Emotion Recognition Through Photographs

An Application of Various Neural Network models on the FER-2013 Dataset

Θεοχαρίδη Φαίδρα ics22069, Βραζάλης Κωνσταντίνος ics22115

Φεβρουάριος - Ιούνιος 2025

Περίληψη

Η αναγνώριση συναισθημάτων από τις εκφράσεις προσώπου παίζει ζωτικό ρόλο στην αλληλεπίδραση ανθρώπου-υπολογιστή. Αυτός ο τομέας της όρασης υπολογιστών έχει αναπτυχθεί ραγδαία την τελευταία δεκαετία, με πολλά μοντέλα να φτάνουν πολύ κοντά στο να πετύχουν 100% ακρίβεια σε συγκεκριμένα σύνολα δεδομένων. Ο στόχος αυτής της εργασίας είναι η συγκριτική αξιολόγηση με πιστή αναπαραγωγή τριών μοντέλων transfer learning και δύο custom μοντέλων από την βιβλιογραφία. Όλα τα μοντέλα εκπαιδεύτηκαν πάνω στο σύνολο δεδομένων FER-2013. Σύμφωνα με τα αποτελέσματα μας το μοντέλο VGG-16 + GAP είχε ακρίβεια 67% ενώ τα μοντέλα InceptionV3 και MobileNetV2 είχαν ακρίβεια 63,8% και 63.3% αντίστοιχα. Τέλος, τα μοντέλα SE-SResNet18 και DSLVM L2 πέτυχαν ακρίβεια, 62% και 51%. Παρόλο που τα αποτελέσματα της παρούσας μελέτης δεν έφτασαν, σε κάθε περίπτωση, τα αποτελέσματα που είχαν επιτύχει οι αρχικές έρευνες, εξακολουθούν να καταδεικνύουν ότι οι προτεινόμενες προσεγγίσεις είναι πολλά υποσχόμενες.

Περιεχόμενα

1 Εισαγωγή **6**

2 Σχετική Βιβλιογραφία **7**

3 Προτεινόμενη Μεθοδολογία **9**
3.1 VGG-16 + GAP 10
3.2 InceptionV3 10
3.3 MobileNetV2 12
3.4 SE-SResNet18 13
3.5 DSLVM L2 15

4 Πειραματικά Αποτελέσματα **16**
4.1 Περιγραφή dataset 16
4.2 Μετρικές 17
4.3 VGG-16 18
4.4 InceptionV3 20
4.5 MobileNetV2 21
4.6 SE-SResNet18 23
4.7 DSLVM L2 24

5 Συμπεράσματα **26**

Κατάλογος Πινάκων

1	Συγκριτική ανάλυση βιβλιογραφίας	8
2	Σύγκριση των αρχιτεκτωνικών VGG vs. VGG + GAP	11
3	Σύγκριση των αρχιτεκτονικών InceptionV3 vs. InceptionV3 + Custom Classifier	12
4	Η αρχιτεκτονική του μοντέλου MobileNetV2	13
5	Σύγκριση της αρχιτεκτονικής του ResNet18 και του προτεινόμενου SE-SResNet18 model	14
6	Η αρχιτεκτονική του μοντέλου DSLVM L2	16
7	Ακρίβεια και μέτρο F1 για κάθε fold του VGG-16.	19
8	Ακρίβεια και μέτρο F1 για κάθε fold του InceptionV3.	20
9	Ακρίβεια και μέτρο F1 για κάθε fold του MobileNetV2.	22
10	Ακρίβεια και μέτρο F1 για κάθε fold του SE-SResNet18.	24
11	Ακρίβεια και μέτρο F1 για κάθε fold του DSLVM L2.	26
12	Σύγκριση αποτελεσμάτων βιβλιογραφίας και παρούσας έρευνας	26

Κατάλογος Σχημάτων

1 Ροή εργασίας ανάλυσης συναισθημάτων βάσει εικόνας 9
2 Δείγματα εικόνων από το FER-2013 dataset για κάθε συναισθηματική κλάση 17
3 Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής VGG + GAP για το FER-2013 dataset στο 2ο fold 19
4 Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής InceptionV3 για το FER-2013 dataset στο 2ο fold 21
5 Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής MobileNetV2 για το FER-2013 dataset στο 2ο fold 22
6 Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής SESResNet18 για το FER-2013 dataset στο 1ο fold 24
7 Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής DSLVM L2 για το FER-2013 dataset στο 3ο fold 25

1 Εισαγωγή

Η κατανόηση των ανθρώπινων συναισθημάτων είναι μια διαδικασία που συναντάται στις ανθρώπινες κοινωνίες πολύ πριν από την εποχή της τεχνητής νοημοσύνης και των νευρωνικών δικτύων. Οι άνθρωποι εκδηλώνουν τα συναισθήματά τους με πολλούς τρόπους, κυρίως με τα λόγια και τις εκφράσεις του προσώπου τους. Στην παρούσα εργασία θα αναλάβουμε το έργο της ταξινόμησης των ανθρώπινων συναισθημάτων από φωτογραφίες με τη χρήση νευρωνικών δικτύων, ένα θέμα που έχει μελετηθεί ευρέως τα τελευταία χρόνια.

Η επιτυχής κατηγοριοποίηση των συναισθημάτων με τη χρήση δεδομένων όπως φωτογραφίες, βίντεο ή καταγραφές ομιλίας θα μπορούσε ενδεχομένως να φέρει μια νέα εποχή σε ανθρωπιστικές επιστήμες όπως η ψυχολογία, με τη δυνατότητα να δημιουργηθούν εργαλεία διάγνωσης και θεραπείας με έγκαιρη ανίχνευση συναισθηματικών διαταραχών. Αντίστοιχα εργαλεία θα μπορούσαν να ενσωματωθούν σε εκπαιδευτικά περιβάλλοντα για την καλύτερη προσαρμογή των μαθησιακών εμπειριών. Επιπλέον, θα μπορούσαν να συμβάλλουν στην αποκλιμάκωση των εντάσεων σε διπλωματικές ή οργανωτικές συγκρούσεις με την ακριβή ανάγνωση συναισθηματικών ενδείξεων και να βελτιώσουν γενικότερα την αλληλεπίδραση ανθρώπου-υπολογιστή.

Δεδομένου ότι, η αναγνώριση συναισθημάτων μέσω φωτογραφιών αποτελεί ένα αρκετά δημοφιλές θέμα τα τελευταία χρόνια, υπήρχαν αρκετοί διαθέσιμοι πόροι, από προηγούμενους διαγωνισμούς που προωθούσαν καλά καθιερωμένα σύνολα δεδομένων [1], μέχρι και ερευνητές που επινόησαν εξειδικευμένους αλγορίθμους [2]. Υπάρχουν πολλά διαθέσιμα datasets που ταιριάζουν με τον ορισμό της παρούσας εργασίας, όπως το JAFFE (Japanese Association of Female Facial Expression) [3], το οποίο διαθέτει 213 εικόνες διαφόρων εκφράσεων προσώπου που μπορούν να ταξινομηθούν σε 7 διαφορετικές συναισθηματικές κατηγορίες, οι οποίες προέρχονται από συνολικά 10 διαφορετικές φοιτήτριες. Αντίστοιχα datasets είναι τα CK [4] και CK+ (Cohn-Kanade) [5] που περιέχουν 593 βίντεο από 123 διαφορετικούς ανθρώπους διαφορετικών ηλικιών, φύλων και εθνικοτήτων. Για τους σκοπούς της παρούσας εργασίας, ωστόσο, επιλέξαμε να χρησιμοποιήσουμε το dataset FER-2013 [1][6], το οποίο, αντιστοίχως με τα προαναφερθέντα datasets, περιέχει περισσότερες από 30.000 φωτογραφίες διαφορετικών υποκειμένων που ταξινομούνται σε 7 διαφορετικές συναισθηματικές καταστάσεις. Επιλέξαμε το dataset αυτό λόγω του μεγέθους του, αλλά και επειδή, σε αντίθεση με άλλα σύνολα, χαρακτηρίζεται ως "in the wild", καθώς οι εικόνες που περιλαμβάνονται δεν έχουν ληφθεί σε εργαστήριο αλλά έχουν προέλθει από το διαδίκτυο.

