}_t$ εισάγεται στη δομή του παραθύρου $W_t = \{\hat{c}_{t-4}, \hat{c}_{t-3}, \hat{c}_{t-2}, \hat{c}_{t-1}, \hat{c}_t\}$ και ως τελικό συναίσθημα $\hat{C}_t$ επιστρέφεται η επικρατέστερη τιμή, δηλαδή η στατιστική δεσπόζουσα τιμή (mode), που φαίνεται στην ακόλουθη σχέση (7.4):

$$\hat{C}_t = \text{mode}(W_t) \quad (7.4)$$

Η προσέγγιση αυτή εξασφαλίζει τη σταθερότητα της οπτικής ανατροφοδότησης χωρίς ανεπιθύμητες μεταπτώσεις, διατηρώντας παράλληλα την απαιτούμενη ευαισθησία στις πραγματικές μεταβολές των εκφράσεων του προσώπου (Chowdary, Nguyen and Hemanth, 2023) [6].

Τέλος, καθορίζεται ένα ελάχιστο κατώφλι εμπιστοσύνης (confidence threshold) ίσο με 0.4, το οποίο λειτουργεί ως φίλτρο ποιότητας των δεδομένων. Αποτελέσματα με βαθμό βεβαιότητας μικρότερο του 40% ($\max_i p_i < 0,40$) απορρίπτονται αυτομάτως από το σύστημα και δεν ενημερώνουν το περιβάλλον αλληλεπίδρασης χρήστη (UI), αποτρέποντας με αυτόν τον τρόπο την προβολή λανθασμένων ή ασταθών ταξινομήσεων που οφείλονται σε ακατάλληλες συνθήκες φωτισμού ή ασαφείς εκφράσεις του χρήστη.

7.4 Υλοποίηση της διεπαφής χρήστη

Η σχεδίαση και η τεχνική υλοποίηση της διεπαφής χρήστη (UI/UX) της εφαρμογής «Express Your Sound» εστιάζει στην παροχή μιας άμεσης, διαισθητικής και αποκρινόμενης εμπειρίας αλληλεπίδρασης. Η κύρια οθόνη της εφαρμογής (*MainActivity*) αναλαμβάνει την άμεση αλληλεπίδραση με τον χρήστη, προβάλλοντας τη ζωντανή ροή της κάμερας σε περιβάλλον πλήρους οθόνης (full-screen preview). Επάνω από τη ροή αυτή επιστρώνεται ένα διαφανές επίπεδο οπτικοποίησης (overlay), το οποίο υλοποιείται ως προσαρμοσμένο γραφικό συστατικό (*FaceOverlayView*) επεκτείνοντας την κλάση *View* του Android Framework. Μέσω του API σχεδίασης Canvas, το συστατικό αυτό σχεδιάζει δυναμικά ένα πλαίσιο οριοθέτησης γύρω από την ανιχνευόμενη περιοχή του προσώπου. Το χρώμα του πλαισίου μεταβάλλεται σε πραγματικό χρόνο ανάλογα με τη διαγνωσμένη κατηγορία συναισθήματος, χρησιμοποιώντας χρωματικές συμβάσεις όπως το κίτρινο για τη χαρά, το μπλε για τη θλίψη και το κόκκινο για τον θυμό, προσφέροντας έτσι άμεση και διαισθητική οπτική ανατροφοδότηση.

Στο κάτω τμήμα της κύριας οθόνης ενσωματώνεται ένα ημιδιαφανές πάνελ πληροφοριών, το οποίο συγκεντρώνει τα βασικά ευρήματα της συναισθηματικής ανάλυσης. Συγκεκριμένα, περιλαμβάνει το αντίστοιχο εικονίδιο (emoji), τη λεκτική ονομασία του ανιχνευόμενου συναισθήματος, μια σύντομη περιγραφή της κατάστασης, καθώς και το ποσοστό εμπιστοσύνης της πρόβλεψης (confidence percentage). Η τιμή αυτή οπτικοποιείται επιπλέον μέσω μιας οριζόντιας μπάρας προόδου (*ProgressBar*). Το στοιχείο αλληλεπίδρασης (κουμπί) για την ανάκτηση μουσικών προτάσεων παραμένει απενεργοποιημένο και ενεργοποιείται μόνο εφόσον ανιχνευθεί έγκυρο συναίσθημα που υπερβαίνει το καθορισμένο κατώφλι εμπιστοσύνης ($\tau \geq 0,40$), διασφαλίζοντας έτσι ότι ο χρήστης δεν μπορεί να μεταβεί στην οθόνη προτάσεων χωρίς την ολοκλήρωση μιας αξιόπιστης ανάλυσης.

Η δευτερεύουσα οθόνη προβολής προτάσεων (*RecommendationsActivity*) αναλαμβάνει την παρουσίαση της παραγόμενης λίστας των δέκα τραγουδιών μέσω του συστατικού κύλισης *RecyclerView*. Για τη βέλτιστη διαχείριση της μνήμης RAM και την αποδοτική ενημέρωση των στοιχείων της λίστας χωρίς την ανάγκη ολικής επανασχεδίασης της οθόνης, αξιοποιείται η αρχιτεκτονική *ListAdapter* σε συνδυασμό με τον αλγόριθμο υπολογισμού διαφορών *DiffUtil*. Κάθε μεμονωμένο στοιχείο της λίστας (*item_track*) προβάλλει το εικαστικό εξωφύλλο του άλμπουμ, το οποίο ανακτάται και αποθηκεύεται προσωρινά ασύγχρονα μέσω της βιβλιοθήκης Glide, καθώς και τους τίτλους του τραγουδιού, του καλλιτέχνη και του άλμπουμ. Παράλληλα, παρέχεται ένας σύνδεσμος άμεσης μετάβασης (deep link) της μορφής *spotify:track:{id}*, ο οποίος ανοίγει το αντίστοιχο κομμάτι απευθείας στην εφαρμογή Spotify. Στην περίπτωση που η εφαρμογή Spotify δεν είναι εγκατεστημένη στη συσκευή του χρήστη, το σύστημα διαχειρίζεται το σφάλμα προωθώντας αυτόματα τον χρήστη στην αντίστοιχη διαδικτυακή πλατφόρμα (web player) μέσω *Implicit Intent*.

Η συνολική κατάσταση της οθόνης προτάσεων μοντελοποιείται με δομημένο τρόπο μέσω μιας περιχαρακωμένης κλάσης (sealed class) της γλώσσας Kotlin, με την ονομασία `RecommendationsState`. Η κλάση αυτή ορίζει τρεις διακριτές και αμοιβαία αποκλειόμενες καταστάσεις διεπαφής (UI states): την κατάσταση αναμονής (Loading) με προβολή δείκτη προόδου κατά τη διάρκεια εκτέλεσης των ασύγχρονων δικτυακών αιτημάτων, κατάσταση επιτυχίας που περιλαμβάνει τη λίστα των ανακτηθέντων μουσικών κομματιών (`List<Track>`) και τροφοδοτεί τα ανακτηθέντα τραγούδια στον `ListAdapter` για την προβολή των αποτελεσμάτων, και την κατάσταση σφάλματος (Error) που περιλαμβάνει το μήνυμα αποτυχίας (`String`) και προβάλλει την κατάλληλη ειδοποίηση προς τον χρήστη σε περίπτωση απώλειας σύνδεσης ή αποτυχίας του API. Η συγκεκριμένη αρχιτεκτονική προσέγγιση εξασφαλίζει μια αυστηρά καθορισμένη και προβλέψιμη διαχείριση των ασύγχρονων κλήσεων στο API, αποτρέποντας ασυνεπείς καταστάσεις της διεπαφής (invalid UI states), ενισχύοντας τη σταθερότητα του λογισμικού και αναβαθμίζοντας την συνολική εμπειρία του χρήστη.

7.5 Διασύνδεση με το Spotify

Η διαδικασία εξουσιοδότησης πρόσβασης στους πόρους του λογαριασμού Spotify του χρήστη υλοποιείται μέσω του πρωτοκόλλου OAuth 2.0 Authorization Code Flow, το οποίο αποτελεί τον διεθνώς ενδεδειγμένο μηχανισμό για εφαρμογές κινητών συσκευών που απαιτούν ασφαλή αλληλεπίδραση με προστατευμένα δεδομένα. Η ροή δικαιωμάτων εκκινεί με την προβολή του σημείου τελικής πρόσβασης εξουσιοδότησης (authorization endpoint) εντός περιβάλλοντος Chrome Custom Tab. Η επιλογή αυτή εξασφαλίζει ότι ο χρήστης καταχωρίζει τα διαπιστευτήριά του σε ένα απομονωμένο και έμπιστο περιβάλλον φυλλομετρητή, αποτρέποντας την έκθεση ευαίσθητων δεδομένων (όπως ο κωδικός πρόσβασης) στην ίδια την εφαρμογή. Μετά την επιτυχή αυθεντικοποίηση, η πλατφόρμα του Spotify ανακατευθύνει την κλήση στο προσαρμοσμένο σχήμα αναγνωριστικού (`expresssound://callback`) μεταφέροντας την παράμετρο του κωδικού εξουσιοδότησης (code), το οποίο συλλαμβάνεται από την εφαρμογή μέσω κατάλληλα διαμορφωμένου φίλτρου προθέσεων (*deep link intent filter*).

Αμέσως μετά, ο κωδικός εξουσιοδότησης ανταλλάσσεται με τα διακριτικά πρόσβασης (`access_token`) και ανανέωσης (`refresh_token`) μέσω μιας σύγχρονης αίτησης τύπου *POST* προς το endpoint `/api/token`. Η κλήση αυτή εκτελείται κάνοντας χρήση του μηχανισμού HTTP Basic Authentication, όπου το αναγνωριστικό (*Client ID*) και το μυστικό κλειδί (*Client Secret*) της εφαρμογής διαβιβάζονται κωδικοποιημένα κατά το πρότυπο *Base64* (`Base64(ClientID:ClientSecret)`). Τα παραληφθέντα διακριτικά αποθηκεύονται με ασφάλεια στο τοπικό σύστημα αποθήκευσης προτιμήσεων (*SharedPreferences*) της συσκευής, ενώ η ειδικευμένη κλάση *TokenManager* αναλαμβάνει την αυτόματη ανανέωση του διακριτικού πρόσβασης μέσω του `refresh_token` όταν πλησιάζει ο χρόνος λήξης του (συγκεκριμένα εντός παραθύρου πέντε λεπτών πριν τη λήξη), διασφαλίζοντας έτσι την αδιάλειπτη εμπειρία του χρήστη χωρίς την ανάγκη συχνής επανεξουσιοδότησης.

Η δικτυακή επικοινωνία με τις διεπαφές του Spotify Web API δομείται πάνω στη βιβλιοθήκη Retrofit, έναν τύπου-ασφαλή (type-safe) πελάτη HTTP για το περιβάλλον Android. Η Retrofit αναλαμβάνει την αυτόματη μετατροπή των αποκρίσεων μορφής JSON σε αντικείμενα δεδομένων της γλώσσας Kotlin (Data Classes) μέσω του μετατροπέα της βιβλιοθήκης Gson. Για τον αυστηρό διαχωρισμό των αρμοδιοτήτων, ορίζονται δύο ανεξάρτητες διεπαφές: η *SpotifyApiService* για τις συνήθεις κλήσεις δεδομένων προς τη διεύθυνση `[api.spotify.com/v1/](https://api.spotify.com/v1/)` και η *SpotifyAuthService* για τις ειδικές κλήσεις αυθεντικοποίησης προς τη διεύθυνση `accounts.spotify.com`. Όλες οι αιτήσεις δικτύου εκτελούνται ασύγχρονα στο παρασκήνιο με τη χρήση των μηχανισμών Kotlin συνρουτινών (Kotlin Coroutines) πάνω στον ρυθμιστή νημάτων εισόδου/εξόδου (Dispatchers.IO),

αποτρέποντας έτσι οποιαδήποτε δέσμευση ή καθυστέρηση του κύριου νήματος σχεδίασης της εφαρμογής (UI thread).

Στο επίπεδο της αρχιτεκτονικής, η κλάση *SpotifyRepository* αποτελεί το κεντρικό σημείο διαχείρισης των αιτημάτων του δικτύου, εφαρμόζοντας το μοτίβο σχεδίασης Repository Pattern. Η κλάση αυτή ενθυλακώνει την πολυπλοκότητα της δικτυακής επικοινωνίας, τη διαχείριση των διακριτικών αυθεντικοποίησης και τη λογική εναλλακτικής πορείας (fallback mechanism), προσφέροντας στα μοντέλα προβολής (ViewModels) μια απλοποιημένη διεπαφή που επιστρέφει αντικείμενα τύπου *Result<T>* για την ασφαλή διαχείριση τόσο των περιπτώσεων επιτυχίας όσο και των πιθανών σφαλμάτων.

