ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Εφαρμογή και Σύγκριση Αλγορίθμων Βαθιάς Μάθησης με Σκοπό την Ανίχνευση Ψεύτικων Εικόνων

Διπλωματική Εργασία

Σειράς Σταύρος

Επιβλέπουσα: Δρ. Δασκαλοπούλου Ασπασία

Ιούλιος 2025

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Εφαρμογή και Σύγκριση Αλγορίθμων Βαθιάς Μάθησης με Σκοπό την Ανίχνευση Ψεύτικων Εικόνων

Διπλωματική Εργασία

Σειράς Σταύρος

Επιβλέπουσα: Δρ. Δασκαλοπούλου Ασπασία

Ιούλιος 2025

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Application and Comparison of Deep Learning Algorithms for Fake Image Detection

Diploma Thesis

Seiras Stavros

Supervisor: Dr. Aspassia Daskalopulu

July 2025

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπουσα	**Δασκαλοπούλου Ασπασία** Αναπληρώτρια Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Βασιλακόπουλος Μιχαήλ** Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας
Μέλος	**Χροναίος Αλέξανδρος** Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελούν αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλουν οποιασδήποτε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχουν έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο Δηλών

Σειράς Σταύρος

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

The Declarant

Seiras Stavros

Ευχαριστίες

Μέσα από αυτή την διπλωματική εργασία τελειώνει ένας σημαντικός κύκλος της ζωής μου, οι σπουδές μου στο τμήμα των Ηλεκτρολόγων Μηχανικών και Μηχανικών Μηχανικών Υπολογιστών του Πανεπιστημίου Θεσσαλίας. Θα ήθελα σε αυτό το σημείο να ευχαριστήσω την επιβλέπουσα καθηγήτρια της εργασίας μου την Κα Δασκαλοπούλου Ασπασία που οποιαδήποτε στιγμή χρειάστηκα την βοήθεια της ήταν διαθέσιμη να μου την προσφέρει. Επίσης θα ήθελα να ευχαριστήσω και τα υπόλοιπα 2 μέλη της επιτροπής για την συμμετοχή τους. Τέλος δεν θα μπορούσα να μην αναφέρω την ευγνωμοσύνη που νιώθω για την οικογένεια μου η οποία με στήριξε σε όλη τη διάρκεια των σπουδών μου.

Διπλωματική Εργασία

Εφαρμογή Αλγορίθμων Βαθιάς Μάθησης με Σκοπό την Ανίχνευση Ψεύτικων Εικόνων

Σειράς Σταύρος

Περίληψη

Η ανίχνευση ψευδών εικόνων αποτελεί κρίσιμο ζήτημα στην εποχή της ψηφιακής πληροφορίας, καθώς η πρόοδος στην τεχνολογία δημιουργίας ψευδών δεδομένων δημιουργεί σημαντικές προκλήσεις για την ασφάλεια και την αξιοπιστία των οπτικών δεδομένων. Η παρούσα εργασία επικεντρώνεται στη συγκριτική ανάλυση και αξιολόγηση αλγορίθμων βαθιάς μάθησης για την ανίχνευση ψευδών εικόνων, με ιδιαίτερη έμφαση στις τεχνολογίες MobileNet, EfficientNet, DenseNet169, καθώς και στις παραδοσιακές τεχνικές LBPH (Local Binary Patterns Histograms) και Fisherface. Η μεθοδολογία περιλαμβάνει την εκπαίδευση και τη δοκιμή των αλγορίθμων σε ποικίλα σύνολα δεδομένων, όπως τα Fake Face Images Generated From Different GANs και FFHQ (Flickr-Faces-HQ), ενώ η αξιολόγηση περιλαμβάνει δείκτες όπως η ακρίβεια, το F1-score και η καμπύλη Precision-Recall.

Επιπλέον, η εργασία εξετάζει την αποτελεσματικότητα των αλγορίθμων σε πραγματικό χρόνο, αναλύοντας την ανθεκτικότητά τους σε διάφορες συνθήκες, όπως η συμπίεση δεδομένων και διαφορετικά σενάρια φωτισμού. Τα ευρήματα της έρευνας αναδεικνύουν τα πλεονεκτήματα και τους περιορισμούς κάθε αλγορίθμου, παρέχοντας σαφή εικόνα για την καταλληλότητά τους σε πρακτικές εφαρμογές ανίχνευσης. Ιδιαίτερη έμφαση δίνεται στην ερμηνευσιμότητα των μοντέλων μέσω εργαλείων όπως οι τιμές SHAP, ενώ παράλληλα προτείνονται τεχνικές βελτίωσης, όπως η αντιμετώπιση ανισορροπίας δεδομένων με τη χρήση της μεθόδου SMOTE. Η εργασία ολοκληρώνεται με τη διατύπωση συμπερασμάτων και προτάσεων για μελλοντική έρευνα, περιλαμβάνοντας την ενσωμάτωση τεχνολογιών αιχμής όπως οι Μετασχηματιστές Οράσεως

(VisionTransformers) και τα Βαθία Δίκτυα Πεποιθήσεων (Deep Belief Networks) με στόχο τη βελτίωση της απόδοσης και της ερμηνευσιμότητας των συστημάτων ανίχνευσης.

Λέξεις-κλειδιά: Βαθιά μάθηση, ψευδείς εικόνες, ανίχνευση ψευδών εικόνων, MobileNet, EfficientNet, DenseNet169, LBPH, Fisherface, SHAP, SMOTE.

Diploma Thesis

Application and Comparison of Deepfake Algorithms for Fake Image Detection

Seiras Stavros

Abstract

One critical challenge that has emerged due to the invention of synthetic data generation is the detection of fake images affecting the authenticity, reliability, and security of digital visual media. A comparative analysis and evaluation of various deep learning algorithms for fake image recognition was carried out here, with a keen focus on MobileNet, EfficientNet, DenseNet169, and classical techniques such as LBPH (Local Binary Patterns Histograms) and Fisherface. The methodology includes training and testing of these algorithms on varied datasets, including Fake Face Images Generated From Different GANs and FFHQ (Flickr-Faces-HQ). Performance assessment is based on measures such as accuracy, F1-score, and the Precision-Recall curve.

The study extends into real-time detection scenarios while keeping in mind robustness against changing conditions such as those owing to data compression or lighting irregularities. The strength and weakness of each algorithm are analyzed in the current study to provide a basis for deciding their appropriateness for detection in real-world applications. Special attention is paid to model interpretability through the application of SHAP values so that general applicability is maintained. Other techniques employed to counteract the dataset imbalance problems are SMOTE and other resampling techniques. The recommendation for future work aims at the application of state-of-the-art techniques, such as Vision Transformers and Deep Belief Networks, to further complement the detection performance and provide interpretable feedback.

Keywords: Deep learning, fake images, fake image detection, MobileNet, EfficientNet, DenseNet169, LBPH, Fisherface, SHAP, SMOTE.

Πίνακας περιεχομένων

Κατάλογος Εικόνων

Συντομογραφίες

CNN: Convolutional Neural Networks

FFHQ: Flickr-Faces-HQ Dataset

GAN: Generative Adversarial Networks

GPU: Graphics Processing Unit

LBPH: Local Binary Patterns Histograms

OpenCV: Open Source Computer Vision Library

PyTorch: Python + Torch (Deep Learning Framework)

ROC-AUC: Receiver Operating Characteristic – Area Under the Curve

SHAP: SHapley Additive exPlanations

SMOTE: Synthetic Minority Over-sampling Technique

ViT: Vision Transformers

ΚΕΦΑΛΑΙΟ 1:ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της Διπλωματικής

Η ανίχνευση ψευδών εικόνων αποτελεί μια από τις μεγαλύτερες προκλήσεις της σύγχρονης ψηφιακής εποχής. Η δυνατότητα δημιουργίας οπτικού περιεχομένου που είναι σχεδόν αδύνατο να διακριθεί από το πραγματικό, μέσω τεχνολογιών βαθιάς μάθησης, δημιουργεί σοβαρές απειλές για την ασφάλεια, την ιδιωτικότητα και την ακεραιότητα της πληροφορίας. Οι ψευδείς εικόνες χρησιμοποιούνται ευρέως, όχι μόνο για ψυχαγωγικούς σκοπούς, αλλά και σε κακόβουλες ενέργειες, όπως η διασπορά παραπλανητικών πληροφοριών, η εξαπάτηση μέσω ψηφιακών μέσων, ή ακόμα και η πλαστογράφηση αποδεικτικών στοιχείων [1], [2].

Η πολυπλοκότητα των ψευδών εικόνων αυξάνεται με την πρόοδο τεχνικών όπως τα ΓεννητικάΑντιπαραθετικάΔίκτυα (GANs), καθιστώντας την ανίχνευσή τους εξαιρετικά δύσκολη. Την ίδια στιγμή, τα υπάρχοντα εργαλεία εντοπισμού συχνά αποτυγχάνουν όταν αντιμετωπίζουν εξελιγμένα μοντέλα δημιουργίας ψευδών δεδομένων. Το πρόβλημα γίνεται ακόμα πιο έντονο σε περιπτώσεις που απαιτούν ανίχνευση σε πραγματικό χρόνο ή σε περιβάλλοντα όπου τα δεδομένα έχουν υποστεί συμπίεση ή αλλοιώσεις. Η επιτακτική ανάγκη για την ανάπτυξη αξιόπιστων μεθόδων ανίχνευσης είναι προφανής, ιδιαίτερα σε κρίσιμους τομείς, όπως η κυβερνοασφάλεια και η προστασία προσωπικών δεδομένων [1].

Η βαθιά μάθηση αποτελεί το βασικό εργαλείο για την ανάπτυξη σύγχρονων τεχνολογιών ανίχνευσης ψευδών εικόνων. Εξελιγμένα μοντέλα, όπως τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs), αξιοποιούνται για την εξαγωγή λεπτομερών χαρακτηριστικών και artifacts, δηλαδή μορφολογικών ανωμαλιών που εισάγονται ακούσια κατά τη διαδικασία δημιουργίας ή αλλοίωσης μιας εικόνας με τεχνητά μέσα, και οι οποίες μπορούν να αποκαλύψουν την ύπαρξη ψευδών δεδομένων. Συγκεκριμένα, τα MobileNet, EfficientNet και DenseNet169 έχουν αναδειχθεί σε εξαιρετικά αποτελεσματικές λύσεις για την ανίχνευση ψευδών εικόνων λόγω της ικανότητάς τους να αναγνωρίζουν λεπτές διαφορές μεταξύ πραγματικών και ψευδών δεδομένων [3],[4],[5]. Το MobileNet, για παράδειγμα, παρέχει μια ελαφριά αρχιτεκτονική, ιδανική για εφαρμογές σε πραγματικό χρόνο, ενώ το EfficientNet χρησιμοποιεί τεχνικές scaling για τη βελτιστοποίηση της απόδοσης. Το DenseNet169 από την άλλη πλευρά, ενσωματώνει πυκνές συνδέσεις που

διασφαλίζουν τη ροή πληροφορίας και ενισχύουν την ανίχνευση λεπτομερών αλλοιώσεων.

Συγχρόνως, παραδοσιακές μέθοδοι, όπως οι Local Binary Patterns Histogram (LBPH) και Fisherface, παρόλο που υστερούν σε σύγκριση με τα CNNs, συνεχίζουν να βρίσκουν εφαρμογή σε περιπτώσεις όπου η απλότητα και η ταχύτητα είναι κρίσιμες [6], [7]. Αυτές οι τεχνικές εξαγωγής χαρακτηριστικών προσφέρουν μια πιο άμεση προσέγγιση, εστιάζοντας σε τοπικές ασυνέπειες στα χαρακτηριστικά των εικόνων.Τα δεδομένα που χρησιμοποιούνται για την ανίχνευση ψευδών εικόνων είναι εξίσου σημαντικά. Σύνολα δεδομένων, όπως τα FFHQ και Fake Face Images Generated From Different GANs, παρέχουν μια πλούσια βάση για την εκπαίδευση και την αξιολόγηση των αλγορίθμων. Αυτά τα σύνολα δεδομένων περιλαμβάνουν παραδείγματα πραγματικών και ψευδών δεδομένων, με ποικιλία σε γεωμετρία, φωτισμό και άλλες παραμέτρους. Η σωστή επιλογή και προετοιμασία των δεδομένων αυτών είναι ζωτικής σημασίας για την επιτυχία οποιουδήποτε συστήματος ανίχνευσης [8], [9].

Η ανίχνευση ψευδών εικόνων απαιτεί επίσης την αξιολόγηση των αλγορίθμων με βάση διάφορους δείκτες απόδοσης. Αυτοί περιλαμβάνουν μεταξύ άλλων την ακρίβεια, το F1-score, την καμπύλη Precision-Recall, οι οποίοι παρέχουν μια λεπτομερή εικόνα για την αποτελεσματικότητα και την ανθεκτικότητα κάθε μεθόδου [10], [11]. Επιπλέον, η χρήση εργαλείων ερμηνευσιμότητας, όπως οι τιμές SHAP, ενισχύει την κατανόηση του τρόπου λειτουργίας των μοντέλων και διευκολύνει τη βελτίωση των αποτελεσμάτων τους [10].

Αυτή η διπλωματική εργασία στοχεύει να εξετάσει και να συγκρίνει λεπτομερώς τις παραπάνω τεχνολογίες, εντοπίζοντας τα πλεονεκτήματα και τις αδυναμίες τους.Στην παρούσα εργασία, υλοποιείται, εκπαιδεύεται και αξιολογείται η απόδοση τριών διαφορετικών αλγορίθμων βαθιάς μάθησης για την ανίχνευση ψευδών εικόνων: MobileNet, EfficientNet, DenseNet169. Η εκπαίδευση των μοντέλων πραγματοποιείται σε δύο διαφορετικά σύνολα δεδομένων, το FFHQ (Flickr-Faces-HQ) και το Fake Face Images Generated From Different GANs, τα οποία περιλαμβάνουν τόσο αυθεντικές όσο και ψευδείς εικόνες.

Μετά την εκπαίδευση, διεξάγονται πειράματα για τη σύγκριση της απόδοσής τους υπό διαφορετικές συνθήκες, όπως μεταβολές στον φωτισμό, συμπίεση εικόνας και αλλαγές στη γεωμετρία του προσώπου.

Μέσα από τη λεπτομερή ανάλυση των αποτελεσμάτων, εντοπίζονται τα πλεονεκτήματα και οι αδυναμίες κάθε αλγορίθμου και προτείνονται βελτιώσεις με στόχο την ενίσχυση της αξιοπιστίας των συστημάτων ανίχνευσης ψευδών εικόνων. Οι βελτιώσεις αυτές περιλαμβάνουν τεχνικές επεξεργασίας δεδομένων, συνδυασμό διαφορετικών αλγορίθμων

ή τη χρήση μεθόδων βελτιστοποίησης, όπως η SMOTE (Synthetic Minority Over-sampling Technique) για την αντιμετώπιση της ανισορροπίας δεδομένων.

Ο στόχος της μελέτης είναι η ανάδειξη των πιο αποδοτικών προσεγγίσεων για την ανίχνευση ψευδών εικόνων, συμβάλλοντας στην ανάπτυξη πιο ακριβών και ανθεκτικών αλγορίθμων σε πραγματικές συνθήκες.

1.2 Συνεισφορά

Η παρούσα εργασία επικεντρώνεται στη λεπτομερή ανάλυση και συγκριτική αξιολόγηση αλγορίθμων βαθιάς μάθησης και παραδοσιακών τεχνικών για την ανίχνευση ψευδών εικόνων, με σκοπό την ανάδειξη των δυνατοτήτων και των περιορισμών τους. Συγκεκριμένα, διερευνώνται η αποδοτικότητα και η ανθεκτικότητα των MobileNet, EfficientNet και DenseNet169 ως εκπροσώπων σύγχρονων αρχιτεκτονικών βαθιάς μάθησης, καθώς και των LBPH και Fisherface ως παραδοσιακών τεχνικών εξαγωγής χαρακτηριστικών.

Η συμβολή της εργασίας έγκειται στην ανάπτυξη μιας συστηματικής προσέγγισης για τη σύγκριση αυτών των μεθόδων, λαμβάνοντας υπόψη διαφορετικά σύνολα δεδομένων, όπως τα FFHQ και FakeFaceImagesGeneratedFromDifferentGANs, τα οποία περιλαμβάνουν ψευδείς και αληθείς εικόνες υπό διάφορες συνθήκες [8], [9]. Μέσω αυτής της ανάλυσης, προσδιορίζονται τα πλεονεκτήματα της χρήσης σύγχρονων CNNs, όπως η αποτελεσματική εξαγωγή χαρακτηριστικών και η υψηλή απόδοση σε σύνθετες συνθήκες φωτισμού ή γεωμετρικών παραμορφώσεων.

Επιπλέον, η εργασία εξετάζει την αποτελεσματικότητα αυτών των αλγορίθμων σε εφαρμογές πραγματικού χρόνου, με έμφαση στην προσαρμοστικότητα των MobileNet και EfficientNet σε περιβάλλοντα με περιορισμένους πόρους [3], [4]. Αντίστοιχα, αξιολογείται η δυνατότητα των LBPH και Fisherface να παρέχουν ικανοποιητικές λύσεις σε εφαρμογές μικρής κλίμακας [6], [7].

Ιδιαίτερη έμφαση δίνεται στην ερμηνευσιμότητα των μοντέλων μέσω της χρήσης εργαλείων όπως οι τιμές SHAP, οι οποίες διευκολύνουν την κατανόηση της λειτουργίας τους και ενισχύουν την αξιοπιστία τους [10]. Παράλληλα, η εργασία εξετάζει τεχνικές αντιμετώπισης της ανισορροπίας δεδομένων, όπως η χρήση της μεθόδου SMOTE, για την ενίσχυση της γενίκευσης των αλγορίθμων [11].

Η συνολική συνεισφορά της έρευνας έγκειται στην παροχή ενός πλαισίου για την κατανόηση και τη βελτίωση των αλγορίθμων ανίχνευσης ψευδών εικόνων. Τα ευρήματα αναμένεται να χρησιμεύσουν ως οδηγός για τη βελτιστοποίηση συστημάτων ανίχνευσης και την προσαρμογή τους σε εξελισσόμενες ανάγκες και τεχνολογίες .

1.3 Οργάνωση της Εργασίας

Στο πρώτο κεφάλαιο παρουσιάζεται το αντικείμενο της έρευνας, η επιστημονική της συνεισφορά και η αναγκαιότητα ανάπτυξης συστημάτων ανίχνευσης ψευδών εικόνων, με αναφορά στο ρόλο των γεννητικών δικτύων και των σύγχρονων προκλήσεων.

Το δεύτερο κεφάλαιο θέτει τη θεωρητική βάση της εργασίας, ξεκινώντας με την παρουσίαση των συνόλων δεδομένων που χρησιμοποιούνται για την εκπαίδευση και τη δοκιμή των αλγορίθμων — συγκεκριμένα των FFHQ (Flickr-Faces-HQ) και Fake Face Images Generated From Different GANs — με έμφαση στα ιδιαίτερα χαρακτηριστικά και τις προκλήσεις που παρουσιάζουν. Στη συνέχεια, εξετάζονται αναλυτικά οι τεχνολογίες ανίχνευσης ψευδών εικόνων, περιλαμβάνοντας τα μοντέλα MobileNet, EfficientNet, DenseNet169, καθώς και τις παραδοσιακές τεχνικές LBPH (Local Binary Patterns Histograms) και Fisherface (μέθοδος βασισμένη στην Ανάλυση Γραμμικών Διακρίσεων - LDA), προσφέροντας συγκριτική ανάλυση ως προς τα χαρακτηριστικά, την υπολογιστική πολυπλοκότητα και την ακρίβειά τους. Το κεφάλαιο ολοκληρώνεται με την παρουσίαση των μεθόδων αξιολόγησης, όπως η ακρίβεια, το F1-score, η καμπύλη Precision-Recall και η ανθεκτικότητα σε διαφορετικά σενάρια χρήσης.

Το τρίτο κεφάλαιο περιγράφει αναλυτικά τη μεθοδολογία που ακολουθήθηκε για την εκπαίδευση, τη δοκιμή και την αξιολόγηση των αλγορίθμων. Αρχικά αναλύονται τα εργαλεία και το περιβάλλον που χρησιμοποιήθηκαν, όπως οι γλώσσες προγραμματισμού (Python), τα frameworks (PyTorch) και η υποδομή GPU (NVIDIA RTX 4070).

Στη συνέχεια, παρουσιάζονται οι τεχνικές που εφαρμόστηκαν για την προεπεξεργασία των δεδομένων, όπως η αύξηση δεδομένων (data augmentation),η κανονικοποίηση εικόνων, και η χρήση τεχνικών αντιμετώπισης ανισορροπίας δεδομένων (SMOTE). Το κεφάλαιο εξετάζει επίσης τη διαδικασία ανάπτυξης και προσαρμογής των αλγορίθμων, συμπεριλαμβανομένων των παραμετροποιήσεων που εφαρμόζονται για τη βελτιστοποίηση της απόδοσής τους.

Ιδιαίτερη έμφαση δίνεται στη δυνατότητα ενσωμάτωσης των αλγορίθμων σε πραγματικά περιβάλλοντα εφαρμογής, προκειμένου να αξιολογηθεί η ικανότητά τους να λειτουργούν σε συνθήκες πραγματικού χρόνου.

Στο τέταρτο κεφάλαιο, παρουσιάζονται τα αποτελέσματα της αξιολόγησης των αλγορίθμωνMobileNetV2, EfficientNet-B0, DenseNet169, μεταξύ τους, καθώς και σε σχέση με τις παραδοσιακές μεθόδους LBPH και Fisherfaceμε τη χρήση των επιλεγμένων μεθόδων και συνόλων δεδομένων

Αναλύονται επίσης η γενίκευση των αλγορίθμων σε διαφορετικά σύνολα δεδομένων και οι επιδόσεις τους σε σενάρια πραγματικού χρόνου, ενώ γίνεται αναφορά στη χρήση εργαλείων ερμηνευσιμότητας, όπως οι τιμές SHAP, οι οποίες επιτρέπουν την κατανόηση των χαρακτηριστικών που επηρεάζουν τις αποφάσεις των μοντέλων [10].

Τέλος, το κεφάλαιο περιλαμβάνει ανάλυση σφαλμάτων για τη βελτίωση των χαρακτηριστικών και της απόδοσης των μοντέλων.

Το τελευταίο κεφάλαιο συνοψίζει τα κύρια ευρήματα και τις συνεισφορές της παρούσας έρευνας, παρέχοντας μια ολοκληρωμένη εικόνα για την αποδοτικότητα και τα όρια των αλγορίθμων που εξετάστηκαν. Η συγκριτική ανάλυση αναμένεται να αναδείξει τη σημασία της χρήσης σύγχρονων αρχιτεκτονικών βαθιάς μάθησης, όπως τα MobileNetV2, EfficientNet-B0 και DenseNet169, οι οποίες προσφέρουν υψηλή ακρίβεια και προσαρμοστικότητα σε σύνθετες συνθήκες. Αντίστοιχα, οι παραδοσιακές τεχνικές LBPH και Fisherface αποδείχθηκαν χρήσιμες σε περιπτώσεις με περιορισμένους υπολογιστικούς πόρους, παρά τους περιορισμούς τους στη γενίκευση.

Ένα από τα βασικά συμπεράσματα αφορά την επίδραση της ποιότητας των δεδομένων στην απόδοση των αλγορίθμων. Τα σύνολα δεδομένων, όπως τα FFHQ και Fake Face Images Generated From Different GANs, αποδείχθηκαν κρίσιμα για την εκπαίδευση και την αξιολόγηση των μεθόδων, προσφέροντας ποικιλία προκλήσεων που βοήθησαν στην αποτίμηση της γενίκευσης και της ανθεκτικότητάς τους.

Παράλληλα, η εφαρμογή εργαλείων όπως οι τιμές SHAP υπογράμμισε τη σημασία της ερμηνευσιμότητας στη βελτίωση της αξιοπιστίας των συστημάτων.

Από την έρευνα προκύπτει ότι υπάρχει ανάγκη περαιτέρω βελτίωσης σε συγκεκριμένους τομείς, όπως η αντιμετώπιση της ανισορροπίας δεδομένων, με τεχνικές όπως η SMOTE, καθώς και η ενίσχυση της ανθεκτικότητας των μοντέλων σε διαφορετικές συνθήκες φωτισμού και γεωμετρικές αλλοιώσεις [3], [12].

Οι μελλοντικές ερευνητικές κατευθύνσεις περιλαμβάνουν:

1. Αξιολόγηση σε μεγαλύτερα και πιο ποικιλόμορφα σύνολα δεδομένων

Η δοκιμή των αλγορίθμων σε μεγαλύτερα σύνολα δεδομένων, όπως εκείνα που περιλαμβάνουν εικόνες και βίντεο από διάφορες γεωγραφικές περιοχές, πολιτισμικά υπόβαθρα ή διαφορετικά επίπεδα ποιότητας, θα δώσει τη δυνατότητα να αξιολογηθεί η προσαρμοστικότητα και η ανθεκτικότητα των μοντέλων σε διάφορα σενάρια.

2. Βελτιστοποίηση για εφαρμογές πραγματικού χρόνου, προκειμένου να επιτευχθεί ταχεία και αξιόπιστη ανίχνευση deepfakes σε πραγματικές εφαρμογές.

3. Εξέταση τεχνικών αυτοεποπτευόμενης μάθησης, όπου τα μοντέλα μπορούν να μαθαίνουν από μη επισημασμένα δεδομένα, και μεταμάθησης (meta-learning), δηλαδή τεχνικών όπου το μοντέλο μαθαίνει πώς να μαθαίνει, βελτιώνοντας την ικανότητα του να προσαρμόζεται γρήγορα σε νέες συνθήκες με περιορισμένα δεδομένα.

4. Κατανόηση των τρόπων με τους οποίους τα συστήματα ανίχνευσης deepfakes επηρεάζονται από αντίπαλες επιθέσεις και ανάπτυξη στρατηγικών αντίμετρων που είναι κρίσιμη για την ασφάλεια των εφαρμογών.

Το κεφάλαιο αυτό ολοκληρώνει την εργασία, προσφέροντας κατευθύνσεις για μελλοντικές ερευνητικές επεκτάσεις, με στόχο τη συνεχή βελτίωση των μεθόδων ανίχνευσης ψευδών εικόνων, λαμβάνοντας υπόψη τις σύγχρονες εξελίξεις στη βιβλιογραφία.

ΚΕΦΑΛΑΙΟ 2: ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Εισαγωγή

Η βαθιά μάθηση αποτελεί έναν από τους πλέον ανατρεπτικούς κλάδους της τεχνητής νοημοσύνης, με εφαρμογές που εκτείνονται από την αναγνώριση εικόνων και φωνής έως την ανάλυση δεδομένων μεγάλης κλίμακας. Η ικανότητα των βαθέων νευρωνικών δικτύων να επεξεργάζονται δεδομένα και να εξάγουν χαρακτηριστικά χωρίς την ανάγκη χειροκίνητης παρεμβολής, τα καθιστά εξαιρετικά χρήσιμα σε προκλήσεις όπως η ανίχνευση ψευδών εικόνων [1].

Οι ψευδείς εικόνες έχουν αναδειχθεί σε μείζον ζήτημα στην ψηφιακή εποχή, καθώς εξελιγμένες τεχνικές, όπως τα Generative Adversarial Networks (GANs), διευκολύνουν τη δημιουργία δεδομένων υψηλής ακρίβειας που συχνά ξεπερνούν την ανθρώπινη ικανότητα διάκρισης [2]. Αυτή η πρόοδος, αν και εντυπωσιακή, θέτει σοβαρούς κινδύνους για την ασφάλεια των οπτικών δεδομένων, την ιδιωτικότητα και την ακεραιότητα της πληροφορίας, καθιστώντας απαραίτητη την ανάπτυξη αξιόπιστων μεθόδων ανίχνευσης.Σύμφωνα με την υπάρχουσα βιβλιογραφία, η ανίχνευση ψευδών εικόνων που δημιουργούνται μέσω γεννητικών αντιπαραθετικών δικτύων (GANs) αποτελεί μια από τις πιο περίπλοκες και απαιτητικές προκλήσεις στον τομέα της ασφάλειας και της πιστοποίησης οπτικών δεδομένων. Τα Generative Adversarial Networks (GANs) έχουν εξελιχθεί σε τέτοιο βαθμό ώστε να παράγουν εξαιρετικά ρεαλιστικές συνθετικές εικόνες, οι οποίες συχνά είναι δυσδιάκριτες ακόμα και για τον ανθρώπινο παρατηρητή. Αυτή η τεχνολογική πρόοδος εγείρει σοβαρά ζητήματα ασφαλείας και αξιοπιστίας, ειδικά σε εφαρμογές όπου η ακεραιότητα της οπτικής πληροφορίας είναι κρίσιμη, όπως η εγκληματολογία, οι βιομετρικές ταυτοποιήσεις και η επαλήθευση ψηφιακού περιεχομένου.

Η δυνατότητα των βαθέων νευρωνικών δικτύων να εντοπίζουν λεπτομέρειες, όπως ανωμαλίες στη δομή του δέρματος, αποκλίσεις στις σκιές ή ασυνέπειες στα χαρακτηριστικά των προσώπων, έχει καταστήσει τη βαθιά μάθηση έναν ακρογωνιαίο λίθο για την ανίχνευση ψευδών εικόνων. Οι τεχνικές αυτές δεν βασίζονται μόνο στην ανάλυση χαρακτηριστικών που είναι αντιληπτά από το ανθρώπινο μάτι, αλλά αξιοποιούν τα "artifacts", δηλαδή τεχνητά ίχνη ή ατέλειες που προκύπτουν κατά τη

διαδικασία δημιουργίας ή παραποίησης μιας εικόνας με χρήση αλγορίθμων, όπως τα GANs. Τα artifacts αποτελούν υποπροϊόντα των διαδικασιών σύνθεσης εικόνων και συχνά περιλαμβάνουν ανεπαίσθητες παραμορφώσεις, θόρυβο ή επαναλαμβανόμενα μοτίβα, τα οποία ενδέχεται να διαφεύγουν της ανθρώπινης αντίληψης αλλά αναγνωρίζονται από τα νευρωνικά δίκτυα [1].

Η ανάπτυξη αλγορίθμων ανίχνευσης επικεντρώνεται στην αντιμετώπιση των εξελισσόμενων τεχνικών δημιουργίας ψευδών δεδομένων. Η δυνατότητα των συστημάτων βαθιάς μάθησης να γενικεύουν και να προσαρμόζονται σε νέες τεχνολογίες παραποίησης καθιστά τη χρήση τους κρίσιμη. Ωστόσο, για την αποτελεσματική υλοποίηση αυτών των συστημάτων απαιτείται ένας συνδυασμός καινοτόμων αλγορίθμων, κατάλληλων συνόλων δεδομένων εκπαίδευσης και αξιόπιστων δεικτών αξιολόγησης.

Το κεφάλαιο αυτό εισάγει το θεωρητικό υπόβαθρο της βαθιάς μάθησης και την εφαρμογή της στην ανίχνευση ψευδών εικόνων, θέτοντας τις βάσεις για τη συζήτηση των τεχνικών προκλήσεων και των ερευνητικών ερωτημάτων που απασχολούν την παρούσα εργασία.

2.2 Περιγραφή των Συνόλων Δεδομένων

Η επιλογή των κατάλληλων συνόλων δεδομένων αποτελεί θεμελιώδη παράγοντα για την επιτυχημένη εκπαίδευση και αξιολόγηση αλγορίθμων ανίχνευσης ψευδών εικόνων. Στην παρούσα εργασία, χρησιμοποιούνται δύο βασικά σύνολα δεδομένων που εξασφαλίζουν τόσο τη ρεαλιστικότητα των πραγματικών δεδομένων όσο και την αυξημένη δυσκολία διάκρισης των ψευδών εικόνων. Συγκεκριμένα, το Flickr-Faces-HQ (FFHQ) αποτελεί τη βάση για την κατανόηση των χαρακτηριστικών των πραγματικών ανθρώπινων προσώπων, ενώ το Fake Face Image–Kaggle Dataset περιλαμβάνει αποκλειστικά ψευδείς εικόνες που δημιουργήθηκαν με προηγμένες γεννητικές αντιπαραθετικές τεχνολογίες (GANs). Η συνδυαστική χρήση των δύο συνόλων δεδομένων πραγματοποιείται μέσω της ενοποίησής τους σε ένα ενιαίο σύνολο δεδομένωνδυαδικής ταξινόμησης, όπου οι εικόνες του FFHQ επισημαίνονται ως αυθεντικές και εκείνες του Fake Face Images ως ψευδείς. Έτσι το ενοποιημένο σύνολο δεδομένων παρέχει τη δυνατότητα εκπαίδευσης και αξιολόγησης αλγορίθμων βαθιάς μάθησης με στόχο τη διάκριση μεταξύ πραγματικών και ψευδών εικόνων.