2 Σχετική Βιβλιογραφία

Όπως αναφέρθηκε παραπάνω, τα τελευταία χρόνια έχει διεξαχθεί εκτεταμένη έρευνα στο θέμα της αναγνώρισης εκφράσεων προσώπου. Μια τέτοια σημαντική συνεισφορά είναι η [7], στην οποία εισήχθη ένα Attentional Convolution Network που εντοπίζει τις πιο σημαντικές περιοχές του προσώπου για κάθε συναίσθημα και τις χρησιμοποιεί για την αποτελεσματικότερη εκτέλεση του classification. Επετεύχθη ακρίβεια έως και 70,02% στο dataset FER-2013 χρησιμοποιώντας 10 convolutional layers. Εκτός από την πρόταση μιας αποτελεσματικής αρχιτεκτονικής, οι συγγραφείς χρησιμοποίησαν επίσης τεχνικές οπτικοποίησης για τον εντοπισμό των πιο σημαντικών περιοχών του προσώπου που συμβάλλουν στις αποφάσεις ταξινόμησης, προσφέροντας ερμηνευσιμότητα παράλληλα με την ακρίβεια.

Μια άλλη αξιοσημείωτη μελέτη είναι η [13] όπου προτείνεται μια αρχιτεκτονική για ένα μοντέλο lightweight CNN που ονομάζεται CLCM. Η προτεινόμενη αρχιτεκτονική έχει πανομοιότυπο πυρήνα με το MobileNetV2, αλλά παγώνει το πρώτο Bottleneck Residual Block. Με αυτόν τον τρόπο εστιάζει στην εκπαίδευση των βαθύτερων επιπέδων, διατηρώντας τα βάρη των προηγούμενων επιπέδων αμετάβλητα. Το CLCM επιτυγχάνει συνολική ακρίβεια 63% στα δεδομένα του FER-2013, ξεπερνώντας το MobileNetV2 και ακόμη και το ShuffleNetV2 για ορισμένες κατηγορίες συναισθημάτων. Αυτή η προσέγγιση δίνει έμφαση στην υπολογιστική αποδοτικότητα επιτυγχάνοντας παρόμοια και καλύτερα αποτελέσματα από αρχιτεκτονικές με πάνω από 3 εκατομμύρια παραμέτρους, με μόλις 2,3 εκατομμύρια.

Από τα πιο συνηθισμένα CNN architectures στην βιβλιογραφία είναι το VGG-16. Μάλιστα, στο [11] προτάθηκε μια παραλλαγή του VGG-16 αντικαθιστώντας απλά τα τελεταία fully connected επίπεδα με ένα Global Average Pooling (GAP) επίπεδο και επιτυγχάνοντας 69,40% accuracy, έχοντας καλύτερη επίδοση από πολλές άλλες αρχιτεκτονικές VGG.

Ένα άλλο μοντέλο που μελετάται σε έρευνες όπως η [12] και αξιοποιεί transfer learning για την ανάλυση συναισθημάτων, είναι το InceptionV3. Στη συγκεκριμένη έρευνα περιλαμβάνεται λεπτομερής ρύθμιση του προ-εκπαιδευμένου μοντέλου InceptionV3 στο σύνολο δεδομένων FER-2013, επιτυγχάνοντας ακρίβεια 73,09%. Η μελέτη αυτή υπογραμμίζει την αποτελεσματικότητα του transfer learning σε διεργασίες FER.

Μία από τις πιο υποσχόμενες προσεγγίσεις είναι η [14], όπου παρουσιάζεται ένας νέος αλγόριθμος FER, ο SE-SResNet18, ο οποίος βελτιστοποιεί την αρχιτεκτονική του Residual Network (ResNet) ενσωματώνοντας μονάδες Squeeze-

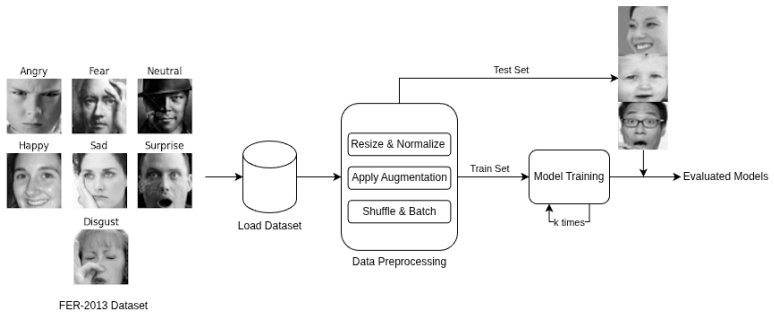
and-Excitation (SENet). Η προτεινόμενη μέθοδος επιτυγχάνει ακρίβειες αναγνώρισης 74.14% στο dataset FER2013 και 95.25% στο dataset CK+, ξεπερνώντας τις υπάρχουσες σύγχρονες τεχνικές και μειώνοντας σημαντικά τις παραμέτρους του μοντέλου. Τα ευρήματα δείχνουν ότι η απλή αύξηση του βάθους του δικτύου δεν βελτιώνει την ακρίβεια, υπογραμμίζοντας τη σημασία της αποδοτικότητας του μοντέλου για την ανάπτυξη σε Ενσωματωμένα Συστήματα.

Τέλος, αν και προ-δεκαετίας, είναι σημαντικό να αναφερθεί η [15], η οποία κατάφερε accuracy 71,20% νικώντας τον αρχικό διαγωνισμό FER-2013[1]. Η πρόταση του άρθρου ήταν μια καινοτόμα προσέγγιση, συνδυάζοντας CNN αρχιτεκτονικές με Support Linear Vector Machines (SVMs), ως αντικατάσταση για το επίπεδο softmax. Στον Πίνακα 1 μάλιστα γίνεται προφανές ότι, αν και παλιά, αυτή η προσέγγιση συνεχίζει να δίνει ένα από τα καλύτερα αποτελέσματα για το dataset FER-2013.

Σε αυτό το σημείο είναι σημαντικό να αναφέρουμε ότι σε διαφορετικά datasets έχει επιτευχθεί πολύ υψηλότερο accuracy, καθώς οι εικόνες έχουν ληφθεί σε εργαστηριακό περιβάλλον, γεγονός που τις καθιστά πολύ πιο καθαρές και συνεπάγεται συγκεκριμένη φυσική στάση των φωτογραφιζόμενων υποκειμένων.

No.	Author	Approach	Dataset	Accuracy(%)	Year
1	Minaee et al. [7]	ACN	FER-2013	70,02%	2021
			FERG	99,3%	
			JAFFE	92,8%	
			CK+	98%	
2	Zhong et al.[14]	SE-SResNet18	FER-2013	74,143%	2020
3	Jonathan & Lim[11]	VGG-16 + GAP		69,40%	2020
4	Meena et al.[12]	InceptionV3		73,09%	2023
5	Gursesli et al. [13]	CLCM		63%	2024
		MobileNetV2		58%	
		ShuffleNetV2		65%	
6	Tang[15]	DSLVM L2		71,2%	2013

Πίνακας 1: Συγκριτική ανάλυση βιβλιογραφίας



Σχήμα 1: Ροή εργασίας ανάλυσης συναισθημάτων βάσει εικόνας

3 Προτεινόμενη Μεθοδολογία

Σε αυτή την ενότητα θα παρουσιαστεί μια λεπτομερής επεξήγηση της μεθοδολογίας που ακολουθήσαμε. Θα δοθεί ο ορισμός τους προβλήματος που αντιμετωπίζουμε καθώς και περιγραφές των μοντέλων που χρησιμοποιήθηκαν .

Η εργασία μας επικεντρώνεται στην αναγνώριση συναισθημάτων μέσω φωτογραφιών από πρόσωπα. Στόχος μας ήταν η πιστή αναπαραγωγή συγκεκριμένων μοντέλων της βιβλιογραφίας και η σύγκριση των αποτελεσμάτων της εκπαίδευσης με τα αντίστοιχα των ερευνητών. Το πρόβλημα που αντιμετωπίζουμε πρόκειται για classification problem, καθώς οι αλγόριθμοι που χρησιμοποιούνται δέχονται ως πληροφορία εικόνες που χαρακτηρίζονται από μια "ετικέτα", το output αντιθέτως κάνει label τις φωτογραφίες του test set με μια από τις εφτά "ετικέτες". Οι ετικέτες αυτές αφορούν τις κλάσεις συναισθημάτων που περιγράφονται από το δατασετ. Στο Σχήμα 1 παρατηρείται η ροή της εργασίας που ακολουθήσαμε κατά την ανάπτυξη των μοντέλων.