Η παραγωγή των μουσικών προτάσεων υλοποιείται πλέον μέσω του σημείου πρόσβασης *GET /search*, καθώς τα σημεία πρόσβασης *GET /recommendations* και *GET /audio-features* αποσύρθηκαν από το Spotify για εφαρμογές σε στάδιο ανάπτυξης (Development Mode), καθιστώντας τη βασισμένη σε αρχικά δείγματα (seed-based) προσέγγιση μη διαθέσιμη. Στο πλαίσιο αυτό, ο πίνακας αντιστοίχισης συναισθήματος σε ηχητικά χαρακτηριστικά της κλάσης *EmotionAudioMapper* διατηρείται ως ο θεωρητικός πυρήνας του συστήματος, αξιοποιούμενος πλέον για τη διαμόρφωση στοχευμένων όρων αναζήτησης βάσει μουσικού είδους και διάθεσης. Για κάθε διαγνωσμένη συναισθηματική κατηγορία ορίζεται μια διαβαθμισμένη ακολουθία έως και τριών διαδοχικών ερωτημάτων αναζήτησης (για παράδειγμα, στην περίπτωση της λύπης εκτελούνται διαδοχικά οι όροι *genre:acoustic*, *genre:acoustic sad* και *acoustic*), η οποία εκτελείται μέχρι την ανάκτηση μη κενών αποτελεσμάτων. Η συγκεκριμένη ιεράρχηση σχεδιάστηκε ώστε τα αμιγώς θεματικά ερωτήματα είδους να προηγούνται της αναζήτησης με ελεύθερο κείμενο συναισθηματικών όρων, αποτρέποντας λανθασμένες αντιστοιχίσεις με τίτλους μουσικών συλλογών. Επιπρόσθετα, στα ανακτηθέντα αποτελέσματα εφαρμόζεται ένας μηχανισμός φιλτραρίσματος για την απόρριψη συλλογών (αποκλείοντας εγγραφές με καλλιτέχνη «Various Artists» ή τίτλους άλμπουμ με όρους όπως «compilation», «greatest hits» και «playlist»), με αυτόματη επαναφορά στα αρχικά αποτελέσματα στην περίπτωση που η λίστα αποβεί κενή. Παράλληλα, η κλήση *GET /me/top/tracks* με τις παραμέτρους *limit=5* και *time_range=short_term*, προκειμένου να ανακτηθούν τα πέντε δημοφιλέστερα τραγούδια που έχει ακούσει ο χρήστης κατά τις τελευταίες τέσσερις εβδομάδες, διατηρείται αποκλειστικά ως μηχανισμός αποτύπωσης του ακροαματικού προφίλ του χρήστη, χωρίς ωστόσο να τροφοδοτεί άμεσα τη διαδικασία παραγωγής των συστάσεων, λόγω των υφιστάμενων περιορισμών της διεπαφής του Spotify API.

7.6 Διαχείριση δικαιωμάτων πρόσβασης και ασφάλεια

Η αρχιτεκτονική ασφάλειας της εφαρμογής βασίζεται στο επίσημο μοντέλο δικαιωμάτων του λειτουργικού συστήματος Android και διαχωρίζει τις απαιτούμενες άδειες σε δύο διακριτές κατηγορίες με βάση το επίπεδο επικινδυνότητάς τους για την ιδιωτικότητα του χρήστη. Η πρώτη κατηγορία αφορά το δικαίωμα πρόσβασης στην κάμερα της συσκευής (*android.permission.CAMERA*), το οποίο κατατάσσεται στα επικίνδυνα δικαιώματα (dangerous permissions). Σύμφωνα με τις προδιαγραφές που εισήχθησαν στην έκδοση Android 6.0 και μεταγενέστερες εκδόσεις, η άδεια αυτή απαιτεί υποχρεωτικά τη ρητή έγκριση του χρήστη κατά τον χρόνο εκτέλεσης της εφαρμογής (runtime permission request).

Το σύστημα ελέγχει δυναμικά την κατάσταση του δικαιώματος πριν από την εκκίνηση της ροής εισόδου της κάμερας CameraX, μέσω των σύγχρονων APIs αλληλεπίδρασης (*ActivityResultLauncher* και *RequestPermission contract*). Σε περίπτωση που ο χρήστης αρνηθεί τη χορήγηση της σχετικής άδειας, η εφαρμογή διαχειρίζεται την εξαίρεση προβάλλοντας ένα σαφές ενημερωτικό μήνυμα μέσω του περιβάλλοντος διεπαφής, ενώ παράλληλα απενεργοποιεί με ασφάλεια τη λειτουργία αναγνώρισης

συναισθήματος χωρίς να προκαλέσει αστάθεια ή τερματισμό του λογισμικού. Αντίθετα, τα δικαιώματα διαδικτυακής σύνδεσης και ελέγχου της κατάστασης δικτύου (*INTERNET* και *ACCESS_NETWORK_STATE*) εντάσσονται στα συνήθη δικαιώματα (normal permissions). Οι άδειες αυτές δηλώνονται στο αρχείο *AndroidManifest.xml* και χορηγούνται αυτόματα από το σύστημα κατά την εγκατάσταση της εφαρμογής, καθώς δεν ενέχουν άμεσο κίνδυνο για τα προσωπικά δεδομένα.

Σχετικά με τη διαχείριση και την προστασία των βιομετρικών δεδομένων, το σύστημα εφαρμόζει αυστηρά την αρχή της ελαχιστοποίησης των δεδομένων (data minimization). Η επεξεργασία των ροών βίντεο, ο εντοπισμός των χαρακτηριστικών προσώπου μέσω MediaPipe και των εικόνων του προσώπου για την εξαγωγή των συναισθηματικών χαρακτηριστικών πραγματοποιείται αποκλειστικά στην τοπική μνήμη RAM της συσκευής (on-device processing) μέσω της βιβλιοθήκης TensorFlow Lite. Κανένα καρτέ βίντεο, αρχείο εικόνας, συντεταγμένη προσώπου ή ενδιάμεσο βιομετρικό χαρακτηριστικό δεν αποθηκεύεται μόνιμα στο σύστημα αρχείων ούτε μεταφέρεται σε εξωτερικούς διακομιστές ή υπηρεσίες υπολογιστικού νέφους (cloud services), μειώνοντας την πιθανότητα υποκλοπής ή διαρροής ευαίσθητων προσωπικών δεδομένων στο δίκτυο.

Τέλος, η αποθήκευση των ευαίσθητων διακριτικών ταυτοποίησης του Spotify (*access_token* και *refresh_token*) υλοποιείται μέσω του τοπικού μηχανισμού SharedPreferences. Για τη διασφάλιση της ακεραιότητας των δεδομένων αυτών χρησιμοποιείται αποκλειστικά η παράμετρος λειτουργίας *MODE_PRIVATE*. Η επιλογή αυτή επιβάλλει στο επίπεδο του πυρήνα του λειτουργικού συστήματος (Linux kernel UID sandboxing) ότι τα συγκεκριμένα αρχεία ρυθμίσεων είναι προσβάσιμα αποκλειστικά και μόνο από το αναγνωριστικό διεργασίας (process ID) της ίδιας της εφαρμογής, αποτρέποντας την ανάγνωση ή την τροποποίησή τους από άλλες εγκατεστημένες εφαρμογές της συσκευής (Google Developers, 2024) [34].

7.7 Δοκιμή και αποσφαλμάτωση της εφαρμογής

Η διαδικασία ποιοτικού ελέγχου και επαλήθευσης της ορθής λειτουργίας της εφαρμογής πραγματοποιήθηκε συνδυαστικά σε εικονικό περιβάλλον Android Virtual Device (AVD) μέσω του ενσωματωμένου εξομοιωτή του Android Studio, καθώς και σε πραγματικές, φυσικές συσκευές Android. Το εικονικό περιβάλλον δοκιμών διαμορφώνεται με επίπεδο διεπαφής προγραμματισμού εφαρμογών API level 30 (Android 11), ενώ ενεργοποιείται υποχρεωτικά η επιτάχυνση υλικού (Hardware Acceleration) για τη βέλτιστη απόδοση του συστήματος και την πιστή αναπαράσταση των συνθηκών εκτέλεσης σε πραγματικές συνθήκες.

Για την αποτελεσματική δοκιμή των λειτουργιών που βασίζονται στην είσοδο της κάμερας εντός του εξομοιωτή, αξιοποιείται η προηγμένη δυνατότητα Virtual Scene Camera. Η τεχνολογία αυτή παρέχει ένα τρισδιάστατο εικονικό περιβάλλον με κινούμενα αντικείμενα και ψηφιακά ανθρώπινα μοντέλα, επιτρέποντας την προσομοίωση της λήψης προσώπων υπό ελεγχόμενες συνθήκες φωτισμού και κίνησης. Παράλληλα, ο εντοπισμός και η ανάλυση των προγραμματιστικών σφαλμάτων (debugging) πραγματοποιείται μέσω του εργαλείου Android Studio Logcat. Το εργαλείο αυτό παρέχει αναλυτική καταγραφή της ροής εκτέλεσης της εφαρμογής σε πραγματικό χρόνο, επιτρέποντας τη λεπτομερή διήθηση των μηνυμάτων συστήματος με βάση το όνομα του πακέτου λογισμικού (*package name*) και το επίπεδο σοβαρότητας των καταγραφών (*DEBUG, INFO, ERROR*).

Ιδιαίτερη έμφαση δίνεται στον έλεγχο και τη διασφάλιση της ακριβούς ακολουθίας αρχικοποίησης και τερματισμού των υποσυστημάτων της εφαρμογής. Το μοντέλο TensorFlow Lite καθώς και ο αγωγός επεξεργασίας δεδομένων MediaPipe pipeline αρχικοποιούνται εφάπαξ (eager initialization) κατά τη δημιουργία του αντικειμένου *FaceAnalyzer*. Η προσέγγιση αυτή αποτρέπει την επανειλημμένη δαπάνη υπολογιστικών πόρων για τη φόρτωση των μοντέλων στη μνήμη κατά την εναλλαγή των καρτέ. Για την αποφυγή των διαρροών μνήμης (memory leaks), οι δεσμευμένοι πόροι των native C++ βιβλιοθηκών και τα αντικείμενα του TFLite Interpreter απελευθερώνονται ρητά και καθαρίζονται εντός της μεθόδου *onDestroy()* του κύκλου ζωής του Activity, διασφαλίζοντας την επιστροφή των πόρων στο λειτουργικό σύστημα.

Επιπλέον, η ασύγχρονη επεξεργασία των καρτέ βίντεο στο νήμα παρασκηνίου (background thread) προστατεύεται αυστηρά από το φαινόμενο των συνθηκών ανταγωνισμού (race conditions). Η προστασία αυτή επιτυγχάνεται με τη χρήση της διακόσμησης *@Volatile* στα κοινά μεταβλητά πεδία (όπως η σημαία επεξεργασίας καρτέ *isProcessing*), η οποία εγγυάται ότι οι αλλαγές στις τιμές των μεταβλητών είναι άμεσα ορατές σε όλα τα ενεργά νήματα εκτέλεσης. Ταυτόχρονα, εφαρμόζονται ατομικές λειτουργίες (atomic operations) μέσω της κλάσης *AtomicInteger* για τη διαχείριση και την ενημέρωση του μετρητή χρονικής εξομάλυνσης (temporal smoothing counter), εξασφαλίζοντας τη μαθηματική ορθότητα των υπολογισμών κατά την παράλληλη εκτέλεση των διεργασιών ανάλυσης (Chowdhury, Idris and Abas, 2024) [3].

Κεφάλαιο 8ο: Αξιολόγηση και συζήτηση αποτελεσμάτων

8.1 Μεθοδολογία αξιολόγησης

Η αξιολόγηση του συστήματος που αναπτύχθηκε στο πλαίσιο της παρούσας διπλωματικής εργασίας οργανώνεται σε δύο ανεξάρτητα αλλά αμοιβαία συμπληρωματικά επίπεδα, καθένα από τα οποία εστιάζει σε διαφορετική πτυχή της λειτουργικότητας και της αποτελεσματικότητάς του.

Το πρώτο επίπεδο επικεντρώνεται στην αξιολόγηση της τεχνικής απόδοσης του εκπαιδευμένου μοντέλου αναγνώρισης συναισθήματος. Η ανάλυση αυτή πραγματοποιείται ποσοτικά με τη χρήση αντικειμενικών δεικτών αξιολόγησης πάνω στο ανεξάρτητο σύνολο δεδομένων δοκιμής του FER-2013, εξετάζοντας παραμέτρους όπως η ακρίβεια ταξινόμησης, η πιστότητα και η υπολογιστική αποδοτικότητα της αρχιτεκτονικής κατά την εκτέλεση στην κινητή συσκευή.

Αντιστοίχως, το δεύτερο επίπεδο αφορά την αξιολόγηση της συνολικής χρηστικότητας, της λειτουργικότητας και της αντιληπτής αποτελεσματικότητας της εφαρμογής από τελικούς χρήστες. Η εμπειρική αυτή αξιολόγηση υλοποιείται μέσω δοκιμών χρηστικότητας με συμμετοχή χρηστών και τη χορήγηση της σταθμισμένης κλίμακας System Usability Scale (SUS), επιτρέποντας τη συλλογή ποιοτικών και ποσοτικών δεδομένων σχετικά με την εμπειρία χρήσης (User Experience).

