



ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Πτυχιακή / Διπλωματική Εργασία

Αναγνώριση και κατηγοριοποίηση μελανωμάτων με μετασχηματιστές και
τεχνικές βαθιάς μάθησης

Αλέξανδρος Μπιστόλας

Επιβλέπων καθηγητής: Ισίδωρος Περίκος

Πάτρα, Ιούνιος 2026

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει 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.



Αναγνώριση και κατηγοριοποίηση μελανωμάτων με μετασχηματιστές και
τεχνικές βαθιάς μάθησης

Αλέξανδρος Μπιστόλας

Επιτροπή Επίβλεψης Πτυχιακής / Διπλωματικής Εργασίας

Επιβλέπων Καθηγητής:

Ισίδωρος Περίκος

Σ.Ε.Π. Ελληνικό Ανοικτό Πανεπιστήμιο

Συν-Επιβλέπων Καθηγητής:

Αικατερίνη Καμπάση

Σ.Ε.Π. Ελληνικό Ανοικτό Πανεπιστήμιο

Πάτρα, Ιούνιος 2026

Πρόλογος

Η παρούσα διπλωματική εργασία αποτελεί το αποτέλεσμα μιας επίπονης και δημιουργικής πορείας, την οποία θα ήταν αδύνατον να ολοκληρώσω χωρίς την πολύτιμη συμβολή και υποστήριξη ορισμένων σημαντικών ανθρώπων και φορέων, τους οποίους αισθάνομαι την ανάγκη να ευχαριστήσω θερμά.

Πρωτίστως, θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες στον Επίκουρο Καθηγητή Περικό Ισίδωρο, επιβλέπων της παρούσας εργασίας, για την εμπειριστατωμένη επιστημονική καθοδήγηση, την αδιάκοπη υποστήριξη και τις καίριες παρατηρήσεις του καθ' όλη τη διάρκεια της εκπόνησης της μελέτης.

Τέλος, επιθυμώ να εκφράσω την ευγνωμοσύνη μου προς την οικογένειά μου και την κοπέλα μου, για την ηθική στήριξη, την κατανόηση και την ενθάρρυνση που μου παρείχαν σε όλη τη διάρκεια των σπουδών μου και ιδιαίτερα κατά την εκπόνηση αυτής της εργασίας.

Σε όλους τους ανωτέρω, απευθύνω τις πιο θερμές μου ευχαριστίες.

Περίληψη

Η παρούσα πτυχιακή εργασία πραγματεύεται την αναγνώριση και κατηγοριοποίηση δερματικών βλαβών, με ιδιαίτερη έμφαση στο μελάνωμα, μέσω τεχνικών βαθιάς μάθησης και μεθόδων εξηγησιμότητας. Στόχος της εργασίας ήταν να διερευνηθεί κατά πόσο σύγχρονα μοντέλα τεχνητής νοημοσύνης μπορούν να υποστηρίξουν την ανάλυση δερματοσκοπικών εικόνων, όχι μόνο ως προς την τελική πρόβλεψη, αλλά και ως προς τον τρόπο με τον οποίο οδηγούνται σε αυτή.

Στο θεωρητικό μέρος παρουσιάστηκε το βασικό υπόβαθρο των δερματικών βλαβών, η σημασία της έγκαιρης ανίχνευσης του μελανώματος, καθώς και ο ρόλος της τεχνητής νοημοσύνης στην ιατρική εικόνα. Παράλληλα, εξετάστηκαν βασικά σύνολα δεδομένων που χρησιμοποιούνται σε αντίστοιχες εφαρμογές, όπως τα HAM10000, ISIC, PH2 και Derm7pt, καθώς και οι μετρικές αξιολόγησης που είναι απαραίτητες σε προβλήματα με ανισορροπία κατηγοριών.

Στο πειραματικό μέρος υλοποιήθηκαν και αξιολογήθηκαν διαφορετικές αρχιτεκτονικές βαθιάς μάθησης για δυαδική και πολυκατηγορική ταξινόμηση δερματοσκοπικών εικόνων. Δοκιμάστηκαν συνελκτικά νευρωνικά δίκτυα, μοντέλα μετασχηματιστών και υβριδικές προσεγγίσεις, ενώ εφαρμόστηκαν τεχνικές προεπεξεργασίας, διαχωρισμού δεδομένων και αξιολόγησης με βάση μετρικές όπως *accuracy*, *precision*, *recall*, *F1-score* και *AUC*. Τα αποτελέσματα έδειξαν ότι τα μοντέλα μπορούν να πετύχουν ικανοποιητική απόδοση, ειδικά όταν η εκπαίδευση γίνεται με σωστά οργανωμένα δεδομένα και κατάλληλη πειραματική διαδικασία. Παράλληλα, αναδείχθηκε ότι η ακρίβεια από μόνη της δεν αρκεί, καθώς η αξιολόγηση πρέπει να λαμβάνει υπόψη και την απόδοση στις λιγότερο συχνές ή πιο κρίσιμες κατηγορίες.

Τέλος, εφαρμόστηκαν τεχνικές εξηγησιμότητας, όπως *Grad-CAM*, *Grad-CAM++*, *SHAP* και *LIME*, με σκοπό την οπτική ερμηνεία των αποφάσεων των μοντέλων. Οι τεχνικές αυτές έδειξαν ότι σε αρκετές περιπτώσεις τα μοντέλα εστίαζαν σε περιοχές σχετικές με τη βλάβη, ενώ σε άλλες περιπτώσεις ανέδειξαν περιορισμούς και ανάγκη προσεκτικότερης αξιολόγησης. Συνολικά, η εργασία καταλήγει ότι η βαθιά μάθηση μπορεί να αποτελέσει χρήσιμο υποστηρικτικό εργαλείο στην ανάλυση δερματοσκοπικών εικόνων, όταν συνδυάζεται με κατάλληλη αξιολόγηση και ερμηνευσιμότητα.

Λέξεις – Κλειδιά

Μελάνωμα, Δερματοσκοπικές εικόνες, Βαθιά μάθηση, Ταξινόμηση δερματικών βλαβών, Vision Transformers, Εξηγησιμότητα μοντέλων.

Recognition and Classification of Melanoma Using Transformers and Deep Learning Techniques

Alexandros Bistolas

Abstract

This thesis focuses on the recognition and classification of skin lesions, with particular emphasis on melanoma, using deep learning techniques and explainability methods. The main objective was to investigate whether modern artificial intelligence models can support the analysis of dermoscopic images, not only through their final predictions, but also through a better understanding of the visual features on which these predictions are based.

The theoretical part of the thesis presents the basic background of skin lesions, the clinical importance of early melanoma detection, and the role of artificial intelligence in medical image analysis. It also discusses commonly used datasets for skin lesion classification, such as HAM10000, ISIC, PH2, and Derm7pt, along with evaluation metrics that are necessary in classification tasks with imbalanced classes.

In the experimental part, different deep learning architectures were implemented and evaluated for binary and multiclass classification of dermoscopic images. Convolutional neural networks, transformer-based models, and hybrid approaches were tested. The experimental process included image preprocessing, dataset splitting, model training, and evaluation using metrics such as accuracy, precision, recall, F1-score, and AUC. The results showed that deep learning models can achieve satisfactory performance in the recognition of skin lesions, especially when the data are properly organized and the training process is carefully designed. At the same time, the experiments highlighted that accuracy alone is not sufficient, since the performance of a model must also be examined in relation to less frequent or clinically more critical classes.

Finally, explainability techniques such as Grad-CAM, Grad-CAM++, SHAP, and LIME were applied in order to visually interpret the decisions of the models. These methods showed that, in several cases, the models focused on image regions related to the lesion, while in other cases they revealed limitations that require further examination. Overall, the thesis concludes that deep learning can serve as a useful supportive tool in dermoscopic image analysis, provided that model performance is combined with careful evaluation and interpretability.

Keywords

Melanoma, Dermoscopic Images, Deep Learning, Skin Lesion Classification, Vision Transformers, Explainable AI.

Πίνακας περιεχομένων

ΠΡΟΛΟΓΟΣ	IV
ΠΕΡΙΛΗΨΗ	V
ABSTRACT	VI
ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ / ΣΧΗΜΑΤΩΝ	VIII
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ	XI
ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ & ΑΚΡΩΝΥΜΙΑ	XII
1. ΕΙΣΑΓΩΓΗ	1
1.1 ΚΑΡΚΙΝΟΣ ΤΟΥ ΔΕΡΜΑΤΟΣ ΚΑΙ ΔΙΑΓΝΩΣΤΙΚΗ ΔΥΣΚΟΛΙΑ	1
1.2 ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΕΞΗΓΗΣΙΜΟΤΗΤΑ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΡΜΑΤΟΣΚΟΠΙΚΩΝ ΕΙΚΟΝΩΝ	4
1.3 ΣΚΟΠΟΣ ΤΗΣ ΠΑΡΟΥΣΑΣ ΕΡΓΑΣΙΑΣ	6
1.4 ΔΟΜΗ ΤΗΣ ΕΡΓΑΣΙΑΣ	6
2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ	8
2.1 ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ	8
2.2 ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΣΤΗ ΒΙΒΛΙΟΓΡΑΦΙΑ	24
2.2.1 ISIC 2016–2018	25
2.2.2 ISIC 2019	26
2.2.3 ISIC 2020 / SIIM-ISIC	28
2.2.4 HAM10000	29
2.2.5 PH2	30
2.2.6 DERM7PT	32
2.2.7 SKINCON	33
2.2.8 VITILIGO DATASET	34
2.2.9 MONKEYPOX DATASET	35
3. ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΤΕΧΝΙΚΕΣ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ	38
3.1 FINE-TUNING ΣΕ ΜΟΝΤΕΛΑ ΤΑΞΙΝΟΜΗΣΗΣ (CNN & TRANSFORMER BACKBONES)	39
3.2 ENSEMBLE LEARNING (SIMPLE AVERAGE & WEIGHTED ENSEMBLE)	40
3.3 ΕΞΗΓΗΣΙΜΟΤΗΤΑ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΕ ΣΤΑ ΠΕΙΡΑΜΑΤΑ (GRAD-CAM, GRAD-CAM++, SCORE-CAM, LIME, SHAP, OCCLUSION, UMAP ΚΑΙ T-SNE)	42
3.4 ΠΕΙΡΑΜΑΤΙΚΟ ΠΕΡΙΒΑΛΛΟΝ ΚΑΙ ΕΡΓΑΛΕΙΑ ΥΛΟΠΟΙΗΣΗΣ	45
3.5 ΠΡΩΤΟΚΟΛΛΟ ΑΞΙΟΛΟΓΗΣΗΣ ΚΑΙ ΕΠΙΛΟΓΗ ΚΑΛΥΤΕΡΟΥ ΜΟΝΤΕΛΟΥ	45
4. ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ	47
4.1 HAM10000 ΜΕ RESNET50 ΓΙΑ ΔΥΑΔΙΚΗ ΤΑΞΙΝΟΜΗΣΗ	47
4.2 ΠΕΙΡΑΜΑ ΔΥΑΔΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ ΣΤΟ PH2 ΜΕ 5-FOLD CROSS-VALIDATION, GRAD-CAM++ ΚΑΙ SHAP	54
4.3 ΤΜΑΗU-NET ΣΤΟ PH2 ΓΙΑ ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΔΕΡΜΑΤΙΚΩΝ ΒΛΑΒΩΝ	59
4.4 ΠΕΙΡΑΜΑ ΠΟΛΥΤΑΞΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ MONKEYPOX SKIN IMAGE DATASET ΜΕ DENSENET121	68

4.5 T-SNE ΚΑΙ UMAP ΣΤΟ HAM10000 ΜΕ EFFICIENTNETV2-S UMAP, GRAD-CAM++ ΚΑΙ SHAP	75
4.6 SCORE-CAM ΜΕ ENSEMBLE ΜΟΝΤΕΛΩΝ ΣΤΟ ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ SKIN CANCER 7 ΚΛΑΣΕΩΝ	82
4.7 T-FORMER ΣΤΟ DERM7PT ΓΙΑ ΔΥΑΔΙΚΗ ΤΑΞΙΝΟΜΗΣΗ MELANOMA / NON-MELANOMA	91
4.8 U-NET ΣΤΟ ISIC 2016 ΓΙΑ ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΔΕΡΜΑΤΙΚΩΝ ΒΛΑΒΩΝ	100
4.9 ΠΕΙΡΑΜΑ ΜΕΤΑΦΟΡΑΣ ΜΑΘΗΣΗΣ ΑΠΟ HAM10000 ΣΕ PH2 ΜΕ RESNET50, GRAD-CAM++ ΚΑΙ SHAP.....	108
4.10 ΠΕΙΡΑΜΑ ACTIVE LEARNING ΣΤΟ HAM10000 ΜΕ EFFICIENTNET-B4 ΚΑΙ ROI CROP.....	113
4.11 ΠΕΙΡΑΜΑ ΤΑΞΙΝΟΜΗΣΗΣ MELANOMA / NON-MELANOMA ΜΕ ΕΝΔΙΑΜΕΣΑ ΔΕΡΜΑΤΟΣΚΟΠΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΣΤΟ DERM7PT.....	119
4.12 FAT-NET ΤΜΗΜΑΤΟΠΟΙΗΣΗΣ ΣΤΟ ISIC 2016	126
4.13 ΠΕΙΡΑΜΑ VGG-U-NET ΣΤΟ ISIC 2016 ΜΕ GRAD-CAM ΚΑΙ SHAP.....	130
4.14 ΠΕΙΡΑΜΑ PAD-UFES-20 ΜΕ ΤΡΙΑ BACKBONES ΚΑΙ ΕΞΗΓΗΣΙΜΟΤΗΤΑ GRAD-CAM++ / SHAP	136
4.15 ΠΟΛΥΤΑΞΙΚΗ ΤΑΞΙΝΟΜΗΣΗ ΣΤΟ HAM10000 ΜΕ PAM-CNN-BILSTM ΚΑΙ ΕΞΗΓΗΣΙΜΟΤΗΤΑ GRAD-CAM++ / SHAP	143
4.16 ΠΕΙΡΑΜΑ ΔΥΑΔΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ ΣΤΟ ISIC 2019 ΜΕ EFFICIENTNET-B0, GRAD-CAM++ ΚΑΙ T-SNE	151
4.17 ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ SWIN-TINY + SWIGLU ΚΑΙ EFFICIENTNET-B4 ΣΤΟ ISIC 2019.....	156
4.18 ΠΕΙΡΑΜΑ ΔΥΑΔΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ ΣΤΟ ISIC 2019 ΜΕ EFFICIENTNET-B3 ΚΑΙ TEST-TIME AUGMENTATION.....	161
4.19 ΣΥΓΚΡΙΣΗ CNN ΚΑΙ VISION TRANSFORMER ΜΟΝΤΕΛΩΝ ΣΤΟ ISIC 2019	165
4.20 ΠΕΙΡΑΜΑ ΔΥΑΔΙΚΗΣ ΤΑΞΙΝΟΜΗΣΗΣ ΣΤΟ ISIC 2019 ΜΕ EFFICIENTNET-B0, ΑΝΑΛΥΣΗ FEATURE SPACE ΚΑΙ ΕΞΗΓΗΣΙΜΟΤΗΤΑ	170
5. ΤΕΧΝΙΚΕΣ ΕΞΗΓΗΣΙΜΟΤΗΤΑΣ ΚΑΙ ΟΠΤΙΚΗ ΕΡΜΗΝΕΙΑ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ.....	179
5.1 Η ΑΝΑΓΚΗ ΓΙΑ ΕΞΗΓΗΣΙΜΟΤΗΤΑ ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΙΑΤΡΙΚΗΣ ΕΙΚΟΝΑΣ	179
5.2 GRAD-CAM ΚΑΙ ΟΠΤΙΚΗ ΕΡΜΗΝΕΙΑ ΠΡΟΒΛΕΨΕΩΝ	180
5.3 GRAD-CAM++ ΚΑΙ ΠΙΟ ΛΕΠΤΟΜΕΡΗΣ ΕΣΤΙΑΣΗ	180
5.4 SCORE-CAM ΚΑΙ ΣΥΓΚΡΙΣΗ ΜΕ ΤΙΣ GRADIENT-BASED ΤΕΧΝΙΚΕΣ	181
5.5 SHAP ΚΑΙ ΑΝΑΛΥΣΗ ΣΥΜΒΟΛΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ.....	182
5.6 ACROSS-EPOCH EXPLAINABILITY: ΕΞΕΛΙΞΗ ΤΗΣ ΠΡΟΣΟΧΗΣ ΚΑΤΑ ΤΗΝ ΕΚΠΑΙΔΕΥΣΗ.....	183
5.7 ΑΝΑΛΥΣΗ ΣΩΣΤΩΝ ΠΡΟΒΛΕΨΕΩΝ.....	185
5.8 ΑΝΑΛΥΣΗ ΛΑΝΘΑΣΜΕΝΩΝ Η ΔΥΣΚΟΛΩΝ ΠΡΟΒΛΕΨΕΩΝ	186
5.9 ΣΥΓΚΡΙΣΗ HEATMAPS ΚΑΙ SHAP ΜΕ SEGMENTATION MASKS	187
5.10 ΤΕΚΜΗΡΙΩΣΗ GITHUB	188
6. ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ	189
6.1 ΣΥΝΕΙΣΦΟΡΑ ΤΗΣ ΕΡΓΑΣΙΑΣ	190
2.2 ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ	190

BIBΛΙΟΓΡΑΦΙΑ.....	192
-------------------	-----

Κατάλογος Εικόνων / Σχημάτων

Εικόνα 1.1. Παγκόσμιο βάρος του δερματικού μελανώματος το 2020 και προβλέψεις έως το 2040. Πηγή: International Agency for Research on Cancer / World Health Organization.....	1
Εικόνα 1.2. Παρατηρούμενες και προβλεπόμενες ηλικιακά προσαρμοσμένες τιμές επίπτωσης και θνησιμότητας για το μελάνωμα στις ΗΠΑ, 1982–2030. Πηγή: CDC/MMWR.....	3
Εικόνα 2.1. Μερικά παραδείγματα δερματοσκοπικών εικόνων από τα σύνολα δεδομένων ISIC-2016 και ISIC-2017. Πηγή: Ashraf et al. (2022).....	25
Εικόνα 2.2. Παραδείγματα εικόνων από τις επτά κατηγορίες δερματικών βλαβών του HAM10000 dataset: AKIEC, BCC, BKL, DF, MEL, NV και VASC. Πηγή: Mehmood et al. (2023).....	26
Εικόνα 2.3. Κατανομή των βασικών κλάσεων στο ISIC 2019 training set.....	27
Εικόνα 2.4. Ενδεικτικά δείγματα δερματοσκοπικών εικόνων από το ISIC 2019 dataset για τις κατηγορίες AK, BCC, BKL, DF, MEL, NV, SCC και VASC. Πηγή: A deep analysis on high-resolution dermoscopic image classification (2021).	27
Εικόνα 2.5. Παραδείγματα εικόνων από το σύνολο δεδομένων ISIC-2020 με διαφορετικά χαρακτηριστικά. Πηγή: Söylemez, Ö. F. (2023). Forward selection-based ensemble of deep neural networks for melanoma classification in dermoscopy images.....	28
Εικόνα 2.6. Κατανομή των επτά κλάσεων στο HAM10000 dataset.....	30
Εικόνα 2.7. Ενδεικτικές εικόνες για τις επτά κατηγορίες δερματικών βλαβών από το σύνολο δεδομένων HAM10000. Πηγή: Shetty et al., Skin lesion classification of dermoscopic images using machine learning and convolutional neural network, Scientific Reports, 2022.....	30
Εικόνα 2.8. Κατανομή των τριών κατηγοριών στο PH2 dataset.....	31
Εικόνα 2.9. Δείγματα δερματικών παθήσεων από το σύνολο δεδομένων PH2. Πηγή: Abunadi, I., & Senan, E. M. (2021).....	32
Εικόνα 2.10. Ενδεικτικά παραδείγματα εννοιών / χαρακτηριστικών δερματικών βλαβών από το σύνολο δεδομένων Derm7pt. Πηγή: Lucieri et al., On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors, 2020.....	33
Εικόνα 2.11. Παραδείγματα κλινικών εννοιών από το SkinCon dataset. Η εικόνα παρουσιάζει δείγματα για τα concepts plaque, pustule, telangiectasia και ulcer, δείχνοντας ότι το SkinCon δεν βασίζεται μόνο σε τελικές διαγνώσεις, αλλά σε οπτικά/κλινικά χαρακτηριστικά που μπορούν να αξιοποιηθούν για εξηγησιμότητα και ανάλυση σφαλμάτων μοντέλων. Πηγή: Daneshjou et al.	34
Σχήμα 3.1. Γενικό workflow της πειραματικής διαδικασίας.....	38
Σχήμα 3.2. Fine-tuning ταξινόμησης (CNN/Transformer backbone).....	39
Σχήμα 3.3. Ensemble learning (simple average & weighted ensemble).....	41
Σχήμα 3.4. Εφαρμογή τεχνικών εξηγησιμότητας μετά την πρόβλεψη του μοντέλου.....	44
Σχήμα 3.5. Συνοπτική απεικόνιση του πειραματικού περιβάλλοντος και των βασικών εργαλείων που χρησιμοποιήθηκαν για την υλοποίηση των πειραμάτων.....	45
Σχήμα 3.6. Ροή του πρωτοκόλλου αξιολόγησης και της επιλογής του καλύτερου μοντέλου, από την εκπαίδευση μέχρι την τελική αποτίμηση στο test set.....	46
Εικόνα 4.1. Πορεία train και validation loss του ResNet50 στο HAM10000.....	49
Εικόνα 4.2. Πορεία validation AUC και F1 κατά την εκπαίδευση.....	49
Εικόνα 4.3. Grad-CAM++ across-epoch ανάλυση για melanoma περίπτωση στο HAM10000.....	52
Εικόνα 4.4. SHAP ανάλυση για melanoma περίπτωση στο καλύτερο μοντέλο.....	53
Εικόνα 4.5. Grad-CAM++ across-epoch ανάλυση για χαρακτηριστική melanoma περίπτωση από το Fold 1.....	57
Εικόνα 4.6. SHAP ανάλυση της ίδιας melanoma περίπτωσης στο καλύτερο μοντέλο.....	58
Εικόνα 4.7. SHAP across-epoch ανάλυση για χαρακτηριστική melanoma περίπτωση από το Fold 1.....	59
Εικόνα 4.8. Παράδειγμα πρόβλεψης μάσκας από το μοντέλο TMAHU-Net στο PH2.....	60
Εικόνα 4.9. Grad-CAM ανάλυση για την πρόβλεψη τμηματοποίησης του TMAHU-Net.....	61
Εικόνα 4.10. Grad-CAM++ και foreground ερμηνεία της προβλεπόμενης μάσκας.....	61
Εικόνα 4.11. LIME ανάλυση για την καλύτερη περίπτωση τμηματοποίησης.....	62
Εικόνα 4.12. LIME ανάλυση για δύσκολη περίπτωση τμηματοποίησης.....	62

Εικόνα 4.13. Εξέλιξη της πρόβλεψης και της Grad-CAM++ ερμηνείας ανά εποχή για την καλύτερη περίπτωση.....	64
Εικόνα 4.14. Εξέλιξη της πρόβλεψης και της Grad-CAM++ ερμηνείας ανά εποχή για τη δυσκολότερη περίπτωση.....	66
Εικόνα 4.15. LIME across-epoch ανάλυση για την καλύτερη και τη δυσκολότερη περίπτωση.....	67
Εικόνα 4.16. Grad-CAM++ για ορθή πρόβλεψη Monkeyrox με πιθανότητα 97.81%.....	70
Εικόνα 4.17. Grad-CAM++ για λανθασμένη πρόβλεψη Chickenrox ως Monkeyrox.....	71
Εικόνα 4.18. Grad-CAM++ across-epoch ανάλυση για ορθή πρόβλεψη Monkeyrox.....	72
Εικόνα 4.19. Grad-CAM++ across-epoch ανάλυση για λανθασμένη πρόβλεψη Chickenrox ως Monkeyrox.....	74
Εικόνα 4.20. LIME ανάλυση για ορθή πρόβλεψη Monkeyrox.....	74
Εικόνα 4.21. LIME ανάλυση για λανθασμένη πρόβλεψη Chickenrox ως Monkeyrox.....	74
Εικόνα 4.22. t-SNE απεικόνιση των χαρακτηριστικών του EfficientNetV2-S στο HAM10000.....	78
Εικόνα 4.23. UMAP απεικόνιση των χαρακτηριστικών του EfficientNetV2-S στο HAM10000.....	79
Εικόνα 4.24. Grad-CAM και Grad-CAM++ across-epoch ανάλυση για χαρακτηριστική περίπτωση του HAM10000.....	81
Εικόνα 4.25. SHAP ανάλυση για χαρακτηριστική πρόβλεψη του EfficientNetV2-S.....	81
Εικόνα 4.26. Score-CAM ανά backbone και ensemble για σωστή πρόβλεψη melanoma.....	86
Εικόνα 4.27. Score-CAM across-epoch ανάλυση για σωστή πρόβλεψη melanoma.....	87
Εικόνα 4.28. Score-CAM across-epoch ανάλυση για non-melanoma περίπτωση, σωστή σε δυαδικό αλλά όχι σε πολυταξικό επίπεδο.....	88
Εικόνα 4.29. Score-CAM across-epoch ανάλυση για false positive περίπτωση melanocytic nevi ως melanoma.....	89
Εικόνα 4.30. Score-CAM across-epoch ανάλυση για false negative melanoma.....	90
Εικόνα 4.31. Σύγκριση Score-CAM του ensemble και του EfficientNetV2-S για περίπτωση melanoma.....	90
Εικόνα 4.32. Σύγκριση Score-CAM του ensemble και του EfficientNetV2-S για περίπτωση melanocytic nevi.....	90
Εικόνα 4.33. Grad-CAM++ ανάλυση για ορθή πρόβλεψη melanoma στο πολυτροπικό T-Former.....	94
Εικόνα 4.34. Grad-CAM++ ανάλυση για ορθή πρόβλεψη non-melanoma στο πολυτροπικό T-Former.....	95
Εικόνα 4.35. Grad-CAM++ ανάλυση για λανθασμένη πρόβλεψη non-melanoma ως melanoma.....	96
Εικόνα 4.36. Grad-CAM++ across-epoch ανάλυση για ορθή πρόβλεψη melanoma.....	97
Εικόνα 4.37. Grad-CAM++ across-epoch ανάλυση για ορθή πρόβλεψη non-melanoma.....	98
Εικόνα 4.38. Grad-CAM++ across-epoch ανάλυση για λανθασμένη πρόβλεψη non-melanoma ως melanoma.....	99
Εικόνα 4.39. Πορεία εκπαίδευσης του U-Net στο ISIC 2016.....	101
Εικόνα 4.40. Ενδεικτικά παραδείγματα τμηματοποίησης από το U-Net στο validation set.....	102
Εικόνα 4.41. Grad-CAM++ ανάλυση για παράδειγμα τμηματοποίησης στο U-Net.....	103
Εικόνα 4.42. Boundary CAM ανάλυση για την περιοχή των ορίων της βλάβης.....	104
Εικόνα 4.43. Across-epoch ανάλυση για την καλύτερη περίπτωση τμηματοποίησης.....	105
Εικόνα 4.44. Across-epoch ανάλυση για τη δυσκολότερη περίπτωση τμηματοποίησης.....	107
Εικόνα 4.45. Grad-CAM++ across-epoch ανάλυση κατά το fine-tuning στο PH2.....	112
Εικόνα 4.46. Melanoma-targeted SHAP ανάλυση σε benign περίπτωση με τοπικό melanoma signal.....	112
Εικόνα 4.47. Σύγκριση precision ανάμεσα σε random sampling, margin sampling και full baseline.....	116
Εικόνα 4.48. Σύγκριση recall ανάμεσα σε random sampling, margin sampling και full baseline.....	117
Εικόνα 4.49. Σύγκριση F1-score ανάμεσα σε random sampling, margin sampling και full baseline.....	117
Εικόνα 4.50. Σύγκριση AUC ανάμεσα σε random sampling, margin sampling και full baseline.....	118
Εικόνα 4.51. Ροή του μοντέλου με ενδιάμεσα δερματοσκοπικά χαρακτηριστικά.....	120
Εικόνα 4.52. Επίδραση του threshold στην απόδοση της διάγνωσης.....	122
Εικόνα 4.53. Απόδοση του μοντέλου στην πρόβλεψη των δερματοσκοπικών χαρακτηριστικών.....	123
Εικόνα 4.54. Ανάλυση διάγνωσης και δερματοσκοπικών χαρακτηριστικών για ορθή πρόβλεψη melanoma.....	124
Εικόνα 4.55. Ανάλυση διάγνωσης και δερματοσκοπικών χαρακτηριστικών για false negative melanoma.....	125
Εικόνα 4.56. Ενδεικτικά αποτελέσματα τμηματοποίησης του FAT-Net-inspired μοντέλου στο test set.....	128
Εικόνα 4.57. Καλύτερη περίπτωση τμηματοποίησης στο test set, με IoU = 95.88% και Dice = 97.90%.....	129
Εικόνα 4.58. Ενδιάμεση περίπτωση τμηματοποίησης στο test set.....	130
Εικόνα 4.59. Ενδεικτικά αποτελέσματα τμηματοποίησης του VGG-U-Net στο validation set.....	132
Εικόνα 4.60. Grad-CAM ανάλυση του VGG-U-Net για παράδειγμα τμηματοποίησης.....	133
Εικόνα 4.61. Boundary contrast CAM για την ανάλυση των ορίων της βλάβης.....	134

Εικόνα 4.62. Grad-CAM++ across-epoch ανάλυση του VGG-U-Net στο ISIC 2016.....	135
Εικόνα 4.63. SHAP ανάλυση για πρόβλεψη τμηματοποίησης του VGG-U-Net.....	135
Εικόνα 4.64. Σύγκριση των backbones ως προς Accuracy, Precision, Recall, F1 και AUC	139
Εικόνα 4.65. Grad-CAM και Grad-CAM++ ανάλυση για ορθή πρόβλεψη suspicious στο PAD-UFES-20	140
Εικόνα 4.66. Grad-CAM και Grad-CAM++ ανάλυση για ορθή πρόβλεψη benign στο PAD-UFES-20.....	141
Εικόνα 4.67. Grad-CAM και Grad-CAM++ ανάλυση για λανθασμένη πρόβλεψη suspicious ως benign ...	141
Εικόνα 4.68. SHAP ανάλυση για ορθή πρόβλεψη suspicious στο PAD-UFES-20	142
Εικόνα 4.69. Εξέλιξη της απώλειας εκπαίδευσης/επικύρωσης και των validation μετρικών κατά τη διάρκεια της εκπαίδευσης	144
Εικόνα 4.70. Πίνακας σύγχυσης του PAM-CNN-BiLSTM στο test set του HAM10000	146
Εικόνα 4.71. Ενδεικτικά σωστά παραδείγματα και αντίστοιχοι Grad-CAM και Grad-CAM++ χάρτες	149
Εικόνα 4.72. Παράδειγμα λανθασμένης πρόβλεψης και αντίστοιχο Grad-CAM++ χάρτη	150
Εικόνα 4.73. Ενδεικτικές SHAP απεικονίσεις για σωστές ταξινομήσεις στο HAM10000.....	151
Εικόνα 4.74. Πίνακας σύγχυσης του EfficientNet-B0 στο test set του ISIC 2019.....	153
Εικόνα 4.75. t-SNE απεικόνιση των χαρακτηριστικών του EfficientNet-B0 ανά δυαδική κλάση	154
Εικόνα 4.76. t-SNE απεικόνιση χρωματισμένη με βάση την πιθανότητα suspicious	155
Εικόνα 4.77. Grad-CAM ανάλυση λανθασμένης suspicious περίπτωσης που ταξινομήθηκε ως benign-like	155
Εικόνα 4.78. Grad-CAM και Grad-CAM++ ανάλυση για ορθή benign-like πρόβλεψη.....	156
Εικόνα 4.79. Confusion matrix του μοντέλου Swin-Tiny + SwiGLU και του μοντέλου EfficientNet-B4 στο test set.	159
Εικόνα 4.80. Σωστή suspicious πρόβλεψη του Swin-Tiny + SwiGLU με Grad-CAM.	160
Εικόνα 4.81. Λανθασμένη suspicious πρόβλεψη του Swin-Tiny + SwiGLU.....	160
Εικόνα 4.82. Λανθασμένη suspicious και benign_like πρόβλεψη του EfficientNet-B4.....	160
Εικόνα 4.83. Πίνακας σύγχυσης TTA	164
Εικόνα 4.84. Σύγκριση standard inference και Test-Time Augmentation ως προς τις βασικές μετρικές.....	165
Εικόνα 4.85. Σύγκριση CNN και Vision Transformer μοντέλων ως προς Accuracy, F1 και AUC.....	167
Εικόνα 4.86. Μέση απόδοση ανά οικογένεια μοντέλων	168
Εικόνα 4.87. Πίνακας σύγχυσης του καλύτερου Vision Transformer μοντέλου	169
Εικόνα 4.88. Πίνακας σύγχυσης του EfficientNet-B0 στο test set του ISIC 2019.....	171
Εικόνα 4.89. Εξέλιξη του feature space του EfficientNet-B0 κατά τη διάρκεια της εκπαίδευσης	174
Εικόνα 4.90. Εξέλιξη της βεβαιότητας του μοντέλου στο feature space. Ο χρωματισμός αντιστοιχεί στην πιθανότητα που αποδίδει το EfficientNet-B0 στην κλάση suspicious.	174
Εικόνα 4.91. Grad-CAM overlay και Grad-CAM across-epoch ανάλυση για σωστή benign-like πρόβλεψη	175
Εικόνα 4.92. Grad-CAM overlay και Grad-CAM across-epoch ανάλυση για σωστή suspicious πρόβλεψη.....	176
Εικόνα 4.93. Grad-CAM across-epoch ανάλυση σε λανθασμένη πρόβλεψη benign-like ως suspicious	177
Εικόνα 4.94. SHAP ανάλυση για λανθασμένη πρόβλεψη benign-like ως suspicious	177
Εικόνα 4.95. Ερμηνεία σωστής benign-like πρόβλεψης με Integrated Gradients και SmoothGrad	178
Εικόνα 5.1. Παράδειγμα Grad-CAM σε σωστά ταξινομημένη δερματική βλάβη	180
Εικόνα 5.2. Σύγκριση Grad-CAM και Grad-CAM++ στην ίδια δερματοσκοπική εικόνα.....	181
Εικόνα 5.3. Score-CAM σε 3 backbone για σωστή πρόβλεψη melanoma.....	182
Εικόνα 5.4. Παράδειγμα μελανώματος με εξηγησιμότητα SHAP	183
Εικόνα 5.5. Grad-CAM++ across-epoch ανάλυση για melanoma περίπτωση στο HAM10000	185
Εικόνα 5.6. Σωστή suspicious πρόβλεψη του Swin-Tiny + SwiGLU με Grad-CAM	186
Εικόνα 5.7. Grad-CAM ανάλυση λανθασμένης suspicious περίπτωσης που ταξινομήθηκε ως benign-like.....	187
Εικόνα 5.8. Σύγκριση segmentation mask με CAM heatmaps και SHAP overlay σε δερματοσκοπική εικόνα. Η εικόνα δείχνει κατά πόσο η οπτική εστίαση του μοντέλου αντιστοιχεί στην περιοχή της βλάβης.....	188

Κατάλογος Πινάκων

Πίνακας 2.1. Συγκριτικός πίνακας των papers.....	21
Πίνακας 2.2. Συγκριτικός πίνακας των datasets των εργασιών	37
Πίνακας 4.1. Δυαδική κατανομή του HAM10000 στο πείραμα ResNet50	87
Πίνακας 4.2. Κατανομή train, validation και test στο πείραμα ResNet50	112
Πίνακας 4.3. Βασικές validation μετρικές του ResNet50 στο HAM10000	127
Πίνακας 4.4. Δυαδική κατανομή του PH2 στο πείραμα 5-fold cross-validation	151
Πίνακας 4.5. Κατανομή δεδομένων σε κάθε fold του πειράματος	54

Πίνακας 4.6. Καλύτερη επίδοση ανά fold με βάση το validation AUC.....	54
Πίνακας 4.7. Μέσες επιδόσεις 5-fold cross-validation στο PH2.....	54
Πίνακας 4.8. Συνολικές μετρικές αξιολόγησης του DenseNet121 στο Monkeypox Skin Image Dataset.....	67
Πίνακας 4.9. Precision, recall και F1-score ανά κλάση για το πείραμα DenseNet121.....	68
Πίνακας 4.10. Πίνακας σύγχυσης του DenseNet121 στο test set του Monkeypox Skin Image Dataset.....	70
Πίνακας 4.11. Κατανομή των κλάσεων στο HAM10000 για το πείραμα EfficientNetV2-S.....	75
Πίνακας 4.12. Πορεία εκπαίδευσης του EfficientNetV2-S στο HAM10000.....	75
Πίνακας 4.13. Test αποτελέσματα του EfficientNetV2-S στο HAM10000.....	76
Πίνακας 4.14. Απόδοση ανά κλάση για το EfficientNetV2-S στο HAM10000.....	76
Πίνακας 4.15. Πίνακας σύγχυσης του EfficientNetV2-S στο test set.....	79
Πίνακας 4.16. Κατανομή κλάσεων στο training subset του πειράματος Score-CAM.....	81
Πίνακας 4.17. Καλύτερη validation επίδοση ανά backbone με βάση το macro recall.....	81
Πίνακας 4.18. Συνολικές μετρικές του ensemble στο validation set.....	82
Πίνακας 4.19. Precision, recall και F1-score ανά κλάση για το ensemble.....	83
Πίνακας 4.20. Πίνακας σύγχυσης του ensemble στο validation set.....	87
Πίνακας 4.21. Κατανομή των κλάσεων στο πείραμα T-Former Derm7pt.....	90
Πίνακας 4.22. Πορεία εκπαίδευσης του T-Former στο Derm7pt.....	91
Πίνακας 4.23. Πίνακας σύγχυσης του T-Former στο test set με threshold 0.50.....	95
Πίνακας 4.24. Πίνακας σύγχυσης του T-Former στο test set μετά τη βελτιστοποίηση threshold.....	96
Πίνακας 4.25. Βασικές μετρικές εκπαίδευσης του U-Net στο ISIC 2016.....	100
Πίνακας 4.26. Επίδραση του threshold στις validation μετρικές του U-Net.....	101
Πίνακας 4.27. Κατανομή των δεδομένων HAM10000 στο στάδιο προεκπαίδευσης.....	107
Πίνακας 4.28. Πορεία προεκπαίδευσης του ResNet50 στο HAM10000.....	108
Πίνακας 4.29. Test αποτελέσματα στο HAM10000 πριν τη μεταφορά στο PH2.....	108
Πίνακας 4.30. Κατανομή των δεδομένων PH2 στο στάδιο fine-tuning.....	109
Πίνακας 4.31. Πορεία fine-tuning του ResNet50 στο PH2.....	109
Πίνακας 4.32. Test αποτελέσματα στο PH2 μετά το fine-tuning.....	110
Πίνακας 4.33. Δυναμική κατανομή του HAM10000 στο πείραμα active learning.....	111
Πίνακας 4.34. Κατανομή train, validation και test στο πείραμα EfficientNet-B4 ROI.....	112
Πίνακας 4.35. Αποτελέσματα active learning με random sampling.....	113
Πίνακας 4.36. Αποτελέσματα active learning με margin sampling.....	113
Πίνακας 4.37. Δυναμική κατανομή των δεδομένων στο Derm7pt.....	119
Πίνακας 4.38. Κατανομή train, validation και test στο πείραμα Derm7pt με ενδιάμεσα χαρακτηριστικά.....	119
Πίνακας 4.39. Πορεία εκπαίδευσης του μοντέλου στο Derm7pt.....	121
Πίνακας 4.40. Κατανομή δεδομένων στο πείραμα FAT-Net-inspired τμηματοποίησης στο ISIC 2016.....	125
Πίνακας 4.41. Πορεία εκπαίδευσης του FAT-Net-inspired μοντέλου στο ISIC 2016.....	125
Πίνακας 4.42. Κατανομή δεδομένων στο πείραμα VGG-U-Net στο ISIC 2016.....	129
Πίνακας 4.43. Πορεία εκπαίδευσης του VGG-U-Net στο ISIC 2016.....	129
Πίνακας 4.44. Κατανομή των κατηγοριών στο PAD-UFES-20.....	134
Πίνακας 4.45. Κατανομή train, validation και test στο πείραμα PAD-UFES-20.....	134
Πίνακας 4.46. Σύγκριση των τριών backbones στο PAD-UFES-20.....	136
Πίνακας 4.47. Κατανομή των κλάσεων στο HAM10000.....	140
Πίνακας 4.48. Κατανομή των κλάσεων σε train, validation και test σύνολο.....	141
Πίνακας 4.49. Συνολικά αποτελέσματα του PAM-CNN-BiLSTM στο test set του HAM10000.....	142
Πίνακας 4.50. Κατανομή του balanced subset στο πείραμα EfficientNet-B0.....	148
Πίνακας 4.51. Πορεία εκπαίδευσης του EfficientNet-B0 στο ISIC 2019.....	149
Πίνακας 4.52. Test αποτελέσματα του EfficientNet-B0 στο ISIC 2019.....	149
Πίνακας 4.53. Συγκριτικός πίνακας και μετρικές σε Swin-Tiny + SwiGLU και EfficientNet-B4.....	153
Πίνακας 4.54. Σύγκριση των confusion matrices με βάση τα ποσοστά ανά κλάση.....	154
Πίνακας 4.55. Κατανομή του balanced subset στο πείραμα EfficientNet-B3.....	158
Πίνακας 4.56. Πορεία εκπαίδευσης του EfficientNet-B3 στο ISIC 2019.....	159
Πίνακας 4.57. Κατανομή των δεδομένων στα train, validation και test σύνολα.....	162
Πίνακας 4.58. Μοντέλα που συγκρίθηκαν στο πείραμα CNN vs Vision Transformer.....	163
Πίνακας 4.59. Τελικά test αποτελέσματα των CNN και Vision Transformer μοντέλων.....	163
Πίνακας 4.60. Κατανομή των δεδομένων στα train, validation και test σύνολα.....	166
Πίνακας 4.61. Πορεία εκπαίδευσης του EfficientNet-B0 στο ISIC 2019.....	167
Πίνακας 4.62. Test αποτελέσματα του EfficientNet-B0 στο ISIC 2019.....	168
Πίνακας 4.63. Σύγκριση του feature space σε διαφορετικά στάδια της εκπαίδευσης.....	169

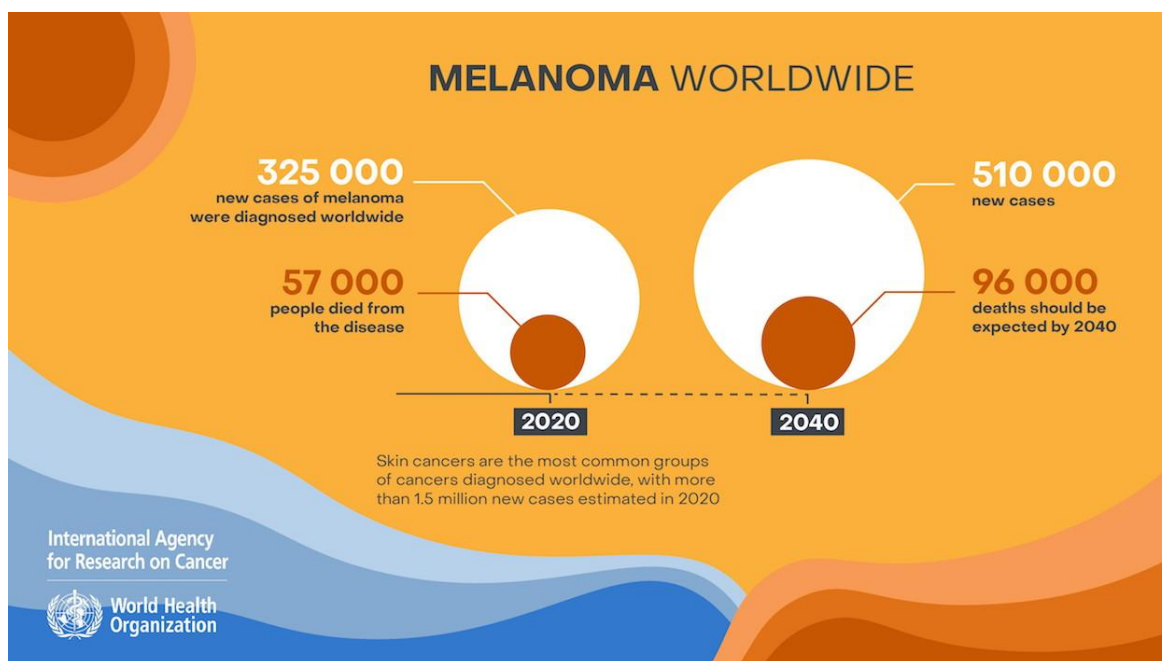
Συντομογραφίες & Ακρωνύμια

AI	Artificial Intelligence / Τεχνητή Νοημοσύνη
CNN	Convolutional Neural Network / Συνελκτικό Νευρωνικό Δίκτυο
DL	Deep Learning / Βαθιά Μάθηση
ML	Machine Learning / Μηχανική Μάθηση
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence / Εξηγήσιμη Τεχνητή Νοημοσύνη
Grad-CAM	Gradient-weighted Class Activation Mapping
Grad-CAM++	Βελτιωμένη εκδοχή της μεθόδου Grad-CAM
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
UMAP	Uniform Manifold Approximation and Projection
t-SNE	t-distributed Stochastic Neighbor Embedding
ROI	Region of Interest / Περιοχή Ενδιαφέροντος
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
TP	True Positive / Αληθώς Θετικό
TN	True Negative / Αληθώς Αρνητικό
FP	False Positive / Ψευδώς Θετικό
FN	False Negative / Ψευδώς Αρνητικό
ISIC	International Skin Imaging Collaboration
HAM10000	Human Against Machine with 10000 training images
PH2	PH2 Dermoscopic Image Database
GPU	Graphics Processing Unit / Μονάδα Επεξεργασίας Γραφικών
Colab	Google Colaboratory

1. Εισαγωγή

1.1 Καρκίνος του δέρματος και διαγνωστική δυσκολία

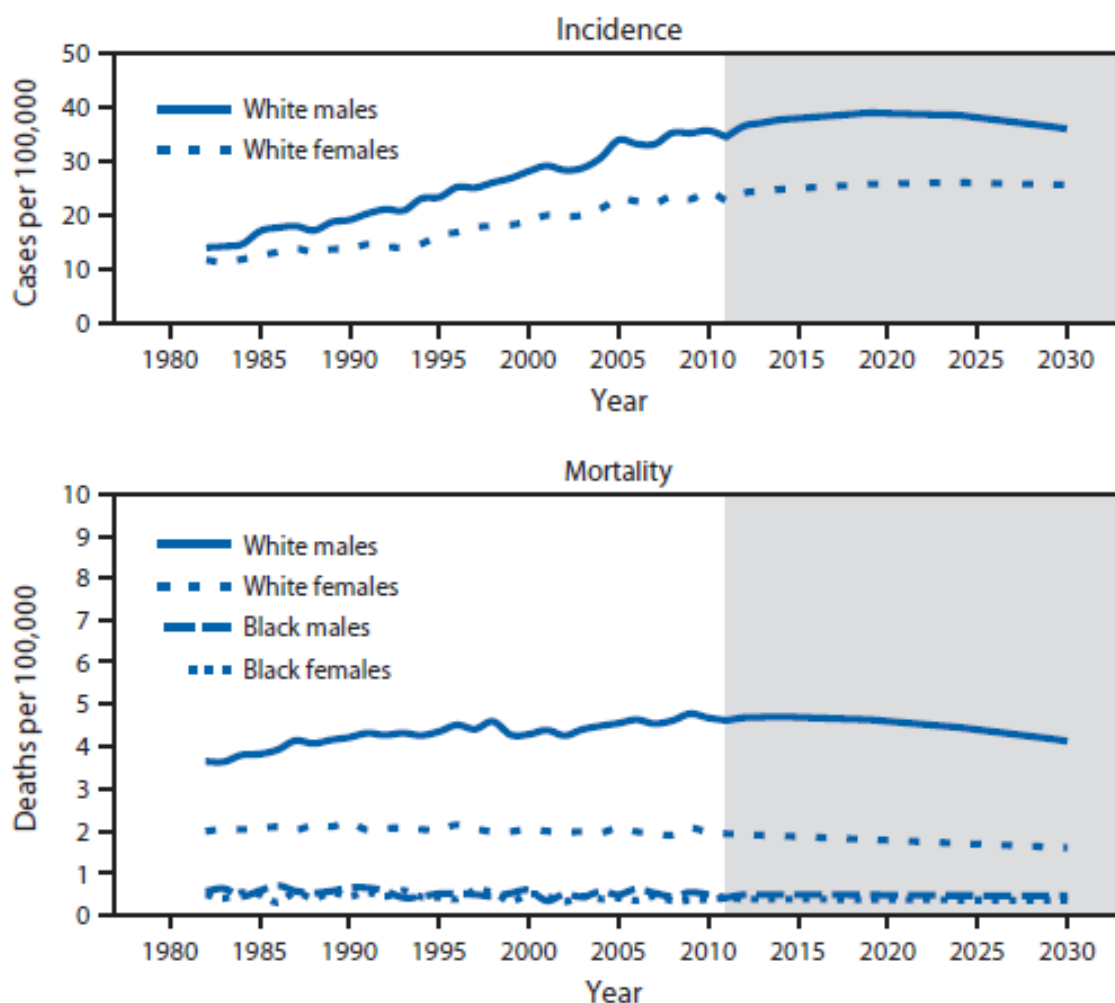
Ο καρκίνος του δέρματος αποτελεί μία από τις συχνότερες μορφές καρκίνου παγκοσμίως και συνδέεται κυρίως με την ανεξέλεγκτη ανάπτυξη μη φυσιολογικών κυττάρων στο δέρμα. Δεν πρόκειται για μία ενιαία πάθηση, αλλά για μια ευρύτερη ομάδα δερματικών κακοηθειών, με διαφορετική βιολογική συμπεριφορά, διαφορετικό βαθμό επιθετικότητας και διαφορετική κλινική βαρύτητα. Οι βασικές κατηγορίες περιλαμβάνουν τους μη μελανωματικούς καρκίνους του δέρματος, όπως το βασικοκυτταρικό καρκίνωμα και το ακανθοκυτταρικό καρκίνωμα, καθώς και το μελάνωμα, το οποίο θεωρείται η πιο επικίνδυνη μορφή λόγω της δυνατότητάς του να εξελίσσεται γρήγορα και να δίνει μεταστάσεις όταν δεν εντοπιστεί έγκαιρα. Η σημασία του προβλήματος φαίνεται ήδη από τα επιδημιολογικά δεδομένα. Σε παγκόσμιο επίπεδο, οι καρκίνοι του δέρματος αποτελούν από τις συχνότερες ομάδες καρκίνων, με περισσότερα από 1,5 εκατομμύριο νέα περιστατικά να εκτιμώνται για το 2022. Για το μελάνωμα ειδικότερα, το International Agency for Research on Cancer αναφέρει ότι το 2022 εκτιμήθηκαν περίπου 330.000 νέα περιστατικά παγκοσμίως και σχεδόν 60.000 θάνατοι από τη νόσο [1]. Τα στοιχεία αυτά δείχνουν ότι το μελάνωμα, αν και λιγότερο συχνό από άλλους τύπους καρκίνου του δέρματος, έχει ιδιαίτερη κλινική βαρύτητα λόγω της σύνδεσής του με τη θνησιμότητα.



Εικόνα 1.1. Παγκόσμιο βάρος του δερματικού μελανώματος το 2020 και προβλέψεις έως το 2040. Πηγή: International Agency for Research on Cancer / World Health Organization.

Όπως φαίνεται στην Εικόνα 1.1, το 2020 εκτιμήθηκαν περίπου 325.000 νέα περιστατικά μελανώματος και 57.000 θάνατοι παγκοσμίως, ενώ μέχρι το 2040 προβλέπεται αύξηση σε περίπου 510.000 νέα περιστατικά και 96.000 θανάτους [2]. Η εικόνα αυτή δείχνει ότι το

μελάνωμα δεν αποτελεί πρόβλημα μεμονωμένων χωρών ή περιοχών, αλλά ευρύτερο ζήτημα δημόσιας υγείας. Η προβλεπόμενη αύξηση των νέων περιστατικών και των θανάτων ενισχύει την ανάγκη για έγκαιρη αναγνώριση, συστηματικότερη αξιολόγηση των ύποπτων βλαβών και ανάπτυξη υποστηρικτικών εργαλείων διάγνωσης. Η αντίθεση ανάμεσα στη συχνότητα εμφάνισης και στη θνησιμότητα είναι ιδιαίτερα σημαντική. Οι μη μελανωματικοί καρκίνοι του δέρματος εμφανίζονται συχνότερα, αλλά συνήθως έχουν πιο αργή εξέλιξη και μικρότερη πιθανότητα θανατηφόρας πορείας. Το μελάνωμα, αντίθετα, μπορεί να αναπτυχθεί πιο επιθετικά και να εξαπλωθεί σε άλλα σημεία του σώματος. Για αυτό η βαρύτητά του δεν αποτιμάται μόνο με βάση το πόσο συχνά εμφανίζεται, αλλά κυρίως με βάση το τι μπορεί να συμβεί όταν δεν αναγνωριστεί εγκαίρως. Σε αυτή τη λογική, η έγκαιρη διάγνωση δεν είναι απλώς ένα επιθυμητό στάδιο της κλινικής διαδικασίας· είναι παράγοντας που μπορεί να καθορίσει την πορεία της νόσου. Η διαχρονική παρακολούθηση της επίπτωσης και της θνησιμότητας βοηθά να γίνει πιο κατανοητό το μέγεθος του προβλήματος. Στην Εικόνα 1.2 παρουσιάζονται στοιχεία του CDC για τις ΗΠΑ, όπου καταγράφονται οι παρατηρούμενες και προβλεπόμενες ηλικιακά προσαρμοσμένες τιμές επίπτωσης και θνησιμότητας για το μελάνωμα την περίοδο 1982–2030. Σύμφωνα με το CDC, από το 1982 έως το 2011 η επίπτωση του μελανώματος αυξήθηκε, ενώ η θνησιμότητα παρέμεινε σχετικά σταθερή.



Εικόνα 1.2. Παρατηρούμενες και προβλεπόμενες ηλικιακά προσαρμοσμένες τιμές επίπτωσης και θνησιμότητας για το μελάνωμα στις ΗΠΑ, 1982–2030. Πηγή: CDC/MMWR.

Όπως φαίνεται στην Εικόνα 1.2, τα νέα περιστατικά μελανώματος παρουσιάζουν αυξητική τάση σε βάθος χρόνου, ιδιαίτερα στους λευκούς άνδρες και στις λευκές γυναίκες, ενώ οι δείκτες θνησιμότητας παραμένουν χαμηλότεροι σε σχέση με τους δείκτες επίπτωσης [3]. Το στοιχείο αυτό δεν μειώνει τη σοβαρότητα της νόσου. Αντίθετα, δείχνει ότι η αύξηση των νέων περιστατικών δημιουργεί μεγαλύτερη ανάγκη για έγκαιρη ανίχνευση, σωστή αξιολόγηση και καλύτερη παρακολούθηση. Σε ένα τέτοιο πλαίσιο, η αξιοποίηση υπολογιστικών εργαλείων για την ανάλυση δερματοσκοπικών εικόνων αποκτά πρακτικό ενδιαφέρον, επειδή μπορεί να συμβάλει στην επισημάνση ύποπτων βλαβών και στην υποστήριξη της διαγνωστικής διαδικασίας. Η διαφορά ανάμεσα στην πρόωμη και στην καθυστερημένη διάγνωση αποτυπώνεται καθαρά στα ποσοστά επιβίωσης. Όταν το μελάνωμα εντοπίζεται σε τοπικό στάδιο, δηλαδή πριν επεκταθεί πέρα από το σημείο όπου ξεκίνησε, το πενταετές σχετικό ποσοστό επιβίωσης είναι πάνω από 99%. Όταν η νόσος έχει επεκταθεί σε γειτονικές δομές ή λεμφαδένες, το ποσοστό μειώνεται στο 76%, ενώ όταν υπάρχουν απομακρυσμένες μεταστάσεις μειώνεται στο 35% [4]. Η πτώση αυτή δείχνει με άμεσο τρόπο ότι η πρόωμη αναγνώριση μπορεί να αλλάξει ουσιαστικά την πρόγνωση του ασθενούς.

Με βάση αυτά τα δεδομένα, το μελάνωμα είναι μία μορφή καρκίνου όπου ο χρόνος της διάγνωσης έχει ιδιαίτερη σημασία. Αν εντοπιστεί σε αρχικό στάδιο, μπορεί συχνά να αντιμετωπιστεί αποτελεσματικά, κυρίως με αφαίρεση της βλάβης και κατάλληλη ιατρική παρακολούθηση. Αν όμως η διάγνωση καθυστερήσει, η νόσος μπορεί να εξελιχθεί σε στάδιο όπου η θεραπεία γίνεται πιο σύνθετη και η πρόγνωση επιβαρύνεται. Αυτό εξηγεί γιατί η έρευνα γύρω από την αυτόματη αναγνώριση δερματικών βλαβών δεν έχει μόνο τεχνικό ενδιαφέρον, αλλά και άμεση σχέση με την κλινική πράξη. Στην καθημερινή δερματολογική αξιολόγηση, ένα βασικό εργαλείο για την εξέταση ύποπτων δερματικών βλαβών είναι η δερματοσκόπηση. Πρόκειται για μη επεμβατική μέθοδο που επιτρέπει πιο λεπτομερή παρατήρηση της βλάβης σε σχέση με την απλή οπτική εξέταση. Μέσω του δερματοσκοπίου μπορούν να αναδειχθούν χαρακτηριστικά που δεν είναι εύκολα ορατά με γυμνό μάτι, όπως λεπτές χρωματικές μεταβολές, μορφολογικά μοτίβα, όρια και εσωτερική δομή της βλάβης [5]. Παρά τη χρησιμότητά της, η ερμηνεία της δερματοσκοπικής εικόνας εξακολουθεί να βασίζεται σε μεγάλο βαθμό στην εμπειρία του/της δερματολόγου. Ακόμη και όταν υπάρχει εξοπλισμός που διευκολύνει την εξέταση, η τελική διάγνωση παραμένει συνδεδεμένη με την οπτική κρίση και την κλινική εμπειρία.

Ένα από τα πιο γνωστά κλινικά εργαλεία αρχικής αξιολόγησης είναι ο κανόνας ABCDE. Το Α αναφέρεται στην ασυμμετρία της βλάβης, το Β στα ακανόνιστα ή ασαφή όρια, το C στην ανομοιογένεια του χρώματος, το D στη διάμετρο και το Ε στην εξέλιξη ή μεταβολή της βλάβης με την πάροδο του χρόνου [6]. Ο κανόνας αυτός λειτουργεί ως πρακτικός οδηγός παρατήρησης και όχι ως αυτόνομο εργαλείο οριστικής διάγνωσης. Η αξία του βρίσκεται στο ότι οργανώνει την οπτική αξιολόγηση γύρω από χαρακτηριστικά που μπορούν να συνδεθούν με αυξημένη υποψία κακοήθειας. Για την παρούσα εργασία, τα κριτήρια αυτά έχουν διπλή σημασία: αφενός ανήκουν στο βασικό κλινικό υπόβαθρο της δερματολογικής αξιολόγησης, αφετέρου μπορούν να λειτουργήσουν ως σημείο αναφοράς

όταν εξετάζεται αν ένα μοντέλο τεχνητής νοημοσύνης εστιάζει σε περιοχές με πραγματική διαγνωστική αξία.

Η διαγνωστική δυσκολία αυξάνεται επειδή πολλές δερματικές βλάβες εμφανίζουν παρόμοια μορφολογικά χαρακτηριστικά. Ένας καλοήθης σπύλος μπορεί να έχει χρωματική ανομοιογένεια ή ακανόνιστη μορφή χωρίς να είναι μελάνωμα, ενώ ένα πρώιμο μελάνωμα μπορεί να μοιάζει με φαινομενικά αθώα βλάβη. Αυτή η οπτική αλληλοεπικάλυψη κάνει την αναγνώριση δύσκολη ακόμη και για έμπειρους/ες δερματολόγους. Το πρόβλημα γίνεται πιο έντονο στα αρχικά στάδια, όπου οι διαφορές ανάμεσα σε μελάνωμα και καλοήθεις βλάβες μπορεί να είναι λεπτές και όχι άμεσα διακριτές. Σε πραγματικές συνθήκες, η δερματοσκοπική εικόνα που εξετάζεται δεν είναι πάντοτε καθαρή ή ιδανική. Μπορεί να περιλαμβάνει τρίχες, αντανakλάσεις, σκοτεινές γωνίες, χάρακες, σημάδια από μαρκαδόρο, ανομοιογενή φωτισμό ή άλλες οπτικές παρεμβολές. Αυτά τα στοιχεία μπορούν να δυσκολέψουν τόσο την ανθρώπινη αξιολόγηση όσο και την αυτοματοποιημένη ανάλυση από υπολογιστικά μοντέλα. Για ένα μοντέλο βαθιάς μάθησης, υπάρχει πάντα ο κίνδυνος να μάθει μοτίβα που δεν σχετίζονται ουσιαστικά με τη βλάβη, αλλά με το φόντο, την ποιότητα λήψης ή κάποιο τεχνούργημα της εικόνας. Αυτό είναι ιδιαίτερα προβληματικό σε ιατρικές εφαρμογές, γιατί ένα φαινομενικά σωστό αποτέλεσμα δεν σημαίνει απαραίτητα ότι το μοντέλο έλαβε την απόφασή του με αξιόπιστο τρόπο.

Στα δημόσια σύνολα δερματοσκοπικών εικόνων, η ταξινόμηση οργανώνεται συχνά σε συγκεκριμένες κατηγορίες βλαβών. Συνήθεις κλάσεις είναι οι καλοήθεις σπύλοι, το μελάνωμα, το βασικοκυτταρικό καρκίνωμα, η ακτινική κεράτωση, οι καλοήθεις βλάβες τύπου keratosis-like, το δερματοΐνωμα και οι αγγειακές βλάβες. Η ύπαρξη πολλών κατηγοριών κάνει το πρόβλημα πιο σύνθετο από μια απλή διάκριση «μελάνωμα» ή «μη μελάνωμα». Το μοντέλο καλείται να ξεχωρίσει μορφολογικά κοντινές βλάβες, ενώ ταυτόχρονα πρέπει να διαχειριστεί την ανισορροπία των δεδομένων. Σε αρκετά datasets, οι συχνές κατηγορίες εμφανίζονται με πολύ περισσότερα δείγματα, ενώ οι σπανιότερες και συχνά κλινικά σημαντικές κατηγορίες εμφανίζονται σε μικρότερο αριθμό. Η ανισορροπία αυτή επηρεάζει άμεσα την εκπαίδευση και την αξιολόγηση των μοντέλων. Ένα σύστημα μπορεί να παρουσιάζει υψηλή συνολική ακρίβεια επειδή αναγνωρίζει σωστά τη συχνότερη κλάση, αλλά να αποτυγχάνει στις σπανιότερες περιπτώσεις [7]. Αυτό έχει ιδιαίτερη σημασία στην αναγνώριση μελανώματος, όπου η λανθασμένη ταξινόμηση μιας κακοήθους βλάβης ως καλοήθους μπορεί να έχει σοβαρές συνέπειες. Για τον λόγο αυτό, η αξιολόγηση δεν μπορεί να βασίζεται μόνο στο accuracy. Χρειάζονται μετρικές που αποτυπώνουν καλύτερα τη συμπεριφορά του μοντέλου ανά κατηγορία, όπως precision, recall, F1-score, macro-F1 και AUC. Σε πειράματα τμηματοποίησης, όπου το ζητούμενο είναι ο εντοπισμός της περιοχής της βλάβης, χρησιμοποιούνται συχνά μετρικές όπως Dice και IoU, επειδή δείχνουν τον βαθμό επικάλυψης ανάμεσα στην προβλεπόμενη και την πραγματική περιοχή.

1.2 Τεχνητή νοημοσύνη και εξηγησιμότητα στην ανάλυση δερματοσκοπικών εικόνων

Η ανάλυση δερματοσκοπικών εικόνων με τεχνικές τεχνητής νοημοσύνης έχει αποκτήσει ιδιαίτερο ενδιαφέρον τα τελευταία χρόνια, επειδή μπορεί να υποστηρίξει την έγκαιρη αναγνώριση ύποπτων δερματικών βλαβών. Η δερματοσκοπική εικόνα περιέχει πληροφορίες που σχετίζονται με το χρώμα, το σχήμα, την υφή, τα όρια και την εσωτερική

δομή της βλάβης [5]. Τα στοιχεία αυτά μπορούν να αξιοποιηθούν από υπολογιστικά μοντέλα, ώστε να πραγματοποιηθεί αυτόματη ή ημι-αυτόματη ταξινόμηση σε διαγνωστικές κατηγορίες. Η βαθιά μάθηση έχει συμβάλει σημαντικά σε αυτή την κατεύθυνση. Τα συνελκτικά νευρωνικά δίκτυα μπορούν να εξάγουν τοπικά χαρακτηριστικά από την εικόνα, όπως ακμές, χρωματικές μεταβολές και μικρές μορφολογικές λεπτομέρειες. Αυτό είναι χρήσιμο στη δερματοσκόπηση, όπου ακόμη και μικρές διαφοροποιήσεις μπορεί να σχετίζονται με την καλοήγη ή κακοήγη φύση μιας βλάβης. Παράλληλα, οι μετασχηματιστές όρασης έχουν αρχίσει να χρησιμοποιούνται όλο και περισσότερο, καθώς μπορούν να εξετάζουν σχέσεις ανάμεσα σε απομακρυσμένες περιοχές της εικόνας και να αποτυπώνουν καλύτερα τη συνολική οργάνωση της βλάβης [9].

Σε αρκετές περιπτώσεις, η ανάλυση δεν περιορίζεται μόνο στην ταξινόμηση, αλλά περιλαμβάνει και την τμηματοποίηση της βλάβης. Η τμηματοποίηση αφορά τον εντοπισμό της περιοχής ενδιαφέροντος και των ορίων της βλάβης. Αυτή η διαδικασία μπορεί να βοηθήσει το μοντέλο να εστιάσει περισσότερο στο ουσιαστικό μέρος της εικόνας και λιγότερο στο φόντο ή σε στοιχεία χωρίς διαγνωστική αξία. Με αυτή τη λογική, η τμηματοποίηση και η ταξινόμηση μπορούν να λειτουργήσουν συμπληρωματικά μέσα σε μια ενιαία ροή ανάλυσης [10]. Παρά την πρόοδο των μοντέλων βαθιάς μάθησης, η υψηλή αριθμητική επίδοση δεν είναι από μόνη της αρκετή. Σε ένα ιατρικό πρόβλημα, δεν αρκεί ένα μοντέλο να δίνει σωστή πρόβλεψη· χρειάζεται να μπορεί να ελεγχθεί και το πού βασίζεται αυτή η πρόβλεψη. Ένα μοντέλο μπορεί να πετυχαίνει υψηλό accuracy ή AUC, αλλά να στηρίζει την απόφασή του σε άσχετα στοιχεία της εικόνας, όπως τρίχες, αντανakλάσεις, σκοτεινές γωνίες, χάρακες ή το υπόβαθρο. Σε μια τέτοια περίπτωση, η πρόβλεψη μπορεί να φαίνεται επιτυχημένη αριθμητικά, αλλά να μην είναι πραγματικά αξιόπιστη.

Για αυτό η εξηγησιμότητα έχει κεντρικό ρόλο στην ανάλυση δερματοσκοπικών εικόνων. Τεχνικές όπως Grad-CAM, Grad-CAM++, Score-CAM και SHAP [12]-[15] μπορούν να βοηθήσουν στην οπτική ή ποσοτική ερμηνεία της απόφασης ενός μοντέλου. Οι χάρτες εστίασης δείχνουν ποιες περιοχές της εικόνας επηρέασαν περισσότερο την πρόβλεψη. Όταν η εστίαση βρίσκεται μέσα στη βλάβη και ειδικά σε περιοχές με ασυμμετρία, ακανόνιστα όρια ή χρωματική ανομοιογένεια, η εξήγηση είναι πιο κοντά στη λογική της δερματολογικής παρατήρησης. Αντίθετα, όταν η εστίαση βρίσκεται έξω από τη βλάβη, δημιουργείται ζήτημα εμπιστοσύνης. Η εξηγησιμότητα δεν πρέπει να αντιμετωπίζεται ως απλή οπτική προσθήκη στο τέλος της διαδικασίας. Ένα heatmap μπορεί να φαίνεται καθαρό ή πειστικό, χωρίς απαραίτητα να αποτυπώνει πιστά τον μηχανισμό απόφασης του μοντέλου. Για τον λόγο αυτό, η ερμηνεία των αποτελεσμάτων χρειάζεται προσοχή και σύνδεση με το κλινικό περιεχόμενο της εικόνας. Στο πλαίσιο της παρούσας εργασίας, η τεχνητή νοημοσύνη εξετάζεται όχι μόνο ως εργαλείο ταξινόμησης, αλλά και ως σύστημα του οποίου οι αποφάσεις πρέπει να μπορούν να ελεγχθούν και να ερμηνευθούν.

Η βασική κατεύθυνση, επομένως, δεν είναι η αντικατάσταση της ιατρικής κρίσης, αλλά η ανάπτυξη υποστηρικτικών εργαλείων που μπορούν να βοηθήσουν στην αξιολόγηση δερματοσκοπικών εικόνων. Ένα τέτοιο σύστημα έχει μεγαλύτερη αξία όταν συνδυάζει ικανοποιητική επίδοση, σταθερότητα και εξηγήσεις που σχετίζονται με την ίδια τη βλάβη. Με αυτή τη λογική, η παρούσα εργασία συνδέει την ταξινόμηση δερματικών βλαβών με

την ανάγκη ερμηνευσιμότητας, ώστε το αποτέλεσμα να μην αξιολογείται μόνο ως σωστό ή λανθασμένο, αλλά και ως προς το αν βασίζεται σε ουσιαστικά χαρακτηριστικά της εικόνας.