Για το classification αποφασίσαμε να ακολουθήσουμε μια τεχνική transfer learning. Το transfer learning είναι μια τεχνική μάθησης η οποία περιλαμβάνει τη χρήση ενός προ-εκπαιδευμένου δικτύου (που έχει εκπαιδευτεί σε ένα μεγάλο σύνολο δεδομένων) ως σημείο εκκίνησης για ένα νέο έργο. Δύο συνηθισμένες προσεγγίσεις του transfer learning είναι το feature-extraction και το fine-tuning. Επιλέξαμε τη μέθοδο fine-tuning, δηλαδή ξεπαγώσαμε όλα τα επίπεδα του προ-εκπαιδευμένου δικτύου και τα εκπαιδεύσαμε από κοινού με το νέο classifier. [10]

Οι εικόνες που χρησιμοποιήθηκαν ήταν αρχικά μεγέθους 48x48, παρ΄ όλ΄ αυτά αποφασίσαμε να χρησιμοποιήσουμε εικόνες μεγέθους 96x96. Αυτό έγινε δίοτι τα μοντέλα transfer learning που χρησιμοποιήθηκαν είναι όλα προεκπεδευ-

μένα σε εικόνες μεγαλύτερου μεγέθους (κυρίως 224x224), λόγω όμως υπολογιστικών περιορισμών, η εκμάθηση δεν μπορούσε να γίνει με εικόνες μεγαλύτερες του 96x96 χωρίς να τερματήσει η εκτέλεση του μοντέλου πρόωρα καθώς και οποιαδήποτε μικρότερη τιμή έδινε λιγότερο βέλτιστα αποτελέσματα.

Τέλος, αποφασίσαμε να εκτελέσουμε ένα ενιαίο pre-processing pipeline για όλα τα μοντέλα transfer learning, με μόνο μικρές αλλαγές μεταξύ τους. Το preprocessing έγινε σε δύο φάσεις: αρχικά οι εικόνες κανονικοποιήθηκαν ώστε η τιμή κάθε pixel να κυμαίνεται στο [0,1], οι φωτογραφίες μετατράπηκαν σε RGB και η διάστασή τους άλλαξε στην επιλεγμένη. Στη δεύτερη φάση εφαρμόστηκε data augmentation, πραγματοποιήθηκε δηλαδή προσθήκη ελαφρώς τροποποιημένων αντιγράφων των δεδομένων για τη μείωση του overfit. Πιο συγκεκριμένα προστέθηκε τυχαίος αντικατοπτρισμός της εικόνας οριζόντια. Με αυτόν τον τρόπο μπορούμε, χωρίς να αλλοιώσουμε την έκφραση προσώπου, να γίνει εκμάθηση του μοντέλου σε εκφράσεις, ανεξαρτήτως της κατεύθυνσης. Έγινε περιστροφή των εικόνων τυχαία κατά $\pm 10\%$ του κύκλου (δηλαδή περίπου $\pm 36^{\circ}$), έτσι το μοντέλο δεν θα εξαρτάται υπερβολικά από τον προσανατολισμό του προσωπου. Τέλος, εφαρμόστηκαν αλλαγές στην αντίθεση και το ζουμ των εικόνων.

3.1 VGG-16 + GAP

Η κλασική αρχιτεκτονική του VGG-16 αποτελείται από 13 convolution layers με 3 πρόσθετα fully connected layers στο τέλος [9]. Στην [11] προτείνεται μια παραλαγμένη αρχιτεκτονική, όπου το τελευταίο επίπεδο Max Pooling αντικαθίσταται με ένα Global Average Pooling επίπεδο και το classifier με ένα Dense/Fully Connected επίπεδο αντί για 3 Fully Connected. Τα βάρη που χρησιμοποιήθηκαν είναι προεκπαιδευμένα στο ImageNet dataset. Οι ερευνητές πραγματοποίησαν πολλαπλά πειράματα, όμως κατέληξαν ότι η καλύτερη απόδοση προέκυπτε χρησιμοποιώντας SGD optimizer και early stopping κατά τη διαδικασία εκμάθησης. Τα hyperparameters που χρησιμοποιούνται σύμφωνα με το άρθρο είναι το learning rate value το οποίο είναι ίσο με 0.001, 0.9 για το momentum του SGD optimizer, και batch size 32. Στον Πίνακα 2 φαίνεται η διαφορά μεταξύ της κλασικής αρχιτεκτονικής και της προτεινόμενης.

3.2 InceptionV3

Στο [12] οι ερευνητές χρησιμοποιούν το μοντέλο InceptionV3 με παραλλαγές μόνο στο επίπεδο classifier, όπως παρατηρείται στον Πίνακα 3. Αυτό έγινε για

Type	VGG Layers	VGG + GAP Layers
Feature Extractor	Convolution	Convolution
	Convolution	Convolution
	Max Pooling	Max Pooling
	Convolution	Convolution
	Convolution	Convolution
	Max Pooling	Max Pooling
	Convolution	Convolution
	Convolution	Convolution
	Convolution	Convolution
	Max Pooling	Max Pooling
	Convolution	Convolution
	Convolution	Convolution
	Convolution	Convolution
	Max Pooling	Max Pooling
	Convolution	Convolution
	Convolution	Convolution
	Convolution	Convolution
	Max Pooling	Global Average Pooling
Classifier	Dense (4096)	Dense (7)
	Dense (4096)	
	Dense (1000)	
	Softmax	

Πίνακας 2: Σύγκριση των αρχιτεκτωνικών VGG vs. VGG + GAP

την προσαρμογή του αλγορίθμου στα συγκεκριμένα δεδομένα τους αντί στα δεδομένα στα οποία προεκπαιδεύτηκε ο αλγόριθμος. Τα επίπεδα που προστίθενται είναι ένα Flatten layer, έπειτα ένα Dropout layer και τέλος ένα Dense layer

με συνάρτηση softmax. Τα βάρη που χρησιμοποιήθηκαν είναι προεκπαιδευμένα στο ImageNet dataset.

Type	**InceptionV3**	**InceptionV3 + Custom Classifier**
Feature Extractor	Input	Input
	Conv + MaxPool	Conv + MaxPool
	Conv	Conv
	Conv	Conv
	Inception Module A ×3	Inception Module A ×3
	Inception Module B ×5	Inception Module B ×5
	Inception Module C ×2	Inception Module C ×2
Classifier	Global Average Pooling	Global Average Pooling
	Dense (1000)	Dense (1024)
	Softmax	ReLU
		Dense (7)
		Softmax

Πίνακας 3: Σύγκριση των αρχιτεκτονικών InceptionV3 vs. InceptionV3 + Custom Classifier

Πρέπει να σημειωθεί ότι οι ερευνητές μετατρέπουν το dataset κατηγοριοποιόντας τις κλάσεις συναισθημάτων σε 3 τύπους, Θετικά, Αρνητικά ή Ουδέτερα συναισθήματα (positive, negative and neutral sentiments). Επιλέξαμε να μην αφομοιώσουμε αυτή τη πρακτική υπέρ την επίτευξη συγκρίσημων αποτλεσμάτων με τα υπόλοιπα μοντέλα.

3.3 MobileNetV2

Η έρευνα που περιγράφεται στο [13] επικεντρώνεται στη δημιουργία Custom Lightweight CNN-based Model (CLCM) βασισμένο στο MobileNetV2, ωστόσο εμείς επιλέξαμε να επικεντρωθούμε στην υλοποίηση του απλού MobileNetV2. Όπως και στο άρθρο η αρχιτεκτονική που χρησιμοποιείται είναι η βασική αρχιτεκτονική του συγκεκριμένου μοντέλου, όπως φαίνεται και στον Πίνακα 4. Δηλαδή, το μοντέλο αποτελείται αρχικά από ένα Convolutional

Layer, έπειτα από 17 Bottleneck Blocks, ένα ακόμα Convolutional Layer, ένα GAP Layer και ένα Classification Layer με softmax.