Η διττή αυτή μεθοδολογική προσέγγιση εξασφαλίζει μια ολοκληρωμένη και σφαιρική εικόνα τόσο της τεχνικής αρτιότητας όσο και της εμπειρικής διάστασης του συστήματος, συνδέοντας την απόδοση του αλγορίθμου μηχανικής μάθησης με την πραγματική του αξία στην καθημερινή αλληλεπίδραση του χρήστη με την εφαρμογή (Bangor, Kortum and Miller, 2008) [\[33\]](#).

8.2 Αξιολόγηση του μοντέλου αναγνώρισης συναισθημάτων

Η αξιολόγηση της ταξινομικής ικανότητας και της δυνατότητας γενίκευσης του προεκπαιδευμένου και προσαρμοσμένου μοντέλου MobileNetV2 πραγματοποιείται αποκλειστικά στο ανεξάρτητο σύνολο δεδομένων δοκιμής του FER-2013. Το συγκεκριμένο υποσύνολο αποτελείται από 7.178 ψηφιοποιημένες εικόνες προσώπου διαστάσεων 48×48 εικονοστοιχείων σε κλίμακα του γκρι. Για τη διασφάλιση της αντικειμενικότητας και της επιστημονικής εγκυρότητας των αποτελεσμάτων, το σύνολο δοκιμής διατηρήθηκε αυστηρά απομονωμένο και δεν χρησιμοποιήθηκε σε κανένα στάδιο της διαδικασίας εκπαίδευσης (training) ή της επικύρωσης και του συντονισμού των υπερπαραμέτρων (validation/hyperparameter tuning). Η διαδικασία αυτή αποσκοπεί στην εξάλειψη του φαινομένου της υπερπροσαρμογής (overfitting) και στην ακριβή αποτίμηση της συμπεριφοράς του μοντέλου κατά την επεξεργασία άγνωστων οπτικών δεδομένων. Οι μετρικές αξιολόγησης που εφαρμόζονται στο παρόν πειραματικό πλαίσιο αντλούν το θεωρητικό τους υπόβαθρο από την ανάλυση της ενότητας 3.4, επιτρέποντας την ποσοτική και ποιοτική αποτίμηση της απόδοσης του δικτύου σε συνθήκες πραγματικής λειτουργίας.

Ως μετρικές αξιολόγησης, η κυριότερη μετρική που χρησιμοποιείται για τη συνολική εκτίμηση της ταξινομικής ισχύος του δικτύου είναι η συνολική ακρίβεια (Overall Accuracy). Η συνολική ακρίβεια ορίζεται και υπολογίζεται ως ο λόγος του αριθμού των ορθώς ταξινομημένων εικόνων προς το συνολικό πλήθος των εικόνων που συνθέτουν το σύνολο δεδομένων δοκιμής. Ως σημείο αναφοράς για την αποτίμηση της απόδοσης του συστήματος, η σύγχρονη βιβλιογραφία αναφέρει τιμή ακρίβειας της τάξης του ~73.82% για αρχιτεκτονικές MobileNetV2 που έχουν υποστεί προσαρμογή μέσω της στρατηγικής λεπτομερούς προσαρμογής στο σύνολο δεδομένων FER-2013 (Sharma, Nagpal and Raja

Praveen, 2025) [2]. Παράλληλα, η ανθρώπινη ακρίβεια αναγνώρισης στο ίδιο ακριβώς σύνολο δεδομένων υπολογίζεται εμπειρικά στο εύρος 65–70% (Goodfellow et al., 2013) [30], γεγονός που αναδεικνύει τον βαθμό δυσκολίας του συγκεκριμένου συνόλου δεδομένων λόγω των αποκλίσεων στις γωνίες λήψης και του θορύβου στις ετικέτες. Τα ποσοτικά αποτελέσματα της παρούσας εκπαιδευτικής διαδικασίας έχουν ενσωματωθεί στον πίνακα 8.1, όπως αυτά εξήχθησαν μετά την ολοκλήρωση των υπολογιστικών κύκλων στο περιβάλλον του Google Colab.

Προκειμένου να εξασφαλιστεί μια ουσιαστικότερη και πιο εις βάθος αποτίμηση της ταξινομικής συμπεριφοράς του μοντέλου, η ανάλυση επεκτείνεται στον υπολογισμό του δείκτη F1-score ανά κατηγορία. Για καθένα από τα επτά βασικά συναισθήματα υπολογίζονται ανεξάρτητα οι δείκτες ακρίβειας (precision), ανακλήσεως (recall) καθώς και το αντίστοιχο F1-score. Βάσει της κατανομής των δεδομένων, είναι αναμενόμενο ότι κατηγορίες με υψηλή αντιπροσώπευση στο σύνολο εκπαίδευσης (training set), όπως η κατηγορία της χαράς με 8.989 εικόνες, θα εμφανίζουν σημαντικά υψηλότερα F1-scores σε σύγκριση με λιγότερο εκπροσωπούμενες κατηγορίες, όπως η κατηγορία της αηδίας με μόλις 547 εικόνες. Η απόκλιση αυτή καταγράφεται παρά τη χρησιμοποίηση σταθμισμένης συνάρτησης σφάλματος για την αντιστάθμιση των κλάσεων. Για την υπέρβαση αυτού του περιορισμού, υπολογίζεται η μέση τιμή της μακρο-μέσου F1-score (macro-averaged F1-score), το οποίο παρέχει μια συνολική μέτρηση της απόδοσης του δικτύου χωρίς να εισάγει προκατάληψη υπέρ των συχνότερα εμφανιζόμενων κατηγοριών (Mahendran et al., 2024) [4].

Συμπληρωματικά, η διαγνωστική ανάλυση των συστηματικών ταξινομικών σφαλμάτων του μοντέλου πραγματοποιείται μέσω της κατασκευής του πίνακα σύγχυσης. Ο πίνακας σύγχυσης διάστασης 7×7 αναπαριστά διαγραμματικά τη συχνότητα όλων των δυνατών συνδυασμών μεταξύ της πραγματικής κατηγορίας (ground truth) και της προβλεπόμενης κατηγορίας από το μοντέλο, επιτρέποντας τον ακριβή εντοπισμό και τη μελέτη των μοτίβων αποκλίσεων. Βάσει της υφιστάμενης βιβλιογραφίας, αναμένονται συγχύσεις κυρίως μεταξύ των συναισθημάτων του φόβου και της έκπληξης, εξαιτίας των κοινών μονάδων δράσης του προσώπου (Action Units: AU1, AU2, AU5) που ενεργοποιούνται ταυτόχρονα και στις δύο εκφράσεις, καθώς και μεταξύ του θυμού και της αηδίας λόγω της παρόμοιας χαμήλωσης των φρυδιών και της ανύψωσης του άνω χείλους (Ekman and Friesen, 1978) [19].

Πίνακας 8-1: Αναφορά ταξινόμησης - σύνολο δοκιμής FER-2013

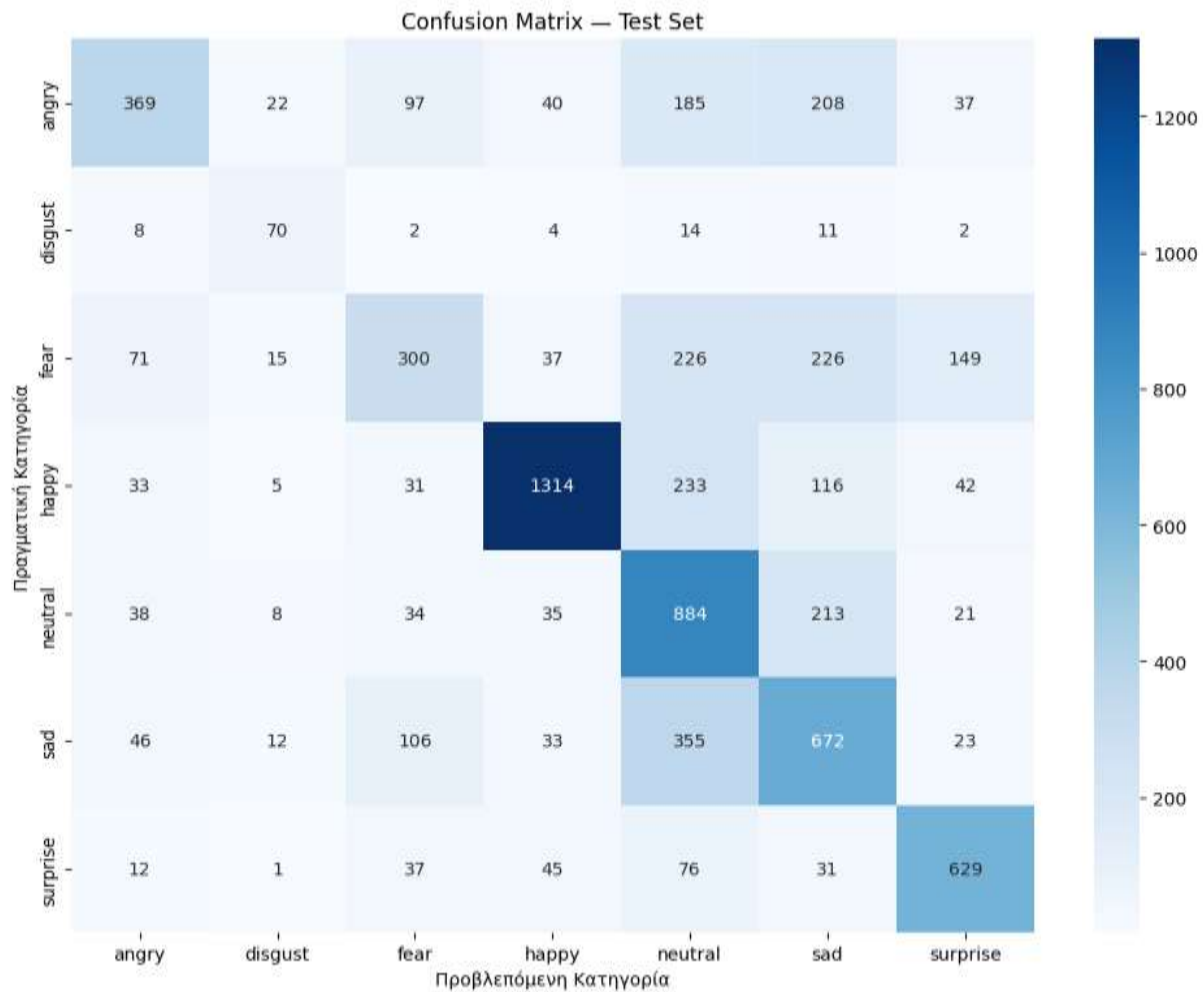
Κατηγορία	Precision	Recall	F1-score	Support
Θυμός	0.64	0.39	0.48	958
Αηδία	0.53	0.63	0.57	111
Φόβος	0.49	0.29	0.37	1024
Χαρά	0.87	0.74	0.80	1774
Λύπη	0.45	0.54	0.49	1247
Έκπληξη	0.70	0.76	0.73	831
Ουδέτερο	0.45	0.72	0.55	1233
Macro avg	0.59	0.58	0.57	7178
Weighted avg	0.62	0.59	0.59	7178
Accuracy			0.59	7178

Η συνολική ακρίβεια (overall accuracy) που πέτυχε το μοντέλο στο σύνολο δοκιμής ανέρχεται σε 59%, ενώ ο αριθμητικός μέσος όρος του μέτρου F1 χωρίς στάθμιση (macro-averaged F1-score) διαμορφώνεται στο 0,57. Η επιλογή του δείκτη macro F1 κρίνεται ιδιαίτερα κρίσιμη για την αμερόληπτη αποτίμηση του συστήματος, καθώς παρέχει μια ισοσταθμισμένη εκτίμηση της γενικευτικής ικανότητας του δικτύου, αποτρέποντας την πλασματική υπερεκτίμηση της επίδοσης από την κυριαρχία των πολυπληθέστερων κλάσεων. Αν και η επιτευχθείσα τιμή υπολείπεται του σημείου αναφοράς (benchmark) του ~73,82% που καταγράφεται στη διεθνή βιβλιογραφία για αρχιτεκτονικές MobileNetV2 κατόπιν εκτεταμένης λεπτομερούς προσαρμογής στο FER-2013 (Sharma, Nagpal and Raja Praveen, 2025) [2], ωστόσο παραμένει συγκρίσιμη με άλλες εργασίες που χρησιμοποιούν την στρατηγική της μεταφοράς της γνώσης χωρίς εκτεταμένη βελτιστοποίηση υπερπαραμέτρων (Chowdhury, Idris and Abas, 2024) [3], ενώ παράλληλα υπερβαίνει το κατώτατο όριο της ανθρώπινης διακριτικής ικανότητας σε συγκεκριμένες απαιτητικές κατηγορίες του συνόλου δεδομένων.