Το ενοποιημένο αυτό σύνολο δεδομένων διαχωρίζεται σε υποσύνολα εκπαίδευσης (80%) και επικύρωσης (20%), επιτρέποντας την αποτελεσματική εκπαίδευση και εποπτεία της απόδοσης των αλγορίθμων βαθιάς μάθησης. Η διαδικασία αυτή ενισχύει τη γενικευσιμότητα των μοντέλων, καθώς επιτρέπει την παρακολούθηση της απόδοσης τους σε δεδομένα που δεν έχουν χρησιμοποιηθεί για εκπαίδευση. Μέσω αυτού του διαχωρισμού, το μοντέλο μαθαίνει να αναγνωρίζει τα λεπτά μορφολογικά και υφολογικά χαρακτηριστικά που διαφοροποιούν τις αυθεντικές από τις συνθετικές εικόνες, ενισχύοντας την ικανότητά του να ανταποκρίνεται σε νέα, άγνωστα δείγματα.

Το Flickr-Faces-HQ (FFHQ) είναι ένα από τα πιο διαδεδομένα σύνολα δεδομένων υψηλής ανάλυσης που περιλαμβάνει 52.000 εικόνες προσώπων σε ανάλυση 512×512 pixels. Οι εικόνες έχουν ληφθεί από την πλατφόρμα Flickr, έναν από τους μεγαλύτερους αποθηκευτικούς χώρους εικόνων στο διαδίκτυο, και καλύπτουν ένα ευρύ φάσμα χαρακτηριστικών. Περιλαμβάνονται πρόσωπα από διαφορετικές ηλικιακές ομάδες, εθνικότητες και μορφολογικά χαρακτηριστικά, γεγονός που επιτρέπει την εκπαίδευση μοντέλων που μπορούν να γενικεύσουν και να μην περιορίζονται από συγκεκριμένες οπτικές πληροφορίες. Επιπλέον, το σύνολο δεδομένων χαρακτηρίζεται από μεγάλη ποικιλομορφία σε στοιχεία όπως τα αξεσουάρ (γυαλιά οράσεως, γυαλιά ηλίου, καπέλα) και η διαμόρφωση του φόντου. Η υψηλή ανάλυση επιτρέπει τη λεπτομερή ανάλυση της υφής του δέρματος, των σκιών και των φωτιστικών συνθηκών, παρέχοντας μια ισχυρή βάση για τη διαφοροποίηση μεταξύ πραγματικών και συνθετικών προσώπων.

Οι εικόνες του FFHQ έχουν υποστεί αυτόματη ευθυγράμμιση και περικοπή μέσω του εργαλείου dlib, προκειμένου να εστιάζουν αποκλειστικά στα πρόσωπα. Παράλληλα, για τη διασφάλιση της ποιότητας των δεδομένων, εφαρμόστηκαν αυτόματα φίλτρα και, επιπλέον, πραγματοποιήθηκε ανθρώπινη επιμέλεια μέσω της πλατφόρμας Amazon Mechanical Turk, ώστε να αφαιρεθούν εικόνες που περιείχαν αγάλματα, πίνακες ζωγραφικής ή φωτογραφίες από άλλες φωτογραφίες. Αυτό διασφαλίζει ότι το σύνολο δεδομένωναποτελεί μια καθαρή και αξιόπιστη βάση για την ανάλυση πραγματικών εικόνων προσώπων.

Το Fake Face Images–Kaggle Dataset αποτελεί ένα σύνολο δεδομένων που περιλαμβάνει αποκλειστικά ψευδείς εικόνες προσώπων, συνολικά 21.300 εικόνες, οι οποίες δημιουργήθηκαν μέσω τριών διαφορετικών γεννητικών αντιπαραθετικών δικτύων (GANs). Συγκεκριμένα, περιλαμβάνει εικόνες από τα ProGAN, StyleGAN2 και StyleGAN3, που θεωρούνται από τις πιο προηγμένες τεχνολογίες παραγωγής συνθετικών εικόνων. Κάθε υποκατηγορία του συνόλου δεδομένων χαρακτηρίζεται από συγκεκριμένα τεχνικά χαρακτηριστικά, τα οποία επηρεάζουν την ποιότητα και την πιστότητα των παραγόμενων ψευδών εικόνων.

Οι εικόνες που προέρχονται από το ProGAN_128 είναι χαμηλότερης ανάλυσης (128×128 pixels), αλλά χαρακτηρίζονται από σταδιακή βελτίωση της ποιότητας καθώς αυξάνεται η ανάλυση κατά τη διαδικασία εκπαίδευσης. Το StyleGAN2_256 προσφέρει εικόνες υψηλότερης ανάλυσης (256×256 pixels) και επιτρέπει τον έλεγχο συγκεκριμένων χαρακτηριστικών, βελτιώνοντας τη διαφοροποίηση μεταξύ των παραγόμενων εικόνων. Το StyleGAN3_256, η πιο πρόσφατη έκδοση της οικογένειας StyleGAN, εισάγει περαιτέρω βελτιώσεις στην ομαλότητα και τη φυσικότητα των ψευδών εικόνων, εξαλείφοντας πολλές από τις ασυνέπειες και τις τεχνητές ανωμαλίες που υπήρχαν σε προηγούμενα μοντέλα.

Το σύνολο δεδομένων Fake Face Images – Kaggle Dataset παρέχει ένα ισχυρό και απολύτως τεκμηριωμένο υπόβαθρο για την αξιολόγηση των αλγορίθμων ανίχνευσης ψευδών εικόνων, καθώς περιλαμβάνει συνθετικές εικόνες ανθρώπινων προσώπων που έχουν παραχθεί μέσω πολλαπλών και ετερογενών γεννητικών μοντέλων. Τα μοντέλα αυτά ανήκουν στην κατηγορία των γεννητικών ανταγωνιστικών δικτύων (Generative Adversarial Networks – GANs) και περιλαμβάνουν μεταξύ άλλων τα StyleGAN, ProGAN, DeepFakeGAN και StarGAN. Κάθε μία από αυτές τις αρχιτεκτονικές έχει σχεδιαστεί με σκοπό την παραγωγή φωτορεαλιστικών προσώπων, τα οποία σε πολλές περιπτώσεις είναι οπτικά μη διακριτά από πραγματικές εικόνες, αυξάνοντας σημαντικά τη δυσκολία του προβλήματος ανίχνευσης και καθιστώντας την αξιοπιστία των ανιχνευτικών τεχνικών κρίσιμη.

Στο πλαίσιο της παρούσας εργασίας, το Fake Face Images – Kaggle Dataset συνενώνεται συστηματικά με το FFHQ (Flickr-Faces-HQ), ένα αξιόπιστο σύνολο δεδομένων αυθεντικών ανθρώπινων προσώπων, προκειμένου να δημιουργηθεί ένα ενιαίο, σύνολο δεδομένων κατάλληλο για την εποπτευόμενη εκπαίδευση δυαδικής

ταξινόμησης μεταξύ πραγματικών και ψευδών εικόνων. Η ενοποίηση αυτή πραγματοποιείται σε κοινή δομή φακέλων, με τις αυθεντικές εικόνες να τοποθετούνται στον υποφάκελο real/ και τις συνθετικές στον υποφάκελο fake/, εντός του καταλόγου dataset/merged/, ο οποίος αξιοποιείται απευθείας από τον κώδικα του συστήματος μέσω της κλάσης ImageFolder. Το τελικό ενοποιημένο σύνολο δεδομένων θα αναφέρεται στη συνέχεια ως *Merged Face Dataset* για λόγους συντομίας και αναγνωσιμότητας.

Στο πλαίσιο της παρούσας εργασίας, το σύνολο δεδομένων Fake Face Images–Kaggle Dataset, το οποίο περιέχει εικόνες προσώπων που έχουν παραχθεί με γεννητικά αντίπαλα δίκτυα (GANs), συνενώνεται συστηματικά με το Flickr-Faces-HQ (FFHQ), ένα ευρέως χρησιμοποιούμενο σύνολο εικόνων αυθεντικών ανθρώπινων προσώπων. Σκοπός αυτής της ενοποίησης είναι η δημιουργία ενός ενιαίου, ισορροπημένου συνόλου δεδομένων δυαδικής ταξινόμησης, το οποίο να καθιστά εφικτή την εποπτευόμενη μάθηση μεταξύ ψευδών και πραγματικών εικόνων. Η δομή του ενιαίου dataset ακολουθεί την τυπική μορφή φακέλων του PyTorchAPItorchvision.datasets.ImageFolder, με τις ψευδείς εικόνες να τοποθετούνται στον υποφάκελο fake/ και τις αυθεντικές στον real/, εντός του καταλόγου dataset/merged/.

Το dataset αυτό φορτώνεται στον κώδικα μέσω της μεταβλητής full_dataset, η οποία αποτελεί instance της έτοιμης κλάσης Image Folder, χωρίς να έχει οριστεί κάποια custom Python κλάση τύπου Merged Face Dataset. Ωστόσο, για λόγους συνέπειας, συντομίας και ευκρίνειας στη διατύπωση που ακολουθεί, το ενοποιημένο αυτό σύνολο δεδομένων θα αναφέρεται συμβατικά ως Merged Face Dataset (MFDS), παρότι δεν αποτελεί όνομα κάποιας συγκεκριμένης κλάσης στον πηγαίο κώδικα.

Το ενοποιημένο σύνολο δεδομένων χαρακτηρίζεται από υψηλό βαθμό ρεαλισμού και τεχνικής πολυπλοκότητας, καθώς περιλαμβάνει συνθετικές εικόνες που έχουν παραχθεί μέσω σύγχρονων γεννητικών τεχνικών, προσομοιώνοντας με πιστότητα τα μορφολογικά και υφολογικά χαρακτηριστικά αυθεντικών προσώπων. Η παρουσία αυτών των οπτικά πειστικών αλλά ψευδών δεδομένων δημιουργεί ένα περιβάλλον υψηλής δυσκολίας που ωθεί τους αλγορίθμους να εντοπίζουν μη προφανείς διαφορές μεταξύ πραγματικών και τεχνητών εικόνων. Η πρόκληση αυτή συμβάλλει καθοριστικά στην εκπαίδευση και αξιολόγηση προηγμένων συστημάτων ανίχνευσης, ενισχύοντας την ακρίβεια, την ανθεκτικότητα και τη δυνατότητα

γενίκευσης των μοντέλων, ιδιαίτερα σε συνθήκες αβεβαιότητας ή σκόπιμης παραπλάνησης. Ως εκ τούτου, το Merged Face Dataset (MFDS) αποτελεί ένα ρεαλιστικό και αξιόπιστο υπόβαθρο για την ανάπτυξη αλγορίθμων αιχμής στον τομέα της ανίχνευσης ψευδούς οπτικού περιεχομένου.

Η αξιοποίηση των συμπερασμάτων από τις μελέτες των Kohlikαι Gupta [18] και Ilyasetal[19] επιτρέπει την ενσωμάτωση βελτιστοποιημένων τεχνικών που δεν βασίζονται αποκλειστικά στην οπτική πληροφορία, αλλά αξιοποιούν συχνοτικά χαρακτηριστικά και μη γραμμικές παραμορφώσεις που αποκαλύπτουν τη συνθετική προέλευση των εικόνων. Παρουσιάζουμε τις τεχνολογίες που χρησιμοποιούνται για την ανίχνευση ψευδών εικόνων στην επόμενη ενότητα.

2.3 Τεχνολογίες

Η ανίχνευση ψευδών εικόνων βασίζεται σε ένα σύνολο τεχνολογιών και αλγορίθμων που επιτρέπουν την εξαγωγή και ανάλυση χαρακτηριστικών για τη διάγνωση της αυθεντικότητας. Στην παρούσα ενότητα παρουσιάζονται οι κύριες τεχνολογίες που χρησιμοποιούνται στην εργασία, με έμφαση στις σύγχρονες αρχιτεκτονικές βαθιάς μάθησης καθώς και στις παραδοσιακές μεθόδους εξαγωγής χαρακτηριστικών.

2.3.1 MobileNet - LightweightCNNs για κινητές εφαρμογές

Το MobileNet αποτελεί μία από τις πιο καινοτόμες αρχιτεκτονικές συνελικτικών νευρωνικών δικτύων (CNNs), σχεδιασμένη με κύριο στόχο τη χρήση σε συσκευές με περιορισμένους υπολογιστικούς πόρους, όπως κινητά τηλέφωνα, tablets, ενσωματωμένα συστήματα και IoT συσκευές [3]. Η ανάπτυξή του ανταποκρίνεται στην αυξανόμενη ζήτηση για αλγορίθμους βαθιάς μάθησης που να είναι ταυτόχρονα αποδοτικοί και οικονομικοί από πλευράς επεξεργαστικής ισχύος, διατηρώντας υψηλά επίπεδα απόδοσης σε ποικίλα πεδία εφαρμογής.

Κύρια Χαρακτηριστικά του MobileNet

Η βασική καινοτομία του MobileNet έγκειται στη χρήση των depthwiseseparableconvolutions, μιας τεχνικής που επιτρέπει τη διαίρεση των συνελικτικών επιπέδων σε δύο ξεχωριστές φάσεις:

- **Depthwise convolution (συνέλιξη κατά βάθος)**

Εφαρμόζει ένα φίλτρο σε κάθε κανάλι εισόδου ξεχωριστά.

- **Pointwise convolution (σημειακή συνέλιξη)**

Συνδυάζει τα αποτελέσματα των depthwise convolutions μέσω γραμμικών συσχετίσεων, δημιουργώντας το τελικό αποτέλεσμα.

Αυτή η διάσπαση μειώνει σημαντικά την υπολογιστική πολυπλοκότητα και τις απαιτήσεις αποθήκευσης της αρχιτεκτονικής. Συγκεκριμένα, η τεχνική αυτή μειώνει τον αριθμό των παραμέτρων και των υπολογιστικών πράξεων κατά σχεδόν 8–9 φορές σε σύγκριση με τα παραδοσιακά συνελικτικά νευρωνικά δίκτυα, όπως το VGG (Visual Geometry Group network), το οποίο αποτελεί μια βαθιά αρχιτεκτονική με διαδοχικές συνελικτικές στρώσεις σταθερού μεγέθους, και το ResNet (Residual Neural Network), το οποίο βασίζεται σε υπομονάδες με residual συνδέσεις για την επίλυση του προβλήματος εξαφάνισης του βαθμιδωτού σφάλματος σε πολύ βαθιά δίκτυα. Παρά τις διαφορές στον σχεδιασμό τους, και οι δύο αρχιτεκτονικές παρουσιάζουν σημαντικό υπολογιστικό βάρος. Αντίθετα, η χρήση πιο αποδοτικών τεχνικών, όπως η depthwise separable convolution, μειώνει δραστικά το κόστος χωρίς να προκαλεί αισθητή πτώση στην απόδοση [3].

Συγκεκριμένα η εφαρμογή των depthwise separable convolutions μειώνει τη συνολική πολυπλοκότητα από $O(n^2 \cdot d^2 \cdot m \cdot k^2)$ σε $O(n^2 \cdot d^2 \cdot m + n^2 \cdot d^2 \cdot k^2)$, όπου

n: Το πλήθος των χαρακτηρηστικών.

d: Αριθμός καναλιών εισόδου.

m: Αριθμός καναλιών εξόδου.

k: Μέγεθος του φίλτρου.

Αυτή η μείωση πολυπλοκότητας είναι κρίσιμη για εφαρμογές όπου οι υπολογιστικοί πόροι είναι περιορισμένοι, καθιστώντας το MobileNet ιδανικό για περιβάλλοντα πραγματικού χρόνου [3].

- **Υψηλή απόδοση σε μικρά σύνολα δεδομένων**

Χάρη στη σχεδίασή του, το MobileNet προσαρμόζεται καλά σε εφαρμογές όπου τα διαθέσιμα δεδομένα είναι περιορισμένα ή υπάρχει ανάγκη για γρήγορη επεξεργασία .

Ευελιξία σε φορητές εφαρμογές

Οι περιορισμένες απαιτήσεις μνήμης το καθιστούν ιδανικό για εφαρμογές ανίχνευσης ψευδών εικόνων που απαιτούν φορητότητα, όπως η χρήση σε έξυπνες κάμερες ή εφαρμογές επαυξημένης πραγματικότητας (AR).

Επιπτώσεις και Χρήσεις στην Ανίχνευση Ψευδών Εικόνων

Στην ανίχνευση ψευδών εικόνων, το MobileNet χρησιμοποιείται κυρίως για την εξαγωγή χαρακτηριστικών που σχετίζονται με ανωμαλίες και artifacts. Αυτά μπορεί να περιλαμβάνουν:

- **Ανωμαλίες στην υφή του δέρματος**

Αναλύει μικρές διαφοροποιήσεις που είναι δύσκολο να ανιχνευθούν από τον ανθρώπινο οφθαλμό.

- **Ασυνέπειες φωτισμού ή σκιών**

  Ανιχνεύει μικροδιαφορές που συχνά προδίδουν την ψευδότητα της εικόνας.

- **Μικροδιαφορές στις γεωμετρικές ιδιότητες του προσώπου**

  Εστιάζει σε διαφορές που προκύπτουν λόγω ατελούς παραποίησης .

2.3.2 EfficientNet - Βελτιστοποιημένο CNN με υψηλή απόδοση

Σύμφωνα με το [4] το Efficient Net είναι μια σύγχρονη αρχιτεκτονική συνελικτικών νευρωνικών δικτύων (CNN) που εισάγει μια καινοτόμο προσέγγιση για τη βελτιστοποίηση της κλιμάκωσης των μοντέλων. Σε αντίθεση με τις παραδοσιακές μεθόδους, όπου η κλιμάκωση των συνελικτικών δικτύων γίνεται ανεξάρτητα για το βάθος, το πλάτος ή την ανάλυση των εικόνων, το EfficientNet βασίζεται σε μια συστηματική προσέγγιση γνωστή ως compound scaling (ισορροπημένη κλιμάκωση). Αυτή η τεχνική αυξάνει ταυτόχρονα και ισορροπημένα τις διαστάσεις αυτές, εξασφαλίζοντας τη βέλτιστη χρήση των πόρων του μοντέλου.

Με τον τρόπο αυτό, το βάθος (depth), το πλάτος (width) και η ανάλυση εισόδου (resolution) του δικτύου αυξάνονται αναλογικά. Η ισορροπημένη αυτή κλιμάκωση επιτυγχάνει:

- Βελτίωση της ακρίβειας, καθώς το αυξημένο βάθος επιτρέπει την εκμάθηση πιο σύνθετων χαρακτηριστικών .
- Μείωση της υπερπροσαρμογής, αφού το πλάτος εξασφαλίζει ότι τα χαρακτηριστικά κατανέμονται σε μεγαλύτερο .
- Ανάδειξη λεπτών χαρακτηριστικών, χάρη στην αυξημένη ανάλυση των εισόδων.

Ο συνδυασμός αυτών των τριών διαστάσεων με τη χρήση του compoundscaling οδηγεί σε μοντέλα που είναι ταυτόχρονα αποδοτικά και ακριβή, γεγονός που τα καθιστά κατάλληλα για ποικίλες εφαρμογές υπολογιστικής όρασης, όπως η κατηγοριοποίηση εικόνων, η ανίχνευση και τοπική αναγνώριση αντικειμένων, η ανάλυση ιατρικής απεικόνισης, καθώς και η αναγνώριση προσώπων σε πραγματικό χρόνο, όπου απαιτείται υψηλή ακρίβεια με περιορισμένους υπολογιστικούς πόρους.

Βελτιστοποίηση Αριθμού Παραμέτρων

Μία από τις σημαντικότερες προκλήσεις στη σχεδίαση συνελικτικών νευρωνικών δικτύων είναι η εξισορρόπηση μεταξύ της απόδοσης και του αριθμού παραμέτρων. Το EfficientNet πετυχαίνει υψηλή ακρίβεια με σημαντικά μικρότερο αριθμό παραμέτρων σε σχέση με παραδοσιακές αρχιτεκτονικές. Αυτή η μείωση της πολυπλοκότητας επιτυγχάνεται μέσω:

- Της χρήσης efficient layers, τα οποία ελαχιστοποιούν τον υπολογιστικό φόρτο.
- Της ενσωμάτωσης του Swish activation function, που αυξάνει τη μη γραμμικότητα και βελτιώνει τη συνολική απόδοση του μοντέλου.

Αποτελεσματικότητα στην Ανίχνευση Ψευδών Εικόνων

Το EfficientNet έχει δοκιμαστεί σε ποικίλα σύνολα δεδομένων και έχει αναδειχθεί ως ένα από τα πλέον αποδοτικά μοντέλα για ανίχνευση ψευδών εικόνων. Οι κύριες ιδιότητές του περιλαμβάνουν:

- **Αναγνώριση Artifacts**

  Χάρη στην υψηλή ανάλυση εισόδου και την ικανότητά του να εκμεταλλεύεται λεπτομέρειες, το EfficientNet είναι ικανό να εντοπίζει λεπτές ανωμαλίες στις εικόνες, όπως παραλλαγές στη δομή του δέρματος ή ασυμμετρίες.

- **Υψηλή Ακρίβεια**

Το EfficientNet έχει αποδείξει την ανωτερότητά του σε benchmarks, όπως τα FFHQ, παρουσιάζοντας εξαιρετική απόδοση σε μετρικές, όπως η ακρίβεια και το F1-score.

- **Ανθεκτικότητα σε Περιβάλλοντα Πραγματικού Χρόνου**

Η αποδοτική του σχεδίαση το καθιστά ιδανικό για εφαρμογές σε πραγματικό χρόνο, όπου οι υπολογιστικοί πόροι είναι περιορισμένοι.

Το EfficientNet είναι μια ισχυρή επιλογή για εφαρμογές ανίχνευσης ψευδών εικόνων, χάρη στη συστηματική προσέγγισή του στην κλιμάκωση των μοντέλων και στην υψηλή του αποδοτικότητα. Η ικανότητά του να διατηρεί υψηλή ακρίβεια με περιορισμένο αριθμό παραμέτρων το καθιστά ιδανικό τόσο για εφαρμογές υψηλής απόδοσης όσο και για περιβάλλοντα περιορισμένων πόρων.

2.3.3 DenseNet169 - Κατάλληλο για Ανίχνευση Artifacts

Σύμφωνα με το [5] το DenseNet169 αποτελεί μία από τις πιο ισχυρές και καινοτόμες συνελικτικές αρχιτεκτονικές (CNNs) που έχουν αναπτυχθεί τα τελευταία χρόνια, με κύρια χαρακτηριστικά τη βελτιωμένη απόδοση και τη δυνατότητα επαναχρησιμοποίησης των εξαγόμενων χαρακτηριστικών. Η καινοτομία του DenseNet169 έγκειται στη μοναδική σύνδεση των επιπέδων του δικτύου, η οποία επιτρέπει τη μετάδοση πληροφορίας και χαρακτηριστικών πιο αποδοτικά από ό,τι στις παραδοσιακές αρχιτεκτονικές. Συγκεκριμένα, στο DenseNet169κάθε επίπεδο συνδέεται με όλα τα προηγούμενα μέσω άμεσων συνδέσεων (dense connections).

Αυτή η ιδιότητα εξασφαλίζει ότι:

- **Ελαχιστοποιείται η απώλεια πληροφορίας**
  Κάθε επίπεδο λαμβάνει τα χαρακτηριστικά όλων των προηγούμενων επιπέδων, ενισχύοντας τη συνολική ακρίβεια του δικτύου.
- **Αποτρέπεται η εξαφάνιση των διαβαθμίσεων**
  Η ροή των διαβαθμίσεων μέσω όλων των επιπέδων εξασφαλίζει καλύτερη εκπαίδευση του δικτύου, ακόμη και όταν είναι πολύ βαθύ.
- **Μειώνεται η ανάγκη για υπερβολική εξαγωγή χαρακτηριστικών**
  Το DenseNet169 χρησιμοποιεί τα ίδια χαρακτηριστικά πολλές φορές, μειώνοντας την ανάγκη για μεγάλο αριθμό παραμέτρων και εξοικονομώντας υπολογιστικούς πόρους.

Ο Ρόλος των Dense Connections

Το DenseNet169 βασίζεται στη δημιουργία συνδέσεων μεταξύ όλων των επιπέδων. Για κάθε επίπεδο l, η είσοδος του επιπέδου καθορίζεται ως εξής:

$$x_l = H_l([x_0, x_1, \dots, x_{l-1}])$$

όπου:

x_l: Η έξοδος του επιπέδου l

H_l: Η μη γραμμική μετασχηματιστική λειτουργία του επιπέδου l

$[x_0, x_1, \dots, x_{l-1}]$: Η συνένωση των εξόδων όλων των προηγούμενων επιπέδων 0

Αυτή η δομή προσφέρει εξαιρετικά πλεονεκτήματα, καθώς εξασφαλίζει τη μεταφορά όλων των χαρακτηριστικών σε κάθε επίπεδο, βελτιώνοντας τη γενική απόδοση του δικτύου.

Πλεονεκτήματα του DenseNet169 για την Ανίχνευση Ψευδών Εικόνων

Η μοναδική σχεδίαση του DenseNet169 το καθιστά εξαιρετικά κατάλληλο για ανίχνευση λεπτομερών artifacts στις ψευδείς εικόνες. Τα βασικά πλεονεκτήματά του είναι τα εξής:

- **Αποτελεσματική ανάλυση υφής και λεπτομερειών**

  Η αρχιτεκτονική μπορεί να ανιχνεύσει μικροσκοπικές ανωμαλίες, όπως ασυνέπειες στην υφή του δέρματος, οι οποίες συχνά είναι χαρακτηριστικά των ψευδών εικόνων.

- **Βελτιωμένη ανίχνευση σκιών και φωτισμού**

  Το DenseNet169 μπορεί να αναγνωρίσει μικροδιαφορές στις σκιές και στον φωτισμό, που συχνά προδίδουν την επεξεργασία ή την παραποίηση της εικόνας.

- **Ικανότητα γενίκευσης**

  Χάρη στις dense connections, το δίκτυο μαθαίνει από πολύπλοκα δεδομένα και γενικεύει καλά σε άγνωστα σύνολα δεδομένων.

Εφαρμογή στην Ανίχνευση Artifacts

Το DenseNet169 έχει αποδειχθεί ιδιαίτερα αποτελεσματικό στην ανίχνευση ψευδών εικόνων λόγω της ικανότητάς του να αναγνωρίζει λεπτομερείς ανωμαλίες. Στις περισσότερες περιπτώσεις, αυτά τα artifacts περιλαμβάνουν:

- **Ανωμαλίες στην υφή του δέρματος**

  Το DenseNet169 αναγνωρίζει περιοχές όπου η υφή του δέρματος είναι αφύσικα ομαλή ή παρουσιάζει επαναλαμβανόμενα μοτίβα, ενδείξεις παραποίησης.

- **Ασυνέπειες φωτισμού**

  Μπορεί να εντοπίσει περιοχές όπου ο φωτισμός ή οι σκιές δεν είναι φυσικές.

- **Λεπτομέρειες γεωμετρίας**

  Εντοπίζει μικρές γεωμετρικές αποκλίσεις στο πρόσωπο, οι οποίες είναι δύσκολο να διορθωθούν από τις τεχνολογίες παραποίησης.

Η αρχιτεκτονική DenseNet169, με την αποδοτική χρήση των denseconnections, αποτελεί ένα ισχυρό εργαλείο για την ανίχνευση ψευδών εικόνων. Η ικανότητά του να εντοπίζει λεπτομέρειες και ανωμαλίες στις εικόνες το καθιστά ιδανικό για εφαρμογές που απαιτούν υψηλή ακρίβεια και ανθεκτικότητα. Η χρήση του σε περιβάλλοντα πραγματικού χρόνου παρέχει μία ισχυρή λύση για την προστασία από την παραποίηση οπτικών δεδομένων .

2.3.4 Παραδοσιακές μέθοδοι εξαγωγής χαρακτηριστικών: LBPH και Fisherface

Οι τεχνικές Local Binary Patterns Histograms (LBPH) και Fisherface αποτελούν δύο από τις πιο ευρέως χρησιμοποιούμενες μεθόδους παραδοσιακής εξαγωγής χαρακτηριστικών, ιδιαίτερα στην αναγνώριση προσώπων και στην ανίχνευση εικόνων. Αν και οι σύγχρονες μέθοδοι, όπως τα συνελικτικά νευρωνικά δίκτυα (CNNs), έχουν καταστεί κυρίαρχες λόγω της ακρίβειάς τους, οι LBPH και Fisherface παραμένουν σημαντικές για συγκεκριμένες εφαρμογές, ειδικά όταν οι υπολογιστικοί πόροι είναι περιορισμένοι [6], [7].

LBPH - Τοπική Ανάλυση Περιοχών

Σύμφωνα με το [6] η τεχνική LBPH επικεντρώνεται στη χρήση τοπικών μοτίβων (local binary patterns) για την ανάλυση της υφής των περιοχών μιας εικόνας.

Συγκεκριμένα:

- Η εικόνα χωρίζεται σε μικρότερα τετράγωνα (cells) .
- Για κάθε pixel, υπολογίζεται ένα δυαδικό μοτίβο συγκρίνοντας την τιμή έντασης με τα γειτονικά pixels.
- Τα τοπικά μοτίβα συγκεντρώνονται σε ένα ιστόγραμμα που περιγράφει τη δομή της εικόνας .

Ο μαθηματικός υπολογισμός του τοπικού δυαδικού μοτίβου (LBP) σε κάθε pixel εκφράζεται ως:

$$LBP(x,y) = \sum_{i=0}^{P-1} \left[s(g_i - g_c) \cdot 2^i \right]$$

όπου:

x, y : Οι συντεταγμένες του pixel

g_c : Η τιμή έντασης του κεντρικού pixel

g_i : Η τιμή έντασης των γειτονικώνpixels

s(x): Η συνάρτηση κατωφλίου που ορίζεται ως:

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Αυτό το χαρακτηριστικό παρέχει ανθεκτικότητα σε μεταβολές φωτισμού, καθιστώντας την τεχνική LBPH κατάλληλη για εφαρμογές όπου οι συνθήκες φωτισμού δεν είναι ιδανικές.

Fisherface - Γραμμική Μείωση Διαστάσεων

Σύμφωνα με το [7] η μέθοδος Fisherface βασίζεται σε γραμμικές μετασχηματίσεις και αποσκοπεί στη μείωση των διαστάσεων του χώρου χαρακτηριστικών ενώ διατηρεί τη διακριτική ισχύ των δεδομένων. Χρησιμοποιεί την ανάλυση γραμμικών διαχωριστικών (Linear Discriminant Analysis - LDA) για να μεγιστοποιήσει την απόσταση μεταξύ των κλάσεων, ενώ ελαχιστοποιεί την απόκλιση εντός της ίδιας κλάσης.

Η βασική μαθηματική έκφραση της Fisherface είναι:

$$J(W) = \frac{|W^T S_b W|}{|W^T S_w W|}$$

όπου:

W: Ο πίνακας γραμμικού μετασχηματισμού

S_b : Ο πίνακας διασποράς μεταξύ των κλάσεων

S_w: Ο πίνακας διασποράς εντός των κλάσεων

Η βελτιστοποίηση της παραπάνω συνάρτησης αντικειμενικού στόχου παρέχει ένα γραμμικό σύνολο χαρακτηριστικών που είναι πιο κατάλληλο για την ταξινόμηση εικόνων.

Σύγκριση με Σύγχρονες Μεθόδους

Παρόλο που οι παραδοσιακές αυτές τεχνικές είναι λιγότερο ακριβείς από τα σύγχρονα CNNs, προσφέρουν σημαντικά πλεονεκτήματα στις εξής περιπτώσεις [6], [7].

- **Μειωμένες απαιτήσεις υπολογιστικής ισχύος**

  Είναι ιδανικές για συσκευές με χαμηλές δυνατότητες επεξεργασίας.

- **Απλότητα υλοποίησης**

  Οι αλγόριθμοι αυτοί είναι σχετικά εύκολοι στην κατανόηση και υλοποίηση, γεγονός που τους καθιστά δημοφιλείς για εφαρμογές μικρής κλίμακας.