Type	**MobileNetV2 Layers**
Feature Extractor	Conv2D (3×3)
	Bottleneck Block x1
	Bottleneck Block x2
	Bottleneck Block x3
	Bottleneck Block x4
	Bottleneck Block x3
	Bottleneck Block x3
	Bottleneck Block x1
	Conv2D (1×1)
	Global Average Pooling
Classifier	Dense (1000) Softmax

Πίνακας 4: Η αρχιτεκτονική του μοντέλου MobileNetV2

Η κανονικοποίηση που πραγματοποιήθηκε στο άρθρο ήταν μεταξυ του [-1,1] και οι εικόνες ήταν μεγέθους 224 και grayscale. Το pre-processing που πραγματοποιήσαν ακολουθεί την ίδια λογική με το pipeline που δημιουργήσαμε. Τα ηψπερπαραμετερς, που αναφέρονται είναι τα εξής: batch size 64,learning rate 0.0001, και Adam optimizer.

3.4 SE-SResNet18

Μια βελτιστοποιημένη αρχιτεκτονική, εξειδικευμένη σε προβλήματα FER, είναι αυτή που προτείνεται στο [14] ονομάζοντας το τελικό μοντέλο SE-SResNet18 στην οποία περίπτωση απομακρυνόμαστε από τα προεκπαιδευμένα μοντέλα και το transfer learning. Το μοντέλο που χρησιμοποιείται ως βάση είναι το ResNet18, από το οποίο αφαίρεσαν το πρώτο Max Pooling Layer, προκειμένου να είναι εφικτή η επεξεργασία μικρών εικόνων 48x48 χωρίς απώλεια των λεπτομερειών του προσώπου. Επίσης, μειώθηκε το μέγεθος του πρώτου convolutional layer kernel από 7x7 με stride=2 σε 3x3 με stride=1, αλλαγή απαραίτη-

τη για εικόνες grayscale χαμηλής αναλυσης. Ακόμη, προστέθηκε ένα dropout layer μετά το επίπεδο Global Average Pooling για την μείωση του overfitting ενώ, τέλος, αντικαταστάθηκε το τελευταίο Fully Connected Layer με ένα FC layer 512 διαστάσεων, ακολουθούμενο από ένα ακόμη FC layer για το τελικό classification.

Type	**ResNet18 Layers**	**SE-SResNet18 Layers**
Feature Extractor	Conv2D (7×7), stride=2	Conv2D (3×3), stride=1
	MaxPooling, stride=2	-
	Res. Block×2 (Conv2_x)	Res. Block×2+SE Block
	Res. Block×2 (Conv3_x)	Res. Block×2+SE Block
	Res. Block×2 (Conv4_x)	Res. Block×2+SE Block
	Res. Block×2 (Conv5_x)	Res. Block×2+SE Block
	Average Pooling	Average Pooling+Dropout
Classifier	FC (1000-d)+Softmax	FC (512-d)+Dropout+Softmax

Πίνακας 5: Σύγκριση της αρχιτεκτονικής του ResNet18 και του προτεινόμενου SE-SResNet18 model

Το επόμενο βήμα ήταν να προστεθεί ένα block SENet, ένα συστατικό που βοηθάει τα convolutional networks να εστιάζουν στα πιο σημαντικά κανάλια χαρακτηριστικών μιας εικόνας. Αυτό γίνεται χρησιμοποιώντας αρχικά την τεχνική Squeeze, δηλαδή την συμπίεση της σημαντικότητας κάθε καναλιού (μέσω Global Average Pooling) σε ένα διάνυσμα μήκους ίσο με το πλήθος των καναλιών. Στη συνέχεια ακολουθεί το Excitation όπου, μέσω δύο μικρών Fully Connected Layers, επιστρέφεται ένα set βαρών, που αντιπροσωπεύει την σημαντικότητα κάθε καναλιού. Τα βάρη αυτά πολλαπλασιάζονται με το ανάλογό τους κανάλι και κρατώνται τα κανάλια που έχουν τιμές βαρών πιο κοντά στο 1, ενώ αυτά που τείνουν στο 0 συμπιέζονται. Η προσθήκη του SENet block οδηγεί στη δημιουργία του τελικού προτεινόμενου μοντέλου. Οι διαφορές της αρχικής αρχιτεκτονικής του ResNet18 και της πρότασης των ερευνητών φαίνεται στον Πίνακα 5.

3.5 DSLVM L2

Μια διαφορετική προσέγγιση από της προηγούμενες είναι αυτη του DSLVM L2, όπου κατασκευάζεται ένα custom μοντέλο, βασισμένο στην υλοποίηση του [15].

Το μοντέλο που προτείνεται χρησιμοποιεί, αντιθέτως με τα προηγούμενα, Support Vector Machines στο τελευταίο επίπεδο, αντί για softmax. Επιπλέον για Loss function επιλέγει το L2 - Quadratic Loss καθώς είναι διαφορίσιμη και επιτρέπει την αποτελεσματική χρήση της μεθόδου της στοχαστικής καθοδικής κλίσης (Stochastic Gradient Descent) για την εκπαίδευση. Η επιλογή αυτή προσφέρει μεγαλύτερη σταθερότητα κατά την εκπαίδευση, αποτρέπει την υπερβολική τιμωρία σε περιπτώσεις ακραίων λαθών και βοηθά στη βελτίωση της γενίκευσης του μοντέλου, ειδικά σε σύνολα δεδομένων μικρού μεγέθους.

Η διαδικασία που ακολουθεί ο Tang περιλαμβάνει αρχικά την προεπεξεργασία των εικόνων. Πιο συγκεκριμένα, από κάθε εικόνα αφαιρείται ο μέσος όρος των τιμών των εικονοστοιχείων της, ώστε να κεντραριστεί γύρω από το μηδέν. Στη συνέχεια, η εικόνα κανονικοποιείται ώστε ο συνολικός της διανυσματικός κανόνας να έχει μέτρο 100. Επιπλέον, εφαρμόζεται και στατιστική κανονικοποίηση ανά pixel, αφαιρώντας τον μέσο όρο κάθε pixel υπολογισμένο σε όλο το σύνολο εκπαίδευσης και διαιρώντας με την αντίστοιχη τυπική απόκλιση. Αυτές οι τεχνικές κανονικοποίησης στοχεύουν στη μείωση της επίδρασης φωτεινότητας και αντίθεσης, ενώ ενισχύουν τη σταθερότητα της εκπαίδευσης και τη γενίκευση του μοντέλου.

Ένα από τα πιο σημαντικά χαρακτηριστικά της προσέγγισης του Tang είναι ότι το μοντέλο δεν περιορίζεται στη χρήση των SVMs ως εξωτερικών ταξινομητών μετά την εξαγωγή χαρακτηριστικών, αλλά ενσωματώνει τα SVMs απευθείας μέσα στο νευρωνικό δίκτυο. Συγκεκριμένα, η συνάρτηση απώλειας του L2-SVM χρησιμοποιείται στο τελευταίο επίπεδο και η παράγωγός της μπορεί να υπολογιστεί κανονικά. Αυτό επιτρέπει τη χρήση του backpropagation, δηλαδή της μεθόδου διάδοσης σφαλμάτων προς τα πίσω, για την εκπαίδευση ολόκληρου του μοντέλου από άκρη σε άκρη.

Με άλλα λόγια, η απώλεια που προκύπτει από το επίπεδο SVM μεταφέρεται μέσω παραγώγων σε όλα τα προηγούμενα επίπεδα του δικτύου. Έτσι, το δίκτυο μπορεί να μάθει, όχι μόνο πώς να ταξινομεί καλύτερα στο τελικό επίπεδο, αλλά και πώς να εξάγει πιο διακριτά και κατάλληλα χαρακτηριστικά στα προηγούμενα επίπεδα (συνελικτικά και πλήρως συνδεδεμένα). Η συνολική εκπαίδευση γίνεται με τη χρήση Stochastic Gradient Descent with Momentum, το οποίο βελτιώνει τη σύγκλιση και αποφεύγει τοπικά ελάχιστα. Στον Πίνακα 6 φαίνεται η

αρχιτεκτονική που προτείνεται.