Η συστηματική εξέταση των αποτελεσμάτων ανά συναισθηματική κατηγορία αποκαλύπτει σημαντική ανομοιογένεια στην ικανότητα ταξινόμησης του δικτύου. Η κατηγορία της χαράς επιτυγχάνει τη βέλτιστη απόδοση με μέτρο F1 ίσο με 0,80 και δείκτη ακρίβειας 0,87. Η υπεροχή αυτή ερμηνεύεται αφενός από την αυξημένη αριθμητική της εκπροσώπηση στο σύνολο εκπαίδευσης (7.215 εικόνες από τις συνολικά 8.989 εικόνες του συνόλου δεδομένων, αποτελώντας την πολυπληθέστερη κλάση) και αφετέρου από τα ιδιαίτερα διακριτά οπτικά και γεωμετρικά χαρακτηριστικά που συνοδεύουν την έκφραση της χαράς. Συγκεκριμένα, η ενεργοποίηση του χαμόγελου Duchenne (ταυτόχρονη συστολή των μυϊκών ομάδων AU6 και AU12) δημιουργεί έντονες και εύκολα ανιχνεύσιμες μεταβολές στην περιοχή των παρειών και των χειλών (Ekman, 2003) [18], τις οποίες τα συνελκτικά φίλτρα του MobileNetV2 αποτυπώνουν με υψηλή πιστότητα. Αντίστοιχα ικανοποιητική συμπεριφορά επιδεικνύει η κατηγορία της έκπληξης με μέτρο F1 στο 0,73 και δείκτη ανάκλησης 0,76, ενώ η ουδέτερη κατάσταση επιτυγχάνει υψηλή ανάκληση ίση με 0,72, αν και συνοδεύεται από χαμηλότερη ακρίβεια (0,45) λόγω της τάσης του μοντέλου να την επιλέγει ως εναλλακτική κατηγορία σε ασαφείς εκφράσεις.

Αντιθέτως, η χαμηλότερη αποτελεσματικότητα του συστήματος καταγράφεται στην κατηγορία του φόβου, η οποία σημειώνει ιδιαίτερα χαμηλό δείκτη ανάκλησης ίση με 0,29 και F1-score στο 0,37. Το ευρύ αυτό ποσοστό σφάλματος υποδηλώνει ότι το δίκτυο αδυνατεί να ταξινομήσει ορθά την έκφραση του φόβου στο 71% των πραγματικών περιπτώσεων. Παρόμοια υποβαθμισμένη εικόνα παρουσιάζει και η κατηγορία του θυμού, με μέτρο F1 να ισούται με 0,48 και δείκτη ανάκλησης 0,39, γεγονός που αναδεικνύει τη δυσκολία εξαγωγής ισχυρών διακριτικών χαρακτηριστικών για συναισθήματα που εκδηλώνονται με ηπιότερες μυϊκές συσπάσεις στο πρόσωπο.

Η εμβάθυνση στα μοτίβα των λανθασμένων ταξινομήσεων πραγματοποιείται μέσω της ανάλυσης του πίνακα σύγχυσης (confusion matrix, σχήμα 8-1), ο οποίος φανερώνει τις δομικές αλληλεπικαλύψεις των παραγόμενων αναπαραστάσεων.



Σχήμα 8-1: Πίνακας σύγχυσης συνόλου δοκιμής

Η πλέον συστηματική σύγχυση παρατηρείται μεταξύ της κατηγορίας του φόβου και της ουδέτερης κατάστασης (226 εικόνες φόβου ταξινομήθηκαν εσφαλμένα ως ουδέτερες), καθώς και μεταξύ φόβου και θλίψης (226 ισόποσες περιπτώσεις εσφαλμένης ταξινόμησης). Η φαινοτυπική αυτή σύγχυση αιτιολογείται επιστημονικά από το γεγονός ότι ο φόβος, η θλίψη και η ουδέτερη έκφραση μοιράζονται παρόμοιες ορθοστατικές κινήσεις των οφρύων και του στόματος, ενώ στις χαμηλής ανάλυσης εικόνες του FER-2013 η ένταση των συσπάσεων δεν είναι πάντοτε επαρκής για τον ασφαλή διαχωρισμό τους. Παράλληλα, έντονη αλληλεπικάλυψη καταγράφεται ανάμεσα στον θυμό και τη θλίψη (208 εικόνες θυμού ταξινομήθηκαν ως θλίψη) καθώς και ανάμεσα στον θυμό και την ουδέτερη κατηγορία (185 εικόνες). Απεναντίας, η κατηγορία της χαράς επιβεβαιώνει την υψηλή διακριτότητά της, εμφανίζοντας τη μέγιστη συγκέντρωση παρατηρήσεων στην κύρια διαγώνιο του πίνακα με 1.314 ορθές ταξινομήσεις. (Ekman and Friesen, 1978) [19].

Σε τελευταία ανάλυση, τα πειραματικά αποτελέσματα που επιτεύχθηκαν από το εκπαιδευμένο μοντέλο τίθενται σε συγκριτική αξιολόγηση με αντίστοιχες δημοσιευμένες εργασίες της βιβλιογραφίας. Η σύγκριση εστιάζει σε μελέτες που αξιοποιούν την αρχιτεκτονική MobileNetV2 με τεχνικές προσαρμοσμένης εκπαίδευσης (fine-tuning) πάνω στο ίδιο σύνολο δεδομένων FER-2013, επιτρέποντας την αποτίμηση της αποτελεσματικότητας της ακολουθούμενης στρατηγικής εκπαίδευσης, της επιλογής

των υπερπαραμέτρων και των τεχνικών επαύξησης δεδομένων. Ο πίνακας 8.2 συνοψίζει τα βασικά σημεία αναφοράς (benchmarks) της διεθνούς βιβλιογραφίας σε σχέση με την παρούσα υλοποίηση:

Πίνακας 8-2: Σύγκριση ακρίβειας

Εργασία	Μοντέλο	Ακρίβεια FER-2013
Sharma, Nagpal and Raja Praveen (2025) [2]	MobileNetV2 fine-tuned	73.82%
Kaur and Kulkarni (2024) [5]	MobileNetV2 fine-tuned (FERFM)	85.70%
Chowdhury, Idris and Abas (2024) [3]	MobileNetV2 transfer learning	~61%
Mahendran et al. (2024) [4]	CNN fine-tuned	~72%
Ανθρώπινη απόδοση	—	65–70%
Παρούσα εργασία	MobileNetV2 fine-tuning	59.04%

Η σύγκριση είναι ενδεικτική, καθώς οι επιμέρους εργασίες ενδέχεται να διαφοροποιούνται ως προς τις συνθήκες εκπαίδευσης και αξιολόγησης.

8.3 Αξιολόγηση της εφαρμογής

Η αξιολόγηση της χρησιμότητας, της λειτουργικότητας και της αντιληπτής αποτελεσματικότητας της αναπτυχθείσας εφαρμογής πραγματοποιήθηκε μέσω μελέτης χρηστών με τη συμμετοχή πέντε ατόμων. Η επιλογή του συγκεκριμένου μεγέθους δείγματος βασίζεται στη θεμελιώδη μεθοδολογική προσέγγιση του Nielsen, σύμφωνα με την οποία η εμπλοκή πέντε συμμετεχόντων σε ποιοτικές δοκιμές χρησιμότητας είναι στατιστικά επαρκής για τον εντοπισμό τουλάχιστον του 85% των ουσιαστικών σχεδιαστικών ατελειών, των εμποδίων αλληλεπίδρασης και των προβλημάτων πλοήγησης μιας διεπαφής χρήστη (Nielsen, 1993) [21]. Οι συμμετέχοντες επιλέχθηκαν βάσει συγκεκριμένων κριτηρίων ένταξης, τα οποία περιλάμβαναν το ηλικιακό εύρος 18–35 ετών και την εξοικείωση με τη χρήση έξυπνων κινητών συσκευών (smartphones) με λειτουργικό σύστημα Android. Όλες οι συνεδρίες αξιολόγησης διεξήχθησαν σε ελεγχόμενο φυσικό περιβάλλον με σταθερό και επαρκή φωτισμό, εξασφαλίζοντας τις απαραίτητες συνθήκες για την αξιόπιστη και απρόσκοπτη λειτουργία του αλγορίθμου ανίχνευσης προσώπου από την κάμερα της συσκευής.

8.3.1 Πρωτόκολλο δοκιμής

Για τη διασφάλιση της αντικειμενικότητας και της επαναληψιμότητας των μετρήσεων, κάθε δράση αξιολόγησης ακολούθησε ένα τυποποιημένο και σαφές πρωτόκολλο πέντε διαδοχικών σταδίων.

Στο πρώτο στάδιο (briefing, 5 λεπτά), ο συμμετέχων ενημερώθηκε για τον σκοπό της εφαρμογής και τη διαδικασία αξιολόγησης, χωρίς να έχει λάβει οδηγίες για τον τρόπο χρήσης, ώστε να αξιολογηθεί η αυτονομία της διεπαφής.

Ως προς το δεύτερο στάδιο (τυποποιημένα σενάρια χρήσης, 15–20 λεπτά), ο συμμετέχων εκτέλεσε τέσσερις προκαθορισμένες ενέργειες. Αρχικά, συνδέθηκε με τον Spotify λογαριασμό του ή δημιούργησε νέο λογαριασμό κατά την εκκίνηση της εφαρμογής. Έπειτα, άφησε την εφαρμογή να

αναγνωρίσει την έκφραση του προσώπου του και παρατήρησε τα αποτελέσματα. Στη συνέχεια, η εφαρμογή ανέκτησε μουσικές προτάσεις και αναπαρήγαγε ένα μουσικό κομμάτι αντίστοιχο της έκφρασης του χρήστη και κατά συνέπεια του συναισθήματος που αντιλήφθηκε η εφαρμογή μέσω της διεπαφής. Τέλος, ο χρήστης είχε την δυνατότητα να αποθηκεύσει την μουσική πρόταση που του παρείχε η εφαρμογή, προτού επιστρέψει στην οθόνη της κάμερας.

Κατά το τρίτο στάδιο (SUS ερωτηματολόγιο, 3-5 λεπτά), ο συμμετέχων συμπλήρωσε το ερωτηματολόγιο αξιολόγησης χρηστικότητας της εφαρμογής (System Usability Scale), πριν συζητήσει με οποιονδήποτε άλλο χρήστη για την προσωπική του εμπειρία.

Στη συνέχεια, στο τέταρτο στάδιο (ερωτηματολόγιο μουσικών προτάσεων, 3 λεπτά), ο συμμετέχων αξιολόγησε με κλίμακα Likert 1-5 τέσσερις πτυχές: ακρίβεια αναγνώρισης συναισθήματος, συναισθηματική συμβατότητα μουσικών προτάσεων, την πιθανότητα χρήσης της εφαρμογής στην καθημερινή ζωή του και την εξατομίκευση προτάσεων σε σχέση με την χειροκίνητη επιλογή.

Το εργαστήριο ολοκληρώθηκε με το πέμπτο στάδιο (σύντομη συνέντευξη, 5 λεπτά), όπου συλλέχθηκαν ποιοτικά σχόλια για τα θετικά και αρνητικά στοιχεία της εφαρμογής, καθώς και προτάσεις βελτίωσης.

Πριν την έναρξη της διαδικασίας, κάθε συμμετέχων ενημερώθηκε αναλυτικά για τον σκοπό και τη φύση της μελέτης και υπέγραψε έντυπο συγκατάθεσης, το οποίο κάλυπτε τόσο τη συμμετοχή στη μελέτη όσο και την ενδεχόμενη χρήση φωτογραφικού υλικού στο πλαίσιο της παρούσας διπλωματικής εργασίας. Ενδεικτικό φωτογραφικό υλικό από τη διαδικασία της μελέτης χρηστών, καθώς και το υπογεγραμμένο έντυπο συγκατάθεσης, παρατίθενται στα παραρτήματα Α και Β.

8.4 Αξιολόγηση χρηστικότητας εφαρμογής

Το System Usability Scale (SUS) αποτελεί το πλέον διαδεδομένο και επιστημονικά επικυρωμένο εργαλείο αξιολόγησης χρηστικότητας, το οποίο αναπτύχθηκε αρχικά από τον Brooke (1996) στο πλαίσιο της αξιολόγησης συστημάτων της Digital Equipment Corporation. Το εργαλείο αποτελείται από δέκα προτάσεις αποτίμησης σε πενταβάθμια κλίμακα Likert (από το 1 που αντιστοιχεί στο «Διαφωνώ απόλυτα» έως το 5 που αντιστοιχεί στο «Συμφωνώ απόλυτα»), οι οποίες είναι εναλλάξ θετικά και αρνητικά διατυπωμένες προκειμένου να αποτρέπεται η συστηματική προκατάληψη των απαντήσεων.