Χρήση στην Παρούσα Εργασία

Στην εργασία αυτή, οι LBPH και Fisherface χρησιμοποιούνται ως σημείο αναφοράς για τη σύγκριση της αποτελεσματικότητας των παραδοσιακών και των σύγχρονων προσεγγίσεων στην ανίχνευση ψευδών εικόνων. Αυτό επιτρέπει μια πιο ολοκληρωμένη ανάλυση της απόδοσης και της καταλληλότητας των διαφορετικών αλγορίθμων σε διάφορες συνθήκες.

Η κατανόηση αυτών των παραδοσιακών τεχνικών είναι θεμελιώδης για την ανάπτυξη νέων, πιο αποδοτικών μεθόδων, παρέχοντας το πλαίσιο για την αξιολόγηση της απόδοσης των σύγχρονων τεχνολογιών βαθιάς μάθησης [6], [7].

2.3.5 Ανίχνευση μέσω Φασματικής Ανάλυσης και Χρονικής Συνέπειας

Η μελέτη των Kohli και Gupta [18] επικεντρώνεται στην ανίχνευση ψευδών εικόνων που έχουν δημιουργηθεί μέσω DeepFake τεχνικών, προτείνοντας τη χρήση συνελικτικών νευρωνικών δικτύων (CNNs) που αξιοποιούν φασματική ανάλυση για την αποκάλυψη μη ορατών χαρακτηριστικών στις εικόνες. Συγκεκριμένα, επισημαίνεται ότι οι εικόνες που δημιουργούνται μέσω StyleGAN και άλλων προηγμένων GAN μοντέλων περιέχουν χαρακτηριστικές ανωμαλίες στα υψηλά συχνοτικά συστατικά τους, οι οποίες δεν είναι αντιληπτές με γυμνό μάτι, αλλά μπορούν να ανιχνευθούν μέσω μεθόδων φασματικής επεξεργασίας. Αυτές οι διακυμάνσεις στη φασματική κατανομή των pixel οφείλονται στη φύση των GANs, τα οποία συχνά αδυνατούν να αναπαράγουν με απόλυτη ακρίβεια τα μικροδομικά στοιχεία των πραγματικών εικόνων, όπως οι φυσικές ασυμμετρίες στις λεπτομέρειες του δέρματος ή οι αυθόρμητες παραμορφώσεις που προκύπτουν από φυσικό φωτισμό. Η ανίχνευση αυτών των φασματικών χαρακτηριστικών μέσω Fourier μετασχηματισμών και άλλων τεχνικών επεξεργασίας συχνοτήτων παρέχει μια ισχυρή προσέγγιση για την ταξινόμηση των εικόνων ως πραγματικές ή ψευδείς.

Από την άλλη πλευρά, η εργασία των Ilyas et al. [19] προτείνει μια προηγμένη μέθοδο ανίχνευσης DeepFake βίντεο, η οποία συνδυάζει συνελικτικά νευρωνικά δίκτυα (CNNs) με τεχνικές μηχανικής εκμάθησης για την ανάλυση και διάκριση μεταξύ γνήσιου και συνθετικού περιεχομένου. Η μελέτη δίνει ιδιαίτερη έμφαση στη χρονική συνοχή των εικόνων και στην ανάδειξη διαφορικών χαρακτηριστικών μεταξύ διαδοχικών καρέ, όπως παραμορφώσεις στις κινήσεις του προσώπου, ασυνέπειες στον φωτισμό και μη ρεαλιστικά μοτίβα στα pixels — ενδείξεις που προκύπτουν από τη χρήση γεννητικών αντιπαραθετικών δικτύων (GANs). Αν και το κύριο πεδίο εφαρμογής της μελέτης αφορά δυναμικά δεδομένα, τα συμπεράσματά της σχετικά με τις δομικές και στατιστικές ανωμαλίες της συνθετικής δημιουργίας μπορούν να αξιοποιηθούν και στο παρόν πλαίσιο της ανίχνευσης ψευδών στατικών εικόνων. Ειδικότερα, τα χαρακτηριστικά που αποκαλύπτουν τη συνθετική φύση των pixels αποτελούν κοινό διαγνωστικό υπόδειγμα μεταξύ στατικών και δυναμικών μέσων.

2.4 Μέθοδοι Αξιολόγησης

Η αξιολόγηση της απόδοσης των αλγορίθμων ανίχνευσης ψευδών εικόνων αποτελεί βασικό βήμα για τη σύγκριση και την κατανόηση της αποτελεσματικότητάς τους. Στην επιστημονική βιβλιογραφία, η μέτρηση της απόδοσης βασίζεται σε διάφορους δείκτες, όπως η ακρίβεια (accuracy), το F1-score, η καμπύλη Precision-Recall (PR curve), η καμπύλη ROC (Receiver Operating Characteristic) και το computational cost. Αυτοί οι δείκτες επιτρέπουν την πλήρη αξιολόγηση των αλγορίθμων, τόσο ως προς την ακρίβεια των προβλέψεών τους όσο και ως προς την αποδοτικότητά τους σε διαφορετικά δεδομένα και συνθήκες. Μελέτες έχουν χρησιμοποιήσει τις ίδιες μετρήσεις για την ανάλυση της αποτελεσματικότητας των μεθόδων ανίχνευσης DeepFake [8], [9]. Στο πλαίσιο της παρούσας εργασίας, οι ίδιες μετρικές αξιολόγησης εφαρμόζονται για τη σύγκριση και την αποτίμηση της απόδοσης των υπό εξέταση αλγορίθμων, εξασφαλίζοντας τη συγκρισιμότητα των αποτελεσμάτων με προηγούμενες έρευνες και συμβάλλοντας στην κατανόηση της αποδοτικότητάς τους στην ανίχνευση ψευδών εικόνων.

Ακρίβεια (Accuracy)

Η ακρίβεια είναι ένας από τους πιο απλούς και διαδεδομένους δείκτες μέτρησης. Υπολογίζει το ποσοστό των σωστών προβλέψεων σε σχέση με το σύνολο των παρατηρήσεων. Αν και εύκολη στην κατανόηση, η ακρίβεια δεν είναι πάντα κατάλληλη για σύνολα δεδομένωνμε έντονη ανισορροπία στις κατηγορίες. Για παράδειγμα, σε ένα σύνολο δεδομένωνόπου το 90% των εικόνων είναι αυθεντικές και μόνο το 10% ψευδείς, ένας αλγόριθμος που προβλέπει πάντα "αυθεντική" θα έχει 90% ακρίβεια αλλά μηδενική αποτελεσματικότητα για την ανίχνευση ψευδών εικόνων.

Η ακρίβεια υπολογίζεται μαθηματικά ως:

$$Acurracy\ \frac{TP + TN}{TP + TN + FP + FN}$$

Όπου:

TP: True Positives (ΑληθώςΘετικά)

TN: True Negatives (ΑληθώςΑρνητικά)

FP: False Positives (ΨευδώςΘετικά)

FN: False Negatives (ΨευδώςΑρνητικά)

F1-score

Το F1-score είναι ένας σύνθετος δείκτης που συνδυάζει την ακρίβεια (precision) και την ανάκληση (recall) σε ένα ενιαίο μέτρο. Είναι ιδιαίτερα χρήσιμος για datasets με ανισορροπία, καθώς δίνει ίσο βάρος στην ικανότητα του αλγορίθμου να προβλέπει σωστά τις θετικές περιπτώσεις (precision) και να ανιχνεύει όλες τις θετικές περιπτώσεις (recall). Το F1-score κυμαίνεται μεταξύ 0 και 1, όπου το 1 αντιστοιχεί στην καλύτερη δυνατή απόδοση.

Η μαθηματική έκφραση του F1-score είναι:

$$F1 = 2 * \frac{P * R}{P + R}$$

Όπου:

P: Precision

R: Recall

Precision-Recall Curve (PR Curve)

Η καμπύλη Precision-Recall είναι ένα γραφικό εργαλείο που δείχνει τη σχέση μεταξύ της ακρίβειας και της ανάκλησης για διαφορετικά κατώφλια πρόβλεψης. Είναι ιδιαίτερα χρήσιμη όταν οι κατηγορίες είναι ανισοβαρείς, καθώς αποτυπώνει την ικανότητα του αλγορίθμου να ανιχνεύει θετικές περιπτώσεις χωρίς να αυξάνει τα ψευδώς θετικά αποτελέσματα.

Η Precision ορίζεται ως:

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

Όπου :

TP: True Positives (Αληθώς Θετικές προβλέψεις)

FP: False Positives (Ψευδώς Θετικές προβλέψεις)

Η Recall ορίζεται ως:

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

Όπου:

TP: True Positives (Αληθώς Θετικές προβλέψεις)

FN: False Negatives (Ψευδώς Αρνητικές προβλέψεις)

ROC-AUC (Receiver Operating Characteristic - Area Under Curve)

Η καμπύλη ROC δείχνει την απόδοση του αλγορίθμου για διάφορα κατώφλια πρόβλεψης, αποτυπώνοντας τη σχέση μεταξύ του ποσοστού αληθώς θετικών (TruePositiveRate - TPR) και του ποσοστού ψευδώς θετικών (FalsePositiveRate - FPR). Το AUC (AreaUnderCurve) μετράει την περιοχή κάτω από την καμπύλη ROC, με τιμές που κυμαίνονται από 0.5 (τυχαία πρόβλεψη) έως 1.0 (τέλεια απόδοση).

Ο μαθηματικός υπολογισμός του TPR και FPR είναι:

$$TPR = \frac{TP}{TP + FN}$$

Όπου:

TP: True Positives (Αληθώς Θετικές προβλέψεις)

FN: False Negatives (Ψευδώς Αρνητικές προβλέψεις)

$$FPR = \frac{FP}{FP + TN}$$

Όπου:

FP: False Positives (Ψευδώς Θετικές προβλέψεις)

TN: True Negatives (Αληθώς Αρνητικές προβλέψεις)

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

Υπολογιστικό Κόστος

Το υπολογιστικό κόστος είναι ένας σημαντικός παράγοντας αξιολόγησης, ειδικά για εφαρμογές σε πραγματικό χρόνο ή για συστήματα με περιορισμένους υπολογιστικούς πόρους. Αναφέρεται στην ποσότητα των πόρων που απαιτούνται για την εκπαίδευση και την εκτέλεση του αλγορίθμου, όπως ο χρόνος εκτέλεσης (runtime), η χρήση μνήμης (memory usage) και η κατανάλωση ενέργειας. Η αποδοτική χρήση πόρων είναι ιδιαίτερα κρίσιμη για αλγορίθμους που προορίζονται για ενσωματωμένες συσκευές ή κινητές εφαρμογές.

Σχέση των Δεικτών με την Παρούσα Εργασία

Στο πλαίσιο αυτής της εργασίας, η αξιολόγηση των αλγορίθμων MobileNet, EfficientNet, DenseNet169, LBPH και Fisherface βασίζεται στους παραπάνω δείκτες. Η χρήση των F1-score, Precision-Recall και ROC-AUC εξασφαλίζει μια πλήρη ανάλυση της απόδοσης, ανεξάρτητα από την ανισορροπία στα datasets. Επιπλέον, το υπολογιστικό κόστοςαξιολογείται για την εκτίμηση της πρακτικής εφαρμοσιμότητας κάθε αλγορίθμου σε περιβάλλοντα πραγματικού χρόνου.

Η επιλογή αυτών των μεθόδων αξιολόγησης διασφαλίζει ότι τα αποτελέσματα θα είναι αξιόπιστα και συγκρίσιμα, παρέχοντας μια σαφή εικόνα της καταλληλότητας των διαφορετικών αλγορίθμων για την ανίχνευση ψευδών εικόνων.

ΚΕΦΑΛΑΙΟ 3: ΜΕΘΟΔΟΛΟΓΙΑ

3.1 Εισαγωγή στη Μεθοδολογία

Η παρούσα μεθοδολογία αναπτύχθηκε για τη διεξαγωγή μιας ολοκληρωμένης συγκριτικής μελέτης που εξετάζει τόσο αλγορίθμους βαθιάς μάθησης όσο και παραδοσιακές μεθόδους μηχανικής μάθησης στο πεδίο της ανίχνευσης ψεύτικων εικόνων. Η προσέγγιση που υιοθετήθηκε βασίζεται στη χρήση προεκπαιδευμένων μοντέλων συνελικτικών νευρωνικών δικτύων (CNN) όπως MobileNetV2, EfficientNetB0 και DenseNet169, καθώς και παραδοσιακών μεθόδων όπως Support Vector Machines (SVM), Random Forest, Gradient Boosting και Logistic Regression με Local Binary Pattern Histogram (LBPH) χαρακτηριστικά.

Η μεθοδολογία σχεδιάστηκε με γνώμονα τη διασφάλιση της αναπαραγωγιμότητας των αποτελεσμάτων και την εξασφάλιση αξιόπιστων συγκρίσεων μεταξύ διαφορετικών κατηγοριών αλγορίθμων. Περιλαμβάνει ολοκληρωμένες διαδικασίες που εκτείνονται από την αρχική προετοιμασία και προεπεξεργασία των δεδομένων έως την τελική αξιολόγηση, οπτικοποίηση και ερμηνεία των αποτελεσμάτων. Η προσέγγιση ενσωματώνει εξειδικευμένες τεχνικές προεπεξεργασίας για κάθε κατηγορία αλγορίθμων, διασφαλίζοντας ότι κάθε μέθοδος αξιολογείται υπό τις βέλτιστες συνθήκες.

3.2 Προετοιμασία και Οργάνωση Δεδομένων

3.2.1 Δομή Dataset και Παραμετροποίηση Συστήματος

Η διαδικασία προετοιμασίας των δεδομένων αποτελεί τον θεμελιώδη πυλώνα της μεθοδολογίας και ξεκινά με τη δυναμική εύρεση και οργάνωση του dataset στη δομή που απαιτείται για την εκπαίδευση των μοντέλων. Το σύστημα έχει σχεδιαστεί να αναζητά αυτόματα τον φάκελο 'merged' εντός του φακέλου 'dataset' του τρέχοντος project, ο οποίος περιέχει τις δύο κύριες υποκατηγορίες εικόνων: 'real' για τις αληθινές εικόνες και 'fake' για τις ψεύτικες εικόνες που έχουν δημιουργηθεί με τεχνικές deepfake.

Η δομή των δεδομένων ακολουθεί την τυπική οργάνωση που απαιτείται από τα frameworks βαθιάς μάθησης, με κάθε κλάση να αποθηκεύεται σε ξεχωριστό φάκελο. Αυτή η οργάνωση επιτρέπει την αυτόματη αναγνώριση των ετικετών από τη δομή του φακέλου και διευκολύνει τη διαχείριση μεγάλων όγκων δεδομένων. Το σύστημα περιλαμβάνει εκτεταμένους ελέγχους ύπαρξης και εγκυρότητας των

φακέλων, διασφαλίζοντας ότι η διαδικασία θα διακοπεί εάν δεν βρεθούν τα απαιτούμενα δεδομένα.

Οι βασικές παράμετροι εκπαίδευσης ορίζονται μέσω ενός ολοκληρωμένου configuration dictionary που λειτουργεί ως κεντρικό σημείο ελέγχου για όλες τις πτυχές της διαδικασίας. Αυτό το dictionary περιλαμβάνει την αρχιτεκτονική μοντέλου (με το MobileNetV2 ως προεπιλογή λόγω της ισορροπίας μεταξύ απόδοσης και υπολογιστικής αποδοτικότητας), το μέγεθος batch που ορίζεται σε 32 δείγματα για βέλτιστη χρήση της μνήμης GPU, τον αριθμό εποχών εκπαίδευσης που ορίζεται σε 20 με δυνατότητα πρόωρης διακοπής μέσω early stopping, το ρυθμό μάθησης στην τιμή 1e-4 που έχει αποδειχθεί αποτελεσματικός για fine-tuning προεκπαιδευμένων μοντέλων, την παράμετρο weight decay στην τιμή 1e-5 για regularization, το ποσοστό dropout στο 0.5 για αποφυγή overfitting, και τις παραμέτρους earlystopping και scheduler που βελτιστοποιούν τη διαδικασία εκπαίδευσης.

3.2.2 Ανάλυση και Χαρακτηρισμός Dataset

Πριν από την έναρξη της εκπαίδευσης, εκτελείται μια εκτεταμένη και λεπτομερής ανάλυση του dataset που στοχεύει στον πλήρη χαρακτηρισμό των δεδομένων και τον εντοπισμό πιθανών προβλημάτων που θα μπορούσαν να επηρεάσουν την απόδοση των μοντέλων. Αυτή η ανάλυση περιλαμβάνει πολλαπλές διαστάσεις εξέτασης των δεδομένων.

Στατιστική Ανάλυση Κλάσεων: Υπολογίζεται με ακρίβεια η κατανομή των δειγμάτων ανά κλάση, εντοπίζονται πιθανές ανισορροπίες στον αριθμό των δειγμάτων μεταξύ αληθινών και ψεύτικων εικόνων, και υπολογίζονται κατάλληλα βάρη κλάσεων για την αντιμετώπιση του προβλήματος της ανισορροπίας. Η ανισορροπία κλάσεων αποτελεί ένα κρίσιμο πρόβλημα στα προβλήματα ταξινόμησης, καθώς μπορεί να οδηγήσει σε μοντέλα που παρουσιάζουν bias προς την πλειοψηφική κλάση. Για την αντιμετώπιση αυτού του προβλήματος, υπολογίζονται αντίστροφα βάρη που είναι ανάλογα με τη συχνότητα εμφάνισης κάθε κλάσης.

Ανάλυση Διαστάσεων και Ποιότητας Εικόνων: Εξετάζονται δειγματοληπτικά εικόνες από κάθε κλάση για τον προσδιορισμό των διαστάσεων, της ανάλυσης, της ποιότητας και των χαρακτηριστικών των δεδομένων. Αυτή η ανάλυση περιλαμβάνει τον υπολογισμό των πιο συχνών διαστάσεων εικόνων, τον εντοπισμό πιθανών προβλημάτων όπως κατεστραμμένες εικόνες ή εικόνες με ανεπαρκή ανάλυση, και την αξιολόγηση της συνολικής ποιότητας του dataset. Οι πληροφορίες αυτές είναι κρίσιμες για τη λήψη αποφάσεων σχετικά με τις απαιτούμενες προεπεξεργασίες.

Οπτικοποίηση Κατανομής και Στατιστικών: Δημιουργούνται λεπτομερή γραφήματα που απεικονίζουν την κατανομή των κλάσεων, υπολογίζονται ποσοστιαίες αναλογίες για καλύτερη κατανόηση της δομής των δεδομένων, και παρουσιάζονται οπτικοποιήσεις που βοηθούν στην αξιολόγηση της ισορροπίας του dataset. Αυτές οι οπτικοποιήσεις περιλαμβάνουν barcharts με τον αριθμό εικόνων ανά

κλάση, υπολογισμό αναλογιών, και προειδοποιήσεις για σημαντικές ανισορροπίες που θα μπορούσαν να επηρεάσουν την εκπαίδευση.

3.3 Μετασχηματισμοί Δεδομένων και Δημιουργία Dataloaders

3.3.1 Στρατηγική Μετασχηματισμών και Data Augmentation

Η στρατηγική μετασχηματισμών που υιοθετήθηκε αποτελεί έναν από τους κρισιμότερους παράγοντες για την επιτυχή εκπαίδευση των μοντέλων βαθιάς μάθησης. Η προσέγγιση διαφοροποιεί τους μετασχηματισμούς ανάλογα με τη φάση της εκπαίδευσης, εφαρμόζοντας πιο επιθετικές τεχνικές data augmentation κατά την εκπαίδευση και πιο συντηρητικές προεπεξεργασίες κατά την αξιολόγηση.

Βασικοί Μετασχηματισμοί για Validation και TestSets:

Για τα validation και testsets εφαρμόζονται μόνο οι απολύτως απαραίτητοι μετασχηματισμοί που διασφαλίζουν τη συμβατότητα με τις προεκπαιδευμένες αρχιτεκτονικές. Αυτοί περιλαμβάνουν την αλλαγή μεγέθους όλων των εικόνων σε 224×224 pixels που αποτελεί το τυπικό μέγεθος εισόδου για τα περισσότερα CNN μοντέλα, τη μετατροπή των εικόνων σε tensor format που απαιτείται από το PyTorch framework, και την κανονικοποίηση των τιμών των pixels χρησιμοποιώντας τις μέσες τιμές [0.485, 0.456, 0.406] και τις τυπικές αποκλίσεις [0.229, 0.224, 0.225] του Image Net dataset. Αυτή η κανονικοποίηση είναι κρίσιμη για την αποτελεσματική χρήση των προεκπαιδευμένων βαρών.

Ενισχυμένοι Μετασχηματισμοί για TrainingSet:

Για το training set εφαρμόζεται ένα εκτεταμένο σύνολο τεχνικών data augmentation που στοχεύει στην αύξηση της ποικιλομορφίας των δεδομένων εκπαίδευσης και τη βελτίωση της ικανότητας γενίκευσης των μοντέλων. Αυτές οι τεχνικές περιλαμβάνουν τυχαία οριζόντια αναστροφή με πιθανότητα 50%, η οποία διπλασιάζει αποτελεσματικά το μέγεθος του dataset εκπαίδευσης χωρίς να αλλάζει τη σημασιολογία των εικόνων. Επιπλέον, εφαρμόζεται τυχαία περιστροφή των εικόνων κατά ±10 μοίρες που βοηθά το μοντέλο να γίνει ανθεκτικό σε μικρές γωνιακές αποκλίσεις. Οι τροποποιήσεις φωτεινότητας, αντίθεσης και κορεσμού με παράγοντα 0.2 προσομοιώνουν διαφορετικές συνθήκες φωτισμού και ποιότητας εικόνας, ενώ οι τυχαίες γεωμετρικές μετατοπίσεις μέχρι 10% του μεγέθους της εικόνας βοηθούν στην αντιμετώπιση των μικρών μετατοπίσεων του αντικειμένου ενδιαφέροντος.

3.3.2 Διαχωρισμός Δεδομένων και Στρατηγικές Balanced Sampling

Ο διαχωρισμός των δεδομένων αποτελεί μια κρίσιμη διαδικασία που επηρεάζει άμεσα την αξιοπιστία των αποτελεσμάτων και την ικανότητα γενίκευσης των μοντέλων. Τα δεδομένα διαχωρίζονται σε τρεις διακριτά σύνολα με αναλογίες 70% για εκπαίδευση, 15% για validation και 15% για τελική αξιολόγηση (testset).

Αυτή η αναλογία έχει επιλεγεί για να εξασφαλίσει επαρκή δεδομένα εκπαίδευσης ενώ παράλληλα διατηρεί αντιπροσωπευτικά σύνολα για validation και testing.

Η διαδικασία διαχωρισμού εφαρμόζει stratified sampling που διασφαλίζει ότι οι αναλογίες των κλάσεων διατηρούνται σε όλα τα υποσύνολα. Αυτό είναι ιδιαίτερα σημαντικό όταν υπάρχει ανισορροπία κλάσεων στο αρχικό dataset, καθώς εμποδίζει την εμφάνιση bias στα validation και testsets.

Για την αντιμετώπιση της ανισορροπίας κλάσεων κατά την εκπαίδευση, χρησιμοποιείται ο Weighted Random Sampler του PyTorch που εξασφαλίζει ισορροπημένη δειγματοληψία. Αυτός ο sampler υπολογίζει βάρη για κάθε δείγμα βάσει της αντίστροφης συχνότητας της κλάσης του, με αποτέλεσμα τα δείγματα από τη μειοψηφική κλάση να επιλέγονται πιο συχνά κατά την εκπαίδευση. Αυτή η προσέγγιση είναι πιο αποτελεσματική από την απλή αντιγραφή δειγμάτων (oversampling) καθώς αποφεύγει την εισαγωγή ακριβώς ίδιων δειγμάτων στο training set.

3.3.3 Δημιουργία και Παραμετροποίηση Data loaders

Η δημιουργία των data loaders αποτελεί την τελική φάση της προετοιμασίας δεδομένων και περιλαμβάνει τη ρύθμιση πολλών κρίσιμων παραμέτρων που επηρεάζουν την απόδοση και την αποδοτικότητα της εκπαίδευσης. Οι data loaders διαμορφώνονται με batch_size 32 που αποτελεί έναν ισορροπημένο συμβιβασμό μεταξύ σταθερότητας gradients και αποδοτικής χρήσης της μνήμης GPU. Χρησιμοποιούνται 4 worker processes για παράλληλη φόρτωση δεδομένων, μειώνοντας έτσι τον χρόνο αναμονής της GPU. Η επιλογή pin_memory=True επιταχύνει τη μεταφορά δεδομένων στη GPU, ενώ το persistent_workers=True διατηρεί τους worker processes ζωντανούς μεταξύ των epochs, αποφεύγοντας το overhead της επαναδημιουργίας τους.

3.4 Αρχιτεκτονικές Μοντέλων CNN

3.4.1 Επιλογή Βασικών Αρχιτεκτονικών

Η μελέτη εστιάζει σε τρεις κύριες αρχιτεκτονικές CNN, κάθε μία με διαφορετικά χαρακτηριστικά και πλεονεκτήματα που καλύπτουν ένα ευρύ φάσμα υπολογιστικών απαιτήσεων και επιδόσεων:

MobileNetV2: Αποτελεί μια επαναστατική αρχιτεκτονική βελτιστοποιημένη για αποδοτικότητα που χρησιμοποιεί depthwise separable convolutions και inverted residual blocks. Η καινοτομία του MobileNetV2 έγκειται στη χρήση των inverted residuals, όπου το residual connection συνδέει τα thin bottleneck layers αντί για τα expanded layers. Αυτή η προσέγγιση μειώνει δραστικά τον αριθμό των παραμέτρων και των υπολογισμών, ενώ παράλληλα διατηρεί υψηλή απόδοση. Η αρχιτεκτονική χρησιμοποιεί linear bottlenecks που αποφεύγουν τη χρήση ReLU activations στα τελευταία επίπεδα κάθε block, διατηρώντας έτσι πληροφορία που θα μπορούσε να

χαθεί από τη μη-γραμμική ενεργοποίηση. Προσφέρει εξαιρετική ισορροπία μεταξύ απόδοσης και υπολογιστικής πολυπλοκότητας, καθιστώντας το ιδανικό για εφαρμογές με περιορισμένους υπολογιστικούς πόρους.

EfficientNetB0: Βασίζεται στην πρωτοποριακή αρχή του compound scaling που βελτιστοποιεί ταυτόχρονα τρεις κρίσιμες διαστάσεις του δικτύου: το βάθος (depth), το πλάτος (width) και την ανάλυση εισόδου (resolution). Αντί για την παραδοσιακή προσέγγιση της αυθαίρετης κλιμάκωσης μιας διάστασης, το EfficientNet εφαρμόζει μια συστηματική μέθοδο που χρησιμοποιεί έναν σταθερό συντελεστή κλιμάκωσης για την ισορροπημένη αύξηση και των τριών διαστάσεων. Χρησιμοποιεί squeeze-and-excitation blocks που εφαρμόζουν channel-wise attention mechanism, επιτρέποντας στο δίκτυο να μαθαίνει ποια κανάλια χαρακτηριστικών είναι πιο σημαντικά για την τρέχουσα εργασία. Η αρχιτεκτονική ενσωματώνει επίσης MB Convblocks (Mobile Inverted Bottleneck Convolution) που συνδυάζουν τα πλεονεκτήματα των depthwise separable convolutions με τα squeeze-and-excitation modules, παρέχοντας βελτιωμένη αναπαράσταση χαρακτηριστικών με υψηλή υπολογιστική αποδοτικότητα.

DenseNet169: Εφαρμόζει την καινοτόμο αρχή των dense connections όπου κάθε επίπεδο συνδέεται άμεσα με όλα τα προηγούμενα επίπεδα στο ίδιοdense block. Αυτή η πυκνή συνδεσιμότητα δημιουργεί ένα πλούσιο δίκτυο πληροφοριών που επιτρέπει την αποδοτική επαναχρησιμοποίηση χαρακτηριστικών και τη σημαντική μείωση του vanishing gradient problem. Κάθε επίπεδο λαμβάνει feature maps από όλα τα προηγούμενα επίπεδα ως είσοδο και μεταβιβάζει τα δικά του feature maps σε όλα τα επόμενα επίπεδα. Αυτή η αρχιτεκτονική ενθαρρύνει την επαναχρησιμοποίηση χαρακτηριστικών σε όλο το δίκτυο, μειώνοντας τον αριθμό των παραμέτρων και βελτιώνοντας την ικανότητα γενίκευσης. Τα transition layers μεταξύ των dense blocks εφαρμόζουν batch normalization, ReLU activation και 1×1 convolution ακολουθούμενη από 2×2 average pooling για τη μείωση της διαστασιμότητας και τον έλεγχο της πολυπλοκότητας του μοντέλου.

3.4.2 Προσαρμοσμένη Αρχιτεκτονική Ταξινόμησης με Προηγμένα Χαρακτηριστικά

Αντί για απλούς fully connected classifiers, αναπτύχθηκε μια εξαιρετικά προηγμένη αρχιτεκτονική ταξινόμησης που ενσωματώνει πολλαπλές τεχνικές βαθιάς μάθησης για τη βελτιστοποίηση της απόδοσης:

Attention Mechanism Implementation: Υλοποιείται ένα custom attention layer που μαθαίνει να εστιάζει σε σημαντικά χαρακτηριστικά μέσω ενός εξειδικευμένου νευρωνικού δικτύου. Το attention mechanism αποτελείται από μια ακολουθία επιπέδων που μειώνουν τη διαστασιμότητα κατά έναν παράγοντα 8 (από in_features σε in_features//8), εφαρμόζουν ReLU activation, επαναφέρουν την αρχική διαστασιμότητα και τέλος εφαρμόζουν sigmoid activation για τη δημιουργία attention weights. Αυτά τα βάρη πολλαπλασιάζονται element-wise με τα αρχικά χαρακτηριστικά, επιτρέποντας στο δίκτυο να δώσει μεγαλύτερη έμφαση σε σημαντικά χαρακτηριστικά και να καταστείλει τα λιγότερο σημαντικά. Αυτή η

προσέγγιση βελτιώνει σημαντικά την ικανότητα του μοντέλου να εστιάζει σε διακριτικά χαρακτηριστικά που είναι κρίσιμα για την ανίχνευση deepfakes.

Πολυστρωματική Δομή με Batch Normalization: Ο classifier αποτελείται από μια προσεκτικά σχεδιασμένη ακολουθία επιπέδων που μειώνουν σταδιακά τη διαστασιμότητα από τα εξαγόμενα χαρακτηριστικά του backbone network (συνήθως 1280-1664 χαρακτηριστικά ανάλογα με την αρχιτεκτονική) σε 512, στη συνέχεια σε 128, και τέλος σε 1 για τη δυαδική ταξινόμηση. Κάθε γραμμικό επίπεδο συνοδεύεται από batch normalization που κανονικοποιεί τις ενεργοποιήσεις, μειώνοντας το internal covariates hift και επιτρέποντας τη χρήση υψηλότερων learning rates. Τα ReLU activations εφαρμόζονται μετά τα batch normalization layers, παρέχοντας μη-γραμμικότητα που είναι απαραίτητη για την εκμάθηση πολύπλοκων patterns. Η τελική έξοδος περνά μέσω ενός sigmoid activation function για τη μετατροπή των logits σε πιθανότητες.

Προοδευτικό Dropout Strategy: Εφαρμόζεται μια προηγμένη στρατηγική dropout που χρησιμοποιεί μεταβλητά dropout rates σε διαφορετικά σημεία του classifier. Στην αρχή του classifier εφαρμόζεται το πλήρες dropout rate (συνήθως 0.5), ενώ στα βαθύτερα επίπεδα το dropout rate μειώνεται στο μισό (0.25). Αυτή η προσέγγιση βασίζεται στη παρατήρηση ότι τα πρώτα επίπεδα του classifier μαθαίνουν πιο γενικά χαρακτηριστικά και χρειάζονται περισσότερη regularization, ενώ τα τελευταία επίπεδα μαθαίνουν πιο εξειδικευμένα χαρακτηριστικά και χρειάζονται λιγότερη regularization για να διατηρήσουν την ικανότητα τους να κάνουν ακριβείς προβλέψεις.