Type	DSLVM L2 Layers
Feature Extractor	Conv2D (5×5, 32 filters, ReLU)
	MaxPooling (2×2)
	Conv2D (4×4, 32 filters, ReLU)
	MaxPooling (2×2)
	Conv2D (5×5, 64 filters, ReLU)
	MaxPooling (2×2)
Classifier	Dense (3072, ReLU)
	Dropout (rate = 0.5)
	Dense (7, Linear)
	L2-SVM Loss (Squared Hinge Loss)

Πίνακας 6: Η αρχιτεκτονική του μοντέλου DSLVM L2

4 Πειραματικά Αποτελέσματα

Σε αυτή την ενότητα θα παρέχουμε λεπτομερειακή ανάλυση των πειραμάτων που πραγματοποιήθηκαν χρησιμοποιώντας τα προαναφερθέντα μοντέλα και των αποτελεσμάτων που προέκυψαν. Θα ξεκινήσουμε με την επισκόπηση του dataset που χρησιμοποιήθηκε και έπειτα θα παρουσιαστούν οι αποδόσεις των διαφορετικών μοντέλων. Τέλος, θα συγκριθούν τα αποτελέσματα μας με αυτά της βιβλιογραφίας βάση επιλεγμένων μετρικών. Όλα τα πειράματα εκτελέστηκαν μέσω της πλατφόρμας Google Colab με τη χρήση της δοθέντας GPU και RAM.

4.1 Περιγραφή dataset

Όπως αναφέρθηκε προηγουμένως, το dataset που χρησιμοποιήθηκε είναι το Facial Expression Recognition 2013 (FER-2013) [6] και πρόκειται για μια συλλογή 32.298 φωτογραφιών που αντλήθηκαν χρησιμοποιώντας το API αναζήτησης εικόνων της Google για την εύρεση φωτογραφιών που αντιστοιχούν σε 7 διαφορετικές καταστάσεις συναισθημάτων: Θυμός (Anger), Αηδία (Disgust),

Σχήμα 2: Δείγματα εικόνων από το FER-2013 dataset για κάθε συναισθηματική κλάση

Φόβος (Fear), Ευτυχία (Happiness), Θλίψη (Sadness), Έκπληξη (Surprise) και Ουδέτερη κατάσταση (Neutral) [1]. Τα δεδομένα χαρακτηρίζονται ως "in the wild" φωτογραφίες, καθώς οι εικόνες δεν είναι στημένες και, σε σύγκριση με άλλα datasets, προσφέρουν μεγαλύτερη ποικιλία σε εικόνες συμπεριλαμβανομένων εικόνων με ημικάλυψη του προσώπου με αντικείμενα όπως γυαλιά ή χέρια και εικόνων με χαμηλή αντίθεση (low-contrast). Οι εικόνες που περιλαμβάνονται στο dataset είναι σε ανάλυση 48x48 και ασπρόμαυρες. Το dataset είναι ήδη χωρισμένο σε test και train sets, με 3.589, 28.709 εικόνες αντίστοιχα. Ενώ το FER-2013 είναι αρκετά μεγάλο, υπάρχει μια αρκετά μεγάλη ανισορροπία μεταξύ των κλάσεων, καθώς ορισμένες κλάσεις, όπως η Ευτυχία, περιέχουν σχεδόν 9.000 εικόνες, ενώ άλλες όπως η Αηδία, περιέχουν μόλις πάνω από 500. Στο Σχήμα 2 παρατηρούνται δείγματα από κάθε κλάση του FER-2013.

4.2 Μετρικές

Οι μετρικές που επιλέχθηκαν ήταν η ακρίβεια (accuracy) και ο δείκτης F1 (F1-score), καθώς παρέχουν μια σφαρική εικόνα της απόδοσης της κάθε μεθόδου. Η ακρίβεια *(1)* είναι απλή και ευρέως χρησιμοποιούμενη, καθώς επίσης υπερισχυέι ως μετρική στην επιλεγμένη βιβλιογραφία. Υπολογίζει το ποσοστό των σωστών προβλέψεων επί του συνολικού αριθμού προβλέψεων, το οποίο αν και σημαντικό μέτρο σύγκρισης μεταξύ αλγορίθμων, μπορεί να είναι παραπλανητικό σε καταστάσεις όπως τη δικιά μας όπου υπάρχει μεγάλη ανισορροπία κλάσεων, διότι μπορεί να δοθεί η εντύπωση υψηλής απόδοσης ακόμη και όταν το μοντέλο αγνοεί τις λιγότερο συχνές κατηγορίες. Από την άλλη, ο δείκτης F1 *(2)* αποτελεί το αρμονικό μέσο της ευκρίνιας (precision) και της ανάκλησης (recall), παρέχοντας έτσι, μια πιο αντιπροσωπευτική εικόνα της απόδοσης του μοντέλου στις επιμέρους κατηγορίες [8]. Με την χρήση των δύο μετρικών επιτυγχάνουμε και την συγκρισημότητα των αποτελεσμάτων μας σε σχέση με αυτά της βιβλιογραφίας αλλά και την σφαιρική αξιολόγηση της υλοποίησης των μοντέλων.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2}$$

4.3 VGG-16

Σύμφωνα με το [11], οι ερευνητές, διεξήγαγαν μια σειρά πειραμάτων βασισμένων στο προεκπαιδευμένο μοντέλο VGG-16, με σκοπό την επίτευξη της καλύτερης δυνατής ακρίβειας στο σύνολο δεδομένων FER-2013. Από τα διάφορα πειράματα που παρουσιάζονται στη μελέτη τους, επιλέξαμε να αναπαράγουμε εκείνο που παρουσίασε την υψηλότερη τελική ακρίβεια.

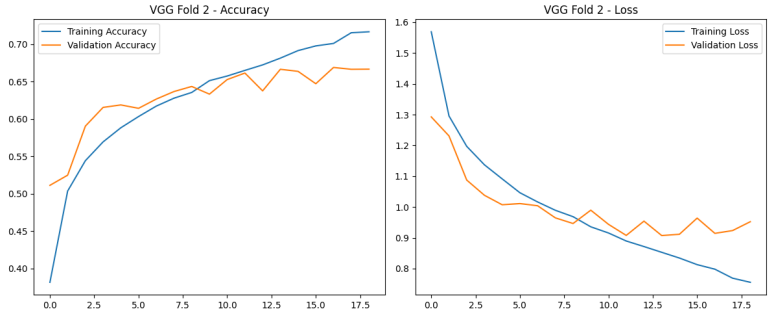
Για την αναπαραγωγή των αποτελεσμάτων του άρθρου ακολουθήσαμε πιστά την αρχιτεκτονική που προτείνεται. Έτσι, αντικαταστήσαμε το τελευταίο επίπεδο με ένα Global Average Pooling επίπεδο ως τελικό pooling layer, και αφαιρέσαμε τα 3 Fully Connected επίπεδα, για ένα με είσοδο 512 και έξοδο 7 νευρώνες. Τέλος, αξιοποιήσαμε τα προαναφερθέντα hyperparameters και χρησιμοποιήσαμε SGD optimizer.

Όπως και στη μελέτη, δεν εφαρμόστηκε καμία τεχνική εξισορρόπησης κλάσεων ή batch normalization. Επιπλέον, δεν παγώσαμε τα βάρη του προεκπαιδευμένου δικτύου και χρησιμοποιήσαμε early stopping με ανοχή 5 εποχών, ακολουθώντας πιστά τη μεθοδολογία της αρχικής έρευνας.

Κάποιες λεπτομέρειες που δεν αναφέρονται στο άρθρο είναι το μέγεθος των εικόνων που χρησιμοποιήθηκαν, αν πραγματοποιήθηκε προσάυξηση δεδομένων και γενικότερα τι προεργασία έγινε πάνω στα δεδομένα. Ωστόσο, κρίναμε απαραίτητη την εφαρμογή τεχνικώνaugmentation για την καλύτερη απόδοση του μοντέλου και την αποφυγή του overfitting.

Για να αξιολογήσουμε την απόδοση των πειραμάτων μας χρησιμοποιήσαμε 5-fold cross-validation. Για την εκπαίδευση με transfer learning χρησιμοποιήσαμε τα προεκπαιδευμένα VGG-16 ImageNet βάρη. Κάθε fold εκπαιδεύτικε για 50 εποχές με early stopping. Ο optimizer που χρησιμοποιήθηκε ήταν ο SGD με learning rate 0.001 και momentum 0.9.