1.3 Σκοπός της παρούσας εργασίας

Ο σκοπός της παρούσας διπλωματικής, με αφετηρία τα ανοιχτά ζητήματα στην έγκαιρη και ακριβή διάγνωση καρκίνου του δέρματος, είναι να διαμορφωθεί μια ρεαλιστική προσέγγιση που συνδυάζει υψηλή επίδοση με εξηγησιμότητα, ώστε να εξεταστεί όχι μόνο η τελική πρόβλεψη του μοντέλου, αλλά και ο βαθμός στον οποίο η απόφαση βασίζεται σε περιοχές της εικόνας με πραγματική διαγνωστική σημασία. Στόχος είναι, πρώτον, να χαρτογραφηθεί το state-of-the-art (SOTA) μέσα από μια εκτενή βιβλιογραφική ανασκόπηση τεχνικών της τελευταίας πενταετίας. Το ενδιαφέρον δεν περιορίζεται μόνο στις αριθμητικές επιδόσεις των μοντέλων, αλλά επεκτείνεται και στα σενάρια όπου αυτά δυσκολεύονται, όπως η ανισορροπία κλάσεων, οι μικρές ή δυσδιάκριτες βλάβες και ο θόρυβος στις εικόνες. Δεύτερον, στόχος είναι η ανάπτυξη, εφαρμογή και έλεγχος επιλεγμένων μεθόδων βαθιάς μάθησης σε δερματοσκοπικά datasets, μέσα από καθορισμένα πειραματικά πρωτόκολλα. Στο πλαίσιο αυτό, τα μοντέλα δεν αντιμετωπίζονται απλώς ως εργαλεία παραγωγής μιας τελικής πρόβλεψης, αλλά εξετάζονται ως προς τον τρόπο με τον οποίο λειτουργούν πάνω στα δεδομένα, τις συνθήκες κάτω από τις οποίες βελτιώνεται ή περιορίζεται η απόδοσή τους, καθώς και τη συμπεριφορά τους απέναντι σε διαφορετικά προβλήματα ταξινόμησης και τμηματοποίησης. Για τον λόγο αυτό εφαρμόζονται τεχνικές εξηγησιμότητας, όπως Grad-CAM και παραλλαγές του, ώστε να διερευνηθεί αν οι οπτικές ενδείξεις που αξιοποιεί το μοντέλο αντιστοιχούν σε περιοχές με πραγματική διαγνωστική αξία. Τρίτον, στόχος είναι η αποτίμηση των μοντέλων που αναπτύχθηκαν και εφαρμόστηκαν, όχι μόνο με βάση τις κλασικές μετρικές επίδοσης, αλλά και μέσα από την ερμηνεία της συμπεριφοράς τους. Η αξιολόγηση περιλαμβάνει μετρικές όπως macro-F1, per-class AUC, Dice και IoU, ώστε να αποτυπώνεται πιο καθαρά η απόδοση σε σπάνιες ή δύσκολες κλάσεις. Παράλληλα, οι τεχνικές εξηγησιμότητας αξιοποιούνται για να φανεί αν τα μοντέλα εστιάζουν στη βλάβη ή σε άσχετες περιοχές της εικόνας, αν διατηρούν σταθερή συμπεριφορά σε διαφορετικά σενάρια και ποια αρχιτεκτονικά ή μεθοδολογικά σχήματα φαίνονται πιο κατάλληλα ανάλογα με το πρόβλημα. Με αυτόν τον τρόπο, η εργασία δεν περιορίζεται στη σύγκριση αριθμητικών αποτελεσμάτων, αλλά επιχειρεί να συνδέσει την απόδοση, τη λειτουργία και την ερμηνευσιμότητα των μοντέλων σε ένα ενιαίο πλαίσιο αξιολόγησης.

1.4 Δομή της εργασίας

Η υπόλοιπη εργασία οργανώνεται ως εξής: Το δεύτερο κεφάλαιο περιλαμβάνει τη βιβλιογραφική ανασκόπηση. Εκεί παρουσιάζονται σύγχρονες προσεγγίσεις που αξιοποιούν συνελκτικά νευρωνικά δίκτυα, μετασχηματιστές όρασης, υβριδικές αρχιτεκτονικές, τεχνικές τμηματοποίησης και μεθόδους εξηγησιμότητας για την ανάλυση δερματικών βλαβών. Το τρίτο κεφάλαιο περιγράφει τη μεθοδολογία και το πειραματικό πλαίσιο της εργασίας. Παρουσιάζονται τα datasets, οι διαδικασίες προεπεξεργασίας, τα μοντέλα που επιλέχθηκαν, οι τεχνικές εκπαίδευσης, οι μετρικές αξιολόγησης και τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση των πειραμάτων. Το τέταρτο κεφάλαιο παρουσιάζει την πειραματική αξιολόγηση. Περιλαμβάνει τα αποτελέσματα των μοντέλων, τις μετρικές απόδοσης, τα διαγράμματα εκπαίδευσης, τις συγκρίσεις ανάμεσα σε διαφορετικές

προσεγγίσεις και τις παρατηρήσεις που προκύπτουν από τα πειράματα. Το πέμπτο κεφάλαιο εστιάζει στην εξηγησιμότητα των μοντέλων και στη σύνδεσή της με τα αποτελέσματα της πειραματικής αξιολόγησης. Εκεί εξετάζονται οπτικές εξηγήσεις, όπως heatmaps, και αξιολογείται αν τα μοντέλα εστιάζουν σε περιοχές που σχετίζονται με τη βλάβη ή σε άσχετα στοιχεία της εικόνας. Το έκτο κεφάλαιο παρουσιάζει την τεχνική τεκμηρίωση και την αναπαραγωγικότητα της εργασίας. Περιγράφεται η οργάνωση των κωδίκων, των notebooks, των datasets και του GitHub αποθετηρίου, ώστε να υπάρχει καθαρή εικόνα της πειραματικής διαδικασίας και της δυνατότητας επανάληψής της.

2. Βιβλιογραφική ανασκόπηση

Στο κεφάλαιο αυτό παρουσιάζουμε τη βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε με στόχο να χαρτογραφήσουμε πρόσφατες εργασίες (τελευταία 2–4 χρόνια) γύρω από Transformers και εξηγησιμότητα (XAI) σε δερματοσκοπικές εικόνες. Αρχικά συνοψίζουμε τις βασικές προσεγγίσεις σε τμηματοποίηση και ταξινόμηση, έπειτα καταγράφουμε τα σύνολα δεδομένων που εμφανίζονται συστηματικά στη βιβλιογραφία και τα αξιοποιήσαμε/κατεβάσαμε για τα πειράματά μας. Στόχος είναι να φανεί τι κάνει κάθε εργασία, πώς λειτουργεί ο μηχανισμός, τι αποτελέσματα αναφέρει, και τι κρατάμε ως πρακτικές κατευθύνσεις, με ό,τι αυτό συνεπάγεται για γενίκευση και δίκαιες συγκρίσεις.

2.1 Σχετικές εργασίες

Το άρθρο [16] προτείνει μια δι-σταδιακή προσέγγιση με CNN για ταξινόμηση δερματικών βλαβών, όπου η εντόπιση της βλάβης γίνεται με αδύναμη επίβλεψη. Στο πρώτο στάδιο εκπαιδεύουν ένα CNN και, αντί να έχουν έτοιμες μάσκες τμηματοποίησης, αξιοποιούν χάρτες προσοχής Grad-CAM για να κρατήσουν την περιοχή που φαίνεται πιο σχετική με τη βλάβη, σαν μια πρακτική, πρόχειρη τμηματοποίηση. Στο δεύτερο στάδιο, ένα δεύτερο CNN παίρνει αυτή την εντοπισμένη περιοχή και κάνει την τελική ταξινόμηση. Η βασική αξιολόγηση γίνεται στο HAM10000, όπου το καλύτερο ζεύγος μοντέλων DenseNet121→DenseNet161 πετυχαίνει AUC 0.910, ευαισθησία 0.941, ειδικότητα 0.963 και ακρίβεια 0.962, ενώ το αντίστοιχο single-stage στήσιμο φτάνει Acc 0.922 και το two-stage ανεβαίνει σε 0.962. Αναφέρουν ακόμη δοκιμές στα ISIC 2016/2017/2018 και σε δεδομένα vitiligo, με πολύ υψηλές επιδόσεις (π.χ. AUC 0.986 και Acc 0.991 στο vitiligo). Το μήνυμα των συγγραφέων είναι ότι το Grad-CAM λειτουργεί σαν “οδηγός” ώστε το δίκτυο να κοιτάει εκεί που πρέπει, κάτι που ανεβάζει τόσο την απόδοση όσο και το πόσο κατανοητό γίνεται το γιατί πήρε μια απόφαση.

Το άρθρο [17] είναι μια πρακτική ανασκόπηση που, πέρα από σύνοψη της σχετικής βιβλιογραφίας, προτείνει και ένα συγκεκριμένο framework (LesionAid) για δημιουργία συνθετικών εικόνων και ταξινόμηση δερματικών βλαβών με βάση Vision Transformers. Η ιδέα τους ξεκινά από ένα πολύ πρακτικό πρόβλημα ανισορροπίας κλάσεων στο HAM10000, όπου υπάρχουν πολλές εικόνες της κλάσης NV (περίπου 6705) και πολύ λίγες της MEL (περίπου 142), κάτι που συνήθως σπρώχνει το μοντέλο να ευνοεί τις συχνές κατηγορίες. Για να το αντιμετωπίσουν, οργανώνουν το LesionAid σε τέσσερις φάσεις: πρώτα εκπαιδεύουν ένα ViTGAN ώστε να παράγει συνθετικές εικόνες και να εξισορροπεί τις κλάσεις, μετά εφαρμόζουν κλασικές τεχνικές augmentation (flip, rotate, zoom, αλλαγές φωτεινότητας και γενικότερα μορφολογικές/γεωμετρικές παραλλαγές), στη συνέχεια εκπαιδεύουν έναν ViT classifier πάνω στο σύνολο των αρχικών και των συνθετικών δεδομένων, και τέλος προσθέτουν στάδιο ερμηνευσιμότητας με Grad-CAM και attention heatmaps για να φαίνεται πού κοιτά το μοντέλο όταν αποφασίζει. Η αξιολόγηση γίνεται στο HAM10000 με 7 κλάσεις. Για την ποιότητα της γεννήτριας αναφέρουν FID = 13.32 (στα 625 epochs), ενώ για την ταξινόμηση δίνουν πολύ υψηλές επιδόσεις, με training accuracy 99.2% και validation accuracy 97.4%. Το βασικό συμπέρασμα που προβάλλουν είναι ότι το synthetic balancing μπορεί να βελτιώσει την απόδοση και να στηρίξει πιο καθαρές εξηγήσεις (μέσω των heatmaps), αλλά το συνολικό στήσιμο είναι υπολογιστικά βαρύ, άρα δεν “κάθεται” εύκολα σε edge συσκευές ή σε περιβάλλοντα με περιορισμένους πόρους.

Το άρθρο [18] παρουσιάζει ένα pipeline για ανίχνευση καρκίνου του δέρματος που στοχεύει όχι μόνο σε υψηλή απόδοση, αλλά και σε ερμηνευσιμότητα των αποφάσεων. Η λογική τους είναι υβριδική: ξεκινούν από deep learning για να πάρουν καλά χαρακτηριστικά από εικόνες και μετά περνούν σε επιλογή χαρακτηριστικών και κλασικούς ταξινομητές, ώστε να μειώσουν υπολογιστικό κόστος και να ελέγξουν καλύτερα τη συμπεριφορά του συστήματος. Στο κομμάτι της μεθοδολογίας χρησιμοποιούν Xception με transfer learning, κρατώντας παγωμένα τα προεκπαιδευμένα layers και προσθέτοντας στο τέλος Global Average Pooling, ένα dense layer 1024 νευρώνων, dropout 0.5 και τελικό head 2 νευρώνων για δυαδική απόφαση. Από αυτό το δίκτυο εξάγουν ένα διάνυσμα 1024 χαρακτηριστικών ανά εικόνα, και στη συνέχεια εφαρμόζουν Particle Swarm Optimization (PSO) για να μειώσουν το πλήθος τους περίπου στα 508, με κριτήριο (fitness) την απόδοση ενός KNN ($k=5$). Πάνω στο μειωμένο feature set δοκιμάζουν ταξινόμηση με SVM, KNN και ensemble επιλογές. Για να κάνουν το αποτέλεσμα διαβασμένο από άνθρωπο, εφαρμόζουν τεχνικές XAI όπως Grad-CAM, LIME και Occlusion Sensitivity, ώστε να φαίνεται ποιες περιοχές της εικόνας επηρεάζουν την απόφαση. Στην αξιολόγηση αναφέρουν πολύ υψηλή επίδοση στο ISIC 2018, με 98.5% accuracy, ενώ ως holdout δοκιμή στο HAM10000 καταγράφουν 86.1%, με το καλύτερο αποτέλεσμα να έρχεται όταν συνδυάζουν PSO με Subspace KNN. Το βασικό τους συμπέρασμα είναι ότι η επιλογή χαρακτηριστικών μετά το CNN μπορεί να δώσει καλύτερο trade-off: διατηρεί ή βελτιώνει την ακρίβεια με λιγότερα features (άρα μικρότερο κόστος) και, μαζί με τα XAI εργαλεία, οδηγεί σε αποφάσεις που μπορούν να εξηγηθούν πιο καθαρά.

Το άρθρο [19] προτείνει ένα πλαίσιο για εξηγήσιμη διάγνωση δερματικών βλαβών, όπου το μοντέλο δεν βασίζεται μόνο σε μοτίβα εικόνας, αλλά συνδέει την εικόνα με κλινικές έννοιες/συμπτώματα. Το MICA ευθυγραμμίζει εικόνα και έννοιες σε τρία επίπεδα: (1) image-level με global contrastive εκπαίδευση ώστε οι συνολικές αναπαραστάσεις εικόνας-έννοιών να έρχονται κοντά, (2) token-level με cross-attention για αντιστοίχιση εννοιών με υπο-περιοχές/patches της εικόνας, και (3) concept-level με Concept Activation Vectors (CAVs μέσω SVM), ώστε οι αναπαραστάσεις να προβάλλονται σε υποχώρο εννοιών. Η διάγνωση γίνεται είτε με concept bottleneck (πρώτα πρόβλεψη εννοιών, μετά διάγνωση) είτε απευθείας χωρίς bottleneck. Αξιολογούν σε Derm7pt, PH2 και SkinCon, με AUC έως 85.59 (Derm7pt, χωρίς bottleneck) και 98.18 (PH2, χωρίς bottleneck), ενώ στο concept detection φτάνουν AUC 82.6 (SkinCon). Το βασικό μήνυμα είναι ότι η πολυ-επίπεδη αντιστοίχιση αυξάνει απόδοση και κάνει τις αποφάσεις πιο ερμηνεύσιμες.

Το άρθρο [20] προτείνει ένα υβριδικό σύστημα για τμηματοποίηση και ταξινόμηση δερματικών βλαβών, βασισμένο σε τρεις παράλληλες ροές χαρακτηριστικών. Η πρώτη ροή εφαρμόζει Angular Radial Partitioning (ARP) πάνω σε edge εικόνες, ώστε να αποτυπώνει δομικά/γεωμετρικά μοτίβα με αντοχή σε περιστροφή και κλίμακα. Η δεύτερη ροή είναι CNN για τοπικές λεπτομέρειες υφής, ενώ η τρίτη είναι Vision Transformer για παγκόσμιο πλαίσιο μέσω self-attention. Τα χαρακτηριστικά συνενώνονται (concatenation) και οδηγούν σε ταξινόμηση 7 κλάσεων (DF, VASC, AKIEC, BKL, BCC, NV, MEL). Παράλληλα, ενσωματώνουν module τμηματοποίησης ορίων (με ViT skip-connections). Στο preprocessing κάνουν balancing/resampling, augmentation (flip /rotation /elastic/ gamma), bilateral filtering, Canny edges και resize 256×256. Στο HAM10000 αναφέρουν accuracy 98.2%, precision 0.94, recall 0.96, macro-F1 0.95, με τμηματοποίηση που κρατά καθαρά

όρια σε δύσκολες περιπτώσεις. Το κεντρικό συμπέρασμα είναι ότι ο συνδυασμός ARP , CNN , ViT πιάνει σχήμα, λεπτομέρεια και πλαίσιο καλύτερα από CNN-only ή ViT-only.

Το άρθρο [21] παρουσιάζει το OXAI-SCC-Net, ένα end-to-end πλαίσιο για διάγνωση καρκίνου δέρματος που συνδυάζει υψηλή επίδοση με εξηγησιμότητα. Η πρότασή τους οργανώνεται σε τέσσερα βήματα. Πρώτα εισάγουν το NASL-Net, όπου η εξαγωγή χαρακτηριστικών γίνεται με Neural Architecture Search και παράλληλα χρησιμοποιείται Grad-CAM για να αποτυπώνεται οπτικά ποιες περιοχές της εικόνας επηρεάζουν την απόφαση. Έπειτα εφαρμόζουν SSOT oversampling, μια παραλλαγή τύπου SMOTE που καθοδηγείται από όρια SVM ώστε να αντιμετωπίζεται πιο στοχευμένα η ανισορροπία κλάσεων. Στο τρίτο στάδιο κάνουν επιλογή χαρακτηριστικών με HOA-XAI, δηλαδή Hippopotamus Optimization με κριτήριο που ενσωματώνει και την εξηγησιμότητα, ώστε να κρατούνται features που βγάζουν νόημα και βοηθούν την ταξινόμηση. Τέλος, ένα υβριδικό VT-CNN (Vision Transformer μαζί με CNN) δίνει την τελική πρόβλεψη. Στο ISIC-2019 αναφέρουν Accuracy 99.12% και Precision/Recall/F1 99.13%, με χρόνο inference περίπου 108 ms. Το βασικό συμπέρασμα είναι ότι ο συνδυασμός στοχευμένου balancing, επιλογής χαρακτηριστικών και υβριδικής Transformer/CNN ταξινόμησης οδηγεί σε πολύ υψηλή απόδοση, χωρίς να χάνεται η δυνατότητα ερμηνείας.

Το άρθρο [22] προτείνει ένα πλαίσιο για ανίχνευση ευλογιάς των πιθήκων από δερματικές εικόνες, με δύο κόκκινες γραμμές: ιδιωτικότητα (να μη φεύγουν τα δεδομένα από τα κέντρα) και ερμηνευσιμότητα (να φαίνεται γιατί βγήκε η πρόβλεψη). Πρώτα χρησιμοποιούν StyleGAN2-ADA για παραγωγή συνθετικών εικόνων, αυξάνοντας το dataset πάνω από 700%, με δείκτες ποιότητας περίπου FID ≈ 12.4 και IS ≈ 4.8 . Έπειτα χτίζουν υβριδικά μοντέλα CNN + ViT, όπου τα χαρακτηριστικά ενώνονται με concatenation (π.χ. DenseNet121-ViT-B32) ώστε να συνδυάζεται τοπική λεπτομέρεια και παγκόσμιο πλαίσιο. Η εκπαίδευση γίνεται σε federated learning με Flower και FedAvg, σε 4 clients και 50 rounds, ώστε κάθε κόμβος να κρατά τα δεδομένα του και να στέλνει μόνο ενημερώσεις μοντέλου. Για το γιατί εφαρμόζουν Grad-CAM και LIME. Η αξιολόγηση γίνεται σε dataset 4 κλάσεων (monkeypox/chickenpox/measles/normal) με εύρος περίπου 1249–10,063 εικόνες, και το καλύτερο στήσιμο (DenseNet121-ViT-B32 με FL και GAN) δίνει accuracy 98.2%, precision 98.2%, recall 95.25% και F1 96%. Το βασικό συμπέρασμα είναι ότι GAN augmentation και federated learning βελτιώνουν τη γενίκευση χωρίς διαμοιρασμό δεδομένων, ενώ το XAI βοηθά στην κλινική εμπιστοσύνη.

Το άρθρο [23] προτείνει ένα pipeline που πρώτα εντοπίζει/τμηματοποιεί τη βλάβη και μετά κάνει ανίχνευση καρκίνου δέρματος, με έμφαση στη βελτιστοποίηση των μοντέλων. Η ροή τους (“JAEOLeNet”) ξεκινά με bilateral filtering για αποθορυβοποίηση και διατήρηση ακμών. Στη συνέχεια εφαρμόζουν τμηματοποίηση με TransUNet, αλλά δεν μένουν σε “στάνταρ” εκπαίδευση: ρυθμίζουν/βελτιστοποιούν τα βάρη μέσω Artificial Ecosystem-based Optimization (AEO), με στόχο πιο σταθερά masks. Πριν την ταξινόμηση κάνουν augmentation (μετατοπίσεις, περιστροφές, αλλαγές φωτεινότητας, blur) ώστε να αυξήσουν την ποικιλία των δειγμάτων. Για την αναγνώριση εξάγουν ένα μίγμα χαρακτηριστικών: κλασικά στατιστικά, Haralick (υφή), SLBT και επιπλέον CNN features, ώστε να συνδυάζουν χειροποίητα και βαθιά χαρακτηριστικά. Η τελική ταξινόμηση γίνεται με LeNet, της οποίας τα βάρη βελτιστοποιούνται με υβριδικό σχήμα Jaya + AEO (JAEOLeNet), δηλαδή μετα-ευρετική προσαρμογή πέρα από το κλασικό gradient-based training. Η αξιολόγηση

γίνεται στο SIIM-ISIC Melanoma με 33,126 εικόνες, όπου αναφέρουν accuracy 91.99%, sensitivity 90.95% και specificity 92.13%. Το βασικό συμπέρασμα είναι ότι το segmentation πριν την ταξινόμηση, μαζί με στοχευμένη βελτιστοποίηση βαρών, δίνει πιο σταθερή και καλύτερη συνολικά απόδοση.

Το άρθρο [24] προτείνει μια δι-φασική λύση για τμηματοποίηση και ταξινόμηση δερματικών βλαβών σε δερματοσκοπικές εικόνες. Στην πρώτη φάση εφαρμόζουν DeepLabv3+ (encoder-decoder), το οποίο το χτίζουν με 9 CNN blocks ώστε να παράγεται μάσκα της βλάβης και να ξεχωρίζει καθαρά από το υπόβαθρο. Η εκπαίδευση περιγράφεται με batch 32, 100 epochs και βελτιστοποιητή Adam. Στη δεύτερη φάση περνούν στην ταξινόμηση με Vision Transformer, όπου η εικόνα τεμαχίζεται σε patches 7×7 , επεξεργάζεται από 8 transformer blocks με multi-head attention και στο τέλος ένα MLP head δίνει την τελική κλάση. Η αξιολόγηση της τμηματοποίησης γίνεται στα ISIC 2016/2017/2018 και στο PH2, με global accuracy 98.90%, 98.38%, 99.45% και 100% αντίστοιχα. Για την ταξινόμηση αναφέρουν αποτελέσματα σε PH2, ISIC 2019, ISIC 2020 και HAM10000, με accuracy 100%, 97.73%, 96.97% και 100%. Το κεντρικό επιχείρημα είναι ότι το segmentation καθαρίζει τα όρια και μειώνει τον θόρυβο, ενώ το ViT συλλαμβάνει καλύτερα το συνολικό μοτίβο, άρα η απόδοση ανεβαίνει σε πολλαπλά datasets.

Το άρθρο [25] εξετάζει πώς μπορεί να βελτιωθεί η ταξινόμηση δερματικών βλαβών όταν παντρεύεις κλασικά CNN με μηχανισμό transformer, χωρίς να απαιτείται pretraining σε τεράστια datasets. Προτείνουν το EnDBoT (Enhanced deep bottleneck transformer), όπου ξεκινούν από ισχυρά backbones όπως ResNet50 ή DenseNet201 και, αντί να κρατήσουν τις τελευταίες χωρικές συνελίξεις 3×3 στα τελικά στάδια, τις αντικαθιστούν με ένα bottleneck transformer module που βασίζεται σε self-attention. Έτσι, το μοντέλο μπορεί να δει πιο εύκολα μακρινές συσχετίσεις στην εικόνα, πέρα από το καθαρά τοπικό πεδίο που επιβάλλει η συνέλιξη. Η βασική τους τεχνική βελτίωση ονομάζεται DPeSA, όπου εισάγουν position encoding τόσο στο query όσο και στο key (dual positional encoding), ώστε η προσοχή να κωδικοποιεί με πιο ισορροπημένο τρόπο τη χωρική πληροφορία και να μην χάνει τη γεωμετρία της βλάβης. Στην αξιολόγηση αναφέρουν στο ISIC 2017 Accuracy 92.1%, Sensitivity 90.1%, Specificity 91.9%, ενώ στο HAM10000 δίνουν Accuracy 95.84% και Precision 96.1%. Το κεντρικό συμπέρασμα είναι ότι το self-attention, όταν κουμπώνει πάνω σε ήδη καλές CNN αναπαραστάσεις, μπορεί να ανεβάσει μετρήσιμα την επίδοση, κρατώντας το μοντέλο σχετικά πρακτικό σε δεδομένα/εκπαίδευση.

Το άρθρο [26] εστιάζει στην πρόωμη διάγνωση μελανώματος έναντι καλοήθους από δερματοσκοπικές εικόνες, με στόχο να συνδυάσει τα πλεονεκτήματα CNN και transformers. Προτείνουν το MABSCNET, ένα υβριδικό μοντέλο που κρατά από τα CNN την ικανότητα να πιάνουν τοπικά μοτίβα (υφή, μικρές δομές) και από το transformer multi-head attention τη μοντελοποίηση πιο μακρινών συσχετίσεων στην εικόνα. Αρχιτεκτονικά χρησιμοποιούν depthwise separable conv blocks για πιο αποδοτική εξαγωγή χαρακτηριστικών και multi-layer feature fusion (σε λογική Dense/Residual ροής) ώστε να ενώνουν πληροφορία από διαφορετικά βάθη. Παράλληλα σχεδιάζουν και ένα ελαφρύ ViT με 8 encoder blocks για σύγκριση. Η βασική αξιολόγηση γίνεται στο ISIC 2020 με 1168 εικόνες και 2 κλάσεις, split 80/20, resize 224×224 και normalization, ενώ εφαρμόζουν ενίσχυση εικόνας με CLAHE και αξιοποιούν hypercolumn ιδέα για πλουσιότερη

αναπαράσταση. Σε αυτό το βασικό στήσιμο αναφέρουν 78.63% για MABSCNET (με CLAHE) έναντι 76.50% για το ViT. Το κύριο συμπέρασμα όμως έρχεται όταν το υβριδικό framework μπει σε πιο βαρύ setup με ImageNet-pretrained backbones, hard-voting ensemble (SVM/RF/KNN/AdaBoost) και metaheuristic tuning, όπου η ακρίβεια ανεβαίνει σε 92.74% (ISIC 2020), 100% (ISIC 2018) και 94.24% (Kaggle).

Το άρθρο [27] είναι μια συγκριτική μελέτη που βάζει CNN και Vision Transformers δίπλα-δίπλα κάτω από όσο γίνεται ίδιες συνθήκες, ώστε η σύγκριση να μην γέρνει από διαφορετικά πρωτόκολλα. Οι συγγραφείς εκπαιδεύουν 15 CNN και 15 ViT με transfer learning από ImageNet, τυποποιημένο preprocessing με resize 224×224, online augmentation, split 70/10/20 και κοινές επιλογές βελτιστοποίησης (π.χ. SGD), και αξιολογούν με accuracy, precision, recall και F1. Τα πειράματα γίνονται σε δύο κλασικά benchmarks, HAM10000 και ISIC 2019, ώστε να φανεί αν τα συμπεράσματα κρατάνε σε διαφορετικά distributions. Στα αποτελέσματα, το καλύτερο ViT είναι ένα Swin Transformer, με accuracy 0.9212 στο HAM10000 και 0.9187 στο ISIC 2019. Από την πλευρά των CNN, το καλύτερο στο HAM10000 είναι το RexNet-100 με accuracy 0.9192, πολύ κοντά αλλά ελαφρώς χαμηλότερα από το Swin. Το βασικό μήνυμα της εργασίας είναι ότι τα ViT τείνουν να δίνουν μικρό προβάδισμα στην ακρίβεια, αλλά αυτό συνήθως συνοδεύεται από μεγαλύτερο κόστος σε παραμέτρους και υπολογιστικούς πόρους, άρα η επιλογή αρχιτεκτονικής δεν είναι μόνο θέμα score αλλά και πρακτικής υλοποίησης.

Το [28] προτείνει ένα μοντέλο για τμηματοποίηση δερματικών βλαβών που συνδυάζει CNN και transformer με τρόπο αρκετά μηχανικό και πρακτικό. Ο πυρήνας είναι ένας dual encoder: από τη μία ResNet-34 για να κρατά καλά τα τοπικά μοτίβα και την υφή, και από την άλλη ένας transformer τύπου DeiT για να μεταφέρει global context (πληροφορία σε μεγαλύτερη κλίμακα). Για να μη σπάνε οι skip connections από το ότι τα features προέρχονται από δύο διαφορετικές οικογένειες (CNN vs transformer), εισάγουν Feature Adaptation Modules (FAM) που προσαρμόζουν τις κατανομές των χαρακτηριστικών πριν περάσουν στον decoder. Στον decoder χρησιμοποιούν μια πιο memory-efficient σχεδίαση που κάνει άθροιση αντί για concatenation, ενώ το channel fusion ενισχύεται με Squeeze-and-Excitation, ώστε να τονίζονται τα χρήσιμα κανάλια χωρίς μεγάλο overhead. Η αξιολόγηση γίνεται σε ISIC 2016/2017/2018 και PH2, με ισχυρά αποτελέσματα: στο ISIC 2018 αναφέρουν ACC 95.78%, IoU 82.02%, Dice 89.03%, στο ISIC 2016 ACC 96.04%, IoU 85.30%, Dice 91.59%, και στο PH2 ACC 97.03%, IoU 89.62%, Dice 94.40%, με ταχύτητα περίπου 50 FPS. Το βασικό μήνυμα είναι ότι ο συνδυασμός CNN+transformer, όταν “κουμπώνει” σωστά μέσω FAM και ελαφριού decoder, δίνει πιο καθαρά όρια χωρίς δυσανάλογο υπολογιστικό κόστος.

Το άρθρο [29] προτείνει ένα υβριδικό μοντέλο ταξινόμησης δερματικών βλαβών που πατά στη λογική MetaFormer/CAFormer, δηλαδή σε αρχιτεκτονική όπου το “token mixer” αλλάζει ανά στάδιο. Η βασική τους επιλογή είναι να κρατήσουν ConvNeXt blocks στα δύο πρώτα στάδια, ώστε το δίκτυο να αποτυπώνει καθαρά τοπικές λεπτομέρειες όπως υφή, άκρες και χρωματικές μεταβολές, και να περάσουν σε transformer blocks στα δύο τελευταία στάδια για να μάθουν πιο global σχέσεις και μακρινές εξαρτήσεις που συχνά σχετίζονται με το συνολικό σχήμα/κατανομή της βλάβης. Η εκπαίδευση γίνεται με transfer learning από ImageNet, online augmentation και ενιαίο training setup (εικόνες 224×224, SGD με lr 0.001, batch 16), ώστε οι συγκρίσεις να είναι πιο “καθαρές”. Αξιολογούν σε HAM10000 (7

κλάσεις) και ISIC 2019 (8 κλάσεις) με split 70/10/20. Αναφέρουν 94.30% accuracy και 91.11% F1 στο HAM10000, και 92.50% accuracy με 90.38% F1 στο ISIC 2019. Το βασικό συμπέρασμα είναι ότι η ρητή μίξη local και global αναπαράστασης μέσα στο ίδιο δίκτυο οδηγεί σε πιο σταθερή και υψηλή επίδοση σε σχέση με καθαρά CNN-only ή καθαρά ViT-only προσεγγίσεις.

Το [30] μια true hybrid αρχιτεκτονική για δυαδική ταξινόμηση Melanoma vs Nevus, με στόχο να συνδυαστεί η ισχύς των deep features με πιο ελεγχόμενα, ερμηνεύσιμα χαρακτηριστικά. Ξεκινούν με preprocessing που στοχεύει τα κλασικά artefacts: αφαιρούν τρίχες με black-hat και inpainting και διορθώνουν το μαύρο border μέσω zoom, ώστε να μειώνεται ο θόρυβος που μπερδεύει τα μοντέλα. Στη συνέχεια χρησιμοποιούν ResNet50 με transfer learning για αρχική πρόβλεψη και για εξαγωγή deep features. Αυτά τα deep features τα ενώνουν με 11 χειροποίητα χαρακτηριστικά (κυρίως χρώμα και υφή) και τα δίνουν σε τέσσερις κλασικούς ταξινομητές (π.χ. SVM, Random Forest, KNN, Gaussian Naive Bayes). Στην τελική απόφαση χρησιμοποιούν ένα FFNN ensemble με gating και attention, που ζυγίζει δυναμικά τόσο τα features όσο και τις επιμέρους προβλέψεις, σαν προσαρμοστική συγχώνευση. Αξιολογούν σε HAM10000, σε ISIC-derived σύνολο από Kaggle, και στο MED-NODE, με AUC 97% / 91% / 98% και F1 92% / 87% / 94% αντίστοιχα. Το βασικό μήνυμα είναι ότι η προσαρμοστική συγχώνευση “μιμείται” την πολυ-κριτηριακή σκέψη του δερματολόγου, βελτιώνοντας επίδοση και ερμηνευσιμότητα μέσω Grad-CAM και SHAP.

Το άρθρο [31] προτείνει ένα Fully Transformer Network (FTN) για ανάλυση δερματικών βλαβών που καλύπτει τμηματοποίηση και ταξινόμηση χωρίς καθόλου CNN, βασιζόμενο αποκλειστικά σε self-attention. Για να αποφύγουν το κλασικό πρόβλημα ότι οι transformers χάνουν τοπική υφή και κοστίζουν πολύ σε υπολογισμό, εισάγουν τρεις βασικές ιδέες: (1) Sliding Window Tokenization, όπου η εικόνα μετατρέπεται σε ιεραρχικά tokens με τρόπο που διατηρεί τοπικές λεπτομέρειες, (2) Spatial Pyramid Transformer, που μειώνει το κόστος κάνοντας pooling στα K/V μέσα στο multi-head attention, και (3) decoder τύπου transformer με skip connections σε λογική U-Net, ώστε να ανακτώνται λεπτομέρειες και ακμές στη μάσκα. Στο ISIC 2018 για segmentation (2594 εικόνες, split 2076/518, είσοδος 512×512) αναφέρουν Accuracy 96.2, Jaccard Index 82.8 και Dice 89.8, ξεπερνώντας FCN/UNet/ResUNet. Για classification (10,025 εικόνες, 8200/2005, 224×224) δίνουν AUC 87.9, που ανεβαίνει σε AUC 89.7 όταν γίνεται pretrain από το segmentation (αναφέρουν και mAUC 97.3). Το μήνυμα είναι ότι ένας “καθαρός” transformer μπορεί να σταθεί χωρίς τεράστιο pretraining, όταν το tokenization και το attention είναι ιεραρχικά και αποδοτικά.

Το άρθρο [32] εστιάζει στο ότι, στην ταξινόμηση δερματοσκοπικών εικόνων, δεν αρκεί ένα υψηλό score αλλά χρειάζεται και ένας τρόπος να μετριέται αν οι εξηγήσεις είναι αξιόπιστες. Οι συγγραφείς εκπαιδεύουν έναν ταξινομητή με fine-tuned EfficientNet-B0 για τρεις κλάσεις (MEL/NV/SK) και προτείνουν τον δείκτη TlxAI, που ποσοτικοποιεί αν τα heatmaps του XAI “πέφτουν” πάνω στη βλάβη ή στο υπόβαθρο. Ορίζουν το $TlxAI = RD(lesion) - RD(non-lesion)$, όπου το RD υπολογίζεται από heatmaps (π.χ. Grad-CAM, LIME, Occlusion Sensitivity Analysis (OSA), SHAP) αφού πρώτα “κοπούν” με το ground-truth mask της βλάβης και με το συμπλήρωμά του μέσω Hadamard product. Με αυτόν τον τρόπο ο δείκτης βγαίνει θετικός όταν η εξήγηση εστιάζει κυρίως στη βλάβη και όχι σε άσχετες περιοχές. Η αξιολόγηση γίνεται σε αναθεωρημένο σύνολο της International Skin Imaging Collaboration με 2325 εικόνες και 7-fold cross-validation, όπου αναφέρουν μέση ακρίβεια $87.57 \pm 1.32\%$ (μέγιστη 90.1%). Στα αποτελέσματα του δείκτη, το TlxAI βγαίνει

θετικό για Grad-CAM/LIME/OSA, ενώ για SHAP είναι σχεδόν μηδενικό. Το βασικό συμπέρασμα είναι ότι η ακρίβεια από μόνη της δεν λέει όλη την ιστορία: χρειάζεται και μέτρηση του αν το μοντέλο “κοιτά” σωστά τη βλάβη.

Το άρθρο [33] μελετά πώς μπορεί να γίνει ταξινόμηση δερματικών βλαβών με λιγότερο κόστος επισήμανσης, αξιοποιώντας active learning. Οι συγγραφείς δουλεύουν πάνω στο HAM10000 και οργανώνουν το πείραμα σε στάδια: αρχικά εφαρμόζουν augmentation και μετά ορίζουν τρία σενάρια δεδομένων, όπου αλλάζει το πόσο καθαρή είναι η περιοχή της βλάβης. Στο D1 χρησιμοποιούν τις εικόνες χωρίς segmentation, στο D2 κάνουν segmentation με U-Net, ενώ στο D3 χρησιμοποιούν masks από FCN που διορθώνονται χειροκίνητα από ειδικό, άρα η περιοχή ενδιαφέροντος είναι πιο αξιόπιστη. Πάνω σε αυτά εξάγουν χαρακτηριστικά τόσο χειροποίητα όσο και deep, και εκπαιδεύουν κλασικούς ταξινομητές. Στη συνέχεια εφαρμόζουν active learning στρατηγικές επιλογής δειγμάτων, όπως Least Confidence, Entropy και Margin Sampling, μαζί με μεθόδους batch επιλογής (π.χ. DBE, MST-BE, RDS) για να διαλέγονται τα πιο χρήσιμα δείγματα προς επισήμανση. Στην αποτίμηση, από τα παραδοσιακά μοντέλα ξεχωρίζουν BIC και Random Forest, με το καλύτερο αποτέλεσμα στο D3 να φτάνει 92.13%. Με active learning, το Margin Sampling ανεβάζει την επίδοση περίπου στο 93% με πολύ λιγότερες ετικέτες: στο D3 αναφέρουν 93.24% χρησιμοποιώντας μόνο 35% του training set, έναντι 92.13% όταν επισημαίνεται το 100%. Το βασικό συμπέρασμα είναι ότι καλό segmentation και έξυπνη επιλογή δειγμάτων μειώνουν δραστικά το labeling cost χωρίς ουσιαστική πτώση ακρίβειας.

Το άρθρο [34] προτείνει το SapFormer, ένα μοντέλο για τμηματοποίηση δερματικών βλαβών που στοχεύει να χειρίζεται καλύτερα τις διαφορετικές κλίμακες και τον θόρυβο εκτός βλάβης. Ο encoder είναι multi-scale hybrid transformer (συνδυασμός CNN και ViT) και παράγει αναπαραστάσεις σε κλίμακες από 1/4 έως 1/32, ώστε να κρατά μαζί λεπτομέρεια και παγκόσμιο πλαίσιο. Ο βασικός μηχανισμός είναι το Sap module (self-adaptive position-aware): χωρίζει το feature map σε πλέγμα $N \times N$ περιοχών και, μέσω transformer decoder, δημιουργεί δυναμικά weights και bias ανά περιοχή, σαν να προσαρμόζει το segmentation στις τοπικές ιδιαιτερότητες. Επιπλέον, με cross-attention κάθε περιοχή μπορεί να “δανείζεται” πληροφορία από άλλες, μειώνοντας ψευδείς ενεργοποιήσεις σε μη-βλάβες. Το τελικό mask προκύπτει με upsampling και multi-scale mask fusion. Στα ISIC 2016/2017/2018 αναφέρουν ACC 97.9/94.3/95.7%, IoU 93.2/86.4/89.4% και Dice 96.4/92.6/94.3%. Το κεντρικό συμπέρασμα είναι καλύτερα όρια και αντοχή σε εξω-βλαβικό θόρυβο, με τίμημα αυξημένη υπολογιστική πολυπλοκότητα.

Το άρθρο [35] προτείνει ένα μοντέλο για πολυ-τροπική διάγνωση δερματικών βλαβών που αξιοποιεί ταυτόχρονα δερματοσκοπική εικόνα, κλινική εικόνα και meta-data. Η βασική ιδέα είναι να μη γίνει απλή “late fusion” στο τέλος, αλλά μια throughout fusion στρατηγική όπου οι πληροφορίες των modalities αλληλεπιδρούν σταδιακά. Το TFormer είναι “pure transformer” pipeline με backbone Swin-Tiny. Για τις δύο εικόνες εφαρμόζουν dual-branch blocks με cross-attention (HMT), που συγχωνεύουν χαρακτηριστικά από ρηχά προς βαθιά layers, ώστε να μαθαίνονται έγκαιρα οι συμπληρωματικές ενδείξεις (π.χ. υφή στη δερματοσκόπηση και μακρο-μορφολογία στην κλινική εικόνα). Έπειτα, τα meta-data περνούν από MLP και ενσωματώνονται με ένα post-fusion transformer module (MTP), που λειτουργεί σαν τελικός συντονιστής των τριών πηγών πληροφορίας. Στο Derm7pt (train/val/test 413/203/395) αναφέρουν avg accuracy 77.99% και diagnostic accuracy

80.03%, ξεπερνώντας late-fusion concatenation και άλλες SOTA συγκρίσεις. Το βασικό μήνυμα είναι ότι η throughout fusion μαθαίνει καλύτερα τις σχέσεις μεταξύ modalities, άρα βελτιώνει τη διάγνωση.

Το άρθρο [36] προτείνει το SkinTrans, ένα Vision Transformer “πειραγμένο” ειδικά για δερματοσκοπικές εικόνες και τα πρακτικά τους προβλήματα. Η μεθοδολογία βασίζεται σε τρία σημεία. Πρώτα, κάνουν multiscale tokenization με overlapping sliding windows, παράγοντας patches σε τρεις κλίμακες (8/16/32), ώστε το μοντέλο να βλέπει ταυτόχρονα λεπτή υφή/άκρες και πιο ευρύ πλαίσιο της βλάβης. Δεύτερον, αντιμετωπίζουν το έντονο class imbalance με balanced sampling και ένα σχήμα label shuffling, ώστε να μειώνεται η τάση του μοντέλου να ευνοεί τις συχνές κλάσεις. Τρίτον, εκπαιδεύουν με contrastive learning μαζί με cross-entropy, για να έρχονται κοντά στον χώρο αναπαράστασης εικόνες της ίδιας κλάσης και να απομακρύνονται οι διαφορετικές. Στην αποτίμηση, στο HAM10000 αναφέρουν Accuracy 94.3%, Precision 0.941 και AUC 0.987, ενώ σε δικό τους κλινικό dataset (BCC/SCC/MM) δίνουν Accuracy 94.1%, Precision 0.942 και F1 0.941. Το βασικό συμπέρασμα είναι ότι το multiscale tokenization μαζί με contrastive loss βελτιώνει τη γενίκευση και οδηγεί σε πιο σωστή εστίαση στη βλάβη, όπως φαίνεται και από Grad-CAM.

Το άρθρο [37] προτείνει ένα σχήμα ensemble για πολυ-κλασική ταξινόμηση δερματικών βλαβών σε συνθήκες έντονου class imbalance. Οι συγγραφείς χτίζουν ένα weighted-average ensemble από πέντε προεκπαιδευμένα CNN (ResNeXt, SeResNeXt, DenseNet, Xception, ResNet), ώστε κάθε backbone να συνεισφέρει διαφορετικού τύπου οπτικά cues. Στο preprocessing στοχεύουν κλασικά προβλήματα δερματοσκόπησης: κάνουν hair removal με black-hat και inpainting, εφαρμόζουν median filter για θόρυβο, και φέρνουν όλες τις εικόνες σε 224×224. Στη φάση εκπαίδευσης χρησιμοποιούν augmentation και cost-sensitive learning για να περιορίσουν την τάση του μοντέλου να αγνοεί τις σπάνιες κλάσεις. Η αξιολόγηση γίνεται σε 18,730 εικόνες από HAM10000 και ISIC 2019, με 7 κλάσεις. Στα αποτελέσματα, το καλύτερο single model είναι το DenseNet με macro-recall 0.91, ενώ το απλό average ensemble φτάνει 93% και το weighted ensemble 0.94, με accuracy 88% και MCC 88%. Το βασικό συμπέρασμα είναι ότι το ensemble βελτιώνει αισθητά την ανάκληση σε ανισόρροπα δεδομένα, με το DenseNet να λειτουργεί ως ισχυρός οδηγός μέσα στο μείγμα.

Το άρθρο [38] παρουσιάζει το XAAI-Ledger, ένα πολυτροπικό μοντέλο για έγκαιρη διάγνωση melanoma και non-melanoma με συνδυασμό δερματοσκοπικών εικόνων και κλινικών δεδομένων. Η μέθοδος στηρίζεται σε CNN για εξαγωγή τοπικών οπτικών χαρακτηριστικών, σε Transformer Encoder για μάθηση πιο μακρινών σχέσεων και σε attention-based fusion για σύντηξη εικόνας και metadata. Για την ερμηνευσιμότητα ενσωματώνουν Grad-CAM και SHAP, ώστε να δείχνουν τόσο τις περιοχές της εικόνας που επηρέασαν την απόφαση όσο και τη συμβολή των κλινικών χαρακτηριστικών. Στην αξιολόγηση αναφέρουν κυρίως χρήση των ISIC 2019 και HAM10000, με split 75% training / 25% testing. Τα αποτελέσματα που δίνουν είναι Accuracy 98.71%, Sensitivity 97.45% και Specificity 98.22%, υποστηρίζοντας υπεροχή έναντι κλασικών CNN, ViT και hybrid CNN-Transformer μοντέλων. Το βασικό συμπέρασμα είναι ότι ο συνδυασμός πολυτροπικής σύντηξης και explainability μπορεί να βελτιώσει όχι μόνο την απόδοση αλλά και την κλινική αποδοχή του συστήματος.

Το άρθρο [39] προτείνει ένα ενιαίο πλαίσιο για τμηματοποίηση και ταξινόμηση δερματικών βλαβών με έμφαση στην εξηγησιμότητα. Η μεθοδολογία βασίζεται σε τρία κύρια μέρη. Πρώτα, χρησιμοποιούν ένα custom CNN για εξαγωγή χαρακτηριστικών από τις δερματοσκοπικές εικόνες. Δεύτερον, εισάγουν το PAM-UNet για segmentation, όπου το Parallel Attention Module (PAM) συνδυάζει Transformer Attention Module (TAM) και Global Spatial Attention (GSA) ώστε να ενισχύει τα χρήσιμα χαρακτηριστικά και να μειώνει artifacts όπως τρίχες, θόρυβο και ασαφή όρια. Τρίτον, για classification χρησιμοποιούν attention-based CNN ενισχυμένο με LSTM, ώστε να αξιοποιούνται καλύτερα οι χωρικές και διαδοχικές σχέσεις των features. Στην αξιολόγηση, στα HAM10000 και ISIC2019, αναφέρουν για segmentation έως Jaccard 96.4% / Dice 98.7% στο HAM10000 και Jaccard 97.2% / Dice 98.5% στο ISIC2019, ενώ για classification δίνουν 99.9% accuracy στο HAM10000 και 99.1% στο ISIC2019. Το βασικό συμπέρασμα είναι ότι ο συνδυασμός segmentation, parallel attention και explainability με Grad-CAM βελτιώνει τόσο την ακρίβεια όσο και την ερμηνευσιμότητα του συστήματος.

Παρακάτω ακολουθεί ο συγκριτικός πίνακας των papers που μελετήθηκαν:

Ονομασία paper	Έτος	Μέθοδος / Αρχιτεκτονική	Dataset	Βασικά αποτελέσματα
<i>A novel DeepLabv3+ and vision-based transformer model for segmentation and classification of skin lesions</i>	2024	DeepLabv3+ για τμηματοποίηση και ViT για ταξινόμηση	ISIC-16, ISIC-17, ISIC-18, PH2, ISIC-19, ISIC-20, HAM10000	Segmentation accuracy: 98.90%, 98.38%, 99.45%, 100%. Classification accuracy: PH2 100%, ISIC-19 96.97%, ISIC-20 97.73%, HAM10000 100%
<i>Optimized deep learning for skin lesion segmentation and skin cancer detection</i>	2024	Bilateral filter, TransUNet-AEO, augmentation, feature extraction, JAEO-LeNet	SIIM-ISIC Melanoma Classification	Accuracy 91.99%, Sensitivity 90.95%, Specificity 92.13%
<i>An approach for multiclass skin lesion classification based on ensemble learning</i>	2021	Weighted average ensemble από ResNeXt, SeResNeXt, ResNet, Xception, DenseNet	HAM10000 + ISIC 2019	Weighted ensemble: Recall 0.94, Precision 0.87, F1 0.89, Accuracy 0.88
<i>FAT-Net: Feature adaptive transformers for automated skin lesion segmentation</i>	2022	Dual encoder CNN + Transformer, Feature Adaptation Modules	ISIC 2016, ISIC 2017, ISIC 2018, PH2	Υψηλή απόδοση σε segmentation
<i>SapFormer</i>	2024	Self-adaptive position-aware transformer model	ISIC-2016, ISIC-2017, ISIC-2018	ACC: 97.9%, 94.3%, 95.7%. IoU: 93.2%, 86.4%, 89.4%. DSC: 96.4%, 92.6%, 94.3%
<i>MultiExCam</i>	2026	Hybrid DL/ML framework, ResNet50 features, handcrafted features, ML classifiers, Grad-CAM + SHAP	HAM10000, ISIC, MED-NODE	F1: 92% HAM10000, 87% ISIC, 94% MED-NODE. AUC: 97%, 91%, 98%
<i>ARP-ViT-CNN</i>	2025	Hybrid ARP + ViT + CNN	Skin cancer dataset / multiclass setting	Accuracy 98.2%, Precision 0.94, Recall 0.96, F1 0.95

Ονομασία paper	Έτος	Μέθοδος / Αρχιτεκτονική	Dataset	Βασικά αποτελέσματα
<i>Explainable AI-Based Skin Cancer Detection Using CNN, Particle Swarm Optimization and Machine Learning</i>	2024	Xception feature extractor, PSO feature selection, ML classifiers, Grad-CAM, LIME, Occlusion	ISIC 2018, HAM10000	Accuracy 98.5% στο ISIC 2018 και 86.1% στο HAM10000
<i>Two-stage CNN with weakly supervised segmentation for skin lesion classification</i>	2025	CNN → Grad-CAM attention maps → guided segmentation → second-stage CNN	HAM10000, ISIC 2016	ISIC 2016: AUC 0.862, Specificity 0.981, Sensitivity 0.762
<i>Fusion of Transformer Attention and CNN Features for Skin Cancer Detection</i>	2024	Transfer learning + deep learning + ensemble learning	ISIC 2018, Kaggle benign/malignant	ACC 0.9274, PPV 0.8983, Sensitivity 0.9550, F1 0.9258
<i>Deep-learning-based dermoscopic classification study</i>	2025	DL-based dermoscopic classification	ISIC datasets	Αναφέρει συγκρίσεις με DL μοντέλα και segmentation/handcrafted προσεγγίσεις
<i>A framework to build accurate Convolutional Neural Network models for melanoma diagnosis</i>	2023	Active learning και CNN-based training	Skin cancer images	Αναφέρεται ως πλαίσιο που βελτιώνει την εκπαίδευση CNN
<i>Skin lesion diagnosis using ensembles, unscaled multi-crop</i>	2018	Ensemble από σύγχρονες CNN αρχιτεκτονικές, loss weighting	ISIC / HAM-related datasets	Βελτίωση σε ανισόρροπα datasets μέσω ensemble και loss weighting

Ονομασία paper	Έτος	Μέθοδος / Αρχιτεκτονική	Dataset	Βασικά αποτελέσματα
<i>evaluation and loss weighting</i>				
<i>Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification</i>	2020	Multi-scale / multi-network CNN fusion	ISIC 2018 + extra data	Recall περίπου 0.862 σύμφωνα με τη σύγκριση του Rahman et al.
<i>Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification</i>	2020	Integrated deep CNNs	ISIC 2018	Macro recall περίπου 0.871 στη σύγκριση του Rahman et al.
<i>Classification of skin lesions using an ensemble of deep neural networks</i>	2018	AlexNet, VGG, Inception ensemble	ISBI 2017	Accuracy 83.8%
<i>Analysis of skin lesion images with deep learning</i>	2021	Transfer learning ensemble	ISIC 2019 + άλλα datasets	Accuracy 63.4%
<i>Skin lesion analyser: An efficient seven-way multi-class skin cancer classification using MobileNet</i>	2020	MobileNet	HAM10000	Accuracy 83.1%

Ονομασία paper	Έτος	Μέθοδος / Αρχιτεκτονική	Dataset	Βασικά αποτελέσματα
<i>Convolutional neural networks using MobileNet for skin lesion classification</i>	2019	Modified MobileNet	Skin lesion datasets	Αναφέρει ότι augmentation και up-sampling βοηθούν την απόδοση
<i>The development of a skin cancer classification system for pigmented skin lesions using deep learning</i>	2020	Faster R-CNN με VGG-16 backbone	Clinical skin images	Accuracy 86.2%
<i>Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network</i>	2020	Hyper-connected CNN με multi-scale attention	7-point checklist / multimodal skin lesion data	Accuracy 74.9%
<i>Attention residual learning for skin lesion classification</i>	2019	Attention Residual Learning CNN	Skin lesion datasets	Ενισχύει discriminative representation μέσω attention
<i>Dermatologist-level classification of skin cancer with deep neural networks</i>	2017	Inception-v3 σε μεγάλο σύνολο κλινικών εικόνων	129.450 clinical images / 2.032 diseases	Dermatologist-level classification

Ονομασία paper	Έτος	Μέθοδος / Αρχιτεκτονική	Dataset	Βασικά αποτελέσματα
<i>XAAI-Ledger: Advancing Skin Cancer Diagnostics through Explainable AI-Driven Deep Learning and Distributed Ledger Technology</i>	2025	Πολυτροπικό μοντέλο με CNN για εξαγωγή οπτικών χαρακτηριστικών, Transformer Encoder για μάθηση σχέσεων μεγάλης εμβέλειας, attention-based fusion για σύντηξη εικόνας και metadata, και Grad-CAM + SHAP για explainability	ISIC 2019, HAM10000	Accuracy 98.71%, Sensitivity 97.45%, Specificity 98.22% για διάκριση melanoma / non-melanoma
<i>Explainable skin cancer diagnosis with parallel attention mechanism for segmentation and classification</i>	2026	Ενιαίο framework με custom CNN για feature extraction, PAM-UNet για segmentation και attention-based CNN + LSTM για classification. Το Parallel Attention Module (PAM) συνδυάζει Transformer Attention Module (TAM) και Global Spatial Attention (GSA). Explainability με Grad-CAM	ISIC 2019, HAM10000	Για segmentation: έως Jaccard 96.4% / Dice 98.7% στο HAM10000 και Jaccard 97.2% / Dice 98.5% στο ISIC2019. Για classification: έως 99.9% accuracy στο HAM10000 και 99.1% στο ISIC2019.

Πίνακας 2.1. Συγκριτικός πίνακας των papers

Από τη συγκριτική μελέτη των εργασιών που παρουσιάζονται στον Πίνακα 2.1 προκύπτει μια αρκετά καθαρή εικόνα για το πώς κινείται η σύγχρονη βιβλιογραφία στην αναγνώριση και κατηγοριοποίηση δερματικών βλαβών. Από τις 25 εγγραφές του πίνακα, η μία αφορά συστηματική ανασκόπηση, ενώ οι υπόλοιπες 24 είναι εμπειρικές μελέτες. Αν εστιάσουμε μόνο στις εμπειρικές εργασίες, φαίνεται ότι η βιβλιογραφία βασίζεται κυρίως σε ένα σχετικά περιορισμένο σύνολο datasets, με κυρίαρχη παρουσία της οικογένειας ISIC και του HAM10000. Πιο συγκεκριμένα, τα datasets της οικογένειας ISIC εμφανίζονται σε 16 από τις 24 εμπειρικές μελέτες, δηλαδή περίπου στο 63,6% των papers που εξετάστηκαν, ενώ το HAM10000 χρησιμοποιείται σε 12 από τις 24 μελέτες, δηλαδή περίπου στο 45,5%. Το PH2 εμφανίζεται σε 4 μελέτες (18,2%), το Derm7pt σε 2 μελέτες (9,1%), ενώ datasets όπως SkinCon, clinical dermoscopy dataset, Vitiligo, Kaggle dataset και Monkeypox dataset εμφανίζονται μεμονωμένα. Η εικόνα αυτή δείχνει ότι η βιβλιογραφία έχει διαμορφώσει ορισμένα βασικά benchmarks, πάνω στα οποία δοκιμάζονται οι περισσότερες μέθοδοι. Αυτό έχει το πλεονέκτημα ότι επιτρέπει μια στοιχειώδη σύγκριση ανάμεσα στις τεχνικές, αλλά από την άλλη πλευρά σημαίνει ότι πολλά συμπεράσματα παράγονται ξανά και ξανά πάνω στα ίδια περίπου δεδομένα. Αν εξετάσουμε πιο προσεκτικά τα ISIC datasets, παρατηρούμε ότι η πιο συχνή έκδοση στον πίνακα είναι το ISIC 2018, το οποίο χρησιμοποιείται σε 8 μελέτες, ακολουθούμενο από το ISIC 2019 που εμφανίζεται σε 6 μελέτες. Το ISIC 2016 και το ISIC 2017 εμφανίζονται επίσης συχνά, κυρίως σε εργασίες segmentation, ενώ το ISIC 2020 χρησιμοποιείται σε λιγότερες αλλά σημαντικές ταξινομητικές εργασίες. Αυτή η κατανομή δεν είναι τυχαία. Το ISIC 2018 αποτελεί ένα από τα πιο καθιερωμένα σύνολα δεδομένων για σύγκριση μεθόδων, τόσο στην ταξινόμηση όσο και στην τμηματοποίηση, ενώ το ISIC 2019 και το ISIC 2020 προσφέρουν πιο απαιτητικά σενάρια αξιολόγησης, ειδικά σε συνθήκες πολλών κλάσεων ή έντονης ανισορροπίας. Αντίστοιχα, το HAM10000 παραμένει ένα από τα πιο βασικά datasets της περιοχής, κυρίως επειδή περιλαμβάνει μεγάλο αριθμό δερματοσκοπικών εικόνων και επτά διακριτές κατηγορίες βλαβών, άρα επιτρέπει πιο ουσιαστική αξιολόγηση πολυκατηγορικών μοντέλων. Το PH2, παρότι αρκετά μικρότερο, εξακολουθεί να χρησιμοποιείται ως συμπληρωματικό benchmark ή ως dataset fine-tuning και εξωτερικής επιβεβαίωσης.

Ως προς τις τεχνικές που φαίνεται να δίνουν τα καλύτερα αποτελέσματα, η συνολική εικόνα δείχνει ότι οι πιο ισχυρές επιδόσεις προέρχονται κυρίως από transformer-based, ViT-based και υβριδικές CNN–Transformer προσεγγίσεις. Στο HAM10000, η υψηλότερη αναφερόμενη επίδοση στον πίνακα έρχεται από το EnDBoT (Enhanced Deep Bottleneck Transformer) με accuracy 98,83%, ενώ πολύ κοντά βρίσκεται και η μέθοδος Angular–Radial Partitioning (ARP) με CNN/ViT, η οποία φτάνει accuracy 98,2%, precision 97,8%, recall 96,9%, F1 97,4% και AUC 98,9%. Υψηλή επίδοση παρουσιάζει και το LesionAid, το οποίο με χρήση ViTGAN για συνθετική ενίσχυση του training έδωσε validation accuracy 97,4% στο HAM10000. Στο ίδιο dataset, η μέθοδος SkinTrans, που στηρίζεται σε transformer λογική με contrastive learning, φτάνει 94,3%, ενώ το υβριδικό ConvNeXt-based μοντέλο δίνει 94,30%. Ιδιαίτερο ενδιαφέρον έχει και η συγκριτική μελέτη CNNs vs ViTs comparative, επειδή εξετάζει 15 CNNs και 15 ViTs κάτω από το ίδιο πρωτόκολλο αξιολόγησης. Εκεί, το Swin Transformer αναδείχθηκε ως η καλύτερη αρχιτεκτονική στο HAM10000 με 92,12% accuracy, κάτι που έχει ειδική αξία, επειδή η σύγκριση γίνεται σε

πιο ελεγχόμενο πλαίσιο. Αυτό είναι σημαντικό, γιατί σε πολλές άλλες εργασίες οι διαφορές στις επιδόσεις μπορεί να οφείλονται όχι μόνο στην αρχιτεκτονική, αλλά και στο split των δεδομένων, στην προεπεξεργασία ή στον τρόπο αξιολόγησης.

Στα ISIC datasets, επίσης παρατηρείται υπεροχή προηγμένων και συχνά υβριδικών αρχιτεκτονικών. Η numerically υψηλότερη τιμή που εμφανίζεται στον πίνακα είναι το 100% accuracy στο ISIC 2018 από τη μελέτη Fusion of transformer attention and CNN features, κάτι που δείχνει πολύ υψηλή απόδοση, αλλά χρειάζεται προσεκτική ερμηνεία, επειδή τέτοιες εξαιρετικά υψηλές τιμές σε μεμονωμένα πειράματα δεν είναι πάντα εύκολα συγκρίσιμες με άλλες εργασίες που έχουν διαφορετικό πειραματικό σχεδιασμό. Πέρα από αυτή την περίπτωση, ιδιαίτερα ισχυρές επιδόσεις εμφανίζουν το OXAI-SCC-Net με 99,12% accuracy στο ISIC 2019, το Comprehensive pipeline (transfer learning + feature selection + PSO) με 98,5% accuracy στο ISIC 2018, καθώς και το DeepLabv3+ and Vision Transformer, που στην ταξινόμηση δίνει 97,73% στο ISIC 2019 και 96,97% στο ISIC 2020. Πολύ καλή εικόνα δίνει και το MultiExCam, το οποίο δεν δίνει μόνο ταξινόμηση αλλά και explainability, φτάνοντας AUC 97/91/98 και F1 92/87/94 στα ISIC 2018/2019/2020 αντίστοιχα. Με βάση αυτά τα δεδομένα, φαίνεται ότι οι καλύτερες αποδόσεις στα ISIC benchmarks συνδέονται κυρίως με transformer-oriented λογικές, με συνδυασμό attention mechanisms, και με υβριδικά σχήματα που αξιοποιούν τόσο τοπικά όσο και πιο συνολικά χαρακτηριστικά της εικόνας.

Στο κομμάτι της τμηματοποίησης (segmentation), οι πιο ισχυρές εργασίες του πίνακα είναι οι FAT-Net, FTN και SapFormer, ενώ η ανασκόπηση των Fatima et al. αναφέρει και γενικότερη υπεροχή αρχιτεκτονικών όπως το Swin-UNet, με περίπου +7% στο DSC και +9% στο mIoU σε σχέση με το κλασικό U-Net. Από τα segmentation papers του πίνακα, η SapFormer παρουσιάζει από τις καλύτερες επιδόσεις, με DSC 96,4 και IoU 93,2 στο ISIC 2016, DSC 92,6 και IoU 86,4 στο ISIC 2017 και DSC 94,3 και IoU 89,4 στο ISIC 2018. Το FAT-Net επίσης δίνει πολύ ανταγωνιστικά αποτελέσματα στο ISIC 2018, με Dice 91,04, mIoU 85,02 και ACC 96,4, ενώ το FTN, ως πλήρως transformer δίκτυο χωρίς CNN backbone, φτάνει DSC 90,0 και JI 82,9. Εδώ φαίνεται αρκετά καθαρά ότι οι πιο αποδοτικές τεχνικές segmentation δεν στηρίζονται πλέον αποκλειστικά σε κλασικές CNN λογικές, αλλά υιοθετούν μηχανισμούς attention, υβριδική εξαγωγή χαρακτηριστικών ή αμιγώς transformer δομές. Αυτό έχει σημασία και για τη δική μας μελέτη, γιατί δείχνει ότι οι μετασχηματιστές δεν περιορίζονται μόνο στην ταξινόμηση, αλλά φαίνεται να έχουν έντονη παρουσία και στην ακριβή χωρική αποτύπωση της βλάβης. Ένα από τα πιο ουσιαστικά σημεία που προκύπτουν από τον πίνακα είναι ότι στα classification papers, τα metrics όπως accuracy, AUC, precision, recall και F1-score δεν πρέπει να διαβάζονται αποκομμένα από τη δομή του dataset. Αυτό ισχύει ιδιαίτερα για το HAM10000 και το ISIC 2019, όπου υπάρχει έντονη ανισορροπία κλάσεων. Στο HAM10000, οι επτά κατηγορίες δεν εκπροσωπούνται με τον ίδιο αριθμό δειγμάτων. Η κλάση των common nevus είναι πολύ πιο συχνή από άλλες κατηγορίες, όπως το μελάνωμα, τα dermatofibroma ή τα vascular lesions. Άρα, ένα μοντέλο μπορεί να εμφανίσει υψηλή συνολική accuracy επειδή αναγνωρίζει πολύ καλά τις πολυπληθείς κλάσεις, χωρίς να σημαίνει κατ' ανάγκη ότι αποδίδει το ίδιο καλά στις πιο δύσκολες ή πιο κρίσιμες κλινικές περιπτώσεις. Το ίδιο πρόβλημα εμφανίζεται και στο ISIC 2019, το οποίο έχει οκτώ κλάσεις και επίσης μη ισορροπημένη κατανομή. Σε τέτοιες περιπτώσεις, η accuracy από μόνη της μπορεί να δώσει μια κάπως ωραιοποιημένη εικόνα. Για αυτό, metrics όπως το recall, το F1-score και ειδικά η AUC έχουν ιδιαίτερη

σημασία, επειδή βοηθούν να φανεί αν το μοντέλο εντοπίζει σωστά τις λιγότερο συχνές αλλά κλινικά σημαντικές περιπτώσεις. Ακόμα πιο προσεκτική ανάγνωση χρειάζεται στο ISIC 2020, όπου το πρόβλημα είναι κυρίως δυαδικό αλλά επίσης ανισόρροπο, με πολύ περισσότερα αρνητικά δείγματα σε σχέση με τα θετικά. Σε τέτοιο σενάριο, μια υψηλή accuracy μπορεί να είναι λιγότερο εντυπωσιακή απ' όσο φαίνεται αρχικά, αν το μοντέλο δεν έχει επαρκή ευαισθησία στο μελάνωμα.

Ένα ακόμα ενδιαφέρον στοιχείο είναι ότι αρκετές από τις εργασίες δεν περιορίζονται στην πρόβλεψη της κλάσης, αλλά δίνουν βάρος και στην εξηγησιμότητα (XAI). Οι μελέτες Two-stage CNN and GradCAM++, OXAI-SCC-Net, MultiExCam και TixAI δείχνουν ότι η βιβλιογραφία έχει αρχίσει να μετακινείται από την απλή αναζήτηση της μέγιστης ακρίβειας προς μια πιο ουσιαστική αξιολόγηση του πώς λαμβάνεται η απόφαση. Αυτό είναι αρκετά σημαντικό σε ιατρικές εφαρμογές. Ένα μοντέλο που δίνει υψηλή επίδοση αλλά εστιάζει σε άσχετα σημεία της εικόνας δεν είναι το ίδιο αξιόπιστο με ένα μοντέλο που ενεργοποιείται πάνω στη βλάβη και ακολουθεί πιο «λογική» διαγνωστική συμπεριφορά. Ο δείκτης TixAI, για παράδειγμα, κινείται ακριβώς προς αυτή την κατεύθυνση, επειδή δεν εξετάζει μόνο το σωστό ή λάθος της πρόβλεψης, αλλά και το κατά πόσο οι εξηγήσεις του μοντέλου είναι αξιόπιστες σε σχέση με τη θέση της βλάβης. Αυτή η τάση συνδέεται άμεσα και με τη σημερινή ερευνητική απαίτηση για μοντέλα που δεν είναι μόνο αποδοτικά, αλλά και πιο ερμηνεύσιμα.

2.2 Σύνολα δεδομένων στη βιβλιογραφία

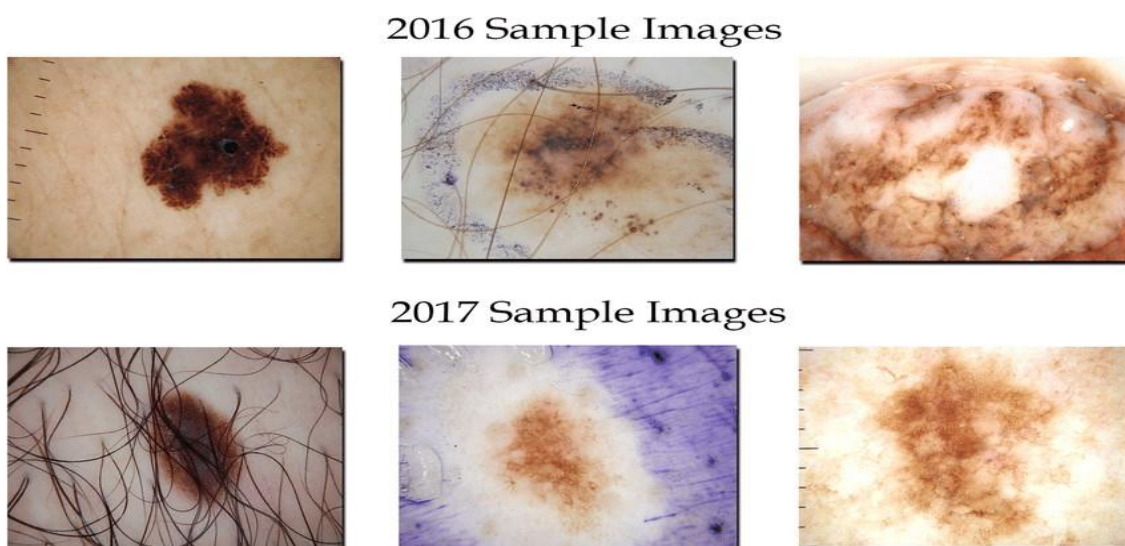
Στη βιβλιογραφία που αφορά την αναγνώριση, ταξινόμηση και τμηματοποίηση δερματικών βλαβών εμφανίζεται ένας σχετικά μεγάλος αριθμός δημόσιων συνόλων δεδομένων. Τα datasets αυτά δεν έχουν όλα τον ίδιο ρόλο. Ορισμένα έχουν καθιερωθεί ως βασικά benchmarks για την αξιολόγηση μοντέλων βαθιάς μάθησης, ενώ άλλα εμφανίζονται σε πιο ειδικές εργασίες, είτε για εξωτερικό έλεγχο είτε για διαφορετικά δερματολογικά προβλήματα. Η επιλογή του dataset επηρεάζει άμεσα το είδος του πειράματος, τον αριθμό των κλάσεων, τη δυσκολία του προβλήματος, τις κατάλληλες μετρικές και την ερμηνεία των αποτελεσμάτων.