Η επεξεργασία των απαντήσεων παράγει μια συνολική βαθμολογία στο εύρος 0 έως 100. Ο υπολογισμός της βαθμολογίας πραγματοποιείται με συγκεκριμένο μαθηματικό κανόνα: για τις μονές ερωτήσεις (1, 3, 5, 7, 9) το επιμέρους σκορ ισούται με την τιμή της απάντησης μείον 1, ενώ για τις ζυγές ερωτήσεις (2, 4, 6, 8, 10) το σκορ ισούται με την τιμή 5 μείον την τιμή της απάντησης. Το άθροισμα των επιμέρους αυτών σκορ πολλαπλασιάζεται στη συνέχεια επί τον συντελεστή 2,5 για την παραγωγή της τελικής βαθμολογίας SUS.

Βάσει εκτεταμένης εμπειρικής μετα-ανάλυσης 500 μελετών χρηστικότητας, καταδείχθηκε ότι βαθμολογία SUS μεγαλύτερη ή ίση του 68 αντιστοιχεί σε αποδεκτό επίπεδο χρηστικότητας πάνω από τον γενικό μέσο όρο, βαθμολογία μεγαλύτερη ή ίση του 74 υποδηλώνει «καλή» χρηστικότητα, ενώ τιμές άνω του 85 αντιστοιχούν σε «εξαιρετική» χρηστικότητα (Bangor, Kortum and Miller, 2008) [33]. Ως αρχικός στόχος αξιολόγησης για την αναπτυχθείσα εφαρμογή ορίστηκε η επίτευξη βαθμολογίας SUS μεγαλύτερης ή ίσης του 68.

Τα 10 ερωτήματα του SUS όπως χορηγήθηκαν στους συμμετέχοντες παρατίθενται παρακάτω στον πίνακα 8-3.

Πίνακας 8-3: Ερωτηματολόγιο SUS

A/A	Πρόταση
1	Θα ήθελα να χρησιμοποιώ αυτή την εφαρμογή συχνά
2	Βρήκα την εφαρμογή περιττά περίπλοκη
3	Η εφαρμογή ήταν εύκολη στη χρήση
4	Χρειάστηκα τη βοήθεια κάποιου για να τη χρησιμοποιήσω
5	Οι διάφορες λειτουργίες της εφαρμογής ήταν καλά ενσωματωμένες
6	Υπήρχαν πολλές ασυνέπειες στην εφαρμογή
7	Πιστεύω ότι οι περισσότεροι άνθρωποι θα μάθαιναν γρήγορα να τη χρησιμοποιούν
8	Βρήκα την εφαρμογή πολύ δύσχρηστη
9	Ένιωσα πολύ σίγουρος/η χρησιμοποιώντας την εφαρμογή
10	Έπρεπε να μάθω πολλά πράγματα πριν αρχίσω να τη χρησιμοποιώ

Συμπληρωματικά του SUS, δόθηκε επιπλέον ένα ειδικά σχεδιασμένο ερωτηματολόγιο τεσσάρων ερωτήσεων Likert 1-5 για την αξιολόγηση της ποιότητας των μουσικών προτάσεων, όπως φαίνεται στον πίνακα 8-4:

Πίνακας 8-4: Ερωτηματολόγιο αξιολόγησης προτάσεων

A/A	Πρόταση
1	Η εφαρμογή αναγνώρισε σωστά τη συναισθηματική μου κατάσταση
2	Τα προτεινόμενα τραγούδια ταίριαζαν στη διάθεσή μου
3	Θα χρησιμοποιούσα αυτές τις μουσικές προτάσεις στην καθημερινή μου ζωή
4	Οι προτάσεις ήταν πιο εξατομικευμένες από ό,τι θα επέλεγα χειροκίνητα

Η αξιολόγηση χρηστικότητας πραγματοποιήθηκε με τη συμμετοχή πέντε ατόμων, τα οποία ακολούθησαν το τυποποιημένο πρωτόκολλο δοκιμής που περιγράφεται στην 8.3.1. Αναγνωρίζεται ότι το μέγεθος δείγματος των πέντε συμμετεχόντων αποτελεί περιορισμό ως προς τη στατιστική γενίκευση των αποτελεσμάτων, ωστόσο παρέχει αξιόπιστες ενδείξεις για την ποιότητα της διεπαφής χρήστη στο πλαίσιο πιλοτικής αξιολόγησης προτύπου εφαρμογής. Ο πίνακας 8-5 παρουσιάζει τις απαντήσεις των πέντε συμμετεχόντων στα δέκα ερωτήματα του SUS ερωτηματολογίου.

Πίνακας 8-5: Απαντήσεις χρηστών στα ερωτήματα SUS (κλίμακα Likert 1–5)

A/A	Ερώτηση	Χρήστης 1	Χρήστης 2	Χρήστης 3	Χρήστης 4	Χρήστης 5	M.O.
1	Θα ήθελα να χρησιμοποιώ αυτή την εφαρμογή συχνά	4	5	4	4	5	4,40
2	Βρήκα την εφαρμογή περιττά περίπλοκη	2	1	2	1	2	1,60
3	Η εφαρμογή ήταν εύκολη στη χρήση	4	5	5	4	4	4,40
4	Χρειάστηκα τη βοήθεια κάποιου για να τη χρησιμοποιήσω	1	1	2	1	1	1,20
5	Οι διάφορες λειτουργίες της εφαρμογής ήταν καλά ενσωματωμένες	4	4	4	5	4	4,20
6	Υπήρχαν πολλές ασυνέπειες στην εφαρμογή	2	1	1	2	2	1,60
7	Πιστεύω ότι οι περισσότεροι άνθρωποι θα μάθαιναν γρήγορα να τη χρησιμοποιούν	5	5	4	5	4	4,60
8	Βρήκα την εφαρμογή πολύ δύσκολη	1	1	2	1	1	1,20
9	Ένιωσα πολύ σίγουρος/η χρησιμοποιώντας την εφαρμογή	4	4	4	4	3	3,80
10	Έπρεπε να μάθω πολλά πράγματα πριν αρχίσω να τη χρησιμοποιώ	2	1	2	2	1	1,60

Μετά την εφαρμογή του τυποποιημένου αλγορίθμου βαθμολόγησης SUS, οι τελικές βαθμολογίες των πέντε συμμετεχόντων διαμορφώθηκαν ως εξής: X1 = 82,50, X2 = 95,00, X3 = 80,00,

$X4 = 87,50$ και $X5 = 82,50$. Η μέση βαθμολογία SUS ανέρχεται σε 85,50 με τυπική απόκλιση 5,97, ελάχιστη τιμή 80,00 και μέγιστη 95,00.

Ο πίνακας 8-6 παρουσιάζει τα αποτελέσματα του ειδικού ερωτηματολογίου αξιολόγησης της ποιότητας των μουσικών προτάσεων. Ο συνολικός μέσος όρος ικανοποίησης του ειδικού ερωτηματολογίου διαμορφώνεται σε 4,05/5,00 (81,0%), τιμή που αντικατοπτρίζει υψηλό επίπεδο αντιληπτής αποτελεσματικότητας του συστήματος σύστασης μουσικής.

Πίνακας 8-6: Αποτελέσματα αξιολόγησης μουσικών προτάσεων

A/A	Ερώτηση	X1	X2	X3	X4	X5	M.O.	Τυπική απόκλιση (SD)
1	Η εφαρμογή αναγνώρισε σωστά τη συναισθηματική μου κατάσταση	4	4	5	4	4	4,20	0,45
2	Τα προτεινόμενα τραγούδια ταίριαζαν στη διάθεσή μου	4	5	4	4	5	4,40	0,55
3	Θα χρησιμοποιούσα αυτές τις μουσικές προτάσεις στην καθημερινή μου ζωή	4	4	3	5	4	4,00	0,71
4	Οι προτάσεις ήταν πιο εξατομικευμένες από ό,τι θα επέλεγα χειροκίνητα	3	4	4	3	4	3,60	0,55

8.5 Συζήτηση αποτελεσμάτων

Συνθέτοντας τα ανωτέρω ευρήματα, η επίτευξη της συνολικής ακρίβειας του 59% αντικατοπτρίζει τους εγγενείς περιορισμούς τόσο του συνόλου δεδομένων όσο και του υπολογιστικού πλαισίου. Το FER-2013 χαρακτηρίζεται από υψηλό ποσοστό θορύβου στις ετικέτες (label noise της τάξης του ~10–15%) λόγω της διαδικασίας συλλογής του μέσω crowdsourcing, ενώ η ανθρώπινη ακρίβεια ταξινόμησης στο συγκεκριμένο σύνολο δεδομένων εκτιμάται περιορισμένη, μεταξύ 65% και 70% (Goodfellow et al., 2013). Υπό αυτό το πρίσμα, η διαπιστωμένη απόκλιση των 15 ποσοστιαίων μονάδων από το σημείο αναφοράς των 73,82% αποδίδεται συνδυαστικά αφενός στους υπολογιστικούς περιορισμούς κατά τη φάση της εκπαίδευσης που δεν επέτρεψαν την εξαντλητική αναζήτηση υπερπαραμέτρων (hyperparameter grid search) και την εφαρμογή σύνθετων τεχνικών τακτοποίησης, αφετέρου στην έντονη στατιστική ανισορροπία των κλάσεων που διατήρησε μια ελαφρά μεροληψία του δικτύου υπέρ των πολυπληθέστερων κατηγοριών παρά την εφαρμογή σταθμισμένης συνάρτησης

απώλειας, και τέλος στην εγγενή αντικειμενική δυσκολία διαχωρισμού των συναισθηματικών εκφράσεων χαμηλής έντασης ή υψηλής ασαφείας σε περιβάλλοντα ανεξέλεγκτων συνθηκών λήψης (in-the-wild facial images) (Li and Deng, 2022) [7].

Σε επίπεδο χρηστικότητας, η οποία πραγματοποιήθηκε με τη συμμετοχή πέντε ατόμων, τα οποία ακολούθησαν το τυποποιημένο πρωτόκολλο δοκιμής που περιγράφεται στην ενότητα 8.3.1. Όπως προαναφέρεται, αναγνωρίζεται ότι το μέγεθος δείγματος των πέντε συμμετεχόντων αποτελεί περιορισμό ως προς τη στατιστική γενίκευση των αποτελεσμάτων, ωστόσο παρέχει αξιόπιστες ενδείξεις για την ποιότητα της διεπαφής χρήστη στο πλαίσιο πιλοτικής αξιολόγησης προτύπου εφαρμογής. Τα αποτελέσματα του ειδικού ερωτηματολογίου αξιολόγησης της ποιότητας των μουσικών προτάσεων παρουσιάζονται στον πίνακα 8-6 (ενότητα 8.4).

Ως προς την επίτευξη μέσης βαθμολογίας 85,50 όπως είδαμε στο κεφάλαιο 8.4, υπερβαίνει σημαντικά τον αρχικό στόχο αξιολόγησης των ≥ 68 μονάδων που είχε τεθεί στην ενότητα 5.1, και εντάσσεται στην κατηγορία «Εξαιρετική Χρηστικότητα» (Excellent / Grade A) σύμφωνα με τα εμπειρικά benchmarks των Bangor, Kortum and Miller (2008) [33], τα οποία ορίζουν βαθμολογία ≥ 85 ως άριστη. Αξιοσημείωτο είναι ότι όλοι οι συμμετέχοντες ξεχωριστά επέτυχαν βαθμολογία άνω των 80 μονάδων, καταδεικνύοντας ομοιογένεια στη θετική εκτίμηση της χρηστικότητας.

Η ανάλυση των επιμέρους ερωτημάτων αναδεικνύει στοιχεία ιδιαίτερου ενδιαφέροντος. Η υψηλότερη βαθμολογία καταγράφεται στο ερώτημα 7 («Οι περισσότεροι άνθρωποι θα μάθαιναν γρήγορα», M.O. 4,60), γεγονός που υποδεικνύει ότι η εφαρμογή κρίνεται ευρέως ως διαισθητική και εύκολα εκμαθητέα από το σύνολο των συμμετεχόντων. Εξίσου υψηλές βαθμολογίες συγκεντρώνουν τα ερωτήματα 1 και 3 (M.O. 4,40 αμφότερα), που αφορούν την επιθυμία συχνής χρήσης και την ευκολία χρήσης αντίστοιχα, ενώ η σχεδόν μηδενική βαθμολογία στα αρνητικά ερωτήματα 4 και 8 (M.O. 1,20 αμφότερα, επί κλίμακας με κατώτατη τιμή το 1) επιβεβαιώνει ότι κανείς συμμετέχων δεν αντιμετώπισε ουσιαστικές δυσκολίες κατά τη χρήση. Η μόνη σχετικά χαμηλότερη βαθμολογία εντοπίζεται στο ερώτημα 9 («Ένιωσα σίγουρος/η», M.O. 3,80), υποδηλώνοντας ότι ορισμένοι χρήστες ένιωθαν μέτρια αβεβαιότητα, πιθανώς λόγω της νέας έννοιας της αυτόματης αναγνώρισης συναισθήματος που απαιτεί εξοικείωση. Τα αποτελέσματα του ειδικού ερωτηματολογίου αξιολόγησης της ποιότητας των μουσικών προτάσεων παρουσιάζονται στον πίνακα 8-6 στο κεφάλαιο 8.4.