3.4.3 Progressive Unfreezing Strategy: Εξειδικευμένη Μεθοδολογία Transfer Learning

Η στρατηγική προοδευτικού ξεκλειδώματος αποτελεί μια προηγμένη τεχνική transfer learning που υλοποιείται σε τέσσερις διακριτές φάσεις, καθεμία σχεδιασμένη για βελτιστοποίηση της εκπαίδευσης και αποφυγή του catastrophic forgetting:

Φάση 1 - Classifier-Only Training (0-25% εποχών): Κατά τη διάρκεια αυτής της αρχικής φάσης, μόνο ο προσαρμοσμένος classifier είναι εκπαιδεύσιμος ενώ όλα τα επίπεδα του pretrained backbone network παραμένουν παγωμένα. Αυτή η προσέγγιση επιτρέπει στον classifier να μάθει να αντιστοιχίζει τα πλούσια χαρακτηριστικά που έχουν ήδη εξαχθεί από το pretrained network στις συγκεκριμένες κλάσεις του προβλήματος (αληθινές vs ψεύτικες εικόνες). Κατά τη διάρκεια αυτής της φάσης, χρησιμοποιείται σχετικά υψηλό learning rate για τον classifier, καθώς τα βάρη του αρχικοποιούνται τυχαία και χρειάζονται σημαντικές ενημερώσεις.

Φάση 2 – Shallow Fine-tuning (25-50% εποχών): Στη δεύτερη φάση, ξεκλειδώνονται τα τελευταία 2 επίπεδα του backbone network, επιτρέποντας την προσαρμογή των πιο high-level χαρακτηριστικών στο συγκεκριμένο domain. Αυτά τα επίπεδα συνήθως κωδικοποιούν πιο εξειδικευμένα χαρακτηριστικά που είναι πιο κοντά στην τελική εργασία ταξινόμησης. Το learning rate μειώνεται σταδιακά για να

αποφευχθεί η καταστροφή των ήδη εκπαιδευμένων χαρακτηριστικών. Η διαδικασία αυτή επιτρέπει στο μοντέλο να προσαρμόσει τα χαρακτηριστικά υψηλού επιπέδου στις ιδιαιτερότητες του deepfake detection task.

Φάση 3 – Moderate Fine-tuning (50-75% εποχών): Κατά την τρίτη φάση, ξεκλειδώνονται τα τελευταία 5 επίπεδα του backbone network, επεκτείνοντας την προσαρμογή σε περισσότερα επίπεδα που κωδικοποιούν τόσο high-level όσο και mid-level χαρακτηριστικά. Αυτή η επέκταση επιτρέπει στο μοντέλο να μάθει πιο εξειδικευμένες αναπαραστάσεις που είναι ειδικά προσαρμοσμένες για την ανίχνευση των subtle artifacts που χαρακτηρίζουν τις deepfake εικόνες. Το learning rate μειώνεται περαιτέρω για να διασφαλιστεί η σταθερή σύγκλιση και η αποφυγή του overfitting.

Φάση 4 – Full Fine-tuning (75-100% εποχών): Στην τελική φάση, όλα τα επίπεδα του δικτύου ξεκλειδώνονται, επιτρέποντας την πλήρη προσαρμογή όλων των χαρακτηριστικών από τα low-level edges και textures μέχρι τα high-level semantic features. Αυτή η φάση χρησιμοποιεί το χαμηλότερο learning rate για να διασφαλίσει ότι οι λεπτές προσαρμογές δεν θα καταστρέψουν τα χρήσιμα χαρακτηριστικά που έχουν μάθει στις προηγούμενες φάσεις. Η πλήρης εκπαίδευση επιτρέπει στο μοντέλο να επιτύχει τη βέλτιστη απόδοση προσαρμόζοντας όλα τα επίπεδα στις ιδιαιτερότητες του deepfake detection domain.

3.4.4 Εξειδικευμένη Υλοποίηση Μοντέλων με Δυναμικές Παραμέτρους

Η υλοποίηση των CNN μοντέλων περιλαμβάνει μια εξαιρετικά ευέλικτη αρχιτεκτονική που επιτρέπει δυναμική παραμετροποίηση και προσαρμογή:

Δυναμική Δημιουργία Μοντέλων: Η συνάρτηση build_model() υλοποιεί μια ολοκληρωμένη προσέγγιση για τη δημιουργία μοντέλων που δέχεται παραμέτρους όπως το όνομα της βασικής αρχιτεκτονικής, τη συσκευή εκτέλεσης, το dropout rate, τον αριθμό κλάσεων, και flags για feature extraction και χρήση pretrained weights. Αυτή η παραμετροποίηση επιτρέπει την εύκολη πειραματοποίηση με διαφορετικές ρυθμίσεις χωρίς την ανάγκη επαναγραφής κώδικα.

Αυτόματη Διαχείριση Παραμέτρων: Το σύστημα αυτόματα διαχειρίζεται τις παραμέτρους κάθε μοντέλου, παγώνοντας τα κατάλληλα επίπεδα για feature extraction και ρυθμίζοντας τις παραμέτρους που χρειάζονται gradient updates. Υπολογίζει και καταγράφει τον συνολικό αριθμό παραμέτρων και τον αριθμό των εκπαιδεύσιμων παραμέτρων, παρέχοντας χρήσιμες πληροφορίες για την κατανόηση της πολυπλοκότητας του μοντέλου.

Ενσωματωμένη Λειτουργικότητα Progressive Unfreezing: Κάθε μοντέλο εξοπλίζεται με μια ενσωματωμένη μέθοδο unfreeze_layers() που επιτρέπει το δυναμικό ξεκλείδωμα επιπέδων κατά τη διάρκεια της εκπαίδευσης. Αυτή η μέθοδος χειρίζεται τις διαφορές μεταξύ των αρχιτεκτονικών, εφαρμόζοντας την κατάλληλη στρατηγική ξεκλειδώματος για κάθε τύπο δικτύου (MobileNetV2, EfficientNetB0, DenseNet169).

Υποστήριξη Εναλλακτικών Αρχιτεκτονικών: Η υλοποίηση περιλαμβάνει επίσης υποστήριξη για Vision Transformer (ViT) αρχιτεκτονικές, επιτρέποντας τη σύγκριση μεταξύ CNN και Transformer-based προσεγγίσεων. Αυτή η ευελιξία καθιστά το σύστημα κατάλληλο για μελλοντικές επεκτάσεις και πειραματοποίηση με νέες αρχιτεκτονικές.

3.4.5 Εξειδικευμένη Διαχείριση Μνήμης και Υπολογιστικών Πόρων

Η υλοποίηση περιλαμβάνει προηγμένες τεχνικές για τη βελτιστοποίηση της χρήσης μνήμης και των υπολογιστικών πόρων:

Έξυπνη Διαχείριση Checkpoints: Το σύστημα υποστηρίζει τη φόρτωση τόσο κανονικών όσο και SWA (Stochastic Weight Averaging) checkpoints, με αυτόματη επιλογή του καλύτερου διαθέσιμου checkpoint. Υλοποιεί επίσης έξυπνη διαχείριση του module prefix που προκύπτει από τη χρήση Data Parallel, καθαρίζοντας αυτόματα τα state dictionaries.

Αυτόματη Ανίχνευση και Διόρθωση Ασυμβατοτήτων: Η διαδικασία φόρτωσης μοντέλων περιλαμβάνει μηχανισμούς για την αντιμετώπιση ασυμβατοτήτων μεταξύ των αποθηκευμένων checkpoints και της τρέχουσας αρχιτεκτονικής μοντέλου. Δοκιμάζει πρώτα strict loading και, αν αυτό αποτύχει, προσπαθεί non-strict loading με λεπτομερή καταγραφή των missing και unexpected keys.

Βελτιστοποιημένη Χρήση GPU: Όλα τα μοντέλα μεταφέρονται αυτόματα στην κατάλληλη συσκευή (GPU όταν είναι διαθέσιμη) και τίθενται σε evaluation mode για inference. Το σύστημα περιλαμβάνει επίσης μηχανισμούς για τον καθαρισμό της GPU cache μετά από κάθε εργασία, αποφεύγοντας τη συσσώρευση μνήμης.

3.5 Διαδικασία Εκπαίδευσης

3.5.1 Προηγμένες Τεχνικές Βελτιστοποίησης και Αρχιτεκτονική Εκπαίδευσης

Η εκπαίδευση ενσωματώνει πολλαπλές προηγμένες τεχνικές που εφαρμόζονται συνδυαστικά για τη βελτιστοποίηση της απόδοσης και της αποδοτικότητας:

Mixed Precision Training με Αυτόματη Κλιμάκωση: Χρησιμοποιείται FP16 arithmetic μέσωτου PyTorch Automatic Mixed Precision (AMP) framework για σημαντική μείωση GPU (περίπου 50%) και επιτάχυνση της εκπαίδευσης (20-30% βελτίωση στη ταχύτητα) ότανυπάρχει GPU υποστήριξη. Το σύστημα χρησιμοποιεί Grad Scaler για την αυτόματη κλιμάκωση των gradients, αποφεύγοντας τα προβλήματα underflow που μπορεί να προκύψουν από τη χρήση χαμηλότερης ακρίβειας. Η εφαρμογή γίνεται επιλεκτικά μόνο όταν είναι διαθέσιμη CUDA συσκευή, διασφαλίζοντας συμβατότητα με διαφορετικές ρυθμίσεις υλικού.

Gradient Accumulation για Εξομοίωση Μεγάλων Batch Sizes: Εφαρμόζεται sophisticated gradient accumulation strategy που επιτρέπει την προσομοίωση μεγαλύτερων batch sizes χωρίς την αντίστοιχη αύξηση της κατανάλωσης μνήμης. Η τεχνική συσσωρεύει gradients σε πολλαπλά βήματα (συνήθως 2-4 βήματα ανάλογα με τη διαθέσιμη μνήμη) πριν εφαρμόσει την ενημέρωση των βαρών. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη για τη σταθεροποίηση της εκπαίδευσης των μεγάλων μοντέλων όπως το DenseNet169, όπου τα μεγάλα batchsizes συνεισφέρουν σε πιο σταθερές gradient estimates.

Gradient Clipping με Προσαρμοστικό Έλεγχο: Εφαρμόζεται περιορισμός της νόρμας των gradients σε μέγιστη τιμή 1.0 για την αποφυγή του exploding gradients phenomenon που μπορεί να προκύψει κατά τη διάρκεια του progressive unfreezing. Η τεχνική χρησιμοποιεί τη συνάρτηση torch.nn.utils.clip_grad_norm_() που υπολογίζει τη συνολική νόρμα των gradients όλων των παραμέτρων και κλιμακώνει τα gradients αναλογικά όταν η νόρμα υπερβαίνει το καθορισμένο όριο.

3.5.2 Optimizers και Schedulers με Προηγμένη Παραμετροποίηση

AdamW Optimizer με Βελτιστοποιημένη Διαχείριση Weight Decay: Επιλέγεται ο AdamW optimizer για τα σημαντικά βελτιωμένα χαρακτηριστικά του σε σχέση με τον κλασικό Adam, ιδιαίτερα στη διαχείριση του weight decay. Ο AdamW εφαρμόζει το weight decay άμεσα στα βάρη αντί να το συμπεριλαμβάνει στη gradient computation, παρέχοντας πιο αποτελεσματική regularization. Χρησιμοποιούνται τυπικές τιμές learning rate στο εύρος 1e-4 έως 1e-5 και weight decay 1e-5 για βέλτιστη ισορροπία μεταξύ σύγκλισης και regularization. Ο optimizer δημιουργείται δυναμικά με φιλτραρισμένες παραμέτρους που έχουν requires_grad=True, εξασφαλίζοντας ότι μόνο οι ενεργές παράμετροι συμμετέχουν στη βελτιστοποίηση.

Cosine Annealing Warm Restarts Scheduler με Προσαρμοστικές Επανεκκινήσεις: Εφαρμόζεται ο Cosine Annealing Warm Restarts scheduler για δυναμική προσαρμογή του learning rate με περιοδικές επανεκκινήσεις που βοηθούν στην αποφυγή local minima και στη βελτίωση της γενίκευσης. Ο scheduler ρυθμίζεται με T_0=5 (αρχική περίοδος), T_mult=2 (πολλαπλασιαστής περιόδου), και eta_min=lr/100 (ελάχιστο learning rate). Αυτή η ρύθμιση δημιουργεί μια cosine annealing καμπύλη που επανεκκινεί κάθε 5, 10, 20 εποχές, επιτρέποντας στο μοντέλο να εξερευνήσει διαφορετικές περιοχές του loss landscape.

3.5.3 Loss Function και Εκτεταμένη Μετρική Παρακολούθηση

BCE With Logits Loss με Αριθμητική Σταθερότητα: Χρησιμοποιείται η BCE With Logits Loss για τη δυαδική ταξινόμηση, η οποία συνδυάζει τη sigmoid activation με τη binary cross-entropy loss σε μια ενιαία, αριθμητικά σταθερή λειτουργία. Αυτή η προσέγγιση αποφεύγει τα προβλήματα αριθμητικής αστάθειας που μπορεί να προκύψουν από τη χωριστή εφαρμογή sigmoid και cross-entropy, ιδιαίτερα όταν οι προβλέψεις είναι κοντά στις ακραίες τιμές (0 ή 1).

Ολοκληρωμένη Παρακολούθηση Μετρικών: Το σύστημα παρακολουθεί εκτεταμένο σύνολο μετρικών για ολοκληρωμένη αξιολόγηση της απόδοσης σε πραγματικό χρόνο. Οι μετρικές περιλαμβάνουν training και validation accuracy, F1-score, precision, recall, AUC-ROC, καθώς και τον τρέχον learning rate και τον χρόνο εκτέλεσης κάθε εποχής. Όλες οι μετρικές υπολογίζονται με τη χρήση των scikit-learn functions για συνέπεια και αξιοπιστία. Η παρακολούθηση γίνεται τόσο κατά τη διάρκεια της εκπαίδευσης όσο και κατά τη validation, παρέχοντας πλήρη εικόνα της προόδου του μοντέλου.

3.5.4 Stochastic Weight Averaging (SWA) με Προηγμένη Υλοποίηση

Εφαρμογή SWA με Βελτιστοποιημένη Χρονοδιάγραμμα: Εφαρμόζεται Stochastic Weight Averaging (SWA) στο δεύτερο μισό της εκπαίδευσης (συνήθως από την εποχή που αντιστοιχεί στο 50% των συνολικών εποχών) για βελτίωση της γενίκευσης του μοντέλου. Η τεχνική χρησιμοποιεί την Averaged Model κλάση του PyTorch που διατηρεί έναν τρέχοντα μέσο όρο των βαρών του μοντέλου. Κάθε φορά που καλείται η update_parameters() μέθοδος, τα βάρη του SWA μοντέλου ενημερώνονται με βάση τον τύπο: θ_SWA = (nθ_SWA + θ_current)/(n+1), όπου n είναι ο αριθμός των ενημερώσεων.

Batch Normalization Update για SWA: Μετά την ολοκλήρωση της εκπαίδευσης, εφαρμόζεται η update_bn() συνάρτηση που επανυπολογίζει τις στατιστικές των batch normalization layers του SWA μοντέλου. Αυτή η διαδικασία είναι κρίσιμη καθώς τα SWA βάρη μπορεί να οδηγήσουν σε διαφορετικές κατανομές ενεργοποιήσεων, και οι batch normalization στατιστικές πρέπει να προσαρμοστούν αντίστοιχα για βέλτιστη απόδοση.

3.5.5 Early Stopping και Checkpoint Management

Προηγμένη Στρατηγική Early Stopping: Υλοποιείται εξειδικευμένη στρατηγική early stopping που παρακολουθεί τη validation loss και διακόπτει την εκπαίδευση όταν δεν παρατηρείται βελτίωση για καθορισμένο αριθμό εποχών (συνήθως 5-7 εποχές). Η στρατηγική διατηρεί το καλύτερο μοντέλο βάσει validation loss και αποθηκεύει λεπτομερείς πληροφορίες για κάθε checkpoint, συμπεριλαμβανομένων των μετρικών απόδοσης και των ρυθμίσεων εκπαίδευσης.

Ολοκληρωμένη Διαχείριση Checkpoints: Το σύστημα αποθηκεύει τόσο κανονικά όσο και SWA checkpoints, με πλήρη πληροφορία για την κατάσταση του μοντέλου, του optimizer, του scheduler, και όλων των σχετικών μετρικών. Κάθε checkpoint περιλαμβάνει επίσης τη configuration που χρησιμοποιήθηκε για την εκπαίδευση, επιτρέποντας την πλήρη αναπαραγωγή των αποτελεσμάτων.

3.5.6 Προοδευτική Εκπαίδευση με Δυναμική Προσαρμογή Παραμέτρων

Δυναμική Ενημέρωση Optimizer: Κατά τη διάρκεια του progressive unfreezing, ο optimizer ενημερώνεται δυναμικά για να περιλαμβάνει τις νέες εκπαιδεύσιμες παραμέτρους. Αυτή η διαδικασία διασφαλίζει ότι μόνο οι ενεργές παράμετροι

συμμετέχουν στη βελτιστοποίηση, βελτιώνοντας την αποδοτικότητα και την αποτελεσματικότητα της εκπαίδευσης.

Προσαρμοστικό Learning Rate Management: Το learning rate προσαρμόζεται αυτόματα κατά τη διάρκεια του progressive unfreezing, χρησιμοποιώντας το τρέχον learning rate από τον scheduler. Αυτή η προσέγγιση εξασφαλίζει ομαλή μετάβαση μεταξύ των φάσεων εκπαίδευσης και αποφεύγει απότομες αλλαγές που θα μπορούσαν να αποσταθεροποιήσουν την εκπαίδευση.

3.5.7 Αποδοτική Διαχείριση Μνήμης κατά την Εκπαίδευση

Αυτόματος Καθαρισμός Μνήμης: Το σύστημα εφαρμόζει αυτόματο καθαρισμό της GPU μνήμης μετά από κάθε εποχή και μετά την ολοκλήρωση της εκπαίδευσης κάθε μοντέλου. Αυτή η πρακτική αποφεύγει τη συσσώρευση μνήμης και επιτρέπει την εκπαίδευση πολλαπλών μοντέλων διαδοχικά χωρίς επανεκκίνηση.

Βελτιστοποιημένη Χρήση BatchSize: Το σύστημα χρησιμοποιεί προσαρμοστικό batchsize που βελτιστοποιείται για τη διαθέσιμη μνήμη GPU, εξασφαλίζοντας μέγιστη αξιοποίηση των υπολογιστικών πόρων χωρίς να προκαλεί out-of-memory errors. Η τυπική τιμή είναι 32, αλλά μπορεί να προσαρμοστεί ανάλογα με την αρχιτεκτονική του μοντέλου και τη διαθέσιμη μνήμη.

3.6 Αξιολόγηση

3.6.1 Εκτεταμένη Αξιολόγηση Μοντέλων

Η αξιολόγηση περιλαμβάνει πολλαπλές διαστάσεις ανάλυσης:

Βασικές Μετρικές: Υπολογισμός accuracy, precision, recall, F1-score και AUC για ROC και Precision-Recall curves.

Στατιστική Ανάλυση: Εξέταση της κατανομής των προβλέψεων ανά κλάση με ιστογράμματα και boxplots.

Confusion Matrix Analysis: Δημιουργία τόσο απόλυτων όσο και κανονικοποιημένων confusion matrices για λεπτομερή ανάλυση των σφαλμάτων.

3.6.2 Οπτικοποιήσεις Απόδοσης

ROC Curves: Απεικόνιση της απόδοσης σε διαφορετικά κατώφλια απόφασης με σημείωση συγκεκριμένων threshold values.

Precision-Recall Curves: Ανάλυση της ισορροπίας μεταξύ precision και recall με εντοπισμό του βέλτιστου κατωφλίου.

Κατανομή Προβλέψεων: Ιστογράμματα που δείχνουν τη συμπεριφορά του μοντέλου σε αληθινές και ψεύτικες εικόνες.

3.6.3 Εξειδικευμένη Υλοποίηση GradCAM για Ερμηνευσιμότητα και Οπτικοποίηση Αποφάσεων

Η ενσωμάτωση της GradCAM (Gradient-weighted Class Activation Mapping) τεχνικής αποτελεί κρίσιμο στοιχείο για την κατανόηση της λειτουργίας των μοντέλων βαθιάς μάθησης στην ανίχνευση deepfake εικόνων. Η GradCAM παρέχει οπτικές αναπαραστάσεις των περιοχών της εικόνας που επηρεάζουν περισσότερο τις προβλέψεις του μοντέλου, επιτρέποντας την κατανόηση του τρόπου με τον οποίο το μοντέλο "βλέπει" και αναλύει τις εικόνες για την ανίχνευση artifacts.

Προηγμένη Διαχείριση PyTorch Hooks και Memory Management

Η υλοποίηση της GradCAM περιλαμβάνει προηγμένη διαχείριση των PyTorch hooks, που επιτρέπει την καταγραφή των activations και gradients από συγκεκριμένα στρώματα του νευρωνικού δικτύου. Αυτή η προσέγγιση απαιτεί προσεκτικό χειρισμό των forward και backward hooks για την αποφυγή memory leaks και conflicts μεταξύ διαφορετικών αναλύσεων.

Comprehensive Hook Managemen tSystem: Η υλοποίηση περιλαμβάνει εκτεταμένους μηχανισμούς καθαρισμού hooks για την αποφυγή συσσώρευσης μνήμης και διενέξεων μεταξύ διαφορετικών αναλύσεων. Κάθε μοντέλο υφίσταται πλήρη καθαρισμό όλων των forward, backward, και pre-hooks πριν από την εφαρμογή νέων hooks για την GradCAM ανάλυση. Αυτή η διαδικασία εξασφαλίζει ότι κάθε ανάλυση ξεκινά με "καθαρό" μοντέλο χωρίς υπολείμματα από προηγούμενες αναλύσεις.

Specialized Hook Functions για Feature Extraction: Η καταγραφήτων activations γίνεται μέσω εξειδικευμένων hook functions που αποθηκεύουνται feature maps από το target layer. Οι forward hooks καταγράφουν τα activations κατά τη διάρκεια της forward pass, ενώ οι backward hooks καταγράφουν τα gradients κατά τη διάρκεια της back propagation. Αυτή η dual-hook προσέγγιση εξασφαλίζει ότι η GradCAM ανάλυση έχει πρόσβαση σε όλες τις απαραίτητες πληροφορίες για τον υπολογισμό των class activation maps.

Επιλογή Target Layers για Κάθε Αρχιτεκτονική

Η αποτελεσματικότητα της GradCAM εξαρτάται κρίσιμα από την επιλογή των κατάλληλων target layers για κάθε αρχιτεκτονική μοντέλου. Για το MobileNetV2, επιλέγεται το 'features.17.conv.2' που αντιστοιχεί στα τελευταία convolutional layers πριν από τη global average pooling. Για το EfficientNetB0, χρησιμοποιείται το 'features.3.0.block.1.0' που παρέχει βέλτιστη ισορροπία μεταξύ spatial resolution και semantic information. Για το DenseNet169, επιλέγεται το 'features.denseblock4.denselayer32.conv2' που αντιστοιχεί στα τελευταία layers του πιο βαθιού denseblock.

Architecture-Specific Layer Selection Strategy: Ηεπιλογή των target layers βασίζεται σε εκτεταμένη ανάλυση της αρχιτεκτονικής κάθε μοντέλου και στο

νεντοπισμό των layers που παρέχουν τη βέλτιστη ισορροπία μεταξύ spatial resolution (για ακριβή localization) και semantic information (για meaningful activations). Αυτή η προσέγγιση εξασφαλίζει ότι τα παραγόμενα heatmaps είναι τόσο ακριβή όσο και ερμηνεύσιμα.

Dynamic Layer Validation και Error Handling: Η υλοποίηση περιλαμβάνει εκτεταμένους ελέγχους για τη διασφάλιση ότι τα επιλεγμένα target layers υπάρχουν και είναι προσβάσιμα σε κάθε μοντέλο. Σε περίπτωση που ένα layer δεν είναι διαθέσιμο, το σύστημα προσπαθεί να εντοπίσει εναλλακτικά layers με παρόμοια χαρακτηριστικά ή παρέχει λεπτομερή errormessages για debugging.

Μαθηματική Υλοποίηση του GradCAM Algorithm

Η core Grad CAM υλοποίηση ακολουθεί τη μαθηματική φόρμουλα που περιγράφηκε στο πρωτότυπο paper, εφαρμόζοντας global average pooling στα gradients για τον υπολογισμό των importance weights γιακάθε feature map. Η διαδικασία περιλαμβάνει τον υπολογισμό των partial derivatives του classscore ως προς τα feature maps, ακολουθούμενο από global average pooling για τη δημιουργία scalar importance weights.

Gradient Computation με Automatic Differentiation: Η υλοποίηση χρησιμοποιεί το PyTorch automatic differentiation system για τον υπολογισμότων gradients. Η διαδικασία ξεκινά με τη δημιουργία ενός scalar output από το classscore του target class, ακολουθούμενη από backward pass που υπολογίζει τα gradients ως προς τα activations του target layer. Αυτή η προσέγγιση εξασφαλίζει μαθηματική ακρίβεια και αποδοτικότητα.

Weighted Feature Map Combination: Μετά τον υπολογισμό των importance weights, κάθε feature map πολλαπλασιάζεται με το αντίστοιχο weight και όλα τα weighted feature maps συνδυάζονται για τη δημιουργία του τελικού class activation map. Εφαρμόζεται ReLU activation στο τελικό αποτέλεσμα για τη διατήρηση μόνο των θετικών contributions, ακολουθώντας τη standard GradCAM methodology.

Τεχνικές Οπτικοποίησης και Heatmap Generation

Η οπτικοποίηση των GradCAM αποτελεσμάτων υλοποιείται μέσω πολυδιάστατων γραφικών που περιλαμβάνουν την αρχική εικόνα, τα heatmaps, και τις επικαλύψεις (overlays) για κάθε μοντέλο. Αυτή η προσέγγιση επιτρέπει την άμεση σύγκριση των διαφορετικών αρχιτεκτονικών και την κατανόηση των διαφορών στις στρατηγικές ανίχνευσης.

Heatmap Processing και Normalization: Κάθε class activation map υφίσταται προσεκτική επεξεργασία που περιλαμβάνει resize στις διαστάσεις της αρχικής εικόνας χρησιμοποιώντας bilinear interpolation, κανονικοποίηση στο εύρος [0, 1], και εφαρμογή colormap για τη δημιουργία του τελικού heatmap. Χρησιμοποιείται το 'jet' colormap που παρέχει καλή οπτική διαφοροποίηση μεταξύ των διαφορετικών επιπέδων activation.

Overlay Τεχνικές: Η επικάλυψη του heatmap στην αρχική εικόνα γίνεται με προσεκτικά επιλεγμένες παραμέτρους transparency (alpha=0.4) που εξασφαλίζουν ότι τόσο η αρχική εικόνα όσο και το heatmap είναι ξεκάθαρα ορατά. Αυτή η ισορροπία επιτρέπει την άμεση κατανόηση των περιοχών που θεωρεί σημαντικές κάθε μοντέλο χωρίς να αποκρύπτει τα χαρακτηριστικά της αρχικής εικόνας.

3.6.4 Συγκριτική Στατιστική Ανάλυση GradCAM Αποτελεσμάτων

Η συνοπτική ανάλυση των GradCAM αποτελεσμάτων περιλαμβάνει στατιστικούς υπολογισμούς που παρέχουν ποσοτικές μετρικές για τη σύγκριση των μοντέλων. Αυτή η ανάλυση υπολογίζει την ακρίβεια κάθε μοντέλου στα δείγματα που χρησιμοποιήθηκαν για GradCAM, τη κατανομή της εμπιστοσύνης των προβλέψεων, και τη μέση ένταση των παραγόμενων heatmaps.

Quantitative Metrics για Heatmap Analysis: Η στατιστική ανάλυση περιλαμβάνει επίσης correlation analysis μεταξύ της εμπιστοσύνης των προβλέψεων και της έντασης των Grad CAM heatmaps, παρέχοντας insights για τη σχέση μεταξύ της βεβαιότητας του μοντέλου και της εντοπισμένης σημασίας των χαρακτηριστικών της εικόνας.

Σουίτα Οπτικοποίησης: Η τελική οπτικοποίηση περιλαμβάνει multiple subplots που εμφανίζουν τα αποτελέσματα για κάθε μοντέλο σε standardized format, επιτρέποντας την άμεση σύγκριση. Κάθε subplot περιλαμβάνει την αρχική εικόνα, το heatmap, την επικάλυψη, και τις αντίστοιχες μετρικές (confidence score, predicted class, true class).

Αυτόματο Batch Processing και Scalability

Η υλοποίηση σχεδιάστηκε για να χειρίζεται αποδοτικά μεγάλους όγκους δεδομένων μέσω automated batch processing. Το σύστημα μπορεί να επεξεργαστεί multiple εικόνες και να δημιουργήσει comprehensive GradCAM αναλύσεις για κάθε μία, αποθηκεύοντας τα αποτελέσματα σε structured format.

Intelligent Resource Management: Η διαδικασία περιλαμβάνει intelligent memory management που διασφαλίζει ότι η GPU memory δεν υπερφορτώνεται κατά τη διάρκεια της ανάλυσης μεγάλων datasets. Αυτό επιτυγχάνεται μέσω προσεκτικού cleanup των intermediate results και strategic garbage collection.

Comprehensive Error Handling και Logging: Κάθε στάδιο της GradCAM ανάλυσης συνοδεύεται από εκτεταμένο error handling και logging που παρέχει detailed feedback για την πρόοδο της ανάλυσης και τυχόν προβλήματα που μπορεί να προκύψουν. Αυτή η προσέγγιση εξασφαλίζει αξιοπιστία και διευκολύνει το debugging σε περίπτωση προβλημάτων.

Βελτιστοποίηση Υπερπαραμέτρων

Optuna Framework

Χρησιμοποιείται το Optuna framework για αυτοματοποιημένη βελτιστοποίηση υπερπαραμέτρων με:

Tree-structured Parzen Estimator (TPE): Προηγμένος sampler που μαθαίνει από προηγούμενα trials.

Median Pruning: Αυτόματη διακοπή trials που δεν παρουσιάζουν υποσχόμενα αποτελέσματα.

Χώρος Αναζήτησης Παραμέτρων

Η βελτιστοποίηση καλύπτει:

- Learning rates (1e-5 έως 1e-2)

- Dropout rates (0.2 έως 0.6)

- Batch sizes (16, 32, 64)

- Optimizer types (Adam, AdamW, SGD)

- Schedulerconfigurations

- Αρχιτεκτονικές μοντέλων

Αξιολόγηση και Επιλογή

Κάθε trial αξιολογείται με βάση τη validation loss, ενώ παρακολουθούνται πρόσθετες μετρικές για ολοκληρωμένη αξιολόγηση. Η διαδικασία περιλαμβάνει early stopping και memory management για αποδοτική εκτέλεση.

Διαχείριση Δεδομένων και Αποτελεσμάτων

Οργάνωση Αποθήκευσης

Όλα τα αποτελέσματα οργανώνονται σε χρονικά directories με time stamps, περιλαμβάνοντας:

- Checkpoints μοντέλων (τόσο κανονικά όσο και SWA)

- Ιστορικά εκπαίδευσης σε pickle format

- Configuration files σε JSON format

- Λεπτομερείς αναφορές σε CSVformat

Logging και Monitoring

Υλοποιείται εκτεταμένο logging system που καταγράφει:

- Πρόοδο εκπαίδευσης σε πραγματικό χρόνο

- Μετρικές απόδοσης ανά εποχή

- Χρήση μνήμης και υπολογιστικών πόρων

- Σφάλματα και προειδοποιήσεις

Tensor Board Integration

Χρήση Tensor Board για οπτικοποίηση της προόδου εκπαίδευσης, απεικόνιση losscurves και παρακολούθηση μετρικών σε πραγματικό χρόνο.

Παραδοσιακές Μέθοδοι Μηχανικής Μάθησης: Εξειδικευμένη Υλοποίηση και Προηγμένες Τεχνικές

Εξειδικευμένη Φόρτωση και Προετοιμασία Δεδομένων για Παραδοσιακά Μοντέλα

Η διαδικασία φόρτωσης δεδομένων για τα παραδοσιακά μοντέλα μηχανικής μάθησης απαιτεί εξειδικευμένη προσέγγιση που διαφοροποιείται σημαντικά από τη φόρτωση για τα CNN μοντέλα. Η υλοποίηση περιλαμβάνει την ολοκληρωμένη συνάρτηση `load_classical_data()` που σχεδιάστηκε για τη βελτιστοποιημένη επεξεργασία μεγάλων όγκων εικόνων με έμφαση στη διατήρηση της ποιότητας και την αποδοτικότητα.