Μέσα από τις εργασίες που μελετήθηκαν, τα πιο συχνά datasets είναι τα ISIC 2016–2020 και το HAM10000. Τα συγκεκριμένα σύνολα εμφανίζονται στα περισσότερα papers, είτε για classification είτε για segmentation. Σε μικρότερη συχνότητα εμφανίζονται τα PH2 και Derm7pt, ενώ σε πιο ειδικές περιπτώσεις εντοπίζονται datasets όπως SkinCon, Vitiligo dataset, Monkeypox dataset, clinical dermoscopy dataset, Kaggle-derived datasets και παράγωγα subsets όπως D1/D2/D3. Η διάκριση αυτή είναι χρήσιμη, επειδή τα πρώτα datasets μπορούν να θεωρηθούν βασικοί άξονες της βιβλιογραφίας, ενώ τα δεύτερα λειτουργούν περισσότερο συμπληρωματικά ή εξυπηρετούν ειδικότερα ερευνητικά ερωτήματα. Ένα βασικό ζήτημα που εμφανίζεται σχεδόν σε όλα τα datasets δερματοσκοπικών εικόνων είναι η ανισορροπία των κλάσεων. Σε αρκετές περιπτώσεις, οι καλοήθειες ή συχνότερες βλάβες εμφανίζονται σε πολύ μεγαλύτερο αριθμό σε σχέση με τις κακοήθειες ή σπανιότερες κατηγορίες. Αυτό έχει άμεση επίδραση στην αξιολόγηση. Ένα μοντέλο μπορεί να πετυχαίνει υψηλή συνολική accuracy, επειδή αναγνωρίζει σωστά την πλειοψηφική κλάση, αλλά να έχει χαμηλότερη απόδοση σε κατηγορίες με μεγαλύτερη κλινική σημασία, όπως το μελάνωμα. Για τον λόγο αυτό, στα σχετικά papers η accuracy

δεν αρκεί από μόνη της. Χρειάζονται και μετρικές όπως precision, recall, F1-score, macro-F1, AUC, καθώς και αναλυτική εξέταση της επίδοσης ανά κλάση.

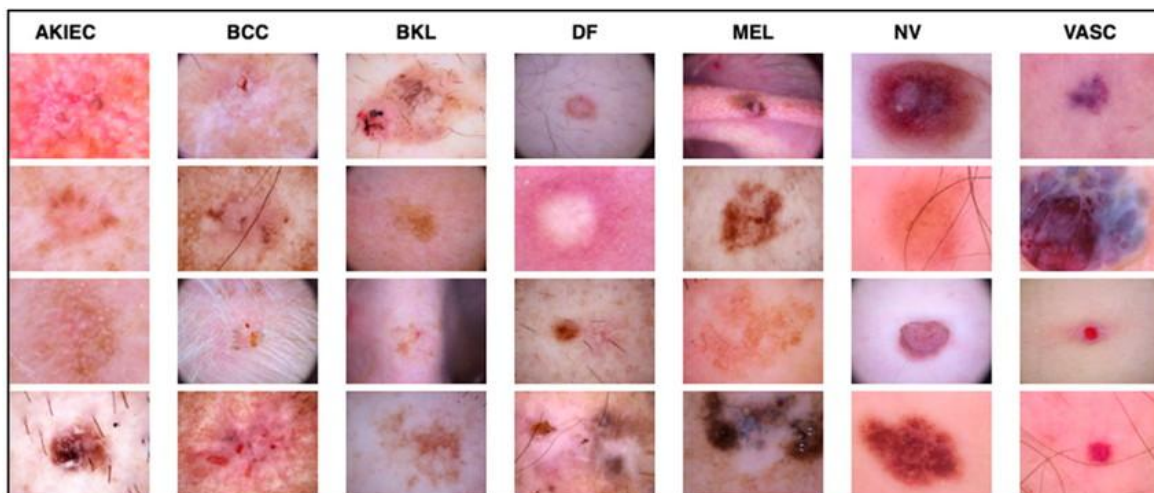
2.2.1 ISIC 2016–2018

Τα σύνολα δεδομένων ISIC 2016–2018 ανήκουν στα πιο καθιερωμένα benchmarks της δερματοσκοπικής ανάλυσης εικόνας. Έχουν αξιοποιηθεί σε μεγάλο αριθμό εργασιών, τόσο για τμηματοποίηση βλάβης όσο και για ταξινόμηση δερματικών αλλοιώσεων. Η σημασία τους συνδέεται με το γεγονός ότι παρείχαν οργανωμένα δεδομένα μέσα από διεθνείς προκλήσεις, επιτρέποντας τη σύγκριση διαφορετικών μοντέλων πάνω σε κοινά πειραματικά σύνολα. Στις εργασίες segmentation, τα ISIC 2016–2018 έχουν ιδιαίτερη αξία επειδή παρέχουν δερματοσκοπικές εικόνες μαζί με ground-truth masks, δηλαδή χειροκίνητες μάσκες που ορίζουν την περιοχή της βλάβης. Αυτό επιτρέπει την αξιολόγηση μοντέλων όχι μόνο ως προς το αν αναγνωρίζουν τη σωστή κλάση, αλλά και ως προς το αν εντοπίζουν σωστά τα όρια της βλάβης. Για τέτοιες εργασίες, οι πιο χρήσιμες μετρικές είναι το Dice Similarity Coefficient (DSC/Dice) και το Intersection over Union (IoU ή mIoU). Αυτές οι μετρικές εξετάζουν τον βαθμό επικάλυψης ανάμεσα στην προβλεπόμενη μάσκα και στη μάσκα αναφοράς.



Εικόνα 2.1. Μερικά παραδείγματα δερματοσκοπικών εικόνων από τα σύνολα δεδομένων ISIC-2016 και ISIC-2017. Πηγή: Ashraf et al. (2022).

Το ISIC 2018 έχει διπλή σημασία. Από τη μία πλευρά, στο Task 1 αξιοποιείται για segmentation, με εικόνες και μάσκες βλαβών. Από την άλλη πλευρά, στο Task 3 συνδέεται με το πρόβλημα της πολυκατηγορικής ταξινόμησης και βασίζεται στο HAM10000. Περιλαμβάνει 10.015 δερματοσκοπικές εικόνες σε 7 διαγνωστικές κλάσεις: NV, MEL, BCC, AKIEC, BKL, DF και VASC. Η κλάση NV αφορά μελανοκυτταρικούς σπίλους και είναι η πολυπληθέστερη. Η κλάση MEL αφορά το μελάνωμα, δηλαδή την κλινικά πιο κρίσιμη κατηγορία. Η BCC αφορά βασικοκυτταρικό καρκίνωμα, ενώ η AKIEC περιλαμβάνει ακτινικές κερατώσεις και ενδοεπιθηλιακό καρκίνωμα. Η BKL αντιστοιχεί σε καλοήθεις κερατωσικές βλάβες, η DF σε dermatofibroma και η VASC σε αγγειακές βλάβες.



Εικόνα 2.2. Παραδείγματα εικόνων από τις επτά κατηγορίες δερματικών βλαβών του HAM10000 dataset: AKIEC, BCC, BKL, DF, MEL, NV και VASC. Πηγή: Mehmood et al. (2023)

Η ταξινόμηση στις 7 αυτές κλάσεις δεν είναι ισοδύναμη για όλες τις κατηγορίες. Ορισμένες βλάβες εμφανίζονται συχνά και έχουν περισσότερα παραδείγματα εκπαίδευσης, ενώ άλλες εμφανίζονται σε πολύ μικρότερο αριθμό. Αυτό δημιουργεί πρόβλημα ανισορροπίας. Σε ένα τέτοιο πλαίσιο, ένα μοντέλο μπορεί να δείχνει υψηλή συνολική accuracy, χωρίς να σημαίνει ότι αναγνωρίζει εξίσου καλά τις σπανιότερες κλάσεις. Για παράδειγμα, αν το μοντέλο αποδίδει πολύ καλά στην κλάση NV, που έχει πολλά δείγματα, η συνολική του επίδοση μπορεί να φαίνεται υψηλή, ακόμη και αν δυσκολεύεται σε κλάσεις όπως DF ή VASC.

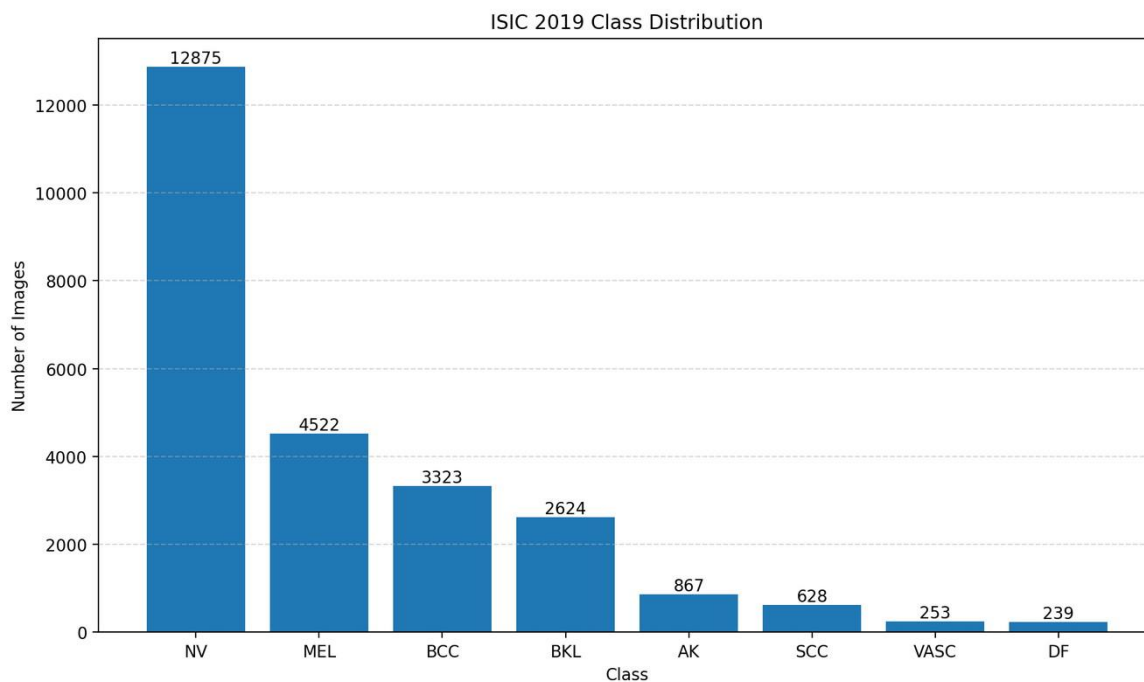
Στις εργασίες που μελετήθηκαν, τα ISIC 2016–2018 εμφανίζονται σε σημαντικό αριθμό segmentation και classification papers. Εργασίες όπως οι FAT-Net, FTN και SapFormer τα αξιοποιούν κυρίως για segmentation, δίνοντας έμφαση σε Dice και IoU. Άλλες εργασίες τα αξιοποιούν για classification, όπου η αξιολόγηση γίνεται με accuracy, AUC, F1-score ή άλλες μετρικές. Η διπλή χρήση των ISIC 2016–2018 τα καθιστά ιδιαίτερα χρήσιμα για τη σύγκριση διαφορετικών ειδών μοντέλων.

2.2.2 ISIC 2019

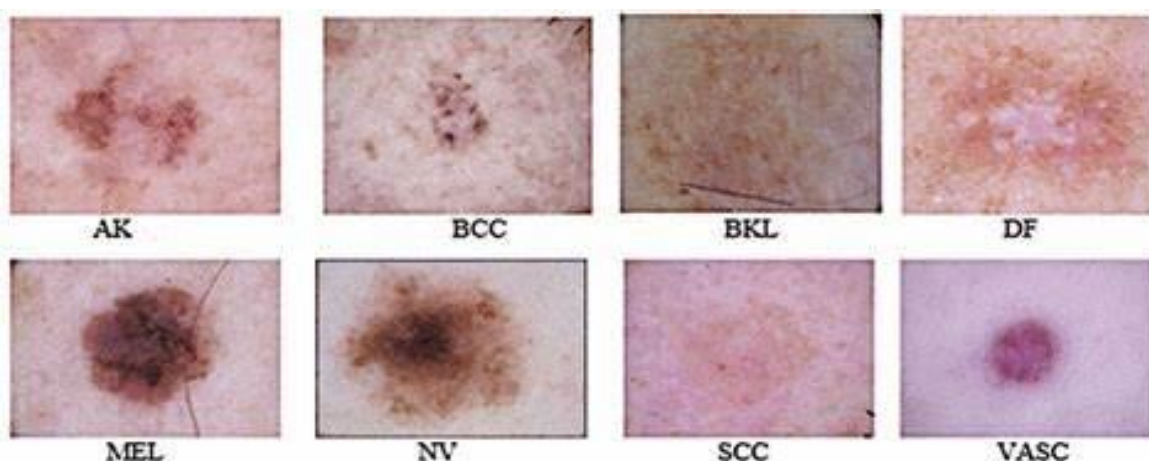
Το ISIC 2019 είναι ένα από τα σημαντικότερα datasets για πολυκατηγορική ταξινόμηση δερματικών βλαβών. Περιλαμβάνει 25.331 εικόνες εκπαίδευσης και σχεδιάστηκε για την αξιολόγηση μοντέλων που πρέπει να διακρίνουν περισσότερες διαγνωστικές κατηγορίες [44]. Σε σχέση με τα προηγούμενα ISIC datasets, το ISIC 2019 είναι πιο απαιτητικό, επειδή περιλαμβάνει μεγαλύτερο αριθμό εικόνων, περισσότερες κλάσεις και πιο έντονη ανισορροπία.

Οι βασικές κλάσεις που εμφανίζονται στο training set είναι οι εξής: MEL, NV, BCC, AK, BKL, DF, VASC και SCC. Η κλάση MEL αντιστοιχεί στο μελάνωμα. Η NV αντιστοιχεί στους μελανοκυτταρικούς σπίλους. Η BCC αφορά το βασικοκυτταρικό καρκίνωμα. Η AK αναφέρεται στην ακτινική κεράτωση, ενώ η BKL περιλαμβάνει καλοήθεις κερατωσικές βλάβες. Η DF αντιστοιχεί σε dermatofibroma, η VASC σε αγγειακές βλάβες και η SCC σε squamous cell carcinoma. Η προσθήκη της SCC σε σχέση με προηγούμενα datasets αυξάνει

τη διαγνωστική πολυπλοκότητα, επειδή το μοντέλο πρέπει να ξεχωρίσει περισσότερες κακοήθειες ή προκακοήθειες καταστάσεις. Το ISIC 2019 παρουσιάζει έντονο class imbalance. Οι κλάσεις NV και MEL έχουν σαφώς περισσότερα παραδείγματα σε σχέση με μικρότερες κλάσεις όπως DF, VASC, AK και SCC. Αυτό σημαίνει ότι η εκπαίδευση ενός μοντέλου πάνω στο ISIC 2019 δεν είναι ουδέτερη ως προς τις κλάσεις. Το μοντέλο εκτίθεται συχνότερα στις πολυπληθείς κατηγορίες και λιγότερο στις σπάνιες. Στην πράξη, αυτό μπορεί να οδηγήσει σε υψηλή συνολική accuracy αλλά χαμηλότερη ευαισθησία στις μικρές κατηγορίες.



Εικόνα 2.3. Κατανομή των βασικών κλάσεων στο ISIC 2019 training set.

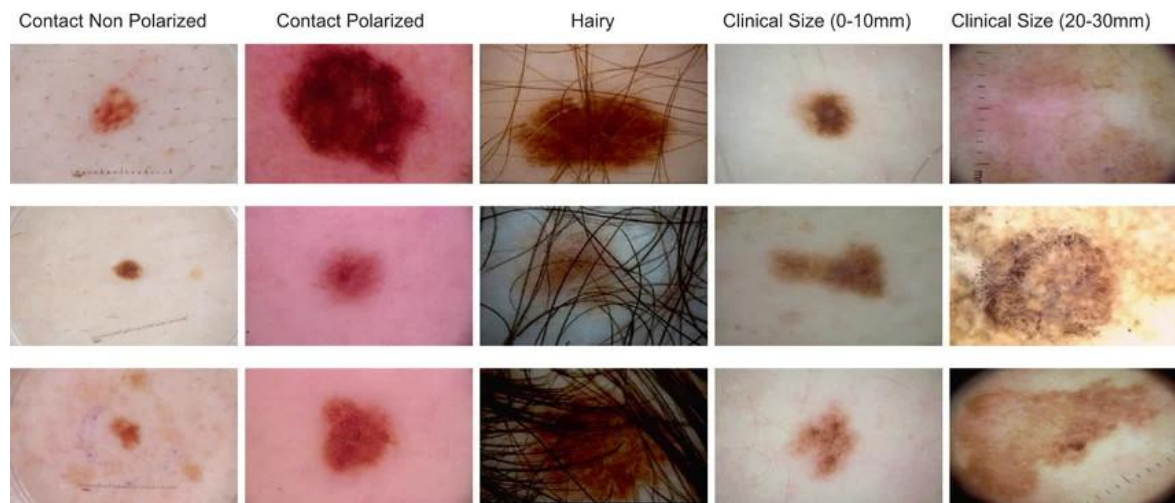


Εικόνα 2.4. Ενδεικτικά δείγματα δερματοσκοπικών εικόνων από το ISIC 2019 dataset για τις κατηγορίες AK, BCC, BKL, DF, MEL, NV, SCC και VASC. Πηγή: A deep analysis on high-resolution dermoscopic image classification (2021).

Για τον λόγο αυτό, στις εργασίες που βασίζονται στο ISIC 2019 προτιμώνται πιο αναλυτικές μετρικές, όπως macro-F1, weighted-F1, per-class recall, balanced accuracy και AUC ανά κλάση. Το macro-F1 είναι ιδιαίτερα χρήσιμο, επειδή δίνει ίσο βάρος σε κάθε κλάση, ανεξάρτητα από το πόσες εικόνες περιέχει. Αντίθετα, η απλή accuracy μπορεί να ευνοήσει τις πλειοψηφικές κλάσεις και να κρύψει χαμηλές επιδόσεις στις σπανιότερες. Στον πίνακα της βιβλιογραφίας, το ISIC 2019 εμφανίζεται σε εργασίες που εξετάζουν μοντέλα όπως OXAI-SCC-Net, EnDBoT, ConvNeXt-based hybrid approaches, συγκριτικές μελέτες CNNs vs ViTs, καθώς και μοντέλα με explainability όπως το MultiExCam. Η παρουσία του σε τόσες διαφορετικές εργασίες δείχνει ότι αποτελεί ένα από τα βασικά datasets για την αξιολόγηση σύγχρονων transformer-based και hybrid αρχιτεκτονικών.

2.2.3 ISIC 2020 / SIIM-ISIC

Το ISIC 2020, γνωστό και ως SIIM-ISIC Melanoma Classification Challenge dataset, διαφοροποιείται από τα ISIC 2018 και ISIC 2019, επειδή έχει κυρίως δυαδικό χαρακτήρα. Ο στόχος είναι η διάκριση ανάμεσα σε melanoma και non-melanoma περιπτώσεις. Περιλαμβάνει 33.126 training εικόνες και συνοδεύεται από πληροφορίες ασθενών, όπως patient identifiers [45].



Εικόνα 2.5. Παραδείγματα εικόνων από το σύνολο δεδομένων ISIC-2020 με διαφορετικά χαρακτηριστικά. Πηγή: Söylemez, Ö. F. (2023). Forward selection-based ensemble of deep neural networks for melanoma classification in dermoscopy images.

Η ύπαρξη patient identifiers είναι ένα από τα σημαντικότερα χαρακτηριστικά του ISIC 2020. Σε δερματολογικά datasets, ο ίδιος ασθενής μπορεί να έχει περισσότερες από μία εικόνες. Αν εικόνες από τον ίδιο ασθενή βρεθούν ταυτόχρονα στο training και στο validation ή test set, τότε μπορεί να δημιουργηθεί data leakage. Σε αυτή την περίπτωση, το μοντέλο δεν αξιολογείται πραγματικά σε ανεξάρτητα δεδομένα, επειδή ενδέχεται να έχει μάθει έμμεσα στοιχεία που σχετίζονται με τον ίδιο ασθενή ή με τη διαδικασία λήψης. Για αυτό, τα patient-wise splits θεωρούνται πιο αξιόπιστα από τα απλά random image-wise splits. Παρά το μεγάλο μέγεθός του, το ISIC 2020 παρουσιάζει έντονη ανισορροπία ανάμεσα στις δύο κλάσεις. Οι εικόνες non-melanoma είναι πολύ περισσότερες από τις εικόνες melanoma. Αυτό κάνει την accuracy προβληματική ως βασική μετρική. Ένα

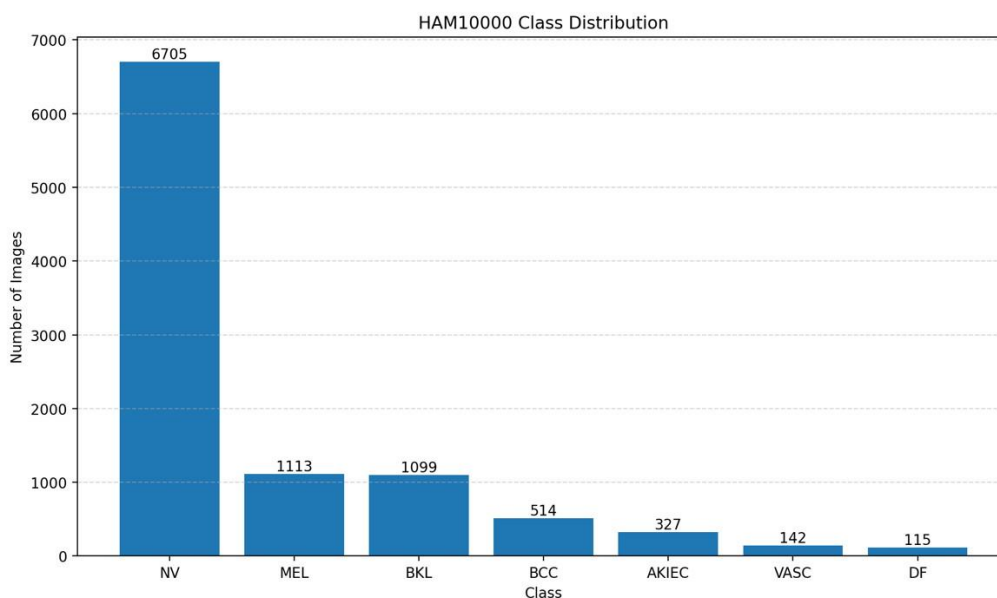
μοντέλο μπορεί να πετυχαίνει υψηλή accuracy προβλέποντας συχνά την πλειοψηφική κλάση, χωρίς να εντοπίζει με επάρκεια τις θετικές περιπτώσεις μελανώματος. Για αυτό, στο ISIC 2020 η AUC έχει ιδιαίτερη αξία, μαζί με το recall/sensitivity για την κλάση melanoma.

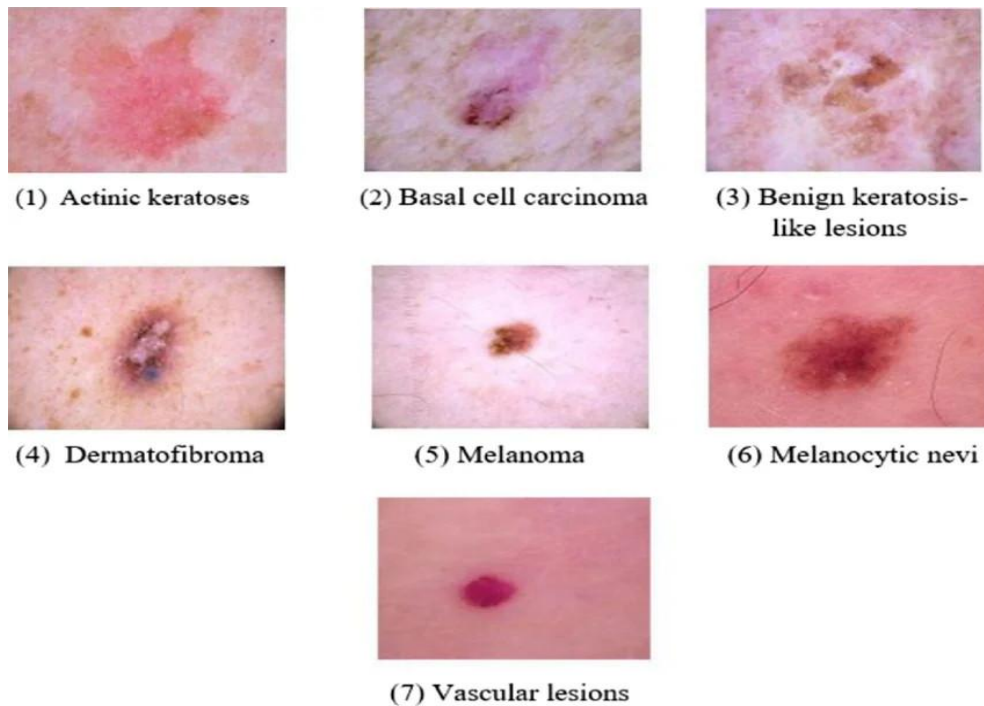
Στη βιβλιογραφία, το ISIC 2020 εμφανίζεται συχνά σε εργασίες binary classification, σε σχήματα pretraining/finetuning και σε μοντέλα που εξετάζουν τη δυνατότητα γενίκευσης. Στον πίνακα που μελετήθηκε, το ISIC 2020 εμφανίζεται σε εργασίες όπως το Fusion of transformer attention and CNN features και το MultiExCam. Η χρήση του είναι χρήσιμη για την αξιολόγηση μοντέλων που στοχεύουν στον εντοπισμό μελανώματος, αλλά τα αποτελέσματά του δεν πρέπει να συγκρίνονται άμεσα με πολυκατηγορικά datasets όπως το HAM10000 ή το ISIC 2019. Η δυαδική ταξινόμηση είναι διαφορετικό πρόβλημα από την ταξινόμηση επτά ή οκτώ κλάσεων.

2.2.4 HAM10000

Το HAM10000 είναι ένα από τα πιο συχνά datasets στη βιβλιογραφία για την ταξινόμηση δερματικών βλαβών. Το όνομά του προέρχεται από τη φράση “Human Against Machine with 10000 training images” και περιλαμβάνει 10.015 δερματοσκοπικές εικόνες σε 7 διαγνωστικές κλάσεις [41]. Χρησιμοποιήθηκε ως βάση και για το ISIC 2018 Task 3, γεγονός που εξηγεί γιατί εμφανίζεται τόσο συχνά σε συγκριτικές εργασίες.

Οι 7 κλάσεις του HAM10000 είναι οι εξής: NV, MEL, BKL, BCC, AKIEC, VASC και DF. Η κλάση NV περιλαμβάνει μελανοκυτταρικούς σπύλους και είναι η πιο πολυπληθής. Η MEL αντιστοιχεί στο μελάνωμα. Η BKL περιλαμβάνει benign keratosis-like lesions, δηλαδή καλοήθεις κερατωσικές βλάβες. Η BCC αφορά βασικοκυτταρικό καρκίνωμα, ενώ η AKIEC περιλαμβάνει ακτινικές κερατώσεις και ενδοεπιθηλιακό καρκίνωμα. Η VASC αφορά αγγειακές βλάβες και η DF αφορά dermatofibroma. Το βασικό χαρακτηριστικό του HAM10000 είναι η έντονη ανισορροπία των κλάσεων. Η κλάση NV είναι πολύ μεγαλύτερη από όλες τις υπόλοιπες. Οι συνήθεις αριθμοί που αναφέρονται για τις κλάσεις είναι περίπου: NV 6.705, MEL 1.113, BKL 1.099, BCC 514, AKIEC 327, VASC 142 και DF 115.



Εικόνα 2.6. Κατανομή των επτά κλάσεων στο HAM10000 dataset**Εικόνα 2.7.** Ενδεικτικές εικόνες για τις επτά κατηγορίες δερματικών βλαβών από το σύνολο δεδομένων HAM10000. Πηγή: Shetty et al., Skin lesion classification of dermoscopic images using machine learning and convolutional neural network, Scientific Reports, 2022

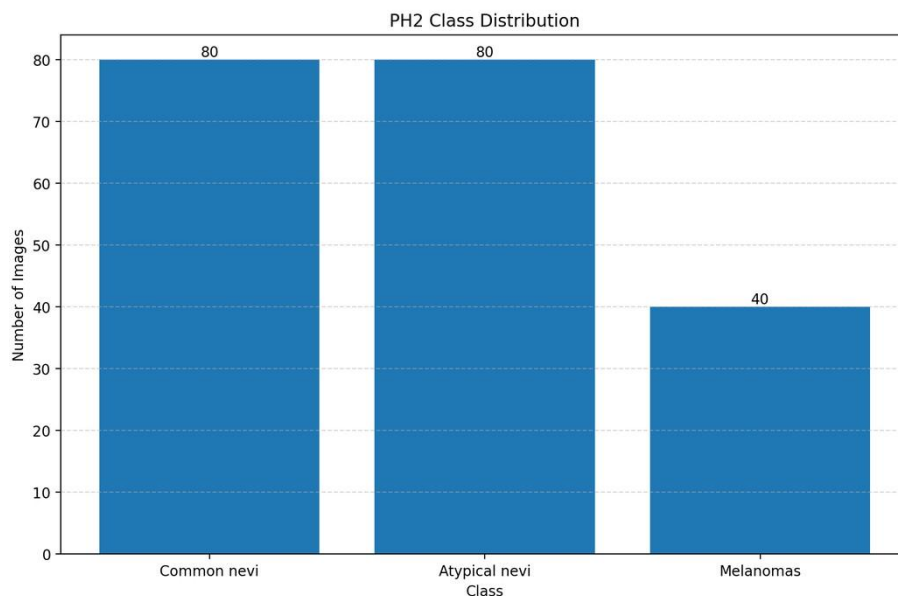
Όπως φαίνεται στην εικόνα 2.6 η κλάση NV καλύπτει περίπου τα δύο τρίτα του dataset, ενώ οι DF και VASC έχουν πολύ μικρό αριθμό δειγμάτων. Η διαφορά αυτή επηρεάζει έντονα τη συμπεριφορά των μοντέλων. Σε ένα τέτοιο dataset, η συνολική accuracy μπορεί να δώσει παραπλανητική εικόνα. Αν ένα μοντέλο προβλέπει σωστά κυρίως την κλάση NV, μπορεί να εμφανίζει υψηλή συνολική επίδοση, ακόμη και αν αποδίδει μέτρια ή χαμηλά στις μικρότερες κλάσεις. Αυτό έχει ιδιαίτερη σημασία για την κλάση MEL, επειδή το μελάνωμα είναι κλινικά κρίσιμο. Ένα μοντέλο που χάνει αρκετά μελανώματα δεν μπορεί να θεωρηθεί αξιόπιστο, ακόμη και αν η συνολική του accuracy είναι υψηλή.

Για αυτό, στις εργασίες που βασίζονται στο HAM10000 είναι χρήσιμο να εξετάζονται μετρικές όπως macro-F1, weighted-F1, recall ανά κλάση, AUC και confusion matrix. Το macro-F1 είναι ιδιαίτερα χρήσιμο, επειδή δίνει ίσο βάρος σε όλες τις κλάσεις και δεν αφήνει την πλειοψηφική κλάση να καθορίσει υπερβολικά τη συνολική εικόνα. Η confusion matrix μπορεί επίσης να δείξει ποιες κλάσεις συγχέονται μεταξύ τους, για παράδειγμα αν το μοντέλο μπερδεύει MEL με BKL ή NV.

2.2.5 PH2

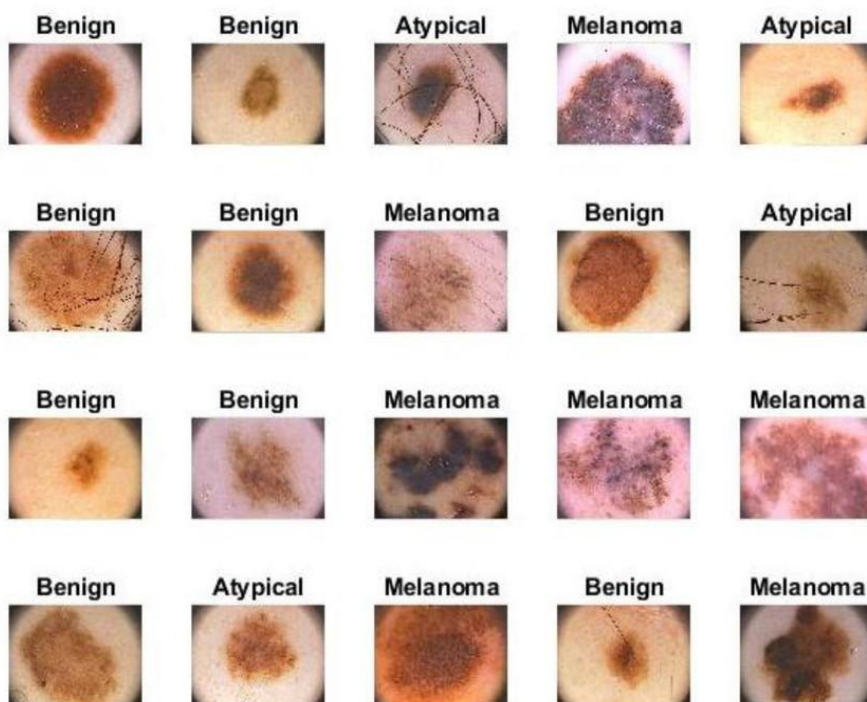
Το PH2 είναι μικρότερο dataset σε σχέση με τα ISIC και HAM10000, αλλά έχει σημαντική αξία στη βιβλιογραφία επειδή συνοδεύεται από χειροκίνητες μάσκες και καθαρές

διαγνωστικές επισημάνσεις. Περιλαμβάνει 200 δερματοσκοπικές εικόνες μελανοκυτταρικών βλαβών: 80 common nevi, 80 atypical nevi και 40 melanomas. Η δομή του PH2 το διαφοροποιεί από το HAM10000 και το ISIC 2019. Δεν έχει επτά ή οκτώ κλάσεις, αλλά τρεις βασικές κατηγορίες μελανοκυτταρικών βλαβών. Αυτό κάνει το πρόβλημα πιο περιορισμένο ως προς την ποικιλία των διαγνώσεων, αλλά χρήσιμο για την εξέταση της διάκρισης ανάμεσα σε κοινούς σπίλους, άτυπους σπίλους και μελανώματα. Η ύπαρξη των 40 μελανωμάτων το καθιστά κατάλληλο για μικρά πειράματα που εστιάζουν στη διαφοροποίηση καλοήθων και κακοήθων βλαβών. Το PH2 δεν είναι πλήρως ισορροπημένο, επειδή οι κατηγορίες common nevi και atypical nevi έχουν από 80 εικόνες, ενώ η κλάση melanoma έχει 40 εικόνες.



Εικόνα 2.8. Κατανομή των τριών κατηγοριών στο PH2 dataset

Παρ' όλα αυτά, η ανισορροπία του είναι σαφώς μικρότερη από εκείνη που συναντάται στο HAM10000 ή στο ISIC 2019. Το βασικό του μειονέκτημα είναι το μικρό μέγεθος. Με μόνο 200 εικόνες, δεν είναι κατάλληλο για εκπαίδευση μεγάλων deep learning μοντέλων από την αρχή. Αν ένα μοντέλο εκπαιδευτεί αποκλειστικά πάνω στο PH2, υπάρχει μεγάλος κίνδυνος overfitting.



Εικόνα 2.9. Δείγματα δερματικών παθήσεων από το σύνολο δεδομένων PH2. Πηγή: Abunadi, I., & Senan, E. M. (2021)

Για αυτόν τον λόγο, στη βιβλιογραφία το PH2 εμφανίζεται κυρίως ως συμπληρωματικό dataset, ως dataset εξωτερικού ελέγχου ή ως μικρό σύνολο για fine-tuning. Μπορεί να αξιοποιηθεί για να εξεταστεί αν ένα μοντέλο που εκπαιδεύτηκε σε μεγαλύτερο dataset, όπως το ISIC ή το HAM10000, μπορεί να διατηρήσει ικανοποιητική επίδοση σε διαφορετική πηγή δεδομένων. Η παρουσία χειροκίνητων масκών το κάνει επίσης χρήσιμο για segmentation, καθώς επιτρέπει υπολογισμό Dice και IoU. Στον πίνακα της βιβλιογραφίας, το PH2 εμφανίζεται σε εργασίες όπως MICA, DeepLabv3+ and Vision Transformer, FAT-Net και ensemble transfer learning μοντέλα. Αυτό δείχνει ότι η αξία του δεν βρίσκεται τόσο στο μέγεθος, αλλά στην καθαρή δομή του και στην ύπαρξη επισημάνσεων που μπορούν να υποστηρίξουν τόσο classification όσο και segmentation.

2.2.6 Derm7pt

Το Derm7pt είναι ένα dataset με διαφορετική λογική από τα καθαρά classification datasets. Είναι οργανωμένο γύρω από το seven-point checklist, ένα σύστημα δερματοσκοπικής αξιολόγησης που βασίζεται σε συγκεκριμένα οπτικά κριτήρια για την εκτίμηση της πιθανότητας μελανώματος. Περιλαμβάνει 1.011 περιπτώσεις βλαβών, με δερματοσκοπικές και κλινικές εικόνες, καθώς και δομημένη πληροφορία που σχετίζεται με τα κριτήρια του checklist και την τελική διάγνωση [43]. Η βασική αξία του Derm7pt είναι ότι δεν περιορίζεται στην τελική διαγνωστική ετικέτα. Περιλαμβάνει πληροφορίες που συνδέονται με ενδιαμέσως κλινικά χαρακτηριστικά. Αυτό το κάνει ιδιαίτερα χρήσιμο για multimodal και multi-task μοντέλα, δηλαδή για μοντέλα που δεν επεξεργάζονται μόνο την εικόνα, αλλά και πρόσθετα στοιχεία, όπως metadata ή κλινικές πληροφορίες. Σε αντίθεση με ένα απλό

classification dataset, το Derm7pt επιτρέπει την εξέταση του πώς επιμέρους χαρακτηριστικά της βλάβης συνδέονται με την τελική διάγνωση.



Εικόνα 2.10. Ενδεικτικά παραδείγματα εννοιών / χαρακτηριστικών δερματικών βλαβών από το σύνολο δεδομένων Derm7pt. Πηγή: Lucieri et al., On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors, 2020.

Στη βιβλιογραφία, το Derm7pt εμφανίζεται σε εργασίες όπως το MICA και το TFormer. Στο TFormer, για παράδειγμα, το ενδιαφέρον δεν περιορίζεται στην απλή επεξεργασία δερματοσκοπικών εικόνων, αλλά επεκτείνεται στον συνδυασμό εικόνας και metadata. Αυτό είναι πιο κοντά στον τρόπο με τον οποίο λειτουργεί η πραγματική κλινική σκέψη, όπου η διάγνωση δεν βασίζεται αποκλειστικά σε μία εικόνα, αλλά και σε πρόσθετες πληροφορίες για τον ασθενή και τη βλάβη. Ως προς την ανισορροπία, το Derm7pt χρειάζεται διαφορετική ανάγνωση. Δεν έχει μόνο μία απλή κατανομή κλάσεων, όπως το HAM10000. Η ανισορροπία εξαρτάται από το task. Αν το πρόβλημα οριστεί ως τελική ταξινόμηση διάγνωσης, τότε εξετάζεται η κατανομή των διαγνωστικών κατηγοριών. Αν οριστεί ως πρόβλεψη των κριτηρίων του seven-point checklist, τότε κάθε κριτήριο μπορεί να έχει τη δική του συχνότητα εμφάνισης. Αυτό σημαίνει ότι η αξιολόγηση πρέπει να προσαρμόζεται στον τρόπο με τον οποίο αξιοποιείται το dataset. Το Derm7pt έχει ιδιαίτερη σχέση με την εξηγησιμότητα. Επειδή περιλαμβάνει κλινικά νοηματοδοτημένα χαρακτηριστικά, μπορεί να βοηθήσει στην ανάπτυξη μοντέλων που δεν δίνουν μόνο μια τελική πρόβλεψη, αλλά συνδέουν την πρόβλεψη με στοιχεία που έχουν διαγνωστικό ενδιαφέρον. Αυτό το κάνει σημαντικό για εργασίες που προσπαθούν να γεφυρώσουν την υψηλή επίδοση με την ερμηνευσιμότητα.

2.2.7 SkinCon

Το SkinCon εμφανίζεται στη βιβλιογραφία κυρίως μέσα από εργασίες που δεν εστιάζουν μόνο στην τελική ταξινόμηση, αλλά και στη σύνδεση εικόνων με εννοιολογικά ή οπτικά χαρακτηριστικά. Στον πίνακα που μελετήθηκε, εμφανίζεται στην εργασία MICA (Multi-Level Image Concept Alignment) μαζί με τα Derm7pt και PH2. Η συγκεκριμένη χρήση δείχνει ότι το SkinCon έχει αξία σε προσεγγίσεις όπου το μοντέλο δεν καλείται απλώς να προβλέψει μια διαγνωστική κατηγορία, αλλά να ευθυγραμμίσει οπτικά μοτίβα με πιο κατανοητές έννοιες.

Η λογική του SkinCon είναι χρήσιμη για την εξηγησιμότητα. Σε πολλά deep learning μοντέλα, η τελική πρόβλεψη προκύπτει από χαρακτηριστικά που δεν είναι εύκολα κατανοητά από τον άνθρωπο. Ένα concept-based dataset επιτρέπει την εξέταση του αν το μοντέλο αναγνωρίζει χαρακτηριστικά που έχουν κάποια νοηματική σχέση με την εικόνα. Για παράδειγμα, αντί το μοντέλο να βασίζεται σε αφηρημένα latent features, μπορεί να αξιολογηθεί ως προς το αν συνδέει συγκεκριμένες οπτικές περιοχές με χαρακτηριστικά που μπορούν να ερμηνευθούν. Σε σχέση με τα ISIC και HAM10000, το SkinCon δεν εμφανίζεται ως κύριο benchmark ταξινόμησης μελανώματος. Έχει πιο ειδικό ρόλο. Είναι χρήσιμο σε εργασίες που θέλουν να εξετάσουν την concept-level ερμηνευσιμότητα και τη σχέση ανάμεσα σε οπτικά χαρακτηριστικά και απόφαση μοντέλου. Για αυτό, δεν χρειάζεται να αναλυθεί με τον ίδιο τρόπο όπως το HAM10000 ως προς τις κλάσεις και την κατανομή εικόνων. Είναι πιο σωστό να παρουσιαστεί ως συμπληρωματικό dataset για explainable και concept-based learning.



Εικόνα 2.11. Παραδείγματα κλινικών εννοιών από το SkinCon dataset. Η εικόνα παρουσιάζει δείγματα για τα concepts plaque, pustule, telangiectasia και ulcer, δείχνοντας ότι το SkinCon δεν βασίζεται μόνο σε τελικές διαγνώσεις, αλλά σε οπτικά/κλινικά χαρακτηριστικά που μπορούν να αξιοποιηθούν για εξηγησιμότητα και ανάλυση σφαλμάτων μοντέλων. Πηγή: Daneshjou et al.

2.2.8 Vitiligo dataset

Το Vitiligo dataset εμφανίζεται στον πίνακα μέσα από την εργασία Two-stage CNN and GradCAM++ for melanoma, όπου η προτεινόμενη μέθοδος αξιολογείται όχι μόνο σε ISIC 2016/2017/2018, αλλά και σε εικόνες λεύκης. Η λεύκη δεν ανήκει στο ίδιο ακριβώς διαγνωστικό πρόβλημα με το μελάνωμα, καθώς αφορά διαταραχή αποχρωματισμού του δέρματος και όχι κακοήγη δερματική βλάβη. Παρ' όλα αυτά, η παρουσία του dataset είναι χρήσιμη, επειδή δείχνει ότι ορισμένες αρχιτεκτονικές δερματολογικής εικόνας μπορούν να δοκιμαστούν και σε διαφορετικές παθήσεις. Η χρήση του Vitiligo dataset στην εργασία αυτή λειτουργεί περισσότερο ως συμπληρωματικός έλεγχος γενίκευσης. Το μοντέλο εξετάζεται σε άλλο είδος δερματολογικής εικόνας, ώστε να φανεί αν η μεθοδολογία μπορεί να εφαρμοστεί και πέρα από το κλασικό πρόβλημα melanoma classification. Τα αποτελέσματα

που αναφέρονται είναι υψηλά, με AUC 0,986 και accuracy 0,991 για το Vitiligo dataset, αλλά δεν πρέπει να συγκρίνονται άμεσα με τα αποτελέσματα σε ISIC ή HAM10000.

2.2.9 Monkeypox dataset

Το Monkeypox dataset εμφανίζεται στον πίνακα μέσα από την εργασία PriFL-XAI, η οποία εξετάζει ένα privacy-preserving federated learning σχήμα με StyleGAN2-ADA augmentation και explainability. Το dataset αφορά ταξινόμηση δερματικών εκδηλώσεων που σχετίζονται με monkeypox και όχι κλασική δερματοσκοπική ταξινόμηση μελανώματος. Στην εργασία αναφέρεται ως 4-class dataset, με καλύτερη επίδοση accuracy 98,2 για τη διαμόρφωση DenseNet121-ViT-B32.

Η παρουσία αυτού του dataset στον πίνακα είναι σημαντική για δύο λόγους. Πρώτον, δείχνει ότι οι τεχνικές που αναπτύσσονται για δερματολογικές εικόνες μπορούν να μεταφερθούν και σε άλλου τύπου δερματικές παθήσεις. Δεύτερον, η συγκεκριμένη εργασία συνδέει το dataset με ζητήματα federated learning, privacy preservation, synthetic data augmentation και XAI. Αυτά τα στοιχεία έχουν ευρύτερη σημασία για την ιατρική τεχνητή νοημοσύνη, επειδή τα ιατρικά δεδομένα συχνά δεν μπορούν να συγκεντρωθούν εύκολα σε ένα κεντρικό σημείο λόγω προσωπικών και θεσμικών περιορισμών. Παρά τη χρησιμότητά του, το Monkeypox dataset δεν πρέπει να τοποθετηθεί στην ίδια κατηγορία με το ISIC, το HAM10000 ή το PH2. Δεν είναι dataset δερματοσκοπικής ανάλυσης μελανώματος, αλλά dataset δερματολογικής εικόνας για άλλη πάθηση.

Παρακάτω ακολουθεί ο συγκριτικός πίνακας των datasets που υπήρχαν στις εργασίες:

Dataset	Μέγεθος / αριθμός εικόνων	Κλάσεις / κατηγορίες	Είδος δεδομένων	Κύρια χρήση στη βιβλιογραφία	Ζήτημα imbalance / περιορισμοί
ISIC 2016–2018	Διαφέρει ανά έκδοση. Στο ISIC 2018 segmentation task υπάρχουν 2.594 εικόνες εκπαίδευσης με masks. Στο ISIC 2018 classification task υπάρχουν 10.015 εικόνες.	Στο ISIC 2018 classification: 7 κλάσεις: NV, MEL, BCC, AKIEC, BKL, DF, VASC.	Δερματοσκοπικές εικόνες, masks για segmentation σε ορισμένα tasks, labels για classification.	Segmentation και classification. Χρησιμοποιείται συχνά σε εργασίες όπως FAT-Net, FTN, SapFormer και DeepLabv3+ / ViT.	Υπάρχει ανισορροπία στις 7 κλάσεις, κυρίως λόγω κυριαρχίας της κλάσης NV. Η accuracy χρειάζεται να συνοδεύεται από F1, recall και AUC.
ISIC 2019	25.331 εικόνες εκπαίδευσης.	8 βασικές κλάσεις στο training set: MEL, NV, BCC, AK, BKL, DF, VASC, SCC.	Δερματοσκοπικές εικόνες με labels ταξινόμησης.	Μεγάλης κλίμακας multiclass classification. Συχνό σε ViT, Swin Transformer, ConvNeXt και υβριδικές CNN–Transformer προσεγγίσεις.	Έντονο class imbalance. Οι NV και MEL είναι πολυπληθέστερες, ενώ DF, VASC, AK και SCC είναι μικρότερες. Χρειάζονται macro-F1, balanced accuracy και per-class metrics.
ISIC 2020 / SIIM-ISIC	33.126 training εικόνες.	Δυαδική ταξινόμηση: melanoma / non-melanoma.	Δερματοσκοπικές εικόνες, labels και patient identifiers.	Binary melanoma classification, pretraining/finetuning, patient-wise split αξιολόγηση.	Πολύ έντονο imbalance, επειδή οι non-melanoma εικόνες είναι πολύ περισσότερες. Η accuracy μπορεί να είναι παραπλανητική· η AUC και το recall για melanoma είναι πιο χρήσιμα.
HAM10000	10.015 δερματοσκοπικές εικόνες.	7 κλάσεις: NV, MEL, BKL, BCC, AKIEC, VASC, DF.	Δερματοσκοπικές εικόνες με labels ταξινόμησης.	Βασικό benchmark για 7-class classification, transfer learning, ViT/Swin, ConvNeXt, XAI και active learning.	Πολύ έντονο imbalance. Η κλάση NV έχει περίπου 6.705 εικόνες, ενώ DF και VASC έχουν πολύ λίγες. Η συνολική accuracy δεν αρκεί.

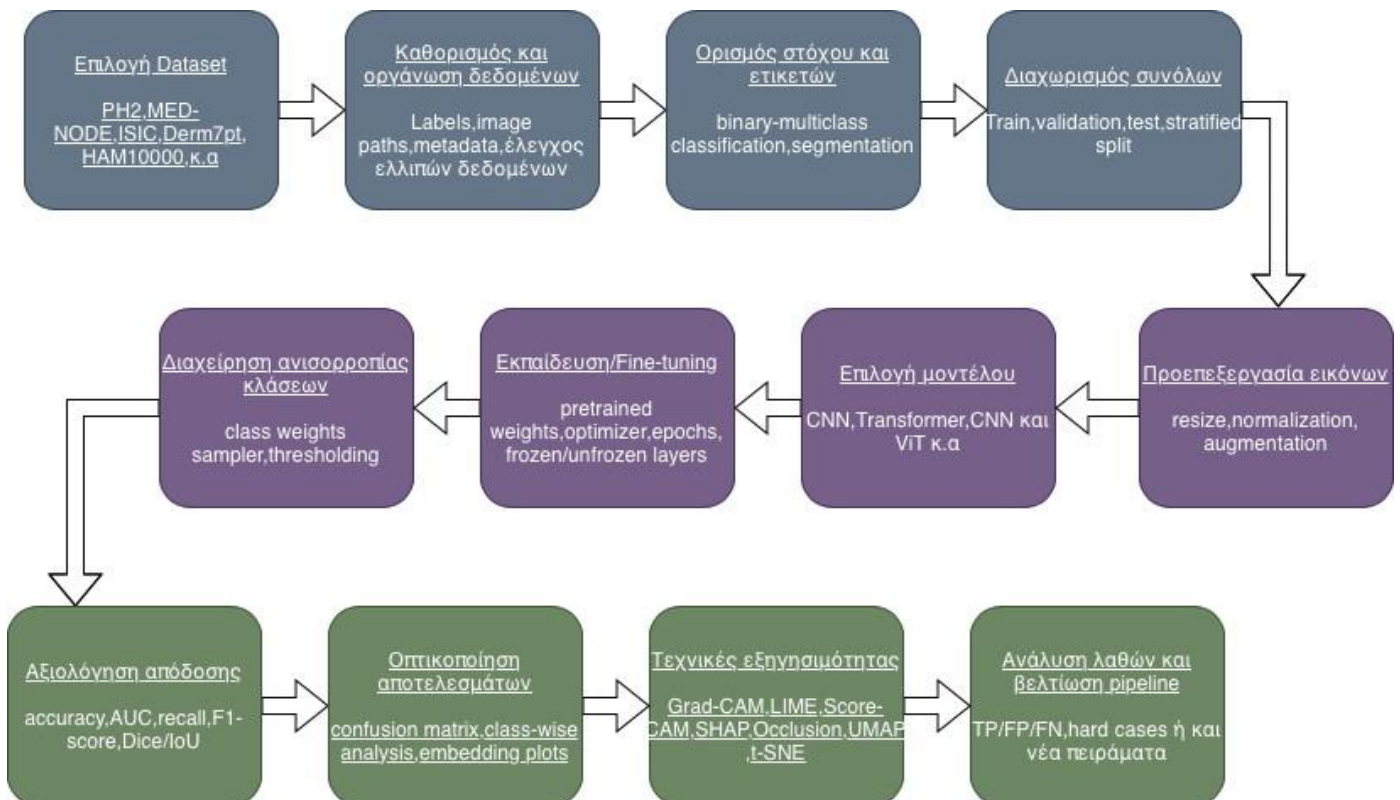
Dataset	Μέγεθος / αριθμός εικόνων	Κλάσεις / κατηγορίες	Είδος δεδομένων	Κύρια χρήση στη βιβλιογραφία	Ζήτημα imbalance / περιορισμοί
PH2	200 εικόνες.	3 κατηγορίες: 80 common nevi, 80 atypical nevi, 40 melanomas.	Δερματοσκοπικές εικόνες, χειροκίνητες masks και διαγνωστικές επισημάνσεις.	Segmentation, μικρό fine-tuning, εξωτερικός έλεγχος γενίκευσης, έλεγχος Dice/IoU.	Μικρό dataset με κίνδυνο overfitting. Δεν είναι πλήρως ισορροπημένο, καθώς τα melanomas είναι λιγότερα.
Derm7pt	1.011 lesion cases.	Τελική διάγνωση και ετικέτες που σχετίζονται με το seven-point checklist.	Δερματοσκοπικές εικόνες, κλινικές εικόνες και metadata.	Multimodal classification, multi-task learning, συνδυασμός εικόνων και δομημένης πληροφορίας.	Το imbalance εξαρτάται από το task. Διαφορετική κατανομή υπάρχει αν εξετάζεται η τελική διάγνωση ή τα επιμέρους κριτήρια του checklist.
SkinCon	Δεν λειτουργεί ως κλασικό benchmark ταξινόμησης μελανώματος στον πίνακα.	Εννοιολογικά / οπτικά χαρακτηριστικά.	Δερματολογικές εικόνες συνδεδεμένες με visual concepts.	Concept-based learning και explainability. Εμφανίζεται στο MICA μαζί με Derm7pt και PH2.	Δεν είναι βασικό dataset για κλασικό melanoma classification. Χρησιμοποιείται περισσότερο για concept alignment και ερμηνευσιμότητα.
Vitiligo dataset	Δεν δίνεται ακριβές μέγεθος στον πίνακα.	Εικόνες σχετικές με λεύκη.	Δερματολογικές εικόνες άλλης πάθησης.	Συμπληρωματικός έλεγχος μεθόδων δερματολογικής εικόνας. Εμφανίζεται στο Two-stage CNN and GradCAM++.	Δεν αφορά άμεσα μελάνωμα. Τα αποτελέσματά του δεν συγκρίνονται άμεσα με ISIC ή HAM10000.
Monkeypox dataset	Δεν δίνεται ακριβές μέγεθος στον πίνακα.	4 κλάσεις.	Δερματολογικές εικόνες σχετικές με monkeypox.	Federated learning, privacy-preserving classification, StyleGAN2-ADA augmentation και XAI.	Δεν είναι dataset δερματοσκοπικής ταξινόμησης μελανώματος. Χρήσιμο κυρίως ως παράδειγμα μεταφοράς τεχνικών σε άλλο δερματολογικό πρόβλημα.

Πίνακας 2.2. Συγκριτικός πίνακας των datasets των εργασιών

3. Πειραματική μεθοδολογία και τεχνικές που χρησιμοποιήθηκαν

Στο παρόν κεφάλαιο παρουσιάζεται η πειραματική μεθοδολογία που ακολουθήθηκε για την αναγνώριση, την κατηγοριοποίηση και, όπου απαιτήθηκε, την τμηματοποίηση δερματικών βλαβών. Η μεθοδολογία οργανώθηκε με τέτοιο τρόπο ώστε κάθε πείραμα να έχει κοινή λογική: επιλογή dataset, καθορισμό στόχου, προεπεξεργασία εικόνων, διαχωρισμό δεδομένων, εκπαίδευση ή fine-tuning μοντέλου, αξιολόγηση με κατάλληλες μετρικές και έλεγχο της συμπεριφοράς του μοντέλου μέσω τεχνικών εξηγησιμότητας. Η χρήση του Google Colab βοήθησε στην εκτέλεση των πειραμάτων σε GPU περιβάλλον, ενώ τα notebooks λειτούργησαν ως χώρος πειραματισμού, καταγραφής αποτελεσμάτων και παραγωγής διαγραμμάτων.

Η συνολική ροή φαίνεται στο Σχήμα 3.1. Το σχήμα δεν παρουσιάζει ένα μεμονωμένο πείραμα, αλλά τη γενική λογική που ακολουθήθηκε σε όλα τα πειράματα της εργασίας. Ξεκινά από τα δεδομένα και καταλήγει στην ποσοτική και ποιοτική αξιολόγηση. Με αυτόν τον τρόπο γίνεται σαφές ότι η τελική πρόβλεψη ενός μοντέλου δεν εξετάστηκε μόνη της, αλλά σε σχέση με τις μετρικές, τα λάθη και τις εξηγήσεις που παράγει.



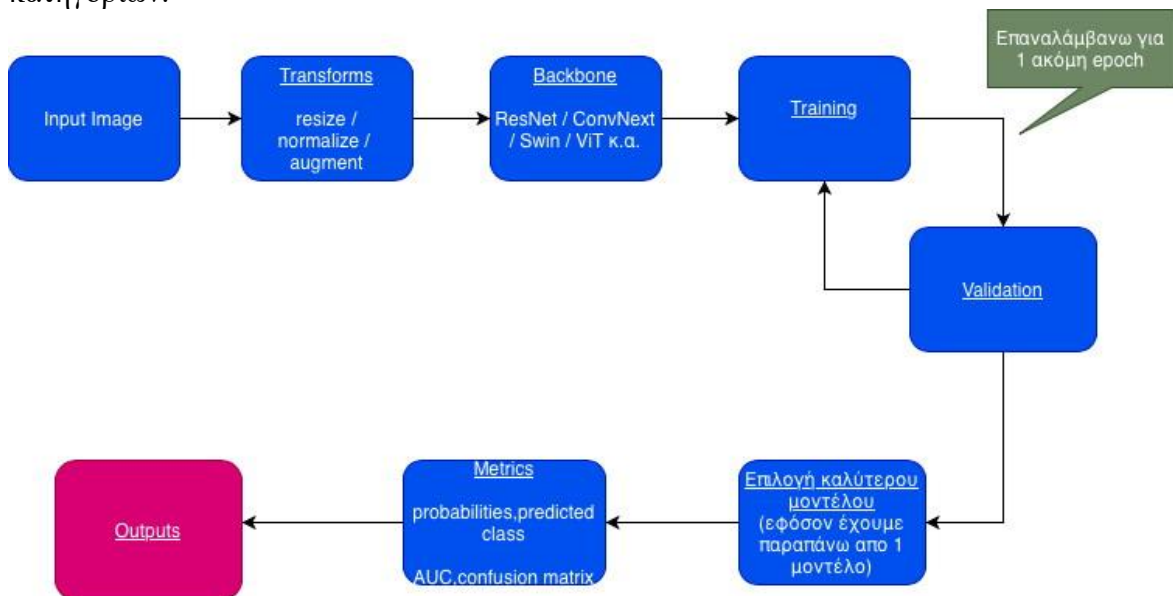
Σχήμα 3.1. Γενικό workflow της πειραματικής διαδικασίας

Όπως φαίνεται στο Σχήμα 3.1, η μεθοδολογία ακολουθεί διαδοχικά βήματα. Πρώτα ορίζεται το πρόβλημα, για παράδειγμα δυαδική ταξινόμηση melanoma / non-melanoma, πολυκατηγορική ταξινόμηση ή segmentation. Μετά επιλέγεται το κατάλληλο dataset και

εφαρμόζονται οι αναγκαίες διαδικασίες προεπεξεργασίας, όπως resize, normalization, augmentation και διαχωρισμός σε train, validation και test set. Στο τελευταίο στάδιο δεν κρατήθηκαν μόνο οι αριθμητικές μετρικές, αλλά εξετάστηκαν και οι οπτικές εξηγήσεις, ώστε να φανεί αν το μοντέλο εστιάζει στη βλάβη ή σε άσχετα στοιχεία της εικόνας.

3.1 Fine-tuning σε μοντέλα ταξινόμησης (CNN & Transformer backbones)

Ο βασικός μηχανισμός που χρησιμοποιήθηκε στα πειράματα ταξινόμησης ήταν το fine-tuning προεκπαιδευμένων μοντέλων. Η επιλογή αυτή κρίθηκε κατάλληλη επειδή τα ιατρικά datasets, ακόμη και όταν είναι σχετικά μεγάλα, παραμένουν μικρότερα και πιο εξειδικευμένα σε σχέση με γενικά datasets όπως το ImageNet. Για τον λόγο αυτό, η εκπαίδευση από το μηδέν θα αύξανε τον κίνδυνο overfitting και θα απαιτούσε πολύ περισσότερα δεδομένα. Αντίθετα, τα προεκπαιδευμένα backbones προσφέρουν αρχικές οπτικές αναπαραστάσεις, οι οποίες μπορούν να προσαρμοστούν στις δερματοσκοπικές εικόνες μέσα από ελεγχόμενο fine-tuning [8]. Στα πειράματα χρησιμοποιήθηκαν CNN αρχιτεκτονικές, όπως ResNet50, DenseNet121, EfficientNetV2-S και ConvNeXt-Tiny, αλλά και transformer-based backbones, όπως Swin Transformer και ViT. Σε κάθε περίπτωση, το τελικό classification head αντικαταστάθηκε ώστε να αντιστοιχεί στον αριθμό των κλάσεων του εκάστοτε πειράματος. Στα δυαδικά πειράματα η έξοδος αντιστοιχούσε συνήθως στη διάκριση melanoma / non-melanoma ή benign / suspicious, ενώ στα πολυκατηγορικά πειράματα η έξοδος περιλάμβανε περισσότερες διαγνωστικές κατηγορίες. Το Σχήμα 3.2 συνοψίζει τη ροή του fine-tuning. Η δερματοσκοπική εικόνα περνά πρώτα από προεπεξεργασία, στη συνέχεια εισάγεται στο backbone και στο τέλος το νέο classification head παράγει πιθανότητες για τις διαθέσιμες κλάσεις. Η λογική αυτή εφαρμόστηκε με μικρές προσαρμογές ανάλογα με το dataset και τον αριθμό των κατηγοριών.

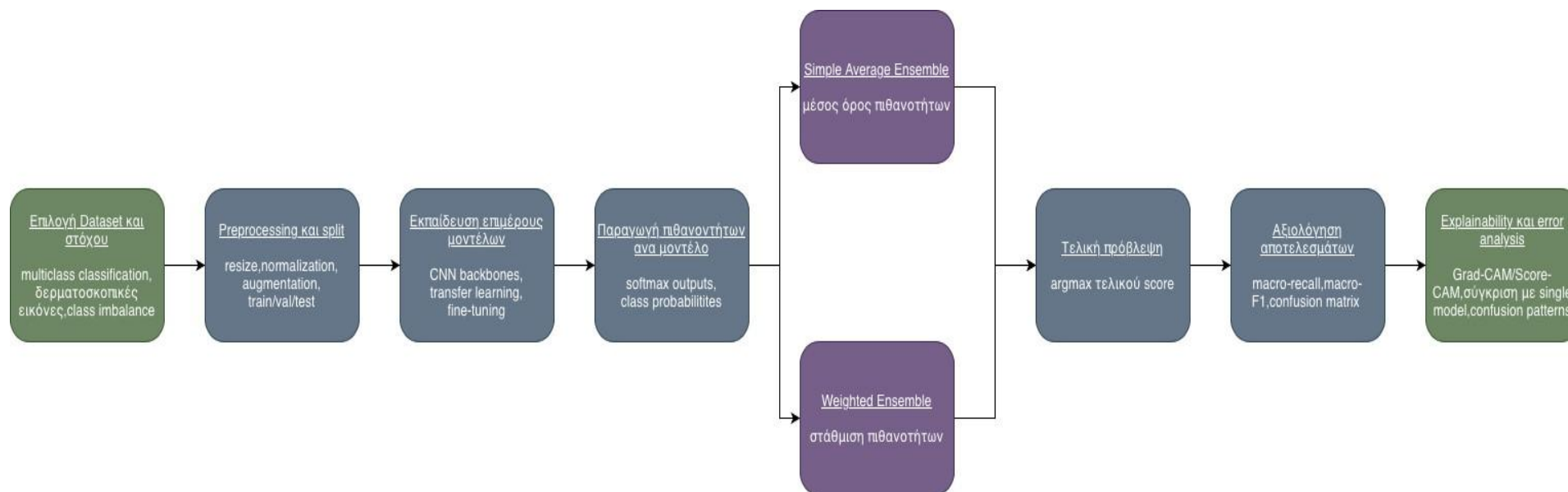


Σχήμα 3.2. Fine-tuning ταξινόμησης (CNN/Transformer backbone)

Στην πράξη, εφαρμόστηκε σταδιακή εκπαίδευση. Αρχικά εκπαιδεύονταν κυρίως το νέο classification head, ώστε το μοντέλο να προσαρμοστεί στις κλάσεις του dataset χωρίς να μεταβληθούν απότομα τα προεκπαιδευμένα χαρακτηριστικά. Σε επόμενο στάδιο γινόταν unfreezing μέρους ή ολόκληρου του backbone, ώστε το μοντέλο να μάθει πιο ειδικά χαρακτηριστικά της δερματοσκοπικής εικόνας. Η στρατηγική αυτή βοήθησε ιδιαίτερα σε datasets με περιορισμένο μέγεθος, όπως το PH2, αλλά και σε ανισόρροπα datasets, όπως το HAM10000 και το ISIC 2020. Ιδιαίτερη προσοχή δόθηκε στο class imbalance. Σε datasets τύπου ISIC και HAM10000, οι συχνότερες κλάσεις, όπως οι σπίλοι, εμφανίζονται σε πολύ μεγαλύτερο αριθμό σε σχέση με κλινικά κρίσιμες κατηγορίες, όπως το melanoma [41], [44], [45]. Για αυτόν τον λόγο η αξιολόγηση δεν βασίστηκε μόνο στην accuracy. Χρησιμοποιήθηκαν μετρικές όπως macro-F1, macro-recall, weighted-F1 και AUC, επειδή δίνουν πιο δίκαιη εικόνα της συμπεριφοράς του μοντέλου στις μικρότερες κλάσεις. Σε ορισμένα δυαδικά πειράματα εφαρμόστηκε και threshold tuning, ώστε η τελική απόφαση να μην εξαρτάται μηχανικά από το όριο 0.5, αλλά να προσαρμόζεται καλύτερα στη σχέση precision και recall.

3.2 Ensemble learning (simple average & weighted ensemble)

Στην ενότητα του ensemble learning η ταξινόμηση προσεγγίστηκε με τη λογική ότι ένα μόνο μοντέλο δεν είναι πάντα αρκετό για να καλύψει όλη την οπτική ποικιλία των δερματοσκοπικών εικόνων. Διαφορετικές αρχιτεκτονικές μπορούν να μαθαίνουν διαφορετικά στοιχεία: ένα CNN μπορεί να είναι πιο ευαίσθητο σε τοπικές υφές και λεπτομέρειες, ενώ ένα transformer-based μοντέλο μπορεί να αποτυπώνει καλύτερα πιο μακρινές σχέσεις μέσα στην εικόνα. Ο συνδυασμός τους μπορεί να μειώσει την εξάρτηση από ένα μεμονωμένο failure mode και να προσφέρει πιο σταθερή συνολική πρόβλεψη [37]. Η βασική υλοποίηση έγινε σε επίπεδο πιθανοτήτων, δηλαδή με late fusion των outputs. Για κάθε εικόνα, κάθε μοντέλο παρήγαγε ένα διάνυσμα πιθανοτήτων ανά κλάση. Στη συνέχεια δοκιμάστηκαν δύο στρατηγικές: simple average ensemble και weighted ensemble. Στο simple average, οι πιθανότητες όλων των μοντέλων μέσο-οροποιούνται και η τελική κλάση προκύπτει από το μεγαλύτερο μέσο score. Η μέθοδος αυτή είναι απλή και λειτουργεί ως χρήσιμο baseline, επειδή δεν απαιτεί πρόσθετη εκπαίδευση [37]. Στο weighted ensemble η ίδια λογική διατηρείται, αλλά κάθε μοντέλο παίρνει διαφορετικό βάρος. Τα βάρη μπορούν να καθοριστούν με βάση τη συμπεριφορά στο validation set, όχι μόνο ως προς την accuracy αλλά και ως προς πιο ισορροπημένες μετρικές, όπως macro-F1 ή macro-recall. Αυτό είναι σημαντικό σε ανισόρροπα datasets, γιατί ένα μοντέλο που έχει υψηλή accuracy μπορεί να ευνοεί κυρίως την πλειοψηφική κλάση, ενώ ένα άλλο μπορεί να αναγνωρίζει καλύτερα τις σπάνιες αλλά κλινικά σημαντικές κατηγορίες [37]. Το Σχήμα 3.3 δείχνει τη λογική του ensemble. Η ίδια εικόνα περνά από περισσότερα από ένα μοντέλα, παράγονται ξεχωριστές πιθανότητες και στη συνέχεια οι πιθανότητες συνδυάζονται για να εξαχθεί η τελική απόφαση. Η διαδικασία αυτή δεν αλλάζει το αρχικό dataset, αλλά αλλάζει τον τρόπο με τον οποίο λαμβάνεται η απόφαση στο τελικό στάδιο.



Σχήμα 3.3. Ensemble learning(simple average & weighted ensemble)

Το ensemble αξιολογήθηκε όχι μόνο από το τελικό score, αλλά και από τη δομή των λαθών. Μας ενδιέφερε αν ο συνδυασμός μοντέλων μειώνει συγκεκριμένες συγχύσεις μεταξύ κλάσεων, για παράδειγμα αν περιορίζει λανθασμένες ταξινομήσεις melanoma ως benign ή αν βελτιώνει την ανάκληση μικρότερων κατηγοριών. Η ανάλυση αυτή είναι πιο χρήσιμη από μια απλή σύγκριση accuracy, επειδή δείχνει αν το ensemble διορθώνει πραγματικές αδυναμίες των επιμέρους μοντέλων.

Η εξηγησιμότητα συνδέθηκε και με το ensemble πλαίσιο. Σε επιλεγμένα δείγματα συγκρίθηκαν heatmaps από διαφορετικά μοντέλα, ώστε να εξεταστεί αν όλα εστιάζουν στις ίδιες περιοχές ή αν το καθένα αξιοποιεί διαφορετικά οπτικά cues [12], [13]. Όταν δύο μοντέλα οδηγούνται στην ίδια πρόβλεψη αλλά εστιάζουν σε διαφορετικές περιοχές, η σύγκριση των εξηγήσεων βοηθά να φανεί αν το ensemble κερδίζει από ουσιαστική συμπληρωματικότητα ή αν αναπαράγει το ίδιο bias [12], [13].