Παράλληλα, η υψηλότερη βαθμολογία καταγράφεται στο ερώτημα 2 («Τα προτεινόμενα τραγούδια ταίριαζαν στη διάθεσή μου», M.O. 4,40, SD = 0,55), γεγονός που επιβεβαιώνει ότι η αντιστοίχιση συναισθήματος-audio features βάσει του Circumplex Model of Affect (Russell, 1980) παράγει μουσικές προτάσεις αντιληπτά συναφείς με τη διάθεση του χρήστη. Ακολουθεί το ερώτημα 1 («Η εφαρμογή αναγνώρισε σωστά τη συναισθηματική μου κατάσταση», M.O. 4,20, SD = 0,45), το οποίο παρουσιάζει και τη μικρότερη τυπική απόκλιση, ένδειξη ομοφωνίας μεταξύ των συμμετεχόντων ως προς την αποτελεσματικότητα της αναγνώρισης συναισθήματος.

Αντίθετα, η χαμηλότερη βαθμολογία εντοπίζεται στο ερώτημα 4 («Οι προτάσεις ήταν πιο εξατομικευμένες από ό,τι θα επέλεγα χειροκίνητα», M.O. 3,60, SD = 0,55). Η τιμή αυτή, αν και παραμένει πάνω από τον μέσο της κλίμακας, υποδηλώνει ότι ορισμένοι χρήστες δεν αντιλήφθηκαν σαφώς την εξατομίκευση των προτάσεων έναντι μιας χειροκίνητης επιλογής, γεγονός που ενδέχεται να σχετίζεται με το σχετικά μικρό ιστορικό ακρόασης ορισμένων χρηστών στο Spotify, το οποίο περιορίζει την αποτελεσματικότητα του μηχανισμού top tracks seeds. Αντίθετα, υψηλή βαθμολογία συγκεντρώνει το ερώτημα 3 («Θα χρησιμοποιούσα αυτές τις μουσικές προτάσεις στην καθημερινή μου ζωή», M.O. 4,00), παρέχοντας θετικές ενδείξεις ως προς την πρόθεση μελλοντικής χρήσης της εφαρμογής.

Συνολικά, τα αποτελέσματα της αξιολόγησης χρηστών παρέχουν θετικές ενδείξεις τόσο για τη χρηστικότητα της εφαρμογής όσο και για την αντιληπτή ποιότητα των μουσικών προτάσεων, επιβεβαιώνοντας ότι το σύστημα εκπληρώνει τους στόχους χρηστικότητας που είχαν τεθεί στην ενότητα 5.1. Η επίτευξη μέσης βαθμολογίας SUS 85,50 τοποθετεί την εφαρμογή σε πολύ καλή κατηγορία χρηστικότητας, ενώ η συνολική ικανοποίηση από τις μουσικές προτάσεις (81,0%) τεκμηριώνει την αποτελεσματικότητα της συνδυαστικής προσέγγισης αναγνώρισης συναισθήματος και εξατομικευμένης σύστασης που υλοποιείται στην παρούσα εργασία (Bangor, Kortum and Miller, 2008) [33].

8.6 Περιορισμοί της εργασίας

Η αξιολόγηση του προτεινόμενου συστήματος υπόκειται σε ορισμένους εγγενείς περιορισμούς, οι οποίοι οφείλουν να λαμβάνονται υπόψη κατά την ερμηνεία και την αξιολόγηση των τελικών αποτελεσμάτων.

Όσον αφορά το μοντέλο αναγνώρισης συναισθήματος, το σύνολο δεδομένων FER-2013 παρουσιάζει γνωστά προβλήματα ποιότητας στην επισημείωση των εικόνων εξαιτίας της διαδικασίας συλλογικής ανάθεσης εργασιών σε ευρύ κοινό (crowdsourcing). Η διαδικασία αυτή έχει οδηγήσει σε ένα εκτιμώμενο ποσοστό λανθασμένων ετικετών της τάξης του 10–15% (Goodfellow et al., 2013) [30]. Επιπροσθέτως, το συγκεκριμένο σύνολο δεδομένων αποτελείται κυρίως από πρόσωπα δυτικής πολιτισμικής προέλευσης, γεγονός που περιορίζει τη δυνατότητα γενίκευσης των προβλέψεων του μοντέλου σε χρήστες διαφορετικών πολιτισμικών υποβάθρων (Li and Deng, 2022) [7]. Τέλος, η ταξινομική απόδοση στις στατικές εικόνες του FER-2013 δεν αντιστοιχεί απαραίτητα στην απόδοση του συστήματος σε πραγματικές συνθήκες χρήσης, όπου παρατηρείται σημαντικά μεγαλύτερη ποικιλομορφία στον φωτισμό, στις γωνίες λήψης της κάμερας και στην αυθόρμητη εκφραστικότητα των χρηστών.

Συγχρόνως, ως προς τη μελέτη χρηστών, το μέγεθος του δείγματος των 5 συμμετεχόντων θέτει περιορισμούς ως προς τη στατιστική ισχύ των αποτελεσμάτων και τη δυνατότητα γενίκευσής τους σε ευρύτερες πληθυσμιακές ομάδες, καθιστώντας τα ευρήματα της αξιολόγησης του δείκτη SUS κυρίως ενδεικτικά. Επιπλέον, η διεξαγωγή των δοκιμών σε ελεγχόμενο περιβάλλον ενδέχεται να μην αντικατοπτρίζει πλήρως την εμπειρία και τη συμπεριφορά χρήσης σε συνθήκες καθημερινής αλληλεπίδρασης. Τέλος, η αξιολόγηση της ποιότητας των μουσικών προτάσεων επηρεάζεται άμεσα από τις υποκειμενικές μουσικές προτιμήσεις και την ακουστική εμπειρία κάθε συμμετέχοντος, εισάγοντας σημαντική ατομική μεταβλητότητα στη διαμόρφωση των τελικών αποτελεσμάτων (Brooke, 1996) [31].

Κεφάλαιο 9ο: Συμπεράσματα και μελλοντική έρευνα

9.1 Συμπεράσματα

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε τον σχεδιασμό και την υλοποίηση κινητής εφαρμογής Android που συνδυάζει αυτόματη αναγνώριση συναισθημάτων από εκφράσεις προσώπου με εξατομικευμένη σύσταση μουσικής μέσω του Spotify Web API. Το σύστημα που αναπτύχθηκε ανταποκρίνεται στους στόχους που τέθηκαν στο κεφάλαιο 1.2 και συμβάλλει στο ευρύτερο πεδίο της συναισθηματικής υπολογιστικής (affective computing) με μια ολοκληρωμένη, on-device εφαρμογή πραγματικού χρόνου (Picard, 1997) [17].

Το εισαγωγικό κεντρικό συμπέρασμα αφορά την τεχνική εφικτότητα της αναγνώρισης συναισθήματος σε πραγματικό χρόνο εντός κινητής συσκευής Android, χωρίς εξάρτηση από εξωτερικούς διακομιστές. Η συνδυαστική αξιοποίηση του MediaPipe Face Landmarker για την ανίχνευση και την προεπεξεργασία του προσώπου και του προσαρμοσμένου MobileNetV2 σε μορφή TensorFlow Lite για την ταξινόμηση αποδεικνύει ότι σύγχρονα, ελαφριά νευρωνικά δίκτυα επιτρέπουν την εκτέλεση ολόκληρης της ροής επεξεργασίας τοπικά, εντός των ορίων απόκρισης που τέθηκαν στην ενότητα 5.1 (Sandler et al., 2018) [15]. Η επιλογή της στρατηγικής διφασικής λεπτομερούς προσαρμογής έναντι εκπαίδευσης από μηδενική βάση αποδεικνύεται τεχνικά ορθή και πρακτικά υλοποιήσιμη στο πλαίσιο διπλωματικής εργασίας, συνδυάζοντας αποδεκτή ακρίβεια με σημαντικά μειωμένο υπολογιστικό κόστος εκπαίδευσης (Chowdhury, Idris and Abas, 2024) [3].

Στο ίδιο πλαίσιο, το δεύτερο κεντρικό συμπέρασμα αφορά την αποτελεσματικότητα της συνδυαστικής προσέγγισης για τη σύσταση μουσικής. Η θεμελίωση της αντιστοίχισης συναισθημάτων σε audio features στο Circumplex Model of Affect (Russell, 1980) [32], παρέχει αξιόπιστη θεωρητική βάση για τον μηχανισμό σύστασης. Ωστόσο, η πορεία της υλοποίησης ανέδειξε και έναν ουσιαστικό πρακτικό περιορισμό: η απόσυρση των σημείων πρόσβασης GET /v1/recommendations και GET /audio-features για εφαρμογές σε στάδιο ανάπτυξης κατέστησε ανέφικτη την άμεση τροφοδότηση των ακουστικών χαρακτηριστικών και των seed tracks στη διαδικασία παραγωγής προτάσεων. Το σύστημα προσαρμόστηκε αναλόγως, αξιοποιώντας τον πίνακα αντιστοίχισης ως θεωρητικό πυρήνα για τη διαμόρφωση στοχευμένων ερωτημάτων αναζήτησης βάσει είδους και διάθεσης, ενώ το ακροαματικό προφίλ του χρήστη διατηρείται ως μηχανισμός καταγραφής. Το εύρημα αυτό αναδεικνύει ότι η εξάρτηση από εμπορικές διεπαφές τρίτων συνιστά ουσιαστικό παράγοντα κινδύνου για συστήματα σύστασης μουσικού περιεχομένου (Schedl et al., 2018) [29].

Σε τελικό επίπεδο, το τρίτο κεντρικό συμπέρασμα αφορά τη σχεδίαση της διεπαφής χρήστη. Η άμεση οπτική ανατροφοδότηση μέσω του FaceOverlayView, με δυναμική χρωματική κωδικοποίηση ανά συναίσθημα και ένδειξη του βαθμού εμπιστοσύνης της πρόβλεψης, ενισχύει τη διαφάνεια της λειτουργίας του συστήματος και εξοικειώνει τον χρήστη με την αυτόματη ανίχνευση. Η αρχιτεκτονική MVVM σε συνδυασμό με το LiveData εξασφαλίζει σταθερή συμπεριφορά της διεπαφής ανεξαρτήτως των μεταβολών του κύκλου ζωής της εφαρμογής.

Συνολικά, η παρούσα εργασία αποδεικνύει ότι ο συνδυασμός αυτόματης αναγνώρισης συναισθήματος και εξατομικευμένης μουσικής σύστασης σε κινητή πλατφόρμα αποτελεί τεχνικά εφικτό, θεωρητικά τεκμηριωμένο και πρακτικά αξιοποιήσιμο σύστημα, το οποίο εντάσσεται στην ευρύτερη ερευνητική κατεύθυνση της συναισθηματικής υπολογιστικής και της συναισθηματικά προσαρμοσμένης εξατομίκευσης (Picard, 1997) [17]. Η ανάπτυξη της εφαρμογής «Express Your Sound» τοποθετείται ουσιαστικά στο σημείο σύγκλισης της αυτόματης αναγνώρισης συναισθήματος,

της εξατομικευμένης μουσικής σύστασης και της ανάπτυξης εφαρμογών για κινητές συσκευές, τριών πεδίων που αποτελούν αντικείμενο ενδιαφέροντος εδώ και δεκαετίες τόσο στην ακαδημαϊκή κοινότητα όσο και στη βιομηχανία τεχνολογίας. Η μουσική, εξάλλου, παραμένει ένας από τους πιο άμεσους και ισχυρούς τρόπους επικοινωνίας και ρύθμισης του ανθρώπινου συναισθήματος (Juslin and Sloboda, 2010) [20], και ένα σύστημα ικανό να αντιλαμβάνεται τη διάθεση του χρήστη και να προτείνει αυτόματα μουσική που της ταιριάζει συνιστά ένα βήμα προς πιο φυσική και ανθρωποκεντρική αλληλεπίδραση ανθρώπου-υπολογιστή, αναδεικνύοντας παράλληλα τόσο τις πραγματικές δυνατότητες όσο και τους περιορισμούς που συνοδεύουν σήμερα την εφαρμοσμένη συναισθηματική υπολογιστική.