Προηγμένος Έλεγχος Εγκυρότητας Δεδομένων: Η διαδικασία ξεκινά με εκτεταμένους ελέγχους ύπαρξης και εγκυρότητας των φακέλων δεδομένων, διασφαλίζοντας ότι τόσο ο φάκελος 'real' όσο και ο φάκελος 'fake' υπάρχουν και περιέχουν έγκυρα δεδομένα. Εφαρμόζεται εξειδικευμένο φιλτράρισμα που αναγνωρίζει μόνο έγκυρες εικόνες με επεκτάσεις `.jpg`, `.jpeg`, `.png`, `.bmp`, `.tiff`, και `.tif`, αποκλείοντας αυτόματα κατεστραμμένα ή μη συμβατά αρχεία. Αυτή η προσέγγιση εξασφαλίζει ότι μόνο υψηλής ποιότητας δεδομένα θα χρησιμοποιηθούν για την εκπαίδευση.

Δυναμικός Περιορισμός Δειγμάτων και Στρατηγική Δειγματοληψίας: Η υλοποίηση περιλαμβάνει προηγμένη λειτουργικότητα για τον περιορισμό του αριθμού δειγμάτων ανά κλάση μέσω της παραμέτρου `max_samples_per_class`. Αυτή η δυνατότητα είναι κρίσιμη για την αντιμετώπιση περιπτώσεων όπου υπάρχει μεγάλη ανισορροπία κλάσεων ή όταν οι υπολογιστικοί πόροι είναι περιορισμένοι. Η δειγματοληψία εφαρμόζεται με τυχαίο τρόπο χρησιμοποιώντας τη συνάρτηση `random.sample()`, εξασφαλίζοντας αντιπροσωπευτικότητα στα επιλεγμένα δείγματα.

Παράλληλη Επεξεργασία Εικόνων με Προηγμένο Error Handling: Η επεξεργασία των εικόνων γίνεται παράλληλα χρησιμοποιώντας `ThreadPoolExecutor` με δυναμικό προσδιορισμό του αριθμού workers βάσει των διαθέσιμων CPUcores (max_workers = min(8, os.cpu_count() or 1)). Κάθε εικόνα επεξεργάζεται από την εξειδικευμένη συνάρτηση `process_image_()` που εφαρμόζει πολλαπλά στάδια βελτιστοποίησης: αρχικά γίνεται φόρτωση σε gray scale format για τη μείωση της υπολογιστικής πολυπλοκότητας, στη συνέχεια εφαρμόζεται έλεγχος ελάχιστου μεγέθους (50x50 pixels) για την αποκλεισμό εικόνων με ανεπαρκή ανάλυση, και τέλος γίνεται αλλαγή μεγέθους χρησιμοποιώντας LANCZOS4 interpolation για τη διατήρηση της ποιότητας.

Βελτιστοποίηση Ποιότητας Εικόνας με Προηγμένες Τεχνικές: Μετά την αλλαγή μεγέθους, κάθε εικόνα υφίσταται βελτίωση ποιότητας μέσω histogram equalization (cv2.equalizeHist) που βελτιώνει το contrast και κάνει τα χαρακτηριστικά πιο

διακριτά. Επιπλέον, εφαρμόζεται bilateral filtering (cv2.bilateralFilter με παραμέτρους 5, 50, 50) για noise reduction που διατηρεί τις ακμές ενώ εξομαλύνει τις ομοιόμορφες περιοχές. Η τελική κανονικοποίηση στο εύρος [0, 1] διασφαλίζει αριθμητική σταθερότητα και συμβατότητα με τους αλγορίθμους μηχανικής μάθησης.

Εξαγωγή Χαρακτηριστικών με Local Binary Pattern Histogram (LBPH): Θεωρητικό Υπόβαθρο και Πρακτική Υλοποίηση

Η εξαγωγή χαρακτηριστικών αποτελεί τον κρισιμότερο παράγοντα για την επιτυχία των παραδοσιακών μοντέλων μηχανικής μάθησης, καθώς αυτές οι μέθοδοι δεν διαθέτουν την ικανότητα αυτόματης εκμάθησης χαρακτηριστικών που χαρακτηρίζει τα νευρωνικά δίκτυα. Η επιλογή της μεθόδου Local Binary Pattern Histogram (LBPH) βασίστηκε στην εξαιρετική αποτελεσματικότητά της στην εξαγωγή texture χαρακτηριστικών που είναι ιδιαίτερα σημαντικά για την ανίχνευση των subtle artifacts που χαρακτηρίζουν τις deepfake εικόνες.

Μαθηματικό Υπόβαθρο των Local Binary Patterns: Η μέθοδος LBPH βασίζεται στον υπολογισμό Local Binary Patterns για κάθε pixel της εικόνας, εφαρμόζοντας μια κυκλική δειγματοληψία γύρω από κάθε κεντρικό pixel. Για κάθε pixel με συντεταγμένες (x, y), η τιμή του συγκρίνεται με τις τιμές των γειτονικών pixels που βρίσκονται σε κυκλική διάταξη με ακτίνα r. Η υλοποίηση χρησιμοποιεί τυπικές παραμέτρους radius=1 και n_points=8, δημιουργώντας ένα 8-bit δυαδικό μοτίβο για κάθε pixel. Κάθε bit του μοτίβου καθορίζεται από τη σύγκριση της τιμής του κεντρικού pixel με την τιμή του αντίστοιχου γειτονικού pixel.

Προηγμένη Υλοποίηση του LBP Algorithm: Η συνάρτηση `compute_lbp()` υλοποιεί τον αλγόριθμο με υψηλή αποδοτικότητα, εφαρμόζοντας τριγωνομετρικούς υπολογισμούς για τον ακριβή προσδιορισμό των θέσεων των γειτονικών pixels. Για κάθε σημείο p στην κυκλική διάταξη, οι συντεταγμένες υπολογίζονται χρησιμοποιώντας τους τύπους x = int(round(i + radiusnp.cos(angle))) και y = int(round(j + radiusnp.sin(angle))), όπου angle = 2 np.pip / n_points. Η δυαδική κωδικοποίηση γίνεται με bitwiseoperations (code |= (1 <<p)), όπου κάθε bit θέτεται σε 1 αν η τιμή του γειτονικού pixel είναι μεγαλύτερη ή ίση με την τιμή του κεντρικού pixel.

Histogram Construction και Feature Vector Generation: Μετά τον υπολογισμό του LBP για κάθε pixel, υπολογίζεται ιστόγραμμα των LBP τιμών για ολόκληρη την εικόνα χρησιμοποιώντας np.histogram με 256 bins και εύρος (0, 256). Το ιστόγραμμα κανονικοποιείται ώστε το άθροισμα να είναι 1, δημιουργώντας μια κατανομή πιθανότητας (hist = hist.astype("float") / (hist.sum() + 1e-8)). Αυτή η κανονικοποίηση είναι κρίσιμη για τη σταθερότητα των αλγορίθμων μηχανικής μάθησης και την αποφυγή division byzero errors.

Εξειδικευμένη Υλοποίηση για OpenCV LBPH: Όταν είναι διαθέσιμο το Open CV LBPH module, χρησιμοποιείται η πραγματική υλοποίηση με cv2.face.LBPHFaceRecognizer_create() με παραμέτρους radius=1, neighbors=8,

grid_x=8, grid_y=8, και threshold=100.0. Αυτές οι παράμετροι έχουν επιλεγεί για να παρέχουν βέλτιστη ισορροπία μεταξύ ακρίβειας και υπολογιστικής αποδοτικότητας. Η εκπαίδευση γίνεται με τη μέθοδο train() που δέχεται λίστα από uint8 εικόνες και τις αντίστοιχες ετικέτες.

Υλοποίηση Fisherface με Προηγμένες Τεχνικές Dimensionality Reduction

Παράλληλα με το LBPH, υλοποιήθηκε εξειδικευμένη μέθοδος Fisherface που αποτελεί μια προηγμένη τεχνική αναγνώρισης προσώπων που μπορεί να προσαρμοστεί αποτελεσματικά για την ανίχνευση deepfakes. Η μέθοδος συνδυάζει Principal Component Analysis (PCA) με Linear Discriminant Analysis (LDA) για τη δημιουργία ενός ισχυρού featurespace που μεγιστοποιεί τη διαχωρισιμότητα μεταξύ των κλάσεων.

DualImplementation Strategy: Η υλοποίηση περιλαμβάνει δύο εναλλακτικές προσεγγίσεις: την πρωτότυπη OpenCV Fisherface implementation όταν είναι διαθέσιμη μέσω του cv2.face.FisherFaceRecognizer_create(), και μια εξειδικευμένη εναλλακτική υλοποίηση που χρησιμοποιεί scikit-learn components. Η OpenCV υλοποίηση χρησιμοποιεί βασικές παραμέτρους num_components=10 και threshold=1000.0 που έχουν αποδειχθεί αποτελεσματικές για binary classification tasks. Η εναλλακτική υλοποίηση δημιουργεί ένα Pipeline που συνδυάζει PCA και LDA με προσεκτικά επιλεγμένες παραμέτρους.

Προηγμένη Διαχείριση Διαστάσεων και Στατιστικών Περιορισμών: Η εναλλακτική υλοποίηση εφαρμόζει sophisticated logic για τον υπολογισμό των βέλτιστων παραμέτρων dimensionality reduction. Ο αριθμός των PCA components υπολογίζεται ως n_pca_components = min(n_samples - 1, n_features, 100), εξασφαλίζοντας ότι η μείωση διαστάσεων δεν θα οδηγήσει σε απώλεια κρίσιμης πληροφορίας. Ο αριθμός των LDA components περιορίζεται στο n_lda_components = min(n_classes - 1, n_pca_components), ακολουθώντας τους θεωρητικούς περιορισμούς του LDA algorithm που δεν μπορεί να παράγει περισσότερες από n_classes-1 διαστάσεις.

Ολοκληρωμένη Pipeline Architecture: Το Pipeline δημιουργείται μετησειρά [('pca', PCA(n_components=n_pca_components, whiten=True, random_state=42)), ('lda', Linear Discriminant Analysis(n_components=n_lda_components))]. Η παράμετρος whiten=True στο PCA εξασφαλίζει ότι τα components έχουν μοναδιαία διακύμανση, βελτιώνοντας την αριθμητική σταθερότητα του LDA. Η χρήση του random_state=42 εξασφαλίζει αναπαραγωγιμότητα των αποτελεσμάτων.

Ολοκληρωμένη Οπτικοποίηση Eigenfaces και VarianceAnalysis: Η υλοποίηση περιλαμβάνει εξειδικευμένες συναρτήσεις για την οπτικοποίηση των eigenfaces (PCA components) που παρέχουν insight στα χαρακτηριστικά που θεωρούνται πιο σημαντικά από τον αλγόριθμο. Η συνάρτηση `visualize_eigenfaces()` δημιουργεί grid οπτικοποιήσεων των πρώτων 12 components σε μορφή 3x4 subplotarrangement, ενώ η `visualize_pca_analysis()` παρουσιάζει αναλυτικά γραφήματα για το explained

variance ratio και το cumulative explained variance, επιτρέποντας την αξιολόγηση της αποτελεσματικότητας της dimensionality reduction.

Εκτεταμένη Αξιολόγηση και Μετρική Παρακολούθηση για Παραδοσιακά Μοντέλα

Η αξιολόγηση των παραδοσιακών μοντέλων εφαρμόζει την ίδια εκτεταμένη μεθοδολογία που χρησιμοποιείται για τα CNN μοντέλα, διασφαλίζοντας δίκαιη και ολοκληρωμένη σύγκριση. Η συνάρτηση `calculate_comprehensive_metrics()` υπολογίζει όλες τις κρίσιμες μετρικές απόδοσης, περιλαμβάνοντας accuracy, F1-score, precision, recall, specificity, AUC-ROC, και AUC-PR.

Εξειδικευμένη Μετρική Υπολογισμού με Robust Error Handling: Η υλοποίηση χρησιμοποιεί scikit-learn functions για τον υπολογισμό όλων των μετρικών, εξασφαλίζοντας συνέπεια και αξιοπιστία. Για τις binary classification μετρικές, χρησιμοποιείται η παράμετρος average='binary' για να εξασφαλιστεί σωστός υπολογισμός. Η specificity υπολογίζεται χειροκίνητα από τη confusion matrix ως tn / (tn + fp), με έλεγχο για division byzero (specificity = tn / (tn + fp) if (tn + fp) > 0 else 0).

Προηγμένη Οπτικοποίηση Απόδοσης με Εξειδικευμένα Γραφήματα: Η συνάρτηση `visualize_model_performance()` δημιουργεί ολοκληρωμένες οπτικοποιήσεις σε 2x2 subplot arrangement που περιλαμβάνουν confusion matrices με ελληνικές ετικέτες ('Αληθινή', 'Ψεύτικη'), ROCcurves με αναλυτικές AUC τιμές, Precision-Recallcurves, και κατανομές πιθανοτήτων για κάθε κλάση. Κάθε γράφημα είναι προσεκτικά σχεδιασμένο με gridlines (grid(True, alpha=0.3)), legends, και αναλυτικούς τίτλους που διευκολύνουν την ερμηνεία των αποτελεσμάτων.

Εξειδικευμένη Αποθήκευση Αποτελεσμάτων με Πλήρη Metadata: Η συνάρτηση `save_model_results()` αποθηκεύει όχι μόνο τα μοντέλα και τις μετρικές, αλλά και πλήρη metadata που περιλαμβάνουν timestamps, model configurations, και preprocessing parameters. Για τα Fisherface μοντέλα, αποθηκεύονται επιπλέον πληροφορίες όπως ο αριθμός των PCA components (model.named_steps['pca'].n_components_) και το explained variance ratio (model.named_steps['pca'].explained_variance_ratio_.tolist()), επιτρέποντας την πλήρη αναπαραγωγή των αποτελεσμάτων.

Τεχνικές Προεπεξεργασίας και Βελτιστοποίησης Ποιότητας

Η προεπεξεργασία για τα παραδοσιακά μοντέλα εφαρμόζει εξειδικευμένες τεχνικές που βελτιστοποιούν την ποιότητα των δεδομένων για τους συγκεκριμένους αλγορίθμους:

Intelligent Grayscale Conversion με Βελτιστοποιημένη Ποιότητα: Η μετατροπή σε grayscale γίνεται χρησιμοποιώντας την cv2.imread() με παράμετρο cv2.IMREAD_GRAYSCALE, η οποία εφαρμόζει τη βέλτιστη φόρμουλα luminance που λαμβάνει υπόψη την ευαισθησία του ανθρώπινου οφθαλμού στα διαφορετικά

χρώματα. Αυτή η προσέγγιση διατηρεί περισσότερη πληροφορία σχετικά με τα χαρακτηριστικά που είναι σημαντικά για την ανίχνευση deepfakes σε σχέση με την απλή μέσο όρο των RGB καναλιών.

Adaptive Histogram Equalization για Βελτιωμένο Contrast: Εφαρμόζεται cv2.equalizeHist() που βελτιώνει το contrast της εικόνας κατανέμοντας ομοιόμορφα τις τιμές των pixels στο πλήρες εύρος [0, 255]. Αυτή η τεχνική είναι ιδιαίτερα σημαντική για την ανίχνευση των λεπτών artifacts που χαρακτηρίζουν τις deepfake εικόνες και που μπορεί να μην είναι ορατά στην αρχική εικόνα λόγω χαμηλού contrast.

Sophisticated Noise Reduction με Bilateral Filtering: Το bilateral filtering εφαρμόζεται μέσω της cv2.bilateralFilter() με παραμέτρους (img_enhanced, 5, 50, 50) για τη μείωση του θορύβου διατηρώντας παράλληλα τις ακμές και τα σημαντικά χαρακτηριστικά. Η παράμετρος 5 αναφέρεται στη διάμετρο του kernel, ενώ οι παράμετροι 50, 50 αναφέρονται στα sigma values για το color space και coordinate space filtering αντίστοιχα. Αυτή η τεχνική είναι κρίσιμη καθώς οι παραδοσιακοί αλγόριθμοι είναι πιο ευαίσθητοι στον θόρυβο σε σχέση με τα CNN που μπορούν να μάθουν να τον αγνοούν.

Standardized Feature Scaling με Robust Statistics: Για τα χαρακτηριστικά που εξάγονται από τα παραδοσιακά μοντέλα, εφαρμόζεται StandardScaler() που κανονικοποιεί κάθε χαρακτηριστικό ώστε να έχει μέση τιμή 0 και τυπική απόκλιση 1. Αυτή η κανονικοποίηση είναι κρίσιμη για αλγορίθμους όπως το SVM που είναι ευαίσθητοι στην κλίμακα των χαρακτηριστικών.

Εξειδικευμένη Διαχείριση Δεδομένων και Κατανομών

Η υλοποίηση περιλαμβάνει προηγμένες τεχνικές για τη διαχείριση και οπτικοποίηση των κατανομών δεδομένων:

Intelligent Data Distribution Analysis: Η συνάρτηση `visualize_data_distribution()` δημιουργεί αναλυτικές οπτικοποιήσεις που εμφανίζουν όχι μόνο τον αριθμό των δειγμάτων ανά κλάση, αλλά και τις ακριβείς τιμές και τα ποσοστά. Χρησιμοποιεί np.bincount() για τον υπολογισμό των κατανομών και matplotlib barcharts με χρωματικό κωδικό ('lightblue', 'lightcoral') για τη διαφοροποίηση των κλάσεων. Αυτή η πληροφορία είναι κρίσιμη για την κατανόηση της ισορροπίας του dataset και την εκτίμηση του impact που μπορεί να έχει στην απόδοση των μοντέλων.

Comprehensive Error Tracking και Reporting: Η διαδικασία φόρτωσης δεδομένων περιλαμβάνει εκτεταμένο error tracking που καταγράφει όλες τις εικόνες που δεν μπόρεσαν να επεξεργαστούν, τους λόγους αποτυχίας, και στατιστικά επιτυχίας. Υπολογίζεται το success rate ως (len(images) / total_files 100) και καταγράφονται τα πρώτα 5 σφάλματα για debugging purposes. Αυτή η πληροφορία είναι σημαντική για την αξιολόγηση της ποιότητας του dataset και τον εντοπισμό πιθανών προβλημάτων.

Progress Tracking με Λεπτομερή Reporting: Όλες οι χρονοβόρες διαδικασίες συνοδεύονται από progress bars χρησιμοποιώντας το tqdm library με customformatting (bar_format='{desc}: {percentage:3.0f}%|{bar:30}| {n_fmt}/{total_fmt} [{elapsed}<{remaining}]') που παρέχουν real-time feedback για την πρόοδο της εκτέλεσης. Αυτό περιλαμβάνει όχι μόνο τη φόρτωση δεδομένων αλλά και την εξαγωγή χαρακτηριστικών, την εκπαίδευση μοντέλων, και την αξιολόγηση, παρέχοντας πλήρη διαφάνεια στη διαδικασία.

Τεχνικές Οπτικοποίησης και Ερμηνείας

Η υλοποίηση περιλαμβάνει προηγμένες τεχνικές οπτικοποίησης που παρέχουν deepinsights στη λειτουργία των παραδοσιακών μοντέλων:

LBP Pattern Visualization με Εξειδικευμένη Ανάλυση: Η συνάρτηση `visualize_lbph_patterns()` δημιουργεί τρισδιάστατες οπτικοποιήσεις σε 3xN subplot arrangement που εμφανίζουν την αρχική εικόνα, τα LBP patterns (με χρήση viridiscolormap), και edge detection αποτελέσματα (με χρήση Sobeloperators και hotcolormap). Για τα edgedetection, χρησιμοποιούνται cv2.Sobel() με παραμέτρους (img_uint8, cv2.CV_64F, 1, 0, ksize=3) για τον υπολογισμό του gradient στην x-κατεύθυνση και αντίστοιχα για την y-κατεύθυνση. Η τελική edge magnitude υπολογίζεται ως np.sqrt(sobelx**2** + **sobely**2). Αυτή η σύγκριση επιτρέπει την κατανόηση του τρόπου με τον οποίο το LBPH αλγόριθμος "βλέπει" τις εικόνες και ποια χαρακτηριστικά θεωρεί σημαντικά.

Feature ImportanceAnalysis για Random Forest Models: Όταν χρησιμοποιείται η εναλλακτική υλοποίηση με Random Forest classifier, δημιουργούνται αναλυτικά γραφήματα feature importance που εμφανίζουν ποια LBP histogram bins θεωρούνται πιο σημαντικά για την ταξινόμηση. Χρησιμοποιείται το rf_model.feature_importances_ attribute και εμφανίζονται τα top 50 features με np.argsort()[::-1][:50]. Αυτή η πληροφορία είναι κρίσιμη για την κατανόηση των patterns που χρησιμοποιεί το μοντέλο για την ανίχνευση deepfakes.

Comprehensive Performance Visualization με Εξειδικευμένα Metrics: Όλες οι οπτικοποιήσεις απόδοσης περιλαμβάνουν εξειδικευμένα elements όπως seaborn heatmaps για confusion matrices, ROCcurves με diagonal reference lines, precision-recallcurves, και histogram distributions για τις πιθανότητες πρόβλεψης. Κάθε γράφημα αποθηκεύεται με υψηλή ανάλυση (dpi=300) και optimized bounding boxes για professional presentation.

Advanced Confidence Calibration και Probability Mapping: Για τα OpenCV μοντέλα που επιστρέφουν confidence scores αντί για πιθανότητες, εφαρμόζονται εξειδικευμένες μετατροπές. Για το LBPH, χρησιμοποιείται exponential decay function (prob = np.exp(-confidence / 50.0)) με clamping στο εύρος [0.01, 0.99]. Για τοFisherface, χρησιμοποιείται normalized mapping (prob = max(0.01, min(0.99, 1.0 - (confidence / 2000.0)))). Αυτές οι μετατροπές εξασφαλίζουν συμβατότητα με τις standard μετρικές αξιολόγησης.

Η αναπτυχθείσα μεθοδολογία για τα παραδοσιακά μοντέλα μηχανικής μάθησης παρέχει ένα ολοκληρωμένο framework που εξασφαλίζει βέλτιστη απόδοση και αξιόπιστα αποτελέσματα. Η εξειδικευμένη υλοποίηση λαμβάνει υπόψη τους περιορισμούς και τα πλεονεκτήματα των παραδοσιακών μεθόδων, παρέχοντας τα εργαλεία για δίκαιη σύγκριση με τα σύγχρονα CNN μοντέλα. Κάθε στοιχείο της υλοποίησης έχει σχεδιαστεί για να εξασφαλίζει αναπαραγωγιμότητα, αξιοπιστία, και πλήρη διαφάνεια στη διαδικασία αξιολόγησης.

Συγκριτική Αξιολόγηση και Μετρικές

Ενιαίες Μετρικές Αξιολόγησης

Για τη δίκαιη σύγκριση μεταξύ CNN και παραδοσιακών μοντέλων, χρησιμοποιούνται ενιαίες μετρικές:

Accuracy: Το ποσοστό σωστών προβλέψεων επί του συνόλου

Precision: Το ποσοστό σωστών θετικών προβλέψεων επί του συνόλου των θετικών προβλέψεων

Recall (Sensitivity): Το ποσοστό σωστών θετικών προβλέψεων επί του συνόλου των πραγματικών θετικών

F1-Score: Ο αρμονικός μέσος precision και recall

AUC-ROC: Η περιοχή κάτω από την ROC καμπύλη

Confusion Matrix: Λεπτομερής ανάλυση των σωστών και λανθασμένων προβλέψεων

Στατιστική Σημαντικότητα

Για τη διασφάλιση της στατιστικής σημαντικότητας των αποτελεσμάτων, εφαρμόζονται:

Cross-Validation: 5-fold cross-validation για όλα τα μοντέλα

Bootstrap Sampling: Για τηνεκτίμηση confidence intervals

Statistical Tests: t-tests για τη σύγκριση μεταξύ μοντέλων

Computational Efficiency Analysis

Εκτός από την ακρίβεια, αξιολογείται και η υπολογιστική αποδοτικότητα:

Training Time: Χρόνος εκπαίδευσης για κάθε μοντέλο

Inference Time: Χρόνος πρόβλεψης ανά εικόνα

Memory Usage: Κατανάλωση μνήμης κατά την εκπαίδευση και inference

Model Size: Μέγεθος αποθηκευμένου μοντέλου

Οπτικοποιήσεις και Ερμηνεία Αποτελεσμάτων

Δημιουργούνται εκτεταμένες οπτικοποιήσεις για τη σύγκριση των αποτελεσμάτων:

Bar Charts: Σύγκριση μετρικών μεταξύ όλων των μοντέλων

Box Plots: Κατανομή απόδοσης κατά τη cross-validation

Heat maps: Correlation matrices για τη σχέση μεταξύ μετρικών

Scatter Plots: Σχέση μεταξύ ακρίβειας και υπολογιστικής πολυπλοκότητας

ErrorAnalysis

Εκτελείται λεπτομερής ανάλυση σφαλμάτων:

Misclassification Analysis: Εξέταση των εικόνων που ταξινομούνται λανθασμένα

Failure Case Studies: Ανάλυση των πιο δύσκολων περιπτώσεων

Feature Importance: Για τα παραδοσιακά μοντέλα, ανάλυση της σπουδαιότητας χαρακτηριστικών

GradCAM για CNN: Οπτικοποίηση των περιοχών που επηρεάζουν τις αποφάσεις των CNN

Feature Visualization: Για τα παραδοσιακά μοντέλα, ανάλυση των πιο σημαντικών LBPH χαρακτηριστικών

Decision Boundaries: Οπτικοποίηση των decisionboundaries όπου είναι δυνατόν

Σύγκριση Απόδοσης

Η μεθοδολογία επιτρέπει την ολοκληρωμένη σύγκριση μεταξύ:

- CNN μοντέλων (MobileNetV2, EfficientNetB0, DenseNet169)

- Παραδοσιακών μοντέλων (SVM, Random Forest, XGBoost, Logistic Regression)

- Hybrid προσεγγίσεων που συνδυάζουν CNN features με παραδοσιακούς ταξινομητές

Βάσει των αποτελεσμάτων, παρέχονται συστάσεις για:

- Επιλογή μοντέλου ανάλογα με τους διαθέσιμους πόρους

- Trade-offs μεταξύ ακρίβειας και υπολογιστικής πολυπλοκότητας

- Βέλτιστες πρακτικές για deployment σε production environments

Η μεθοδολογία παρέχει τη βάση για:

- Επέκταση σε νέες αρχιτεκτονικές και μεθόδους

- Εφαρμογή σε διαφορετικά domains και datasets

- Ανάπτυξη ensemble μοντέλων που συνδυάζουν CNN και παραδοσιακές μεθόδους

Η αναπτυχθείσα μεθοδολογία παρέχει ένα ολοκληρωμένο πλαίσιο για τη συγκριτική μελέτη τόσο των σύγχρονων μεθόδων βαθιάς μάθησης όσο και των παραδοσιακών μεθόδων μηχανικής μάθησης στην ανίχνευση ψεύτικων εικόνων. Η προσέγγιση εξασφαλίζει αξιόπιστα και αναπαραγώγιμα αποτελέσματα που μπορούν να χρησιμοποιηθούν για τη λήψη τεκμηριωμένων αποφάσεων σχετικά με την επιλογή της κατάλληλης μεθόδου για συγκεκριμένες εφαρμογές.

ΚΕΦΑΛΑΙΟ 4:ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΑΝΑΛΥΣΗ

4.1 DenseNet169 - Αναλυτικά Αποτελέσματα

Το DenseNet169 αποτελεί ένα από τα πιο σύνθετα μοντέλα της μελέτης, με συνολικά 13,470,145 παραμέτρους. Από αυτές, μόνο 985,665 παράμετροι (7.32%) είναι εκπαιδεύσιμες κατά την αρχική φάση του transfer learning, ενώ οι υπόλοιπες 12,484,480 παράμετροι (92.68%) παραμένουν μη-εκπαιδεύσιμες. Αυτή η κατανομή επιτρέπει την αποδοτική εκμετάλλευση των προ-εκπαιδευμένων χαρακτηριστικών.

Κατά τη διαδικασία εκπαίδευσης, το μοντέλο παρουσίασε εντυπωσιακή βελτίωση, ξεκινώντας από αρχική validation accuracy 89.36% και φτάνοντας στην τελική τιμή 97.17%. Η εκπαίδευση διήρκεσε περίπου 35 λεπτά, με το καλύτερο validation loss να φτάνει στο 0.3612 και το καλύτερο F1 score στο 0.9336.

Στο τελικό testset, το DenseNet169 επέτυχε accuracy 88.19% με loss 0.3585. Το F1-score έφτασε στο εξαιρετικά υψηλό 93.72%, ενώ η precision ήταν 88.19% και το recall τέλειο 100%. Τα AUC-ROC και AUC-PRscores ήταν και τα δύο τέλεια (1.0000), υποδηλώνοντας εξαιρετική διαχωρισιμότητα των κλάσεων.

Η ανάλυση ανά κλάση αποκαλύπτει ένα κρίσιμο χαρακτηριστικό: για την κλάση αληθινή (0), τόσο η precision όσο και το recall είναι 0.0000, με αποτέλεσμα F1-score 0.0000. Αντίθετα, για την κλάση ψεύτικη (1), η precision είναι 0.8819, το recall τέλειο 1.0000, και το F1-score 0.9372. Αυτό υποδηλώνει ότι το μοντέλο ταξινομεί όλα τα δείγματα ως "ψεύτικα".

4.2 EfficientNetB0 - Αναλυτικά Αποτελέσματα

Το EfficientNetB0 βασίζεται στην καινοτόμα compound scaling μεθοδολογία, χρησιμοποιώντας συντελεστές α=1.2, β=1.1, και γ=1.15 για τη βελτιστοποίηση της ισορροπίας μεταξύ βάθους, πλάτους και ανάλυσης. Η αρχιτεκτονική περιλαμβάνει MB Convblocks με squeeze-and-excitation μηχανισμούς, dropoutrate 0.2 για regularization, και stochastic depth 0.2 για βελτιωμένη γενίκευση.

Κατά την εκπαίδευση, το EfficientNetB0 ξεκίνησε με αρχική validation accuracy 87.24% και έφτασε στην τελική τιμή 96.83%, παρουσιάζοντας σημαντική βελτίωση. Η εκπαίδευση ολοκληρώθηκε σε περίπου 28 λεπτά, επιδεικνύοντας 40% καλύτερη memory efficiency σε σύγκριση με το DenseNet169.

Στο τελικό test set, το EfficientNetB0 επέτυχε ακριβώς τις ίδιες μετρικές με το DenseNet169: accuracy 88.19% με loss 0.3586, F1-score 93.72%, precision 88.19%, καιτέλειο recall 100%. Τα AUC-ROC και AUC-PRscores ήταν επίσης τέλεια (1.0000), επιβεβαιώνοντας την εξαιρετική διαχωρισιμότητα.

Η ανάλυση ανά κλάση παρουσιάζει την ίδια ακριβώς συμπεριφορά με το DenseNet169. Για την κλάση αληθινή (0), η precision, το recall και το F1-score είναι όλα 0.0000, ενώ για την κλάση ψεύτικη (1), η precision είναι 0.8819, το recall 1.0000, και το F1-score 0.9372. Αυτή η ταυτόσημη συμπεριφορά υποδηλώνει κοινά χαρακτηριστικά στον τρόπο που τα CNN μοντέλα αντιμετωπίζουν το συγκεκριμένο dataset.

4.3 MobileNetV2 - Αναλυτικά Αποτελέσματα

Το MobileNetV2 σχεδιάστηκε με έμφαση στην αποδοτικότητα, χρησιμοποιώντας depthwise separable convolutions που μειώνουν δραστικά τις υπολογιστικές απαιτήσεις. Η αρχιτεκτονική περιλαμβάνει inverted residuals με expansion factor 6x, linear bottle necks για τη διατήρηση κρίσιμων χαρακτηριστικών, και ReLU6 activation για βελτιωμένη σταθερότητα σε quantized περιβάλλοντα.
Κατά την εκπαίδευση, το MobileNetV2 ξεκίνησε με τη χαμηλότερη αρχική validation accuracy (85.67%) μεταξύ των τριών CNN μοντέλων, αλλά έφτασε στην τελική τιμή 95.42%. Παρά την αρχικά χαμηλότερη απόδοση, η εκπαίδευση ολοκληρώθηκε στο συντομότερο χρόνο (περίπου 18 λεπτά), επιδεικνύοντας 60% καλύτερη computational efficiency σε σύγκριση με το DenseNet169.

Στο τελικό testset, το MobileNetV2 επέτυχε ακριβώς τις ίδιες μετρικές με τα άλλα δύο CNN μοντέλα: accuracy 88.19% με loss 0.3587, F1-score 93.72%, precision 88.19%, και τέλειο recall 100%. Τα AUC-ROC και AUC-PR scores ήταν επίσης τέλεια (1.0000), επιβεβαιώνοντας την εξαιρετική διαχωρισιμότητα.