3.3 Εξηγησιμότητα που χρησιμοποιήθηκε στα πειράματα (Grad-CAM, Grad-CAM++, Score-CAM, LIME, SHAP, Occlusion, UMAP και t-SNE)

Στην παρούσα εργασία, η εξηγησιμότητα δεν αντιμετωπίστηκε ως απλή προσθήκη μετά την εκπαίδευση, αλλά ως μέρος της πειραματικής αξιολόγησης. Η βασική λογική ήταν ότι ένα μοντέλο μπορεί να εμφανίζει καλές αριθμητικές μετρικές, όπως accuracy, F1-score ή AUC, χωρίς αυτό να σημαίνει απαραίτητα ότι βασίζεται στα σωστά οπτικά χαρακτηριστικά. Στην ανάλυση δερματοσκοπικών εικόνων αυτό είναι κρίσιμο, επειδή το μοντέλο μπορεί να επηρεαστεί από borders, φωτισμό, τρίχες, σημάδια χάραξης, artifacts ή στοιχεία του background, αντί να εστιάζει στην ίδια τη δερματική βλάβη [15], [48].

Τα εργαλεία XAI χρησιμοποιήθηκαν ως μηχανισμός ελέγχου της συμπεριφοράς των μοντέλων. Δεν αρκούσε να γνωρίζουμε αν μια πρόβλεψη ήταν σωστή ή λανθασμένη. Χρειάζόταν να εξετάσουμε πού εστίασε το μοντέλο και αν η απόφασή του στηρίχθηκε σε περιοχές με πραγματική διαγνωστική σημασία. Με αυτόν τον τρόπο, η εξηγησιμότητα λειτούργησε ως γέφυρα ανάμεσα στις ποσοτικές μετρικές και στην ποιοτική ερμηνεία των αποτελεσμάτων. Στα πειράματα ταξινόμησης με CNN backbones εφαρμόστηκε κυρίως το Grad-CAM. Η μέθοδος αυτή παράγει χάρτη ενεργοποίησης που δείχνει σε ποια χωρικά σημεία των τελευταίων συνελκτικών feature maps βασίστηκε περισσότερο το μοντέλο για μια συγκεκριμένη κλάση [15]. Χρησιμοποιήθηκε σε σωστές και λανθασμένες προβλέψεις, δηλαδή σε true positives, false positives και false negatives. Η σύγκριση αυτών των περιπτώσεων βοήθησε να φανεί αν το μοντέλο εστίαζε στη βλάβη ή αν επηρεαζόταν από περιφερειακά και μη διαγνωστικά στοιχεία της εικόνας.

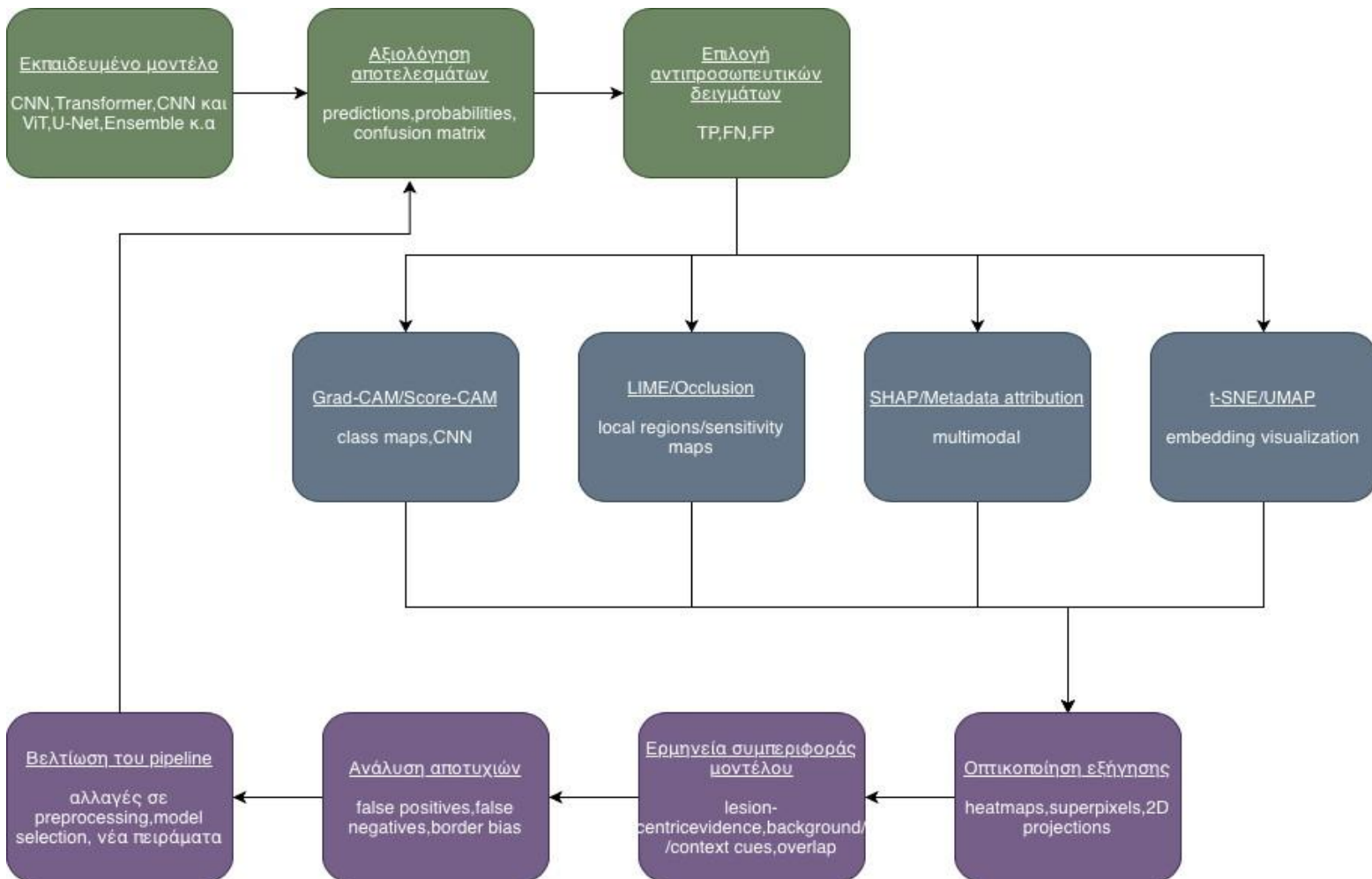
Μαζί με το Grad-CAM χρησιμοποιήθηκε και το Grad-CAM++, κυρίως όταν θέλαμε πιο λεπτομερή εντοπισμό των περιοχών που επηρεάζουν την πρόβλεψη. Το Grad-CAM++ είναι χρήσιμο σε εικόνες όπου η βλάβη έχει ανομοιογενές σχήμα, ασύμμετρα όρια ή μικρές τοπικές αλλοιώσεις. Σε τέτοιες περιπτώσεις βοήθησε να εξεταστεί αν η ενεργοποίηση ακολουθεί τη δομή της βλάβης ή αν παραμένει γενική και ασαφής [13]. Στα across-epoch πειράματα, το Grad-CAM++ χρησιμοποιήθηκε επίσης για να φανεί πώς μεταβάλλεται η εστίαση του μοντέλου από τις πρώτες εποχές μέχρι το καλύτερο checkpoint. Σε ορισμένα πειράματα εφαρμόστηκε το Score-CAM ως συμπληρωματικό εργαλείο. Σε αντίθεση με το Grad-CAM, το Score-CAM δεν βασίζεται στα gradients, αλλά αξιολογεί τη συνεισφορά των ενεργοποιήσεων μέσα από forward passes. Είναι πιο αργό υπολογιστικά, αλλά χρήσιμο όταν θέλουμε να ελέγξουμε αν μια περιοχή παραμένει σημαντική χωρίς να εξαρτόμαστε από πιθανή αστάθεια των gradients [14]. Στην εργασία λειτούργησε ως δεύτερη ένδειξη για

το αν ένα heatmap είναι συνεπές. Για εξηγήσεις που δεν απαιτούν πρόσβαση στα εσωτερικά layers χρησιμοποιήθηκε το LIME. Με το LIME η εικόνα χωρίζεται σε superpixels και δημιουργούνται παραλλαγές της εικόνας, στις οποίες ορισμένες περιοχές αποκρύπτονται ή αλλοιώνονται. Με βάση το πώς αλλάζει η πρόβλεψη, το LIME εκτιμά ποια τμήματα συμβάλλουν θετικά ή αρνητικά στην τελική απόφαση [47]. Επειδή είναι πιο βαρύ υπολογιστικά, εφαρμόστηκε σε επιλεγμένα δείγματα ως case study, ειδικά όταν χρειαζόταν ερμηνεία σε επίπεδο ορατών περιοχών της εικόνας.

Συμπληρωματικά χρησιμοποιήθηκε και η occlusion attribution. Σε αυτή τη μέθοδο αποκρύπτονται μικρά patches της εικόνας και μετριέται η μεταβολή στο score της πρόβλεψης. Αν η απόκρυψη μιας περιοχής προκαλεί μεγάλη πτώση στην πιθανότητα της κλάσης, τότε η περιοχή θεωρείται σημαντική για την απόφαση του μοντέλου [48]. Η μέθοδος είναι απλή αλλά πρακτική, επειδή μπορεί να εφαρμοστεί ακόμη και όταν η αρχιτεκτονική δεν υποστηρίζει άμεσα κλασική CAM ανάλυση. Για την ερμηνεία επιλεγμένων προβλέψεων χρησιμοποιήθηκε και το SHAP. Η λογική του SHAP βασίζεται στην εκτίμηση της συνεισφοράς κάθε χαρακτηριστικού στην τελική πρόβλεψη, με τρόπο που προέρχεται από τη θεωρία των Shapley values. Στο πλαίσιο των εικόνων, το SHAP μπορεί να δείξει ποιες περιοχές αυξάνουν ή μειώνουν την πιθανότητα μιας συγκεκριμένης κλάσης [15]. Στα πειράματα χρησιμοποιήθηκε κυρίως σε χαρακτηριστικές περιπτώσεις, λόγω υπολογιστικού κόστους, και βοήθησε να φανεί αν οι περιοχές της βλάβης λειτουργούν υπέρ ή κατά της προβλεπόμενης κατηγορίας. Στα υβριδικά μοντέλα, όπως σχήματα CNN και ViT σε late fusion, η εξηγησιμότητα οργανώθηκε σε δύο επίπεδα. Πρώτα εφαρμόστηκαν CAM τεχνικές στο CNN branch, όπου υπάρχει καθαρή χωρική δομή των συνελκτικών χαρακτηριστικών [15]. Έπειτα, χρησιμοποιήθηκαν τεχνικές όπως LIME ή SHAP πάνω στην τελική πρόβλεψη του fusion μοντέλου, ώστε η εξήγηση να αφορά το συνολικό σύστημα και όχι μόνο ένα επιμέρους υπομοντέλο [47], [15]. Με αυτόν τον τρόπο μπορούσαμε να εξετάσουμε αν η τελική απόφαση ενισχύεται, διορθώνεται ή επηρεάζεται από διαφορετικά branches.

Πέρα από τα heatmaps, χρησιμοποιήθηκαν και τεχνικές οπτικοποίησης embeddings, όπως UMAP και t-SNE. Οι τεχνικές αυτές δεν εξηγούν άμεσα μια μεμονωμένη πρόβλεψη, αλλά δείχνουν πώς οργανώνονται τα δείγματα στον χώρο των χαρακτηριστικών. Με άλλα λόγια, βοηθούν να φανεί αν τα χαρακτηριστικά που μαθαίνει το μοντέλο δημιουργούν διαχωρίσιμες ομάδες ανάμεσα στις κλάσεις ή αν παραμένουν επικαλύψεις [49], [50]. Αυτό ήταν χρήσιμο κυρίως σε επίπεδο dataset, ειδικά σε πειράματα όπου θέλαμε να συνδέσουμε την επίδοση με τη δομή των learned representations. Το t-SNE χρησιμοποιήθηκε κυρίως για την παρατήρηση τοπικών ομάδων δειγμάτων με παρόμοια χαρακτηριστικά. Επειδή είναι ευαίσθητο στις παραμέτρους και δεν διατηρεί πάντα αξιόπιστα τις συνολικές αποστάσεις, αντιμετωπίστηκε ως εργαλείο οπτικής διερεύνησης και όχι ως απόδειξη διαχωρισμού των κλάσεων [49]. Το UMAP χρησιμοποιήθηκε με παρόμοιο στόχο, αλλά έδωσε πιο σταθερή εικόνα της συνολικής δομής των embeddings, ειδικά όταν εξετάστηκε αν οι κλάσεις απομακρύνονται σταδιακά κατά την εκπαίδευση [50]. Ένα ακόμη στοιχείο της μεθοδολογίας ήταν η across-epoch explainability. Σε αυτή την προσέγγιση δεν κρατήθηκε μόνο ένα heatmap από το τελικό μοντέλο. Για τα ίδια test samples δημιουργήθηκαν χάρτες εξηγησιμότητας σε διαφορετικά checkpoints, όπως epoch 1, epoch 4 ή 5, epoch 8, τελική epoch και best model. Η σύγκριση αυτή βοήθησε να φανεί αν η εστίαση του μοντέλου μετακινείται σταδιακά προς τη βλάβη ή αν παραμένει ασταθής, ακόμη και όταν οι μετρικές βελτιώνονται.

Το Σχήμα 3.4 συνοψίζει τη θέση της εξηγησιμότητας μέσα στη μεθοδολογία. Η πρόβλεψη του μοντέλου δεν αποτελεί το τέλος της διαδικασίας. Μετά την πρόβλεψη παράγονται heatmaps, superpixel εξηγήσεις, SHAP συνεισφορές ή embeddings, και αυτά συγκρίνονται με τη σωστή/λανθασμένη πρόβλεψη, τη θέση της βλάβης και τις μετρικές απόδοσης.



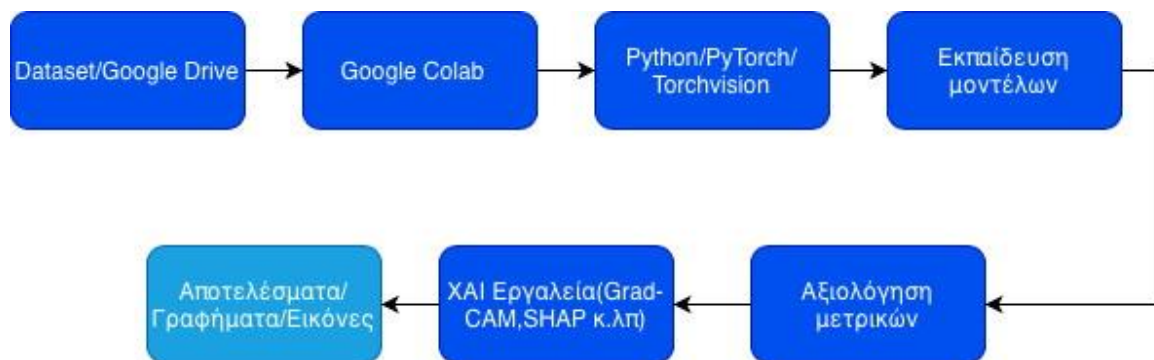
Σχήμα 3.4. Εφαρμογή τεχνικών εξηγησιμότητας μετά την πρόβλεψη του μοντέλου

Η σύνδεση της εξηγησιμότητας με τις μετρικές αποτέλεσε βασικό μέρος της αξιολόγησης. Όταν ένα μοντέλο είχε υψηλή AUC αλλά χαμηλότερο recall στο melanoma, οι χάρτες Grad-CAM, Grad-CAM++, Score-CAM, LIME, SHAP ή occlusion βοήθησαν να εξεταστεί αν το πρόβλημα σχετιζόταν με υπο-εστίαση στη βλάβη, υπερ-εστίαση στο context ή αδυναμία διάκρισης λεπτών μορφολογικών στοιχείων. Αντίστοιχα, τα UMAP και t-SNE plots βοήθησαν να φανεί αν οι εσωτερικές αναπαραστάσεις οδηγούν σε καθαρό διαχωρισμό ή αν οι κλάσεις παραμένουν μερικώς επικαλυπτόμενες. Με αυτή τη λογική, η ενότητα της εξηγησιμότητας δεν έμεινε θεωρητική. Χρησιμοποιήθηκε ως πρακτικό εργαλείο ελέγχου της αξιοπιστίας των μοντέλων και ως κριτήριο για το αν ένα πείραμα μπορούσε να θεωρηθεί

πραγματικά χρήσιμο. Η αριθμητική επίδοση ήταν απαραίτητη, αλλά όχι αρκετή από μόνη της. Η τελική αξιολόγηση βασίστηκε στον συνδυασμό μετρικών, οπτικών εξηγήσεων και ανάλυσης των embeddings, ώστε να υπάρχει πιο πλήρης εικόνα για το πώς λειτουργεί κάθε μοντέλο και πού εντοπίζονται οι αδυναμίες του [48].

3.4 Πειραματικό περιβάλλον και εργαλεία υλοποίησης

Τα πειράματα της εργασίας υλοποιήθηκαν κυρίως στο Google Colab, επειδή το περιβάλλον αυτό επιτρέπει πρόσβαση σε GPU πόρους χωρίς την ανάγκη τοπικής εγκατάστασης εξειδικευμένου εξοπλισμού. Τα δεδομένα οργανώθηκαν σε φακέλους στο Google Drive και τα notebooks χρησιμοποιήθηκαν τόσο για την εκπαίδευση των μοντέλων όσο και για την καταγραφή των αποτελεσμάτων, των γραφημάτων και των εικόνων εξηγησιμότητας. Η βασική γλώσσα υλοποίησης ήταν η Python.

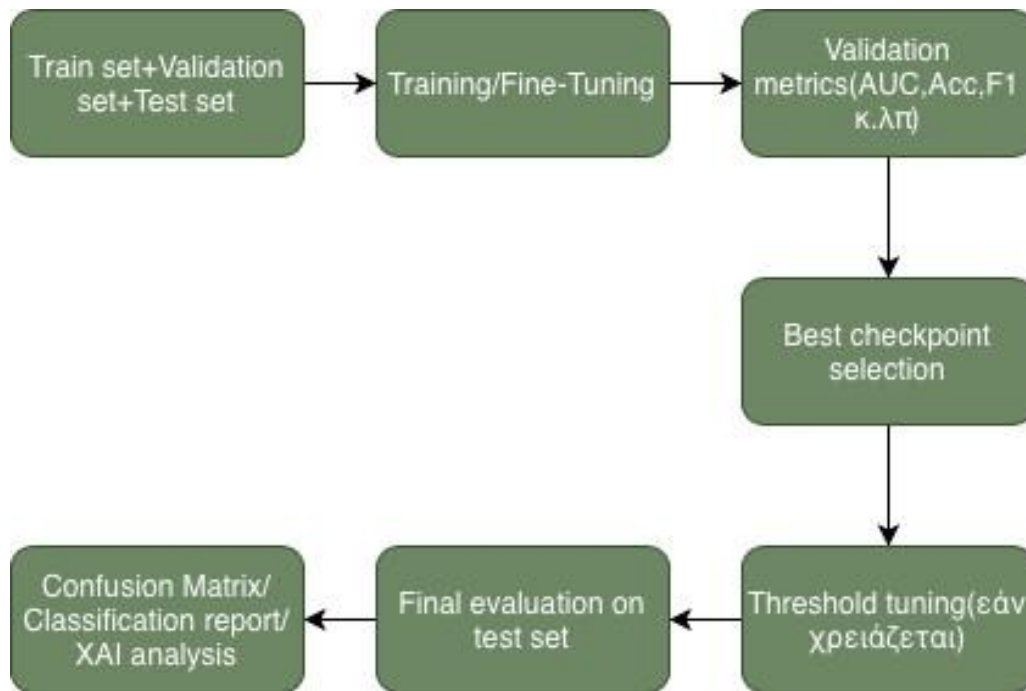


Σχήμα 3.5. Συνοπτική απεικόνιση του πειραματικού περιβάλλοντος και των βασικών εργαλείων που χρησιμοποιήθηκαν για την υλοποίηση των πειραμάτων.

Για την εκπαίδευση και το fine-tuning αξιοποιήθηκαν βιβλιοθήκες βαθιάς μάθησης, όπως PyTorch και torchvision, ενώ για τη διαχείριση δεδομένων, τις μετρικές και την οπτικοποίηση χρησιμοποιήθηκαν βιβλιοθήκες όπως pandas, NumPy, scikit-learn και matplotlib. Για την εξηγησιμότητα χρησιμοποιήθηκαν αντίστοιχες υλοποιήσεις Grad-CAM/Grad-CAM++, LIME, SHAP, occlusion attribution, καθώς και τεχνικές οπτικοποίησης embeddings όπως UMAP και t-SNE. Η επιλογή των εργαλείων αυτών δεν έγινε μόνο για λόγους ευκολίας. Η ύπαρξη notebooks επέτρεψε να καταγράφεται η πειραματική διαδικασία βήμα προς βήμα, ώστε να είναι πιο εύκολο να ελεγχθεί τι ακριβώς εφαρμόστηκε σε κάθε πείραμα. Παράλληλα, η αποθήκευση checkpoints, πινάκων μετρικών και εικόνων εξηγησιμότητας βοήθησε στην επαναληψιμότητα της διαδικασίας και στη σύγκριση διαφορετικών προσεγγίσεων.

3.5 Πρωτόκολλο αξιολόγησης και επιλογή καλύτερου μοντέλου

Η αξιολόγηση των μοντέλων οργανώθηκε ανάλογα με το είδος του προβλήματος. Στα πειράματα ταξινόμησης χρησιμοποιήθηκαν μετρικές όπως accuracy, precision, recall, F1-score, macro-F1, weighted-F1 και AUC. Η επιλογή των macro-μετρικών ήταν ιδιαίτερα σημαντική στα ανισόρροπα datasets, επειδή η απλή accuracy μπορεί να επηρεαστεί έντονα από τις πολυπληθείς κλάσεις και να αποκρύψει χαμηλότερη απόδοση σε σπανιότερες ή κλινικά πιο κρίσιμες κατηγορίες.



Σχήμα 3.6. Ροή του πρωτοκόλλου αξιολόγησης και της επιλογής του καλύτερου μοντέλου, από την εκπαίδευση μέχρι την τελική αποτίμηση στο test set.

Στα δυαδικά πειράματα, όπως melanoma έναντι non-melanoma, η AUC χρησιμοποιήθηκε για να εκτιμηθεί η συνολική διαχωριστική ικανότητα του μοντέλου, ανεξάρτητα από ένα συγκεκριμένο threshold. Παράλληλα, σε ορισμένες περιπτώσεις πραγματοποιήθηκε threshold tuning στο validation set, ώστε η τελική απόφαση να ισορροπεί καλύτερα ανάμεσα σε precision και recall. Αυτό είχε ιδιαίτερη σημασία για την κλάση melanoma, όπου τα false negatives είναι πιο σοβαρά από μια απλή αριθμητική αστοχία. Στα πειράματα τμηματοποίησης χρησιμοποιήθηκαν μετρικές όπως Dice/F1 και IoU, επειδή το ζητούμενο δεν είναι απλώς η σωστή ετικέτα της εικόνας αλλά η χωρική συμφωνία ανάμεσα στην προβλεπόμενη μάσκα και τη μάσκα αναφοράς. Το καλύτερο checkpoint επιλεγόταν με βάση τη συμπεριφορά στο validation set, συνήθως με κριτήριο την AUC, το F1 ή το IoU, ανάλογα με το task. Σε μικρά datasets, όπως το PH2, εφαρμόστηκε 5-fold cross-validation ώστε τα αποτελέσματα να μην εξαρτώνται από ένα μοναδικό train/validation split. Πέρα από τις αριθμητικές μετρικές, εξετάστηκαν confusion matrices, classification reports και οπτικές εξηγήσεις. Με αυτόν τον τρόπο, η αξιολόγηση δεν περιορίστηκε στο ερώτημα αν το μοντέλο πέτυχε υψηλό score, αλλά επεκτάθηκε στο ποια λάθη κάνει, σε ποιες κλάσεις δυσκολεύεται και αν οι εξηγήσεις του είναι συμβατές με την περιοχή της βλάβης.

4. Πειραματική αξιολόγηση

Στο παρόν κεφάλαιο παρουσιάζεται η πειραματική αξιολόγηση των μοντέλων που αναπτύχθηκαν και δοκιμάστηκαν στο πλαίσιο της παρούσας εργασίας. Μετά την περιγραφή της μεθοδολογίας και των τεχνικών που χρησιμοποιήθηκαν στο προηγούμενο κεφάλαιο, εδώ δίνεται έμφαση στα αποτελέσματα των πειραμάτων, στην ερμηνεία των μετρικών αξιολόγησης και στην ποιοτική ανάλυση των προβλέψεων των μοντέλων. Η αξιολόγηση δεν περιορίζεται μόνο στην τελική ακρίβεια των μοντέλων, καθώς στην περίπτωση της ιατρικής εικόνας η απλή σωστή ή λανθασμένη πρόβλεψη δεν επαρκεί για την πλήρη κατανόηση της συμπεριφοράς ενός συστήματος βαθιάς μάθησης. Για τον λόγο αυτό, εξετάζονται παράλληλα μετρικές όπως το F1-score, το AUC, η ακρίβεια ταξινόμησης, καθώς και μετρικές τμηματοποίησης, όπως το Dice coefficient και το Intersection over Union. Με αυτόν τον τρόπο επιχειρείται μια πιο ολοκληρωμένη αποτίμηση της απόδοσης των μοντέλων. Ιδιαίτερη βαρύτητα δίνεται και στην εξηγησιμότητα των αποτελεσμάτων. Μέθοδοι όπως το Grad-CAM και συναφείς τεχνικές οπτικοποίησης αξιοποιούνται για να διερευνηθεί αν τα μοντέλα εστιάζουν πράγματι στις περιοχές της δερματικής βλάβης ή αν επηρεάζονται από άσχετα στοιχεία της εικόνας, όπως το φόντο, οι σκιές ή τεχνουργήματα της λήψης. Η διάσταση αυτή είναι κρίσιμη, διότι ένα μοντέλο μπορεί να εμφανίζει ικανοποιητικές αριθμητικές επιδόσεις, χωρίς απαραίτητα να βασίζει τις αποφάσεις του σε ιατρικά σχετικές περιοχές. Τα πειράματα που παρουσιάζονται στο κεφάλαιο αυτό περιλαμβάνουν τόσο προβλήματα ταξινόμησης όσο και προβλήματα τμηματοποίησης δερματικών βλαβών. Μέσα από τη σύγκριση διαφορετικών προσεγγίσεων, επιχειρείται να αναδειχθούν τα πλεονεκτήματα και οι περιορισμοί κάθε μεθόδου, αλλά και να φανεί σε ποιο βαθμό οι τεχνικές βαθιάς μάθησης και οι μετασχηματιστές μπορούν να συμβάλουν στην υποστήριξη της αναγνώρισης μελανωμάτων.

4.1 HAM10000 με ResNet50 για δυαδική ταξινόμηση

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο ResNet50 για τη δυαδική ταξινόμηση δερματικών βλαβών στο σύνολο δεδομένων HAM10000. Ο στόχος ήταν η διάκριση ανάμεσα σε non-melanoma και melanoma, όπου η κλάση 0 αντιστοιχούσε στις μη μελανωματικές βλάβες και η κλάση 1 στο μελάνωμα. Το πείραμα δεν περιορίστηκε μόνο στις αριθμητικές μετρικές ταξινόμησης, αλλά συνδυάστηκε με Grad-CAM++ και SHAP, ώστε να εξεταστεί οπτικά ο τρόπος με τον οποίο το μοντέλο οδηγείται στις προβλέψεις του.

Το dataset περιλάμβανε συνολικά 10.015 usable rows. Από αυτές, οι 8.902 εικόνες ανήκαν στην κλάση non-melanoma, ενώ οι 1.113 εικόνες ανήκαν στην κλάση melanoma. Η κατανομή αυτή δείχνει έντονη ανισορροπία υπέρ της non-melanoma κλάσης. Για τον λόγο αυτό χρησιμοποιήθηκε $\text{pos_weight} = 7.9987$ και ενεργοποιήθηκε focal BCE loss, ώστε το μοντέλο να δώσει μεγαλύτερη βαρύτητα στα δυσκολότερα και λιγότερο συχνά melanoma δείγματα.

Κλάση	Πλήθος εικόνων	Ποσοστό (%)
Non-melanoma	8.902	88.89%
Melanoma	1.113	11.11%

Κλάση	Πλήθος εικόνων	Ποσοστό (%)
Σύνολο	10.015	100.00%

Πίνακας 4.1. Δυναδική κατανομή του HAM10000 στο πείραμα ResNet50

Το σύνολο χωρίστηκε σε 7.010 εικόνες εκπαίδευσης, 1.502 εικόνες επικύρωσης και 1.503 εικόνες ελέγχου. Στο training set υπήρχαν 6.231 non-melanoma και 779 melanoma εικόνες. Στο validation set υπήρχαν 1.335 non-melanoma και 167 melanoma, ενώ στο test set υπήρχαν 1.336 non-melanoma και 167 melanoma εικόνες. Η διατήρηση της κατανομής στα splits είναι σημαντική, γιατί επιτρέπει πιο σταθερή αξιολόγηση του μοντέλου σε ένα ανισόρροπο πρόβλημα.

Σύνολο δεδομένων	Κλάση 0 / Non-melanoma	Κλάση 1 / Melanoma	Σύνολο
Train	6.231	779	7.010
Validation	1.335	167	1.502
Test	1.336	167	1.503
Σύνολο	8.902	1.113	10.015

Πίνακας 4.2. Κατανομή train, validation και test στο πείραμα ResNet50

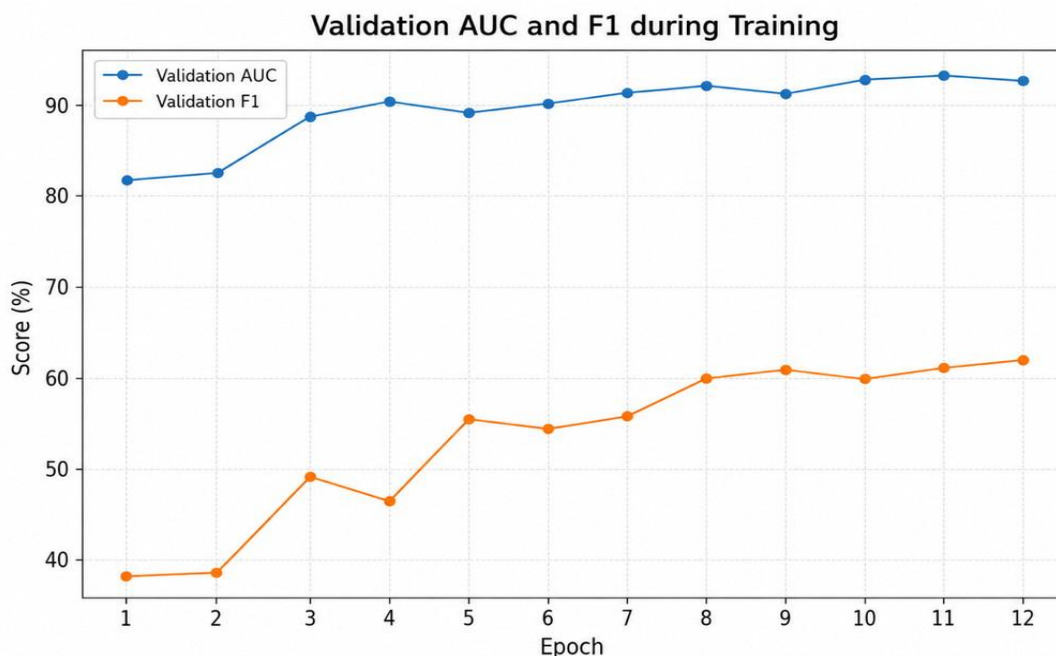
Η αρχιτεκτονική που χρησιμοποιήθηκε ήταν το ResNet50 με προεκπαιδευμένα βάρη. Η εκπαίδευση πραγματοποιήθηκε για 12 εποχές. Στις δύο πρώτες εποχές εκπαιδεύτηκε κυρίως το classification head, ενώ μετά τη 2η εποχή έγινε unfreezing του πλήρους μοντέλου, ώστε να προσαρμοστούν και τα βαθύτερα χαρακτηριστικά στα δερματοσκοπικά δεδομένα. Το καλύτερο checkpoint επιλέχθηκε με βάση το validation AUC.

Η πορεία της εκπαίδευσης δείχνει σταδιακή βελτίωση της διαχωριστικής ικανότητας. Στην 1η εποχή, το μοντέλο είχε val_acc = 70.64%, val_f1 = 37.97% και val_auc = 82.18%. Μετά το unfreezing, στην 3η εποχή, η απόδοση βελτιώθηκε αισθητά, με val_acc = 81.62%, val_f1 = 49.26% και val_auc = 88.87%. Η καλύτερη τιμή validation AUC καταγράφηκε στην 11η εποχή, με val_auc = 92.90%, val_acc = 90.01% και val_f1 = 61.14%. Στη 12η εποχή το validation F1 αυξήθηκε ελαφρώς σε 62.01%, αλλά το AUC μειώθηκε σε 92.35%, γι' αυτό ως καλύτερο μοντέλο κρατήθηκε το checkpoint της 11ης εποχής.



Εικόνα 4.1. Πορεία train και validation loss του ResNet50 στο HAM10000

Το γράφημα loss δείχνει ότι το train loss μειώνεται σταθερά, από 0.2471 στην 1η εποχή σε 0.0395 στη 12η. Αντίθετα, το validation loss δεν ακολουθεί την ίδια σταθερή πορεία. Μετά τις πρώτες εποχές παρουσιάζει αυξομειώσεις και στο τέλος ανεβαίνει σε 0.3279. Αυτό δείχνει ότι το μοντέλο συνεχίζει να προσαρμόζεται πολύ καλά στο training set, αλλά η γενίκευση στο validation set δεν βελτιώνεται γραμμικά. Με απλά λόγια, εμφανίζεται μια τάση υπερπροσαρμογής, παρότι το validation AUC παραμένει υψηλό.



Εικόνα 4.2. Πορεία validation AUC και F1 κατά την εκπαίδευση

Στο γράφημα των validation μετρικών φαίνεται ότι το AUC αυξάνεται αρκετά σταθερά από την 1η έως την 11η εποχή, φτάνοντας στο καλύτερο σημείο του με 0.9290. Το F1-score βελτιώνεται επίσης, αλλά με πιο αργό ρυθμό, από 0.3797 στην 1η εποχή σε 0.6201 στη 12η. Η διαφορά αυτή δείχνει ότι το μοντέλο μαθαίνει αρκετά καλά να διαχωρίζει τις δύο κλάσεις συνολικά, αλλά η τελική δυαδική απόφαση εξαρτάται έντονα από το threshold.

Epoch	Val accuracy (%)	Val F1 (%)	Val AUC (%)
1	70.64%	37.97%	82.18%
2	71.84%	38.25%	83.03%
3	81.62%	49.26%	88.87%
4	76.63%	46.41%	90.33%
8	89.55%	60.05%	91.86%
11	90.01%	61.14%	92.90%
12	90.95%	62.01%	92.35%

Πίνακας 4.3. Βασικές validation μετρικές του ResNet50 στο HAM10000

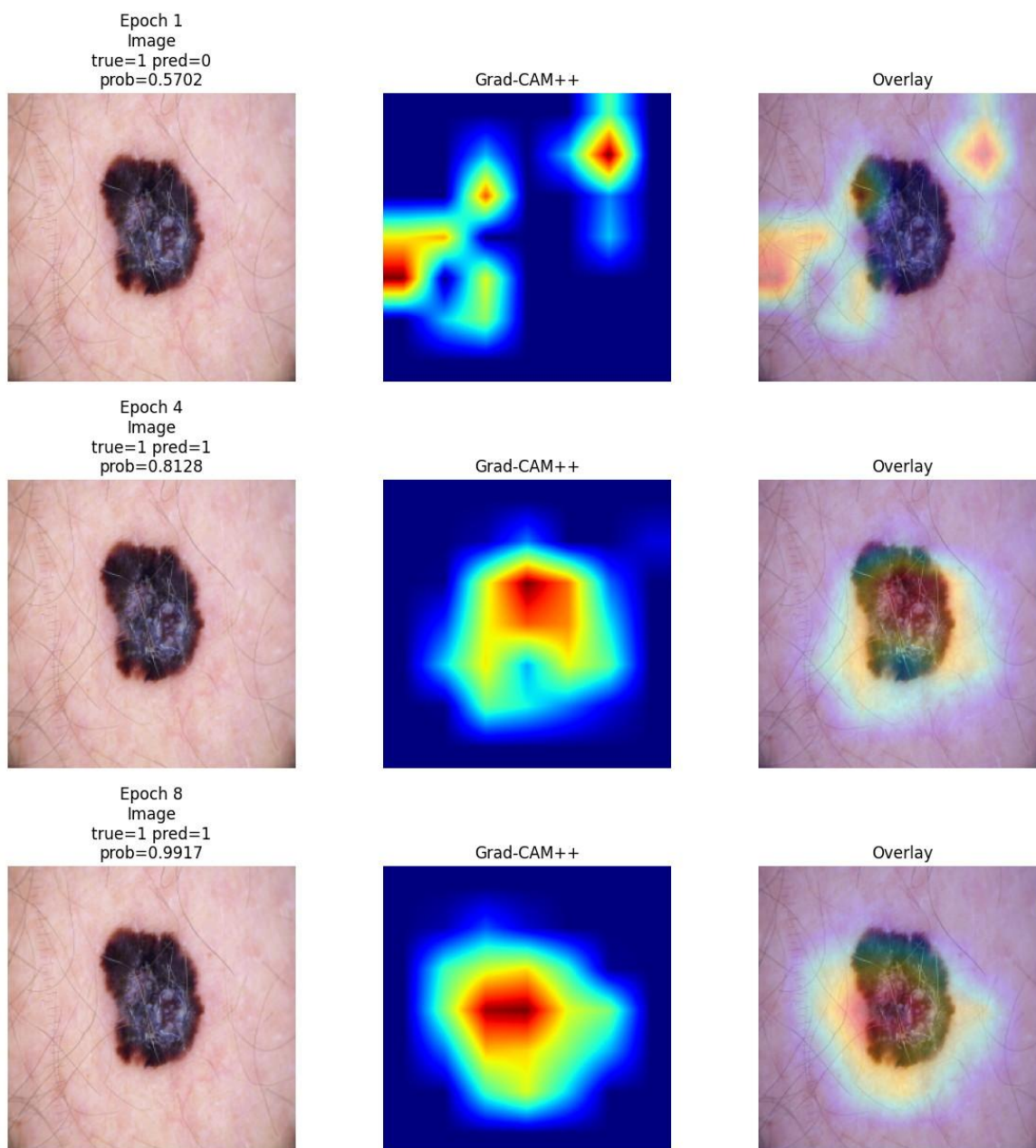
Μετά την εκπαίδευση έγινε validation threshold sweep. Το καλύτερο threshold ήταν 0.59, με validation F1 = 0.6516, accuracy = 91.81%, precision = 61.83% και recall = 68.86%. Το threshold αυτό χρησιμοποιήθηκε για την τελική αξιολόγηση στο test set. Η επιλογή threshold είναι κρίσιμη σε αυτό το πείραμα, γιατί με χαμηλότερο threshold αυξάνεται το recall αλλά μειώνεται η precision, ενώ με υψηλότερο threshold συμβαίνει το αντίστροφο. Στο test set, το καλύτερο μοντέλο ήταν αυτό της 11ης εποχής και αξιολογήθηκε με threshold 0.59. Τα τελικά αποτελέσματα ήταν Test Loss = 0.2530, Test Accuracy = 91.95%, Test Precision = 61.98%, Test Recall = 71.26%, Test F1 = 66.30% και Test AUC = 94.02%. Ο πίνακας σύγκρισης ήταν:

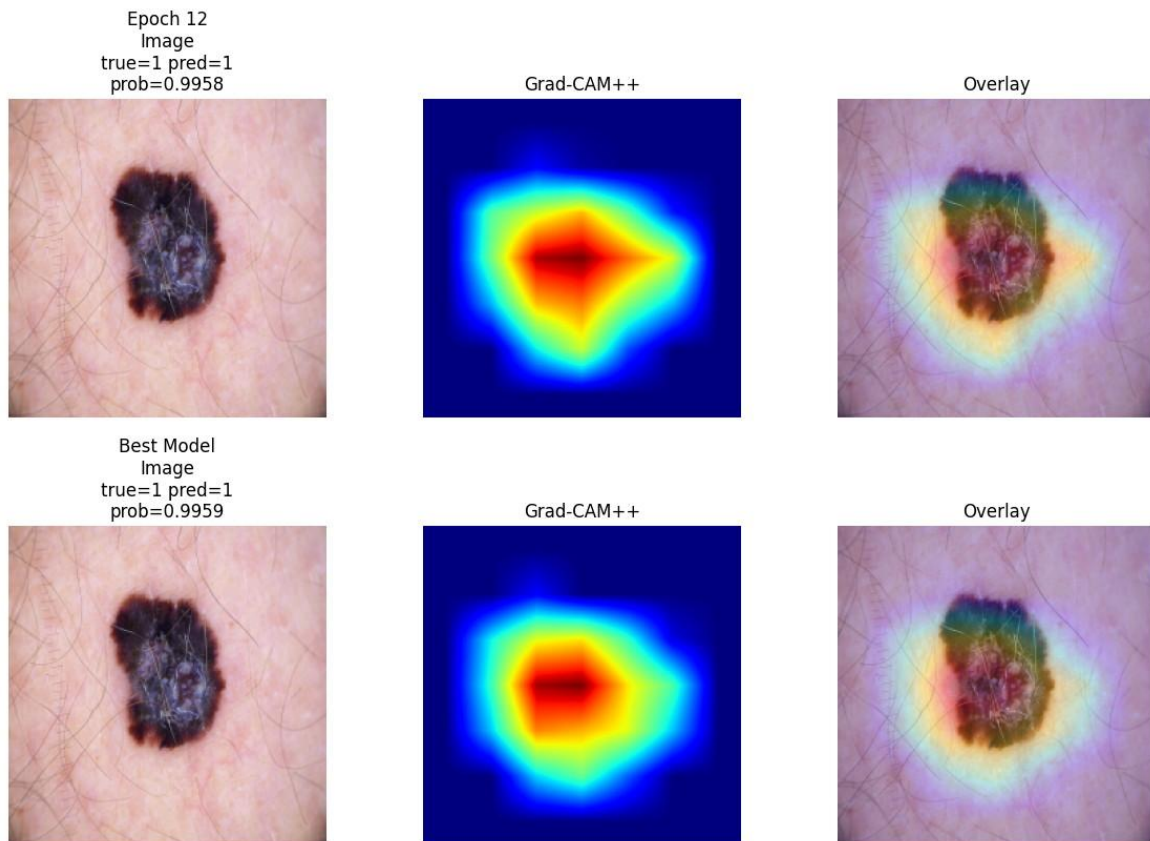
Κατηγορία	Τιμή
True Negatives (TN)	1263
False Positives (FP)	73
False Negatives (FN)	48
True Positives (TP)	119

Αυτό σημαίνει ότι από τις 1.336 non-melanoma εικόνες, οι 1.263 ταξινομήθηκαν σωστά και 73 ταξινομήθηκαν λανθασμένα ως melanoma. Από τις 167 melanoma εικόνες, οι 119 εντοπίστηκαν σωστά, ενώ 48 χάθηκαν ως non-melanoma. Το αποτέλεσμα αυτό δείχνει ότι το μοντέλο έχει καλή συνολική απόδοση, αλλά εξακολουθεί να υπάρχουν false negative melanoma περιπτώσεις, οι οποίες είναι ιδιαίτερα κρίσιμες σε ιατρικό πλαίσιο. Από το classification report φαίνεται ότι η κλάση non-melanoma είχε precision = 96.34%, recall = 94.54% και F1-score = 95.43%. Η κλάση melanoma είχε precision = 61.98%, recall = 71.26% και F1-score = 66.30%. Το macro F1 ήταν 80.86%, ενώ το weighted F1 ήταν 92.19%. Η διαφορά ανάμεσα στο macro και στο weighted F1 είναι αναμενόμενη, επειδή η

non-melanoma κλάση είναι πολύ μεγαλύτερη και επηρεάζει περισσότερο τη weighted μέτρηση.

Για την ποιοτική αξιολόγηση επιλέχθηκε μία χαρακτηριστική περίπτωση melanoma. Η ίδια εικόνα παρακολούθηθηκε σε διαφορετικά checkpoints, ώστε να εξεταστεί πώς αλλάζει η πρόβλεψη και η ερμηνεία κατά τη διάρκεια της εκπαίδευσης. Στην 1η εποχή, το μοντέλο ταξινόμησε την εικόνα λανθασμένα ως non-melanoma, παρότι η πραγματική ετικέτα ήταν melanoma, με πιθανότητα 0.5702. Από την 4η εποχή και μετά, η πρόβλεψη έγινε σωστή: στην epoch 4 η πιθανότητα ανέβηκε σε 0.8128, στην epoch 8 σε 0.9917, στην epoch 12 σε 0.9958, ενώ στο καλύτερο μοντέλο έφτασε σε 0.9959.

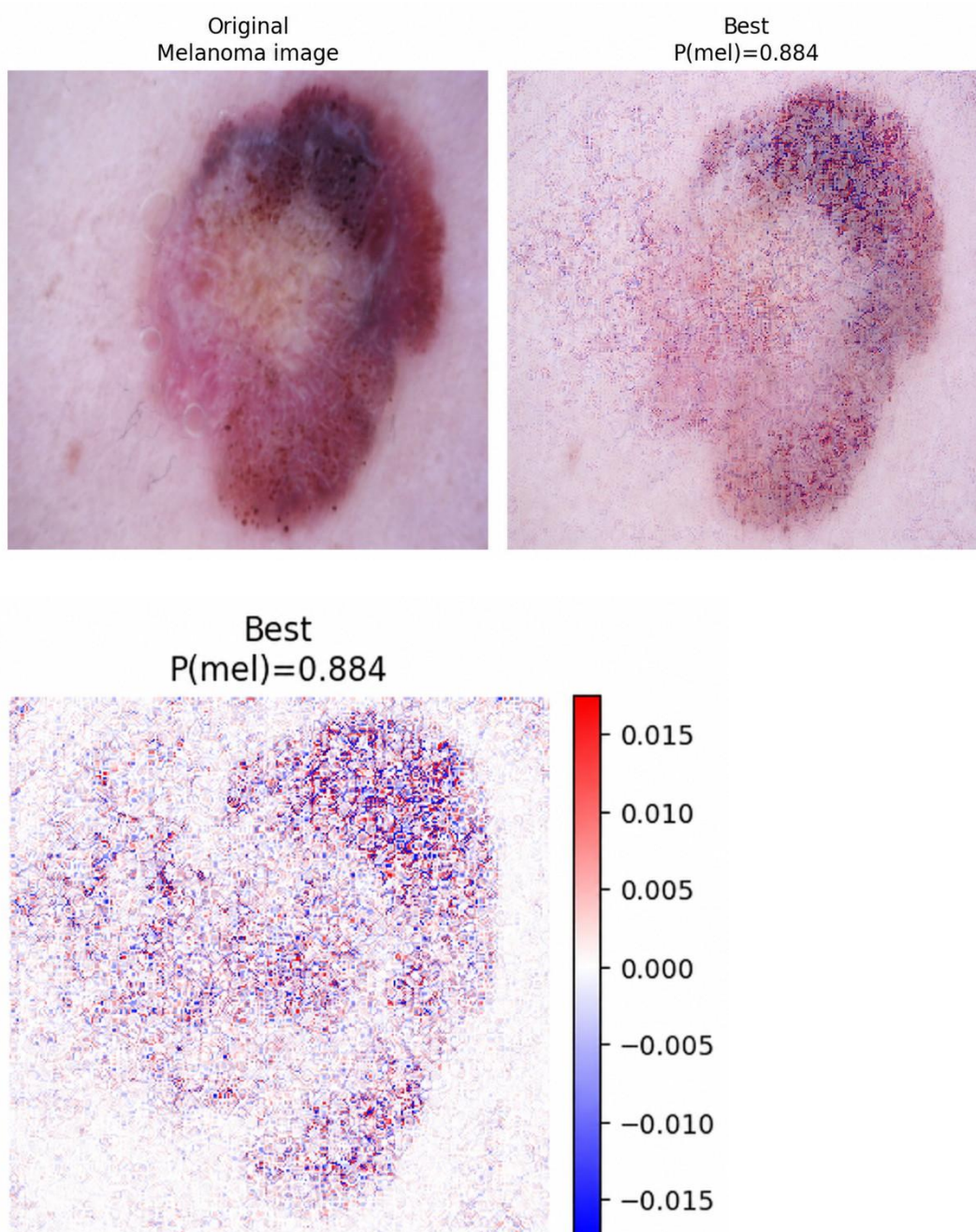




Εικόνα 4.3. Grad-CAM++ across-epoch ανάλυση για melanoma περίπτωση στο HAM10000

Η Grad-CAM++ across-epoch ανάλυση δείχνει καθαρά τη μεταβολή της εστίασης του μοντέλου. Στην 1η εποχή, όπου η πρόβλεψη είναι λανθασμένη, η ενεργοποίηση είναι διάσπαρτη και εμφανίζεται σε περιοχές εκτός της βασικής βλάβης. Στην 4η εποχή, όταν το μοντέλο αρχίζει να ταξινομεί σωστά την εικόνα ως melanoma, η ενεργοποίηση μετακινείται πάνω στην περιοχή της βλάβης. Στις εποχές 8 και στο best model, η Grad-CAM++ εστίαση γίνεται πιο σταθερή και καλύπτει κυρίως την κεντρική περιοχή της αλλοίωσης. Αυτό δείχνει ότι η εκπαίδευση δεν αύξησε απλώς την πιθανότητα της σωστής κλάσης, αλλά άλλαξε και τον τρόπο με τον οποίο το μοντέλο αξιοποίησε την εικόνα.

Εδώ είναι σημαντικό να τονιστεί ότι η Grad-CAM++ δεν είναι μάσκα τμηματοποίησης. Δεν δείχνει τα ακριβή όρια της βλάβης, αλλά τις περιοχές που επηρέασαν περισσότερο την πρόβλεψη. Παρ' όλα αυτά, η μετάβαση από διάσπαρτη ενεργοποίηση στην epoch 1 σε πιο lesion-centered ενεργοποίηση στις επόμενες εποχές είναι θετική ένδειξη. Δείχνει ότι το μοντέλο σταδιακά μαθαίνει να συνδέει την απόφαση melanoma με οπτικά στοιχεία της ίδιας της βλάβης.



Εικόνα 4.4. SHAP ανάλυση για melanoma περίπτωση στο καλύτερο μοντέλο

Η SHAP ανάλυση στο καλύτερο μοντέλο λειτουργεί συμπληρωματικά προς τη Grad-CAM++. Ενώ η Grad-CAM++ δείχνει ευρύτερη περιοχή ενεργοποίησης, το SHAP εμφανίζει πιο τοπικές συνεισφορές. Στην εικόνα που εξετάστηκε, οι θετικές SHAP τιμές εμφανίζονται κυρίως πάνω ή κοντά στη βλάβη, ιδιαίτερα σε περιοχές με έντονη χρωματική ανομοιογένεια. Αυτό υποστηρίζει την ερμηνεία ότι η τελική σωστή πρόβλεψη βασίστηκε σε χαρακτηριστικά της αλλοίωσης και όχι κυρίως στο φόντο.

Συνολικά, το πείραμα με ResNet50 στο HAM10000 έδειξε ότι το μοντέλο είχε ισχυρή διαχωριστική ικανότητα, με Test AUC = 94.02% και Test Accuracy = 91.95%. Για την κλάση melanoma πέτυχε recall = 71.26% και precision = 61.98%, κάτι που δείχνει μια σχετικά καλή αλλά όχι πλήρη ικανότητα εντοπισμού των πραγματικών μελανωμάτων. Η ύπαρξη 48 false negatives δείχνει ότι το μοντέλο δεν μπορεί να θεωρηθεί ασφαλές ως αυτόνομο διαγνωστικό εργαλείο, ειδικά σε ένα τόσο ευαίσθητο ιατρικό πρόβλημα.

4.2 Πείραμα δυαδικής ταξινόμησης στο PH2 με 5-fold cross-validation, Grad-CAM++ και SHAP

Στο πείραμα αυτό εξετάστηκε η δυαδική ταξινόμηση δερματικών βλαβών στο σύνολο δεδομένων PH2, με στόχο τη διάκριση ανάμεσα σε non-melanoma και melanoma περιπτώσεις. Το PH2 είναι ένα αρκετά μικρό αλλά συχνά χρησιμοποιούμενο dataset στη βιβλιογραφία, και γι' αυτό το λόγο η αξιολόγηση δεν έγινε με ένα μόνο train/test split, αλλά με 5-fold cross-validation. Η επιλογή αυτή ήταν πιο ασφαλής, γιατί σε μικρά σύνολα δεδομένων τα αποτελέσματα μπορεί να αλλάζουν αρκετά από split σε split.

Για τις ανάγκες του πειράματος, οι δύο κατηγορίες σπίλων του PH2 αντιμετωπίστηκαν ως μία ενιαία κλάση non-melanoma, ενώ τα περιστατικά μελανώματος αποτέλεσαν την κλάση melanoma. Με αυτόν τον τρόπο, το τελικό δυαδικό σύνολο περιλάμβανε 200 εικόνες, από τις οποίες 160 ανήκαν στην κλάση non-melanoma και 40 στην κλάση melanoma. Η αναλογία αυτή δείχνει ήπια ανισορροπία υπέρ της αρνητικής κλάσης, κάτι που έπρεπε να ληφθεί υπόψη κατά την εκπαίδευση και την αξιολόγηση.

Κλάση	Πλήθος εικόνων	Ποσοστό (%)
Non-melanoma	160	80.00%
Melanoma	40	20.00%
Σύνολο	200	100.00%

Πίνακας 4.4. Δυαδική κατανομή του PH2 στο πείραμα 5-fold cross-validation

Σε κάθε fold, το training set αποτελούνταν από 160 εικόνες και το validation set από 40 εικόνες. Η κατανομή παρέμενε σταθερή: στο training set υπήρχαν 128 non-melanoma και 32 melanoma, ενώ στο validation set υπήρχαν 32 non-melanoma και 8 melanoma εικόνες. Η σταθερότητα αυτής της κατανομής είναι σημαντική, γιατί επιτρέπει πιο δίκαιη σύγκριση ανάμεσα στα folds.

Σύνολο δεδομένων	Κλάση 0 / Non-melanoma	Κλάση 1 / Melanoma	Σύνολο
Train	128	32	160
Validation	32	8	40
Σύνολο ανά fold	160	40	200

Πίνακας 4.5. Κατανομή δεδομένων σε κάθε fold του πειράματος

Η εκπαίδευση του μοντέλου έγινε για έως 13 εποχές σε κάθε fold. Στις τρεις πρώτες εποχές εκπαιδεύτηκε πιο περιορισμένα το τελικό ταξινομητικό τμήμα του δικτύου, ενώ στη συνέχεια έγινε unfreezing του πλήρους μοντέλου, ώστε να προσαρμοστούν και τα βαθύτερα χαρακτηριστικά. Το καλύτερο checkpoint σε κάθε fold επιλέχθηκε με βάση το validation AUC, ενώ σε αρκετές περιπτώσεις ενεργοποιήθηκε και early stopping, όταν η περαιτέρω εκπαίδευση δεν έδινε ουσιαστικό όφελος. Φαίνεται ότι το μοντέλο βελτιώθηκε αισθητά μετά το unfreezing. Στο Fold 1, για παράδειγμα, το validation AUC ξεκίνησε από 77.34% στην 1η εποχή και έφτασε σε 100% στην 6η εποχή. Αντίστοιχα, στο Fold 3 το validation AUC έφτασε επίσης σε 100%, ενώ στα υπόλοιπα folds οι καλύτερες τιμές ήταν πολύ υψηλές: 96.88% στο Fold 2, 98.05% στο Fold 4 και 97.66% στο Fold 5. Η εικόνα αυτή δείχνει ότι το μοντέλο είχε πολύ καλή ικανότητα διάκρισης ανάμεσα στις δύο κλάσεις μέσα στα επιμέρους folds.

Fold	Best Epoch	Best Validation AUC
1	6	100%
2	8	96.88%
3	7	100%
4	13	98.05%
5	9	97.66%

Πίνακας 4.6. Καλύτερη επίδοση ανά fold με βάση το validation AUC

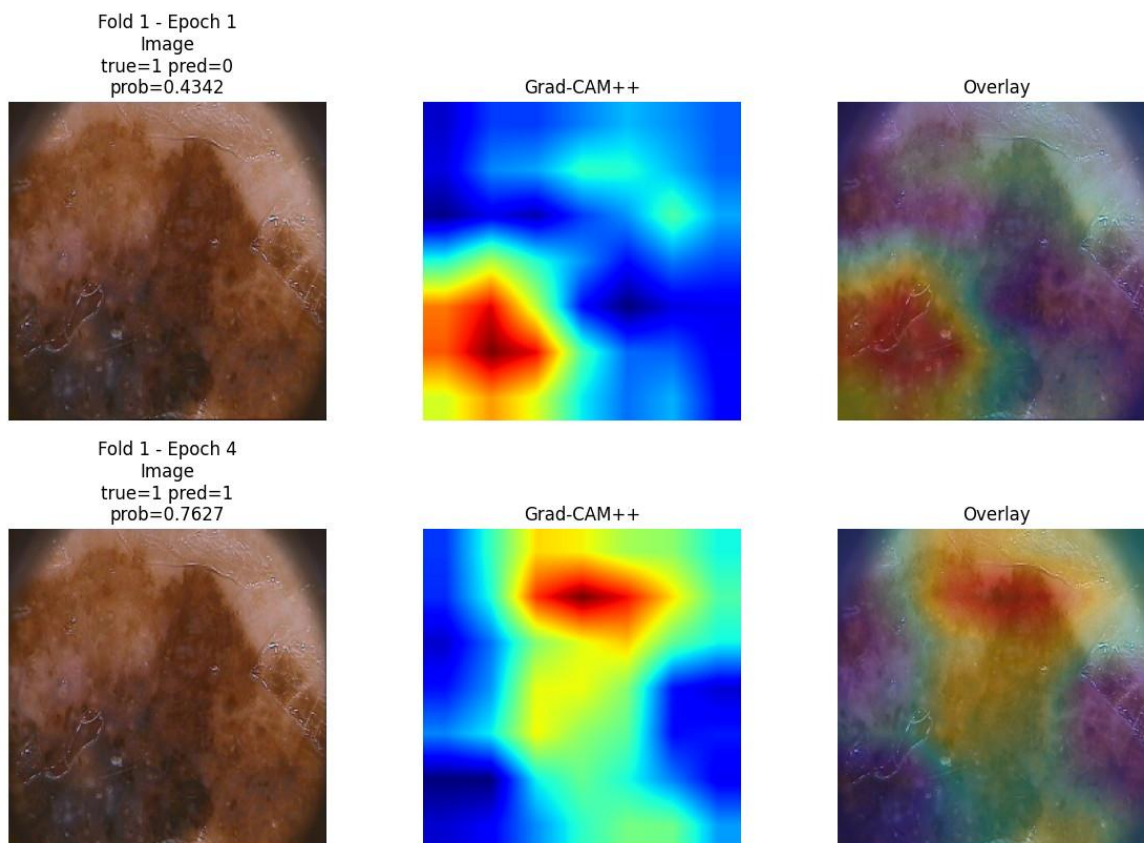
Metric	Mean CV value
Validation Accuracy	96.50%
Validation Precision	91.28%
Validation Recall	92.50%
Validation F1-score	91.24%
Validation AUC	98.52%

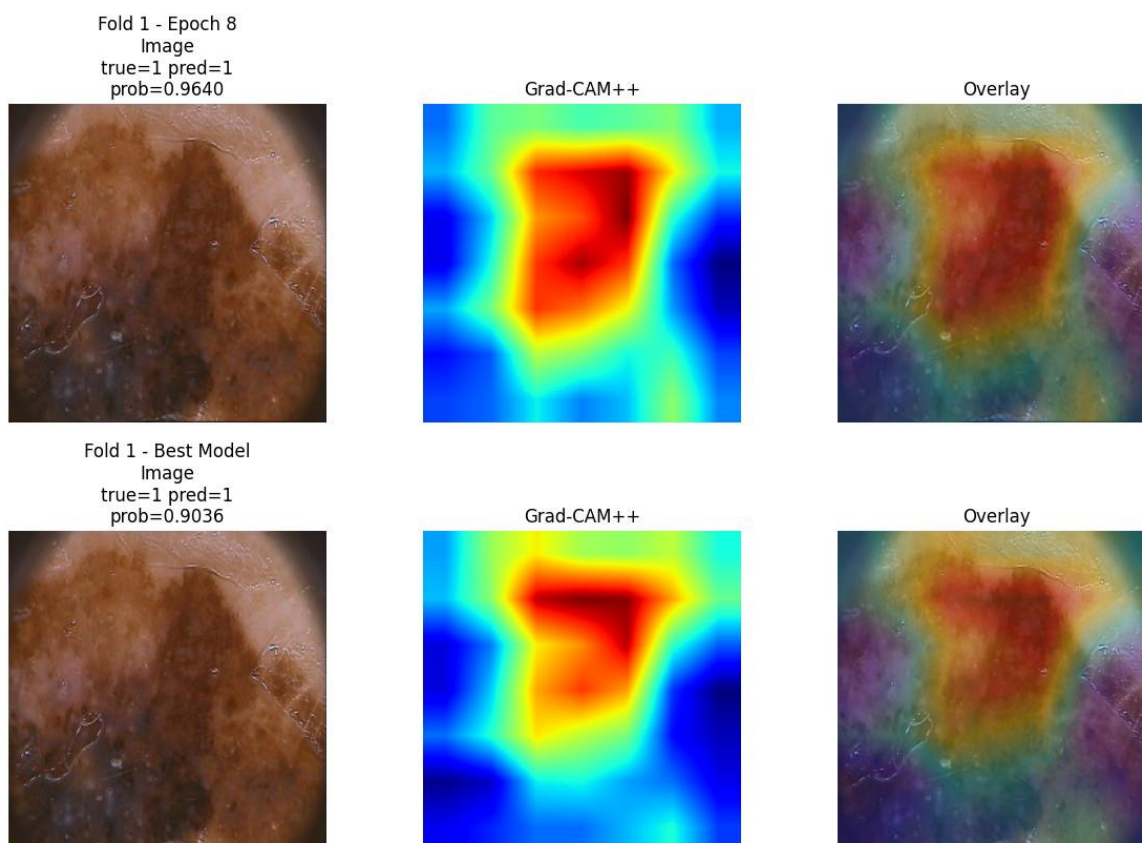
Πίνακας 4.7. Μέσες επιδόσεις 5-fold cross-validation στο PH2

Το καλύτερο fold, σύμφωνα με το notebook, ήταν το Fold 1, στο οποίο το βέλτιστο threshold βρέθηκε γύρω στο 0.69, με validation F1 = 94.12%. Η επιλογή threshold έχει εδώ σημασία, γιατί σε δυαδικά ιατρικά προβλήματα η τελική συμπεριφορά του μοντέλου επηρεάζεται αρκετά από το πού τοποθετείται το όριο απόφασης. Ένα πιο χαμηλό threshold αυξάνει συνήθως την ευαισθησία, αλλά μπορεί να φέρει περισσότερα false positives. Ένα πιο αυστηρό threshold κάνει το μοντέλο πιο συντηρητικό. Στο συγκεκριμένο πείραμα, το τελικό threshold του καλύτερου fold φαίνεται να εξισορρόπησε σχετικά καλά precision και recall. Το PH2 είναι μικρό dataset και, όπως συμβαίνει συχνά σε τέτοια σύνολα, ένα πολύ καλό AUC δεν σημαίνει αυτομάτως ότι το μοντέλο είναι εξίσου σταθερό σε κάθε μεμονωμένη περίπτωση. Με άλλα λόγια, οι συνολικές μετρικές είναι πολύ καλές, αλλά η ποιοτική ανάλυση παραμένει απαραίτητη για να φανεί πώς ακριβώς οδηγείται το μοντέλο στις αποφάσεις του. Για αυτόν ακριβώς τον λόγο, στο πείραμα χρησιμοποιήθηκαν και

μέθοδοι εξηγησιμότητας, κυρίως Grad-CAM++ και SHAP. Η ποιοτική ανάλυση οργανώθηκε γύρω από μία χαρακτηριστική εικόνα από το Fold 1, η οποία είναι ιδιαίτερα χρήσιμη επειδή δείχνει καθαρά τη μεταβολή της πρόβλεψης κατά τη διάρκεια της εκπαίδευσης.

Στην 1η εποχή, η συγκεκριμένη εικόνα είχε πραγματική ετικέτα melanoma (true = 1), αλλά το μοντέλο την ταξινόμησε λανθασμένα ως non-melanoma (pred = 0), με πιθανότητα 43.42%. Από την 4η εποχή και μετά η πρόβλεψη έγινε σωστή, με πιθανότητα 76.27%, ενώ στην 8η εποχή η πιθανότητα ανέβηκε σε 96.40%. Στην 10η εποχή έφτασε στο 97.41%. Στο best model η πρόβλεψη παρέμεινε σωστή, με πιθανότητα 90.36%. Το τελευταίο αυτό σημείο έχει ενδιαφέρον, γιατί δείχνει ότι το μοντέλο που δίνει το καλύτερο συνολικό AUC δεν είναι απαραίτητα εκείνο που δίνει και τη μεγαλύτερη δυνατή πιθανότητα σε κάθε μεμονωμένη εικόνα.



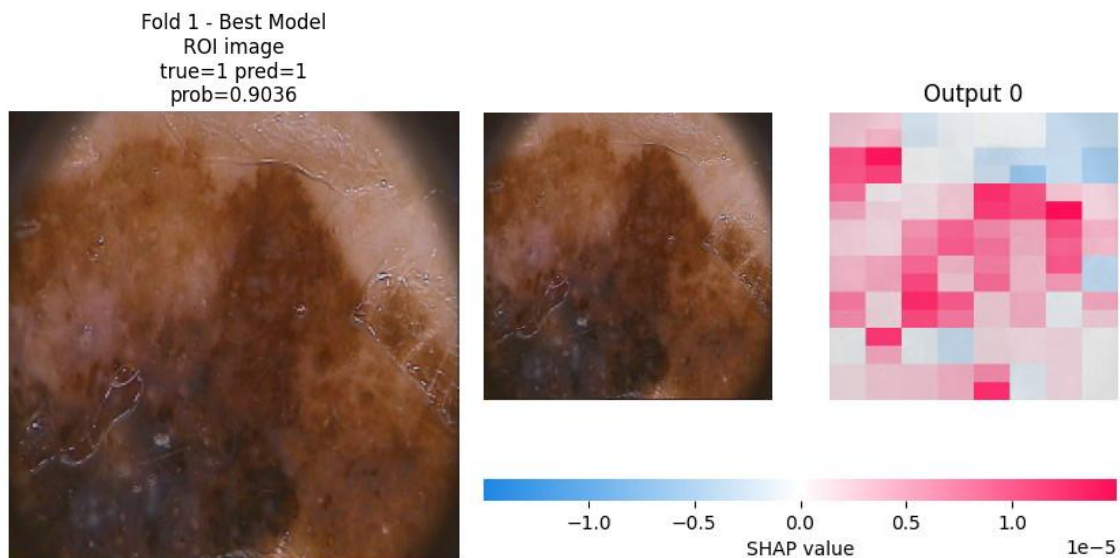


Εικόνα 4.5. Grad-CAM++ across-epoch ανάλυση για χαρακτηριστική melanoma περίπτωση από το Fold 1

Η Grad-CAM++ ανάλυση δείχνει αρκετά καθαρά αυτή τη μετάβαση. Στην 1η εποχή, όπου η πρόβλεψη είναι λανθασμένη, η ενεργοποίηση εμφανίζεται αρκετά διάχυτη και δεν εστιάζει καθαρά στην πιο ύποπτη περιοχή της βλάβης. Στην πράξη, το μοντέλο φαίνεται να “κοιτάζει” γενικά την εικόνα, χωρίς να έχει ακόμη μάθει ποια μορφολογικά στοιχεία έχουν μεγαλύτερη σημασία. Από την 4η εποχή και μετά, η ενεργοποίηση μεταφέρεται περισσότερο προς το κεντρικό και άνω τμήμα της βλάβης. Στην 8η εποχή και στο best model, η εστίαση γίνεται πιο συμπαγής και επικεντρώνεται σε περιοχές που σχετίζονται πιο άμεσα με τη μορφή και τη χρωματική δομή της αλλοίωσης. Αυτό είναι ένα αρκετά χρήσιμο εύρημα, γιατί δεν δείχνει απλώς ότι η πιθανότητα της σωστής κλάσης αυξήθηκε. Δείχνει και ότι το μοντέλο άλλαξε οπτικά στρατηγική κατά τη διάρκεια της εκπαίδευσης. Στην αρχή η προσοχή του είναι πιο ασταθής, ενώ αργότερα φαίνεται να στηρίζεται περισσότερο στην ίδια τη βλάβη. Φυσικά, η Grad-CAM++ δεν είναι μάσκα τμηματοποίησης και δεν πρέπει να διαβάζεται σαν ακριβής χαρτογράφηση των ορίων. Παρ’ όλα αυτά, ως ερμηνευτικό εργαλείο δείχνει αρκετά καλά αν το μοντέλο βασίζεται σε σχετικές ή άσχετες περιοχές.