Κατά τη διάρκεια των 5 folds το μοντέλο διατήρησε μια σταθερή και ανοδική πορεία όσον αφορά το accuracy καταλήγοντας με 67% μέγιστη ακρίβεια και 66% μέση ακρίβεια σε όλα τα folds. Στο Σχήμα 3 παρατηρείται η πορεία του accuracy και του loss στο 2ο fold. Όπως φαίνεται στο αριστερό διάγραμμα η ακρίβεια του validation ξεκινάει πιο ψηλά από αυτή του train, όμως μετά την 7η εποχή, ενώ



Σχήμα 3: Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής VGG + GAP για το FER-2013 dataset στο 2ο fold

η πορεία του train accuracy είναι ομαλή και ανοδική το validation accuracy, αν και συνεχίζει ανοδικά, παρουσιάζει κάποιες αστάθειες. Το 2ο fold επιλέχθηκε συμβολικά. Η απώλεια παρουσιάζει μια παρόμοια συμεριφορά καθώς φαίνεται να πέφτει σε πολλαπλές επαναλήψεις κάτω από το 1.0 χωρίς όμως να μπορεί να διατηρηθεί σε αυτό το σημείο.

Τα αντίστοιχα νούμερα για το μέτρο f1 σε κάθε fold ακολουθούν την ίδια τροχιά με αυτή της ακρίβειας καθώς, όπως παρατηρείται από τον Πίνακα 7 έχουν πανομοιότυπες τιμές σε κάθε fold. Αυτό δείχνει ότι το μοντέλο καταφέρνει να διατηρήσει μια, αν και όχι ιδανική, καλύτερη από μέτρια και ισορροπημένη απόδοση σε όλες τις κατηγορίες συναισθημάτων.

Fold	**Accuracy**	**F1 Score**
1	0.65896	0.65705
2	0.65185	0.65085
3	0.67024	0.66318
4	0.65965	0.65071
5	0.65938	0.65937

Πίνακας 7: Ακρίβεια και μέτρο F1 για κάθε fold του VGG-16.

Σε σχέση με το μοντέλο που αναπαράγουμε, τα αποτελέσματά μας ήρθαν πολύ κοντά. Οι ερευνητές κατάφεραν να πετύχουν μια ακρίβεια 69.4% στο test set ενώ το δικό μας μέγιστο ήταν λίγο πάνω από 67%. Η μικρή απόκλιση μεταξύ των αποτελεσμάτων μας και αυτών των ερευνητών ενδέχεται να οφείλεται στις διαφορές στην προεπεξεργασία των δεδομένων και τις ακριβείς συνθήκες εκπαίδευσης. Παρ' όλ' αυτά, η συνολική απόδοση του μοντέλου μας δείχνει ότι

η αρχιτεκτονική είναι αναπαραγώγιμη και επαρκώς αποδοτική, ακόμη και υπό διαφορετικές πειραματικές συνθήκες.

4.4 InceptionV3

Οι ερευνητές στο [12] πρότειναν μια παρόμοια προσέγγιση transfer learning για το μοντέλο InceptionV3 πάνω στο dataset FER-2013. Επιλέξαμε να αναπαράγουμε τη μεθοδολογία με βάση τις αρχές και τη δομή μοντέλου που παρουσιάστηκαν στο άρθρο.

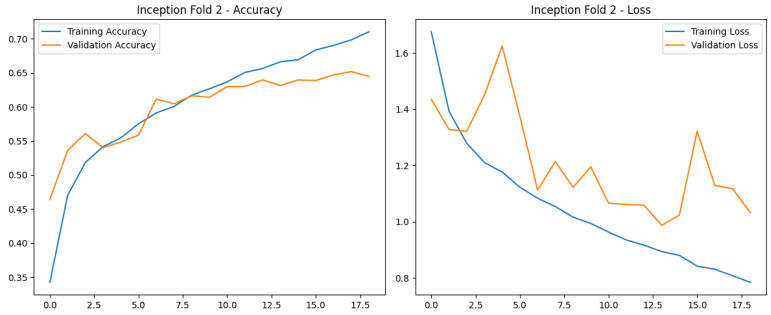
Το μοντέλο βασίστηκε στην αρχιτεκτονική InceptionV3, αλλά με την αλλαγή των τελευταίου Fully Connected Layer με τρια επίπεδα, Flatten, Dropout & Dense. Αν και στο άρθρο δεν αναφέρεται συγκεκριμένη μορφή επεξεργασίας των δεδομένων κρίναμε απαραίτητο να εφαρμόσουμε τεχνικές προσαύξησης δεδομένων, για την ενίσχυση της γενίκευσης του μοντέλου και τη μείωση του overfitting.

Ακολουθήσαμε και πάλι 5-fold cross-validation με χρήση προεκπαιδευμένων βαρών InceptionV3 στο ImageNet. Κάθε fold εκπαιδεύτηκε για 50 επόχες με early stopping και ανοχή 5 εποχών. Για την εκπαίδευση χρησιμοποιήθηκε ο SGD optimizer με learning rate 0.001 και momentum 0.9, όπως και στο πείραμα με το VGG-16.

Fold	**Accuracy**	**F1 Score**
1	0.62413	0.62255
2	0.62636	0.63098
3	0.62928	0.62062
4	0.63750	0.63097
5	0.63207	0.62935

Πίνακας 8: Ακρίβεια και μέτρο F1 για κάθε fold του InceptionV3.

Σύμφωνα με τα αποτελέσματα, το μοντέλο παρουσίασε ελαφρώς χαμηλότερη επίδοση σε σχέση με το αντίστοιχο VGG-16, καθώς το μέγιστο accuracy που επιτεύχθηκε σε fold ήταν 63%, ενώ ο μέσος όρος των accuracies ήταν περίπου 62%. Η επίδοση του μοντέλου παρέμεινε σταθερή, χωρίς έντονες αποκλίσεις μεταξύ των διαφορετικών folds. Στο Σχήμα 4 απεικονίζεται η πορεία του accuracy και του loss κατά την εκπαίδευση στο 2ο fold, όπου φαίνεται ότι το validation accuracy κυμαίνεται κοντά στο training accuracy με μικρές διακυμάνσεις, ένδειξη ότι το μοντέλο διατηρεί σχετικά καλή γενίκευση.



Σχήμα 4: Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής InceptionV3 για το FER-2013 dataset στο 2ο fold

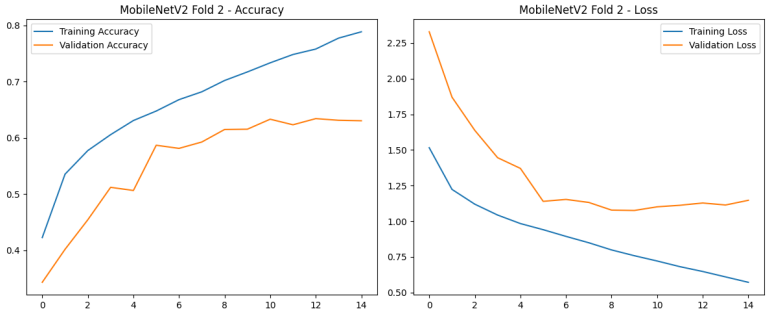
Οι τιμές του F1 score σε κάθε fold ήταν αντίστοιχες με τις τιμές του accuracy, όπως παρουσιάζονται στον Πίνακα 8, κάτι που φανερώνει ισορροπημένη απόδοση ανάμεσα στις κλάσεις. Ωστόσο, η γενική επίδοση ήταν κατώτερη από αυτή που αναφέρεται στο άρθρο (περίπου 72% accuracy στο test set), γεγονός που πιθανώς οφείλεται σε διαφορές στην προεπεξεργασία των δεδομένων ή άλλες παραμέτρους του πειραματισμού που δεν αναφέρονται λεπτομερώς στη δημοσίευση.

4.5 MobileNetV2

Για το μοντέλο MobileNetV2, οι ερευνητές δεν κάνουν ιδιαίτερες αλλαγές στην αρχιτεκτονική. Όπως και στα προηγούμενα μοντέλα η διαδικασία του transfer learning γίνεται με τη χρήση των προεκπαιδευμένων βαρών από το ImageNet. Μετά τα βασικά blocks έχει προστεθεί ένα επίπεδο Global Average Pooling και ένα τελικό Dense επίπεδο με softmax ενεργοποίηση για ταξινόμηση στις κατηγορίες του FER-2013 dataset. Όπως και στις προηγούμενες υλοποιήσεις, εφαρμόστηκε προσαύξηση δεδομένων με τυχαία αναστροφή, περιστροφή, zoom και contrast για την ενίσχυση της γενίκευσης όμως, αντιθέτως με τους ερευνητές, λόγω περιορισμένων υπολογιστικών πόρων, οι εικόνες που χρησιμοποιήσαμε ήταν μικρότερων διαστάσεων.