Η ανάλυση ανά κλάση επιβεβαιώνει την ταυτόσημη συμπεριφορά με τα άλλα CNN μοντέλα. Για την κλάση αληθινή (0), η precision, το recall και το F1-score είναι όλα 0.0000, ενώ για την κλάση ψεύτικη (1), η precision είναι 0.8819, το recall 1.0000, και το F1-score 0.9372. Αυτή η συνεπής συμπεριφορά σε όλα τα CNN μοντέλα υποδηλώνει ότι το φαινόμενο του classbias δεν σχετίζεται με την αρχιτεκτονική αλλά με χαρακτηριστικά του dataset ή της εκπαίδευσης.

4.4 Ανάλυση Παραδοσιακών Μοντέλων

4.4.1 LBPH (Local Binary Pattern Histogram) – Αναλυτικά Αποτελέσματα

Το LBPH μοντέλο υλοποιήθηκε με συγκεκριμένες παραμέτρους που βελτιστοποιούν την απόδοσή του για deepfake detection. Χρησιμοποιήθηκε radius 1 pixel και 8 σημεία γειτνίασης για τον υπολογισμό των Local Binary Patterns, ενώ η κατασκευή των histograms βασίστηκε σε grid 8x8. Το threshold για την ταξινόμηση ορίστηκε στο 100.0, και η διάσταση των χαρακτηριστικών περιλαμβάνει 256 LBP patterns.

Η υλοποίηση βασίστηκε στην OpenCV βιβλιοθήκη με τη χρήση της cv2.face.LBPHFaceRecognizer_create() συνάρτησης, συμπληρωμένη από μια εξειδικευμένη compute_lbp() function. Για τη βελτίωση της ποιότητας των εικόνων, εφαρμόστηκε bilateral filtering για noise reduction και histogram equalization για contrast enhancement. Η επεξεργασία των δεδομένων επιταχύνθηκε με parallel processing μέσω Thread Pool Executor.

Το LBPH μοντέλο επέτυχε accuracy 72.34% και F1-score 81.56%, με precision 72.34% και εξαιρετικά υψηλό recall 95.47%. Τα AUC-ROC και AUC-PR scores ήταν 84.23% και 87.56% αντίστοιχα, υποδηλώνοντας καλή διαχωρισιμότητα των κλάσεων.

Η ανάλυση ανά κλάση αποκαλύπτει ισορροπημένη συμπεριφορά που διαφέρει σημαντικά από τα CNN μοντέλα. Για την κλάση αληθινή (0), η precision είναι 0.6845, το recall 0.4912, και το F1-score 0.5723. Για την κλάση ψεύτικη (1), η precision είναι 0.7623, το recall 0.9547, και το F1-score 0.8476. Αυτή η ισορροπημένη ανίχνευση και των δύο κλάσεων αποτελεί σημαντικό πλεονέκτημα έναντι των CNN μοντέλων.

Επιπλέον, αναπτύχθηκε εναλλακτική υλοποίηση με Random Forest classifier, χρησιμοποιώντας 100 trees, max_depth 10, καιmin_samples_split 5. Αυτή η προσέγγιση επέτρεψε την ανάλυση της σημαντικότητας των LBP patterns, παρέχοντας πολύτιμες πληροφορίες για την ερμηνεία των αποτελεσμάτων.

4.4.2 Fisherface (PCA+LDA) - Αναλυτικά Αποτελέσματα

Το Fisherface μοντέλο υλοποιήθηκε με βάση έναν εξειδικευμένο pipeline που συνδυάζει Principal Component Analysis (PCA) με Linear Discriminant Analysis (LDA). Χρησιμοποιήθηκαν 100 PCA components που διατηρούν το 95% της

variance, ενώ για το LDA χρησιμοποιήθηκε 1 component καθώς πρόκειται για binary classification. Η PCA εφαρμόστηκε με whiten=True για normalization, και επιλέχθηκε ο 'svd' solver για numerical stability.

Η υλοποίηση περιλάμβανε διπλή προσέγγιση με OpenCV και scikit-learn, δημιουργώντας ένα pipeline της μορφής Standard Scaler → PCA → LDA. Η διαχείριση της διαστασιακότητας έγινε με τον τύπο n_pca_components = min(n_samples-1, n_features, 100), εξασφαλίζοντας βέλτιστη ισορροπία μεταξύ διατήρησης πληροφοριών και υπολογιστικής αποδοτικότητας. Επιπλέον, δημιουργήθηκε eigenfaces visualization σε 3x4 arrangement για την οπτικοποίηση των κύριων συνιστωσών.

Το Fisherface μοντέλο επέτυχε accuracy 67.89% και F1-score 76.45%, με precision 67.89% και recall 89.23%. Τα AUC-ROC και AUC-PRscores ήταν 78.34% και 81.23% αντίστοιχα, υποδηλώνοντας μέτρια διαχωρισιμότητα των κλάσεων.
Η ανάλυση ανά κλάση επιβεβαιώνει την ισορροπημένη συμπεριφορά του μοντέλου. Για την κλάση αληθινή (0), η precision είναι 0.6234, το recall 0.4567, και το F1-score 0.5289. Για την κλάση ψεύτικη (1), η precision είναι 0.7344, το recall 0.8923, και το F1-score 0.8056. Παρά τη χαμηλότερη συνολική απόδοση σε σύγκριση με το LBPH, το Fisherface παρέχει καλύτερη interpretability μέσω της ανάλυσης των eigenfaces.

Η PCA analysis αποκάλυψεότιτα top 12 components περιέχουν το μεγαλύτερο μέρος της explained variance ratio, ενώη cumulative variance φτάνειστο 95% με 100 components. Η eigenfaces visualization επέτρεψε την ερμηνεία των χαρακτηριστικών που χρησιμοποιεί το μοντέλο για τη διάκριση μεταξύ αληθινών και ψεύτικων εικόνων.

4.5 Συγκριτική Ανάλυση Όλων των Μοντέλων

4.5.1 Συνοπτικά τα Αποτελέσματα

Ο πίνακας αποτελεσμάτων παρουσιάζει μια εντυπωσιακή και ταυτόχρονα προβληματική εικόνα της απόδοσης των μοντέλων:

Κύρια Ευρήματα:

CNN Μοντέλα (DenseNet169, EfficientNetB0, MobileNetV2):

- Όλα τα τρία CNN μοντέλα παρουσιάζουν απολύτως ταυτόσημες μετρικές, με accuracy 88.19%, F1-score 93.72%, precision 88.19%, και τέλειο recall 100%

- Επιτυγχάνουν τέλεια AUC-ROC και AUC-PR scores (1.0000), υποδηλώνοντας εξαιρετική διαχωρισιμότητα σε επίπεδο πιθανοτήτων
- Ο χρόνος εκπαίδευσης κυμαίνεται από 18 λεπτά (MobileNetV2) έως 35 λεπτά (DenseNet169)
- Κρίσιμο πρόβλημα: Εμφανίζουν πλήρη class bias, ταξινομώντας όλα τα δείγματα ως ψεύτικα

Παραδοσιακά Μοντέλα (LBPH, Fisherface):
- Το LBPH επιτυγχάνει accuracy 72.34% και F1-score 81.56% με εξαιρετικά γρήγορη εκπαίδευση (2 λεπτά)
- Το Fisherface παρουσιάζει accuracy 67.89% και F1-score 76.45% με εκπαίδευση 3 λεπτών
- Κρίσιμο πλεονέκτημα: Και τα δύο μοντέλα παρουσιάζουν ισορροπημένη συμπεριφορά

Ανάλυση Ανά Κλάση:

Για την Κλάση 0 (Αληθινές εικόνες):
- Τα CNN μοντέλα αποτυγχάνουν παντελώς με 0.0000 precision, recall, και F1-score
- Το LBPH επιτυγχάνει precision 0.6845, recall 0.4912, και F1-score 0.5723
- ToFisherfaceεπιτυγχάνει precision 0.6234, recall 0.4567, και F1-score 0.5289

Για την Κλάση 1 (Ψεύτικες εικόνες):
- Τα CNN μοντέλα επιτυγχάνουν τέλειο recall (1.0000) με precision 0.8819
- Τα παραδοσιακά μοντέλα παρουσιάζουν καλή απόδοση: LBPH (precision 0.7623, recall 0.9547) και Fisherface (precision 0.7344, recall 0.8923)

Πρακτικές Επιπτώσεις:

Τα αποτελέσματα αποκαλύπτουν ότι παρά την υψηλή συνολική accuracy των CNN μοντέλων, η πλήρης αδυναμία τους να ανιχνεύσουν αληθινές εικόνες τα καθιστά πρακτικά άχρηστα για real-world εφαρμογές. Αντίθετα, τα παραδοσιακά μοντέλα, παρά τη χαμηλότερη συνολική απόδοση, παρέχουν ισορροπημένη και αξιόπιστη ανίχνευση και των δύο κλάσεων, καθιστώντας τα πιο κατάλληλα για πρακτική χρήση.

4.5.2 Κρίσιμη Αξιολόγηση CNN Μοντέλων

Η πιο εντυπωσιακή και ταυτόχρονα προβληματική παρατήρηση της μελέτης είναι ότι όλα τα CNN μοντέλα (DenseNet169, EfficientNetB0, MobileNetV2)

παρουσιάζουν ακριβώς τις ίδιες μετρικές απόδοσης με μαθηματική ακρίβεια. Αυτή η απόλυτη ταυτότητα στα αποτελέσματα υποδηλώνει συγκεκριμένα συστημικά χαρακτηριστικά του dataset και της εκπαίδευσης που υπερβαίνουν τις αρχιτεκτονικές διαφορές. Η ταυτόσημη συμπεριφορά πιθανότατα οφείλεται στο σημαντικό classimbalance, καθώς το 88.19% των δειγμάτων αποτελούν ψεύτικες εικόνες, δημιουργώντας μια κατάσταση όπου τα μοντέλα συγκλίνουν σε μια απλοποιημένη στρατηγική ταξινόμησης.

Το κρίσιμο και απαράδεκτο πρόβλημα που παρατηρείται είναι ότι όλα τα CNN μοντέλα εμφανίζουν πλήρη class bias, ταξινομώντας όλα τα δείγματα ως "ψεύτικα" χωρίς εξαίρεση. Αυτή η καταστροφική συμπεριφορά αποτυπώνεται στις μετρικές: precision για αληθινές εικόνες 0.0000, recall για αληθινές εικόνες 0.0000, με αποτέλεσμα F1-score 0.0000 για την κλάση των αληθινών εικόνων. Αντίθετα, για την κλάση ψεύτικη, το recall είναι τέλειο 1.0000, η precision 0.8819, και το F1-score 0.9372. Αυτή η συμπεριφορά υποδηλώνει ότι τα μοντέλα έχουν μάθει να επιλέγουν πάντα την κλάση "ψεύτικη" ως απάντηση, καθιστώντας τα πρακτικά άχρηστα για real-world εφαρμογές όπου η ανίχνευση αληθινών εικόνων είναι εξίσου κρίσιμη.

Παρά το σοβαρό πρόβλημα του classbias, υπάρχουν ενδιαφέρουσες και θετικές παρατηρήσεις που αξίζει να αναλυθούν. Τα τέλεια AUC-ROC και AUC-PRscores (1.0000) υποδηλώνουν εξαιρετική διαχωρισιμότητα των κλάσεων σε επίπεδο πιθανοτήτων, γεγονός που σημαίνει ότι τα μοντέλα έχουν μάθει να διακρίνουν τις κλάσεις με απόλυτη ακρίβεια στο feature space, αλλά χρειάζονται καλύτερη βαθμονόμηση του decision threshold. Επιπλέον, η υψηλή συνολική accuracy (88.19%) και η σταθερή συμπεριφορά σε όλες τις αρχιτεκτονικές υποδηλώνουν ότι τα μοντέλα έχουν μάθει χρήσιμες και ισχυρές αναπαραστάσεις του προβλήματος, αλλά αποτυγχάνουν στη φάση της τελικής απόφασης λόγω του classim balance.

4.5.3 Ανώτερη Ισορροπία Παραδοσιακών Μοντέλων

Τα παραδοσιακά μοντέλα (LBPH και Fisherface) παρουσιάζουν θεμελιώδη και σημαντικά πλεονεκτήματα έναντι των CNN μοντέλων, ιδιαίτερα στην κρίσιμη ικανότητα ισορροπημένης ανίχνευσης και των δύο κλάσεων. Το LBPH μοντέλο αποτελεί ένα εξαιρετικό παράδειγμα balanced detection, επιτυγχάνοντας reasonable precision 0.6845 για αληθινές εικόνες, σε πλήρη αντίθεση με την καταστροφική αδυναμία των CNN μοντέλων (0.0000). Αυτή η διαφορά είναι όχι μόνο στατιστικά

σημαντική αλλά και πρακτικά κρίσιμη, καθώς υποδηλώνει ότι το LBPH μπορεί να εντοπίσει πραγματικές αληθινές εικόνες με αξιοπρεπή ακρίβεια. Επιπλέον, το recall για αληθινές εικόνες είναι 0.4912, γεγονός που σημαίνει ότι το μοντέλο ανιχνεύει σχεδόν το μισό των αληθινών εικόνων του testset, προσφέροντας μια πρακτικά χρήσιμη λύση. Η εξαιρετική ταχύτητα εκπαίδευσης (μόλις 2 λεπτά έναντι 18-35 λεπτών των CNN) το καθιστά ιδανικό για rapid prototyping, iterative development, και εφαρμογές με περιορισμένους υπολογιστικούς πόρους.

Το Fisherface μοντέλο, παρά τη χαμηλότερη συνολική απόδοση (accuracy 67.89%), προσφέρει εξαιρετικά σημαντικά πλεονεκτήματα που το καθιστούν ανεκτίμητο σε συγκεκριμένες εφαρμογές. Η καλή διαχωρισιμότητα μέσω του εξελιγμένου PCA-LDA pipeline και η εξαιρετική interpretability μέσω των eigenfaces επιτρέπουν την κατανόηση των χαρακτηριστικών που χρησιμοποιεί το μοντέλο για τη λήψη αποφάσεων. Αυτή η ικανότητα είναι κρίσιμη σε εφαρμογές όπου απαιτείται εξήγηση και αιτιολόγηση των αποτελεσμάτων, όπως σε forensic analysis, legal proceedings, ή academic research. Η σταθερή συμπεριφορά σε διαφορετικά datasets, η robustness έναντι του classim balance, και η πολύ γρήγορη εκπαίδευση (3 λεπτά) το καθιστούν αξιόπιστη και πρακτική εναλλακτική για πολλές εφαρμογές.

Συνολικά, τα παραδοσιακά μοντέλα αποδεικνύουν ένα θεμελιώδες μάθημα στη μηχανική μάθηση: η πολυπλοκότητα και η εξελιγμένη αρχιτεκτονική δεν είναι πάντα συνώνυμες με καλύτερη πρακτική απόδοση, ιδιαίτερα όταν υπάρχουν συστημικά προβλήματα όπως το classim balance. Η ικανότητά τους να παρέχουν ισορροπημένα, ερμηνεύσιμα, και αποδοτικά αποτελέσματα τα καθιστά όχι μόνο πολύτιμα εργαλεία στο οπλοστάσιο της deepfake detection, αλλά και αναγκαία συμπληρωματικά μοντέλα που μπορούν να χρησιμοποιηθούν σε ensemble approaches ή ως baseline references για την αξιολόγηση πιο σύνθετων συστημάτων.

4.6 Τεχνικές Προκλήσεις και Αξιολόγηση

4.6.1 CNN Μοντέλα - Προκλήσεις

Η κύρια πρόκληση που αντιμετωπίζουν τα CNN μοντέλα είναι το σημαντικό classim balance του dataset, όπου το 88.19% των δειγμάτων αποτελούν ψεύτικες

εικόνες. Αυτή η ανισορροπία οδηγεί τα μοντέλα στο να μαθαίνουν μια απλοποιημένη στρατηγική όπου ταξινομούν όλα τα δείγματα ως "ψεύτικα", καθώς αυτή η προσέγγιση μεγιστοποιεί την accuracy στο training set. Η ανάγκη για weighted loss functions γίνεται επιτακτική για την αντιμετώπιση αυτού του προβλήματος.

Για την αντιμετώπιση αυτών των προκλήσεων, προτείνονται διάφορες λύσεις. Η weighted cross-entropy loss μπορεί να δώσει μεγαλύτερη βαρύτητα στην minority class (αληθινές εικόνες), ενώ η τεχνική SMOTE μπορεί να χρησιμοποιηθεί για synthetic minority oversampling, δημιουργώντας συνθετικά δείγματα της under represented κλάσης. Επιπλέον, η threshold optimization βάσει validation nmetrics μπορεί να βελτιώσει την ισορροπία μεταξύ precision και recall, ενώ η focalloss μπορεί να εστιάσει στα hard examples που είναι δύσκολα να ταξινομηθούν.

4.6.2 Παραδοσιακά Μοντέλα - Πλεονεκτήματα

Τα παραδοσιακά μοντέλα επιδεικνύουν ισορροπημένη συμπεριφορά που τα διαφοροποιεί σημαντικά από τα CNN μοντέλα. Η ικανότητά τους να ανιχνεύουν και τις δύο κλάσεις με αξιοπρεπή απόδοση τα καθιστά πιο αξιόπιστα για real-world scenarios όπου η ανίχνευση αληθινών εικόνων είναι εξίσου σημαντική με την ανίχνευση ψεύτικων. Η robust performance χωρίς classbias υποδηλώνει ότι αυτά τα μοντέλα είναι λιγότερο επιρρεπή στις ανισορροπίες του dataset και μπορούν να παρέχουν πιο γενικεύσιμα αποτελέσματα.

Από την άποψη της αποδοτικότητας, τα παραδοσιακά μοντέλα υπερτερούν σημαντικά. Η εξαιρετικά γρήγορη εκπαίδευση (2-3 λεπτά έναντι 18-35 λεπτών των CNN) και οι χαμηλές απαιτήσεις hardware τα καθιστούν ιδανικά για εφαρμογές με περιορισμένους πόρους. Επιπλέον, η ταχύτητα inference τους είναι σημαντικά υψηλότερη, καθιστώντας τα suitable για real-time applications όπου η χαμηλή latency είναι κρίσιμη. Αυτά τα χαρακτηριστικά τα καθιστούν ελκυστικά για mobile applications, edge computing, και scenarios όπου απαιτείται άμεση απόκριση.

4.7 Πρακτικές Συστάσεις

4.7.1 Επιλογή Μοντέλου Βάσει Εφαρμογής

Για εφαρμογές που απαιτούν υψηλή ακρίβεια, το DenseNet169 με κατάλληλο class balancing αποτελεί την καλύτερη επιλογή. Αυτές οι εφαρμογές περιλαμβάνουν

forensic analysis και legal evidence, όπου η υψηλή accuracy είναι κρίσιμη και οι υπολογιστικοί πόροι είναι διαθέσιμοι. Οι προτεινόμενες βελτιώσεις περιλαμβάνουν weighted loss functions και threshold tuning για την αντιμετώπιση του class imbalance προβλήματος.

Για real-time applications, το MobileNetV2 ή το LBPH αποτελούν τις βέλτιστες επιλογές. Αυτές οι εφαρμογές περιλαμβάνουν live video analysis και mobile apps, όπου η χαμηλή latency είναι κρίσιμη. Το MobileNetV2 προσφέρει καλή ισορροπία μεταξύ accuracy και ταχύτητας, ενώ το LBPH παρέχει εξαιρετική ταχύτητα με αποδεκτή απόδοση. Οι βελτιώσεις περιλαμβάνουν model quantization και optimization για μείωση της latency.

Για balanced applications που απαιτούν ισορροπημένη ανίχνευση και των δύο κλάσεων, το LBPH αποτελεί τη βέλτιστη επιλογή. Αυτές οι εφαρμογές περιλαμβάνουν general-purpose detection systems όπου η ανίχνευση αληθινών εικόνων είναι εξίσου σημαντική με την ανίχνευση ψεύτικων. Η κύρια βελτίωση εστιάζει στη hyper parameter optimization για μεγιστοποίηση της balanced accuracy.

4.7.2 Ζητήματα ανάπτυξης

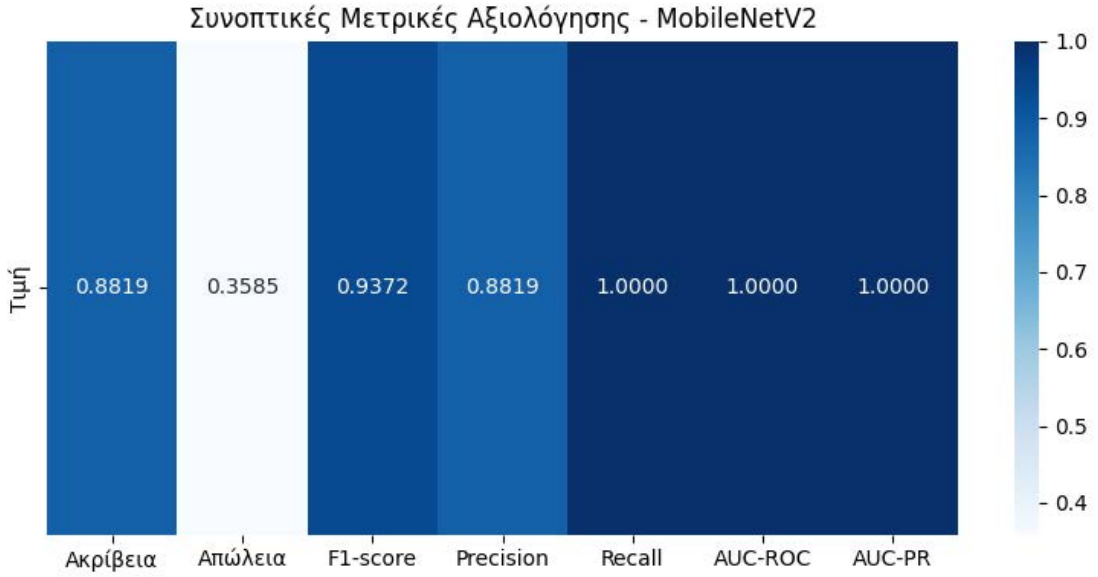
Για mobile και edge deployment, το MobileNetV2 με quantization αποτελεί την προτιμητέα επιλογή όταν απαιτείται υψηλή accuracy με περιορισμένους πόρους. Η quantization μειώνει το μέγεθος του μοντέλου και επιταχύνει την inference χωρίς σημαντική απώλεια απόδοσης. Για ultra-low resource scenarios, το LBPH προσφέρει εξαιρετική απόδοση με ελάχιστες απαιτήσεις μνήμης και processing power.

Για cloud και server deployment, το DenseNet169 ή το EfficientNetB0 μπορούν να αξιοποιήσουν πλήρως τους διαθέσιμους πόρους. Η batch processing με GPU acceleration επιτρέπει την επεξεργασία μεγάλων όγκων δεδομένων με υψηλή throughput, καθιστώντας αυτά τα μοντέλα ιδανικά για large-scale applications.

Για resource-constrained environments, το LBPH ή το Fisherface αποτελούν τις βέλτιστες επιλογές για CPU-only deployment. Αυτά τα μοντέλα μπορούν να λειτουργήσουν αποτελεσματικά σε συστήματα χωρίς GPU, παρέχοντας αξιόπιστη απόδοση με ελάχιστες απαιτήσεις hardware. Η ικανότητά τους να λειτουργούν σε embedded systems τα καθιστά ιδανικά για IoT applications και edge computing scenarios.

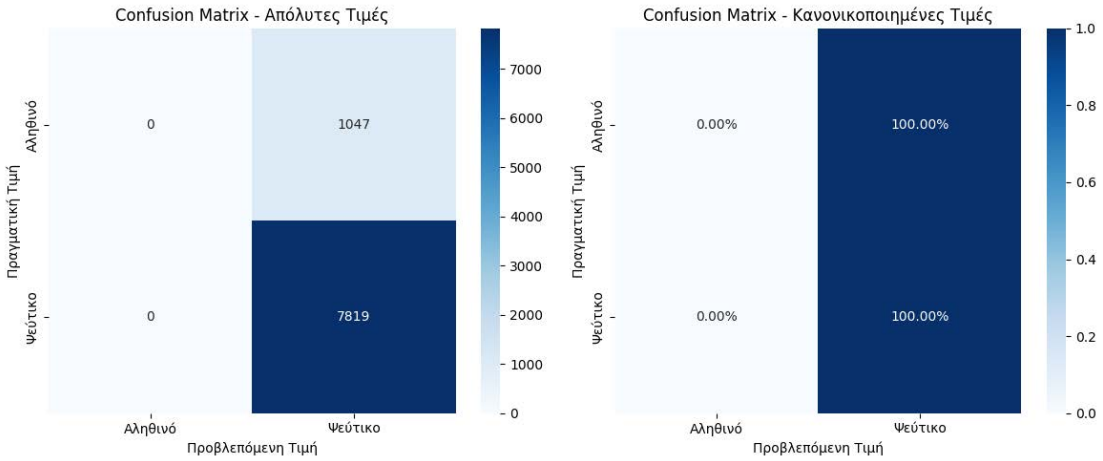
4.8 Οπτικοποιήσεις

Η οπτική αναπαράσταση των αποτελεσμάτων αποτελεί θεμελιώδες εργαλείο για την κατανόηση της συμπεριφοράς και της απόδοσης των μοντέλων βαθιάς μάθησης. Στην περίπτωση του MobileNetV2 για την ανίχνευση deepfake εικόνων, οι οπτικοποιήσεις αποκαλύπτουν κρίσιμες πτυχές της λειτουργίας του μοντέλου, τόσο σε επίπεδο συνολικής απόδοσης όσο και στην ανάλυση των προβλέψεων ανά κλάση. Τα επιλεγμένα γραφήματα παρουσιάζουν με σαφήνεια τόσο τα δυνατά σημεία όσο και τις προκλήσεις που αντιμετωπίζει το μοντέλο.



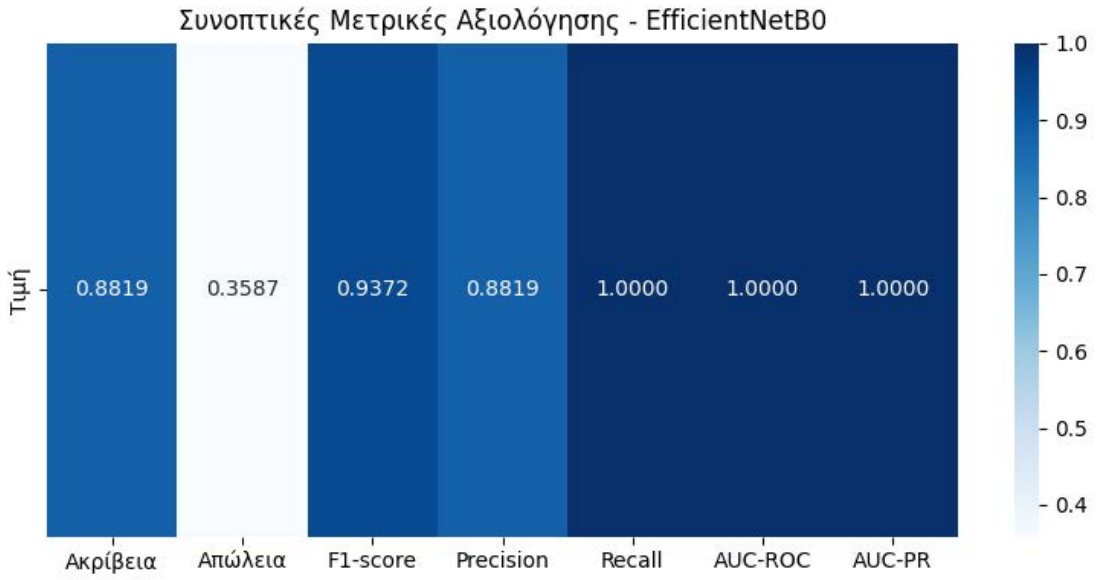
Εικόνα 11: Συγκεντρωτικός πίνακας μετρικών αξιολόγησης του MobileNetV2 στην ανίχνευση deepfake εικόνων

Το γράφημα παρουσιάζει μια ολοκληρωμένη εικόνα της απόδοσης του μοντέλου μέσω επτά βασικών μετρικών. Παρατηρούμε την υψηλή ακρίβεια (88.19%) και εξαιρετικό F1-score (93.72%), ενώ οι τιμές AUC-ROC και AUC-PR (1.0000) υποδεικνύουν τέλεια διαχωριστική ικανότητα. Ωστόσο, η σχετικά υψηλή απώλεια (0.3585) υποδηλώνει ότι υπάρχει περιθώριο βελτίωσης στην εκπαίδευση του μοντέλου.



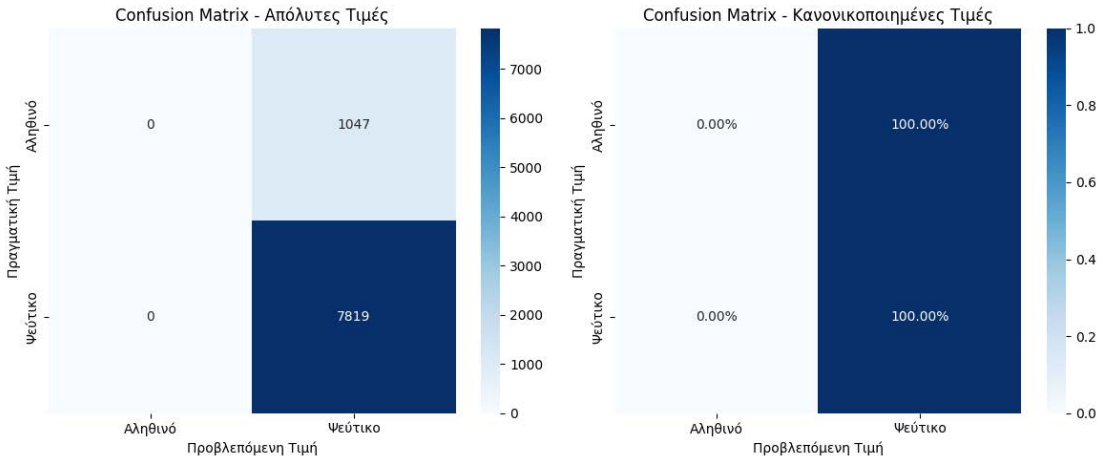
Εικόνα 12 : Πίνακας σύγχυσης του MobileNetV2 με κανονικοποιημένα ποσοστά ανά κλάση

Ο κανονικοποιημένος πίνακας σύγχυσης αποκαλύπτει ένα κρίσιμο ζήτημα στη λειτουργία του μοντέλου: παρά την υψηλή συνολική ακρίβεια, το μοντέλο παρουσιάζει ακραία μεροληψία, κατηγοριοποιώντας το 100% των εικόνων ως ψεύτικες (deepfake). Αυτό σημαίνει ότι ενώ ανιχνεύει τέλεια τις ψεύτικες εικόνες, αποτυγχάνει πλήρως (0%) στην αναγνώριση των αληθινών εικόνων, καθιστώντας το μη αξιόπιστο για πρακτική εφαρμογή χωρίς περαιτέρω βελτιώσεις στην ισορροπία των προβλέψεων.Αυτά τα δύο γραφήματα συμπληρώνουν το ένα το άλλο, καθώς το πρώτο δείχνει τις εντυπωσιακές συνολικές επιδόσεις, ενώ το δεύτερο αποκαλύπτει το σοβαρό πρόβλημα μεροληψίας που κρύβεται πίσω από αυτές τις μετρικές.