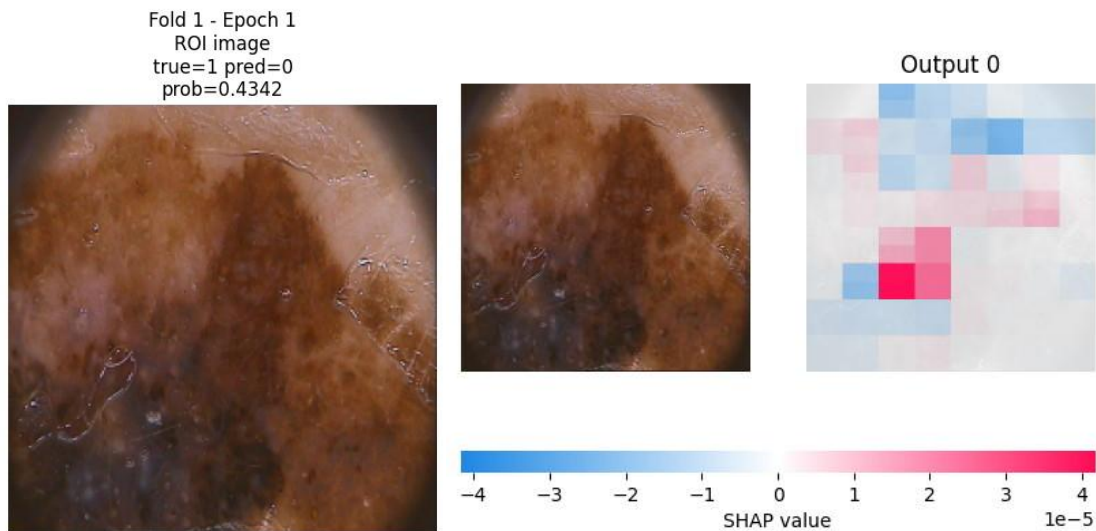
Για να υπάρχει και μια δεύτερη, πιο τοπική ματιά στην ίδια εικόνα, χρησιμοποιήθηκε και SHAP. Η SHAP ανάλυση είναι χρήσιμη γιατί δεν δίνει μόνο ένα θερμικό πεδίο προσοχής, αλλά δείχνει πιο συγκεκριμένα ποια τμήματα της εικόνας συνεισφέρουν θετικά ή αρνητικά στην τελική πρόβλεψη. Στη συγκεκριμένη περίπτωση, στην 1η εποχή οι SHAP συνεισφορές

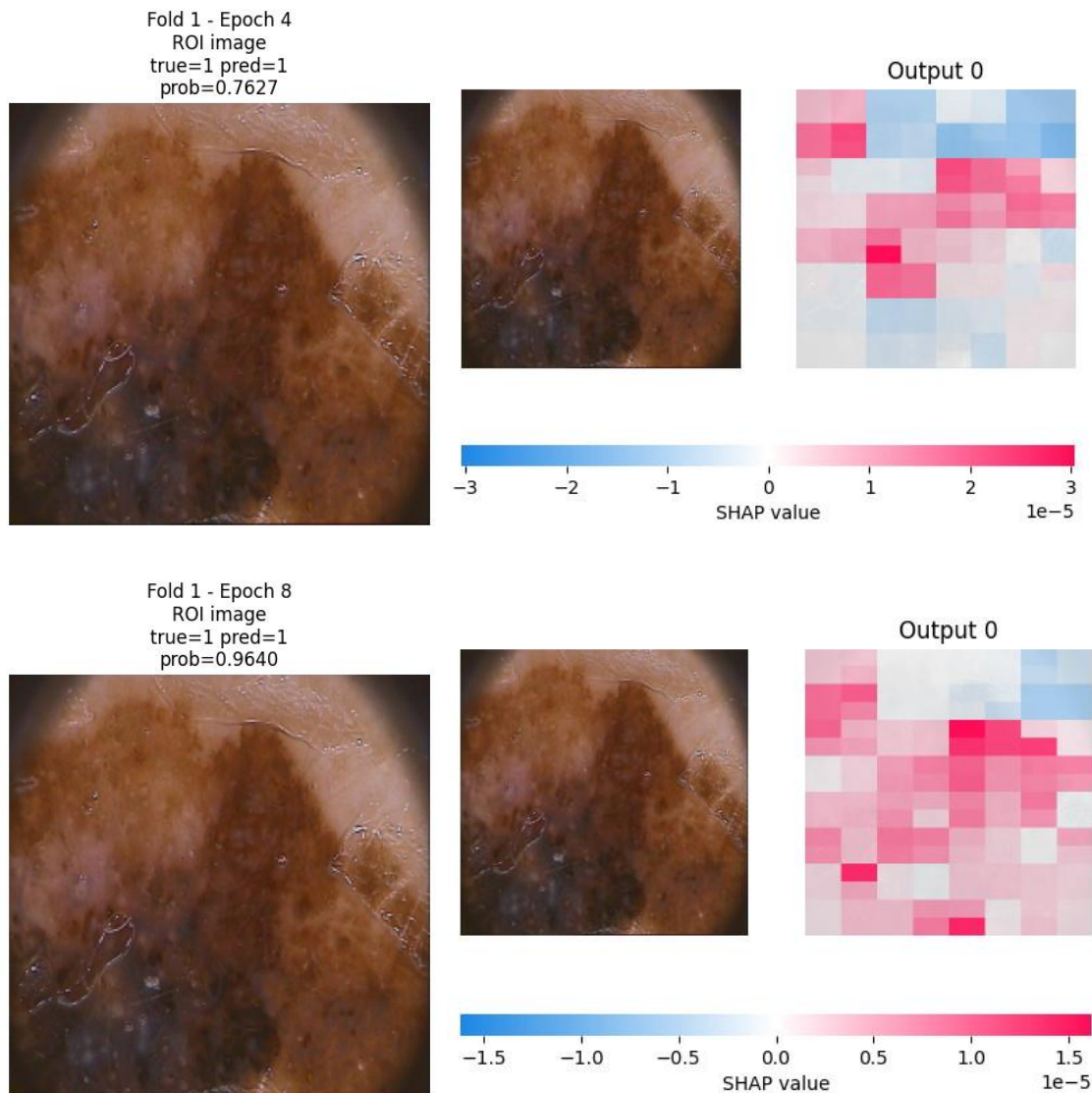
είναι πιο περιορισμένες και ασθενείς, ενώ από την 4η εποχή και μετά εμφανίζονται πιο καθαρά θετικές περιοχές πάνω στη βλάβη.



Εικόνα 4.6. SHAP ανάλυση της ίδιας melanoma περίπτωσης στο καλύτερο μοντέλο

Στο best model, η SHAP απεικόνιση δείχνει ότι οι πιο θετικές συνεισφορές εντοπίζονται κυρίως στο εσωτερικό της βλάβης και ιδιαίτερα σε περιοχές με πιο έντονη χρωματική διαφοροποίηση. Αυτό ενισχύει την ερμηνεία που δίνει και η Grad-CAM++: όσο το μοντέλο μαθαίνει καλύτερα, μεταφέρει όλο και περισσότερο την προσοχή του σε στοιχεία που συνδέονται με τη βλάβη και όχι με το φόντο.





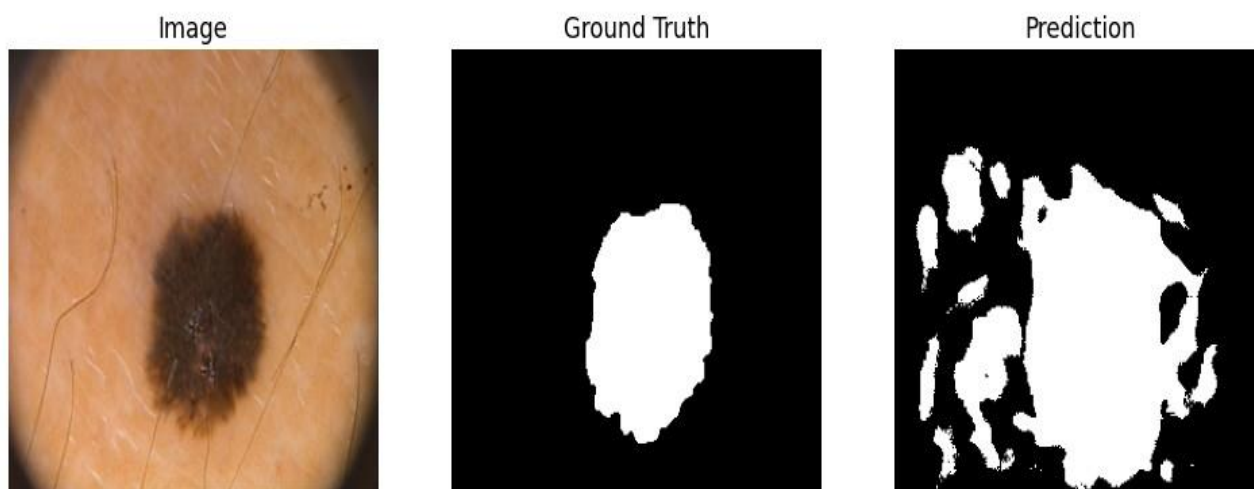
Εικόνα 4.7. SHAP across-epoch ανάλυση για χαρακτηριστική melanoma περίπτωση από το Fold 1

Συνολικά, το πείραμα αυτό έδειξε πολύ καλή απόδοση στη δυαδική ταξινόμηση του PH2, με ιδιαίτερα υψηλή μέση AUC 98.52% και πολύ καλή ισορροπία ανάμεσα σε precision (91.28%), recall (92.50%) και F1-score (91.24%). Το βασικό του ενδιαφέρον δεν βρίσκεται μόνο στους αριθμούς, αλλά και στο ότι η 5-fold cross-validation έδωσε μια πιο σταθερή εικόνα της συμπεριφοράς του μοντέλου σε ένα μικρό dataset. Ακόμη πιο χρήσιμη είναι η ποιοτική ανάγνωση του πειράματος. Η χαρακτηριστική εικόνα από το Fold 1 δείχνει ότι η μετάβαση από λανθασμένη σε σωστή πρόβλεψη συνοδεύεται και από μεταβολή της προσοχής του μοντέλου. Το μοντέλο δεν γίνεται απλώς πιο βέβαιο, αλλά φαίνεται να στηρίζεται σταδιακά σε πιο ουσιαστικά σημεία της βλάβης.

4.3 TMAHU-Net στο PH2 για τμηματοποίηση δερματικών βλαβών

Στο πείραμα αυτό αξιολογήθηκε ένα αναβαθμισμένο μοντέλο τύπου TMAHU-Net για την τμηματοποίηση δερματικών βλαβών στο σύνολο δεδομένων PH2. Σε αντίθεση με τα

προηγούμενα πειράματα ταξινόμησης, όπου το ζητούμενο ήταν η απόδοση μίας ετικέτας σε κάθε εικόνα, εδώ το μοντέλο έπρεπε να προβλέψει με ακρίβεια την περιοχή της βλάβης σε επίπεδο pixel. Το πρόβλημα, δηλαδή, αντιμετωπίστηκε ως binary segmentation, όπου κάθε pixel ταξινομείται είτε ως περιοχή βλάβης είτε ως υπόβαθρο. Το συγκεκριμένο πείραμα έχει ιδιαίτερη σημασία για την εργασία, επειδή η τμηματοποίηση αποτελεί ένα πιο λεπτομερές στάδιο ανάλυσης σε σχέση με την απλή ταξινόμηση. Ένα μοντέλο μπορεί να προβλέπει σωστά την κατηγορία μίας εικόνας, αλλά να μη βασίζεται πάντα στην πραγματική περιοχή της βλάβης. Στην τμηματοποίηση, η αξιολόγηση είναι πιο αυστηρή, καθώς εξετάζεται αν η προβλεπόμενη μάσκα συμπίπτει χωρικά με τη χειροκίνητα επισημασμένη μάσκα αναφοράς.

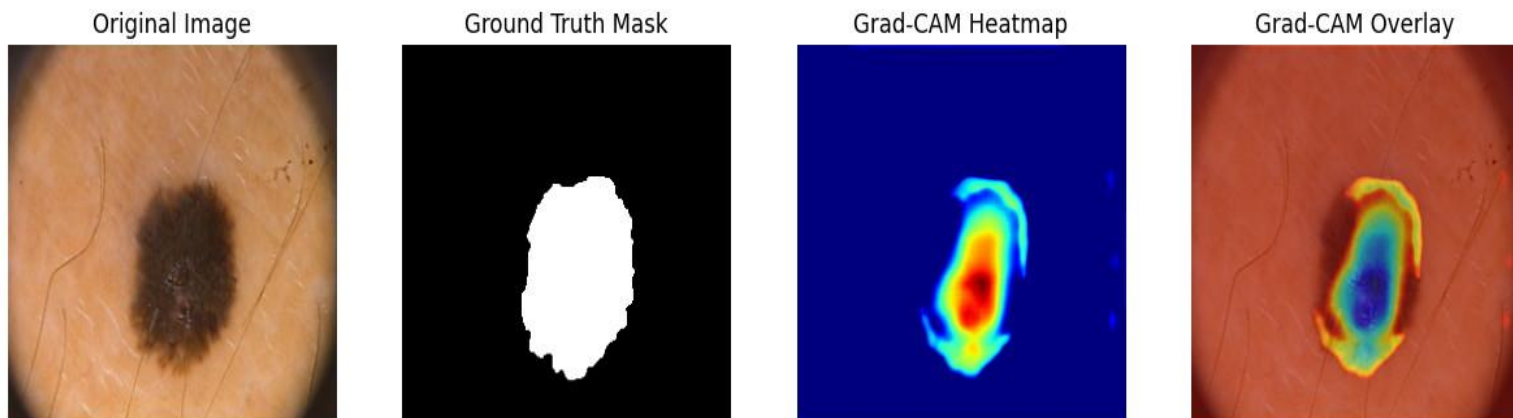


Εικόνα 4.8. Παράδειγμα πρόβλεψης μάσκας από το μοντέλο TMAHU-Net στο PH2

Το μοντέλο εκπαιδεύτηκε για 20 εποχές. Η καλύτερη επίδοση στο validation set καταγράφηκε στην 20ή εποχή, με Val F1 = 93.08% και Val IoU = 87.25%. Η τιμή αυτή δείχνει ότι το μοντέλο κατάφερε να μάθει αρκετά καλά τα όρια και την εσωτερική περιοχή των δερματικών βλαβών. Η μετρική F1 δείχνει την ισορροπία ανάμεσα στο precision και το recall, ενώ η μετρική IoU είναι πιο αυστηρή, επειδή μετρά τον βαθμό επικάλυψης ανάμεσα στην προβλεπόμενη μάσκα και τη μάσκα αναφοράς. Στο test set, το μοντέλο πέτυχε Test Loss = 0.2337, Test Precision = 92.94%, Test Recall = 91.73%, Test F1 = 91.81% και Test IoU = 85.33%. Οι τιμές αυτές δείχνουν ότι η απόδοση του μοντέλου παρέμεινε υψηλή και σε δεδομένα που δεν είχαν χρησιμοποιηθεί στην εκπαίδευση. Το precision 92.94% δείχνει ότι οι περιοχές που το μοντέλο χαρακτήρισε ως βλάβη ήταν σε μεγάλο βαθμό σωστές, ενώ το recall 91.73% δείχνει ότι το μοντέλο κατάφερε να εντοπίσει το μεγαλύτερο μέρος της πραγματικής βλάβης. Η τιμή IoU 85.33% είναι αρκετά ικανοποιητική για segmentation σε δερματοσκοπικές εικόνες, καθώς δείχνει μεγάλη χωρική συμφωνία ανάμεσα στην πρόβλεψη και στη μάσκα αναφοράς.

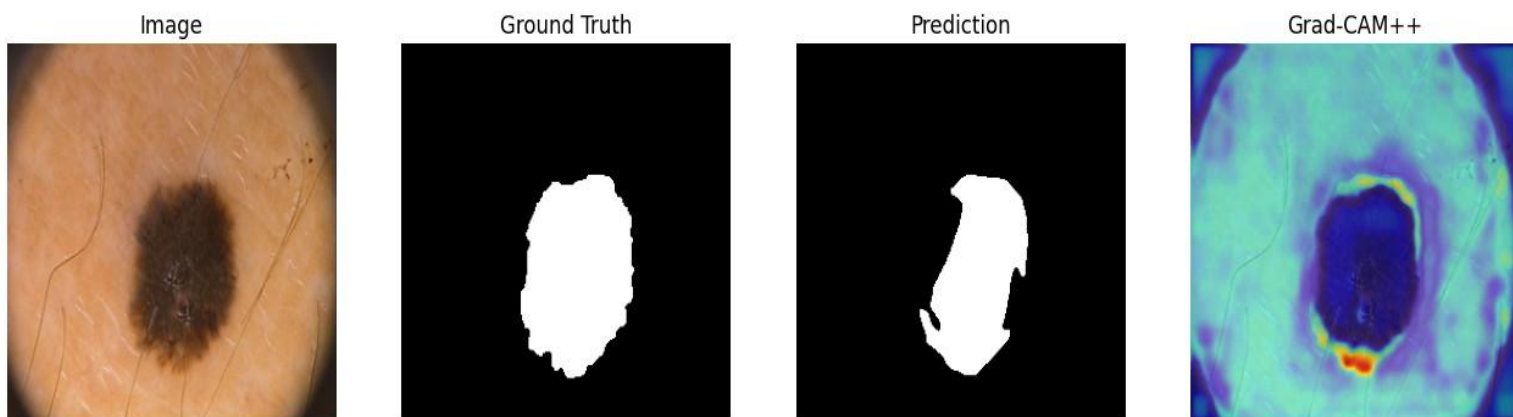
Test Loss	Test Precision	Test Recall	Test F1	Test IoU
0.2337	92.94%	91.73%	91.81%	85.33%

Για την οπτική αξιολόγηση του μοντέλου χρησιμοποιήθηκαν τεχνικές εξηγησιμότητας, ώστε να εξεταστεί όχι μόνο αν η μάσκα είναι σωστή, αλλά και ποιες περιοχές της εικόνας επηρεάζουν περισσότερο την απόφαση του μοντέλου. Αρχικά εφαρμόστηκε Grad-CAM για segmentation. Η μέθοδος αυτή ανέδειξε τις περιοχές στις οποίες εστίασε το μοντέλο κατά την παραγωγή της μάσκας. Στην περίπτωση του συγκεκριμένου παραδείγματος, η εστίαση φαίνεται να συνδέεται με την περιοχή της βλάβης και όχι με τυχαία σημεία του φόντου. Αυτό είναι θετικό, επειδή δείχνει ότι το μοντέλο δεν παράγει απλώς μία αριθμητικά καλή μάσκα, αλλά βασίζεται σε οπτικά σχετική πληροφορία.



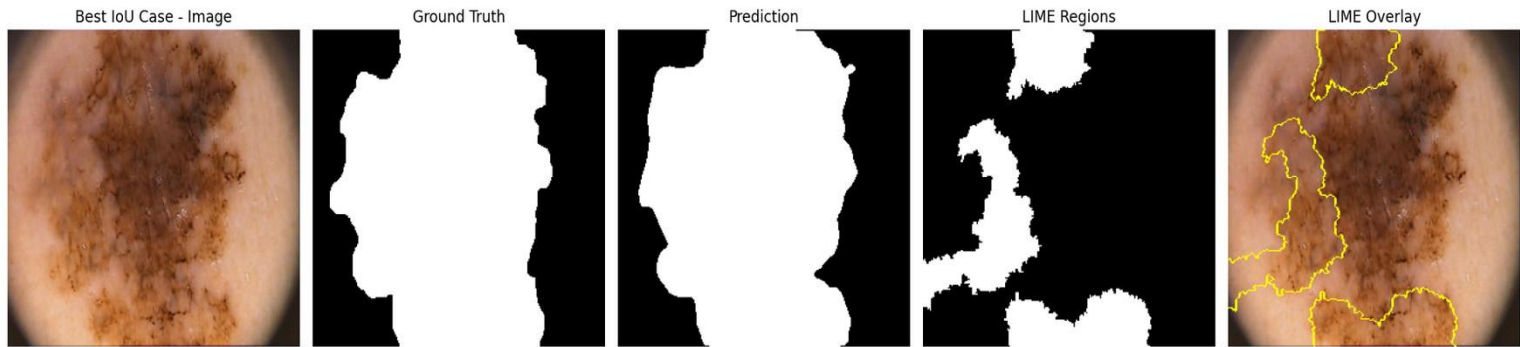
Εικόνα 4.9. Grad-CAM ανάλυση για την πρόβλεψη τμηματοποίησης του TMAHU-Net

Στη συνέχεια εφαρμόστηκε Grad-CAM++, το οποίο προσφέρει πιο λεπτομερή οπτική ερμηνεία σε σχέση με το απλό Grad-CAM. Στην αντίστοιχη απεικόνιση Grad-CAM++, η ενεργοποίηση εμφανίζεται κυρίως γύρω από την περιοχή της δερματικής βλάβης, χωρίς όμως να ταυτίζεται πλήρως με τη μάσκα αναφοράς ή με την προβλεπόμενη μάσκα. Αυτό δείχνει ότι η μέθοδος λειτουργεί περισσότερο ως ένδειξη της περιοχής που επηρέασε το μοντέλο και όχι ως ακριβής χάρτης τμηματοποίησης. Η foreground εκδοχή του Grad-CAM++ βοηθά να φανεί πιο καθαρά η σχέση ανάμεσα στην ενεργοποίηση και την περιοχή που το μοντέλο θεωρεί σημαντική. Παρατηρείται ότι η ερμηνεία συγκεντρώνεται κοντά στη βλάβη, αλλά υπάρχουν και περιοχές στα όρια ή περιφερειακά της βλάβης που ενεργοποιούνται, κάτι που δείχνει ότι η εξήγηση πρέπει να διαβαστεί συμπληρωματικά προς τις μετρικές Dice/IoU και όχι ως ανεξάρτητη απόδειξη ακρίβειας.



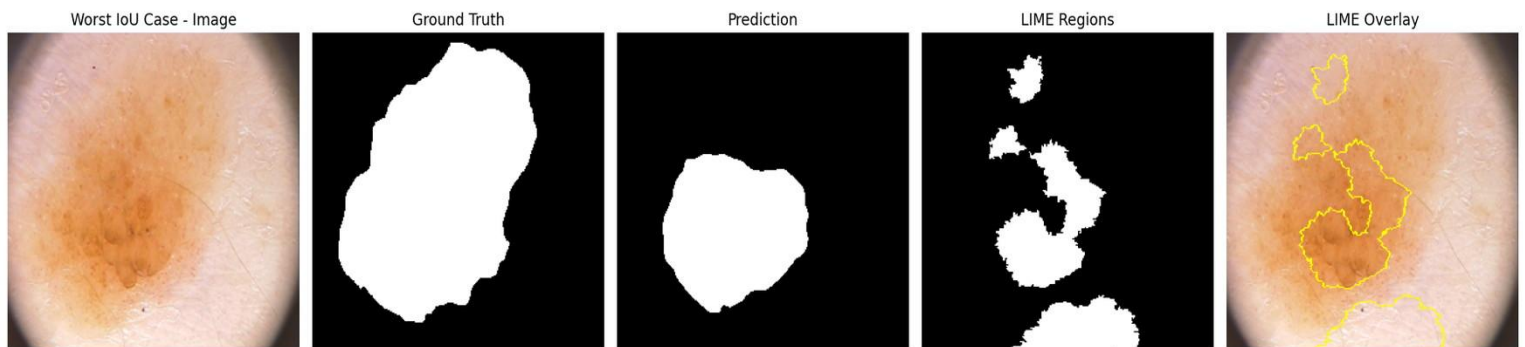
Εικόνα 4.10. Grad-CAM++ και foreground ερμηνεία της προβλεπόμενης μάσκας

Συμπληρωματικά χρησιμοποιήθηκε και η μέθοδος LIME, επίσης για best και worst case. Στην καλύτερη περίπτωση, το LIME ανέδειξε περιοχές που σχετίζονται με τη βλάβη και υποστήριξε την καλή πρόβλεψη της μάσκας. Σε αυτή την περίπτωση, η εξήγηση είναι συμβατή με το αριθμητικό αποτέλεσμα, καθώς το υψηλό IoU συνοδεύεται από οπτική εστίαση στην περιοχή ενδιαφέροντος.



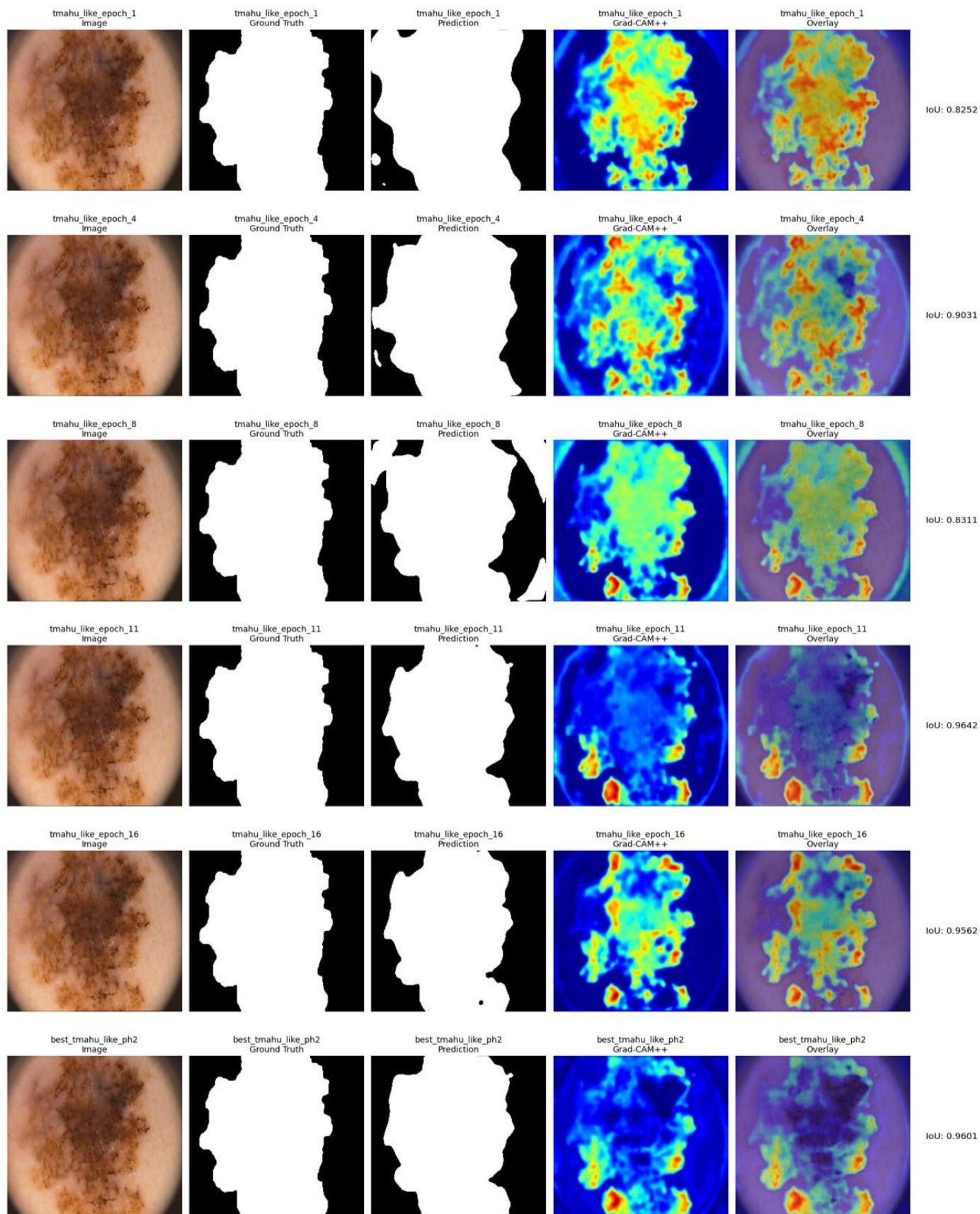
Εικόνα 4.11. LIME ανάλυση για την καλύτερη περίπτωση τμηματοποίησης

Στη δυσκολότερη περίπτωση, το LIME δείχνει με πιο καθαρό τρόπο ότι το μοντέλο δεν καταφέρνει να απομονώσει πλήρως τη σωστή περιοχή. Η εξήγηση μπορεί να παρουσιάζει διάσπαρτες ή λιγότερο συνεκτικές περιοχές ενδιαφέροντος, κάτι που συνδέεται με τη χαμηλότερη τιμή IoU. Με αυτόν τον τρόπο, το LIME λειτουργεί ως πρόσθετο εργαλείο ελέγχου, καθώς επιτρέπει να δούμε αν η λανθασμένη μάσκα οφείλεται σε ασάφεια των ορίων, σε ελλιπή κάλυψη της βλάβης ή σε εστίαση σε δευτερεύουσες περιοχές.



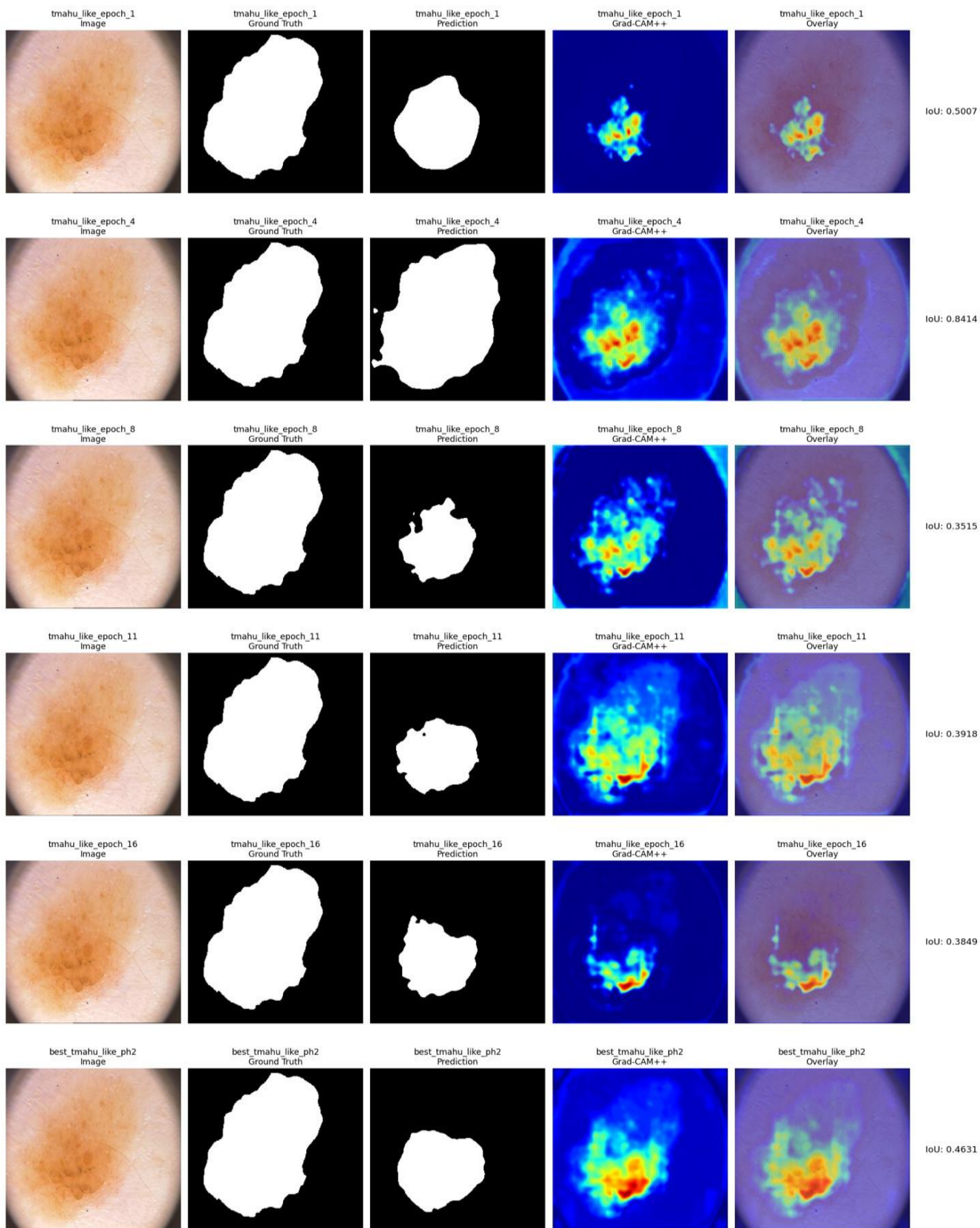
Εικόνα 4.12. LIME ανάλυση για δύσκολη περίπτωση τμηματοποίησης

Ένα από τα πιο χρήσιμα σημεία του αναβαθμισμένου πειράματος είναι η ανάλυση across epochs. Η συγκεκριμένη ανάλυση δεν περιορίζεται στην τελική κατάσταση του μοντέλου, αλλά εξετάζει πώς εξελίσσεται η πρόβλεψη και η εξήγηση κατά τη διάρκεια της εκπαίδευσης. Για την καλύτερη περίπτωση, η τιμή IoU ανά εποχή ήταν: 82.52% στην epoch 1, 90.31% στην epoch 4, 83.11% στην epoch 8, 96.42% στην epoch 11, 95.62% στην epoch 16 και 96.01% στο best model. Οι τιμές αυτές δείχνουν ότι η μάσκα βελτιώνεται σημαντικά από τις πρώτες εποχές και σταθεροποιείται σε πολύ υψηλό επίπεδο μετά την epoch 11.



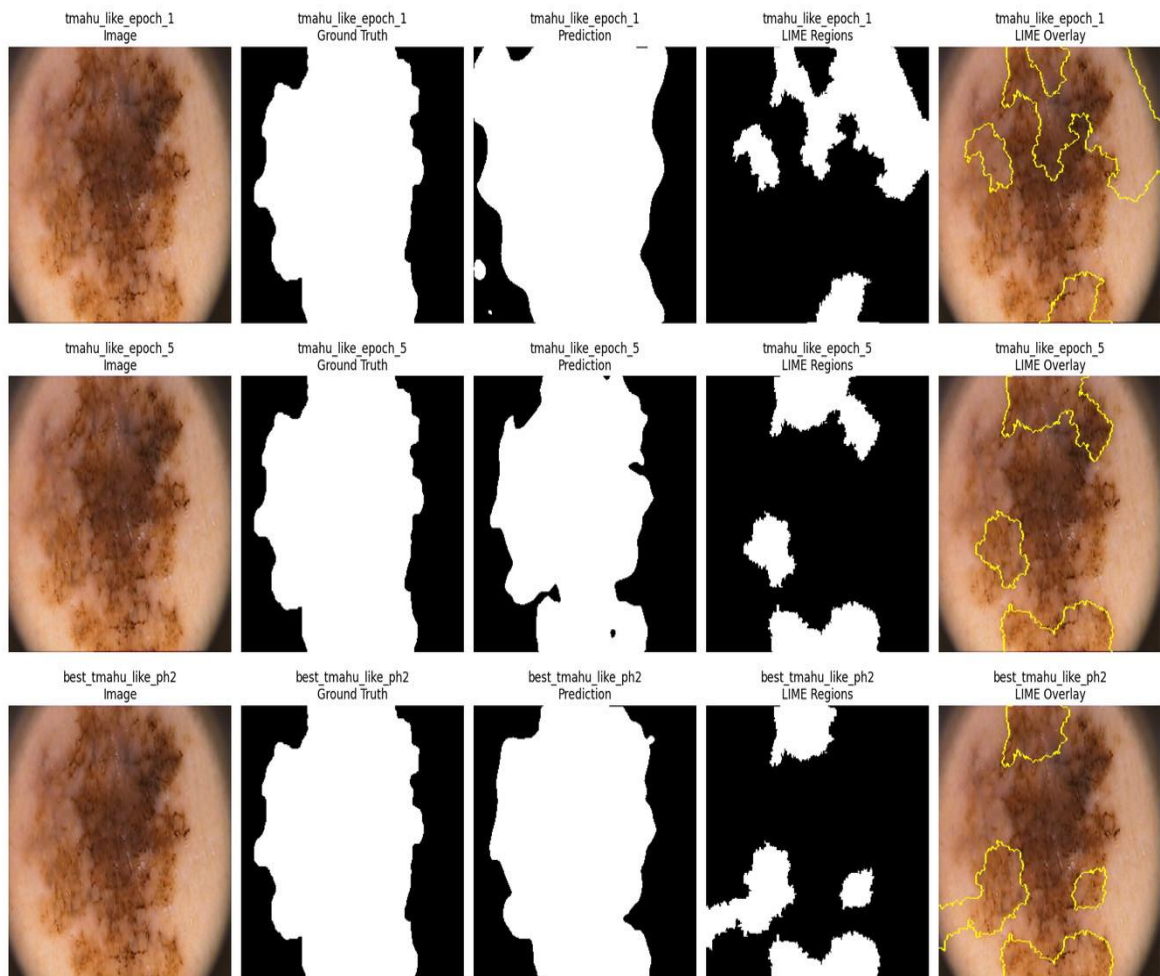
Εικόνα 4.13. Εξέλιξη της πρόβλεψης και της Grad-CAM++ ερμηνείας ανά εποχή για την καλύτερη περίπτωση

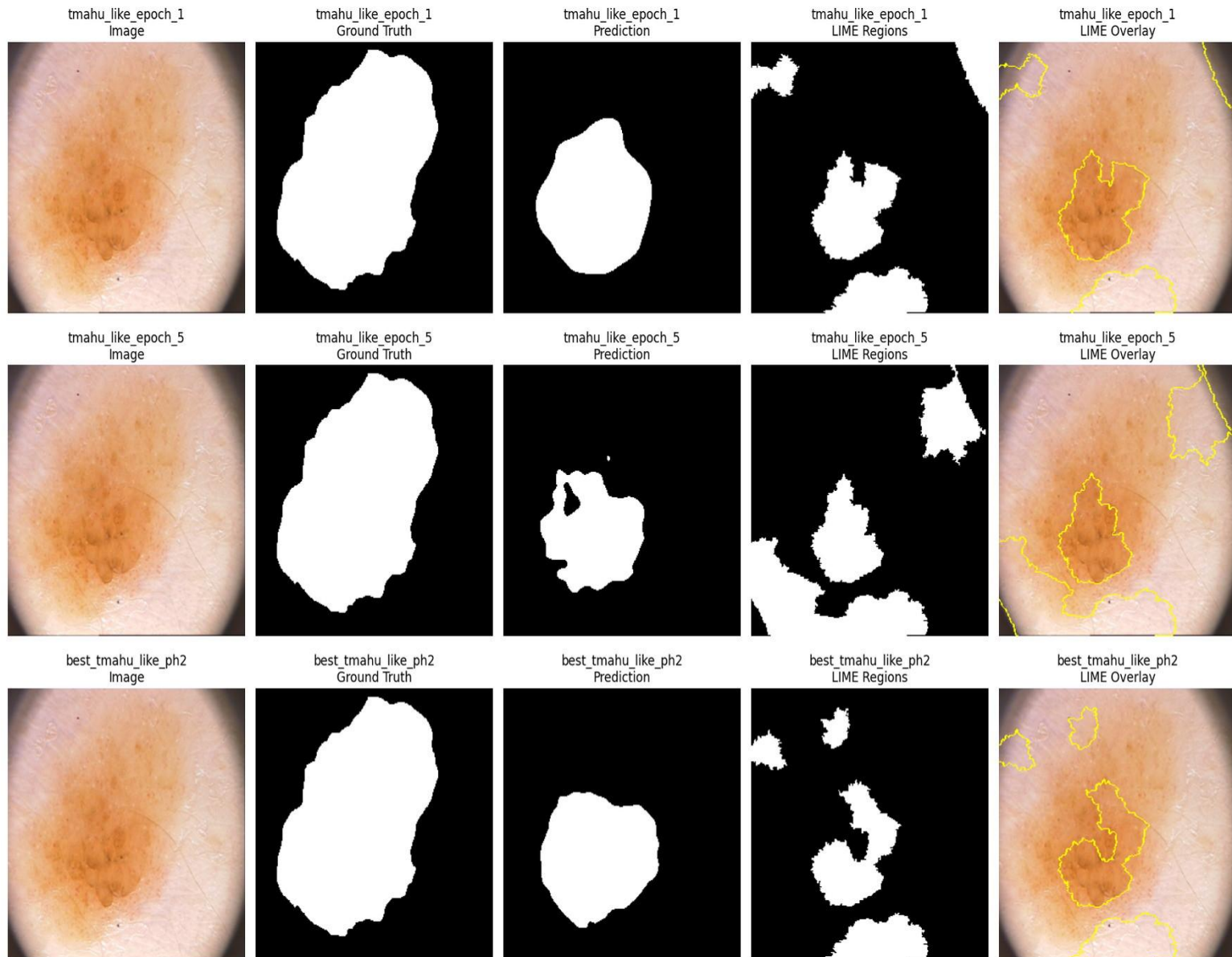
Η ίδια λογική εφαρμόστηκε και στη δύσκολη περίπτωση. Εκεί, οι τιμές IoU ήταν 50.07% στην epoch 1, 84.14% στην epoch 4, 35.15% στην epoch 8, 39.18% στην epoch 11, 38.49% στην epoch 16 και 46.31% στο best model. Η πορεία αυτή δείχνει ότι η δύσκολη περίπτωση δεν βελτιώθηκε σταθερά με την εκπαίδευση. Αντί να υπάρχει συνεχής αύξηση της χωρικής συμφωνίας, η απόδοση παρουσίασε έντονες διακυμάνσεις. Αυτό δείχνει ότι ορισμένες εικόνες παραμένουν δύσκολες για το μοντέλο, ακόμη και όταν η συνολική απόδοση στο test set είναι υψηλή.



Εικόνα 4.14. Εξέλιξη της πρόβλεψης και της Grad-CAM++ ερμηνείας ανά εποχή για τη δυσκολότερη περίπτωση

Η αντίστοιχη LIME across epochs ανάλυση χρησιμοποιήθηκε για να εξεταστεί η σταθερότητα των τοπικών εξηγήσεων στις epochs 1, 5 και στο best model. Στην καλύτερη περίπτωση, η σταδιακή συγκέντρωση της ερμηνείας γύρω από τη βλάβη δείχνει ότι η εκπαίδευση οδηγεί το μοντέλο σε πιο στοχευμένη χρήση των οπτικών πληροφοριών. Στη δύσκολη περίπτωση, η LIME απεικόνιση μπορεί να δείξει μεγαλύτερη αστάθεια, δηλαδή περιοχές ενδιαφέροντος που δεν συμπίπτουν πλήρως με τη μάσκα αναφοράς.





Εικόνα 4.15. LIME across-epoch ανάλυση για την καλύτερη και τη δυσκολότερη περίπτωση

Συνολικά, το TMAHU-Net στο PH2 έδειξε πολύ καλή απόδοση στην τμηματοποίηση δερματικών βλαβών. Το Test F1 = 91.81% και το Test IoU = 85.33% δείχνουν ότι το μοντέλο μπορεί να παράγει μάσκες με υψηλή χωρική ακρίβεια. Παράλληλα, οι τεχνικές Grad-CAM, Grad-CAM++ και LIME πρόσθεσαν ένα δεύτερο επίπεδο αξιολόγησης, πέρα από τις αριθμητικές μετρικές. Η ανάλυση έδειξε ότι στις επιτυχημένες περιπτώσεις το μοντέλο εστιάζει κυρίως στην περιοχή της βλάβης, ενώ στις δύσκολες περιπτώσεις η ερμηνεία γίνεται λιγότερο σταθερή και η χωρική συμφωνία με τη μάσκα αναφοράς μειώνεται. Το βασικό συμπέρασμα από το πείραμα είναι ότι το TMAHU-Net μπορεί να λειτουργήσει αποτελεσματικά ως μοντέλο τμηματοποίησης στο PH2, ενώ η across-epoch εξηγησιμότητα δίνει μια πιο ουσιαστική εικόνα της διαδικασίας μάθησης. Δεν αξιολογείται μόνο το τελικό αποτέλεσμα, αλλά και η πορεία με την οποία το μοντέλο μαθαίνει να

εντοπίζει τη βλάβη. Αυτό έχει ιδιαίτερη αξία σε ιατρικές εικόνες, όπου η ερμηνευσιμότητα της απόφασης είναι σχεδόν εξίσου σημαντική με την ίδια την απόδοση.

4.4 Πείραμα πολυταξικής ταξινόμησης Monkeypox Skin Image Dataset με DenseNet121

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο DenseNet121 για την πολυταξική ταξινόμηση εικόνων δερματικών αλλοιώσεων από το Monkeypox Skin Image Dataset. Σε αντίθεση με τα δυαδικά πειράματα, όπου το μοντέλο έπρεπε να διακρίνει ανάμεσα σε δύο μόνο κατηγορίες, εδώ το πρόβλημα ήταν πιο σύνθετο, καθώς περιλάμβανε τέσσερις κλάσεις: Chickenpox, Measles, Monkeypox και Normal. Ο στόχος ήταν να εξεταστεί αν ένα συνελκτικό μοντέλο, προεκπαιδευμένο σε ImageNet, μπορεί να διαχωρίσει όχι μόνο τις παθολογικές από τις φυσιολογικές εικόνες, αλλά και παθήσεις που παρουσιάζουν οπτικές ομοιότητες μεταξύ τους. Το μοντέλο βασίστηκε στην αρχιτεκτονική DenseNet121, με προεκπαιδευμένα βάρη από το ImageNet και μέγεθος εισόδου 224×224 pixels. Το σύνολο δεδομένων χωρίστηκε σε 539 εικόνες εκπαίδευσης, 115 εικόνες επικύρωσης και 116 εικόνες ελέγχου. Στο test set υπήρχαν 44 εικόνες Normal, 42 εικόνες Monkeypox, 16 εικόνες Chickenpox και 14 εικόνες Measles. Η κατανομή αυτή δείχνει ότι το test set δεν ήταν πλήρως ισορροπημένο, καθώς οι κλάσεις Normal και Monkeypox είχαν περισσότερα δείγματα από τις κλάσεις Chickenpox και Measles. Η εκπαίδευση πραγματοποιήθηκε για 10 εποχές, με καλύτερη επίδοση στην 10η εποχή, όπου το μοντέλο πέτυχε Validation Macro F1 = 92.27%. Η επιλογή του macro F1 ως βασικής validation μετρικής είναι χρήσιμη σε αυτό το πείραμα, επειδή λαμβάνει υπόψη την απόδοση σε κάθε κλάση ανεξάρτητα από το πλήθος των δειγμάτων της. Με αυτόν τον τρόπο, η αξιολόγηση δεν επηρεάζεται υπερβολικά από τις πολυπληθέστερες κλάσεις. Στο σύνολο ελέγχου, το μοντέλο πέτυχε Test Loss = 0.2423, Test Accuracy = 89.66%, Test Macro F1 = 86.11% και Test Weighted F1 = 89.60%. Η συνολική ακρίβεια δείχνει ότι το μοντέλο ταξινόμησε σωστά σχεδόν εννέα στις δέκα εικόνες. Παράλληλα, το weighted F1 είναι κοντά στην ακρίβεια, κάτι που δείχνει καλή συνολική συμπεριφορά στο test set. Το macro F1 είναι χαμηλότερο, στοιχείο που φανερώνει ότι οι μικρότερες ή δυσκολότερες κλάσεις επηρέασαν περισσότερο την ισορροπία της απόδοσης.

Test Loss	Test Accuracy	Test Macro F1	Test Weighted F1
89.66%	89.66%	86.11%	89.60%

Πίνακας 4.8. Συνολικές μετρικές αξιολόγησης του DenseNet121 στο Monkeypox Skin Image Dataset

Αναλυτικότερα, η καλύτερη απόδοση εμφανίστηκε στην κλάση Normal, όπου το μοντέλο πέτυχε precision 96%, recall 100% και F1-score 98%. Αυτό σημαίνει ότι οι φυσιολογικές εικόνες διαχωρίστηκαν σχεδόν πλήρως από τις παθολογικές εικόνες. Η κλάση Monkeypox είχε επίσης ικανοποιητική απόδοση, με precision 92%, recall 86% και F1-score 89%. Η χαμηλότερη επίδοση παρατηρήθηκε στην κλάση Chickenpox, με precision 75%, recall 75% και F1-score 75%. Η κλάση Measles είχε ενδιάμεση συμπεριφορά, με precision 80%, recall 86% και F1-score 83%.

Τα αποτελέσματα αυτά δείχνουν ότι το μοντέλο μπορεί να αναγνωρίσει αρκετά καλά την κλάση Monkeypox, αλλά δυσκολεύεται περισσότερο στις περιπτώσεις όπου οι δερματικές

αλλοιώσεις μοιάζουν οπτικά μεταξύ τους. Η χαμηλότερη επίδοση στο Chickenpox είναι αναμενόμενη, καθώς οι εικόνες αυτής της κατηγορίας μπορεί να εμφανίζουν ερυθρότητα, βλατίδες ή διάσπαρτες αλλοιώσεις που μοιάζουν με ορισμένες περιπτώσεις Monkeypox. Στην πράξη, αυτό το σημείο είναι κρίσιμο, γιατί η αριθμητική επίδοση πρέπει να εξετάζεται μαζί με το είδος των λαθών που κάνει το μοντέλο.

Classification Report	Precision	Recall	F1-score	Support
Chickenpox	75%	75%	75%	16
Measles	80%	86%	83%	14
Monkeypox	92%	86%	89%	42
Normal	96%	100%	98%	44
Accuracy	-	-	90%	116
Macro avg	86%	87%	86%	116
Weighted avg	90%	90%	90%	116

Πίνακας 4.9. Precision, recall και F1-score ανά κλάση για το πείραμα DenseNet121

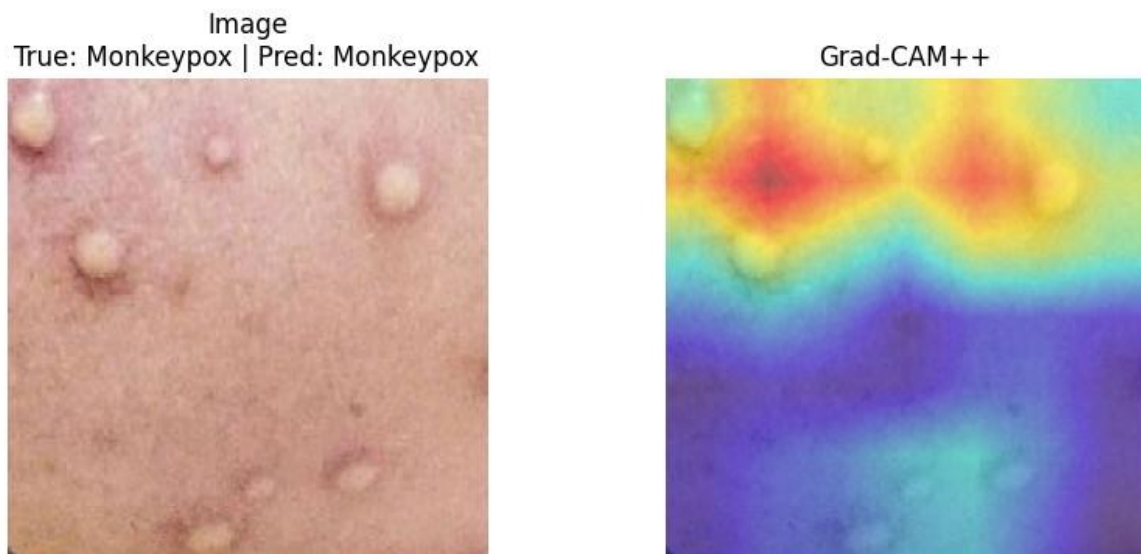
Η εικόνα αυτή επιβεβαιώνεται και από τον πίνακα σύγχυσης. Από τις 16 εικόνες Chickenpox, οι 12 ταξινομήθηκαν σωστά, ενώ 2 ταξινομήθηκαν ως Measles και 2 ως Monkeypox. Από τις 14 εικόνες Measles, οι 12 ταξινομήθηκαν σωστά, ενώ 1 ταξινομήθηκε ως Monkeypox και 1 ως Normal. Από τις 42 εικόνες Monkeypox, οι 36 ταξινομήθηκαν σωστά, ενώ 4 ταξινομήθηκαν ως Chickenpox, 1 ως Measles και 1 ως Normal. Τέλος, και οι 44 εικόνες Normal ταξινομήθηκαν σωστά. Το σημαντικότερο εύρημα είναι ότι η βασική σύγχυση εμφανίζεται ανάμεσα στις παθολογικές κατηγορίες, κυρίως μεταξύ Monkeypox και Chickenpox, ενώ η φυσιολογική κλάση αναγνωρίζεται χωρίς λάθη στο test set.

Confusion Matrix:

12	2	2	0
0	12	1	1
4	1	36	1
0	0	0	44

Πίνακας 4.10. Πίνακας σύγχυσης του DenseNet121 στο test set του Monkeypox Skin Image Dataset

Για την ερμηνεία των προβλέψεων εφαρμόστηκε αρχικά η μέθοδος Grad-CAM++. Σε μία ορθή περίπτωση Monkeypox, το μοντέλο προέβλεψε σωστά την κλάση Monkeypox με πιθανότητα 97.81%. Οι υπόλοιπες πιθανότητες ήταν αρκετά χαμηλές: Chickenpox 0.71%, Measles 1.45% και Normal 0.03%. Η υψηλή πιθανότητα δείχνει ότι το μοντέλο ήταν ιδιαίτερα βέβαιο για την πρόβλεψή του. Η Grad-CAM++ απεικόνιση επιτρέπει να εξεταστεί αν αυτή η βεβαιότητα σχετίζεται με οπτικά χαρακτηριστικά των δερματικών αλλοιώσεων και όχι με άσχετα σημεία της εικόνας.



Εικόνα 4.16. Grad-CAM++ για ορθή πρόβλεψη Monkeypox με πιθανότητα 97.81%

Στην ορθή αυτή περίπτωση, η ενεργοποίηση του Grad-CAM++ φαίνεται να συγκεντρώνεται κυρίως σε περιοχές όπου υπάρχουν εμφανείς δερματικές αλλοιώσεις. Αυτό είναι θετικό, γιατί δείχνει ότι η πρόβλεψη του μοντέλου συνδέεται με τα οπτικά στοιχεία που είναι σχετικά με την παθολογική εικόνα. Παρ' όλα αυτά, η Grad-CAM++ απεικόνιση δεν πρέπει να διαβάζεται ως ακριβής ιατρική επισημάνση της βλάβης. Λειτουργεί περισσότερο ως ένδειξη των περιοχών που επηρέασαν το μοντέλο κατά την ταξινόμηση.

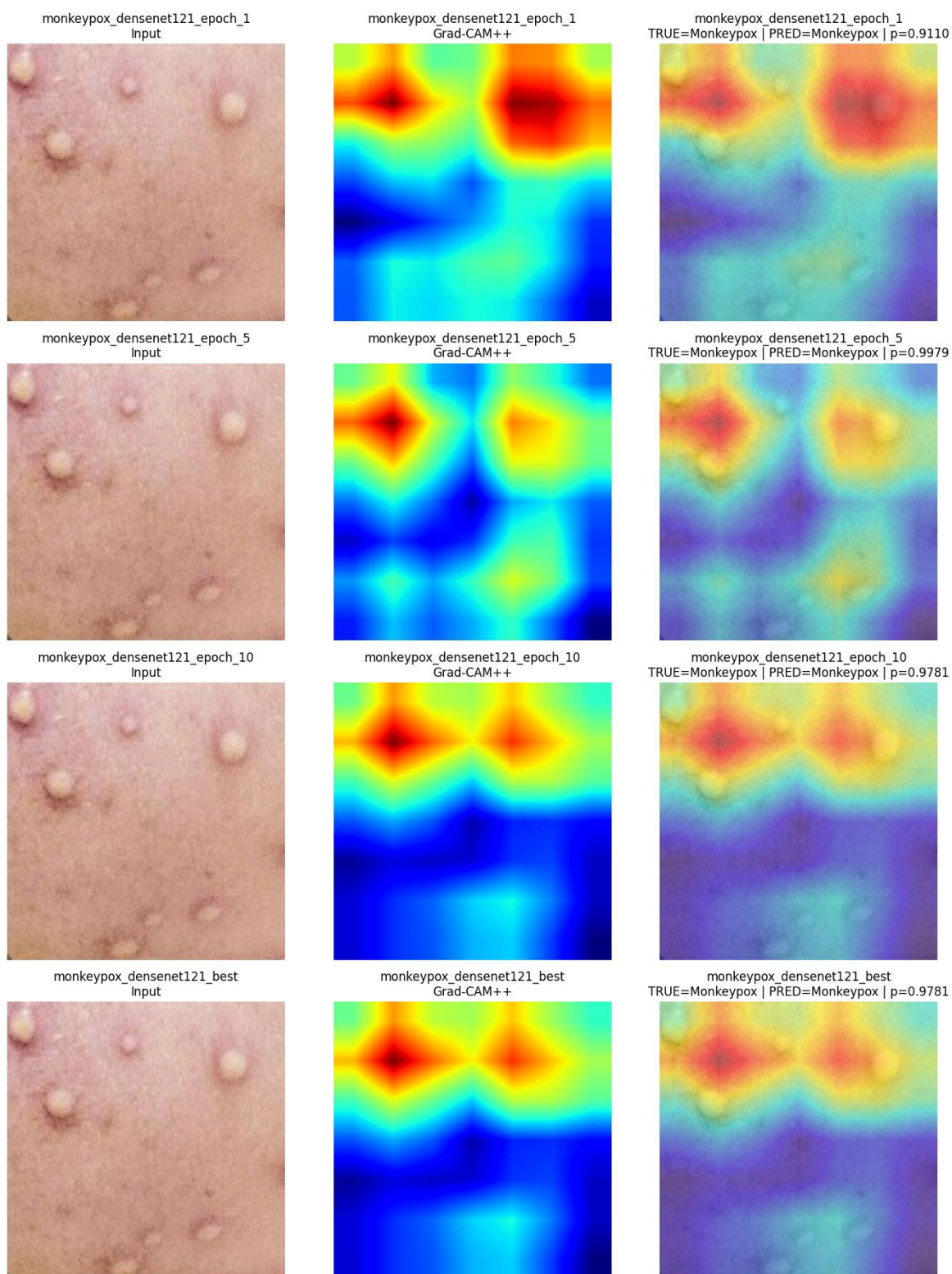
Σε μία λανθασμένη περίπτωση, η πραγματική κλάση ήταν Chickenpox, αλλά το μοντέλο προέβλεψε Monkeypox με πιθανότητα 67.41%. Η πιθανότητα για Chickenpox ήταν 27.87%, για Measles 3.63% και για Normal 1.09%. Η περίπτωση αυτή είναι ιδιαίτερα χρήσιμη για την ανάλυση, επειδή δείχνει ότι το μοντέλο δεν έκανε μια τυχαία ή εντελώς αβέβαιη πρόβλεψη, αλλά έδωσε σχετικά υψηλή πιθανότητα στην κλάση Monkeypox. Αυτό φανερώνει ότι ορισμένα οπτικά χαρακτηριστικά της εικόνας Chickenpox πιθανόν ενεργοποίησαν μοτίβα που το μοντέλο είχε συνδέσει με Monkeypox.



Εικόνα 4.17. Grad-CAM++ για λανθασμένη πρόβλεψη Chickenpox ως Monkeypox

Η λανθασμένη αυτή πρόβλεψη συνδέεται με το βασικό πρόβλημα του πειράματος: οι παθολογικές εικόνες δεν διαφέρουν πάντα καθαρά μεταξύ τους. Η σύγχυση ανάμεσα σε Chickenpox και Monkeypox φάνηκε και στον πίνακα σύγχυσης, όπου 4 εικόνες Monkeypox ταξινομήθηκαν ως Chickenpox και 2 εικόνες Chickenpox ταξινομήθηκαν ως Monkeypox. Με άλλα λόγια, το λάθος αυτό δεν είναι μεμονωμένο, αλλά αντιπροσωπεύει μια ευρύτερη δυσκολία του μοντέλου στην οριοθέτηση οπτικά συγγενικών δερματικών καταστάσεων.

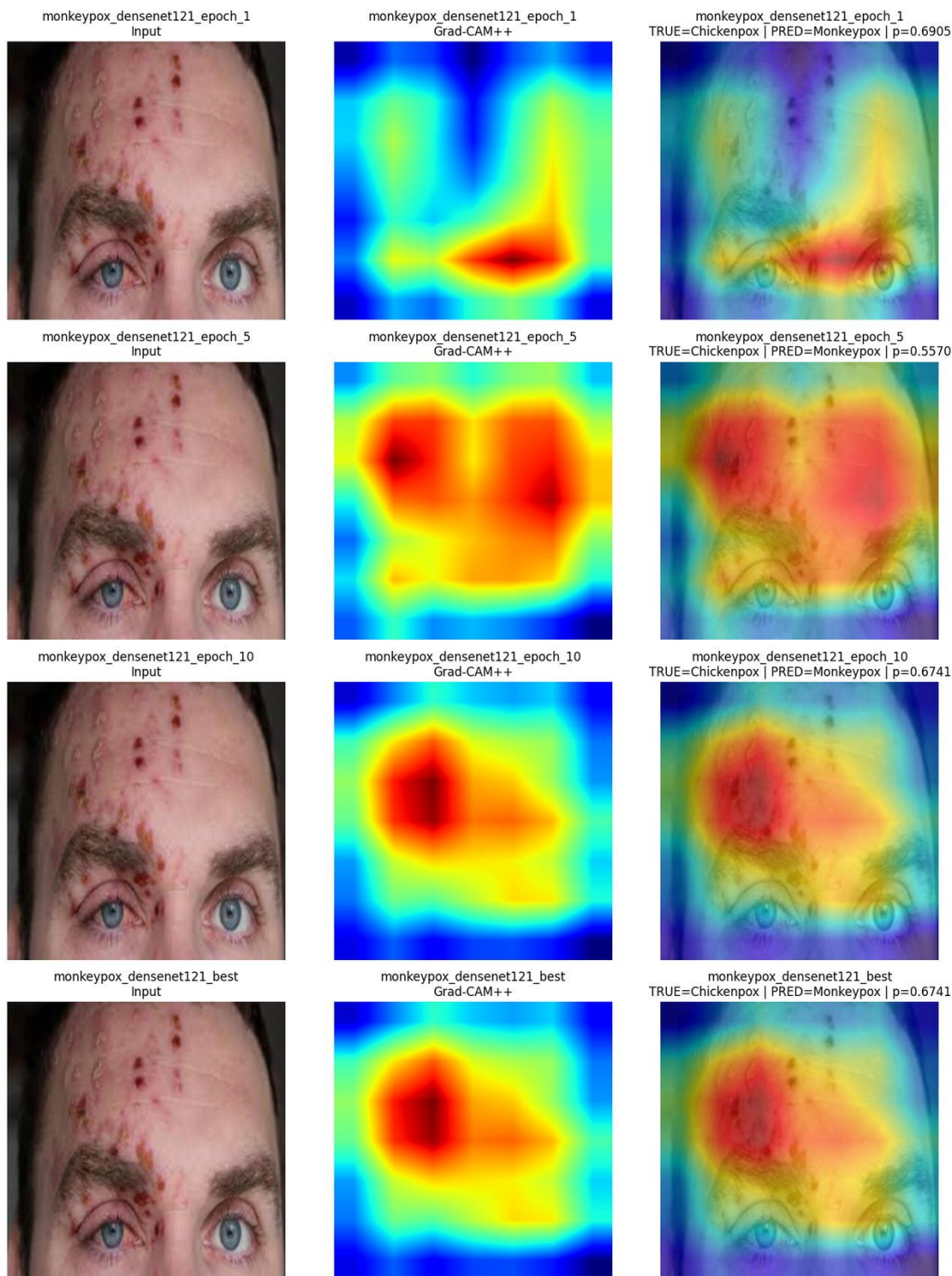
Στη συνέχεια πραγματοποιήθηκε Grad-CAM++ across-epoch ανάλυση, ώστε να εξεταστεί πώς μεταβάλλεται η οπτική εστίαση του μοντέλου στη διάρκεια της εκπαίδευσης. Η ανάλυση έγινε για checkpoints της epoch 1, epoch 5, epoch 10 και για το best model. Στην ορθή περίπτωση Monkeypox, η σύγκριση των checkpoints μπορεί να δείξει αν το μοντέλο σταδιακά μετακινεί την εστίασή του προς τις δερματικές αλλοιώσεις ή αν από νωρίς βασίζεται σε παρόμοιες περιοχές. Αυτή η across-epoch παρακολούθηση είναι χρήσιμη, γιατί δεν αξιολογεί μόνο το τελικό αποτέλεσμα, αλλά και την πορεία με την οποία το μοντέλο διαμορφώνει την απόφασή του.



Εικόνα 4.18. Grad-CAM++ across-epoch ανάλυση για ορθή πρόβλεψη Monkeypox

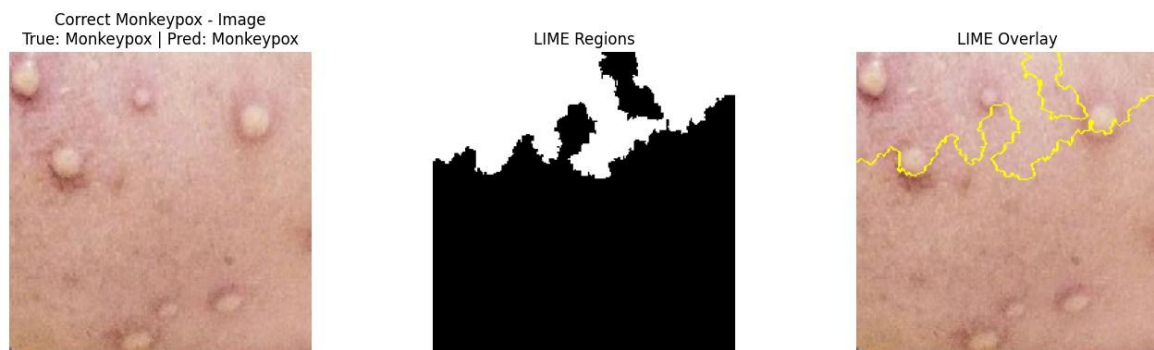
Η ίδια διαδικασία εφαρμόστηκε και στη λανθασμένη περίπτωση Chickenpox → Monkeypox. Σε αυτή την περίπτωση, η across-epoch ανάλυση είναι ακόμη πιο σημαντική,

γιατί βοηθά να φανεί αν το λάθος δημιουργείται σταδιακά, δηλαδή αν το μοντέλο από κάποια εποχή και μετά αρχίζει να δίνει μεγαλύτερη έμφαση σε περιοχές που το οδηγούν προς την κλάση Monkeypox. Αν οι ενεργοποιήσεις παραμένουν σταθερά πάνω σε παθολογικές περιοχές αλλά η πρόβλεψη εξακολουθεί να είναι λανθασμένη, τότε το πρόβλημα δεν είναι απλώς ότι το μοντέλο κοιτάζει άσχετα σημεία, αλλά ότι δυσκολεύεται να ερμηνεύσει σωστά τα οπτικά χαρακτηριστικά της βλάβης.



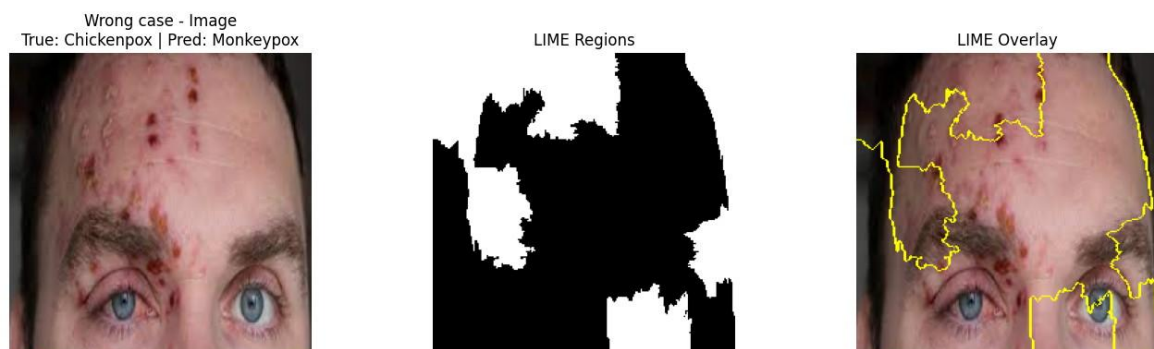
Εικόνα 4.19. Grad-CAM++ across-epoch ανάλυση για λανθασμένη πρόβλεψη Chickenpox ως Monkeypox

Συμπληρωματικά εφαρμόστηκε και η μέθοδος LIME στις ίδιες χαρακτηριστικές περιπτώσεις. Στην ορθή πρόβλεψη Monkeypox, το LIME ανέδειξε περιοχές που σχετίζονται με τις δερματικές αλλοιώσεις και υποστηρίζουν την τελική απόφαση του μοντέλου. Η σύγκριση με το Grad-CAM++ και το SHAP είναι χρήσιμη, επειδή οι τρεις μέθοδοι δεν εξηγούν την απόφαση με τον ίδιο ακριβώς τρόπο. Το Grad-CAM++ δείχνει ευρύτερες περιοχές ενεργοποίησης, το SHAP εξετάζει τοπικές συνεισφορές, ενώ το LIME απομονώνει περιοχές της εικόνας που φαίνεται να επηρεάζουν την πρόβλεψη.



Εικόνα 4.20. LIME ανάλυση για ορθή πρόβλεψη Monkeypox

Στη λανθασμένη πρόβλεψη Chickenpox ως Monkeypox, το LIME βοηθά να ελεγχθεί αν η απόφαση του μοντέλου βασίστηκε σε συγκεκριμένες αλλοιώσεις ή σε πιο διάσπαρτα σημεία της εικόνας. Αν οι περιοχές που αναδεικνύονται βρίσκονται πάνω στις βλάβες, τότε το λάθος φαίνεται να οφείλεται περισσότερο σε μορφολογική ομοιότητα μεταξύ των κλάσεων. Αντίθετα, αν οι περιοχές είναι άσχετες με τις αλλοιώσεις, τότε το λάθος θα μπορούσε να συνδέεται με θόρυβο, φόντο ή τεχνικά στοιχεία της εικόνας.



Εικόνα 4.21. LIME ανάλυση για λανθασμένη πρόβλεψη Chickenpox ως Monkeypox

Συνολικά, το πείραμα με το DenseNet121 στο Monkeypox Skin Image Dataset έδειξε ικανοποιητική απόδοση στην πολυταξική ταξινόμηση δερματικών εικόνων. Το Test Accuracy = 89.66% και το Test Weighted F1 = 89.60% δείχνουν ότι το μοντέλο είχε καλή συνολική συμπεριφορά, ενώ το Test Macro F1 = 86.11% δείχνει ότι η απόδοση μειώνεται

όταν όλες οι κλάσεις υπολογίζονται ισότιμα. Η κλάση Normal αναγνωρίστηκε με πολύ υψηλή ακρίβεια, ενώ η μεγαλύτερη δυσκολία εμφανίστηκε στις οπτικά συγγενικές παθολογικές κλάσεις, κυρίως ανάμεσα σε Monkeyrox και Chickenrox.

Οι μέθοδοι Grad-CAM++, SHAP και LIME πρόσθεσαν ένα δεύτερο επίπεδο αξιολόγησης. Δεν περιορίστηκαν στην απλή επιβεβαίωση των αριθμητικών αποτελεσμάτων, αλλά επέτρεψαν να εξεταστεί πού εστιάζει το μοντέλο και πώς μπορούν να ερμηνευθούν τόσο οι σωστές όσο και οι λανθασμένες προβλέψεις. Το σημαντικότερο εύρημα είναι ότι τα λάθη δεν φαίνεται να οφείλονται απαραίτητα σε τυχαία εστίαση, αλλά συχνά σε οπτική ομοιότητα ανάμεσα στις παθολογικές κατηγορίες. Αυτό δείχνει ότι, για τέτοιου τύπου προβλήματα, η υψηλή ακρίβεια πρέπει πάντα να συνοδεύεται από ποιοτική ανάλυση των προβλέψεων και των εξηγήσεων.