Η εκπαίδευση πραγματοποιήθηκε και πάλι με 5-fold cross-validation για 50 επόχες με early stopping και ανοχή 5 εποχών. Ο optimizer που χρησιμοποιήθηκε ήταν ο Adam με learning rate 0.0001, όπως και στο άρθρο, καθώς η επιλογή αυτή προσφέρει καλύτερη σύγκλιση σε μοντέλα τύπου MobileNetV2 .

Από τα πειράματα προέκυψε ότι ενώ η τελική μέγιστη ακρίβεια που πετύχαμε



Σχήμα 5: Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής MobileNetV2 για το FER-2013 dataset στο 2ο fold

με το μοντέλο MobileNet (63%) είναι αρκετά μεγαλύτερη από αυτή των ερευνητών (58%), όπως παρατηρείται και στο Σχήμα 5 η ακρίβεια του validation δεν κατάφερε ποτέ να ξεπεράσει, ή να έρθει κοντά σε αυτή του train. Αυτό αποτελεί ένδειξη ότι το μοντέλο εμφάνισε κάποιο βαθμό overfitting στο training set, δηλαδή έμαθε καλύτερα τις ιδιαιτερότητες των εκπαιδευτικών δειγμάτων απ΄ ό,τι τις γενικές αναπαραστάσεις των συναισθημάτων. Αυτό είναι αναμενόμενο σε lightweight μοντέλα όπως το MobileNetV2 που έχουν περιορισμένη χωρητικότητα και συχνά δυσκολεύονται να διαχειριστούν datasets με μεγάλο θόρυβο όπως το FER-2013.

Τελικά, οι τιμές του μέτρου F1 κινήθηκαν παρόμοια με αυτές της ακρίβειας, παρουσιάζοντας σταθερή και σχετικά ισορροπημένη συμπεριφορά σε όλα τα folds.

Fold	**Accuracy**	**F1 Score**
1	0.59877	0.59399
2	0.61396	0.61201
3	0.63360	0.62675
4	0.61828	0.62253
5	0.62148	0.60990

Πίνακας 9: Ακρίβεια και μέτρο F1 για κάθε fold του MobileNetV2.

Η μικρή διαφορά μεταξύ των δύο δεικτών σε κάθε fold, όπως φαίνεται και στον Πίνακα 9, δείχνει ότι το μοντέλο δεν έκανε overfit σε ορισμένες κατηγορίες αλλά κατάφερε να διατηρήσει μια συγκριτικά καλή ισορροπία μεταξύ precision και recall για όλες τις κλάσεις συναισθημάτων. Το γεγονός ότι το F1

score ακολουθεί παρόμοια πορεία ενισχύει την παρατήρηση ότι το μοντέλο αποδίδει ικανοποιητικά σε όλες τις κατηγορίες και δεν ευνοεί περισσότερο κάποια συγκεκριμένη.

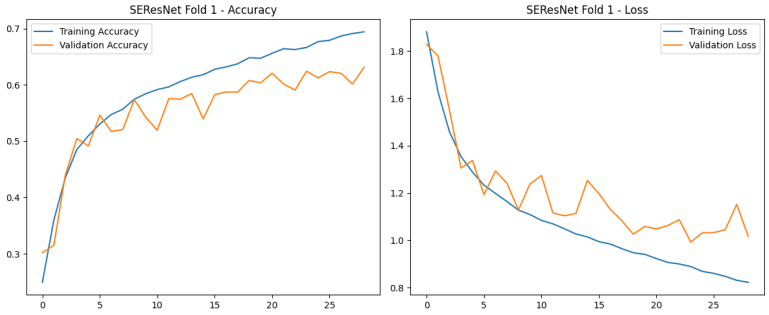
4.6 SE-SResNet18

Ακολουθώντας πιστά το υπόδειγμα της έρευνας [14], δημιουργήσαμε ένα μοντέλο ResNet18, από το οποίο αφαιρέσαμε το αρχικό Max Pooling Layer και μειώσαμε το μέγεθος του kernel στο πρώτο convolutional layer, προσθέτοντας ένα επίπεδο dropout και αντικαθιστώντας το τελευταίο Fully Connected Layer με ένα ίδιο 512 διαστάσεων, ακολουθούμενο από το τελικό classification layer. Ακόμη, εισάγαμε ένα SENet block στα residual blocks για την ενίσχυση της απόδοσης.Με βάση το πρότυπο που ακολουθήθηκε στην εκπαίδευση των προηγούμενων μοντέλων, εφαρμόσαμε 5-fold cross-validation, με 50 εποχές σε κάθε fold και ανοχή 5 εποχών, ενώ ως optimizer χρησιμοποιήθηκε το Adam με learning rate ίσο με 0.0001.

Στο σχήμα 6, απεικονίζονται ενδεικτικά τα αποτελέσματα εκπαίδευσης του πρώτου fold. Τα ποσοστά στο training set παρουσιάζουν σταθερή άνοδο, προσεγγίζοντας το 70% με ρυθμό που δεν συνιστά την παρουσία overfitting. Αν και με σημαντικές διακυμάνσεις, η εξέλιξη του validation φαίνεται να είναι επίσης ανοδική και χωρίς σημαντικές αποκλίσεις από το training. Το loss στο training set παρουσιάζει επίσης αναμενόμενη συμπεριφορά, πέφτοντας κάτω από το 0.9 ενώ το αντίστοιχο validation, αν και διαφοροποιείται ελαφρώς προς τις τελευταίες εποχές, ακολυθεί αντίστοιχη πορεία, υποδεικνύοντας πως το μοντέλο έχει επαρκή δυνατότητα γενίκευσης.

Πρέπει επίσης να σημειωθεί πως, λόγω της περιορισμένου χρόνου runtime της πλατφόρμας Google Colab, το μοντέλο δεν κατάφερε να ολοκληρώσει το πέμπτο fold, σταματώντας στην εποχή 26, χωρίς ωστόσο το συγκεκριμένο fold να παρουσιάζει σημαντικές διαφοροποιήσεις από τα προηγούμενα, με αποτέλεσμα τα καταγεγραμμένα ποσοστά να είναι ενδεικτικά της ικανότητας του μοντέλου. Στον Πίνακα 10 παρουσιάζονται συγκεντρωτικά οι τιμές του accuracy για τα folds που πρόλαβαν να ολοκληρωθούν, καθώς και του F1 score, το οποίο ακολουθεί με μικρές αποκλίσεις, γεγονός που υποδεικνύει αμεροληψία του classifier ως προς τις ανισότητες κατανομής των κατηγοριών.

Γενικα, αποτελέσματα της εκπαίδευσης εμφανίζουν σημαντικές διαφοροποιήσεις από την έρευνα, επιτυγχάνοντας μέση ακρίβεια ίση με 61.3% σε κάθε fold, ενώ η υψηλότερη τιμή που παρατηρήθηκε ήταν 62.2% (οι ερευνητές αναφέρουν accuracy έως και 74.14%). Και πάλι, η πιθανότερη αιτία είναι διαφοροποίησεις



Σχήμα 6: Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής SE-SResNet18 για το FER-2013 dataset στο 1ο fold

στο preprocessing των εικόνων, αλλά και στα hyperparameters που χρησιμοποιήθηκαν. Ωστόσο το μοντέλο που υλοποιήσαμε φαίνεται να έχει επαρκή ικανότητα να μάθει χωρίς να παρουσιάζει overfitting και τα αποτελέσματα κινούνται σε ένα εύρος αποδεκτό για το dataset FER-2013.

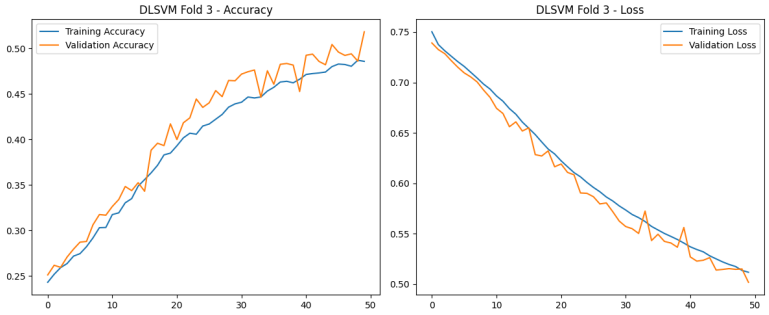
Fold	**Accuracy**	**F1 Score**
1	0.6215	0.6194
2	0.6091	0.6050
3	0.5970	0.5871
4	0.6241	0.6203

Πίνακας 10: Ακρίβεια και μέτρο F1 για κάθε fold του SE-SResNet18.