9.2 Συνεισφορά της εργασίας

Η βασική συνεισφορά της εργασίας έγκειται στον σχεδιασμό και την ολοκληρωμένη υλοποίηση ενός συστήματος αναγνώρισης συναισθήματος και προσαρμοσμένης μουσικής σύστασης σε περιβάλλον φορητών συσκευών. Αρχικά, σε θεωρητικό επίπεδο, τεκμηριώνει την εφαρμοσιμότητα του Circumplex Model of Affect ως θεωρητικής γέφυρας μεταξύ της κατηγορικής ταξινόμησης συναισθήματος κατά Ekman και της ποσοτικής αντιστοίχισης σε μουσικές παραμέτρους (Russell, 1980) [32]. Στο τεχνικό επίπεδο, η εργασία συνεισφέρει στην ανάπτυξη μιας αυτόνομης αρχιτεκτονικής επεξεργασίας (on-device pipeline) στο λειτουργικό σύστημα Android, η οποία συνδυάζει ελαφριά μοντέλα βαθιάς μάθησης (MobileNetV2, MediaPipe) για την εκτέλεση της αναγνώρισης συναισθήματος απευθείας στη συσκευή, εξασφαλίζοντας χαμηλή υπολογιστική επιβάρυνση και προστασία της ιδιωτικότητας των οπτικών δεδομένων του χρήστη. Τέλος, σε εφαρμοσμένο επίπεδο, η εργασία ολοκληρώνεται με την ανάπτυξη της λειτουργικής Android εφαρμογής «Express Your Sound» και την πιλοτική αξιολόγηση της χρησιμότητάς της μέσω δοκιμών σε πραγματικούς χρήστες (System Usability Scale), προσφέροντας ένα ολοκληρωμένο πρωτότυπο που μπορεί να αποτελέσει βάση για περαιτέρω ακαδημαϊκή έρευνα.

9.3 Μελλοντική έρευνα

Η παρούσα εργασία ανοίγει αρκετές ενδιαφέρουσες κατευθύνσεις για μελλοντική έρευνα και ανάπτυξη, οι οποίες αντιμετωπίζουν τους περιορισμούς που αναλύθηκαν στην ενότητα 8.6.

Στο επίκεντρο βρίσκεται η βελτίωση ακρίβειας αναγνώρισης συναισθήματος. Μελλοντικές εργασίες θα μπορούσαν να διερευνήσουν την αξιοποίηση αρχιτεκτονικών Vision Transformer (ViT) ή EfficientNet που επιτυγχάνουν υψηλότερη ακρίβεια στο FER-2013 (>85%), εφόσον βελτιστοποιηθούν για on-device εκτέλεση. Εναλλακτικά, η αξιοποίηση των 52 blendshapes του MediaPipe ως πρόσθετου χαρακτηριστικού εισόδου στον ταξινομητή (αντί αποκλειστικής χρήσης εικόνας) ενδέχεται να βελτιώσει την ακρίβεια ιδίως για λεπτές εκφράσεις χαμηλής έντασης (Li and Deng, 2022) [7].

Μια ουσιαστική ποιοτική αναβάθμιση θα μπορούσε να επιτευχθεί μέσω της πολυτροπικής αναγνώρισης συναισθήματος. Η ενσωμάτωση πρόσθετων πηγών πληροφορίας πέραν των εκφράσεων προσώπου, όπως η ανάλυση φωνής μέσω μικροφώνου ή τα φυσιολογικά σήματα μέσω αισθητήρων smartwatch, θα επέτρεπε πολυτροπική αναγνώριση συναισθήματος με υψηλότερη αξιοπιστία, ιδιαίτερα σε περιπτώσεις χαμηλής εκφραστικότητας προσώπου (Zhang et al., 2023) [10].

Ιδιαίτερη έμφαση ενδείκνυται να δοθεί στην δυναμική προσαρμογή βάσει ανατροφοδότησης. Η ενσωμάτωση μηχανισμού ανατροφοδότησης χρήστη, π.χ. αξιολόγηση κάθε πρότασης με thumbs up/down, θα επέτρεπε τη δυναμική προσαρμογή των παραμέτρων αντιστοίχισης συναισθήματος–audio features με βάση τις προτιμήσεις κάθε χρήστη, υλοποιώντας ένα σύστημα ενισχυτικής μάθησης που βελτιώνεται με την πάροδο του χρόνου χρήσης (Deldjoo et al., 2022) [13].

Συμπληρωματικά, προτείνεται η επέκταση με ανίχνευση συναισθήματος από κείμενο. Η προσθήκη δυνατότητας ανίχνευσης συναισθήματος από κείμενο εισόδου χρήστη, π.χ. μέσω sentiment analysis μικρής έκτασης κειμένου, θα επέτρεπε υβριδική σύσταση που συνδυάζει οπτική και γλωσσική συναισθηματική πληροφορία, διευρύνοντας τα σενάρια χρήσης σε περιπτώσεις όπου η κάμερα δεν είναι διαθέσιμη.

Η εγκυρότητα της παρούσας μεθοδολογίας προτείνεται να επεκταθεί μέσω της αξιολόγησης σε ευρύτερο πληθυσμό. Μελλοντική αξιολόγηση με μεγαλύτερο και πιο αντιπροσωπευτικό δείγμα χρηστών, συμπεριλαμβανομένης μεγαλύτερης ποικιλίας ηλικιών, πολιτισμικών υποβάθρων και τύπων κινητών συσκευών. Έτσι θα παρείχε εκτενέστερη και γενικεύσιμη αξιολόγηση της χρηστικότητας και της αντιληπτής αποτελεσματικότητας (Bangor, Kortum and Miller, 2008) [\[33\]](#).

Τέλος, σε θεσμικό και τεχνολογικό επίπεδο, κρίνεται σκόπιμη η προστασία ιδιωτικότητας με Federated Learning. Ιδιαίτερο ερευνητικό ενδιαφέρον παρουσιάζει η αξιοποίηση τεχνικών ομοσπονδιακής μάθησης (Federated Learning) για τη βελτίωση του μοντέλου αναγνώρισης συναισθήματος με δεδομένα πραγματικών χρηστών, χωρίς αποκέντρωση βιομετρικών δεδομένων από τη συσκευή, εξισορροπώντας έτσι τη συνεχή βελτίωση του μοντέλου με την προστασία της ιδιωτικότητας σύμφωνα με τις αρχές του GDPR (Picard, 1997) [\[17\]](#).

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] A. R. Khan, “Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges,” *Information*, vol. 13, no. 6, p. 268, 2022. Available: <https://doi.org/10.3390/info13060268>.
- [2] K. P. Sharma, T. Nagpal, and K. N. Raja Praveen, “Evaluating MobileNetV2 Architecture for Resource-Efficient Facial Emotion Recognition,” *National Academy Science Letters*, 2025. Available: <https://doi.org/10.1007/s40009-025-01671-w>.
- [3] R. R. Chowdhury, A. C. Idris, and P. E. Abas, “A Transfer Learning Approach for Facial Emotion Recognition Using a Deep Learning Model,” *International Journal of Research and Scientific Innovation*, vol. XI, no. IV, pp. 274–284, 2024. Available: <https://doi.org/10.51244/IJRSI.2024.1104022>.
- [4] J. K. Mahendran et al., “Image-based facial emotion recognition using convolutional neural network on emognition dataset,” *Scientific Reports*, vol. 14, art. no. 14765, 2024. Available: <https://doi.org/10.1038/s41598-024-65276-x>.
- [5] S. Kaur and K. Kulkarni, “FERFM: An Enhanced Facial Emotion Recognition System Using Fine-tuned MobileNetV2 Architecture,” *IETE Journal of Research*, vol. 70, no. 4, pp. 3723–3737, 2024. Available: <https://doi.org/10.1080/03772063.2023.2202158>.
- [6] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, “Deep learning-based facial emotion recognition for human–computer interaction applications,” *Neural Computing and Applications*, vol. 35, no. 32, pp. 23311–23328, 2023. Available: <https://doi.org/10.1007/s00521-021-06012-8>.
- [7] S. Li and Deng, W., “Deep Facial Expression Recognition: A Survey,” *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2022. Available: <https://doi.org/10.1109/TAFFC.2020.2981446>.
- [8] Z.-Y. Huang et al., “A study on computer vision for facial emotion recognition,” *Scientific Reports*, 2023. Available: <https://www.nature.com/articles/s41598-023-35446-4>.
- [9] K. Sarvakar and K. Rana, “Revolutionizing facial emotion recognition: in-depth analysis of cutting-edge models, methodologies, and datasets,” *Discover Artificial Intelligence*, Springer Nature, 2025. Available: <https://link.springer.com/article/10.1007/s44163-025-00553-w>.
- [10] S. Zhang et al., “A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face,” *Entropy*, vol. 25, no. 10, p. 1440, 2023. Available: <https://doi.org/10.3390/e25101440>.
- [11] P. Ekman, “An argument for basic emotions,” *Cognition & Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992. Available: <https://doi.org/10.1080/02699939208411068>.
- [12] M. V. Thoma et al., “The Effect of Music on the Human Stress Response,” *PLOS ONE*, 2013. Available: <https://doi.org/10.1371/journal.pone.0070156>.
- [13] Y. Deldjoo et al., “Induced Emotion-Based Music Recommendation through Reinforcement Learning,” *Applied Sciences*, vol. 12, no. 21, p. 11209, 2022. Available: <https://doi.org/10.3390/app122111209>.
- [14] M. Mollahosseini et al., “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing,” *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019.

- [15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, June 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.
- [16] X. Lowe-Brown et al., "Would You Tell Spotify How You're Feeling? Exploring Acceptability and Ethics of Emotion-Regulation Plugins for Music Streaming Apps," In Proc. ACM Conference, 2024. Available: <https://dl.acm.org/doi/10.1145/3726986.3726998>.
- [17] R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.
- [18] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York: Times Books, 2003.
- [19] P. Ekman and W. V. Friesen, *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.
- [20] P. N. Juslin and J. A. Sloboda, Eds., *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford, UK: Oxford University Press, 2010.
- [21] J. Nielsen, *Usability Engineering*. San Francisco, CA: Morgan Kaufmann, 1993.
- [22] Google AI Edge, "Face Landmarker - MediaPipe Solutions," 2023. Available: https://developers.google.com/mediapipe/solutions/vision/face_landmarker.
- [23] Google AI Edge, "MediaPipe Face Mesh – Technical Documentation," 2023. Available: https://github.com/google-ai-edge/mediapipe/blob/master/docs/solutions/face_mesh.md.
- [24] H. J. Chong, H. J. Kim, and B. Kim, "Scoping Review on the Use of Music for Emotion Regulation," *PMC / PubMed Central*, 2024. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11428991/>.
- [25] R. G. Raikar et al., "Survey on Facial Emotion Recognition using Deep Learning," *International Journal of Research in Applied Science & Engineering Technology (IJRASET)*, 2023. Available: <https://doi.org/10.22214/ijraset.2023.51433>.
- [26] B. K. Kiran et al., "Emotion-Based Music Recommendation System using VGG16-CNN Architecture," *IJRASET*, 2024. Available: <https://www.ijraset.com/research-paper/emotion-based-music-recommendation-system->.
- [27] S. Kumar, "Emotion-Based Music Recommendation System," *ResearchGate*, 2024. Available: <https://www.researchgate.net/publication/387652550>.
- [28] S. Tidke et al., "Mood-Based Song Recommendation System," *INRIA / HAL Open Science*, 2022. Available: <https://inria.hal.science/hal-03771292v1/document>.
- [29] M. Schedl et al., "Current Challenges and Visions in Music Recommender Systems Research," *International Journal of Multimedia Information Retrieval*, 2018. Available: <https://arxiv.org/pdf/1710.03208>.
- [30] I. Goodfellow et al., "Challenges in Representation Learning: A report on three machine learning contests (FER2013 dataset)," arXiv preprint arXiv:1307.0414, 2013.
- [31] J. Brooke, *SUS: A 'quick and dirty' usability scale*. London: Taylor & Francis, 1996. Available: <https://doi.org/10.1201/9781498710411-35>.

- [32] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.
- [33] A. Bangor, P. T. Kortum, and J. T. Miller, *An empirical evaluation of the System Usability Scale*. *International Journal of Human-Computer Interaction*, 24(6), pp. 574–594, 2008. Available: <https://www.tandfonline.com/doi/full/10.1080/10447310802205776>.
- [34] Google Developers, “Android Security Architecture & Application Sandboxing,” *Android Open-Source Project (AOSP) Security Documentation*, 2024. Available: <https://source.android.com/docs/security/app-sandbox>.