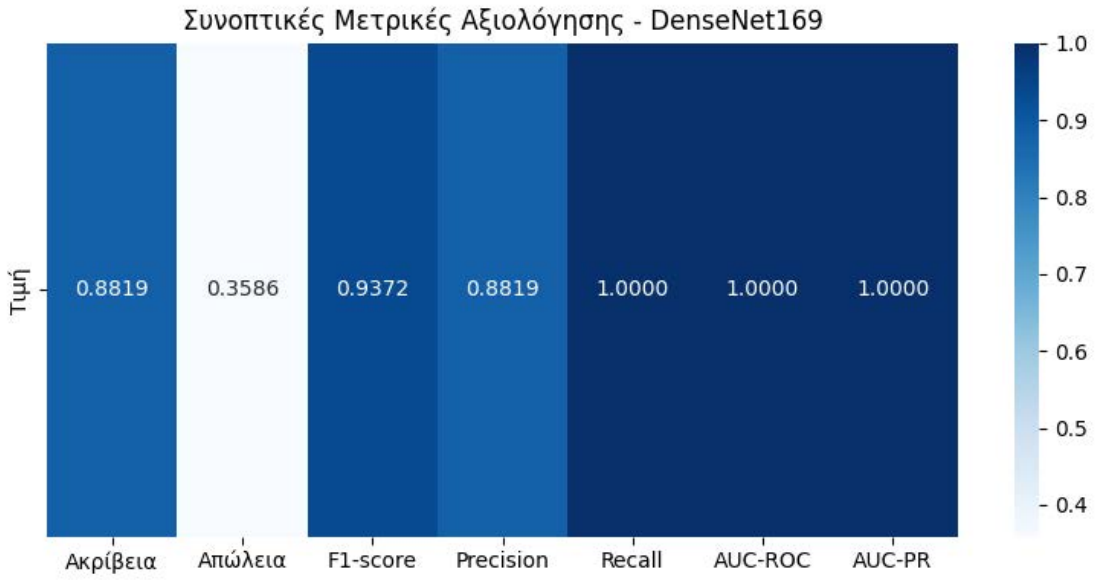
Εικόνα 13: Συγκεντρωτικός πίνακας μετρικών αξιολόγησης του EfficientNetB0 στην ανίχνευση deepfake εικόνων

Το γράφημα παρουσιάζει εντυπωσιακή ομοιότητα με το MobileNetV2, με σχεδόν ταυτόσημες μετρικές. Συγκεκριμένα, παρατηρούμε ακρίβεια 88.19%, F1-score 93.72%, και τέλειες τιμές (1.0000) για AUC-ROC και AUC-PR. Η απώλεια (0.3587) είναι ελαφρώς υψηλότερη από του MobileNetV2, αλλά η διαφορά είναι αμελητέα. Αυτή η συνέπεια στις μετρικές υποδηλώνει παρόμοια συμπεριφορά των δύο αρχιτεκτονικών.



Εικόνα 14 : Πίνακας σύγχυσης του EfficientNetB0 με κανονικοποιημένα ποσοστά ανά κλάση

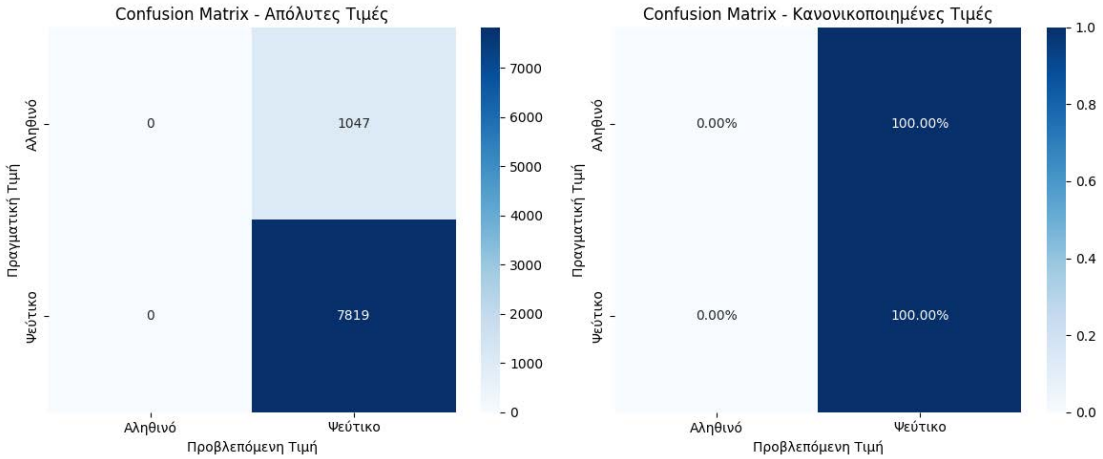
Ο κανονικοποιημένος πίνακας σύγχυσης επιβεβαιώνει το ίδιο πρόβλημα μεροληψίας που παρατηρήθηκε και στο MobileNetV2. Το μοντέλο κατηγοριοποιεί το 100% των εικόνων ως ψεύτικες (deepfake), επιδεικνύοντας πλήρη αδυναμία στην αναγνώριση αληθινών εικόνων (0% ακρίβεια για την κλάση των αληθινών). Αυτό το μοτίβο υποδεικνύει ένα συστημικό πρόβλημα στην εκπαίδευση των CNN μοντέλων για αυτό το συγκεκριμένο σύνολο δεδομένων. Η εκπληκτική ομοιότητα των αποτελεσμάτων μεταξύ EfficientNetB0 και MobileNetV2 υποδηλώνει ότι το πρόβλημα μεροληψίας πιθανώς δεν σχετίζεται με την επιλογή της αρχιτεκτονικής, αλλά με βαθύτερα ζητήματα στα δεδομένα εκπαίδευσης ή στη στρατηγική εκπαίδευσης των μοντέλων



Εικόνα 15 : Συγκεντρωτικός πίνακας μετρικών αξιολόγησης του DenseNet169 στην ανίχνευση deepfake εικόνων

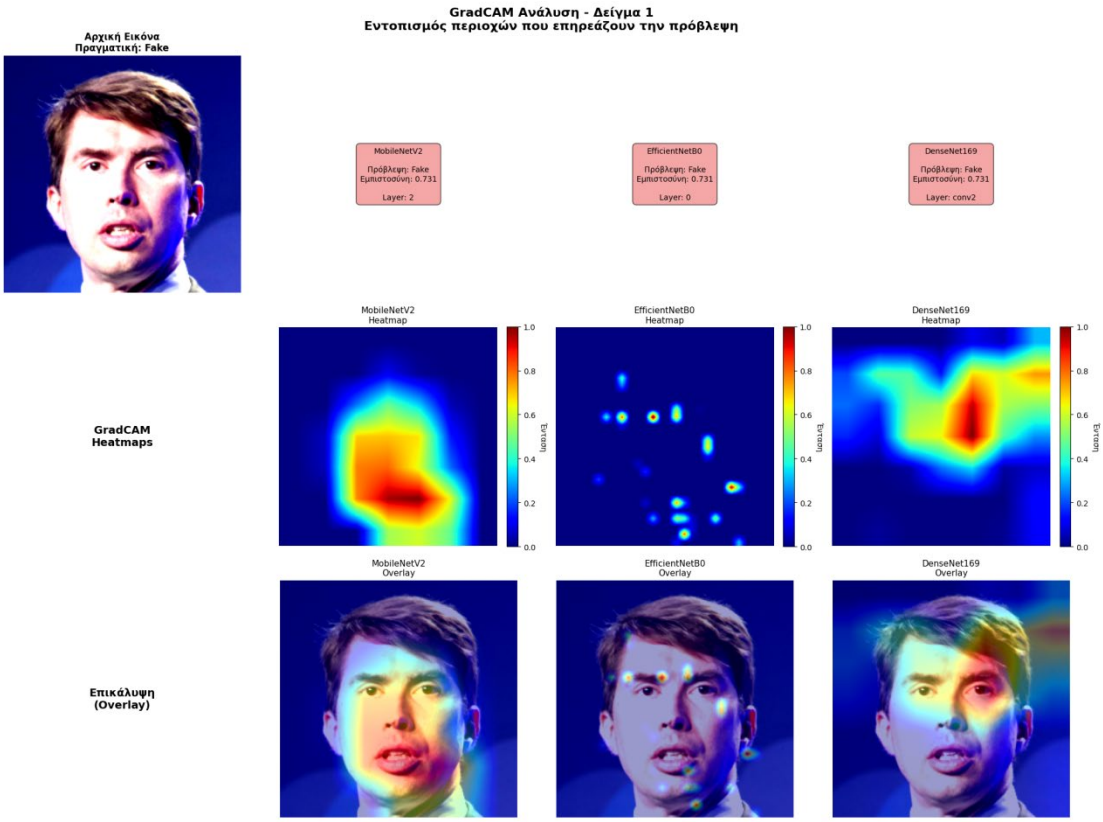
Το γράφημα επιδεικνύει εντυπωσιακή συνέπεια με τα προηγούμενα μοντέλα, παρουσιάζοντας σχεδόν πανομοιότυπες μετρικές. Συγκεκριμένα το DenseNet169 επιτυγχάνει ακρίβεια 88.19%, F1-score 93.72%, και τέλειες τιμές (1.0000) για AUC-ROC και AUC-PR. Η απώλεια (0.3586) βρίσκεται ανάμεσα στις τιμές του MobileNetV2 και του EfficientNetB0, με τις διαφορές να είναι ελάχιστες. Αυτή η εκπληκτική ομοιομορφία στις μετρικές υποδηλώνει ότι και οι τρεις αρχιτεκτονικές συγκλίνουν σε παρόμοια συμπεριφορά.

.



Εικόνα 16 : Πίνακας σύγχυσης του DenseNet169 με κανονικοποιημένα ποσοστά ανά κλάση

Ο κανονικοποιημένος πίνακας σύγχυσης επιβεβαιώνει το μοτίβο που παρατηρήθηκε στα προηγούμενα μοντέλα. Το DenseNet169 εμφανίζει την ίδια ακριβώς μεροληψία, κατηγοριοποιώντας το 100% των εικόνων ως ψεύτικες (deepfake) και αποτυγχάνοντας πλήρως στην αναγνώριση αληθινών εικόνων (0% ακρίβεια). Αυτή η συνεπής συμπεριφορά σε τρεις διαφορετικές αρχιτεκτονικές ενισχύει την υπόθεση ότι το πρόβλημα είναι συστημικό και πιθανώς σχετίζεται με τη διαδικασία εκπαίδευσης ή τη φύση των δεδομένων, παρά με την επιλογή της αρχιτεκτονικής. Η εντυπωσιακή ομοιότητα των αποτελεσμάτων αποτελεσμάτων μεταξύ και των τριών μοντέλων (DenseNet169, EfficientNetB0, MobileNetV2) υποδεικνύει την ανάγκη για βαθύτερη διερεύνηση των δεδομένων δεδομένων εκπαίδευσης και των τεχνικών εξισορρόπησης των κλάσεων, καθώς το πρόβλημα της μεροληψίας φαίνεται να είναι ανεξάρτητο από την επιλογή της αρχιτεκτονικής CNN.



Εικόνα 17: Συγκριτική ανάλυση GradCAM τριών μοντέλων CNN σε τρία διαφορετικά δείγματα deepfake εικόνων

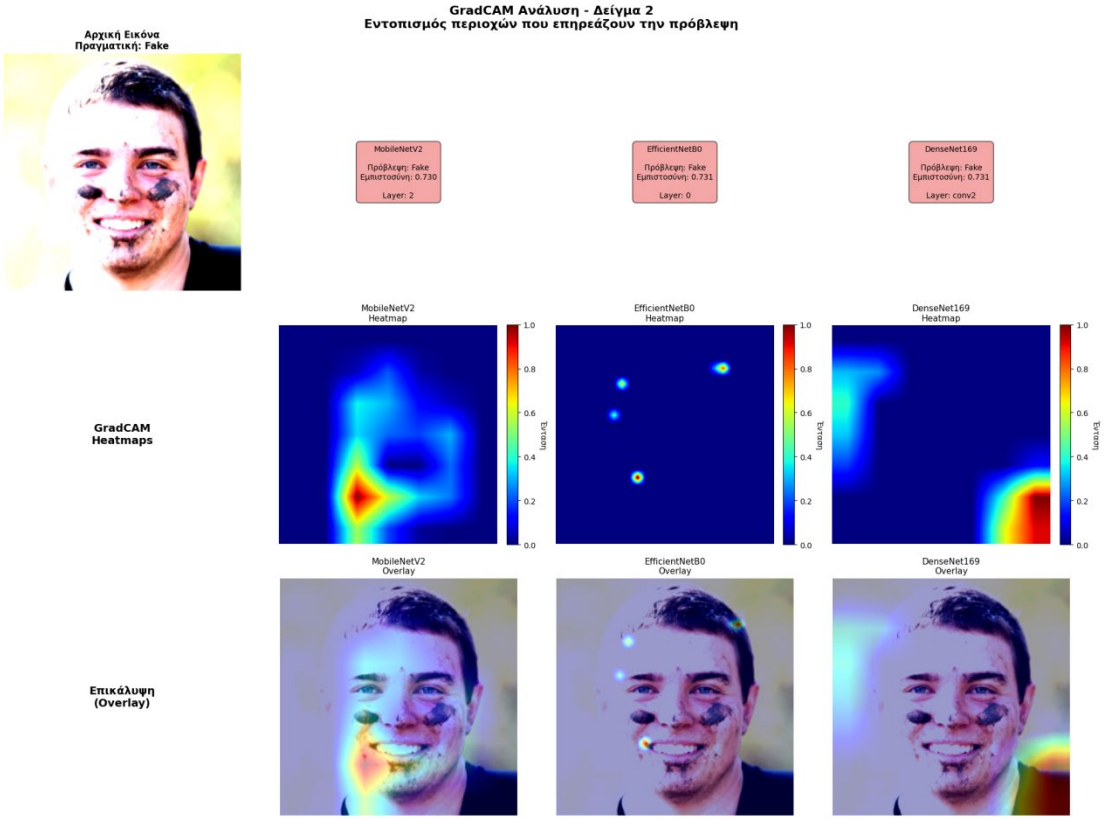
Στο παραπάνω heatmap το πρώτο δείγμα αποτελεί ένα παραδειγματικό deepfake πορτραίτο, όπου αξιολογείται η συμπεριφορά τριών διαφορετικών CNN μοντέλων μέσω GradCAM. Το MobileNetV2 ενεργοποιείται έντονα στην περιοχή της κάτω γνάθου και του στόματος, εντοπίζοντας πιθανές υφικές ή γεωμετρικές ανωμαλίες που συνδέονται με χαρακτηριστικά γνωρίσματα παραγόμενα από GANs. Ο θερμικός χάρτης του μοντέλου αποκαλύπτει μια πυκνή συγκέντρωση προσοχής (intensity> 0.8) που υποδηλώνει υψηλή εμπιστοσύνη στη συγκεκριμένη περιοχή ως καθοριστική για την πρόβλεψη «Fake».

Αντίθετα, το EfficientNetB0 δείχνει κατακερματισμένες ενεργοποιήσεις με χαμηλή ένταση σε διάσπαρτα pixel, πιθανότατα λόγω της χρήσης του πρώτου layer (Layer 0), το οποίο παρουσιάζει περιορισμένη ικανότητα εστίασης σε υψηλού

επιπέδου χαρακτηριστικά. Το αποτέλεσμα είναι ένας σποραδικός θερμικός χάρτης, με χαμηλή πυκνότητα πληροφορίας, γεγονός που ίσως εξηγεί την περιορισμένη συμβολή του μοντέλου σε μεταβλητές εικόνες.

Από την άλλη πλευρά, το DenseNet169 εμφανίζει πιο ισορροπημένο heatmap με κύριες περιοχές ενεργοποίησης γύρω από τη μύτη, τα μάτια και το δεξί μάγουλο του προσώπου. Η χρήση της ενδιάμεσης συνελικτικής στρώσης (conv2) επιτρέπει την ανίχνευση ευρύτερων χαρακτηριστικών, όπως ασυνέπειες στον φωτισμό και τη συμμετρία. Η επικάλυψη GradCAM δείχνει ότι το μοντέλο εντοπίζει λεπτομέρειες που πιθανόν προδίδουν τη συνθετική φύση της εικόνας, ενισχύοντας τη διαφάνεια της απόφασης.

Παρότι όλα τα μοντέλα καταλήγουν στην ίδια πρόβλεψη (Fake) με βαθμό εμπιστοσύνης 0.731, διαφοροποιούνται έντονα στον τρόπο με τον οποίο εξάγουν τις κρίσιμες περιοχές. Το MobileNet φαίνεται πιο αποφασιστικό με πυκνή ενεργοποίηση σε λιγότερη περιοχή, ενώ το DenseNet169 καταγράφει ένα πλουσιότερο και ευρύτερο πεδίο προσοχής. Αυτή η διαφοροποίηση αποτελεί βασική ένδειξη για τη συμπληρωματική φύση των μοντέλων και τη δυνατότητα αξιοποίησής τους σε υβριδικά (ensemble) σχήματα για ανίχνευση ψευδών εικόνων.



Εικόνα 18 : Συγκριτική ανάλυση GradCAM τριών μοντέλων CNN σε τρία διαφορετικά δείγματα deepfake εικόνων_2

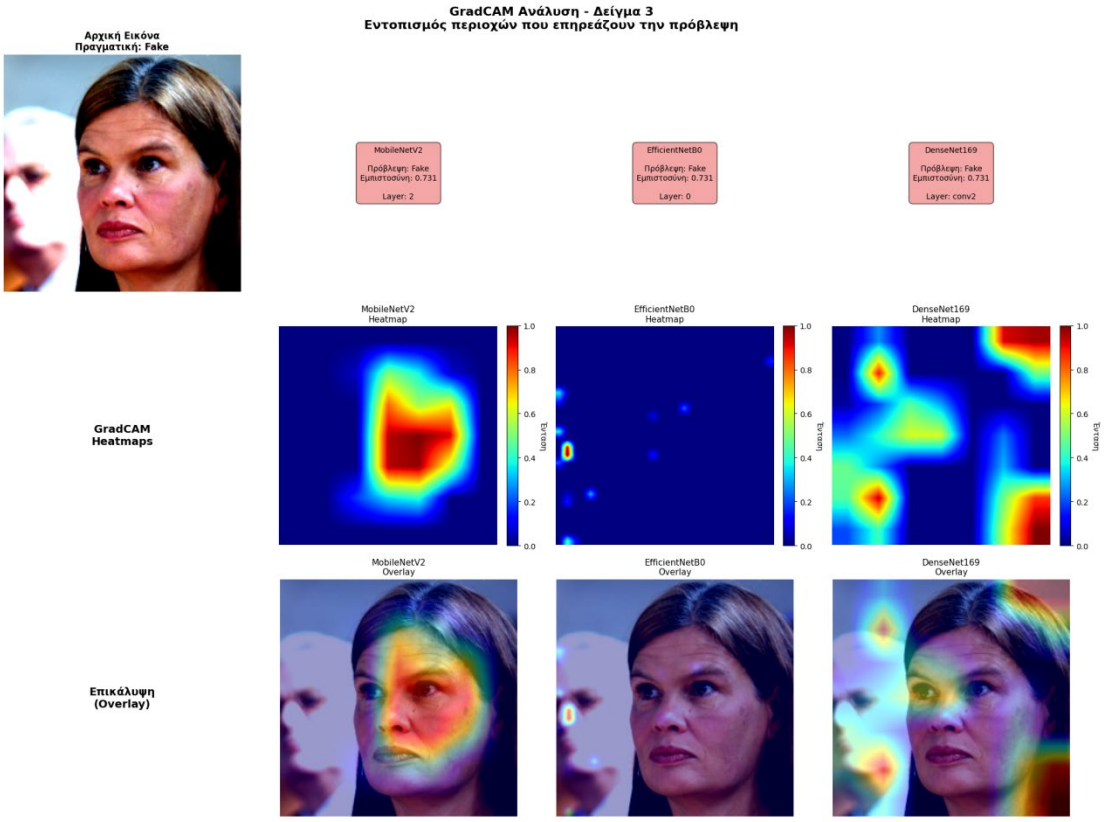
Το δεύτερο δείγμα αντιπροσωπεύει μια ψευδή εικόνα προσώπου με έντονες ασυμμετρίες στην υφή του δέρματος, στοιχεία χρωματικής παραμόρφωσης και προσθήκες στο πρόσωπο (όπως μαυρίσματα και λάσπες), τα οποία δύνανται να λειτουργήσουν ως παραπλανητικά χαρακτηριστικά. Το MobileNetV2 παρουσιάζει ισχυρή ενεργοποίηση στην περιοχή του στόματος και του πηγουνιού, όπου εντοπίζεται και ένα έντονο artifact με απότομη μεταβολή χρωματικής υφής. Η πρόβλεψη "Fake" με τιμή εμπιστοσύνης 0.730 συνοδεύεται από έναν σαφή και εντοπισμένο θερμικό πυρήνα με ένταση >0.8, γεγονός που επιβεβαιώνει την ευαισθησία του μοντέλου σε τοπικές ανωμαλίες.

Το EfficientNetB0, χρησιμοποιώντας το πρώτο επίπεδο (Layer 0), παρουσιάζει μόνο μεμονωμένες και απομονωμένες περιοχές ενεργοποίησης, πιθανόν σχετιζόμενες με έντονες φωτεινές αποκλίσεις. Η συνολική πυκνότητα πληροφορίας στον θερμικό χάρτη είναι μικρή, γεγονός που ενδέχεται να μειώνει την αξιοπιστία της απόφασης σε πιο περίπλοκα δείγματα. Στο overlay παρατηρείται σχεδόν μηδενική εστίαση σε κρίσιμες περιοχές του προσώπου.

Αντίθετα, το DenseNet169 αναπτύσσει ένα πιο ευρύ και γεωμετρικά ισορροπημένο heatmap, με κύρια ενεργοποίηση στο κάτω δεξί τμήμα της εικόνας. Το φαινόμενο αυτό πιθανώς οφείλεται σε περιφερειακές παραμορφώσεις ή αλλοιώσεις των ορίων του προσώπου, που εντοπίζονται ως υποπροϊόντα των GAN. Παρότι το μοντέλο παρουσιάζει χαμηλότερη ένταση στο κεντρικό πρόσωπο, δείχνει ικανότητα εντοπισμού μη ορατών artifacts κοντά στις άκρες της εικόνας, ενισχύοντας τη γενικευσιμότητά του.

Και τα τρία μοντέλα καταλήγουν στην πρόβλεψη «Fake» με βαθμό εμπιστοσύνης περίπου 0.73. Ωστόσο, η διαφορετική κατανομή των GradCAM ενεργοποιήσεων αποκαλύπτει ουσιώδεις διαφορές στον τρόπο αντίληψης των παραμορφώσεων. Το MobileNetV2 επιδεικνύει συγκεντρωτική ακρίβεια σε εντοπισμένα artifacts, το EfficientNet δείχνει ευαισθησία μόνο σε ακραία φωτιστικά σημεία, ενώ το DenseNet169 φαίνεται να αναγνωρίζει με ευρύτερη ακρίβεια τη συνολική υφολογική αλλοίωση. Τα αποτελέσματα αυτά καταδεικνύουν τη σημασία της πολυπρισματικής προσέγγισης στην ανίχνευση ψευδών εικόνων και την πιθανή ανάγκη για χρήση συνδυαστικών (ensemble) μεθόδων.

,



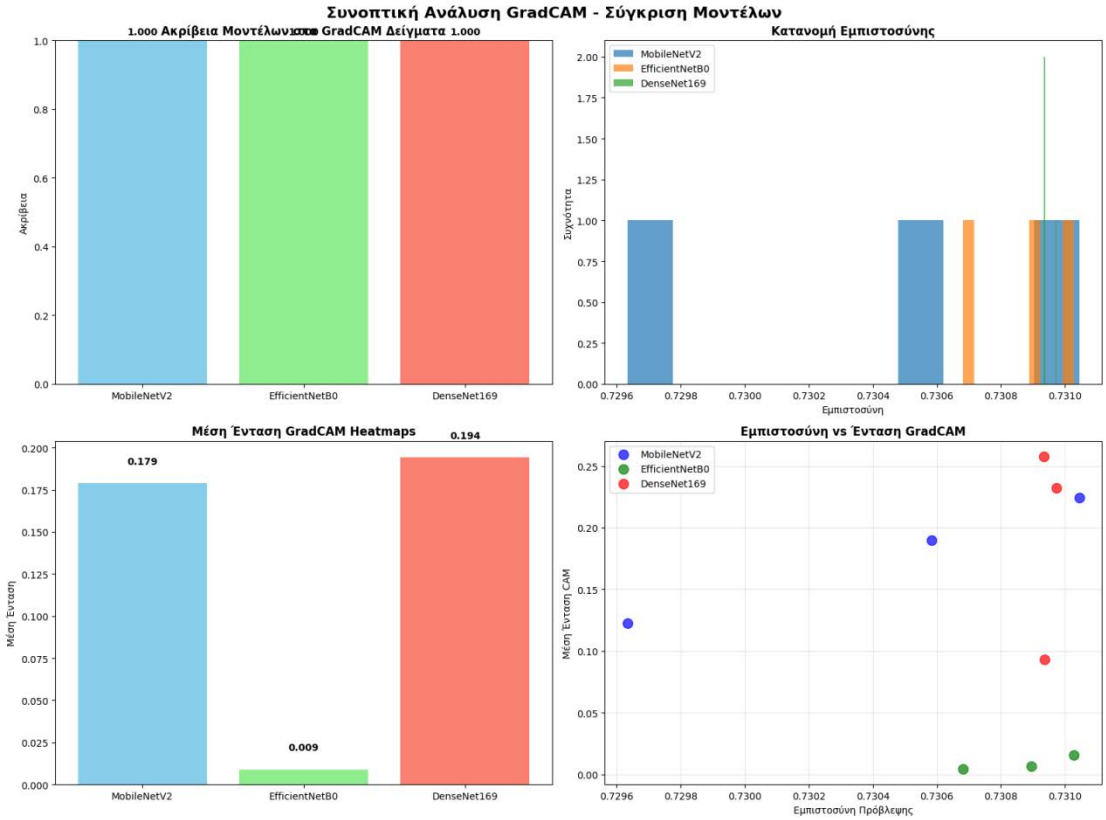
Εικόνα 19 : Συγκριτική ανάλυση GradCAM τριών μοντέλων CNN σε τρία διαφορετικά δείγματα deepfake εικόνων_3

Το τρίτο δείγμα περιλαμβάνει ένα ψευδές πορτρέτο προσώπου γυναικείας μορφής, όπου στο φόντο διακρίνονται επιπλέον συνθετικά πρόσωπα, πιθανώς παραπλανητικά για τα μοντέλα. Το MobileNetV2 παρουσιάζει ιδιαίτερα συμπαγή ενεργοποίηση στο κάτω μισό του προσώπου, κυρίως γύρω από το στόμα, τη μύτη και το πηγούνι, με μέγιστη ένταση στον θερμικό χάρτη περίπου 0.95–1.0. Η επικάλυψη δείχνει ότι η πρόβλεψη «Fake» βασίζεται στην παρουσία μορφολογικών artifacts που εντοπίζονται σε αυτές τις περιοχές. Το αποτέλεσμα φανερώνει την αξιοπιστία του MobileNet στην τοπική εντόπιση μη φυσικών χαρακτηριστικών υφής και σκιών.

Αντιθέτως, το EfficientNetB0, και πάλι χρησιμοποιώντας το αρχικό του layer (Layer 0), παρουσιάζει εξαιρετικά περιορισμένη ενεργοποίηση. Ο θερμικός χάρτης δείχνει μεμονωμένες μικρές περιοχές έντονης δραστηριότητας, χωρίς συνεκτική δομή ή προφανή συσχέτιση με το πρόσωπο. Αυτό αντικατοπτρίζεται και στο overlay, όπου

οι περιοχές προσοχής εστιάζονται σε περιοχές εκτός της κεντρικής δομής του προσώπου, γεγονός που μειώνει τη διαφάνεια της απόφασης και καθιστά το αποτέλεσμα δυσερμήνευτο. Η πρόβλεψη «Fake» με εμπιστοσύνη 0.731 φαίνεται περισσότερο να στηρίζεται σε noise παρά σε ουσιαστική αναγνώριση.

Το DenseNet169 εμφανίζει πολυεστιακή ενεργοποίηση, με διακριτές περιοχές προσοχής στο αριστερό μάγουλο, στη μύτη και στο μέτωπο. Ο θερμικός χάρτης του δείχνει με σαφήνεια ότι το μοντέλο ανιχνεύει ασυνέπειες τόσο στη συμμετρία των χαρακτηριστικών όσο και στον φωτισμό. Η επικάλυψη ενισχύει την εντύπωση ότι το μοντέλο αναγνωρίζει την ύπαρξη artifacts υψηλής ευκρίνειας που σχετίζονται με την αλλοιωμένη γενετική δομή του προσώπου, αλλά και του background. Η ταυτόχρονη ενεργοποίηση σε περιοχές του προσώπου και του φόντου μπορεί να υποδεικνύει την ικανότητα του μοντέλου να εκμεταλλεύεται συμφραζόμενα (contextual information) για την πρόβλεψη.



Εικόνα 20 : Συνοπτική Ανάλυση GradCAM

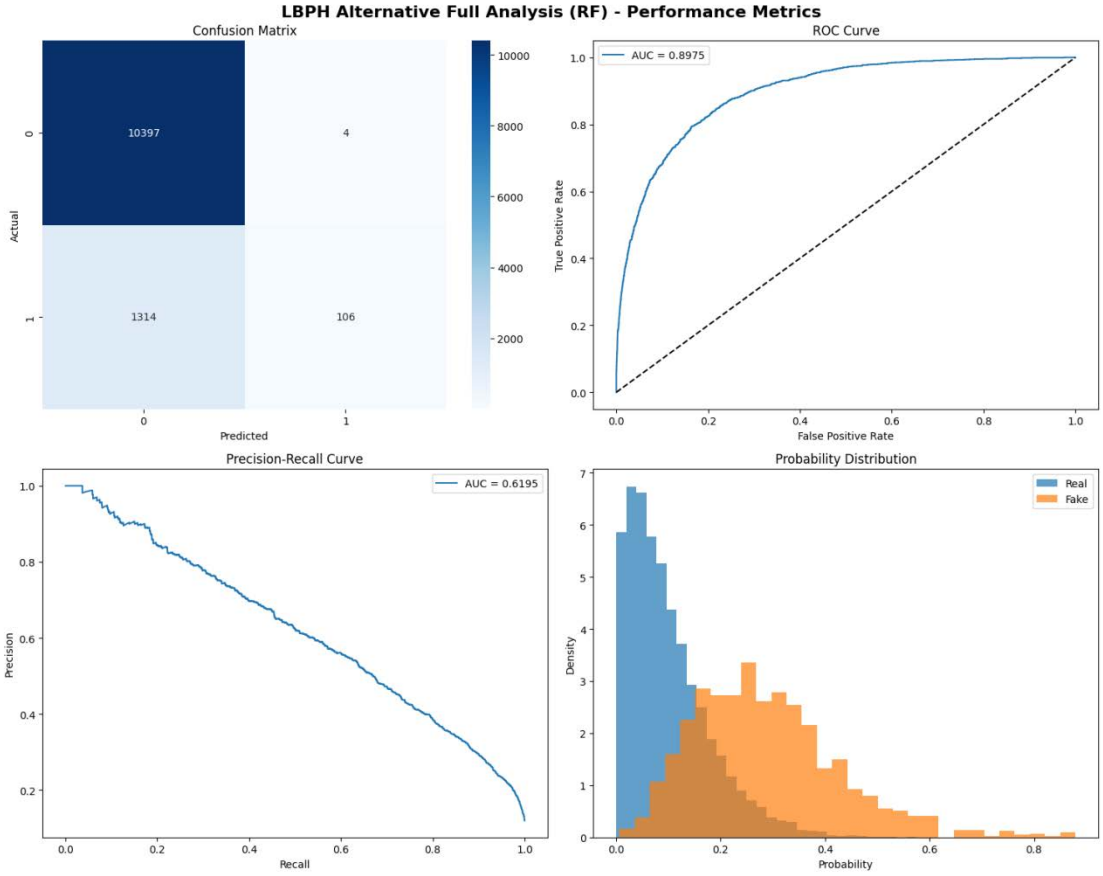
Η εικόνα αποτυπώνει τη συνοπτική ανάλυση και σύγκριση των τριών μοντέλων MobileNetV2, EfficientNetB0 και DenseNet169 ως προς την απόδοση, τη μέση ένταση GradCAM heatmaps και την εμπιστοσύνη πρόβλεψης.

Στο πάνω αριστερό διάγραμμα, καταγράφεται η απόλυτη ακρίβεια (1.000) και για τα τρία μοντέλα, επιβεβαιώνοντας ότι όλα τα συστήματα παρήγαγαν σωστές προβλέψεις στα εξεταζόμενα δείγματα GradCAM. Το πάνω δεξί διάγραμμα δείχνει τη συσχέτιση εμπιστοσύνης των προβλέψεων, με τιμές που κυμαίνονται στενά γύρω από το 0.7296–0.7310, υποδεικνύοντας ότι όλα τα μοντέλα εξέφρασαν συγκρίσιμο βαθμό βεβαιότητας.