4.7 DSLVM L2

Για την αναπαραγωγή του πειράματος ακολουθήσαμε πλήρως την αρχιτεκτονική που περιγράφεται στο άρθρο. Χρησιμοποιήσαμε ένα CNN τριών συνελικτικών επιπέδων ακολουθούμενο από ένα Dense layer μεγέθους 3072 και γραμμική έξοδο. Το μοντέλο εκπαιδεύτηκε από το μηδέν με χρήση SGD optimizer με momentum και Nesterov acceleration, όπως προτείνεται στο άρθρο. Ως τεχνική προεπεξεργασίας εφαρμόσαμε mean subtraction, κανονικοποίηση και στατιστική κανονικοποίηση ανά pixel, σύμφωνα με την μεθοδολογία του Tang.

Οι εικόνες του dataset είναι ασπρόμαυρες με ανάλυση 48 x 48, οπότε δεν πραγματοποιήσαμε κάποια αλλαγή, εφόσον αυτές τις ρυθμίσεις ακολουθεί και ο Tang στην ερευνά του. Για την αύξηση του όγκου των δεδομένων (data



Σχήμα 7: Ακρίβεια (accuracy) και Απώλεια (loss) της αρχιτεκτονικής DSLVM L2 για το FER-2013 dataset στο 3ο fold

augmentation), εφαρμόζονται τεχνικές όπως image mirroring και τυχαίοι γεωμετρικοί μετασχηματισμοί, όπως μικρές περιστροφές, μετατοπίσεις και αλλαγές κλίμακας. Από την άλλη, οι ετικέτες κωδικοποιήθηκαν με τον τρόπο που απαιτεί το DLSVM σχήμα (κωδικοποίηση +1 / -1). Κατά την εκπαίδευση χρησιμοποιήσαμε και ήπια προσαύξηση των δεδομένων για τη βελτίωση της γενίκευσης του μοντέλου.

Για την αξιολόγηση εφαρμόσαμε 5-fold cross-validation, με early stopping για την αποφυγή υπερεκπαίδευσης.

Κατά τη διάρκεια της εκπαίδευσης του μοντέλου παρατηρήσαμε ότι, ενώ το accuracy είχε ανοδική πορεία, όπως και στο Σχήμα 7, υπάρχουν πολλές και ραγδαίες διακυμάνσεις μετά τη 10η εποχή. Πέρα από τις διακυμάνσεις, η μεγαλύτερη ακρίβεια που επιτεύχθηκε κατά την εκπαίδευση ήταν 51.21% ενώ η μέση ακρίβεια ανά fold ήταν 50%. Η διαφορά μεταξύ training accuracy και validation accuracy δεν ήταν ιδιαίτερα έντονη, γεγονός που δείχνει ότι το μοντέλο δεν παρουσίασε σοβαρά φαινόμενα υπερπροσαρμογής, αλλά μάλλον περιορίστηκε από την ίδια του την εκφραστική ικανότητα. Το F1 score ακολούθησε αντίστοιχη πορεία με την ακρίβεια (Πίνακας 11), ενισχύοντας την εικόνα ότι το μοντέλο διατηρούσε ισορροπημένη απόδοση σε όλες τις κατηγορίες χωρίς να ευνοεί κάποια συγκεκριμένα.

Ωστόσο, σε σχέση με τα αποτελέσματα του αρχικού άρθρου (70%), η ακρίβεια που επιτύχαμε ήταν αρκετά χαμηλότερη. Αυτό μπορεί να οφείλεται σε διάφορους λόγους που πηγάζουν κυρίως από την έλλειψη λεπτομερειών σε διάφορες πτυχές της υλοποίησης της αρχικής έρευνας.

Παρά την απόκλιση από τα αποτελέσματα της αρχικής έρευνας, το γεγονώς ότι καταφέραμε να φτάσουμε μια ακρίβεια έστω και 50% αποδεικνύει ότι το

Fold	Accuracy	F1 Score
1	0.49568	0.47213
2	0.49067	0.46781
3	0.51184	0.49700
4	0.51212	0.49955
5	0.49735	0.46928

Πίνακας 11: Ακρίβεια και μέτρο F1 για κάθε fold του DSLVM L2.

μοντέλο είναι ικανό να μάθει, όμως με περιορισμένη εκφραστική ικανότητα σε σύγκριση με πιο μοντέρνες τεχνικές.

5 Συμπεράσματα

Η παρούσα έρευνα εστίασε στην συγκριτική αξιολόγηση 5 διαφορετικών CNN-based μοντέλων, για την αναγνώριση ανθρώπινων συναισθημάτων απο φωτογραφίες. Για την εκπαίδευση των μοντέλων αντλήσαμε κάθε δυνατή πληροφορία για την αρχιτεκτονική τους από τη σχετική βιβλιογραφία. Όσον αφορα τις επιδόσεις, υπήρξαν δικές μας υλοποιήσεις που προσέγγισαν αλλά και ξεπέρασαν τις αντίστοιχες προτεινόμενες, ενώ σε άλλες περιπτώσεις είδαμε μεγάλη απόκλιση στην ακρίβεια.

Model	Bibliography	Implementation
VGG16	0.694	0.670
InceptionV3	0.739	0.638
MobileNetV2	0.580	0.633
SE-SResNet18	0.741	0.624
DSLVM L2	0.712	0.512

Πίνακας 12: Σύγκριση αποτελεσμάτων βιβλιογραφίας και παρούσας έρευνας

Στην δική μας έρευνα το υψηλότερο ποσοστό ακρίβειας επετεύχθη μέσω του VGG-16 (67%), ενώ στη βιβλιογραφία παρουσιάζονται υψηλότερα ποσοστά, όπως αυτά του InceptionV3 και του SE-SResNet18, που ξεπερνούν το 70%, όπως φαίνεται στον Πίνακα 12. Θεωρούμε την διαφορά αυτή αποδεκτή, λόγω των περιορισμένων διαθέσιμων υπολογιστικών πόρων, αλλά και της πολλές φορές ελλιπούς πληροφόρησης που είχαμε για την αρχιτεκτονική των μοντέλων και την προεπεξεργασία που δέχθηκε το dataset. Σε κάθε περίπτωση,

τα διεξαχθέντα πειράματα καταδεικνύουν ότι οι σύγχρονες προσεγγίσεις στην αναγνώριση συναισθημάτων μέσω φωτογραφιών γίνονται όλο και πιο ακριβείς, ολοκληρωμένες και ικανές να αναταπεξέλθουν στις απαιτήσεις του πραγματικού κόσμου.

Αναφορές

[1] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, Nov. 3–7, 2013, Proceedings, Part III*, vol. 20, pp. 117–124, Springer, 2013.

[2] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1195–1215, 2020.

[3] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 1998, pp. 200–205.

[4] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 2000, pp. 46–53.

[5] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 94–101.

[6] M. Sambare, "FER-2013," Kaggle, 2020. [Online]. Available: `https://www.kaggle.com/datasets/msambare/fer2013`

[7] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *Sensors*, vol. 21, no. 9, p. 3046, 2021.

[8] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[9] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[10] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 1–40, 2016.

[11] G. P. Kusuma, J. Jonathan, and A. P. Lim, "Emotion recognition on FER-2013 face images using fine-tuned VGG-16," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 5, no. 6, pp. 315–322, 2020.

[12] G. Meena, K. K. Mohbey, S. Kumar, R. K. Chawda, and S. V. Gaikwad, "Image-based sentiment analysis using InceptionV3 transfer learning approach," *SN Comput. Sci.*, vol. 4, no. 3, p. 242, 2023.

[13] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini, and A. Lanata, "Facial emotion recognition (FER) through custom lightweight CNN model: Performance evaluation in public datasets," *IEEE Access*, 2024.

[14] Y. Zhong, S. Qiu, X. Luo, Z. Meng, and J. Liu, "Facial expression recognition based on optimized ResNet," in *Proc. 2nd World Symp. Artif. Intell. (WSAI)*, Jun. 2020, pp. 84–91.

[15] Y. Tang, "Deep learning using linear support vector machines," arXiv:1306.0239, 2013. [Online]. Available: https://arxiv.org/abs/1306.0239