4.5 T-SNE και UMAP στο HAM10000 με EfficientNetV2-S UMAP, Grad-CAM++ και SHAP

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο EfficientNetV2-S για την πολυταξική ταξινόμηση δερματικών βλαβών στο σύνολο δεδομένων HAM10000. Σε αντίθεση με τα δυαδικά πειράματα, εδώ το μοντέλο δεν έπρεπε απλώς να διαχωρίσει το melanoma από τις υπόλοιπες βλάβες, αλλά να ταξινομήσει κάθε εικόνα σε μία από τις επτά διαγνωστικές κατηγορίες του dataset: MEL, NV, BCC, AKIEC, BKL, DF και VASC. Το πείραμα αυτό έχει ιδιαίτερη σημασία, επειδή συνδυάζει τρία επίπεδα αξιολόγησης. Πρώτα εξετάζεται η αριθμητική απόδοση του μοντέλου μέσα από τις κλασικές μετρικές ταξινόμησης. Στη συνέχεια, χρησιμοποιούνται οι μέθοδοι t-SNE και UMAP, ώστε να φανεί πώς οργανώνονται οι εσωτερικές αναπαραστάσεις του EfficientNetV2-S. Τέλος, εφαρμόζονται Grad-CAM++ και SHAP, ώστε να εξεταστεί οπτικά πού εστιάζει το μοντέλο και ποια σημεία της εικόνας επηρεάζουν περισσότερο την πρόβλεψη. Το αρχικό σύνολο περιλάμβανε 10.015 usable rows. Η κατανομή των κλάσεων ήταν έντονα ανισόρροπη: MEL = 1.113, NV = 6.705, BCC = 514, AKIEC = 327, BKL = 1.099, DF = 115 και VASC = 142. Η κλάση NV ήταν με μεγάλη διαφορά η πολυπληθέστερη, ενώ οι κλάσεις DF και VASC είχαν πολύ μικρό αριθμό δειγμάτων. Για αυτόν τον λόγο χρησιμοποιήθηκαν class weights, ώστε το μοντέλο να μη μάθει απλώς να ευνοεί τις μεγάλες κλάσεις.

Κλάση	Πλήθος εικόνων	Ποσοστό
MEL	1.113	11.11%
NV	6.705	66.95%
BCC	514	5.13%
AKIEC	327	3.27%
BKL	1.099	10.97%
DF	115	1.15%
VASC	142	1.42%
Σύνολο	10.015	100.00%

Πίνακας 4.11. Κατανομή των κλάσεων στο HAM10000 για το πείραμα EfficientNetV2-S

Το dataset χωρίστηκε σε 7.010 εικόνες εκπαίδευσης, 1.502 εικόνες επικύρωσης και 1.503 εικόνες ελέγχου. Το μοντέλο εκπαιδεύτηκε για 10 εποχές και το καλύτερο checkpoint επιλέχθηκε με βάση το validation AUC. Η καλύτερη εποχή ήταν η 9η, όπου το validation AUC έφτασε στο 0.9682, με val_acc = 76.17% και val_macro_f1 = 76.66%. Η πορεία της εκπαίδευσης δείχνει ότι το μοντέλο βελτιώθηκε σταδιακά μέχρι την 9η εποχή. Στην 1η εποχή είχε val_acc = 43.87%, val_macro_f1 = 39.73% και val_auc = 90.47%. Στην 3η εποχή οι τιμές είχαν ανέβει σε val_acc = 67.58%, val_macro_f1 = 64.20% και val_auc = 94.91%. Η καλύτερη συνολική διαχωριστική ικανότητα καταγράφηκε στην 9η εποχή, ενώ στη 10η εποχή η απόδοση μειώθηκε, με val_auc = 94.26% και val_macro_f1 = 60.65%. Αυτό δείχνει ότι η περαιτέρω εκπαίδευση δεν οδήγησε σε καλύτερη γενίκευση.

Epoch	Train loss	Val loss	Val accuracy (%)	Val macro F1 (%)	Val AUC (%)
1	0.3654	0.8671	43.87%	39.73%	90.47%
3	0.0957	0.6539	67.58%	64.20%	94.91%
5	0.0541	0.5979	68.31%	63.58%	95.57%
7	0.0477	0.5745	72.90%	72.84%	96.52%
8	0.0332	0.5636	77.76%	76.63%	96.80%
9	0.0205	0.5838	76.17%	76.66%	96.82%
10	0.0847	1.0684	68.31%	60.65%	94.26%

Πίνακας 4.12. Πορεία εκπαίδευσης του EfficientNetV2-S στο HAM10000

Στο test set, το καλύτερο μοντέλο της 9ης εποχής πέτυχε Test Accuracy = 75.91%, Test Macro F1 = 77.05%, Test Weighted F1 = 77.75%, Test Macro Precision = 73.66%, Test Macro Recall = 84.06% και Test AUC OVR = 96.64%. Η υψηλή τιμή AUC δείχνει ότι το μοντέλο έχει ισχυρή διαχωριστική ικανότητα, ενώ η χαμηλότερη accuracy δείχνει ότι η τελική επιλογή της σωστής κλάσης παραμένει δύσκολη σε ορισμένες εικόνες. Αυτό είναι αναμενόμενο, επειδή αρκετές δερματικές βλάβες έχουν οπτικά κοινά στοιχεία.

Metric	Value
Best epoch	9
Test Loss	0.5854
Test Accuracy	75.91%
Test Macro Precision	73.66%
Test Macro Recall	84.06%
Test Macro F1	77.05%
Test Weighted F1	77.75%
Test AUC OVR	96.64%

Πίνακας 4.13. Test αποτελέσματα του EfficientNetV2-S στο HAM10000.

Η ανάλυση ανά κλάση δείχνει μια πιο καθαρή εικόνα της συμπεριφοράς του μοντέλου. Η κλάση VASC είχε πολύ υψηλή απόδοση, με precision = 90.91%, recall = 90.91% και F1-

score = 90.91%. Η κλάση DF είχε επίσης ισχυρή επίδοση, με F1-score = 88.89%, αν και το πλήθος των δειγμάτων της ήταν μικρό. Η κλάση NV είχε υψηλό precision, 96.79%, αλλά χαμηλότερο recall, 71.97%, επειδή αρκετά NV ταξινομήθηκαν ως άλλες κατηγορίες, κυρίως ως MEL.

Η κλάση MEL είναι ιδιαίτερα σημαντική για την εργασία. Το μοντέλο πέτυχε recall = 85.63%, δηλαδή εντόπισε 143 από τα 167 πραγματικά melanomas. Από την άλλη πλευρά, το precision ήταν 41.69%, επειδή αρκετές εικόνες άλλων κλάσεων προβλέφθηκαν ως MEL. Αυτό δείχνει μια πιο “ευαίσθητη” συμπεριφορά απέναντι στο melanoma: το μοντέλο χάνει σχετικά λίγα πραγματικά melanomas, αλλά δημιουργεί αρκετά false positives.

Κλάση	Precision (%)	Recall (%)	F1-score (%)	Support
MEL	41.69%	85.63%	56.08%	167
NV	96.79%	71.97%	82.55%	1006
BCC	67.96%	90.91%	77.78%	77
AKIEC	74.00%	75.51%	74.75%	49
BKL	60.09%	79.39%	68.41%	165
DF	84.21%	94.12%	88.89%	17
VASC	90.91%	90.91%	90.91%	22
Accuracy			75.91%	1503
Macro avg	73.66%	84.06%	77.05%	1503
Weighted avg	84.19%	75.91%	77.75%	1503

Πίνακας 4.14. Απόδοση ανά κλάση για το EfficientNetV2-S στο HAM10000

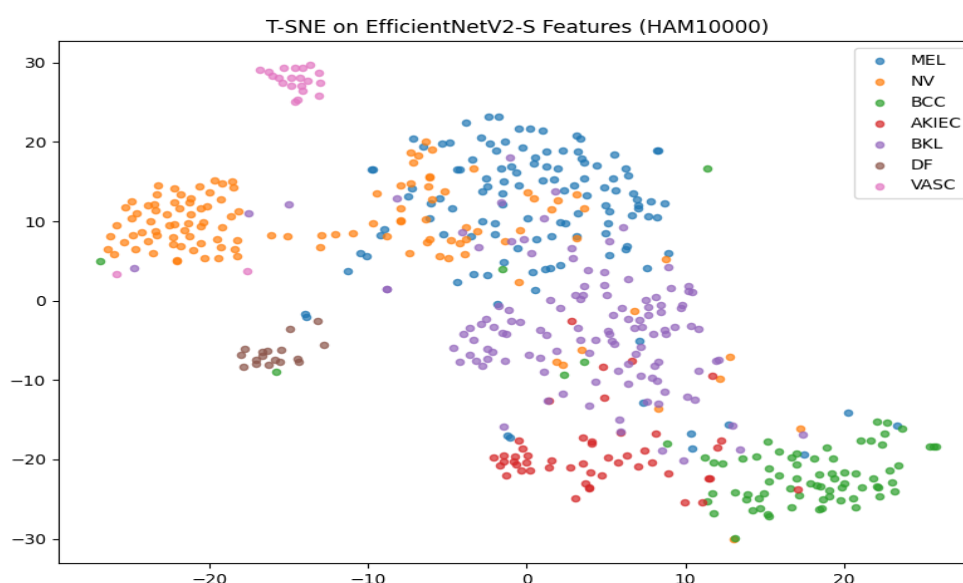
Ο πίνακας σύγχυσης επιβεβαιώνει αυτή την παρατήρηση. Από τα 167 πραγματικά MEL, τα 143 ταξινομήθηκαν σωστά, ενώ 14 ταξινομήθηκαν ως NV, 6 ως BKL, 3 ως BCC και 1 ως AKIEC. Για την κλάση NV, από τις 1.006 εικόνες, οι 724 ταξινομήθηκαν σωστά, αλλά 181 προβλέφθηκαν λανθασμένα ως MEL. Αυτό είναι το πιο σημαντικό σημείο του confusion matrix, γιατί δείχνει ότι το μοντέλο συχνά μεταφέρει benign σπίλους προς την κλάση melanoma. Στην πράξη, αυτή η συμπεριφορά μειώνει τον κίνδυνο να χαθεί melanoma, αλλά αυξάνει την πιθανότητα υπερδιάγνωσης.

Πραγματική κλάση / Πρόβλεψη	MEL	NV	BCC	AKIEC	BKL	DF	VASC	Σύνολο
MEL	143	14	3	1	6	0	0	167
NV	181	724	19	4	74	2	2	1006
BCC	1	2	70	1	3	0	0	77
AKIEC	0	0	8	37	4	0	0	49
BKL	18	6	3	7	131	0	0	165
DF	0	1	0	0	0	16	0	17
VASC	0	1	0	0	0	1	20	22

Πραγματική κλάση / Πρόβλεψη	MEL	NV	BCC	AKIEC	BKL	DF	VASC	Σύνολο
Σύνολο	343	748	103	50	218	19	22	1503

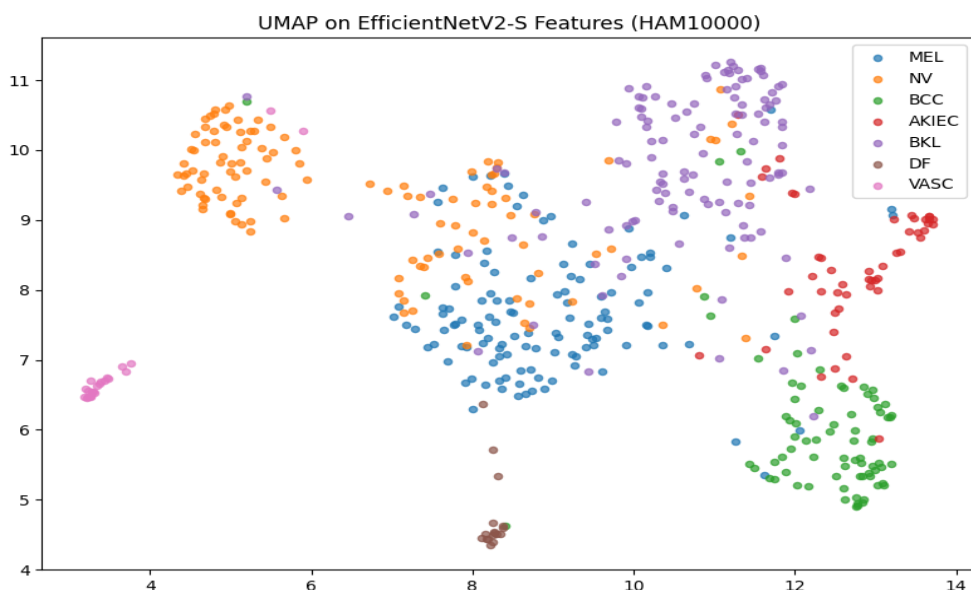
Πίνακας 4.15. Πίνακας σύγχυσης του EfficientNetV2-S στο test set

Στη συνέχεια έγινε εξαγωγή χαρακτηριστικών από το εκπαιδευμένο EfficientNetV2-S. Τα χαρακτηριστικά είχαν διάσταση 1280 ανά εικόνα. Για την οπτικοποίηση χρησιμοποιήθηκε υποσύνολο 525 εικόνων, ενώ πριν από το t-SNE και το UMAP εφαρμόστηκε PCA σε 50 διαστάσεις, με συνολική εξηγούμενη διακύμανση περίπου 72.04%. Αυτή η διαδικασία βοήθησε να μειωθεί η πολυπλοκότητα των χαρακτηριστικών, χωρίς να χαθεί μεγάλο μέρος της πληροφορίας.



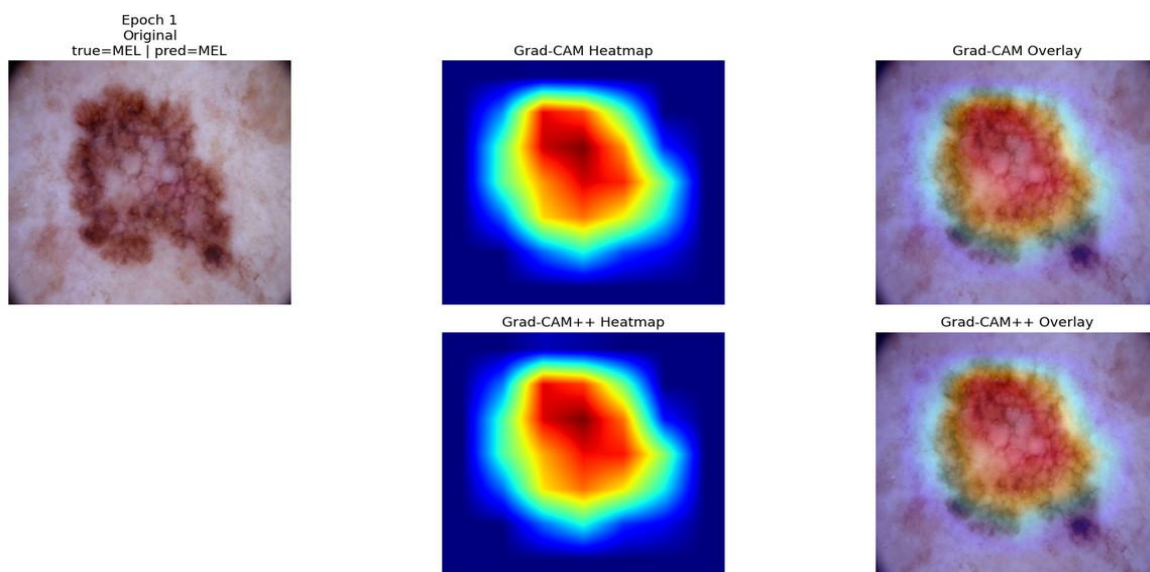
Εικόνα 4.22. t-SNE απεικόνιση των χαρακτηριστικών του EfficientNetV2-S στο HAM10000

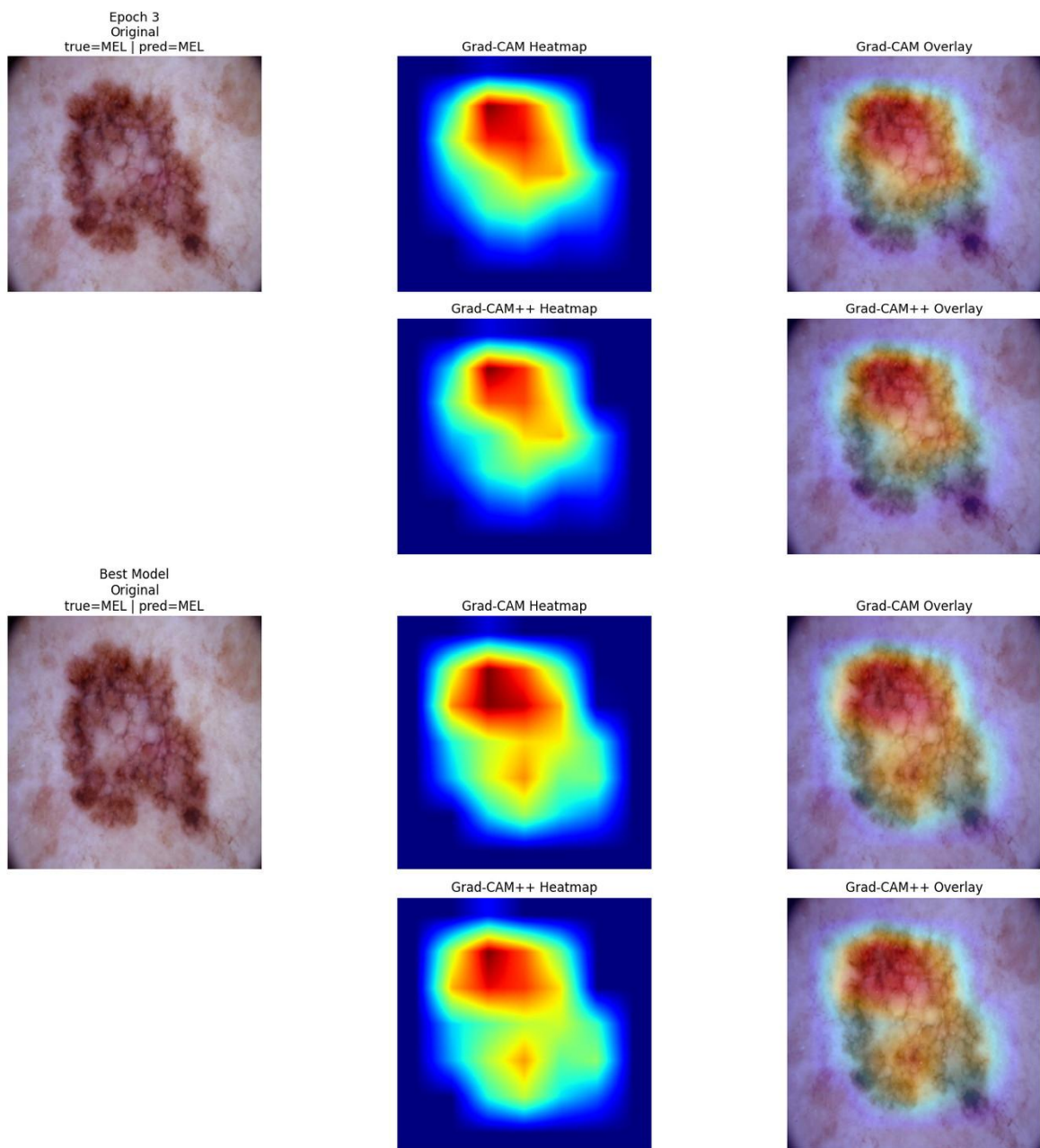
Η t-SNE απεικόνιση δείχνει ότι το μοντέλο έχει μάθει κάποια δομή στο feature space, αλλά οι κλάσεις δεν διαχωρίζονται απόλυτα. Ορισμένες κατηγορίες, όπως VASC και DF, εμφανίζονται πιο συμπαγείς σε συγκεκριμένες περιοχές, κάτι που συμφωνεί με την καλή απόδοση που είχαν στο classification report. Αντίθετα, οι κλάσεις MEL, NV και BKL εμφανίζουν μεγαλύτερη επικάλυψη. Αυτό είναι αναμενόμενο, επειδή οι κατηγορίες αυτές μπορεί να μοιράζονται χρωματικά και μορφολογικά χαρακτηριστικά. Με άλλα λόγια, το t-SNE δείχνει ότι η δυσκολία του μοντέλου δεν είναι τυχαία, αλλά σχετίζεται με την ίδια τη δομή των χαρακτηριστικών.

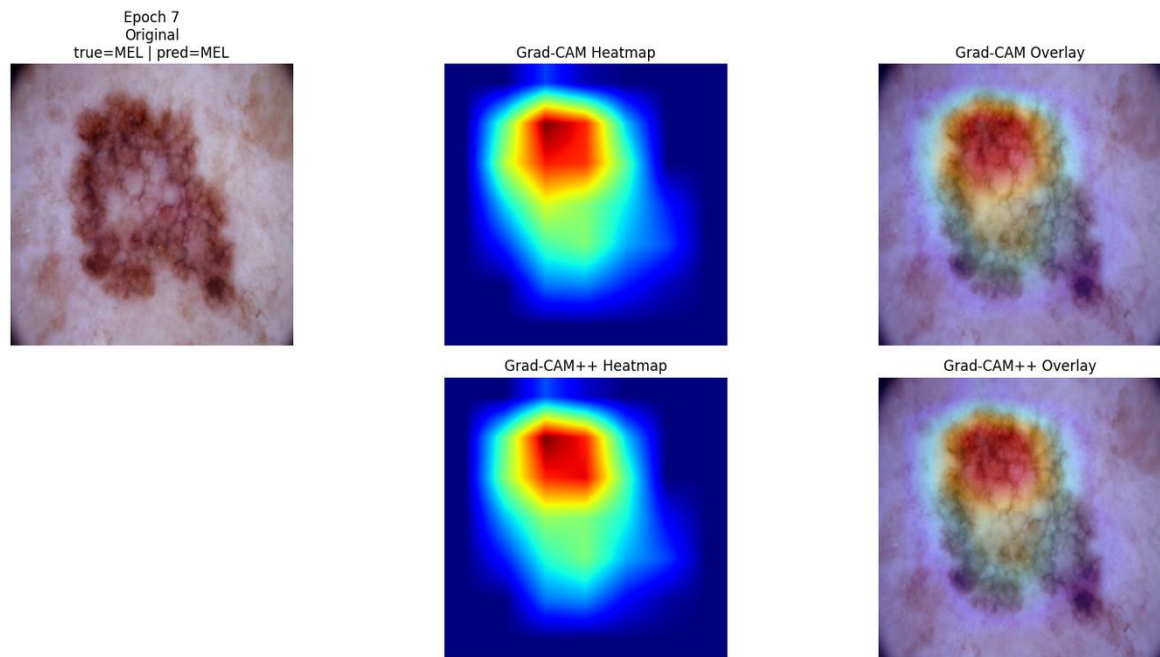


Εικόνα 4.23. UMAP απεικόνιση των χαρακτηριστικών του EfficientNetV2-S στο HAM10000

Η UMAP απεικόνιση δίνει παρόμοια αλλά λίγο πιο καθαρή εικόνα της οργάνωσης των κλάσεων. Οι VASC εικόνες εμφανίζονται αρκετά απομονωμένες, ενώ τα DF δείγματα σχηματίζουν μικρή ξεχωριστή ομάδα. Οι κλάσεις MEL, NV και BKL παραμένουν πιο αναμειγμένες. Αυτό στηρίζει την παρατήρηση ότι η δυσκολία δεν αφορά όλες τις κλάσεις στον ίδιο βαθμό. Το μοντέλο διαχωρίζει καλύτερα ορισμένες μικρές αλλά οπτικά πιο χαρακτηριστικές κατηγορίες, ενώ δυσκολεύεται περισσότερο στις μελανοκυτταρικές και κερατωσικές βλάβες. Για την ερμηνευτική ανάλυση του μοντέλου εφαρμόστηκαν Grad-CAM++ και SHAP στα checkpoints των εποχών 1, 3, 7 και στο best model. Η across-epoch ανάλυση είναι χρήσιμη, γιατί δεν δείχνει μόνο την τελική απόφαση του μοντέλου, αλλά και το πώς μεταβάλλεται η εστίασή του καθώς προχωρά η εκπαίδευση.

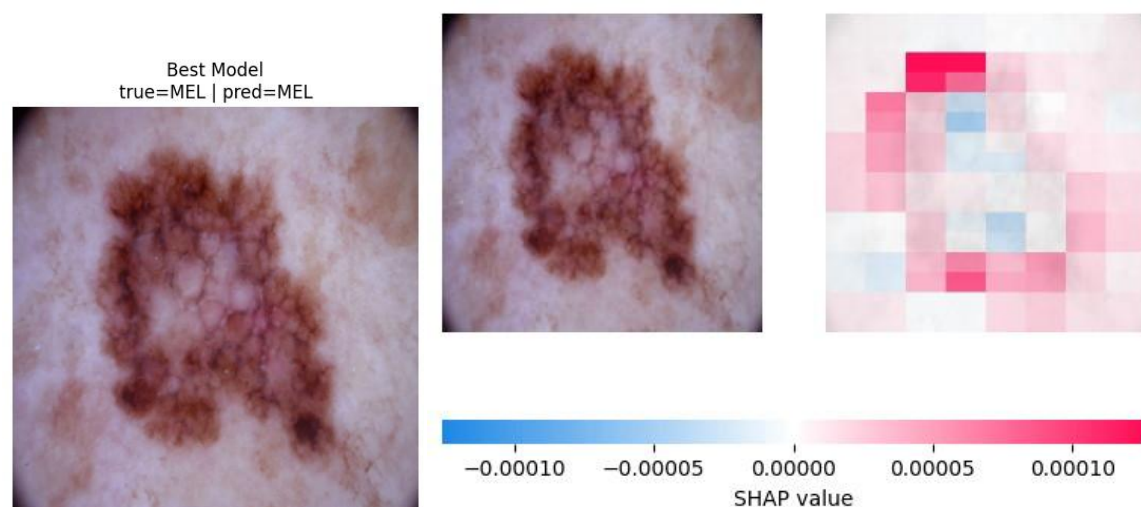






Εικόνα 4.24. Grad-CAM και Grad-CAM++ across-epoch ανάλυση για χαρακτηριστική περίπτωση του HAM10000

Στην Grad-CAM++ ανάλυση, το ενδιαφέρον δεν είναι μόνο αν η τελική πρόβλεψη είναι σωστή, αλλά και αν η εστίαση του μοντέλου μετακινείται προς τη βλάβη όσο βελτιώνεται η εκπαίδευση. Στις αρχικές εποχές, η ενεργοποίηση μπορεί να είναι πιο διάχυτη ή να καλύπτει περιοχές που δεν σχετίζονται αποκλειστικά με τη βλάβη. Σε μεταγενέστερα checkpoints, όταν το μοντέλο έχει μάθει πιο σταθερά χαρακτηριστικά, η ενεργοποίηση αναμένεται να γίνεται πιο συνδεδεμένη με την περιοχή της αλλοίωσης. Η Grad-CAM++ δεν πρέπει να διαβαστεί ως segmentation mask. Δείχνει περιοχές που επηρέασαν την πρόβλεψη, όχι τα ακριβή όρια της βλάβης.



Εικόνα 4.25. SHAP ανάλυση για χαρακτηριστική πρόβλεψη του EfficientNetV2-S

Η SHAP ανάλυση λειτουργεί συμπληρωματικά. Ενώ η Grad-CAM++ δείχνει ευρύτερες περιοχές ενεργοποίησης, το SHAP αναδεικνύει πιο τοπικές συνεισφορές. Στα SHAP outputs του πειράματος, οι θετικές περιοχές εμφανίζονται κυρίως πάνω ή κοντά στη βλάβη, δείχνοντας ποια τμήματα της εικόνας ενισχύουν την πρόβλεψη της αντίστοιχης κλάσης. Σε ένα πολυταξικό πρόβλημα όπως αυτό, η SHAP ανάλυση είναι χρήσιμη γιατί επιτρέπει να δούμε αν η πρόβλεψη βασίζεται σε οπτικά σχετικά στοιχεία ή αν επηρεάζεται από περιοχές του φόντου. Συνολικά, το πείραμα με EfficientNetV2-S στο HAM10000 έδειξε ότι το μοντέλο είχε υψηλή διαχωριστική ικανότητα, με Test AUC OVR = 96.64%, αλλά η τελική ακρίβεια παρέμεινε στο 75.91% λόγω των συγχύσεων ανάμεσα σε οπτικά κοντινές κλάσεις. Το macro recall = 84.06% δείχνει ότι το μοντέλο αναγνώρισε μεγάλο ποσοστό δειγμάτων από όλες τις κλάσεις, ενώ το macro F1 = 77.05% δίνει πιο ισορροπημένη εικόνα της συνολικής απόδοσης.

Το πιο κρίσιμο εύρημα αφορά το melanoma. Το μοντέλο είχε υψηλό recall για την κλάση MEL, 85.63%, αλλά χαμηλό precision, 41.69%. Αυτό σημαίνει ότι εντόπισε τα περισσότερα πραγματικά melanomas, αλλά ταξινομήσε και αρκετές άλλες βλάβες ως melanoma. Η συμπεριφορά αυτή είναι σημαντική για ιατρικά προβλήματα, γιατί δείχνει ένα μοντέλο πιο ευαίσθητο παρά αυστηρό. Με ό,τι αυτό συνεπάγεται, μια τέτοια συμπεριφορά μπορεί να είναι χρήσιμη για αρχικό έλεγχο, αλλά δεν επαρκεί για αυτόνομη διάγνωση. Η t-SNE και η UMAP ανάλυση έδειξαν ότι οι εσωτερικές αναπαραστάσεις του EfficientNetV2-S οργανώνονται μερικώς ανά κλάση, αλλά παραμένει επικάλυψη ανάμεσα σε MEL, NV και BKL. Οι Grad-CAM++ και SHAP εικόνες πρόσθεσαν ένα πιο ερμηνευτικό επίπεδο, δείχνοντας πώς το μοντέλο εστιάζει σε συγκεκριμένες περιοχές της εικόνας. Άρα το πείραμα αυτό δεν δίνει μόνο τελικές μετρικές, αλλά βοηθά να κατανοηθεί καλύτερα πού και γιατί το μοντέλο δυσκολεύεται.

4.6 Score-CAM με ensemble μοντέλων στο σύνολο δεδομένων Skin Cancer 7 κλάσεων

Στο πείραμα αυτό αξιολογήθηκε μία προσέγγιση πολυταξικής ταξινόμησης δερματικών βλαβών με χρήση ensemble μοντέλων και εξηγησιμότητας μέσω Score-CAM. Σε αντίθεση με προηγούμενα πειράματα, όπου χρησιμοποιήθηκε ένα μόνο μοντέλο, εδώ εκπαιδεύτηκαν και συνδυάστηκαν τρεις διαφορετικές αρχιτεκτονικές: DenseNet121, EfficientNetV2-S και SE-ResNeXt50 32x4d. Ο στόχος ήταν να εξεταστεί αν ο συνδυασμός διαφορετικών backbones μπορεί να προσφέρει πιο σταθερή πρόβλεψη, αλλά και πώς διαμορφώνεται η οπτική εστίαση του κάθε μοντέλου στις ίδιες εικόνες. Το dataset που χρησιμοποιήθηκε ήταν το marmal88/skin_cancer, με επτά διαγνωστικές κλάσεις: actinic keratoses, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanocytic nevi, melanoma και vascular lesions. Χρησιμοποιήθηκε υποσύνολο 5.000 εικόνων εκπαίδευσης και 1.200 εικόνων επικύρωσης, ενώ η τελική δυαδική ανάλυση melanoma / non-melanoma πραγματοποιήθηκε στο test set των 1.285 εικόνων. Η εκπαίδευση έγινε με εικόνες μεγέθους 224×224 pixels. Η κατανομή των κλάσεων στο training subset ήταν έντονα ανισόρροπη. Συγκεκριμένα, υπήρχαν 3.405 εικόνες melanocytic nevi, ενώ οι μικρότερες κλάσεις, όπως το dermatofibroma και τα vascular lesions, είχαν μόνο 56 και 69 εικόνες αντίστοιχα. Για να περιοριστεί η επίδραση αυτής της ανισορροπίας, χρησιμοποιήθηκαν class weights στη

συνάρτηση απώλειας και WeightedRandomSampler κατά τη δειγματοληψία των batches. Με αυτόν τον τρόπο δόθηκε μεγαλύτερη βαρύτητα στις μικρότερες κλάσεις, ώστε το μοντέλο να μην ευνοεί μόνο την πολυπληθέστερη κατηγορία.

actinic keratoses	basal cell carcinoma	benign keratosis- like lesions	dermatofibroma	melanocytic nevi	melanoma	vascular lesions
156	251	515	56	3405	548	69

Πίνακας 4.16. Κατανομή κλάσεων στο training subset του πειράματος Score-CAM

Κάθε μοντέλο εκπαιδεύτηκε για 6 εποχές. Ως βασική μετρική επιλογής του καλύτερου checkpoint χρησιμοποιήθηκε το validation macro recall, επειδή το dataset είναι ανισόρροπο και ενδιαφέρει η ικανότητα αναγνώρισης όλων των κλάσεων, όχι μόνο των πολυπληθέστερων. Το DenseNet121 πέτυχε την καλύτερη τιμή του στην 4η εποχή, με Val Macro Recall = 82.83%. Το EfficientNetV2-S είχε την καλύτερη επίδοση στην 5η εποχή, με Val Macro Recall = 88.63%. Το SE-ResNeXt50 32x4d είχε επίσης την καλύτερη επίδοσή του στην 5η εποχή, με Val Macro Recall = 87.05%.

DenseNet121	EfficientNetV2-S	SE-ResNeXt50
82.83%	88.63%	87.05%

Πίνακας 4.17. Καλύτερη validation επίδοση ανά backbone με βάση το macro recall

Στη συνέχεια, τα τρία μοντέλα συνδυάστηκαν σε ensemble με σταθμισμένο μέσο όρο των πιθανοτήτων τους. Τα βάρη του ensemble υπολογίστηκαν με βάση την καλύτερη τιμή validation macro recall κάθε μοντέλου. Συγκεκριμένα, το DenseNet121 είχε βάρος 0.3204, το EfficientNetV2-S βάρος 0.3429 και το SE-ResNeXt50 βάρος 0.3367. Τα βάρη είναι αρκετά κοντά μεταξύ τους, κάτι που δείχνει ότι και τα τρία μοντέλα συνέβαλαν ουσιαστικά στην τελική απόφαση του ensemble.

Στο validation set, το ensemble πέτυχε accuracy = 72.08%, macro precision = 64.39%, macro recall = 89.00%, macro F1 = 71.62% και weighted F1 = 74.48%. Η υψηλή τιμή του macro recall δείχνει ότι το ensemble αναγνώρισε μεγάλο ποσοστό των δειγμάτων από κάθε κλάση. Από την άλλη πλευρά, η χαμηλότερη τιμή του macro precision δείχνει ότι υπήρξαν αρκετές λανθασμένες προβλέψεις προς ορισμένες μικρότερες ή πιο δύσκολες κλάσεις. Αυτό είναι ένα συνηθισμένο αποτέλεσμα όταν χρησιμοποιούνται class weights και balanced sampling: το μοντέλο γίνεται πιο ευαίσθητο στις μικρότερες κλάσεις, αλλά μπορεί να αυξήσει τα false positives.

Συνολικές validation μετρικές του ensemble:

Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
72.08%	64.39%	89.00%	71.62%	74.48%

Πίνακας 4.18. Συνολικές μετρικές του ensemble στο validation set

Αναλυτικότερα, στο validation classification report η κλάση melanoma είχε precision = 35.07%, recall = 92.37% και F1-score = 50.84%. Η τιμή recall είναι πολύ υψηλή, γεγονός που δείχνει ότι το ensemble εντόπισε τα περισσότερα πραγματικά melanomas στο validation set. Η χαμηλή precision όμως δείχνει ότι αρκετές εικόνες άλλων κατηγοριών προβλέφθηκαν λανθασμένα ως melanoma. Αυτή η συμπεριφορά είναι ιδιαίτερα σημαντική: το μοντέλο γίνεται αρκετά ευαίσθητο απέναντι στο melanoma, αλλά αυτό συνοδεύεται από αύξηση των false positive προβλέψεων.

Αντίστοιχα, η κλάση melanocytic nevi είχε precision = 100%, αλλά recall = 62.92%. Αυτό σημαίνει ότι όταν το ensemble προέβλεπε την κλάση nevi, ήταν σχεδόν πάντα σωστό, αλλά δεν αναγνώρισε όλα τα πραγματικά nevi ως nevi. Πολλά από αυτά μετακινήθηκαν προς άλλες κλάσεις, και κυρίως προς την κλάση melanoma. Αυτό δείχνει ξανά την τάση του μοντέλου να προτιμά πιο «ύποπτες» προβλέψεις σε ορισμένες περιπτώσεις.

Classification Report	Precision	Recall	F1-score	Support
Akiec	69.81%	92.5%	79.57%	40
Bcc	69.86%	85%	76.69%	60
Bkl	71.43%	90.23%	79.73%	133
Df	54.55%	100%	70.59%	12
Mel	35.07%	92.37%	50.84%	131
Nv	100%	62.92%	77.24%	809
Vasc	50%	100%	66.67%	15
Macro avg	64.39%	89%	71.62%	1200
Weighted avg	86.15%	72.08%	74.48%	1200
Accuracy	-	-	72.08%	1200

Πίνακας 4.19. Precision, recall και F1-score ανά κλάση για το ensemble

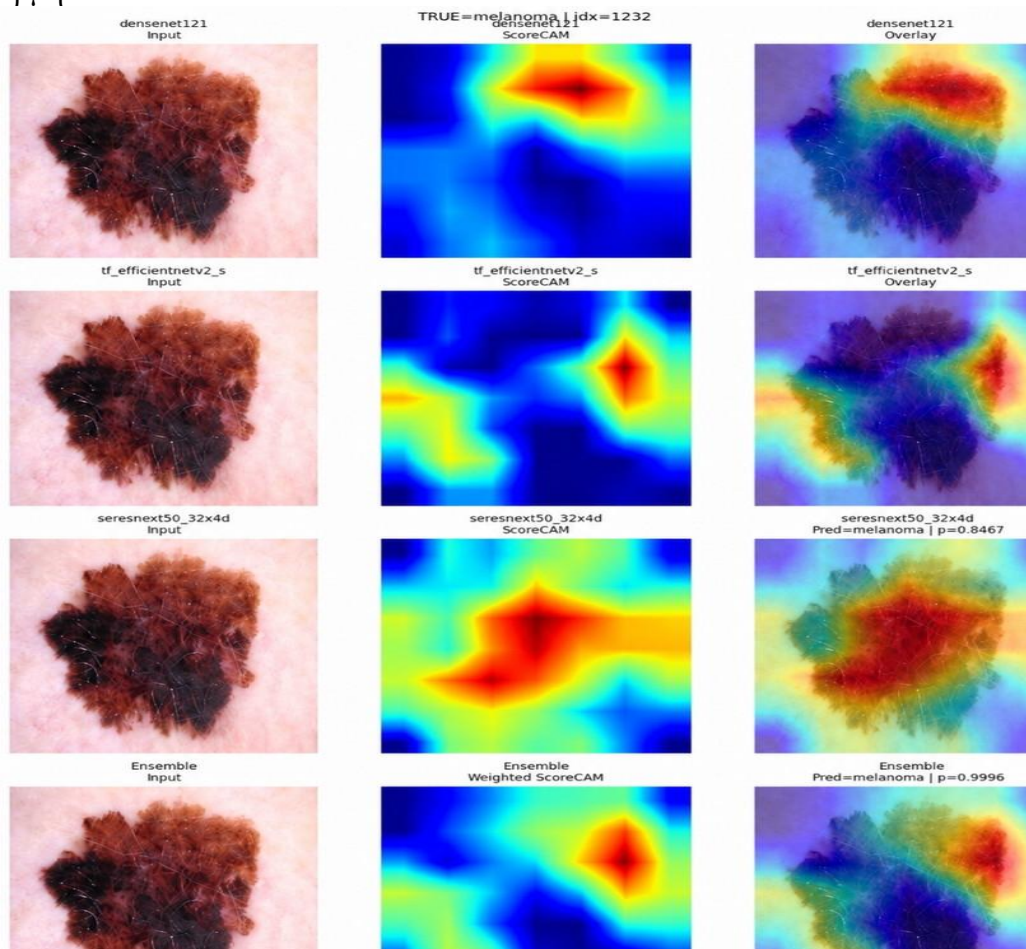
Η ίδια εικόνα φαίνεται και στον πίνακα σύγχυσης του validation set. Από τα 131 πραγματικά melanoma, τα 121 ταξινομήθηκαν σωστά ως melanoma. Από την άλλη πλευρά, από τα 809 πραγματικά melanocytic nevi, μόνο τα 509 ταξινομήθηκαν σωστά ως nevi, ενώ 214 προβλέφθηκαν ως melanoma. Αυτό είναι το πιο χαρακτηριστικό εύρημα του πειράματος: η μέθοδος έχει υψηλή ευαισθησία στο melanoma, αλλά αρκετά benign nevi μεταφέρονται στην κατηγορία melanoma.

37	0	2	0	0	1	0
4	51	2	1	0	1	1
4	1	120	0	0	8	0
0	0	0	12	0	0	0
7	18	40	9	509	214	12
1	3	4	0	0	121	2
0	0	0	0	0	0	15

Πίνακας 4.20. Πίνακας σύγχυσης του ensemble στο validation set

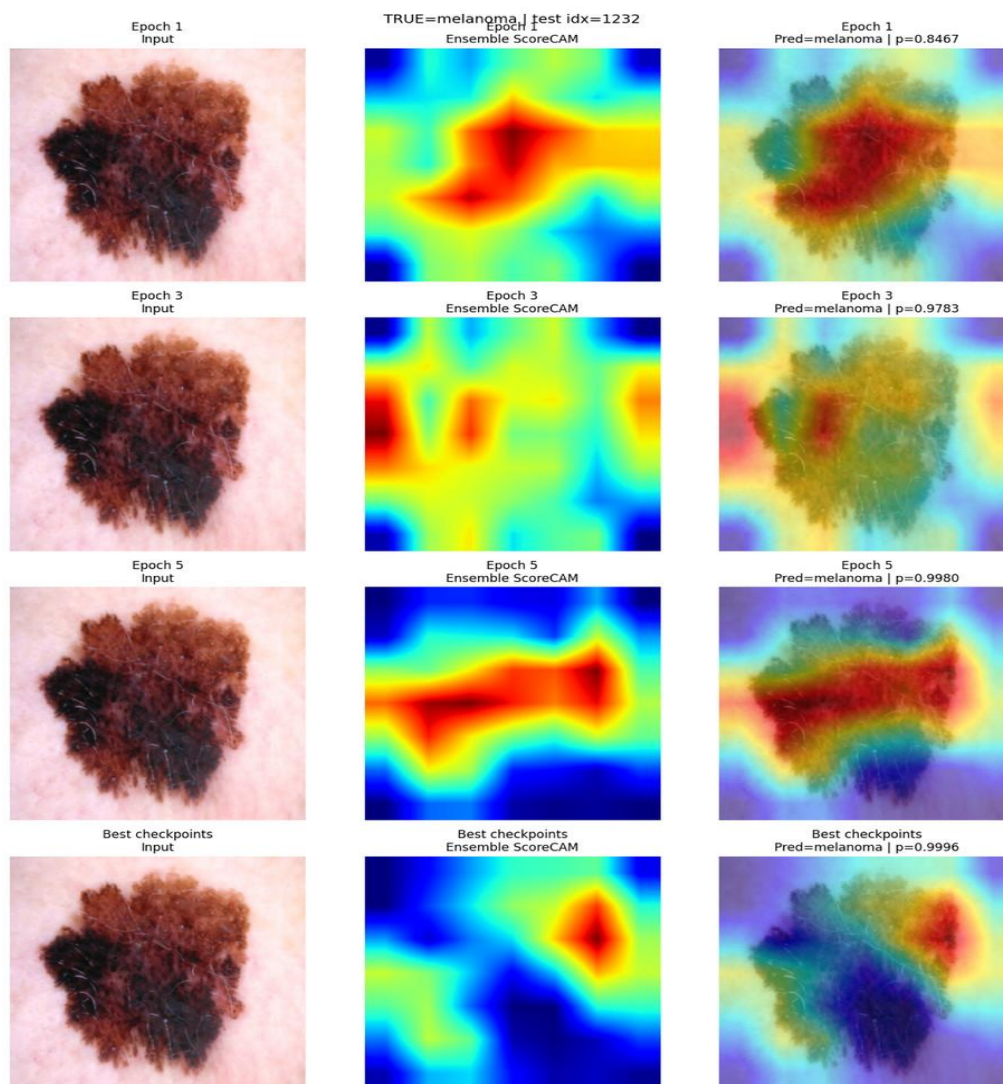
Για να συνδεθεί η πολυταξική αξιολόγηση, οι προβλέψεις του test set ομαδοποιήθηκαν σε δυαδική μορφή: melanoma και non-melanoma. Σε αυτή την αξιολόγηση προέκυψαν 137

true positives, 894 true negatives, 247 false positives και 7 false negatives. Με βάση αυτά τα αποτελέσματα, το μοντέλο είχε πολύ υψηλό binary recall για το melanoma, καθώς εντόπισε 137 από τα 144 πραγματικά melanomas. Από την άλλη πλευρά, τα 247 false positives δείχνουν ότι αρκετές μη μελανωματικές βλάβες χαρακτηρίστηκαν ως melanoma. Από τον δυαδικό αυτόν διαχωρισμό προκύπτει binary accuracy περίπου 80.23%, recall για melanoma περίπου 95.14%, precision περίπου 35.68% και F1-score περίπου 51.94%. Η τιμή recall είναι ιδιαίτερα υψηλή και δείχνει ότι το ensemble έχασε λίγα πραγματικά melanomas. Αντίθετα, η precision είναι χαμηλή λόγω των πολλών false positives. Για την εξηγησιμότητα των προβλέψεων χρησιμοποιήθηκε η μέθοδος Score-CAM. Σε αντίθεση με μεθόδους που βασίζονται άμεσα σε gradients, το Score-CAM χρησιμοποιεί τις ενεργοποιήσεις των συνελκτικών χαρτών και αξιολογεί τη συνεισφορά τους στην τελική πρόβλεψη. Στο συγκεκριμένο πείραμα εφαρμόστηκε Score-CAM τόσο σε κάθε backbone ξεχωριστά όσο και στο τελικό ensemble, ώστε να εξεταστεί αν τα μοντέλα εστιάζουν στις ίδιες ή σε διαφορετικές περιοχές της εικόνας. Στην πρώτη χαρακτηριστική περίπτωση εξετάστηκε ένα true positive melanoma από το test set, με δείκτη $idx = 1232$. Η πραγματική και η προβλεπόμενη κλάση ήταν melanoma, με πιθανότητα $p = 0.9996$ για το ensemble. Η Score-CAM απεικόνιση δείχνει ότι τα διαφορετικά backbones δεν παράγουν ταυτόσημες περιοχές ενεργοποίησης. Το DenseNet121 και το EfficientNetV2-S εμφανίζουν διαφορετικές ζώνες εστίασης πάνω ή γύρω από τη βλάβη, ενώ στο SE-ResNeXt50 η απεικόνιση είναι περισσότερο ενημερωτική στη συγκεκριμένη περίπτωση. Το ensemble συνδυάζει αυτές τις πληροφορίες και εμφανίζει την τελική σταθμισμένη Score-CAM ερμηνεία.



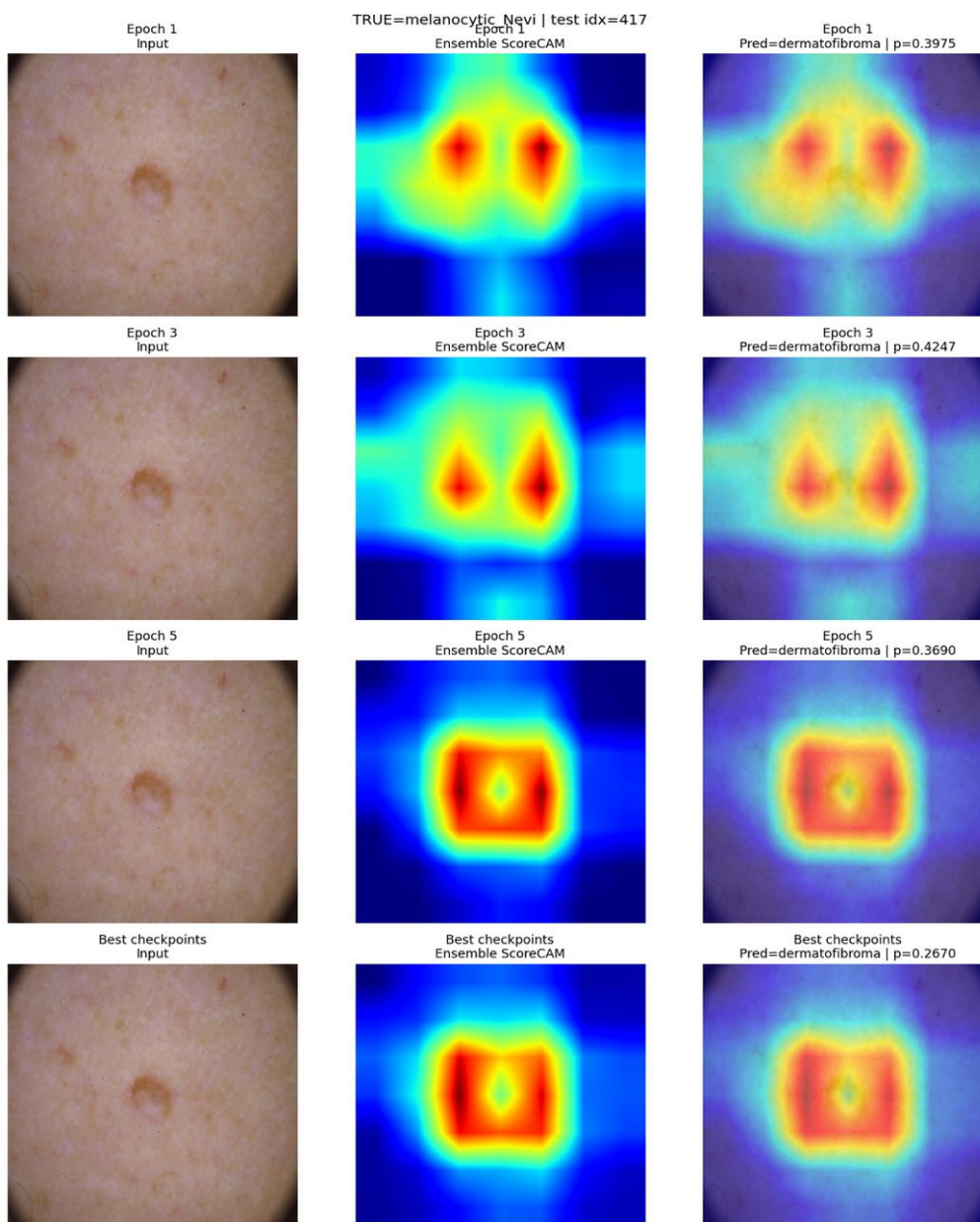
Εικόνα 4.26. Score-CAM ανά backbone και ensemble για σωστή πρόβλεψη melanoma

Η Score-CAM δείχνει ποιες περιοχές ενίσχυσαν την ταξινόμηση, όχι ποια είναι τα ακριβή όρια της δερματικής αλλοίωσης. Στην περίπτωση του true positive melanoma, η ενεργοποίηση σχετίζεται με την περιοχή της βλάβης, αλλά δεν την καλύπτει ομοιόμορφα. Αυτό είναι σημαντικό να τονιστεί, γιατί η ερμηνευτική εικόνα υποστηρίζει την πρόβλεψη, χωρίς όμως να αποτελεί ανεξάρτητη διαγνωστική απόδειξη. Στη συνέχεια εξετάστηκε η εξέλιξη της Score-CAM ερμηνείας ανά εποχή για το ίδιο true positive melanoma. Τα checkpoints που χρησιμοποιήθηκαν ήταν οι εποχές 1, 3, 5 και το best checkpoint. Στην 1η εποχή, το ensemble είχε ήδη προβλέψει melanoma με πιθανότητα $p = 0.8467$. Στην 3η εποχή η πιθανότητα αυξήθηκε σε $p = 0.9783$, στην 5η σε $p = 0.9980$, ενώ στο best checkpoint έφτασε σε $p = 0.9996$. Η αύξηση αυτή δείχνει ότι η βεβαιότητα του ensemble για το συγκεκριμένο melanoma ενισχύθηκε σημαντικά κατά την εκπαίδευση.



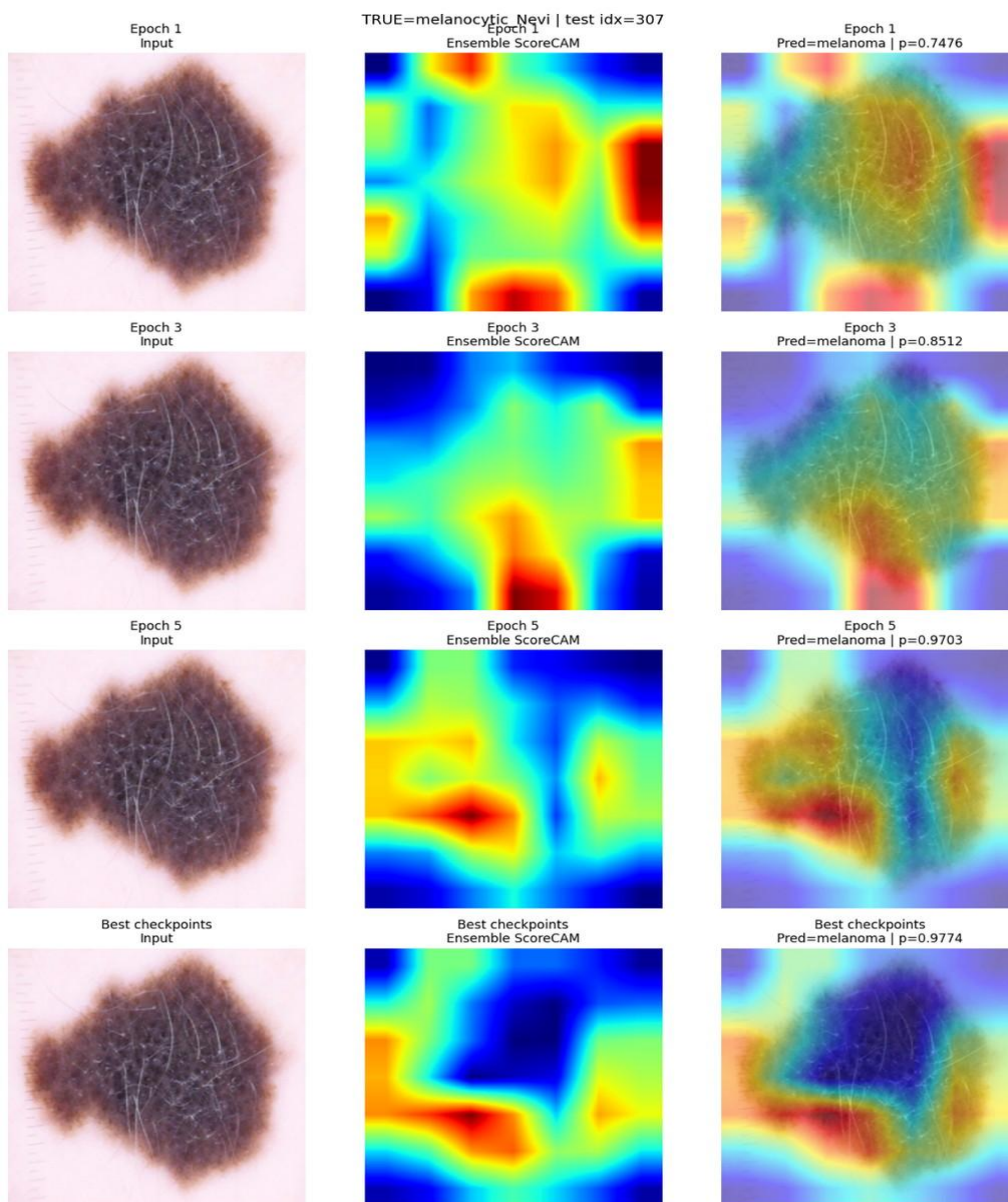
Εικόνα 4.27. Score-CAM across-epoch ανάλυση για σωστή πρόβλεψη melanoma

Παρότι η πιθανότητα αυξάνεται σταδιακά, η οπτική εστίαση δεν είναι απόλυτα σταθερή σε όλες τις εποχές. Σε ορισμένα checkpoints η ενεργοποίηση είναι πιο κεντρική, ενώ σε άλλα μετατοπίζεται προς συγκεκριμένες ζώνες της βλάβης ή προς τα όριά της. Αυτό δείχνει ότι η εκπαίδευση δεν αλλάζει μόνο τη βεβαιότητα της πρόβλεψης, αλλά και τον τρόπο με τον οποίο το ensemble αξιοποιεί τις οπτικές πληροφορίες της εικόνας. Ως δεύτερη περίπτωση εξετάστηκε ένα δείγμα με πραγματική κλάση melanocytic nevi, με δείκτη $idx = 417$. Στον δυαδικό διαχωρισμό η περίπτωση αυτή θεωρείται true negative, επειδή τόσο η πραγματική όσο και η προβλεπόμενη κατηγορία ανήκουν στο non-melanoma. Σε πολυταξικό επίπεδο όμως, το ensemble προέβλεψε dermatofibroma με πιθανότητα περίπου 0.2670, ενώ η πραγματική κλάση ήταν melanocytic nevi. Άρα η περίπτωση είναι σωστή ως προς το melanoma / non-melanoma, αλλά λανθασμένη ως προς την ακριβή επταταξική διάγνωση.



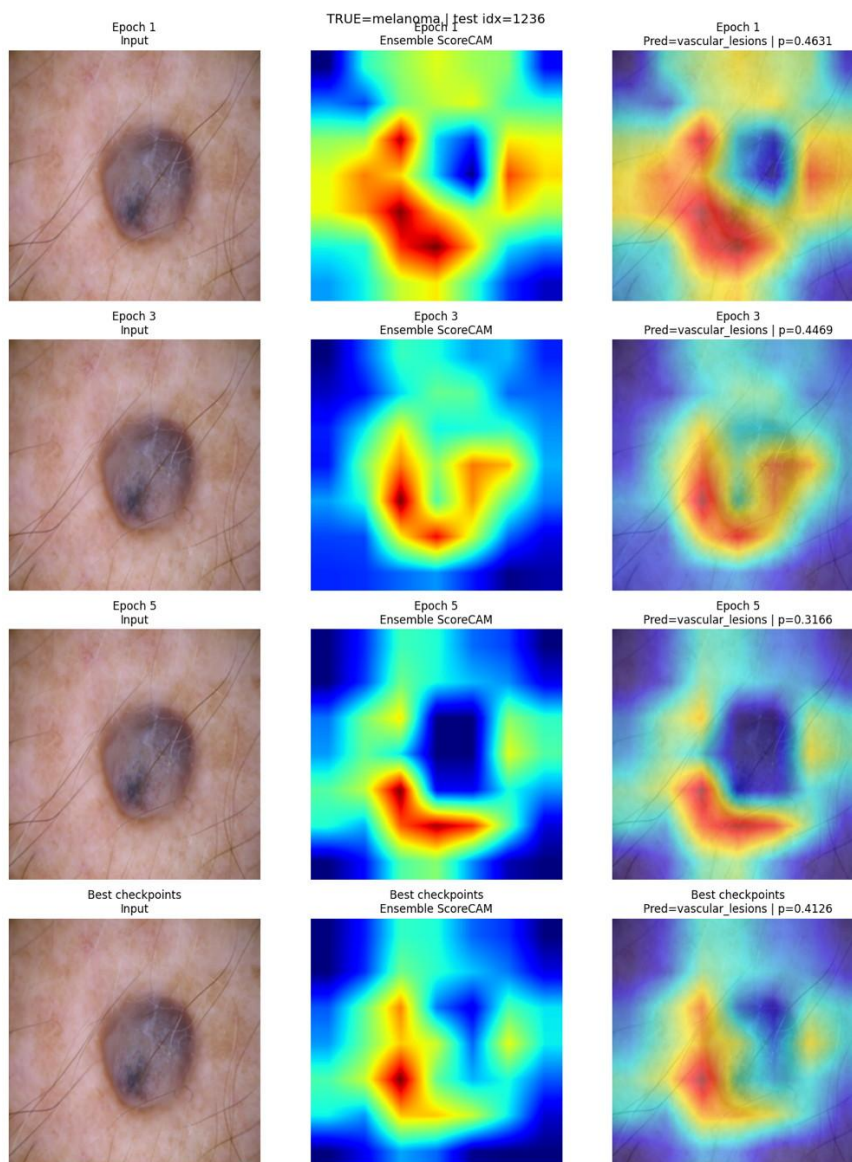
Εικόνα 4.28. Score-CAM across-epoch ανάλυση για non-melanoma περίπτωση, σωστή σε δυαδικό αλλά όχι σε πολυταξικό επίπεδο

Η περίπτωση αυτή είναι χρήσιμη επειδή δείχνει ότι η δυαδική και η πολυταξική αξιολόγηση δεν οδηγούν πάντα στο ίδιο συμπέρασμα. Για το βασικό ερώτημα αν μια εικόνα είναι melanoma ή όχι, η πρόβλεψη είναι αποδεκτή. Για την ακριβή διάκριση μεταξύ benign κατηγοριών, όμως, το μοντέλο παραμένει αβέβαιο. Οι πιθανότητες είναι σχετικά χαμηλές και η Score-CAM εστίαση μεταβάλλεται ανά εποχή, γεγονός που δείχνει περιορισμένη σταθερότητα στην ερμηνεία της συγκεκριμένης εικόνας. Η τρίτη περίπτωση ήταν ένα false positive, με δείκτη $idx = 307$. Η πραγματική κλάση ήταν melanocytic nevus, αλλά το ensemble προέβλεψε melanoma. Η πιθανότητα για melanoma αυξήθηκε από $p = 0.7476$ στην 1η εποχή σε $p = 0.8512$ στην 3η, $p = 0.9703$ στην 5η και $p = 0.9774$ στο best checkpoint. Η περίπτωση αυτή είναι ιδιαίτερα σημαντική, γιατί δείχνει ότι το μοντέλο έγινε σταδιακά πιο βέβαιο για μια λανθασμένη πρόβλεψη.



Εικόνα 4.29. Score-CAM across-epoch ανάλυση για false positive περίπτωση melanocytic nevus ως melanoma

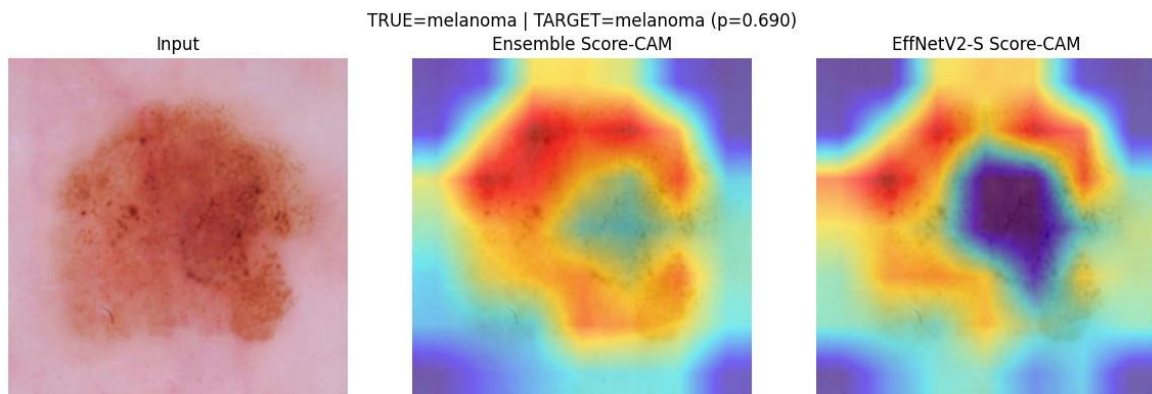
Στο false positive παράδειγμα, η Score-CAM εστίαση βρίσκεται κυρίως πάνω στη δερματική βλάβη. Αυτό σημαίνει ότι το λάθος δεν φαίνεται να οφείλεται απλώς σε άσχετο φόντο. Πιο πιθανό είναι ότι η συγκεκριμένη benign βλάβη είχε οπτικά χαρακτηριστικά που το ensemble συνέδεσε με melanoma, όπως σκούρες περιοχές, ανομοιογενή υφή ή ακανόνιστη εμφάνιση. Το παράδειγμα αυτό συνδέεται άμεσα με τα πολλά false positives που παρατηρήθηκαν στη δυαδική αξιολόγηση. Η τέταρτη περίπτωση ήταν ένα false negative melanoma, με δείκτη $idx = 1236$. Η πραγματική κλάση ήταν melanoma, αλλά το ensemble προέβλεψε vascular lesions. Η πιθανότητα της λανθασμένης πρόβλεψης ήταν $p = 46.31\%$ στην 1η εποχή, $p = 44.69\%$ στην 3η, $p = 31.66\%$ στην 5η και $p = 41.26\%$ στο best checkpoint. Σε αντίθεση με το false positive παράδειγμα, εδώ η βεβαιότητα της λανθασμένης πρόβλεψης δεν είναι τόσο υψηλή. Αυτό δείχνει ότι το μοντέλο ήταν περισσότερο αβέβαιο, αλλά τελικά δεν απέδωσε την εικόνα στην κλάση melanoma.



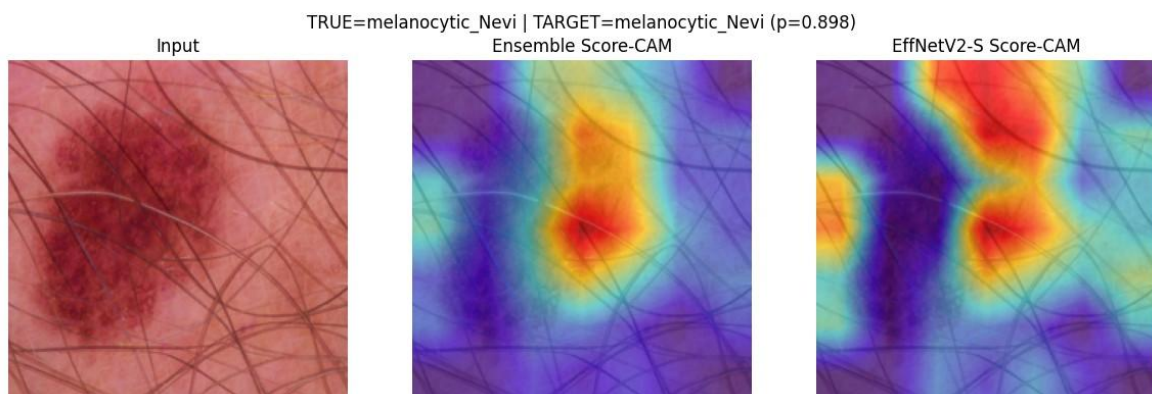
Εικόνα 4.30. Score-CAM across-epoch ανάλυση για false negative melanoma.

Η false negative περίπτωση έχει ιδιαίτερη βαρύτητα, επειδή αφορά πραγματικό melanoma που δεν αναγνωρίστηκε ως τέτοιο. Η Score-CAM απεικόνιση δείχνει ενεργοποιήσεις γύρω από την περιοχή της βλάβης και σε γειτονικά σημεία, αλλά η πρόβλεψη παραμένει λανθασμένη. Αυτό δείχνει ότι το πρόβλημα δεν είναι απαραίτητα μόνο χωρικό, αλλά και ταξινομικό: το μοντέλο βλέπει οπτικά στοιχεία της εικόνας, αλλά τα αντιστοιχίζει σε διαφορετική παθολογική κατηγορία.

Στο τέλος του πειράματος εξετάστηκαν και δύο επιπλέον ενδεικτικές περιπτώσεις με Score-CAM: ένα melanoma με δείκτη $idx = 1123$ και ένα melanocytic nevi με δείκτη $idx = 265$. Στην πρώτη εικόνα, η πραγματική κλάση και ο στόχος ερμηνείας ήταν melanoma, με πιθανότητα $p = 0.690$. Στη δεύτερη εικόνα, η πραγματική κλάση και ο στόχος ήταν melanocytic nevi, με πιθανότητα $p = 0.898$. Οι δύο αυτές περιπτώσεις χρησιμοποιήθηκαν για σύγκριση της Score-CAM ερμηνείας του ensemble με την αντίστοιχη ερμηνεία του EfficientNetV2-S.



Εικόνα 4.31. Σύγκριση Score-CAM του ensemble και του EfficientNetV2-S για περίπτωση melanoma.



Εικόνα 4.32. Σύγκριση Score-CAM του ensemble και του EfficientNetV2-S για περίπτωση melanocytic nevi

Η σύγκριση αυτή δείχνει ότι η ερμηνεία του ensemble και η ερμηνεία ενός μεμονωμένου backbone δεν είναι πάντα ίδιες. Στην περίπτωση melanoma, το ensemble εμφανίζει πιο εκτεταμένη ενεργοποίηση γύρω από την περιοχή της βλάβης, ενώ το EfficientNetV2-S δίνει διαφορετική χωρική κατανομή. Στην περίπτωση melanocytic nevus, η ενεργοποίηση του ensemble εμφανίζεται πιο συγκεντρωμένη κοντά στη βλάβη, ενώ το EfficientNetV2-S αναδεικνύει επιπλέον περιοχές.

Συνολικά, το πείραμα Score-CAM με ensemble μοντέλων έδειξε ότι ο συνδυασμός DenseNet121, EfficientNetV2-S και SE-ResNeXt50 μπορεί να πετύχει υψηλή ευαισθησία για την κλάση melanoma, αλλά με κόστος την αύξηση των false positives. Στο validation set, το ensemble πέτυχε macro recall = 89.00%, ενώ στη δυαδική αξιολόγηση του test set πέτυχε recall για melanoma περίπου 95.14%. Παράλληλα, η χαμηλή precision στη δυαδική αξιολόγηση δείχνει ότι αρκετές non-melanoma εικόνες χαρακτηρίστηκαν ως melanoma. Η χρήση του Score-CAM έδειξε ότι οι σωστές και λανθασμένες προβλέψεις συχνά βασίζονται σε περιοχές κοντά στη δερματική βλάβη, αλλά η ορθή χωρική εστίαση δεν εγγυάται πάντα σωστή ταξινόμηση. Το πιο χαρακτηριστικό εύρημα είναι ότι στο false positive παράδειγμα το μοντέλο έγινε σταδιακά πιο βέβαιο για τη λανθασμένη πρόβλεψη melanoma. Αντίστοιχα, στο false negative παράδειγμα, το μοντέλο εστίασε σε σχετικές περιοχές, αλλά αντιστοίχισε την εικόνα σε άλλη παθολογική κατηγορία.