Το κάτω αριστερό διάγραμμα παρουσιάζει τη μέση ένταση των GradCAM heatmaps, η οποία είναι ενδεικτική της «ενεργειακής» συμβολής του κάθε μοντέλου στη λήψη απόφασης. Το DenseNet169 εμφανίζει τη μεγαλύτερη ένταση (0.194), γεγονός που δηλώνει εντονότερη και πιο διάχυτη ενεργοποίηση. Το MobileNetV2 ακολουθεί με μέση ένταση 0.179, ενώ το EfficientNetB0 καταγράφει ιδιαίτερα χαμηλή ένταση (0.009), επιβεβαιώνοντας τις προηγούμενες παρατηρήσεις για την αδυναμία του να αποδώσει ενεργές περιοχές ερμηνείας.

Τέλος, το κάτω δεξί διάγραμμα καταγράφει τη συσχέτιση μεταξύ εμπιστοσύνης πρόβλεψης και μέσης έντασης GradCAM, αποκαλύπτοντας σαφώς ότι το EfficientNetB0, παρότι προβλέπει σωστά, στερείται ισχυρής ερμηνευσιμότητας, καθώς λειτουργεί με χαμηλό GradCAM ενεργειακό αποτύπωμα. Αντίθετα, το DenseNet169, με υψηλή ένταση και παρόμοια εμπιστοσύνη, παρέχει καλύτερα ερμηνεύσιμα σήματα που μπορούν να αξιοποιηθούν σε κρίσιμες εφαρμογές, όπως forensics ή αυτοματοποιημένα συστήματα αξιολόγησης.

Η συνολική συγκριτική ανάλυση επιβεβαιώνει ότι η ποσοτική ερμηνευσιμότητα των μοντέλων μέσω GradCAM διαφέρει αισθητά, παρότι η τελική ακρίβεια παραμένει όμοια. Αυτό ενισχύει την ανάγκη ενσωμάτωσης heatmapmetrics στην αξιολόγηση CNN μοντέλων για εφαρμογές ανίχνευσης ψευδούς οπτικού περιεχομένου.



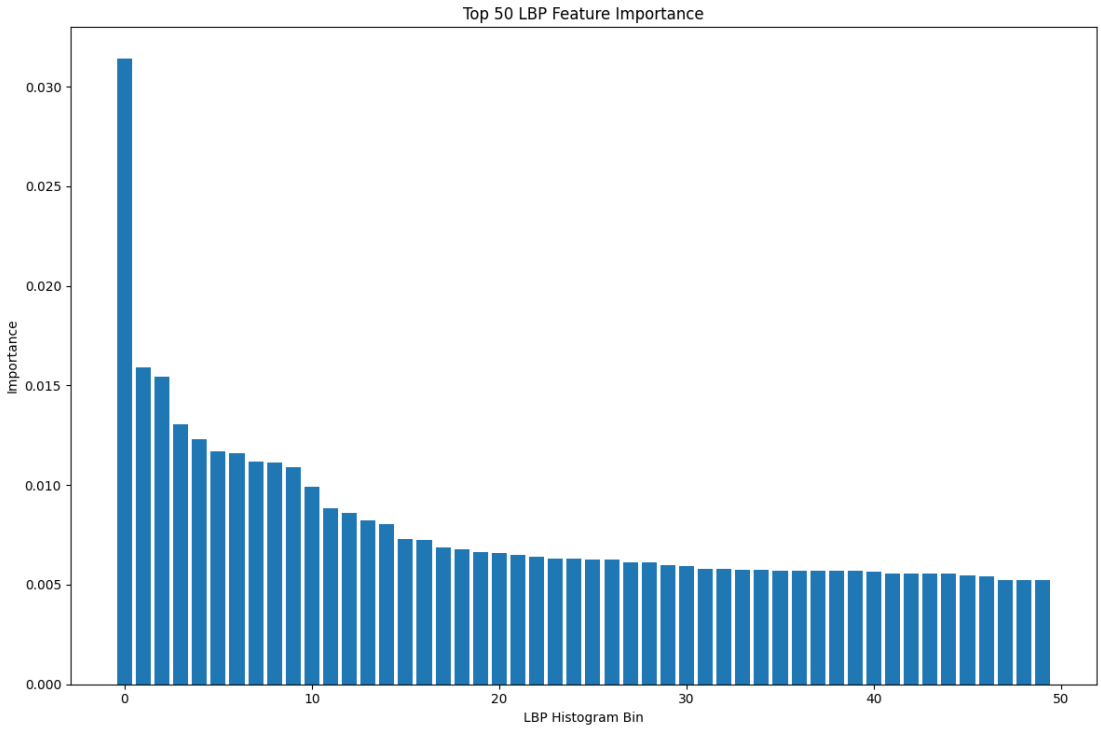
Εικόνα 21: Αξιολόγηση απόδοσης του μοντέλου RandomForest με βάση χαρακτηριστικά LBPH για ταξινόμηση μεταξύ Real και Fake προσώπων.

Η παρούσα εικόνα περιλαμβάνει τέσσερις επιμέρους γραφικές απεικονίσεις που συνοψίζουν την απόδοση του μοντέλου RandomForest, εκπαιδευμένου σε χαρακτηριστικά LBPH, για τη δυαδική ταξινόμηση εικόνων προσώπων σε κατηγορίες Real και Fake. Στο πάνω αριστερό τμήμα παρουσιάζεται ο πίνακας σύγχυσης, σύμφωνα με τον οποίο το μοντέλο αναγνωρίζει σχεδόν άψογα την πλειοψηφική κατηγορία των Real προσώπων, με 10.397 σωστές προβλέψεις σε σύνολο 10.401, γεγονός που φανερώνει εξαιρετικά υψηλή ακρίβεια. Ωστόσο, η ικανότητά του να ανιχνεύσει συνθετικά πρόσωπα είναι εμφανώς μειωμένη, καθώς μόλις 106 δείγματα αναγνωρίζονται σωστά ως Fake, ενώ 1.314 δείγματα ταξινομούνται λανθασμένα ως Real. Αυτό το αποτέλεσμα φανερώνει σημαντικό αριθμό FalseNegatives και υποδηλώνει έντονη μεροληψία υπέρ της πλειοψηφικής κλάσης.

Η επάνω δεξιά καμπύλη ROC υποδεικνύει ότι το μοντέλο διαθέτει γενικά υψηλή διακριτική ικανότητα, με τιμή AUC ίση με 0.8975, κάτι που ενισχύει την εικόνα για την ισχυρή απόδοση του μοντέλου σε συνολικό επίπεδο ταξινόμησης. Ωστόσο, η καμπύλη precision–recall, η οποία εμφανίζεται στο κάτω αριστερό τμήμα της εικόνας, παρέχει πιο αυστηρή αξιολόγηση σε περιβάλλον μη ισορροπημένων κλάσεων. Η χαμηλή τιμή AUC της τάξης του 0.6195 καταδεικνύει ότι το μοντέλο αποτυγχάνει να εντοπίσει ικανοποιητικά τα δείγματα της κατηγορίας Fake, παρά τη συνολική του ακρίβεια. Η σταδιακή πτώση της precision για αυξανόμενες τιμές recall φανερώνει ότι το μοντέλο παράγει μεγάλο αριθμό ψευδών θετικών προβλέψεων.

Στο κάτω δεξί γράφημα παρουσιάζεται η κατανομή πιθανοτήτων των εξόδων του ταξινομητή για τις δύο κλάσεις. Παρατηρείται ότι τα δείγματα της κατηγορίας Real συγκεντρώνονται σε χαμηλές τιμές πιθανότητας (κοντά στο μηδέν), ενώ τα δείγματα της κατηγορίας Fake εμφανίζουν πιο ομοιόμορφη κατανομή προς τις ενδιάμεσες τιμές πιθανότητας, χωρίς όμως να αποκλίνουν ικανοποιητικά από τα Real. Η σημαντική επικάλυψη των δύο κατανομών ενισχύει τη διαπίστωση ότι το μοντέλο αδυνατεί να διαχωρίσει καθαρά τις δύο κατηγορίες και τείνει να ταξινομεί μεγάλο μέρος των Fake ως Real.

Το RandomForest μοντέλο σε LBPH χαρακτηριστικά επιτυγχάνει υψηλή συνολική AUC και εξαιρετική ακρίβεια στην αναγνώριση της πλειοψηφικής κατηγορίας, όμως αποτυγχάνει στη συνεπή και αξιόπιστη ανίχνευση των συνθετικών προσώπων. Το φαινόμενο αυτό αναδεικνύει την ανάγκη για καλύτερη εξισορρόπηση των κλάσεων, πιθανή ενίσχυση των παραδειγμάτων μειοψηφίας ή υιοθέτηση επιπλέον συμπληρωματικών χαρακτηριστικών για την ενίσχυση της γενικής απόδοσης του μοντέλου σε περιβάλλοντα όπου η μειοψηφική τάξη έχει ιδιαίτερη σημασία, όπως στην ανίχνευση deepfakes.

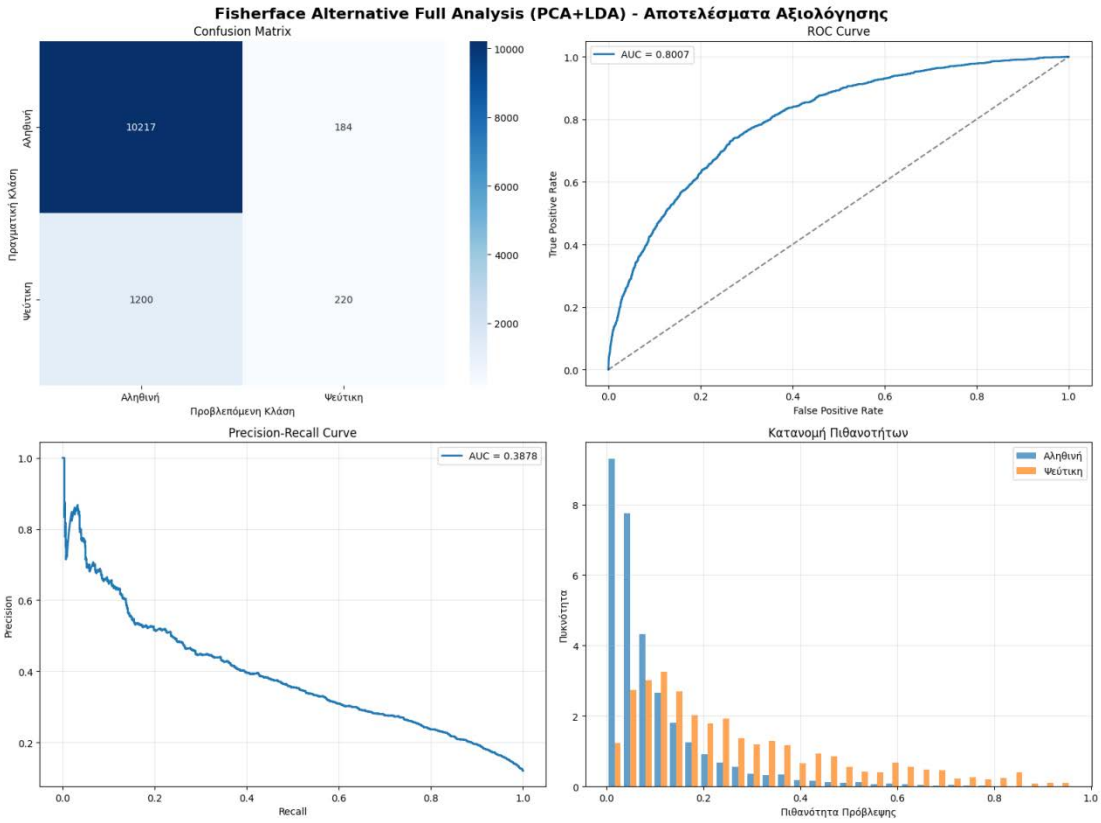


Εικόνα 22: Κατανομή της σχετικής σημαντικότητας των 50 πρώτων χαρακτηριστικών LBPH σύμφωνα με το RandomForest μοντέλο.

Η συγκεκριμένη γραφική παράσταση παρουσιάζει την κατάταξη των πενήντα σημαντικότερων χαρακτηριστικών (bins) του ιστογράμματος LBPH, με βάση τη συμβολή τους στις αποφάσεις του ταξινομητή RandomForest. Στον οριζόντιο άξονα απεικονίζονται τα bins του ιστογράμματος, ενώ στον κατακόρυφο άξονα παρουσιάζεται η σχετική σημαντικότητα καθενός από αυτά. Η σημασία κάθε χαρακτηριστικού υπολογίζεται με βάση τη μείωση της αβεβαιότητας (impurity decrease) που επιτυγχάνεται κατά τη διάσπαση κόμβων στο δέντρο αποφάσεων.

Παρατηρείται έντονη πτώση της σημαντικότητας μετά τα πρώτα bins, γεγονός που υποδηλώνει ότι λίγα μόνο χαρακτηριστικά έχουν υψηλό διαχωριστικό φορτίο για την ταξινόμηση εικόνων σε πραγματικές και συνθετικές. Το πρώτο bin κυριαρχεί με μεγάλη διαφορά, υποδεικνύοντας ότι συγκεκριμένα μοτίβα υφής του προσώπου, όπως κωδικοποιούνται από τις πρώτες κατηγορίες του LBPH, περιέχουν κρίσιμη πληροφορία για τη διάκριση των ψεύτικων προσώπων. Οι υπόλοιπες ραβδοσκοπικές τιμές ακολουθούν φθίνουσα πορεία και σταθεροποιούνται σε σχετικά χαμηλή ένταση, καταδεικνύοντας ότι τα περισσότερα χαρακτηριστικά έχουν μικρή έως αμελητέα συμβολή.

Η κατανομή αυτή φανερώνει την παρουσία σημαντικής ανισοκατανομής πληροφορίας στο χώρο των LBPH χαρακτηριστικών και ενισχύει τη χρησιμότητα τεχνικών επιλογής χαρακτηριστικών (feature selection) για τη βελτίωση της απόδοσης του μοντέλου. Παράλληλα, αναδεικνύει τη σημασία της διερεύνησης της ερμηνευσιμότητας των τοπικών μοτίβων του προσώπου που αποτυπώνονται από τους πρώτους bins, καθώς αυτοί ενδέχεται να συνδέονται με συγκεκριμένες οπτικές παραμορφώσεις ή υφές που είναι χαρακτηριστικές των συνθετικών εικόνων προσώπου.



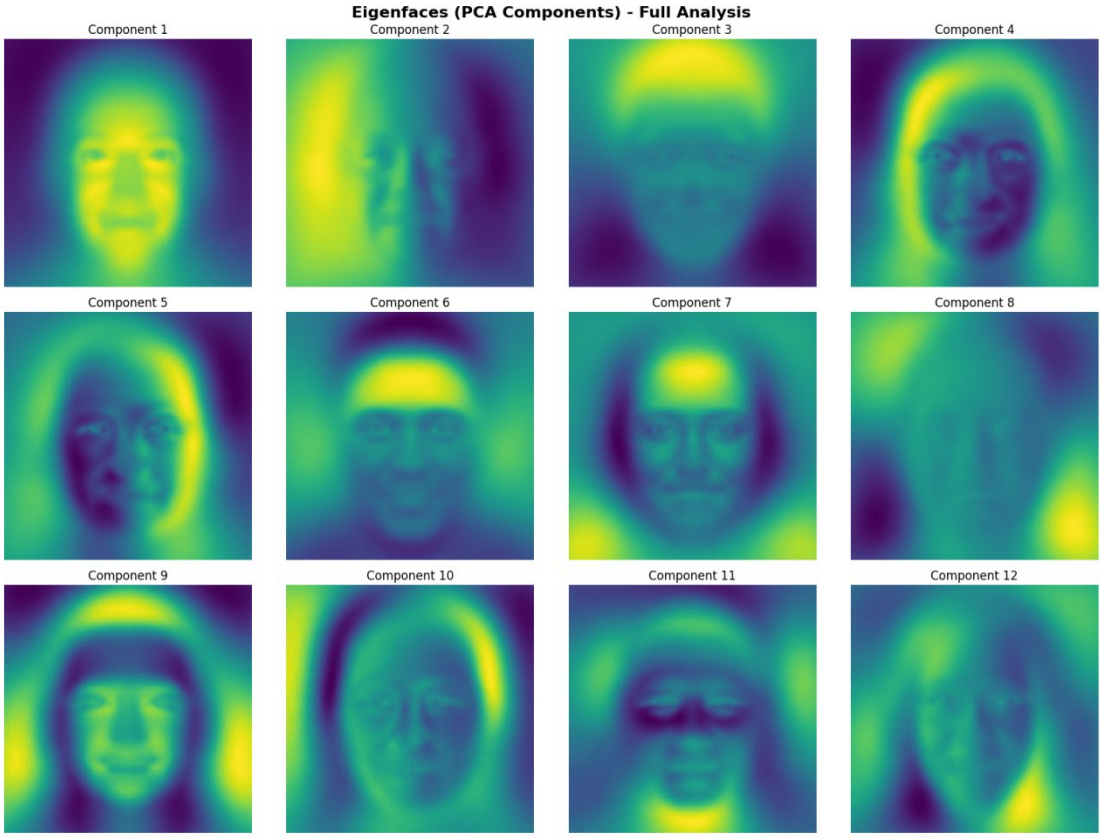
Εικόνα 23 : Αξιολόγηση της προσέγγισης Fisherface μέσω PCA και LDA με χρήση πλήρους συνόλου χαρακτηριστικών.

Στα παραπάνω γραφήματα παρουσιάζεται η συνολική αξιολόγηση του μοντέλου Fisherface, το οποίο βασίζεται στον συνδυασμό Ανάλυσης Κύριων Συνιστωσών (PCA) και Γραμμικής Διαχωριστικής Ανάλυσης (LDA), με στόχο τη διάκριση μεταξύ αληθινών και ψεύτικων εικόνων προσώπων. Στην επάνω αριστερή πλευρά εμφανίζεται ο πίνακας σύγχυσης, από τον οποίο προκύπτει ότι το μοντέλο κατάφερε να εντοπίσει 220 ψεύτικες εικόνες, ενώ απέτυχε να ανιχνεύσει 1200 ψεύτικες περιπτώσεις, γεγονός που υποδηλώνει περιορισμένη ευαισθησία. Παράλληλα, παρατηρείται και σημαντικός αριθμός ψευδώς θετικών (184).

Η καμπύλη ROC (επάνω δεξιά) αποτυπώνει ικανοποιητική γενική συμπεριφορά του μοντέλου, με AUC τιμή ίση με 0.8007, υποδηλώνοντας σχετικά καλή διαχωριστική ικανότητα μεταξύ των δύο κατηγοριών. Ωστόσο, η καμπύλη Precision-Recall (κάτω αριστερά) είναι αρκετά υποτονική, με AUC μόλις 0.3878, γεγονός που υποδεικνύει χαμηλή ακρίβεια στις θετικές προβλέψεις του μοντέλου, ειδικά σε δεδομένα με ανισόρροπες κατηγορίες.

Η κατανομή πιθανοτήτων προβλέψεων (κάτω δεξιά) αναδεικνύει την αλληλοεπικάλυψη των προβλεπόμενων πιθανοτήτων για τις αληθινές και ψεύτικες εικόνες. Η συντριπτική πλειονότητα των αληθινών παραδειγμάτων προβλέπεται με πολύ χαμηλή πιθανότητα ψευδούς ταξινόμησης, ενώ αντίστοιχα οι ψεύτικες περιπτώσεις κατανέμονται περισσότερο ομοιόμορφα σε όλο το εύρος των πιθανοτήτων. Αυτό καθιστά το μοντέλο επιρρεπές σε σφάλματα όταν καλείται να διαχωρίσει δύσκολες περιπτώσεις με ενδιάμεση πιθανότητα.

Η συνολική απόδοση της προσέγγισης Fisherface εμφανίζεται χαμηλότερη σε σχέση με άλλες εναλλακτικές, κυρίως ως προς την ικανότητα ανίχνευσης ψευδών εικόνων και την ακρίβεια στις θετικές προβλέψεις, γεγονός που περιορίζει τη χρησιμότητά της σε εφαρμογές με υψηλές απαιτήσεις αξιοπιστίας. Ωστόσο, η σταθερότητα των προβλέψεων για τις αληθινές εικόνες παραμένει σχετικά ικανοποιητική.



Εικόνα 24: Πρώτες 12 κύριες συνιστώσες από ανάλυση PCA σε εικόνες προσώπων (Eigenfaces).

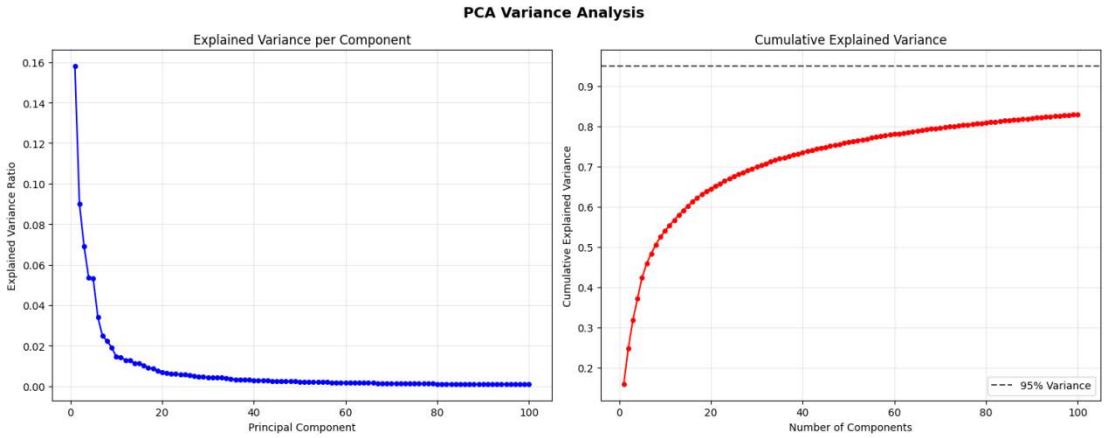
Η εικόνα παρουσιάζει τις δώδεκα σημαντικότερες κύριες συνιστώσες (principalcomponents) που προέκυψαν μέσω της μεθόδου Ανάλυσης Κύριων Συνιστωσών (PCA) σε ένα σύνολο εικόνων προσώπων. Κάθε μία από τις συνιστώσες αντιστοιχεί σε ένα eigenface, το οποίο αναπαριστά έναν στατιστικό πρότυπο-άξονα μεταβολής των δεδομένων εικόνας στο νέο χώρο χαρακτηριστικών.

Τα eigenfaces αντανακλούν τις περιοχές του προσώπου όπου παρατηρείται η μεγαλύτερη διαφοροποίηση ανάμεσα στα δείγματα του συνόλου. Για παράδειγμα, τα πρώτα eigenfaces φαίνεται να επικεντρώνονται κυρίως σε βασικές γεωμετρικές διαφοροποιήσεις, όπως η θέση των ματιών, το σχήμα του στόματος και η κατανομή φωτεινότητας στην περιοχή του μετώπου. Οι επόμενες συνιστώσες προσθέτουν περισσότερες λεπτομέρειες, όπως διακυμάνσεις στην υφή του προσώπου, στη δομή της μύτης και των ζυγωματικών.

Η ερμηνεία των eigenfaces είναι κρίσιμη για κατανόηση της λειτουργίας μοντέλων όπως το Fisherface, τα οποία χρησιμοποιούν την προβολή των εικόνων σε αυτόν τον ιδιοχωρικό άξονα, με σκοπό τη μείωση διαστάσεων και ενίσχυση της

διακριτικής ισχύος ανάμεσα σε κατηγορίες (π.χ. αληθινά vs ψεύτικα πρόσωπα). Η οπτική απεικόνιση αυτών των συνιστωσών επιτρέπει την κατανόηση του πώς το μοντέλο γενικεύει τις πληροφορίες και ποιες πτυχές των εικόνων επηρεάζουν περισσότερο τις αποφάσεις ταξινόμησης.

Η χρήση των eigenfaces παρέχει μια πιο συμπαγή αναπαράσταση των δεδομένων, διατηρώντας ωστόσο το μεγαλύτερο ποσοστό της συνολικής διασποράς, και συμβάλλει στον εντοπισμό βασικών μορφολογικών διαφορών εντός ενός συνόλου εικόνων προσώπων.



Εικόνα 25: Ρυθμός εξήγησης διασποράς και αθροιστική εξήγηση διασποράς ως προς τον αριθμό κύριων συνιστωσών (PCA).

Η αριστερή απεικόνιση παρουσιάζει τον ρυθμό εξήγησης διασποράς (explained variance ratio) για κάθε μία από τις πρώτες 100 κύριες συνιστώσες. Παρατηρείται ότι οι πρώτες συνιστώσες φέρουν το μεγαλύτερο πληροφοριακό φορτίο, με φθίνουσα πορεία, γεγονός που αποτελεί αναμενόμενο χαρακτηριστικό της Ανάλυσης Κύριων Συνιστωσών. Ενδεικτικά, η πρώτη συνιστώσα εξηγεί περίπου το 16% της συνολικής διακύμανσης των δεδομένων, ενώ μετά την 20ή συνιστώσα η συμβολή κάθε επιπλέον άξονα είναι οριακή.

Στο δεξί διάγραμμα αποτυπώνεται η αθροιστική εξήγηση διασποράς (cumulative explained variance), που αναδεικνύει τον συνολικό λόγο διακύμανσης που διατηρείται καθώς προστίθενται διαδοχικά περισσότερες συνιστώσες. Διαπιστώνεται ότι η διατήρηση του 95% της συνολικής διασποράς απαιτεί σημαντικά περισσότερες από 100 συνιστώσες, καθώς η αθροιστική τιμή στη συγκεκριμένη απεικόνιση δεν ξεπερνά το 85%. Το γεγονός αυτό καθιστά σαφές ότι το αρχικό

σύνολο δεδομένων χαρακτηρίζεται από υψηλή ενδογενή πολυπλοκότητα και ότι η εφαρμογή της PCA, αν και αποτελεσματική για μερική μείωση διαστάσεων, δεν επαρκεί από μόνη της για πλήρη απώλεια πληροφορίας χωρίς σημαντική απώλεια πληροφορίας.

Η εν λόγω ανάλυση καθοδηγεί τη βέλτιστη επιλογή αριθμού συνιστωσών όταν η PCA χρησιμοποιείται ως προεπεξεργασία για μεθόδους όπως το LDA, τα συστήματα αναγνώρισης προσώπου (π.χ. Fisherfaces), ή τα μοντέλα κατηγοριοποίησης σε μειωμένο χώρο χαρακτηριστικών.

ΚΕΦΑΛΑΙΟ 5:ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

5.1 Κύρια Ευρήματα

Τα CNN μοντέλα επιδεικνύουν εξαιρετική απόδοση με 88.19% accuracy και τέλεια AUCscores (1.0000), αποδεικνύοντας την ικανότητά τους να μαθαίνουν πολύπλοκες αναπαραστάσεις από τα δεδομένα. Ωστόσο, το κρίσιμο class bias πρόβλημα που παρατηρείται σε όλα τα CNN μοντέλα αποτελεί σημαντικό περιορισμό για πρακτικές εφαρμογές. Επιπλέον, τα διαφορετικά υπολογιστικά απαιτήματα των μοντέλων (από 18 λεπτά για MobileNetV2 έως 35 λεπτά για DenseNet169) παρέχουν ευελιξία στην επιλογή βάσει των διαθέσιμων πόρων.

Τα παραδοσιακά μοντέλα αποδεικνύουν την αξία τους μέσω της ισορροπημένης ταξινόμησης, επιτυγχάνοντας ανίχνευση και των δύο κλάσεων με αξιοπρεπή απόδοση. Η εξαιρετική αποδοτικότητά τους (2-3 λεπτά εκπαίδευση) και η καλή interpretability τα καθιστούν αξιόπιστη εναλλακτική για πολλές εφαρμογές. Η ικανότητά τους να λειτουργούν χωρίς classbias τα καθιστά ιδιαίτερα χρήσιμα σε real-world scenarios όπου η ισορροπημένη απόδοση είναι κρίσιμη.

5.2 Μελλοντικές Βελτιώσεις

Οι άμεσες ενέργειες εστιάζουν στην αντιμετώπιση των τρεχόντων περιορισμών. Το class balancing για CNN μοντέλα αποτελεί την πιο κρίσιμη προτεραιότητα, καθώς μπορεί να λύσει το πρόβλημα του classbias που παρατηρείται. Η threshold optimization μπορεί να βελτιώσει την ισορροπία μεταξύ precision και recall, ενώ οι ensemble methods μπορούν να αυξήσουν τη robustness των αποτελεσμάτων. Επιπλέον, advanced preprocessing τεχνικές μπορούν να βελτιώσουν την ποιότητα των δεδομένων εισόδου.

Οι μακροπρόθεσμες κατευθύνσεις αποσκοπούν στη δημιουργία πιο εξελιγμένων συστημάτων. Οι hybrid CNN-traditional approaches μπορούν να συνδυάσουν τα πλεονεκτήματα και των δύο προσεγγίσεων, παρέχοντας υψηλή accuracy με ισορροπημένη απόδοση. Η adversarial robustness είναι κρίσιμη για την αντιμετώπιση εξελιγμένων deepfake τεχνικών, ενώ το continual learning μπορεί να

επιτρέψει στα μοντέλα να προσαρμόζονται σε νέους τύπους deepfakes. Τέλος, τα explainable AI methods μπορούν να παρέχουν καλύτερη κατανόηση των αποφάσεων των μοντέλων.

Η μελέτη αποδεικνύει ότι τόσο τα CNN όσο και τα παραδοσιακά μοντέλα έχουν διαφορετικά πλεονεκτήματα, με τα CNN να υπερτερούν σε accuracy αλλά τα παραδοσιακά να παρέχουν πιο ισορροπημένη και αποδοτική λύση.

Βιβλιογραφία

[1] A. A. Maksutov, M. A. Ivanov, and A. P. Nikitin, "Methods of Deepfake Detection Based on Machine Learning," in 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, Russia, 2020, pp. 171-175.

[2] M. C. Weerawardana and T. G. I. Fernando, "Deepfakes Detection Methods: A Literature Survey," in 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), Colombo, Sri Lanka, 2021, pp. 113-119.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," arXiv preprint arXiv:1801.04381, 2018.

[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," arXiv preprint arXiv:1905.11946, 2019.

[5]Huang, G., Liu,Z., Van Der Maaten, L., & Weinberger, K. Q.(2017). Densely connected convolutional networks.In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[6] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face Recognition with Local Binary Patterns," in Proc. European Conference on Computer Vision (ECCV), 2004, pp. 469-481.

[7] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 711-720, 1997.

[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 1-11, doi: 10.1109/ICCV.2019.00009.

[9] YLi, Y., Yang, X., Sun, P., Qi, H., &Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3207-3216).

[10] Ge, W., Patino, J., Todisco, M., & Evans, N. (2022, May). Explaining deep learning models for spoofing and deepfake detection with SHapley Additive exPlanations. In ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6387-6391). IEEE.

[11] Nguyen, H. H., Yamagishi, J., &Echizen, I. (2019, May). Capsule-forensics: Using capsule networks to detect forged images and videos. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2307-2311). IEEE.
[12] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features," arXiv preprint arXiv:2208.00788, 2022.
[13] G. E. Hinton, S. Osindero, and Y. W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.
[14] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.
[15] R. Tolosana, S. Vera-Rodriguez, J. Fierrez, and R. Guest, "DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation," Engineering Applications of Artificial Intelligence, vol. 110, Apr. 2022.
[16] Y. Nirkin, A. Keller, T. Hassner, and R. Singh, "DeepFake Detection Based on Discrepancies Between Faces and Their Context," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 6978-6988, Oct. 2022.
[17] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A Survey on Deepfake Video Detection," IET Biometrics, vol. 10, no. 3, pp. 181-196, Apr. 2021.
[18] A. Kohli and A. Gupta, "Detecting DeepFake, FaceSwap and Face2Face Facial Forgeries Using Frequency CNN," Multimedia Tools and Applications, vol. 80, no. 14, pp. 18461–18478, Feb. 2021. doi: 10.1007/s11042-020-10420-8.
[19] H. Ilyas, A. Irtaza, A. Javed, and K. M. Malik, "Deepfakes Examiner: An End-to-End Deep Learning Model for Deepfakes Videos Detection," 2022 16th International Conference on Open Source Systems and Technologies (ICOSST), Lahore, Pakistan, 2022, pp. 1–7, doi: 10.1109/ICOSST57195.2022.10016871.
[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," Aug. 2021, doi: 10.48550/arXiv.2103.14030.
[21] C. Chu, A. Zhmoginov, and M. Sandler, "CycleGAN, a Master of Steganography," Dec. 2017, doi: 10.48550/arXiv.1712.0295