ΠΑΡΑΡΤΗΜΑ Α : ΦΩΤΟΓΡΑΦΙΚΟ ΥΛΙΚΟ ΤΟΥ USER STUDY

ΧΡΗΣΤΗΣ 1

ΧΡΗΣΤΗΣ 2

ΧΡΗΣΤΗΣ 3

ΧΡΗΣΤΗΣ 4

ΧΡΗΣΤΗΣ 5

ΧΑΡΑ



ΛΥΠΗ



ΘΥΜΟΣ



ΟΥΛΕΤΕΡΟ



ΑΝΔΙΑ



ΧΡΗΣΤΗΣ 1

ΧΡΗΣΤΗΣ 2

ΧΡΗΣΤΗΣ 3

ΧΡΗΣΤΗΣ 4

ΧΡΗΣΤΗΣ 5

ΕΚΠΛΗΞΗ



ΦΟΒΟΣ



ΠΑΡΑΡΤΗΜΑ Β : ΒΕΒΑΙΩΣΗ ΣΥΜΜΕΤΟΧΗΣ

ΛΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ ΕΠΙΤΡΟΠΗ ΗΘΙΚΗΣ ΚΑΙ ΔΕΟΝΤΟΛΟΓΙΑΣ ΤΗΣ ΕΡΕΥΝΑΣ

Έντυπο Συγκατάθεσης Κατόπιν Ενημέρωσης

Καλείστε να συμμετέχετε σε μια έρευνα που διεξάγεται από (ονοματεπώνυμο Ευριπίδης Τσιαρδακλής, ιδιότητα προπτυχιακός φοιτητής, Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων, Πολυτεχνική Σχολή, Διεθνές Πανεπιστήμιο της Ελλάδος). Πρέπει να είστε τουλάχιστον 18 χρονών για να συμμετέχετε στην έρευνα. Η συμμετοχή σας είναι εθελοντική. Μπορείτε να αφιερώσετε όσο χρόνο χρειάζεστε για να διαβάσετε το «Έντυπο Συγκατάθεσης Κατόπιν Ενημέρωσης». Μπορείτε επίσης να αποφασίσετε να το συζητήσετε με την οικογένειά ή τους φίλους σας. Αντίγραφο αυτού του εντύπου θα σας δοθεί.

1. Σκοπός της έρευνας

Σας ζητάμε να συμμετέχετε στην παρούσα έρευνα γιατί ο σκοπός της είναι η αξιολόγηση της χρηστικότητας και της αποτελεσματικότητας μιας εφαρμογής κινητού τηλεφώνου (Android) που αναγνωρίζει το συναίσθημα του χρήστη μέσω της κάμερας και προτείνει μουσική ανάλογα με αυτό, στο πλαίσιο εκπόνησης διπλωματικής εργασίας.

Η συμπλήρωση και επιστροφή του ερωτηματολογίου ή οι απαντήσεις στις ερωτήσεις της συνέντευξης αποτελούν συγκατάθεση συμμετοχής στην παρούσα έρευνα.

2. Ενδεχόμενοι κίνδυνοι

Δεν υπάρχουν προβλέψιμοι κίνδυνοι που προκύπτουν από τη συμμετοχή σας στην παρούσα έρευνα. Εάν αισθανθείτε δυσφορία κατά την απάντησή σας σε συγκεκριμένες ερωτήσεις, παρακαλούμε μη διστάσετε να ζητήσετε να παραιτηθούν.

3. Ενδεχόμενα οφέλη για τα άτομα και την κοινωνία

Η συμμετοχή σας συμβάλλει στην αξιολόγηση και βελτίωση μιας πραγματικής εφαρμογής συναισθηματικής υπολογιστικής, με πιθανό όφελος για την ευρύτερη έρευνα στον τομέα της εξατομικευμένης μουσικής σύστασης βάσει συναισθήματος.

4. Αποζημίωση για τη συμμετοχή

Δεν θα έχετε κάποιο άμεσο ή μελλοντικό σημαντικό οικονομικό όφελος από τη συμμετοχή σας στην παρούσα έρευνα.

5. Ενδεχόμενη σύγκρουση συμφερόντων

Κατά τη δήλωση των ερευνητών της έρευνας δεν υπάρχει σύγκρουση συμφερόντων από τη διεξαγωγή της παρούσας έρευνας.

6. Εμπιστευτικότητα

Οποιοδήποτε πληροφορίες αποκτηθούν σχετικά με την παρούσα έρευνα, οι οποίες θα μπορούσαν να σας ταυτοποιήσουν προσωπικά, θα παραμείνουν απόρρητες και θα αποκαλυφθούν μόνο με την άδειά σας ή όπως προβλέπεται από τον νόμο. Οι πληροφορίες εκείνες που σας ταυτοποιούν προσωπικά, θα διατηρηθούν ξεχωριστά από τα υπόλοιπα δεδομένα που σας αφορούν. Τα δεδομένα θα φυλάσσονται με ευθύνη του ερευνητή. Σε περίπτωση ηχογράφησης, θα ζητηθεί η συγκατάθεσή σας. Μπορείτε να αρνηθείτε να ηχογραφηθείτε. Ο ερευνητής θα μεταγράψει τις ηχογραφήσεις και μπορεί να σας προμηθεύσει με ένα αντίγραφο του απομαγνητοφωνημένου κειμένου κατόπιν αιτήσεώς σας. Έχετε το δικαίωμα να ελέγξετε και να επεξεργαστείτε την απομαγνητοφώνηση. Προτάσεις οι οποίες έχετε ζητήσει από τον ερευνητή να παραληφθούν δεν θα χρησιμοποιηθούν και θα σβηστούν από όλα τα αντίστοιχα αρχεία. Σε περίπτωση που τα αποτελέσματα της έρευνας δημοσιευτούν ή παρουσιαστούν σε συνέδρια δεν θα συμπεριληφθούν πληροφορίες που θα αποκαλύπτουν την ταυτότητά σας. Σε περίπτωση που φωτογραφίες σας, βίντεο ή ακουστικές ηχογραφήσεις χρησιμοποιηθούν για εκπαιδευτικούς σκοπούς, η ταυτότητά σας θα προστατεύεται ή θα συγκαλύπτεται, εκτός εάν συναινέσετε ρητά διαφορετικά παρακάτω.

Σύμφωνα με το άρθρο 5 στοιχ. α του Κανονισμού 2016/679 (GDPR), τα δεδομένα προσωπικού χαρακτήρα μπορούν να αποθηκεύονται για διάστημα μεγαλύτερο αυτού κατά το οποίο διεξάγεται η ερευνητική διαδικασία, εφόσον εφαρμόζονται τα κατάλληλα τεχνικά και οργανωτικά μέτρα που απαιτεί ο Κανονισμός 2016/679 (GDPR) για τη διασφάλιση των δικαιωμάτων και ελευθεριών του υποκειμένου (ή των υποκειμένων) των δεδομένων που έχουν συλλεχθεί στο πλαίσιο της έρευνας. Σε αυτήν τη βάση, τα δεδομένα της παρούσας έρευνας θα διατηρηθούν για τρία χρόνια μετά το πέρας αυτής.

Επιπρόσθετα, όσον αφορά στην ανάγκη περαιτέρω (δευτερογενούς) επεξεργασίας των προσωπικών δεδομένων, σύμφωνα με το άρθρο 5 στοιχ. β του Κανονισμού 2016/679 (GDPR), για λόγους ερευνητικούς, η δευτερογενής επεξεργασία των προσωπικών δεδομένων θεωρείται συμβατή με τους αρχικούς σκοπούς της επεξεργασίας χωρίς να γίνεται λόγος περί της ανάγκης λήψης τεχνικών και οργανωτικών μέτρων. Κατά συνέπεια, τα προσωπικά δεδομένα της παρούσας έρευνας δύνανται να χρησιμοποιηθούν και από άλλες έρευνες που θα εγκριθούν αρμοδίως χωρίς να χρειαστεί να δοθεί εκ νέου η συγκατάθεση των συμμετεχόντων.

7. Συμμετοχή και αποχώρηση

Μπορείτε να επιλέξετε να συμμετέχετε ή όχι στην παρούσα έρευνα. Αν συμμετέχετε εθελοντικά σε αυτή την έρευνα, μπορείτε να αποχωρήσετε οποιαδήποτε στιγμή χωρίς καμία συνέπεια. Μπορείτε επίσης να αρνηθείτε να απαντήσετε σε οποιοδήποτε ερωτήσεις δεν επιθυμείτε να απαντήσετε και να παραμείνετε στην έρευνα. Ο ερευνητής μπορεί να σας ζητήσει να αποσυρθείτε από την έρευνα, αν ανακύψουν περιστάσεις που το απαιτούν.

8. Δικαιώματα των συμμετεχόντων στην έρευνα

Μπορείτε να αποσύρετε τη συγκατάθεσή σας οποιαδήποτε στιγμή και να διακόψετε τη συμμετοχή σας χωρίς να υποστείτε καμία κύρωση.

9. Τρόποι καταγγελίας

Για οποιαδήποτε παράπονα ή καταγγελίες σχετικά με τη διεξαγωγή της έρευνας μπορείτε να προσφύγετε στην Επιτροπή Ηθικής και Δεοντολογίας του Διεθνούς Πανεπιστημίου της Ελλάδος στο email: ehde@ihu.gr

Για θέματα προσωπικών δεδομένων βάσει του Κανονισμού (ΕΕ) 2016/679 του Ευρωπαϊκού Κοινοβουλίου και του Συμβουλίου της 27ης Απριλίου 2016 για την προστασία των φυσικών προσώπων έναντι της επεξεργασίας των δεδομένων προσωπικού χαρακτήρα και για την ελεύθερη κυκλοφορία των δεδομένων αυτών και την κατάργηση της οδηγίας 95/46/ΕΚ, μπορείτε να επικοινωνείτε με τον Υπεύθυνο Προστασίας Δεδομένων (ΥΠΔ /DPO) στο email: dpo@ihu.gr.

Για οποιαδήποτε παράβαση αφορά την επεξεργασία των προσωπικών σας δεδομένων, έχετε το δικαίωμα υποβολής καταγγελίας στην Αρχή Προστασίας Δεδομένων Προσωπικού Χαρακτήρα, Κηφισίας 1-3, Τ.Κ. 115 23, Αθήνα, τηλ.: +30-210 6475600, e-mail: contact@dpa.gr.

10. Ταυτότητα των ερευνητών

(Αν έχετε οποιοδήποτε ερωτήσεις ή ανησυχίες σε σχέση με την έρευνα, μη διστάσετε να επικοινωνήσετε με τον ερευνητή.)

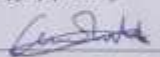
Ευριπίδης Τσιαρδακλής, προπτυχιακός φοιτητής, Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων, Διεθνές Πανεπιστήμιο της Ελλάδος, email: evripidistsiardaklis@gmail.com. Επιβλέπων Καθηγητής: Παναγιώτης Αδαμίδης, ίδιο Τμήμα.

☒ Η συμμετέχουσα διάβασα τα παραπάνω και αποδέχομαι τη συμμετοχή μου στην έρευνα.

Ονοματεπώνυμο Ευφία Μοκρίδου Υπογραφή 

Ημερομηνία 20/08/2026

Συναίνω επιπλέον στη συμπεριλήψη φωτογραφικού υλικού στο οποίο εμφανίζομαι αναγνωρίσιμη/-η στο Παράρτημα της διπλωματικής εργασίας.

Ονοματεπώνυμο Ευφία Μοκρίδου Υπογραφή 

Ημερομηνία 20/08/2026

Ο/Η συμμετέχων/ουσα διάβασα τα παραπάνω και αποδέχομαι τη συμμετοχή μου στην έρευνα.

Όνοματεπώνυμο Αλέξανδρα Μακρίδου Υπογραφή [Signature]
Ημερομηνία 20/08/2026

Συναίνω επιπλέον στη συμπερίληψη φωτογραφικού υλικού στο οποίο εμφανίζομαι αναγνωρίσιμα/ή στο Παράρτημα της διπλωματικής εργασίας.

Όνοματεπώνυμο Αλέξανδρα Μακρίδου Υπογραφή [Signature]
Ημερομηνία 20/08/2026

Ο/Η συμμετέχων/ουσα διάβασα τα παραπάνω και αποδέχομαι τη συμμετοχή μου στην έρευνα.

Όνοματεπώνυμο Γιώργος Χατζηγεωργίου Υπογραφή [Signature]
Ημερομηνία 20/08/2026

Συναίνω επιπλέον στη συμπερίληψη φωτογραφικού υλικού στο οποίο εμφανίζομαι αναγνωρίσιμα/ή στο Παράρτημα της διπλωματικής εργασίας.

Όνοματεπώνυμο Γιώργος Χατζηγεωργίου Υπογραφή [Signature]
Ημερομηνία 20/08/2026

Ο/Η συμμετέχων/ουσα διάβασα τα παραπάνω και αποδέχομαι τη συμμετοχή μου στην έρευνα.

Όνοματεπώνυμο Γιάννη Ρωμανίδου Υπογραφή [Signature]
Ημερομηνία 20/08/2026

Συναίνω επιπλέον στη συμπερίληψη φωτογραφικού υλικού στο οποίο εμφανίζομαι αναγνωρίσιμα/ή στο Παράρτημα της διπλωματικής εργασίας.

Όνοματεπώνυμο Γιάννη Ρωμανίδου Υπογραφή [Signature]
Ημερομηνία 20/08/2026

Ο/Η συμμετέχων/ονσα διάβασα τα παραπάνω και αποδέχομαι τη συμμετοχή μου στην έρευνα.

Ονοματεπώνυμο Βασίλης Δάτσης Υπογραφή 

Ημερομηνία 20/08/2026

Συναινώ επιπλέον στη συμπερίληψη φωτογραφικού υλικού στο οποίο εμφανίζομαι αναγνωρίσιμα/-η στο Παράρτημα της διπλωματικής εργασίας.

Ονοματεπώνυμο Βασίλης Δάτσης Υπογραφή 

Ημερομηνία 20/08/2026