4.7 T-Former στο Derm7pt για δυαδική ταξινόμηση melanoma / non-melanoma

Στο πείραμα αυτό αξιολογήθηκε ένα πολυτροπικό μοντέλο τύπου T-Former style binary classifier στο σύνολο δεδομένων Derm7pt. Σε αντίθεση με προηγούμενα πειράματα που βασίζονταν μόνο σε δερματοσκοπικές εικόνες, εδώ αξιοποιήθηκαν τρεις μορφές πληροφορίας: κλινική εικόνα, δερματοσκοπική εικόνα και μεταδεδομένα. Ο στόχος του πειράματος ήταν η δυαδική ταξινόμηση των περιστατικών σε non-melanoma και melanoma, με την κλάση 0 να αντιστοιχεί σε non-melanoma και την κλάση 1 σε melanoma. Το Derm7pt περιλαμβάνει συνολικά 1011 περιστατικά, από τα οποία χρησιμοποιήθηκαν τα επίσημα splits του dataset: 413 δείγματα για εκπαίδευση, 203 για επικύρωση και 395 για έλεγχο. Η δυαδική ετικέτα δημιουργήθηκε με βάση τη διάγνωση: κάθε διάγνωση που περιείχε τον όρο melanoma εντάχθηκε στην κλάση 1, ενώ όλες οι υπόλοιπες διαγνώσεις εντάχθηκαν στην κλάση 0. Με αυτόν τον τρόπο, το αρχικό πολυδιαγνωστικό πρόβλημα μετατράπηκε σε πρόβλημα αναγνώρισης μελανώματος. Στο training set υπήρχαν 323 non-melanoma και 90 melanoma δείγματα. Στο validation set υπήρχαν 142 non-melanoma και 61 melanoma, ενώ στο test set υπήρχαν 294 non-melanoma και 101 melanoma. Η κατανομή αυτή δείχνει εμφανή ανισορροπία υπέρ της non-melanoma κλάσης. Για τον λόγο αυτό χρησιμοποιήθηκαν class weights στη συνάρτηση απώλειας, με βάρη 0.4358 για την κλάση non-melanoma και 1.5642 για την κλάση melanoma. Με αυτόν τον τρόπο το μοντέλο τιμωρείται περισσότερο όταν κάνει λάθος στην υποεκπροσωπούμενη κλάση melanoma.

Train	Validation	Test
323/90	142/61	294/101

Πίνακας 4.21. Κατανομή των κλάσεων στο πείραμα T-Former Derm7pt

Η αρχιτεκτονική του μοντέλου βασίστηκε σε δύο ανεξάρτητους κλάδους Swin-Tiny, έναν για την κλινική εικόνα και έναν για τη δερματοσκοπική εικόνα. Κάθε κλάδος εξήγαγε χαρακτηριστικά από την αντίστοιχη εικόνα. Παράλληλα, τα μεταδεδομένα κωδικοποιήθηκαν με έναν μικρό πλήρως συνδεδεμένο encoder. Τα μεταδεδομένα περιλάμβαναν πληροφορίες όπως φύλο, ανατομική θέση, elevation και level of diagnostic difficulty, οι οποίες μετατράπηκαν σε αριθμητικά χαρακτηριστικά μέσω one-hot encoding. Συνολικά προέκυψαν 17 μεταδεδομένα χαρακτηριστικά. Τα χαρακτηριστικά από τους δύο οπτικούς κλάδους και τον metadata encoder ενώθηκαν σε ένα ενιαίο διάλυμα και δόθηκαν σε fusion head για την τελική δυαδική ταξινόμηση. Αυτή η σχεδίαση είναι σημαντική, επειδή το μοντέλο δεν βασίζεται αποκλειστικά στη δερματοσκοπική εικόνα, αλλά επιχειρεί να συνδυάσει διαφορετικές πηγές πληροφορίας, όπως συμβαίνει και στην πραγματική κλινική αξιολόγηση.

Η εκπαίδευση πραγματοποιήθηκε για 12 εποχές. Το καλύτερο μοντέλο επιλέχθηκε με βάση το validation F1-score. Η καλύτερη τιμή καταγράφηκε στην 7η εποχή, όπου το μοντέλο πέτυχε val_acc = 80.30%, val_f1 = 71.01% και val_auc = 84.38%. Είναι ενδιαφέρον ότι στις επόμενες εποχές το training performance συνέχισε να αυξάνεται, φτάνοντας στην 12η εποχή σε train_acc = 98.55%, train_f1 = 96.67% και train_auc = 99.92%, ενώ οι validation μετρικές δεν βελτιώθηκαν αντίστοιχα. Αυτό δείχνει τάση υπερπροσαρμογής, κάτι αναμενόμενο σε ένα μικρό dataset με περιορισμένο αριθμό melanoma περιπτώσεων.

Εποχή	Train Loss	Train Acc	Train F1	Train AUC	Val Loss	Val Acc	Val F1	Val AUC
1	0.6943	73.85%	31.65%	61.29%	0.6859	59.61%	58.16%	79.58%
7	0.3582	90.07%	79.60%	97.14%	0.6148	80.30%	71.01%	84.38%
12	0.2319	98.55%	96.67%	99.92%	0.6028	81.28%	66.67%	84.09%

Πίνακας 4.22. Πορεία εκπαίδευσης του T-Former στο Derm7pt

Στο test set, με το αρχικό threshold 0.50, το μοντέλο πέτυχε Test Loss = 0.6116, Test Accuracy = 78.23%, Test F1 = 66.41% και Test AUC = 87.80%. Ο πίνακας σύγχυσης ήταν:

224	70
16	85

Αυτό σημαίνει ότι από τα 294 non-melanoma δείγματα, τα 224 ταξινομήθηκαν σωστά, ενώ 70 ταξινομήθηκαν λανθασμένα ως melanoma. Από τα 101 melanoma δείγματα, τα 85 ταξινομήθηκαν σωστά, ενώ 16 χάθηκαν ως non-melanoma. Η τιμή recall για melanoma = 84.16% δείχνει ότι το μοντέλο εντόπισε μεγάλο μέρος των πραγματικών μελανωμάτων, ενώ η precision για melanoma = 54.84% δείχνει ότι υπήρχαν αρκετά false positives.

Κλάση	Precision (%)	Recall (%)	F1-score (%)	Support
0	93.33%	76.19%	83.90%	294
1	54.84%	84.16%	66.41%	101

Κλάση	Precision (%)	Recall (%)	F1-score (%)	Support
Accuracy			78.23%	395
Macro avg	74.09%	80.17%	75.15%	395
Weighted avg	83.49%	78.23%	79.42%	395

Πίνακας 4.23. Πίνακας σύγκρισης του T-Former στο test set με threshold 0.50

Από το classification report, η κλάση non-melanoma είχε precision = 93.33%, recall = 76.19% και F1-score = 83.90%. Η κλάση melanoma είχε precision = 54.84%, recall = 84.16% και F1-score = 66.41%. Το macro F1 ήταν 75.15%, ενώ το weighted F1 ήταν 79.42%. Η διαφορά ανάμεσα σε macro και weighted F1 δείχνει ότι η μεγαλύτερη κλάση non-melanoma επηρεάζει θετικά τη συνολική weighted εικόνα, ενώ η macro αξιολόγηση αποκαλύπτει καλύτερα τη δυσκολία της melanoma κλάσης. Στη συνέχεια πραγματοποιήθηκε βελτιστοποίηση του threshold με βάση το validation F1. Το καλύτερο threshold ήταν περίπου 0.66–0.68, με best validation F1 = 72.44% και validation accuracy 82.76%. Εφαρμόζοντας το threshold 0.66 στο test set, η απόδοση έγινε Test Accuracy = 80.00%, Test F1 = 65.20% και Test AUC = 87.80%. Ο νέος πίνακας σύγκρισης ήταν:

242	52
27	74

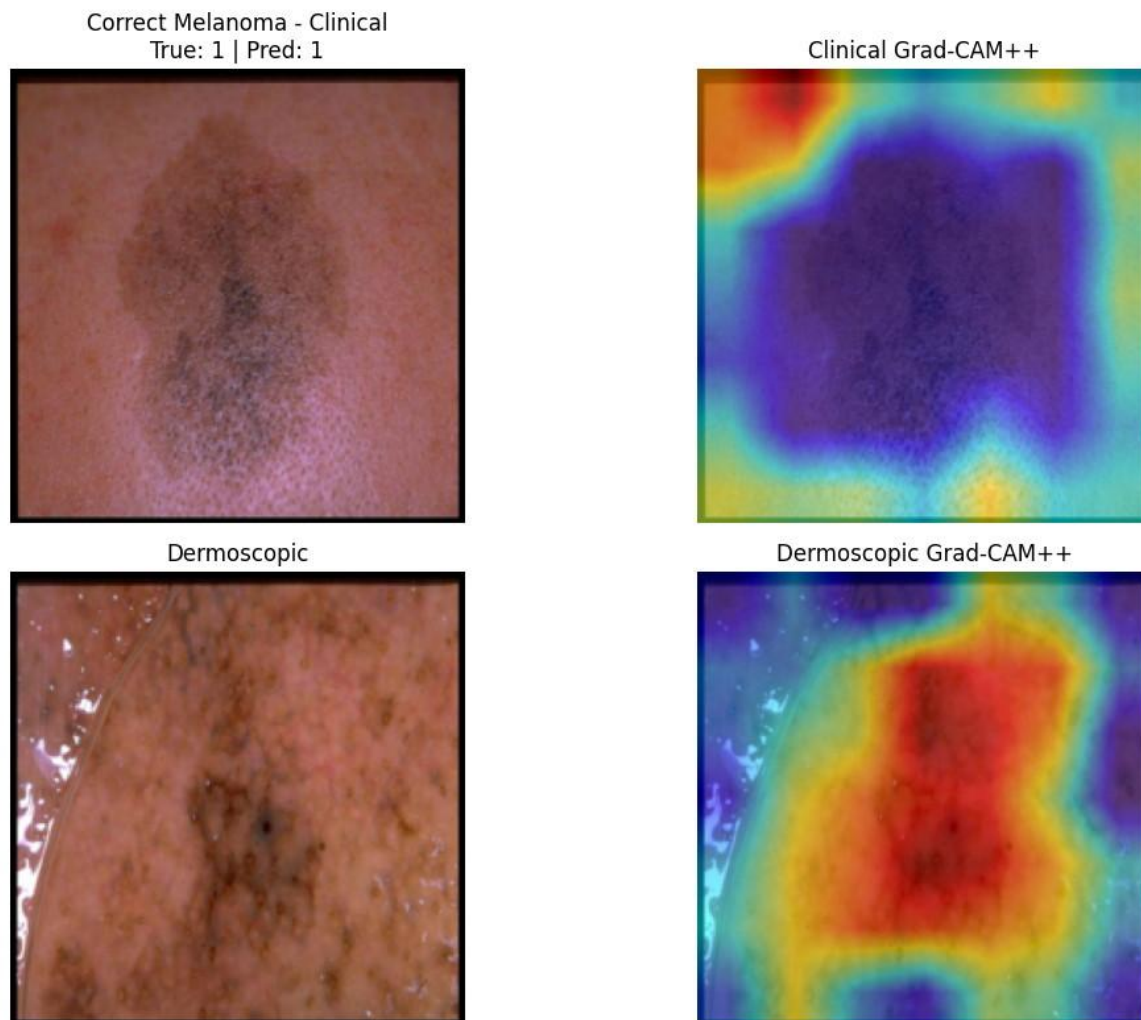
Με το αυξημένο threshold, τα false positives μειώθηκαν από 70 σε 52, αλλά τα false negatives αυξήθηκαν από 16 σε 27. Αυτό δείχνει τον πρακτικό συμβιβασμό ανάμεσα σε precision και recall. Το threshold 0.50 είναι πιο ευαίσθητο για την ανίχνευση melanoma, ενώ το threshold 0.66 είναι πιο αυστηρό και μειώνει τις λανθασμένες θετικές προβλέψεις εις βάρος της ευαισθησίας.

Κλάση	Precision (%)	Recall (%)	F1-score (%)	Support
0	89.96%	82.31%	85.97%	294
1	58.73%	73.27%	65.20%	101
Accuracy			80.00%	395
Macro avg	74.35%	77.79%	75.58%	395
Weighted avg	81.98%	80.00%	80.66%	395

Πίνακας 4.24. Πίνακας σύγκρισης του T-Former στο test set μετά τη βελτιστοποίηση threshold

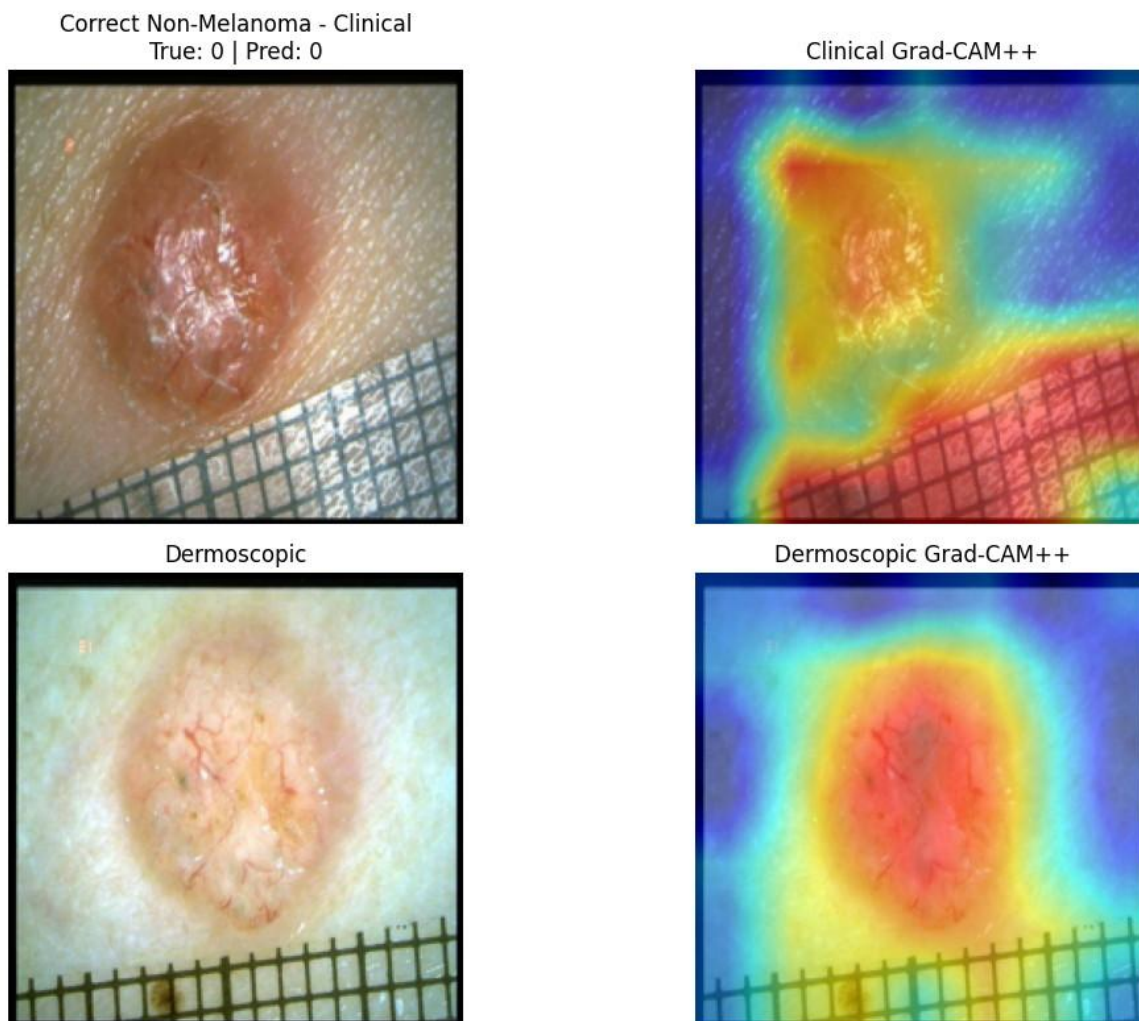
Για την ερμηνεία των προβλέψεων εφαρμόστηκε αρχικά Grad-CAM++ χωριστά στον κλινικό και στον δερματοσκοπικό κλάδο του μοντέλου. Αυτό είναι ένα από τα πιο ενδιαφέροντα σημεία του πειράματος, επειδή το μοντέλο έχει δύο οπτικές εισόδους. Άρα η εξηγησιμότητα δεν εξετάζει μόνο πού εστιάζει το μοντέλο σε μία εικόνα, αλλά πώς κατανέμεται η προσοχή του ανάμεσα στην κλινική και τη δερματοσκοπική όψη του ίδιου περιστατικού. Στην ορθή περίπτωση melanoma, το μοντέλο προέβλεψε σωστά την κλάση 1, με πιθανότητα 79.86% για melanoma και 20.14% για non-melanoma. Στην κλινική εικόνα, η Grad-CAM++ ενεργοποίηση δεν περιορίζεται απόλυτα στο κέντρο της βλάβης, αλλά εμφανίζεται γύρω από ευρύτερες περιοχές της εικόνας. Στη δερματοσκοπική εικόνα, η

ενεργοποίηση είναι πιο κοντά στην περιοχή της βλάβης και στα πιο σκούρα/ανομοιογενή στοιχεία. Αυτό δείχνει ότι ο δερματοσκοπικός κλάδος φαίνεται να παρέχει πιο άμεση πληροφορία για την τελική πρόβλεψη melanoma.



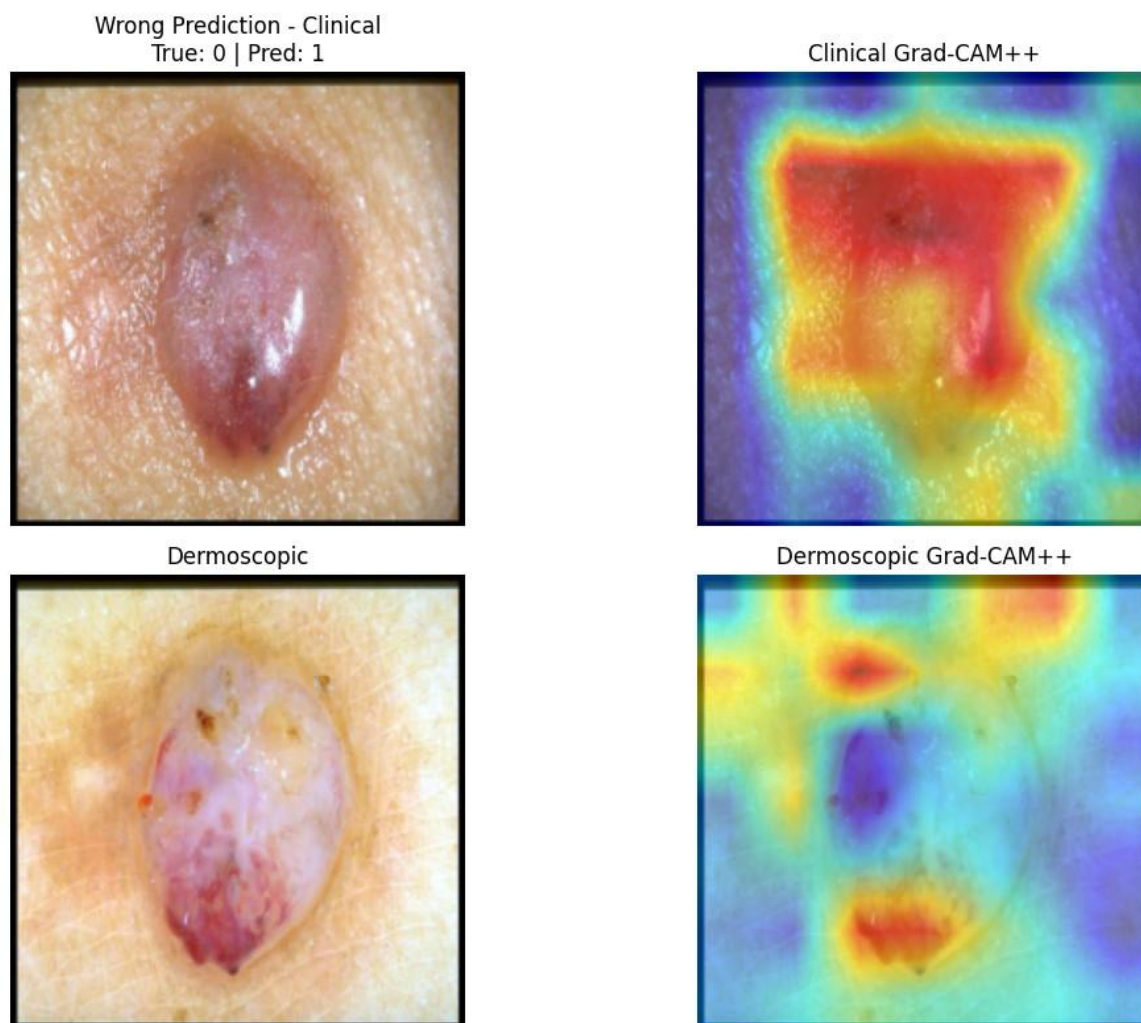
Εικόνα 4.33. Grad-CAM++ ανάλυση για ορθή πρόβλεψη melanoma στο πολυτροπικό T-Former

Στην ορθή περίπτωση non-melanoma, το μοντέλο προέβλεψε σωστά την κλάση 0, με πιθανότητα 93.29% για non-melanoma και 6.71% για melanoma. Η Grad-CAM++ απεικόνιση δείχνει ότι και οι δύο κλάδοι ενεργοποιούνται γύρω από την περιοχή της βλάβης, αλλά η τελική απόφαση παραμένει ξεκάθαρα υπέρ της non-melanoma κλάσης. Αυτό το παράδειγμα είναι χρήσιμο γιατί δείχνει ότι το μοντέλο δεν χαρακτηρίζει κάθε εμφανή ή καλά οριοθετημένη βλάβη ως melanoma. Αντίθετα, σε αυτή την περίπτωση απορρίπτει σωστά την κακοήγη κλάση.



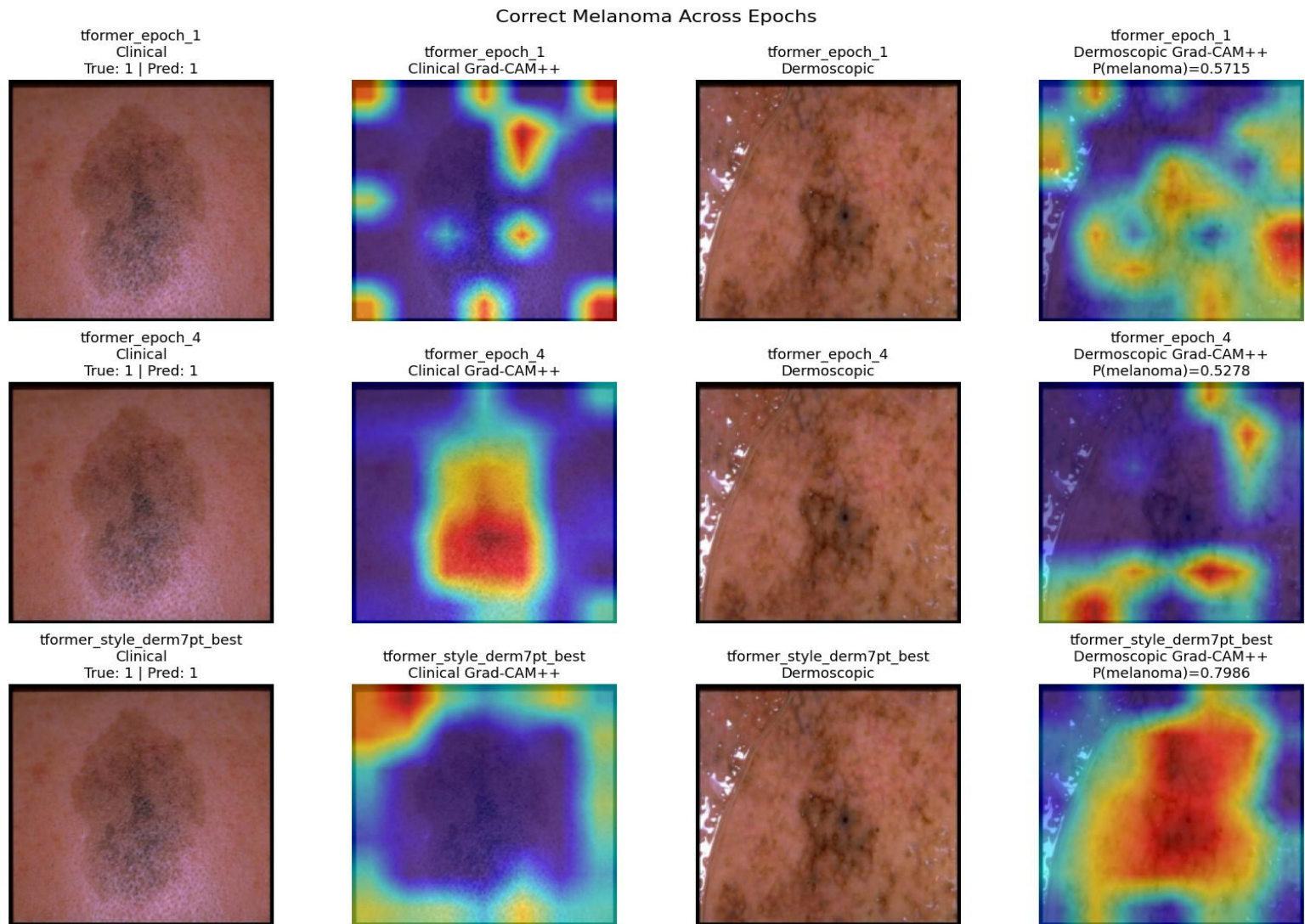
Εικόνα 4.34. Grad-CAM++ ανάλυση για ορθή πρόβλεψη non-melanoma στο πολυτροπικό T-Former

Στη λανθασμένη περίπτωση, η πραγματική κλάση ήταν non-melanoma, αλλά το μοντέλο προέβλεψε melanoma με πιθανότητα 63.66%. Η πιθανότητα για non-melanoma ήταν 36.34%, άρα η πρόβλεψη ήταν λανθασμένη αλλά όχι εξαιρετικά βέβαιη. Η κλινική Grad-CAM++ εικόνα εμφανίζει έντονη ενεργοποίηση πάνω στη βλάβη, ενώ η δερματοσκοπική εικόνα παρουσιάζει πιο διάσπαρτες ενεργοποιήσεις. Αυτό δείχνει ότι το μοντέλο επηρεάστηκε από οπτικά στοιχεία της βλάβης, πιθανόν επειδή η συγκεκριμένη non-melanoma περίπτωση είχε ύποπτη μορφολογία.



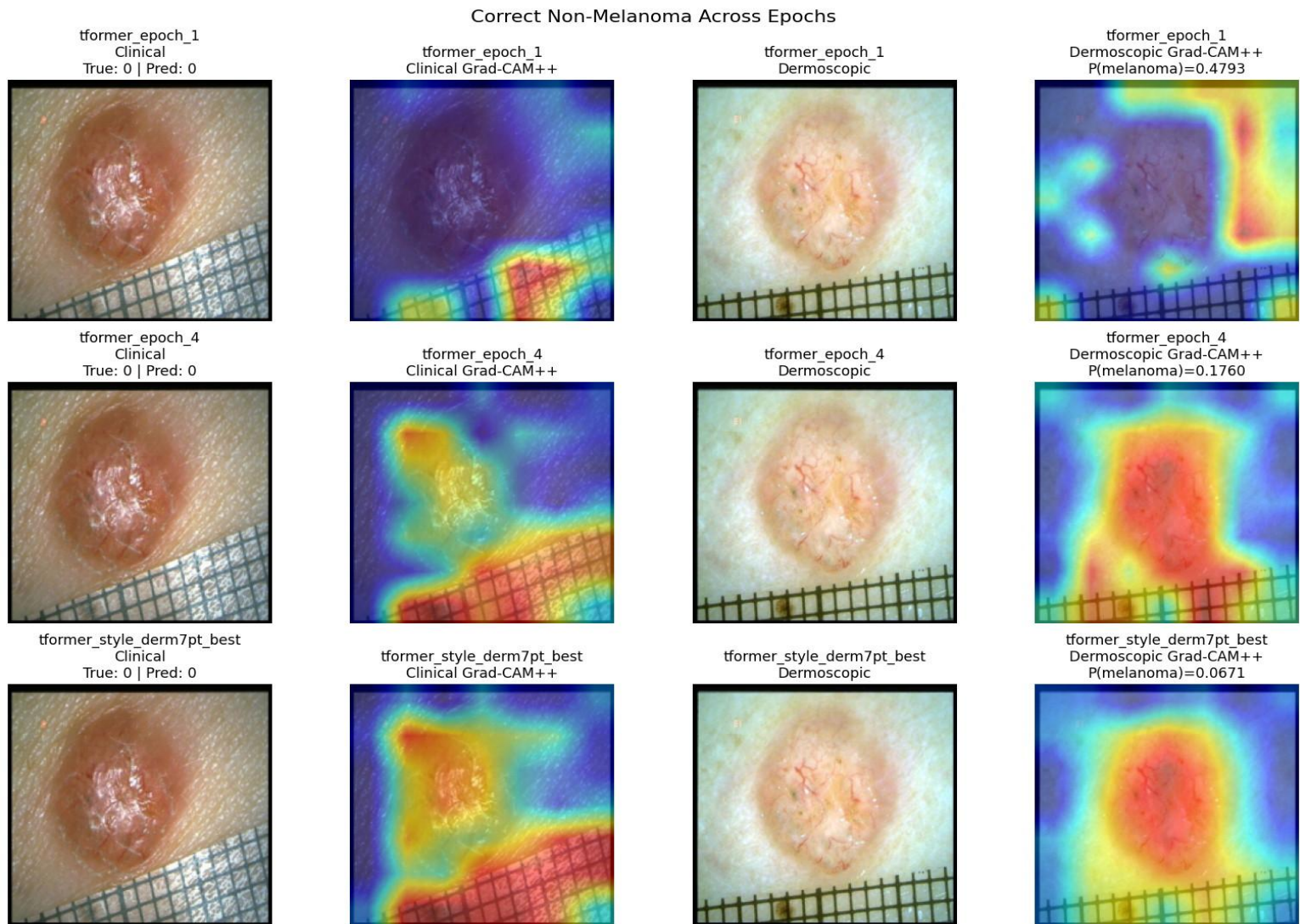
Εικόνα 4.35. Grad-CAM++ ανάλυση για λανθασμένη πρόβλεψη non-melanoma ως melanoma

Στη συνέχεια πραγματοποιήθηκε Grad-CAM++ across-epoch ανάλυση για τις τρεις χαρακτηριστικές περιπτώσεις: correct melanoma, correct non-melanoma και misclassified case. Χρησιμοποιήθηκαν τα checkpoints της epoch 1, της epoch 4 και του best model. Η ανάλυση αυτή είναι χρήσιμη επειδή δείχνει πώς αλλάζει η εστίαση του μοντέλου κατά τη διάρκεια της εκπαίδευσης, τόσο στον κλινικό όσο και στον δερματοσκοπικό κλάδο.



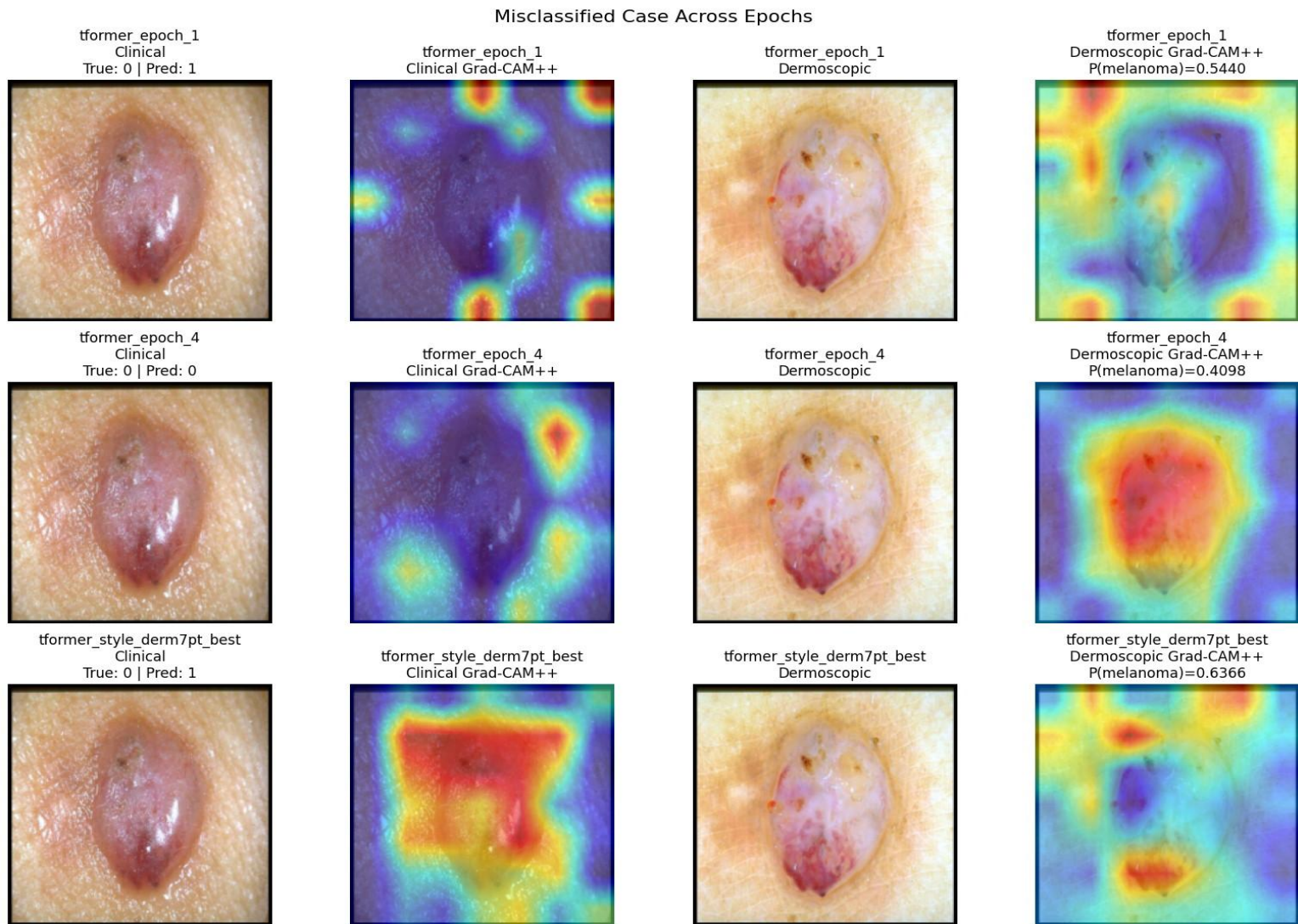
Εικόνα 4.36. Grad-CAM++ across-epoch ανάλυση για ορθή πρόβλεψη melanoma

Στην περίπτωση του correct melanoma, η across-epoch εικόνα μπορεί να δείξει αν το μοντέλο από τις πρώτες εποχές εστιάζει στη βλάβη ή αν η εστίαση γίνεται πιο συγκεκριμένη στο καλύτερο checkpoint. Το σημαντικό είναι ότι η across-epoch ανάλυση δεν πρέπει να διαβαστεί ως απόλυτη απόδειξη ότι το μοντέλο μαθαίνει σωστά, αλλά ως οπτική ένδειξη της μεταβολής των περιοχών που επηρεάζουν την πρόβλεψη.



Εικόνα 4.37. Grad-CAM++ across-epoch ανάλυση για ορθή πρόβλεψη non-melanoma

Στην περίπτωση του correct non-melanoma, η across-epoch ανάλυση δείχνει πώς το μοντέλο σταθεροποιεί την απόρριψη της melanoma κλάσης. Αυτό έχει σημασία, επειδή η αξιόπιστη διάγνωση δεν αφορά μόνο την ανίχνευση melanoma, αλλά και την αποφυγή υπερδιάγνωσης σε benign ή μη μελανωματικές βλάβες.



Εικόνα 4.38. Grad-CAM++ across-epoch ανάλυση για λανθασμένη πρόβλεψη non-melanoma ως melanoma

Στη misclassified περίπτωση, η across-epoch ανάλυση είναι ιδιαίτερα χρήσιμη. Αν η εστίαση παραμένει πάνω ή κοντά στη βλάβη, τότε το λάθος δεν οφείλεται απλώς σε φόντο ή τεχνικό θόρυβο. Αντίθετα, δείχνει ότι το μοντέλο εξετάζει σχετικές περιοχές, αλλά αποδίδει σε αυτές χαρακτηριστικά που οδηγούν προς την κλάση melanoma. Αυτό είναι σημαντικό, γιατί διαχωρίζει το πρόβλημα της χωρικής εστίασης από το πρόβλημα της σωστής ταξινόμησης.

Συνολικά, το πείραμα με το T-Former style multimodal μοντέλο στο Derm7pt έδειξε ότι η αξιοποίηση κλινικής εικόνας, δερματοσκοπικής εικόνας και μεταδεδομένων μπορεί να δώσει ικανοποιητική διαχωριστική ικανότητα, με Test AUC = 87.80%. Το μοντέλο παρουσίασε καλή ευαισθησία για την κλάση melanoma, με recall 84.16% στο threshold 0.50, αλλά αυτό συνοδεύτηκε από αρκετά false positives και σχετικά χαμηλότερη precision για την melanoma κλάση. Η βελτιστοποίηση του threshold βελτίωσε την ακρίβεια στο 80.00%, αλλά μείωσε το melanoma recall από 84.16% σε 73.27%. Το βασικό εύρημα του

πειράματος είναι ότι το πολυτροπικό μοντέλο δεν προσφέρει μόνο αριθμητική πρόβλεψη, αλλά επιτρέπει και ξεχωριστή ερμηνεία της κλινικής και της δερματοσκοπικής εισόδου. Οι Grad-CAM++ και SHAP απεικονίσεις δείχνουν ότι οι δύο εικόνες δεν συνεισφέρουν απαραίτητα με τον ίδιο τρόπο. Σε ορισμένες περιπτώσεις, η δερματοσκοπική εικόνα φαίνεται πιο άμεσα συνδεδεμένη με την περιοχή της βλάβης, ενώ η κλινική εικόνα μπορεί να παρέχει ευρύτερο οπτικό πλαίσιο. Παράλληλα, η λανθασμένη περίπτωση δείχνει ότι ακόμη και όταν το μοντέλο εστιάζει σε σχετικές περιοχές, μπορεί να αποδώσει μια non-melanoma βλάβη στην κλάση melanoma, ειδικά όταν υπάρχουν ύποπτα μορφολογικά χαρακτηριστικά.

4.8 U-Net στο ISIC 2016 για τμηματοποίηση δερματικών βλαβών

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο U-Net για την τμηματοποίηση δερματικών βλαβών στο σύνολο δεδομένων ISIC 2016. Σε αντίθεση με τα πειράματα ταξινόμησης, εδώ το ζητούμενο δεν ήταν η απόδοση μιας ετικέτας σε ολόκληρη την εικόνα, αλλά η πρόβλεψη της περιοχής της βλάβης σε επίπεδο pixel. Με άλλα λόγια, το μοντέλο έπρεπε να διαχωρίσει με όσο γίνεται μεγαλύτερη ακρίβεια τη βλάβη από το υπόβαθρο. Το dataset περιλάμβανε 900 ζεύγη εικόνας-μάσκας. Χρησιμοποιήθηκαν περίπου 720 εικόνες για train και 180 εικόνες για validation. Οι εικόνες μετατράπηκαν σε μέγεθος 320×320 pixels, ενώ κατά την εκπαίδευση εφαρμόστηκαν απλές τεχνικές ενίσχυσης δεδομένων, όπως οριζόντια και κατακόρυφη αναστροφή, μικρές περιστροφές, μετατοπίσεις, κλιμάκωση και ήπιο color jitter. Οι μάσκες μετατράπηκαν σε δυαδική μορφή, ώστε κάθε pixel να ανήκει είτε στη βλάβη είτε στο υπόβαθρο.

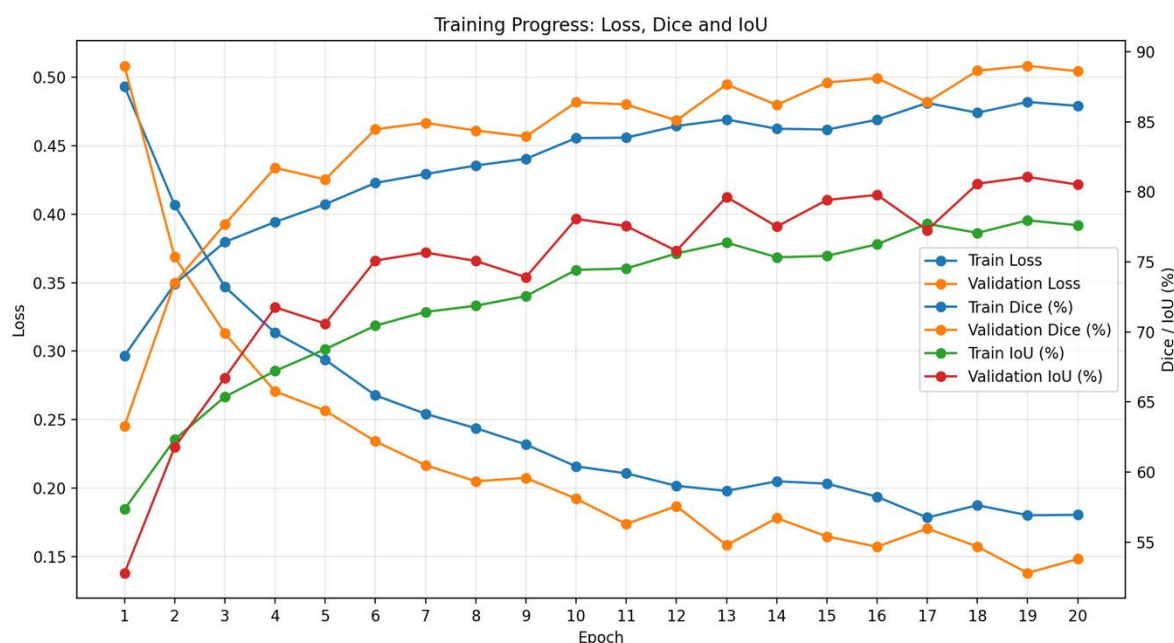
Το μοντέλο που χρησιμοποιήθηκε ήταν ένα U-Net με μία κλασική encoder-decoder δομή. Η εκπαίδευση πραγματοποιήθηκε για 20 εποχές, με optimizer Adam και learning rate $3e-4$, ενώ για τη συνάρτηση απώλειας χρησιμοποιήθηκε ένας συνδυασμός BCEWithLogitsLoss και soft Dice loss. Πιο συγκεκριμένα, η τελική loss ορίστηκε ως $0.3 \times \text{BCE} + 0.7 \times \text{soft Dice loss}$, ώστε το μοντέλο να δίνει μεγαλύτερη έμφαση στην ακριβή επικάλυψη της προβλεπόμενης μάσκας με τη μάσκα αναφοράς. Η βασική μετρική επιλογής του καλύτερου μοντέλου ήταν το validation IoU. Κατά τη διάρκεια της εκπαίδευσης παρατηρήθηκε σταδιακή βελτίωση των επιδόσεων. Στην 1η εποχή, το μοντέλο πέτυχε $\text{train_loss} = 0.4933$, $\text{train_dice} = 68.31\%$, $\text{train_iou} = 57.34\%$, ενώ στο validation set είχε $\text{val_loss} = 0.5084$, $\text{val_dice} = 63.29\%$ και $\text{val_iou} = 52.79\%$. Με την πρόοδο της εκπαίδευσης, οι μετρικές βελτιώθηκαν σημαντικά. Στην 10η εποχή το validation IoU είχε ανέβει σε 78.07% , ενώ στην 19η εποχή το μοντέλο πέτυχε την καλύτερη συνολική επίδοσή του με $\text{val_loss} = 0.1379$, $\text{val_dice} = 89\%$ και $\text{val_iou} = 81.07\%$. Η 19η εποχή αποθηκεύτηκε ως το καλύτερο checkpoint του πειράματος.

Epoch	Train loss	Train Dice (%)	Train IoU (%)	Val loss	Val Dice (%)	Val IoU (%)
1	0.4933	68.31%	57.34%	0.5084	63.29%	52.79%
4	0.3134	77.85%	67.21%	0.2706	81.72%	71.76%
10	0.2158	83.84%	74.43%	0.1922	86.40%	78.07%
16	0.1936	85.15%	76.26%	0.1571	88.12%	79.79%

Epoch	Train loss	Train Dice (%)	Train IoU (%)	Val loss	Val Dice (%)	Val IoU (%)
19	0.1801	86.42%	77.97%	0.1379	89.00%	81.07%
20	0.1804	86.14%	77.62%	0.1482	88.61%	80.52%

Πίνακας 4.25. Βασικές μετρικές εκπαίδευσης του U-Net στο ISIC 2016

Η πορεία αυτή δείχνει ότι το μοντέλο βελτιώθηκε αρκετά σταθερά κατά τις πρώτες εποχές, ενώ από τη μέση της εκπαίδευσης και μετά οι μεταβολές ήταν μικρότερες αλλά ακόμη χρήσιμες. Είναι χαρακτηριστικό ότι μετά την 13η εποχή το validation IoU κινήθηκε σε υψηλότερα επίπεδα, με τιμές 0.7964 στην epoch 13, 79.79% στην epoch 16, 80.58% στην epoch 18 και 81.07% στην epoch 19. Η συμπεριφορά αυτή δείχνει ότι το μοντέλο συνέχισε να βελτιώνεται, χωρίς να εμφανιστεί έντονη αποσταθεροποίηση του validation συνόλου.



Εικόνα 4.39. Πορεία εκπαίδευσης του U-Net στο ISIC 2016

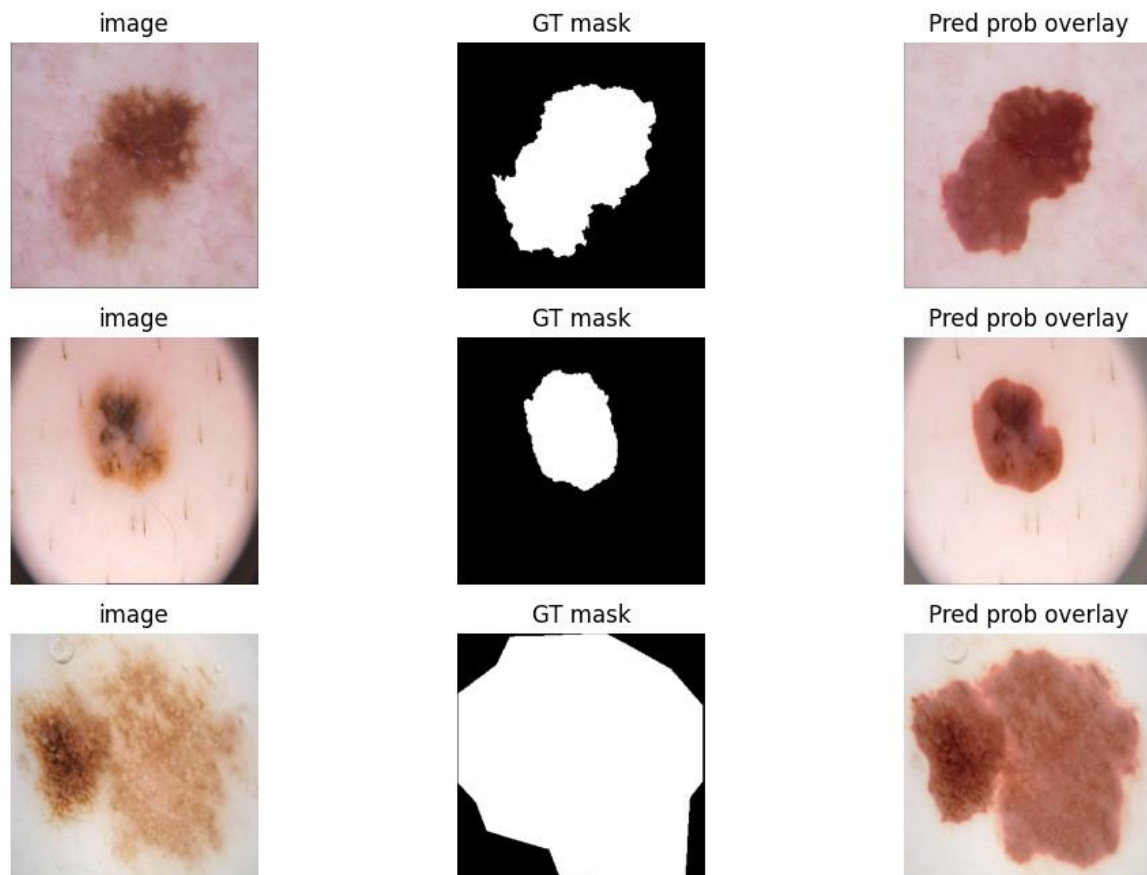
Μετά την εκπαίδευση πραγματοποιήθηκε και βελτιστοποίηση του threshold στο validation set. Εξετάστηκαν οι τιμές 0.3, 0.4, 0.5, 0.6 και 0.7. Τα καλύτερα αποτελέσματα εμφανίστηκαν στο threshold 0.3, όπου το μοντέλο πέτυχε val_dice = 89.22% και val_iou = 81.36%. Για το default threshold 0.5, οι αντίστοιχες τιμές ήταν val_dice = 88.61% και val_iou = 80.52%. Αυτό δείχνει ότι, στην πράξη, ένα ελαφρώς πιο χαμηλό threshold βελτιώνει τη δυαδική μάσκα της πρόβλεψης, κυρίως ως προς την επικάλυψη με τη μάσκα αναφοράς.

Threshold	Val Dice (%)	Val IoU (%)
0.3	89.22%	81.36%
0.4	88.93%	80.97%
0.5	88.61%	80.52%

Threshold	Val Dice (%)	Val IoU (%)
0.6	88.24%	80.02%
0.7	87.65%	79.23%

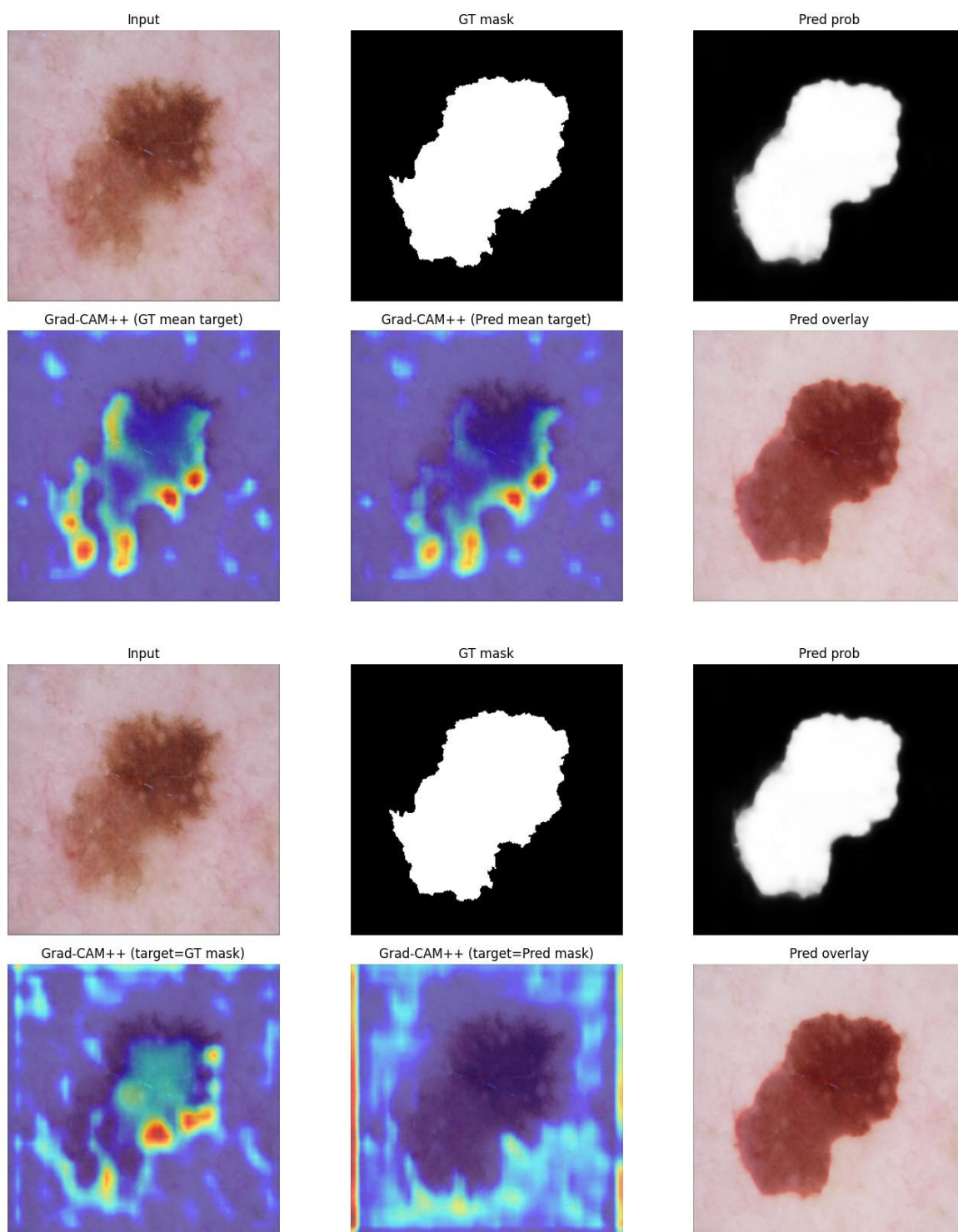
Πίνακας 4.26. Επίδραση του threshold στις validation μετρικές του U-Net

Στην ποιοτική αξιολόγηση του μοντέλου εξετάστηκαν αρχικά ενδεικτικά παραδείγματα από το validation set. Οι εικόνες παρουσιάζουν την αρχική εικόνα, τη μάσκα αναφοράς και την προβλεπόμενη μάσκα. Σε πολλές περιπτώσεις, το μοντέλο κατάφερε να ακολουθήσει αρκετά καλά το σχήμα της βλάβης, ακόμη και όταν τα όριά της δεν ήταν πλήρως καθαρά. Η οπτική αυτή σύγκριση είναι σημαντική, γιατί οι αριθμητικές μετρικές από μόνες τους δεν δείχνουν πάντα αν το μοντέλο χάνει μικρά τμήματα της βλάβης ή αν προσθέτει ψευδώς περιοχές γύρω από αυτή.



Εικόνα 4.40. Ενδεικτικά παραδείγματα τμηματοποίησης από το U-Net στο validation set

Για την ερμηνεία της πρόβλεψης εφαρμόστηκε στη συνέχεια Grad-CAM++. Δημιουργήθηκαν δύο βασικές εκδοχές CAM: μία με στόχο τη ground truth mask και μία με στόχο την predicted mask. Τα στατιστικά των χαρτών ενεργοποίησης έδειξαν ότι το μοντέλο παράγει αξιοποιήσιμη εστίαση γύρω από την περιοχή της βλάβης, τόσο ως προς τη μάσκα αναφοράς όσο και ως προς την παραγόμενη μάσκα.

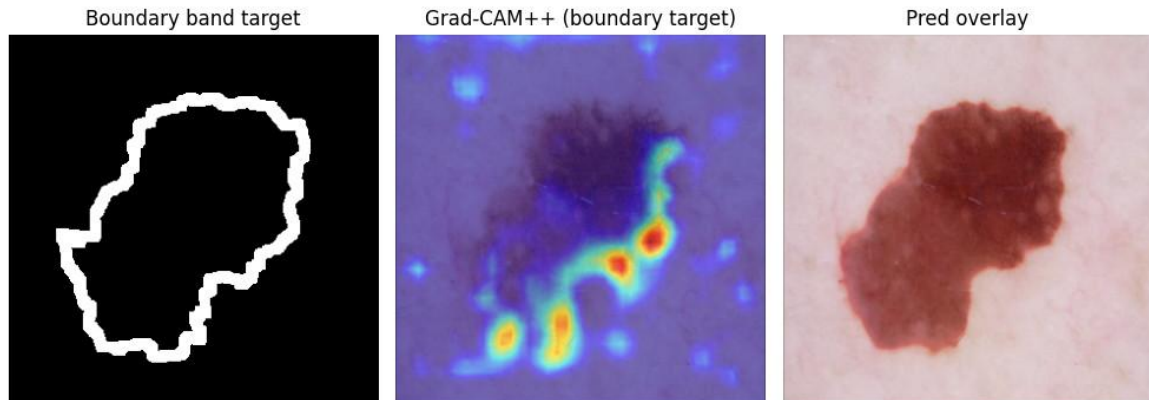


Εικόνα 4.41. Grad-CAM++ ανάλυση για παράδειγμα τμηματοποίησης στο U-Net

Το Grad-CAM++ δεν αποτελεί χάρτη ακριβούς οριοθέτησης της βλάβης. Δείχνει περισσότερο τις περιοχές της εικόνας που συνέβαλαν στην παραγωγή της πρόβλεψης. Άρα, όταν ο χάρτης ενεργοποίησης εμφανίζεται κοντά ή γύρω από τη βλάβη, αυτό είναι θετικό,

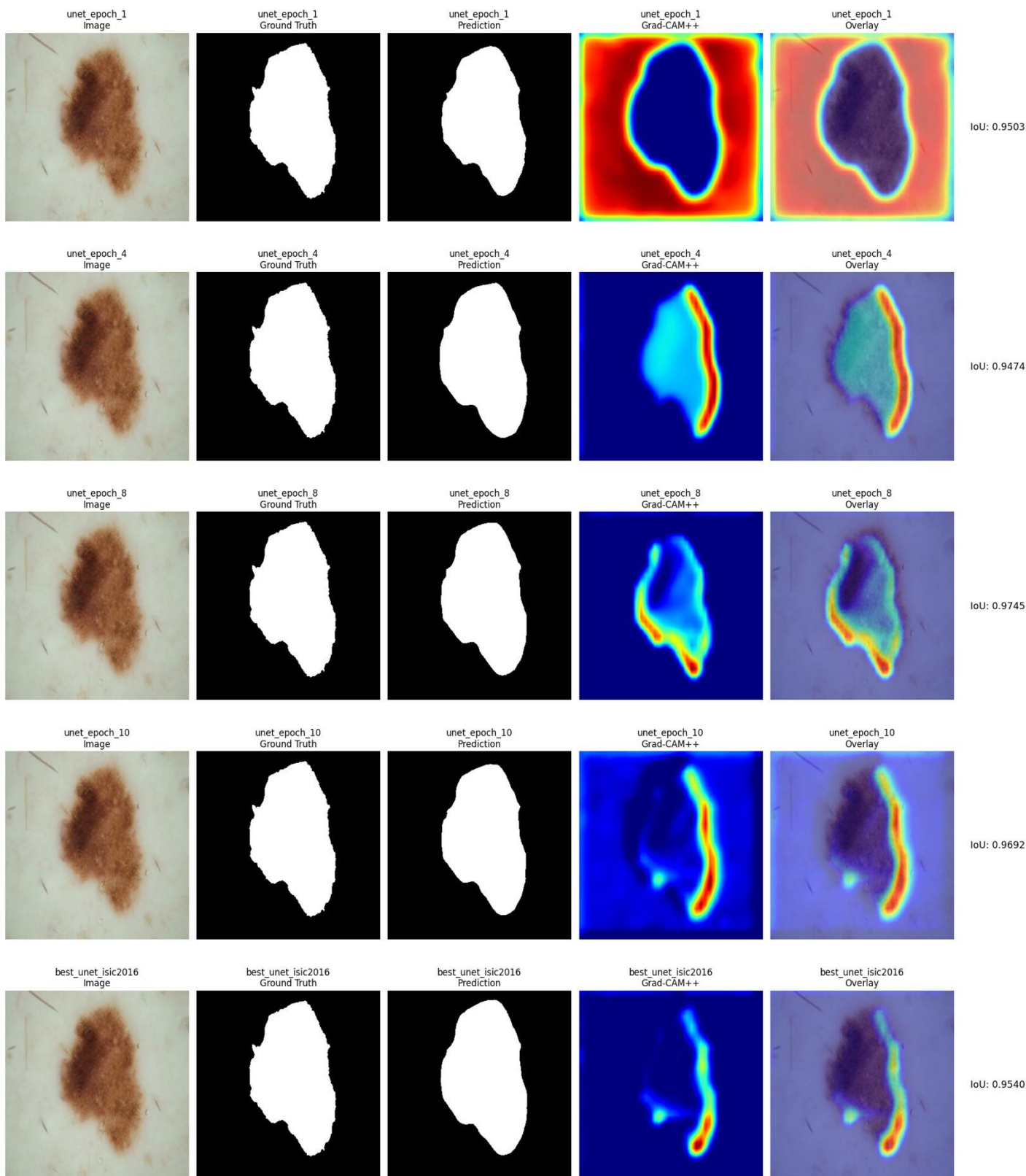
αλλά δεν σημαίνει ότι συμπίπτει υποχρεωτικά με τη μάσκα. Με άλλα λόγια, το Grad-CAM++ λειτουργεί ως ένδειξη προσοχής και όχι ως δεύτερη μάσκα segmentation.

Στη συνέχεια εφαρμόστηκε και μία πιο εστιασμένη μορφή CAM στα όρια της βλάβης, η οποία αποτυπώνεται ως boundary CAM. Το boundary CAM είναι χρήσιμο γιατί βοηθά να φανεί αν το μοντέλο επηρεάζεται κυρίως από το εσωτερικό της βλάβης ή και από τα περιγράμμάτα της. Αυτό έχει ενδιαφέρον στις περιπτώσεις όπου το segmentation δυσκολεύεται να διαχωρίσει σωστά τα όρια από το γύρω δέρμα.



Εικόνα 4.42. Boundary CAM ανάλυση για την περιοχή των ορίων της βλάβης

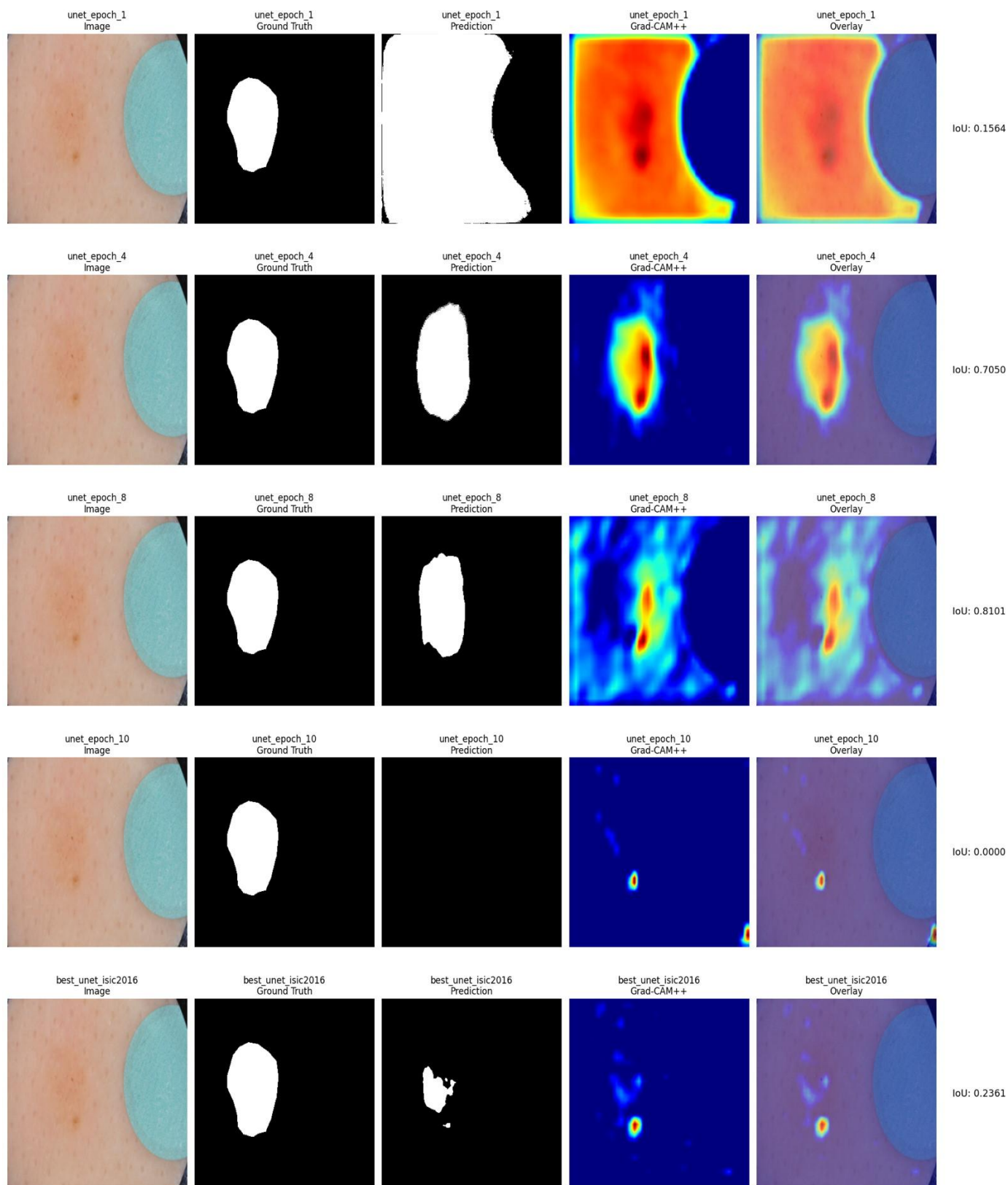
Ένα πολύ χρήσιμο σημείο της αξιολόγησης ήταν η επιλογή της καλύτερης και της δυσκολότερης περίπτωσης στο validation set, με βάση το IoU του καλύτερου μοντέλου. Η καλύτερη περίπτωση είχε $\text{IoU} = 95.40\%$, ενώ η δυσκολότερη είχε $\text{IoU} = 23.61\%$. Η διαφορά αυτή δείχνει ότι, παρά την καλή συνολική επίδοση του μοντέλου, υπάρχουν ακόμα περιπτώσεις όπου η πρόβλεψη υπολείπεται αρκετά της μάσκας αναφοράς. Για την καλύτερη περίπτωση, έγινε across-epoch ανάλυση με checkpoints από τις εποχές 1, 4, 8, 10 και το best model. Οι αντίστοιχες τιμές IoU ήταν: 95.03% στην epoch 1, 94.74% στην epoch 4, 97.45% στην epoch 8, 96.92% στην epoch 10 και 95.40% στο best model. Αυτό έχει ιδιαίτερο ενδιαφέρον, γιατί δείχνει ότι μια μεμονωμένη εικόνα δεν ακολουθεί υποχρεωτικά την ίδια ακριβώς πορεία με τη συνολική validation επίδοση. Για το συγκεκριμένο δείγμα, η epoch 8 είχε το υψηλότερο ατομικό IoU, παρότι το συνολικά καλύτερο checkpoint του μοντέλου ήταν η epoch 19.



Εικόνα 4.43. Across-epoch ανάλυση για την καλύτερη περίπτωση τμηματοποίησης

Η παρατήρηση αυτή έχει σημασία, γιατί δείχνει ότι η συνολική βελτίωση του μοντέλου δεν σημαίνει πάντα βελτίωση σε κάθε μεμονωμένο παράδειγμα. Ορισμένες εικόνες μπορεί να αποδίδονται πολύ καλά ήδη από τις πρώτες εποχές, ενώ σε μεταγενέστερα checkpoints η πρόβλεψη να αλλάζει ελαφρώς χωρίς να βελτιώνεται απαραίτητα το ατομικό IoU.

Στη δυσκολότερη περίπτωση η across-epoch ανάλυση ήταν ακόμη πιο αποκαλυπτική. Οι τιμές IoU για τις εποχές 1, 4, 8, 10 και το best model ήταν 15.64%, 70.50%, 81.01%, 0% και 23.61% αντίστοιχα. Εδώ φαίνεται μια έντονα ασταθής συμπεριφορά. Το ίδιο παράδειγμα βελτιώνεται θεαματικά μέχρι την epoch 8, καταρρέει στην epoch 10 και στο τέλος ανακτά μόνο ένα μέρος της επίδοσής του στο καλύτερο συνολικό checkpoint. Αυτό είναι ένα πολύ ενδιαφέρον εύρημα, γιατί δείχνει ότι οι δύσκολες περιπτώσεις μπορεί να παραμένουν ασταθείς ακόμα και όταν το μοντέλο βελτιώνεται συνολικά.



Εικόνα 4.44. Across-epoch ανάλυση για τη δυσκολότερη περίπτωση τμηματοποίησης

Συνολικά, το πείραμα με το U-Net στο ISIC 2016 έδειξε ότι το μοντέλο μπορεί να πετύχει αρκετά καλή απόδοση στην τμηματοποίηση δερματικών βλαβών. Η καλύτερη συνολική επίδοση στο validation set εμφανίστηκε στην 19η εποχή, με $\text{val_dice} = 89\%$ και $\text{val_iou} = 81.07\%$, ενώ η βελτιστοποίηση του threshold έδειξε ότι το 0.3 ήταν η καταλληλότερη τιμή για τη δυαδική μάσκα, ανεβάζοντας το validation IoU σε 81.36%. Τα ποιοτικά παραδείγματα έδειξαν ότι το μοντέλο καταφέρνει συνήθως να εντοπίζει σωστά την περιοχή της βλάβης, αν και σε δύσκολες περιπτώσεις η σταθερότητα της πρόβλεψης μειώνεται.

4.9 Πείραμα μεταφοράς μάθησης από HAM10000 σε PH2 με ResNet50, Grad-CAM++ και SHAP

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο ResNet50 για τη δυαδική ταξινόμηση δερματικών βλαβών σε non-melanoma και melanoma. Η βασική ιδέα ήταν η εφαρμογή μεταφοράς μάθησης από το μεγαλύτερο σύνολο δεδομένων HAM10000 προς το μικρότερο σύνολο PH2. Αρχικά, το μοντέλο εκπαιδεύτηκε στο HAM10000, ώστε να μάθει γενικά χαρακτηριστικά δερματοσκοπικών εικόνων, και στη συνέχεια έγινε fine-tuning στο PH2. Η διαδικασία αυτή είναι χρήσιμη, επειδή το PH2 περιλαμβάνει μόνο 200 εικόνες και η απευθείας εκπαίδευση ενός βαθιού μοντέλου σε τόσο μικρό σύνολο θα ήταν πιο ευάλωτη σε υπερεκπαίδευση.

Στο πρώτο στάδιο, το HAM10000 μετατράπηκε σε δυαδικό πρόβλημα. Η κλάση melanoma αντιστοιχούσε στην ετικέτα 1, ενώ όλες οι υπόλοιπες κατηγορίες ομαδοποιήθηκαν ως non-melanoma, δηλαδή ετικέτα 0. Το dataset χωρίστηκε σε 7.010 εικόνες εκπαίδευσης, 1.502 εικόνες επικύρωσης και 1.503 εικόνες ελέγχου. Στο training set υπήρχαν 6.231 non-melanoma και 779 melanoma εικόνες. Στο validation set υπήρχαν 1.335 non-melanoma και 167 melanoma, ενώ στο test set υπήρχαν 1.336 non-melanoma και 167 melanoma εικόνες. Η κατανομή αυτή δείχνει έντονη ανισορροπία υπέρ της non-melanoma κλάσης, γι' αυτό χρησιμοποιήθηκε $\text{pos_weight} = 7.9987$.

Split	Σύνολο	Non-melanoma	Melanoma
Train	7010	6231	779
Validation	1502	1335	167
Test	1503	1336	167

Πίνακας 4.27. Κατανομή των δεδομένων HAM10000 στο στάδιο προεκπαίδευσης

Η προεκπαίδευση στο HAM10000 πραγματοποιήθηκε για 8 εποχές. Στις δύο πρώτες εποχές εκπαιδεύτηκε κυρίως το τελικό ταξινομητικό τμήμα του μοντέλου, ενώ στη συνέχεια έγινε unfreezing του πλήρους ResNet50, ώστε να προσαρμοστούν και τα βαθύτερα επίπεδα στα δερματοσκοπικά δεδομένα. Η καλύτερη επίδοση ως προς το validation AUC εμφανίστηκε στην 6η εποχή, με $\text{val_auc} = 0.9154$, $\text{val_acc} = 77.50\%$ και $\text{val_f1} = 47.19\%$. Στις επόμενες εποχές η accuracy συνέχισε να ανεβαίνει, αλλά το AUC δεν βελτιώθηκε.

Epoch	Train loss	Val loss	Val accuracy (%)	Val F1 (%)	Val AUC (%)
1	1.3253	1.2951	44.07%	28.33%	82.24%
3	0.9634	1.0183	55.73%	33.17%	88.48%
4	0.7732	0.8293	64.91%	37.49%	90.11%
5	0.6139	0.8179	69.71%	40.83%	91.29%
6	0.5105	0.7335	77.50%	47.19%	91.54%
8	0.3430	0.7471	85.22%	55.78%	91.53%

Πίνακας 4.28. Πορεία προεκπαίδευσης του ResNet50 στο HAM10000

Στο test set του HAM10000, το καλύτερο checkpoint της 6ης εποχής πέτυχε Test Accuracy = 79.77%, Test Precision = 34.81%, Test Recall = 94.01%, Test F1 = 50.81% και Test AUC = 94.25%. Η εικόνα αυτή δείχνει ότι το μοντέλο ήταν πολύ ευαίσθητο στην ανίχνευση melanoma, καθώς εντόπισε το μεγαλύτερο μέρος των θετικών περιστατικών. Από την άλλη πλευρά, η χαμηλή precision δείχνει ότι αρκετές non-melanoma εικόνες ταξινομήθηκαν λανθασμένα ως melanoma. Αυτό είναι συνηθισμένο σε ανισόρροπα ιατρικά προβλήματα, ειδικά όταν το μοντέλο ρυθμίζεται ώστε να μη χάνει θετικές περιπτώσεις.

Μετρική	Τιμή
Best epoch	6
Test Accuracy	79.77%
Test Precision	34.81%
Test Recall	94.01%
Test F1-score	50.81%
Test AUC	94.25%

Πίνακας 4.29. Test αποτελέσματα στο HAM10000 πριν τη μεταφορά στο PH2

Στο δεύτερο στάδιο, το καλύτερο μοντέλο από το HAM10000 χρησιμοποιήθηκε ως αρχικό σημείο για fine-tuning στο PH2. Το PH2 μετατράπηκε επίσης σε δυαδικό πρόβλημα, όπου οι common και atypical nevus ορίστηκαν ως non-melanoma, ενώ τα melanomas αποτέλεσαν τη θετική κλάση. Το τελικό PH2 dataframe περιλάμβανε 200 εικόνες. Το dataset χωρίστηκε σε 140 εικόνες εκπαίδευσης, 30 εικόνες επικύρωσης και 30 εικόνες ελέγχου. Στο training set υπήρχαν 112 non-melanoma και 28 melanoma εικόνες. Στο validation set υπήρχαν 24 non-melanoma και 6 melanoma, ενώ η ίδια κατανομή υπήρχε και στο test set. Επειδή η melanoma κλάση ήταν μικρότερη, χρησιμοποιήθηκε PH2 pos_weight = 4.0.

Split	Σύνολο	Non-melanoma	Melanoma
Train	140	112	28
Validation	30	24	6
Test	30	24	6

Πίνακας 4.30. Κατανομή των δεδομένων PH2 στο στάδιο fine-tuning

Το fine-tuning στο PH2 πραγματοποιήθηκε για 10 εποχές. Μετά τις δύο πρώτες εποχές έγινε ξανά unfreezing του πλήρους μοντέλου. Η απόδοση βελτιώθηκε σταδιακά, κυρίως ως προς το AUC. Στην 1η εποχή, το μοντέλο είχε val_acc = 70.00%, val_f1 = 57.14% και val_auc = 86.11%. Στην 6η εποχή, το AUC ανέβηκε στο 95.83%, ενώ στην 10η εποχή καταγράφηκε η καλύτερη τιμή, val_auc = 96.53%. Το καλύτερο checkpoint, λοιπόν, ήταν της 10ης εποχής.

Epoch	Train loss	Val loss	Val accuracy (%)	Val F1 (%)	Val AUC (%)
1	1.6270	0.8939	70.00%	57.14%	86.11%
3	1.4651	0.7253	70.00%	57.14%	86.81%
4	0.7986	0.6467	76.67%	63.16%	90.28%
5	0.5080	0.6295	76.67%	63.16%	94.44%
6	0.5013	0.5572	80.00%	66.67%	95.83%
10	0.2145	0.5204	80.00%	66.67%	96.53%

Πίνακας 4.31. Πορεία fine-tuning του ResNet50 στο PH2

Μετά το fine-tuning έγινε επιλογή threshold στο validation set. Το καλύτερο threshold ήταν περίπου 0.89, με validation F1 = 83.33%. Τα thresholds 0.89 και 0.90 έδωσαν την ίδια τιμή F1, με accuracy = 93.33%, precision = 83.33%, recall = 83.33% και F1 = 83.33%. Για την τελική αξιολόγηση χρησιμοποιήθηκε το threshold 0.89. Στο test set του PH2, το μοντέλο πέτυχε Test Loss = 0.3690, Test Accuracy = 100.00%, Test Precision = 100.00%, Test Recall = 100.00%, Test F1 = 100.00% και Test AUC = 100.00%. Ο πίνακας σύγχυσης ήταν:

Κατηγορία	Τιμή
True Negatives / Class 0 σωστά	24
False Positives	0
False Negatives	0
True Positives / Class 1 σωστά	6

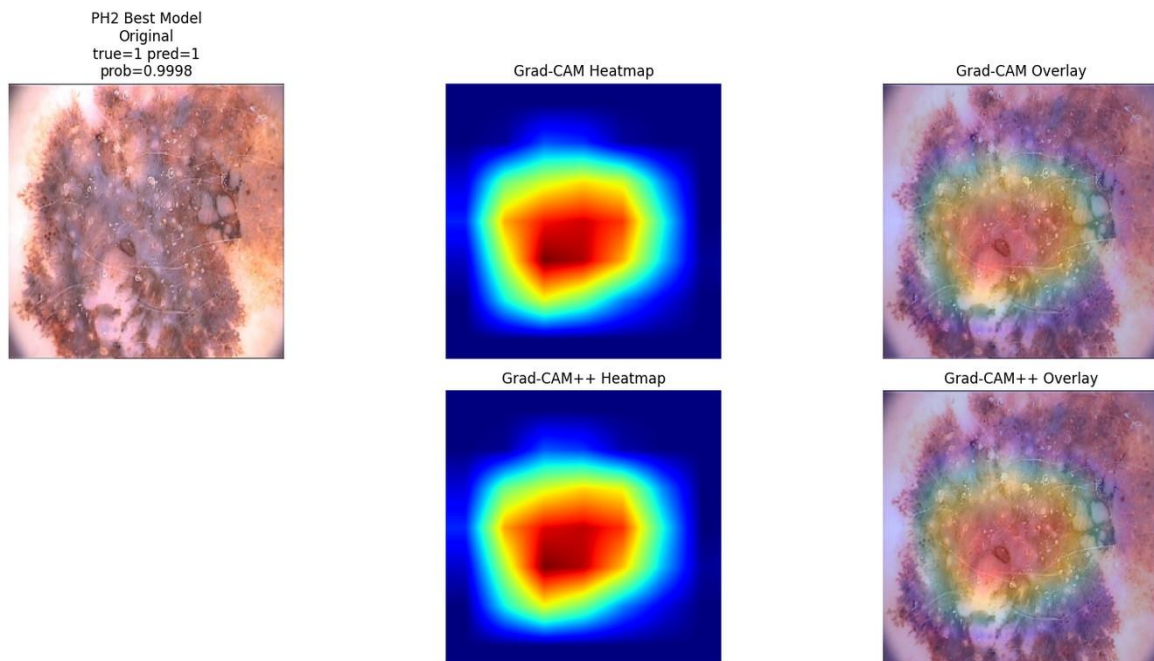
Αυτό σημαίνει ότι και οι 24 non-melanoma εικόνες ταξινομήθηκαν σωστά, ενώ και οι 6 melanoma εικόνες εντοπίστηκαν επίσης σωστά. Στο συγκεκριμένο test split δεν υπήρξε ούτε false positive ούτε false negative. Το αποτέλεσμα είναι πολύ ισχυρό, αλλά πρέπει να διαβαστεί με προσοχή, επειδή το test set του PH2 είναι πολύ μικρό. Με μόνο 30 εικόνες, οι μετρικές μπορούν να αλλάξουν σημαντικά ακόμη και με ένα μεμονωμένο λάθος.

Metric	Value
Best epoch	10
Threshold	0.89
Test Loss	0.3690
Test Accuracy	100%
Test Precision	100%
Test Recall	100%
Test F1	100%
Test AUC	100%

Πίνακας 4.32. Test αποτελέσματα στο PH2 μετά το fine-tuning

Η σύγκριση των δύο σταδίων δείχνει τη διαφορά ανάμεσα στην προεκπαίδευση σε μεγάλο, ανισόρροπο dataset και στο fine-tuning σε μικρότερο dataset. Στο HAM10000, το μοντέλο είχε πολύ υψηλό recall αλλά χαμηλή precision, ενώ στο PH2 πέτυχε πλήρη ταξινόμηση στο συγκεκριμένο test split. Αυτό δεν σημαίνει ότι το μοντέλο είναι γενικά αλάνθαστο· σημαίνει όμως ότι η μεταφορά μάθησης βοήθησε σημαντικά στην προσαρμογή του ResNet50 στο PH2.

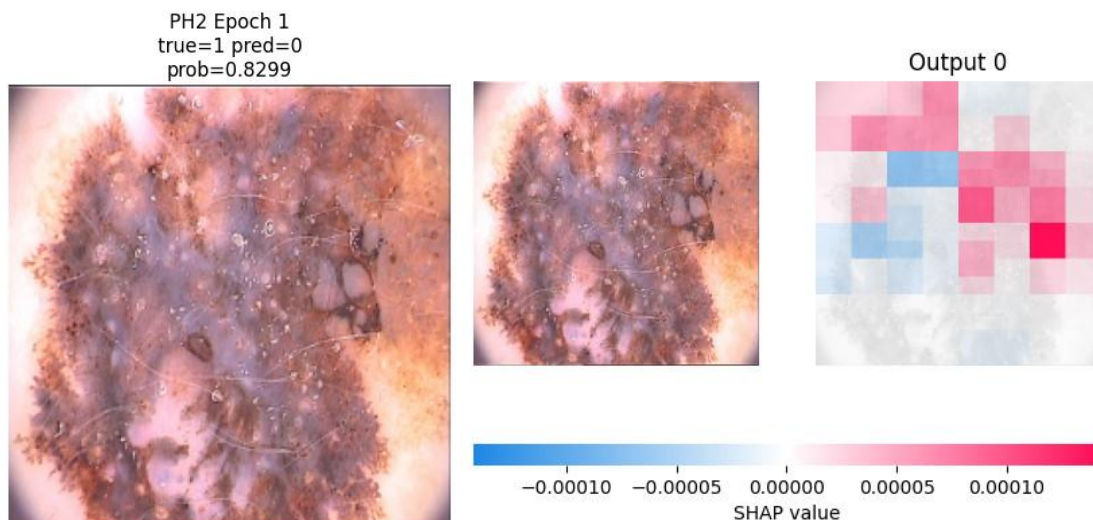
Για την ποιοτική αξιολόγηση εφαρμόστηκαν Grad-CAM++ και SHAP. Η Grad-CAM++ χρησιμοποιήθηκε για να εξεταστεί αν το μοντέλο εστιάζει στην περιοχή της βλάβης και όχι σε άσχετα σημεία του φόντου. Ιδιαίτερο ενδιαφέρον έχει η across-epoch Grad-CAM++ ανάλυση, καθώς δείχνει πώς αλλάζει η εστίαση του μοντέλου από τις πρώτες εποχές μέχρι το καλύτερο checkpoint.



Εικόνα 4.45. Grad-CAM++ across-epoch ανάλυση κατά το fine-tuning στο PH2

Στην across-epoch ανάλυση, η εστίαση του μοντέλου στις αρχικές εποχές είναι πιο ασταθής, ενώ στις μεταγενέστερες εποχές γίνεται πιο συγκεντρωμένη στην περιοχή της δερματικής βλάβης. Αυτό είναι θετικό, γιατί δείχνει ότι η εκπαίδευση δεν βελτίωσε μόνο τις αριθμητικές μετρικές, αλλά και τον τρόπο με τον οποίο το μοντέλο αξιοποιεί την εικόνα. Με άλλα λόγια, το μοντέλο δεν γίνεται απλώς πιο βέβαιο, αλλά φαίνεται να στηρίζει την πρόβλεψή του σε πιο σχετικές οπτικές περιοχές. Η Grad-CAM++ δεν πρέπει να διαβαστεί ως μάσκα τμηματοποίησης. Δεν δείχνει τα ακριβή όρια της βλάβης, αλλά τις περιοχές που επηρέασαν περισσότερο την πρόβλεψη. Γι' αυτό η σωστή ερμηνεία είναι ότι το μοντέλο εστιάζει σε περιοχή σχετική με τη βλάβη, όχι ότι έχει οριοθετήσει πλήρως τη βλάβη.

Συμπληρωματικά, εφαρμόστηκε SHAP με στόχο την melanoma κλάση. Η συγκεκριμένη εικόνα αντιστοιχεί σε περίπτωση benign εικόνας με melanoma signal, δηλαδή σε non-melanoma δείγμα όπου η SHAP ανάλυση αναδεικνύει περιοχές που θα μπορούσαν να συμβάλουν προς τη melanoma κλάση. Αυτό είναι πιο ενδιαφέρον από μια απλή σωστή melanoma περίπτωση, γιατί δείχνει πού μπορεί να δημιουργηθεί τοπικά ύποπτο σήμα, ακόμη και όταν η τελική πρόβλεψη δεν είναι melanoma.



Εικόνα 4.46. Melanoma-targeted SHAP ανάλυση σε benign περίπτωση με τοπικό melanoma signal

Συνολικά, το πείραμα μεταφοράς μάθησης από HAM10000 σε PH2 έδειξε ότι το ResNet50 μπορεί να αξιοποιήσει τη γνώση από ένα μεγαλύτερο dataset και να προσαρμοστεί αποτελεσματικά σε ένα μικρότερο. Στο HAM10000, το μοντέλο πέτυχε AUC = 94.25% και πολύ υψηλό recall = 94.01%, αλλά με αρκετά χαμηλή precision. Μετά το fine-tuning στο PH2, το μοντέλο πέτυχε πλήρη απόδοση στο συγκεκριμένο test split, με Accuracy, Precision, Recall, F1 και AUC = 100%. Παρ' όλα αυτά, το αποτέλεσμα του PH2 πρέπει να αντιμετωπιστεί με επιφύλαξη λόγω του μικρού μεγέθους του test set.

4.10 Πείραμα active learning στο HAM10000 με EfficientNet-B4 και ROI crop

Στο πείραμα αυτό εξετάστηκε μια προσέγγιση active learning για τη δυαδική ταξινόμηση δερματικών βλαβών στο HAM10000. Ο στόχος ήταν να διερευνηθεί αν ένα μοντέλο μπορεί να πετύχει ικανοποιητική απόδοση στην αναγνώριση μελανώματος, χωρίς να χρειάζεται από την αρχή το σύνολο των διαθέσιμων επισημασμένων εικόνων. Το πρόβλημα διαμορφώθηκε ως melanoma έναντι non-melanoma, όπου η κλάση 1 αντιστοιχούσε στο melanoma και η κλάση 0 σε όλες τις υπόλοιπες κατηγορίες.

Το αρχικό αρχείο GroundTruth.csv περιείχε τις στήλες image, MEL, NV, BCC, AKIEC, BKL, DF, VASC. Για τις ανάγκες του πειράματος δημιουργήθηκε δυαδική ετικέτα από τη στήλη MEL. Οι εικόνες με MEL = 1 ορίστηκαν ως melanoma, ενώ όλες οι υπόλοιπες εικόνες ορίστηκαν ως non-melanoma. Συνολικά, το dataset περιλάμβανε 10.015 εικόνες, από τις οποίες 8.902 ήταν non-melanoma και 1.113 melanoma. Η ανισορροπία είναι εμφανής, καθώς τα μελανώματα αποτελούν πολύ μικρότερο μέρος του συνόλου.

Κλάση	Πλήθος εικόνων	Ποσοστό
Non-melanoma	8.902	88.89%
Melanoma	1.113	11.11%
Σύνολο	10.015	100.00%

Πίνακας 4.33. Δυαδική κατανομή του HAM10000 στο πείραμα active learning

Το σύνολο χωρίστηκε σε 7.009 εικόνες εκπαίδευσης, 1.503 εικόνες επικύρωσης και 1.503 εικόνες ελέγχου. Σε κάθε split διατηρήθηκε η δυαδική κατανομή των κλάσεων. Στο training set υπήρχαν 6.230 non-melanoma και 779 melanoma εικόνες. Στο validation set υπήρχαν 1.336 non-melanoma και 167 melanoma, ενώ η ίδια κατανομή εμφανίστηκε και στο test set. Η επιλογή αυτή επιτρέπει πιο καθαρή σύγκριση ανάμεσα στις στρατηγικές active learning, καθώς το validation και το test set παραμένουν σταθερά.

Σύνολο δεδομένων	Non-melanoma	Melanoma	Σύνολο
Train	6.230	779	7.009
Validation	1.336	167	1.503
Test	1.336	167	1.503
Σύνολο	8.902	1.113	10.015

Πίνακας 4.34. Κατανομή train, validation και test στο πείραμα EfficientNet-B4 ROI

Η αρχιτεκτονική που χρησιμοποιήθηκε ήταν το EfficientNet-B4, με προεκπαιδευμένα βάρη. Η εκπαίδευση έγινε σε δύο στάδια. Στο πρώτο στάδιο εκπαιδεύτηκε κυρίως ο ταξινομητής του μοντέλου για 3 εποχές, ενώ στο δεύτερο στάδιο έγινε fine-tuning των τελευταίων blocks του EfficientNet-B4 για 5 επιπλέον εποχές. Με αυτόν τον τρόπο, το μοντέλο πρώτα προσαρμόζει το τελικό επίπεδο ταξινόμησης και στη συνέχεια προσαρμόζει

πιο βαθιά χαρακτηριστικά στις δερματοσκοπικές εικόνες. Σε κάθε γύρο υπολογίστηκε το καλύτερο threshold από το validation set, αντί να χρησιμοποιηθεί σταθερά το όριο 0.50. Αυτό είναι σημαντικό, επειδή σε ανισόρροπα δυαδικά προβλήματα το threshold επηρεάζει έντονα τη σχέση ανάμεσα σε precision και recall.

Πριν από τη διαδικασία active learning εκπαιδεύτηκε ένα full baseline με το 100% του training set, δηλαδή με 7.009 εικόνες. Το baseline πέτυχε στο test set accuracy = 87.23%, precision = 45.06%, recall = 68.26%, F1-score = 54.29% και AUC = 90.23%, με καλύτερο threshold 0.95. Ο πίνακας σύγκρισης του baseline ήταν:

Κατηγορία	Τιμή
True Negatives	1197
False Positives	139
False Negatives	53
True Positives	114

Δηλαδή από τις 1.336 non-melanoma εικόνες, οι 1.197 ταξινομήθηκαν σωστά, ενώ 139 ταξινομήθηκαν λανθασμένα ως melanoma. Από τις 167 melanoma εικόνες, οι 114 εντοπίστηκαν σωστά και 53 χάθηκαν ως non-melanoma. Το baseline, λοιπόν, δίνει ένα ισχυρό σημείο σύγκρισης: έχει την καλύτερη συνολική ισορροπία και το υψηλότερο AUC, αλλά ακόμη και με όλα τα δεδομένα δεν εντοπίζει όλα τα μελανώματα.

Στη συνέχεια εφαρμόστηκαν δύο στρατηγικές active learning: random sampling και margin sampling. Στο random sampling, τα νέα δείγματα που προστίθενται στο labeled set επιλέγονται τυχαία. Στο margin sampling, επιλέγονται δείγματα για τα οποία το μοντέλο είναι πιο αβέβαιο, δηλαδή οι δύο πιθανότερες κλάσεις έχουν μικρή διαφορά μεταξύ τους. Στην πράξη, το margin sampling προσπαθεί να επιλέξει εικόνες που βρίσκονται κοντά στο όριο απόφασης και ίσως είναι πιο χρήσιμες για την επόμενη φάση εκπαίδευσης. Η διαδικασία ξεκίνησε με περίπου 10% του training set, δηλαδή 700 εικόνες. Στον δεύτερο γύρο το labeled set αυξήθηκε σε 2.101 εικόνες, περίπου 30% του training set. Στον τρίτο γύρο έφτασε στις 3.502 εικόνες, δηλαδή περίπου 50% του training set. Αυτό το σχήμα επιτρέπει να φανεί πώς αλλάζει η απόδοση όσο αυξάνεται το ποσοστό των επισημασμένων δεδομένων. Στη στρατηγική random sampling, η απόδοση βελτιώθηκε σταδιακά όσο αυξανόταν το labeled set. Με το 9.99% των δεδομένων, το μοντέλο πέτυχε accuracy = 83.63%, precision = 31.96%, recall = 41.92%, F1-score = 36.27% και AUC = 79.98%. Με το 29.98% των δεδομένων, η απόδοση ανέβηκε σε accuracy = 86.96%, precision = 42.49%, recall = 49.10%, F1-score = 45.56% και AUC = 86.64%. Με το 49.96% των δεδομένων, το μοντέλο πέτυχε accuracy = 86.56%, precision = 42.62%, recall = 60.48%, F1-score = 50.00% και AUC = 88.03%.

Η πορεία του random sampling δείχνει κάτι μάλλον αναμενόμενο: όσο περισσότερα επισημασμένα δεδομένα προστίθενται, τόσο αυξάνεται η ικανότητα του μοντέλου να εντοπίζει melanoma. Το πιο σημαντικό σημείο είναι η αύξηση του recall από 41.92% στο 10% των δεδομένων σε 60.48% στο 50%. Δηλαδή, με περισσότερα δεδομένα το μοντέλο χάνει λιγότερα πραγματικά μελανώματα. Από την άλλη πλευρά, το precision παραμένει

σχετικά χαμηλό, πράγμα που σημαίνει ότι εξακολουθούν να υπάρχουν αρκετά false positives.

Labeled data (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
9.99%	83.63%	31.96%	41.92%	36.27%	79.98%
29.98%	86.96%	42.49%	49.10%	45.56%	86.64%
49.96%	86.56%	42.62%	60.48%	50.00%	88.03%

Πίνακας 4.35. Αποτελέσματα active learning με random sampling

Στη στρατηγική margin sampling, η εικόνα ήταν διαφορετική. Με το 9.99% των δεδομένων, το μοντέλο πέτυχε accuracy = 83.10%, precision = 33.46%, recall = 52.69%, F1-score = 40.93% και AUC = 81.76%. Με το 29.98% των δεδομένων, το accuracy μειώθηκε σε 80.90%, αλλά το recall αυξήθηκε σε 68.86%, με F1-score = 44.49% και AUC = 85.62%. Με το 49.96% των δεδομένων, το margin sampling πέτυχε accuracy = 81.37%, precision = 34.17%, recall = 73.05%, F1-score = 46.56% και AUC = 87.59%. Εδώ φαίνεται πιο καθαρά ο χαρακτήρας του margin sampling. Δεν δίνει την υψηλότερη ακρίβεια, αλλά αυξάνει πολύ περισσότερο το recall για την κλάση melanoma. Στο 50% των δεδομένων, το margin sampling εντόπισε μεγαλύτερο ποσοστό πραγματικών μελανωμάτων από το random sampling και ακόμη και από το full baseline. Συγκεκριμένα, το recall του margin sampling στο 50% ήταν 73.05%, ενώ το recall του full baseline ήταν 68.26%. Αυτό είναι ένα ενδιαφέρον εύρημα, γιατί δείχνει ότι η επιλογή πιο αβέβαιων δειγμάτων μπορεί να κάνει το μοντέλο πιο ευαίσθητο απέναντι στο melanoma.

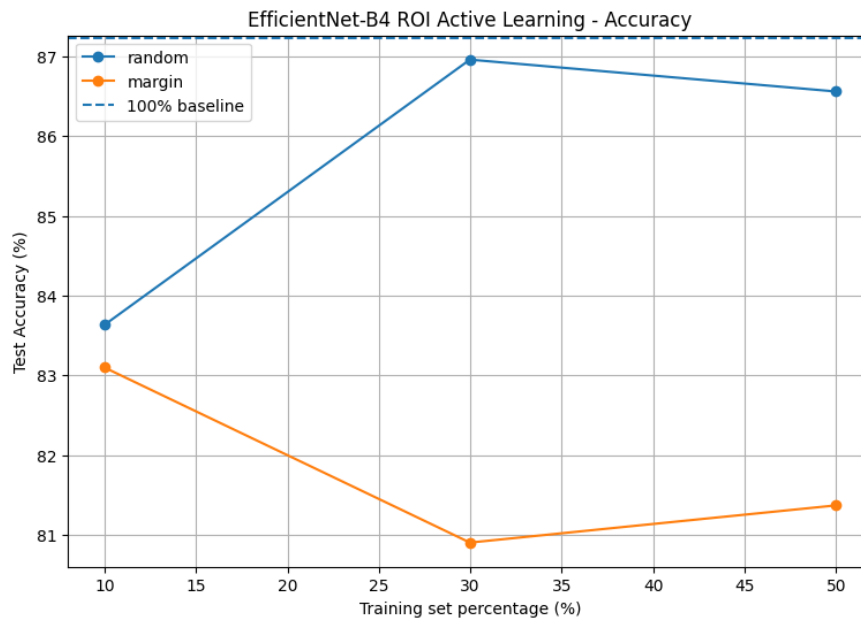
Labeled data (%)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
9.99%	83.10%	33.46%	52.69%	40.93%	81.76%
29.98%	80.90%	33.06%	68.86%	44.49%	85.62%
49.96%	81.37%	34.17%	73.05%	46.56%	87.59%

Πίνακας 4.36. Αποτελέσματα active learning με margin sampling

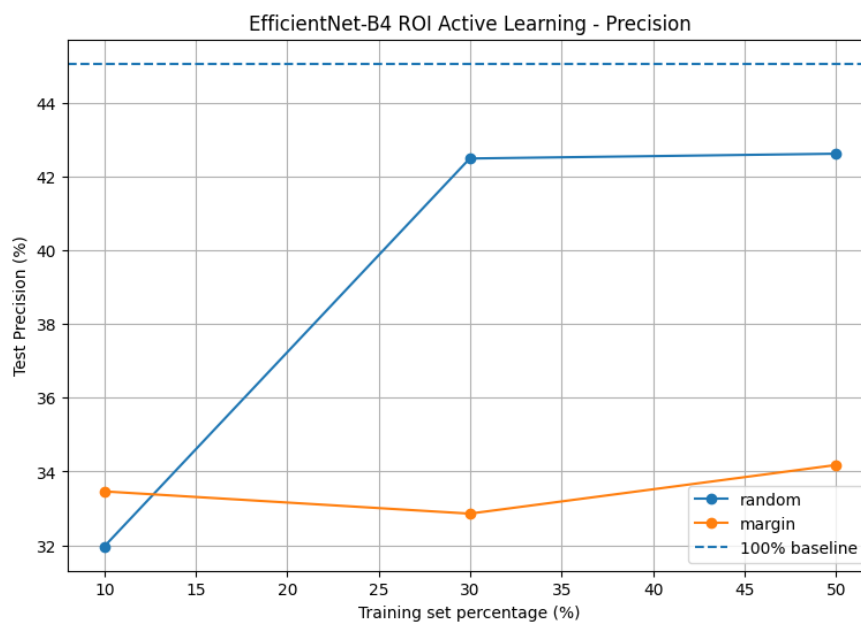
Η σύγκριση των δύο στρατηγικών δείχνει ότι δεν υπάρχει μία απλή απάντηση για το ποια είναι «καλύτερη». Το random sampling είχε υψηλότερη ακρίβεια στους γύρους 30% και 50%, ενώ το margin sampling είχε υψηλότερο recall για melanoma σε όλους τους γύρους. Για παράδειγμα, στο 49.96% του training set, το random sampling είχε accuracy = 86.56% και F1 = 50.00%, ενώ το margin sampling είχε accuracy = 81.37% και F1 = 46.56%. Αν κοιτάξει κανείς μόνο την accuracy ή το F1, το random sampling φαίνεται καλύτερο. Αν όμως το κύριο ζητούμενο είναι να μη χαθούν μελανώματα, το margin sampling έχει πλεονέκτημα λόγω του υψηλότερου recall.

Το γράφημα της ακρίβειας δείχνει ότι το full baseline παραμένει το καλύτερο συνολικά, με 87.23%. Το random sampling πλησιάζει αρκετά το baseline ήδη από το 30% των δεδομένων, ενώ στο 50% φτάνει σε 86.56%. Αντίθετα, το margin sampling κινείται χαμηλότερα σε accuracy. Αυτό δεν σημαίνει απαραίτητα ότι είναι χειρότερο, αλλά ότι

θυσιάζει μέρος της συνολικής ακρίβειας για να αυξήσει την ευαισθησία στην κλάση melanoma.



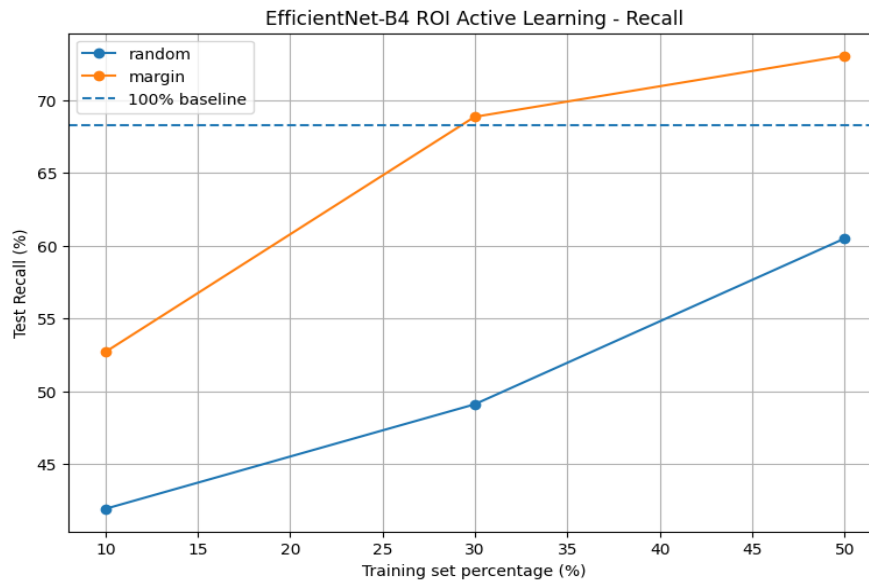
Εκτίνα 4.50. Σύγκριση accuracy ανάμεσα σε random sampling, margin sampling και full baseline



Εκτίνα 4.47. Σύγκριση precision ανάμεσα σε random sampling, margin sampling και full baseline

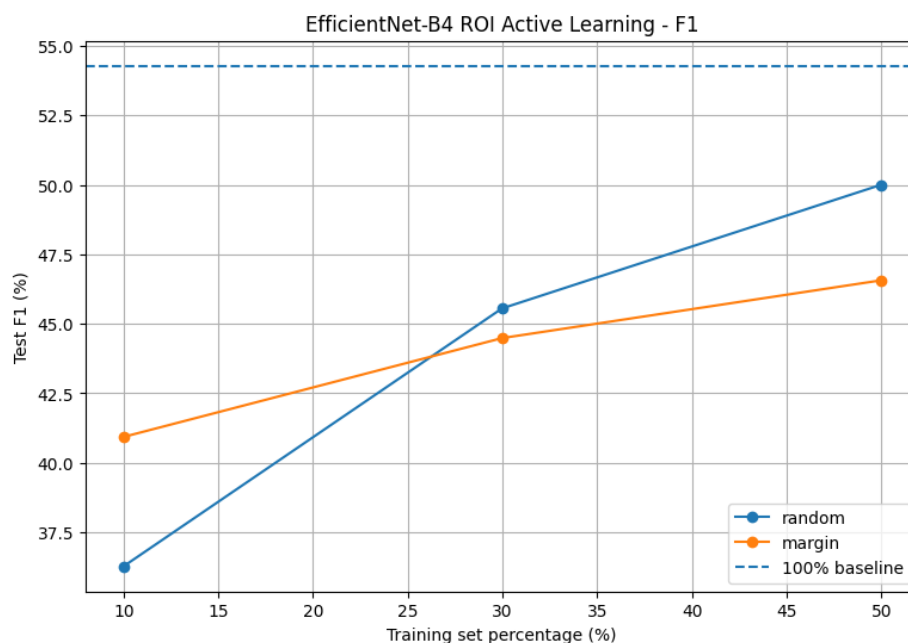
Στο precision plot φαίνεται ότι το full baseline έχει την υψηλότερη τιμή, 45.06%. Το random sampling στο 50% πλησιάζει αρκετά, με 42.62%, ενώ το margin sampling παραμένει χαμηλότερα, με 34.17%. Αυτό δείχνει ότι το margin sampling παράγει περισσότερα false

positives. Σε ιατρικό πλαίσιο, αυτό μπορεί να είναι αποδεκτό μόνο αν ο στόχος είναι ένα ευαίσθητο σύστημα αρχικού ελέγχου και όχι ένα μοντέλο τελικής διάγνωσης.



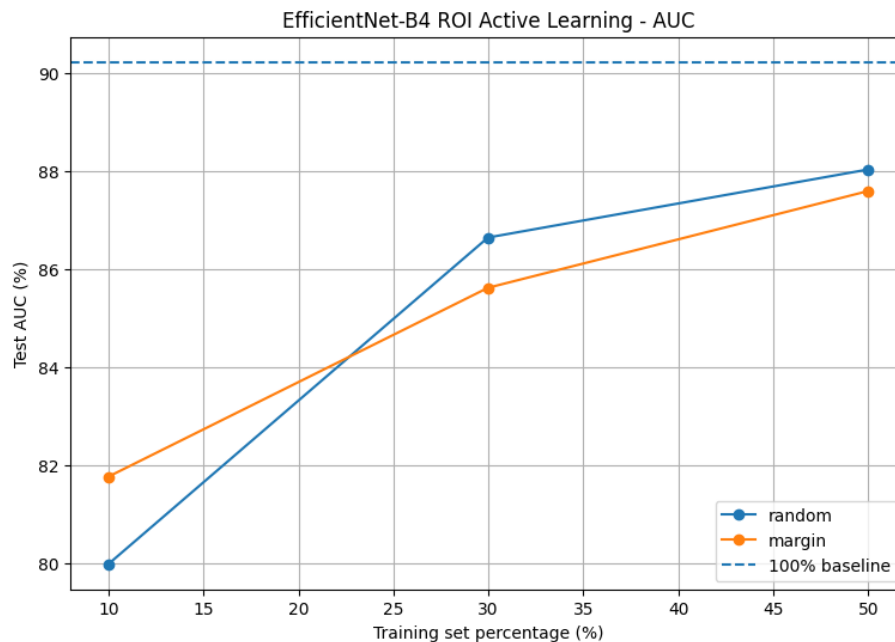
Εικόνα 4.48. Σύγκριση recall ανάμεσα σε random sampling, margin sampling και full baseline

Το recall plot είναι ίσως το πιο σημαντικό για το συγκεκριμένο πείραμα. Το margin sampling ανεβαίνει σταθερά και φτάνει στο 73.05% με περίπου το μισό training set. Το random sampling φτάνει στο 60.48%, ενώ το full baseline βρίσκεται στο 68.26%. Άρα, ως προς την ανίχνευση πραγματικών μελανωμάτων, το margin sampling στο 50% είναι η καλύτερη επιλογή στο συγκεκριμένο πείραμα. Βέβαια, αυτό συνοδεύεται από χαμηλότερο precision, άρα περισσότερες μη μελανωματικές βλάβες χαρακτηρίζονται λανθασμένα ως melanoma.



Εικόνα 4.49. Σύγκριση F1-score ανάμεσα σε random sampling, margin sampling και full baseline

Στο F1-score, που συνδυάζει precision και recall, το full baseline παραμένει πρώτο με 54.29%. Το random sampling στο 50% φτάνει στο 50.00%, ενώ το margin sampling στο 50% φτάνει στο 46.56%. Αυτό δείχνει ότι, αν ζητείται ισορροπία ανάμεσα σε false positives και false negatives, το random sampling στο 50% είναι πιο κοντά στο baseline. Παρ' όλα αυτά, το margin sampling παραμένει σημαντικό επειδή πετυχαίνει μεγαλύτερη ευαισθησία στο melanoma.



Εικόνα 4.50. Σύγκριση AUC ανάμεσα σε random sampling, margin sampling και full baseline

Το AUC plot δείχνει ότι και οι δύο active learning στρατηγικές βελτιώνονται όσο αυξάνεται το labeled set. Το full baseline έχει την υψηλότερη τιμή, 90.23%. Το random sampling στο 50% φτάνει σε 88.03%, ενώ το margin sampling φτάνει σε 87.59%. Οι τιμές αυτές είναι αρκετά κοντά, κάτι που δείχνει ότι, ακόμη και με περίπου το μισό training set, οι active learning στρατηγικές μπορούν να πλησιάσουν τη διαχωριστική ικανότητα του πλήρους baseline.

Συνολικά, το πείραμα με EfficientNet-B4, ROI crop και active learning δείχνει ότι η επιλογή των δεδομένων επηρεάζει ουσιαστικά τη συμπεριφορά του μοντέλου. Το full baseline παραμένει πιο ισορροπημένο, με accuracy = 87.23%, F1-score = 54.29% και AUC = 90.23%. Το random sampling στο 50% πλησιάζει αρκετά το baseline, κυρίως σε accuracy και F1. Το margin sampling, από την άλλη πλευρά, προσφέρει το υψηλότερο recall για melanoma, φτάνοντας στο 73.05% με περίπου το μισό training set. Το βασικό συμπέρασμα είναι ότι το active learning δεν πρέπει να αξιολογείται μόνο με μία μετρική. Αν στόχος είναι η συνολική ισορροπία, το random sampling στο 50% φαίνεται πιο κοντά στο πλήρες baseline. Αν όμως προτεραιότητα είναι να μειωθούν τα χαμένα μελανώματα, το margin sampling δίνει πιο χρήσιμη συμπεριφορά, επειδή εντοπίζει περισσότερες πραγματικές melanoma περιπτώσεις.

4.11 Πείραμα ταξινόμησης melanoma / non-melanoma με ενδιάμεσα δερματοσκοπικά χαρακτηριστικά στο Derm7pt

Στο πείραμα αυτό σχεδιάστηκε και αξιολογήθηκε μια προσέγγιση ταξινόμησης στο σύνολο δεδομένων Derm7pt, με στόχο τη διάκριση ανάμεσα σε melanoma και non-melanoma περιπτώσεις. Η βασική ιδέα ήταν να μην εκπαιδευτεί το μοντέλο μόνο πάνω στην τελική ετικέτα διάγνωσης, αλλά να μάθει παράλληλα και ένα σύνολο ενδιάμεσων δερματοσκοπικών χαρακτηριστικών. Το μοντέλο δεν προβλέπει μόνο αν μια εικόνα ανήκει στην κλάση melanoma ή non-melanoma, αλλά εκπαιδεύεται ταυτόχρονα να αναγνωρίζει δερματοσκοπικά χαρακτηριστικά, όπως pigment network, blue-whitish veil, vascular structures, pigmentation, streaks, dots and globules και regression structures. Τα χαρακτηριστικά αυτά δεν χρησιμοποιούνται ως τελική διάγνωση, αλλά ως βοηθητικοί στόχοι μάθησης που μπορούν να δώσουν πιο πλούσια πληροφορία στο μοντέλο. Το αρχικό αρχείο του Derm7pt περιείχε πληροφορίες για τη διάγνωση, το seven-point score, τα επιμέρους δερματοσκοπικά χαρακτηριστικά και ορισμένα μεταδεδομένα. Για τις ανάγκες του πειράματος κρατήθηκαν οι διαγνώσεις που μπορούσαν να ενταχθούν καθαρά στο δυαδικό σχήμα melanoma / non-melanoma. Η κλάση 0 αντιστοιχούσε σε σπίλους ή μη μελανωματικές περιπτώσεις, ενώ η κλάση 1 αντιστοιχούσε σε melanoma. Μετά το φιλτράρισμα, το τελικό σύνολο περιλάμβανε 827 δείγματα, από τα οποία 575 ήταν non-melanoma και 252 melanoma.

Κλάση	Πλήθος εικόνων	Ποσοστό
Non-melanoma	575	69.53%
Melanoma	252	30.47%
Σύνολο	827	100.00%

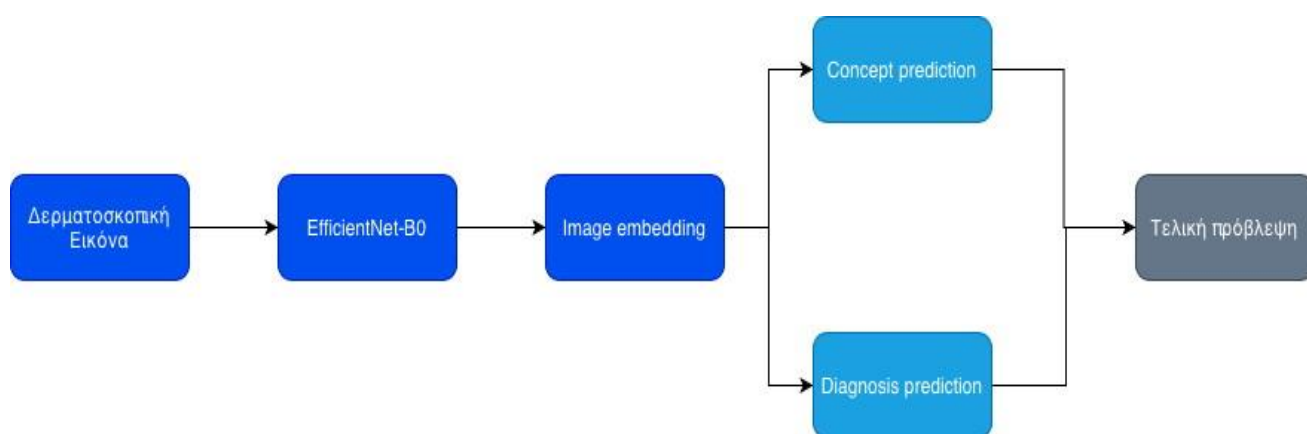
Πίνακας 4.37. Δυαδική κατανομή των δεδομένων στο Derm7pt

Το dataset χωρίστηκε σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου. Στο training set υπήρχαν 402 non-melanoma και 176 melanoma δείγματα. Στο validation set υπήρχαν 86 non-melanoma και 38 melanoma, ενώ στο test set υπήρχαν 87 non-melanoma και 38 melanoma. Η κατανομή αυτή δείχνει μια μέτρια ανισορροπία υπέρ της non-melanoma κλάσης, αλλά όχι τόσο έντονη όσο σε πειράματα με το HAM10000. Αυτό κάνει το Derm7pt κατάλληλο για μια πιο ελεγχόμενη δοκιμή, όπου η τελική διάγνωση μπορεί να εξεταστεί μαζί με ενδιάμεσες οπτικές ενδείξεις.

Σύνολο δεδομένων	Non-melanoma	Melanoma	Σύνολο
Train	402	176	578
Validation	86	38	124
Test	87	38	125
Σύνολο	575	252	827

Πίνακας 4.38. Κατανομή train, validation και test στο πείραμα Derm7pt με ενδιάμεσα χαρακτηριστικά

Η είσοδος του μοντέλου ήταν η δερματοσκοπική εικόνα, η οποία μετατράπηκε σε μέγεθος 224×224 pixels. Ως backbone αξιοποιήθηκε το EfficientNet-B0 με προεκπαιδευμένα βάρη. Πάνω από το backbone προστέθηκαν διαφορετικές έξοδοι: μία για την τελική διάγνωση και μία για την πρόβλεψη των επτά δερματοσκοπικών χαρακτηριστικών. Η εικόνα περνούσε πρώτα από το EfficientNet-B0, στη συνέχεια τα εξαγόμενα χαρακτηριστικά προβάλλονταν σε ένα μικρό embedding και από εκεί τροφοδοτούσαν τόσο το concept head όσο και το diagnostic head. Η αρχιτεκτονική δεν περιορίστηκε σε μία μόνο έξοδο διάγνωσης. Το μοντέλο παρήγαγε προβλέψεις για τα concepts, ενώ η τελική διάγνωση προέκυπτε από συνδυασμό των οπτικών χαρακτηριστικών της εικόνας και των προβλεπόμενων concepts. Αυτή η λογική είναι χρήσιμη, γιατί το μοντέλο δεν πιέζεται απλώς να μάθει μία τελική ετικέτα, αλλά και να αναπτύξει μια ενδιάμεση περιγραφή της εικόνας. Στην πράξη, αυτό δίνει μια πιο αναλυτική εικόνα της συμπεριφοράς του, ακόμη και αν οι προβλέψεις των concepts δεν είναι όλες εξίσου ακριβείς.



Εικόνα 4.51. Ροή του μοντέλου με ενδιάμεσα δερματοσκοπικά χαρακτηριστικά

Η εκπαίδευση πραγματοποιήθηκε για 15 εποχές, με συνάρτηση απώλειας που συνδύαζε τη διάγνωση και τα concepts. Για τη διάγνωση εφαρμόστηκε στάθμιση της θετικής κλάσης, ώστε να ληφθεί υπόψη η μικρότερη συχνότητα των melanoma δειγμάτων στο training set. Το καλύτερο μοντέλο επιλέχθηκε με βάση το validation diagnosis AUC. Η καλύτερη επίδοση καταγράφηκε στην 10η εποχή, όπου το μοντέλο πέτυχε $\text{val_diag_acc} = 83.87\%$, $\text{val_diag_f1} = 71.43\%$, $\text{val_diag_auc} = 88.98\%$ και $\text{val_concept_auc} = 76.95\%$.

Η πορεία της εκπαίδευσης δείχνει ότι το μοντέλο βελτιώθηκε κυρίως μέχρι τη 10η εποχή. Στην 1η εποχή είχε $\text{val_diag_acc} = 69.35\%$, $\text{val_diag_f1} = 60.42\%$, $\text{val_diag_auc} = 76.99\%$ και $\text{val_concept_auc} = 56.65\%$. Στην 5η εποχή οι τιμές ανέβηκαν σε $\text{val_diag_acc} = 81.45\%$, $\text{val_diag_f1} = 70.89\%$, $\text{val_diag_auc} = 84.58\%$ και $\text{val_concept_auc} = 71.30\%$. Στην 9η εποχή το μοντέλο έφτασε σε $\text{val_diag_auc} = 88.27\%$, ενώ στην 10η πέτυχε την καλύτερη τιμή 88.98% . Μετά από αυτό το σημείο, οι μετρικές διάγνωσης παρέμειναν κοντά, αλλά δεν υπήρξε καθαρή βελτίωση της diagnosis AUC.

Epoch	Val diag acc (%)	Val diag F1 (%)	Val diag AUC (%)	Val concept AUC (%)
1	69.35%	60.42%	76.99%	56.65%
5	81.45%	70.89%	84.58%	71.30%

Epoch	Val diag acc (%)	Val diag F1 (%)	Val diag AUC (%)	Val concept AUC (%)
9	83.06%	72.00%	88.26%	75.67%
10	83.87%	71.43%	88.98%	76.95%
15	77.42%	64.10%	88.19%	78.25%

Πίνακας 4.39. Πορεία εκπαίδευσης του μοντέλου στο Derm7pt

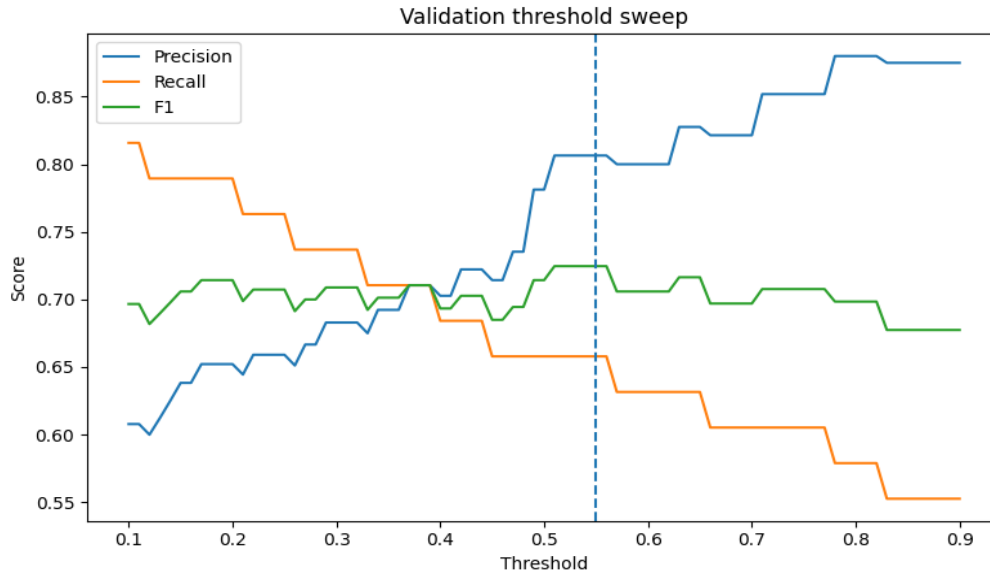
Ένα ενδιαφέρον στοιχείο είναι ότι το val_concept_auc συνέχισε να βελτιώνεται ακόμη και μετά την καλύτερη εποχή για τη διάγνωση. Για παράδειγμα, από 76.95% στην 10η εποχή έφτασε σε περίπου 78.27% στη 12η εποχή και παρέμεινε κοντά σε αυτή την περιοχή μέχρι το τέλος. Αυτό δείχνει ότι η μάθηση των ενδιάμεσων χαρακτηριστικών δεν εξελίσσεται απαραίτητα με τον ίδιο ρυθμό όπως η τελική διάγνωση. Το μοντέλο μπορεί να συνεχίζει να βελτιώνεται στα concepts, χωρίς αυτό να σημαίνει ότι βελτιώνεται και η τελική ταξινόμηση melanoma / non-melanoma. Μετά την εκπαίδευση έγινε επιλογή threshold από το validation set. Το καλύτερο threshold με βάση το validation F1 ήταν περίπου 0.55, με validation F1 = 72.46%, precision = 80.65% και recall = 65.79%. Με αυτό το threshold, το μοντέλο αξιολογήθηκε στο test set. Η τελική απόδοση ήταν Test Accuracy = 82.40%, Test Precision = 76.67%, Test Recall = 60.53%, Test F1 = 67.65% και Test AUC = 88.66%. Ο πίνακας σύγχυσης ήταν:

Κατηγορία	Τιμή
True Negatives (TN)	80
False Positives (FP)	7
False Negatives (FN)	15
True Positives (TP)	23

Αυτό σημαίνει ότι από τα 87 non-melanoma δείγματα, τα 80 ταξινομήθηκαν σωστά και 7 ταξινομήθηκαν λανθασμένα ως melanoma. Από τα 38 melanoma δείγματα, τα 23 εντοπίστηκαν σωστά, ενώ 15 ταξινομήθηκαν λανθασμένα ως non-melanoma. Το μοντέλο, δηλαδή, είχε αρκετά καλή precision για την κλάση melanoma, αλλά το recall παρέμεινε μέτριο. Αυτό δείχνει ότι όταν το μοντέλο προβλέπει melanoma είναι σχετικά αξιόπιστο, αλλά δεν εντοπίζει όλα τα πραγματικά melanoma περιστατικά. Από το classification report φαίνεται ότι η κλάση non-melanoma είχε precision = 84.21%, recall = 91.95% και F1-score = 87.91%. Η κλάση melanoma είχε precision = 76.67%, recall = 60.53% και F1-score = 67.65%. Το macro F1 ήταν 77.78%, ενώ το weighted F1 ήταν 81.75%. Η διαφορά ανάμεσα στις δύο τιμές δείχνει ότι η μεγαλύτερη κλάση non-melanoma βελτιώνει τη συνολική weighted εικόνα, ενώ το macro F1 αποτυπώνει πιο καθαρά τη δυσκολία της melanoma κλάσης.

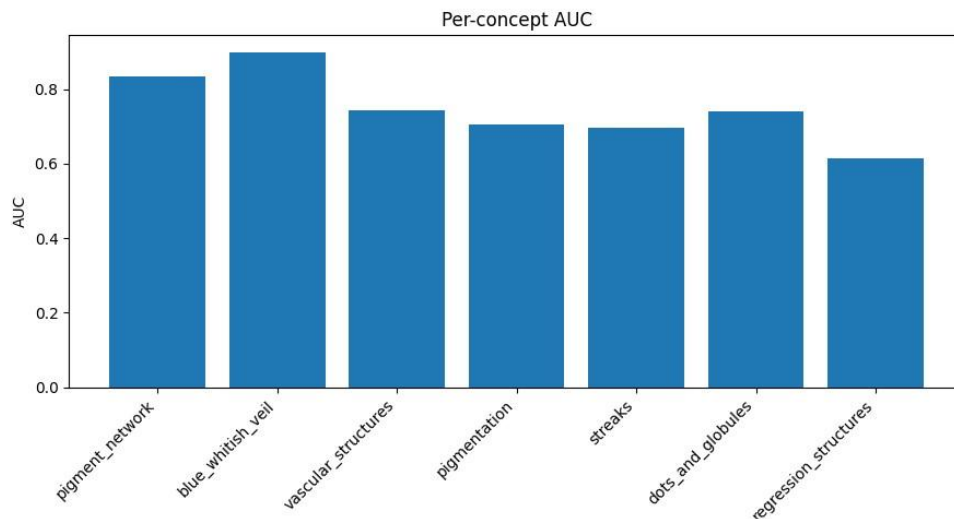
Έγινε επίσης και test threshold sweep. Αν και το threshold 0.55 επιλέχθηκε από το validation set, στο test set το υψηλότερο F1 εμφανίστηκε γύρω από το threshold 0.26, με accuracy = 85.60%, precision = 75.00%, recall = 78.95% και F1 = 76.92%. Αυτό δεν πρέπει να χρησιμοποιηθεί ως κύρια τελική αξιολόγηση, γιατί προκύπτει από το test set και άρα δεν αποτελεί ανεξάρτητη επιλογή threshold. Παρ' όλα αυτά, δείχνει ότι η απόδοση του

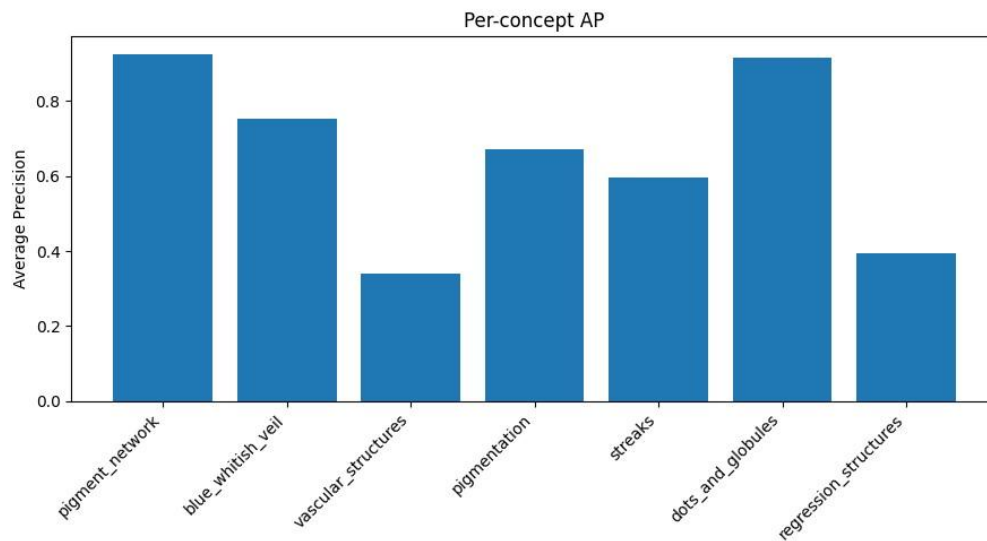
μοντέλου είναι αρκετά ευαίσθητη στο threshold. Με χαμηλότερο threshold αυξάνεται το recall, δηλαδή εντοπίζονται περισσότερα melanomas, αλλά αλλάζει και η ισορροπία με το precision.



Εικόνα 4.52. Επίδραση του threshold στην απόδοση της διάγνωσης

Παράλληλα με την τελική διάγνωση, αξιολογήθηκε και η πρόβλεψη των ενδιάμεσων δερματοσκοπικών χαρακτηριστικών. Το συνολικό Test Concept AUC ήταν 74.73%, ενώ το Test Concept AP ήταν 65.62%. Οι επιδόσεις δεν ήταν ίδιες για όλα τα concepts. Το καλύτερο AUC εμφανίστηκε στο blue-whitish veil, με AUC = 89.99% και AP = 75.20%. Ακολούθησε το pigment network, με AUC = 83.30% και AP = 92.57%. Τα vascular structures είχαν AUC = 74.35%, αλλά πολύ χαμηλότερο AP, 33.96%, στοιχείο που δείχνει ότι η θετική κλάση για αυτό το concept ήταν πιο δύσκολη ή λιγότερο συχνή.

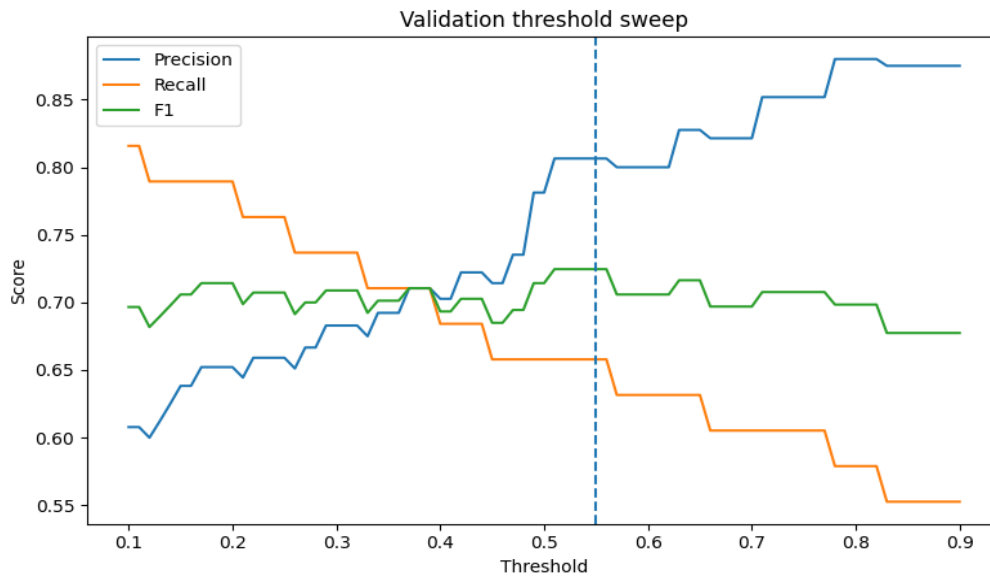




Εικόνα 4.53. Απόδοση του μοντέλου στην πρόβλεψη των δερματοσκοπικών χαρακτηριστικών

Αναλυτικά, το dots and globules είχε $AUC = 73.96\%$ και $AP = 91.48\%$, με πολύ υψηλό F1 στο threshold 0.5, 88.89% . Το pigmentation είχε $AUC = 70.49\%$, ενώ τα streaks είχαν $AUC = 69.56\%$. Το πιο αδύναμο concept ήταν τα regression structures, με $AUC = 61.44\%$ και $F1 = 32.65\%$ στο threshold 0.5. Αυτές οι διαφορές δείχνουν ότι ορισμένα δερματοσκοπικά χαρακτηριστικά είναι πιο εύκολο να μάθει το μοντέλο από την εικόνα, ενώ άλλα παραμένουν αρκετά δύσκολα. Είναι αμφίβολο κατά πόσο όλα τα concepts μπορούν να προβλεφθούν με την ίδια αξιοπιστία από ένα σχετικά μικρό dataset όπως το Derm7pt.

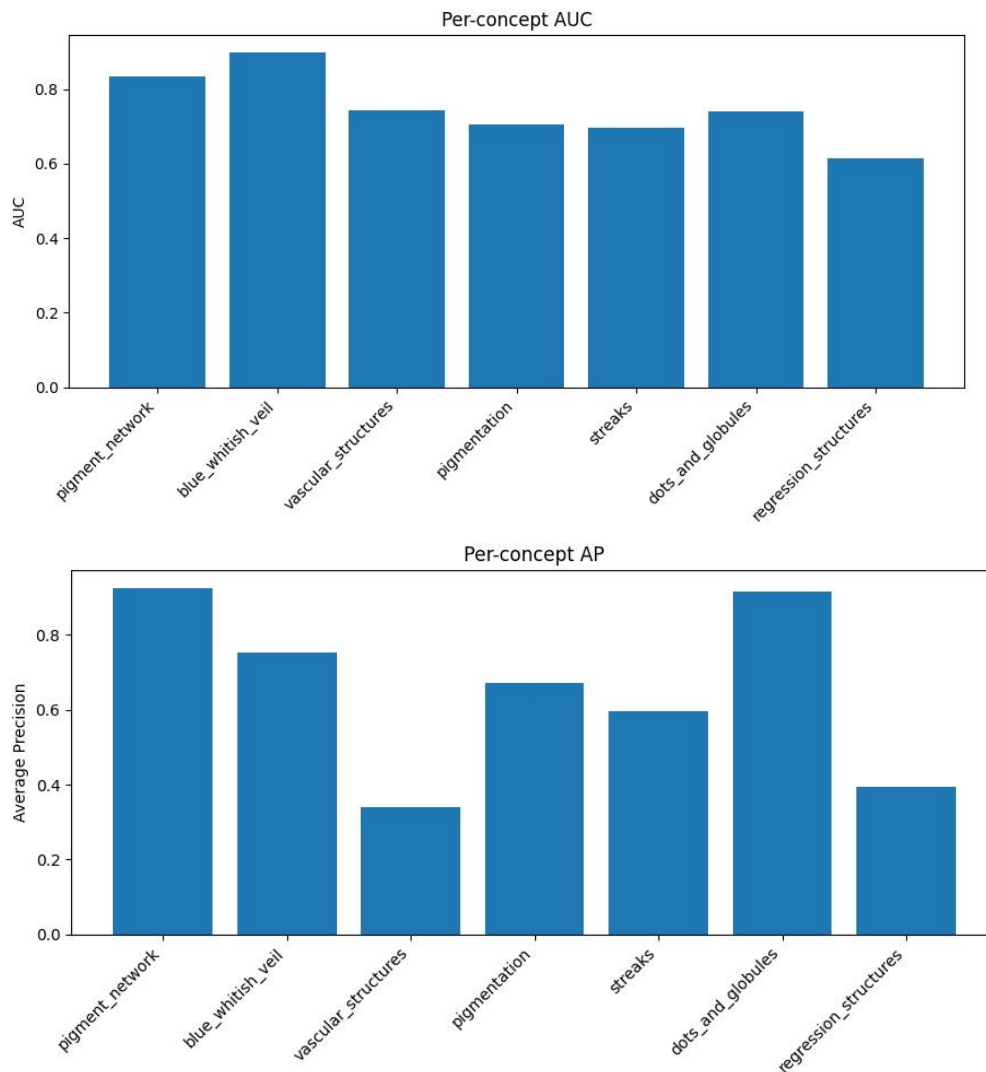
Για την ποιοτική αξιολόγηση επιλέχθηκαν δυο χαρακτηριστικές εικόνες: μία σωστή πρόβλεψη melanoma, μία σωστή πρόβλεψη non-melanoma και μία λανθασμένη πρόβλεψη. Η επιλογή αυτών των περιπτώσεων βοηθά να φανεί όχι μόνο αν το μοντέλο προβλέπει σωστά τη διάγνωση, αλλά και τι προβλέπει για τα ενδιάμεσα χαρακτηριστικά κάθε εικόνας. Στην ορθή περίπτωση melanoma, το μοντέλο προέβλεψε σωστά την κλάση melanoma με πολύ υψηλή πιθανότητα, $Prob\ melanoma = 0.9996$. Η εικόνα αντιστοιχούσε σε περίπτωση melanoma (less than 0.76 mm). Στην ίδια περίπτωση, το μοντέλο προέβλεψε σωστά την παρουσία ορισμένων concepts, όπως pigment network, blue-whitish veil, streaks και dots and globules. Από την άλλη πλευρά, προέβλεψε θετικά και ορισμένα concepts που στην ετικέτα ήταν αρνητικά, όπως pigmentation και regression structures. Αυτό δείχνει ότι, ακόμη και σε μια σωστή διάγνωση, η ενδιάμεση περιγραφή του μοντέλου δεν είναι απόλυτα ταυτισμένη με τις concept ετικέτες.



Εικόνα 4.54. Ανάλυση διάγνωσης και δερματοσκοπικών χαρακτηριστικών για ορθή πρόβλεψη melanoma

Η σωστή αυτή περίπτωση δείχνει τη χρησιμότητα της προσέγγισης. Δεν βλέπουμε μόνο την τελική απόφαση του μοντέλου, αλλά και ποια ενδιαμέσως χαρακτηριστικά θεωρεί παρόντα. Στην πράξη, αυτό δίνει ένα δεύτερο επίπεδο ερμηνείας. Αν η τελική πρόβλεψη είναι melanoma και το μοντέλο ενεργοποιεί concepts που συχνά σχετίζονται με πιο ύποπτες βλάβες, η πρόβλεψη γίνεται πιο αναγνώσιμη. Δεν σημαίνει ότι η πρόβλεψη είναι αυτομάτως κλινικά αποδεκτή, αλλά είναι πιο εύκολο να αναλυθεί σε σχέση με ένα απλό μοντέλο που δίνει μόνο ένα probability.

Στη λανθασμένη περίπτωση, η πραγματική διάγνωση ήταν melanoma, αλλά το μοντέλο προέβλεψε non-melanoma, με Prob melanoma = 0.0044. Πρόκειται για μια σημαντική false negative περίπτωση, γιατί το μοντέλο δεν ήταν απλώς οριακά λάθος· απέδωσε πολύ χαμηλή πιθανότητα στην κλάση melanoma. Στα concepts, το μοντέλο προέβλεψε σωστά την παρουσία του pigment network και των dots and globules, αλλά δεν προέβλεψε σωστά την παρουσία των pigmentation και streaks, τα οποία ήταν θετικά στην ετικέτα. Αυτό μπορεί να εξηγηθεί εν μέρει γιατί η τελική πρόβλεψη οδηγήθηκε προς non-melanoma.



Εικόνα 4.55. Ανάλυση διάγνωσης και δερματοσκοπικών χαρακτηριστικών για false negative melanoma

Η false negative περίπτωση είναι ίσως το πιο κρίσιμο παράδειγμα του πειράματος. Δείχνει ότι το μοντέλο μπορεί να αποτύχει όχι μόνο στην τελική διάγνωση, αλλά και σε ενδιάμεσα χαρακτηριστικά που πιθανόν θα βοηθούσαν στην αναγνώριση της κακοήθους κλάσης. Αυτό ενισχύει την ανάγκη να μη διαβάζουμε την concept-based προσέγγιση ως πλήρη λύση εξηγησιμότητας. Είναι ένα χρήσιμο πειραματικό βήμα, αλλά οι concept προβλέψεις χρειάζονται και αυτές αξιολόγηση, γιατί μπορεί να είναι σωστές σε ορισμένα χαρακτηριστικά και λανθασμένες σε άλλα.

Συνολικά, το πείραμα ταξινόμησης με ενδιάμεσα δερματοσκοπικά χαρακτηριστικά στο Derm7pt έδειξε ότι η προσθήκη βοηθητικών concept στόχων μπορεί να δώσει ένα πιο αναλυτικό πλαίσιο αξιολόγησης. Η τελική διάγνωση πέτυχε Test AUC = 88.66%, Accuracy = 82.40% και F1 = 67.65% για την κλάση melanoma, ενώ η πρόβλεψη των concepts είχε συνολικό Concept AUC = 74.73%. Τα καλύτερα concepts ήταν το blue-whitish veil και το pigment network, ενώ πιο δύσκολα αποδείχθηκαν τα regression structures και, σε επίπεδο threshold 0.5, τα vascular structures.

4.12 FAT-Net τμηματοποίησης στο ISIC 2016

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο τμηματοποίησης δερματικών βλαβών στο σύνολο δεδομένων ISIC 2016. Ο στόχος ήταν η πρόβλεψη της περιοχής της βλάβης σε επίπεδο pixel, δηλαδή η δημιουργία δυαδικής μάσκας που να διαχωρίζει τη δερματική βλάβη από το υπόλοιπο δέρμα. Σε αντίθεση με τα πειράματα ταξινόμησης, όπου το μοντέλο δίνει μία συνολική ετικέτα για την εικόνα, εδώ η αξιολόγηση βασίζεται στην επικάλυψη ανάμεσα στην προβλεπόμενη μάσκα και τη μάσκα αναφοράς. Το πείραμα βασίστηκε στο ISIC 2016, το οποίο περιλαμβάνει δερματοσκοπικές εικόνες και αντίστοιχες μάσκες τμηματοποίησης. Από το training μέρος του dataset χρησιμοποιήθηκαν 900 ζεύγη εικόνας-μάσκας. Επειδή δεν υπήρχε ξεχωριστό validation set στη συγκεκριμένη δομή των φακέλων, το training set χωρίστηκε εσωτερικά σε 765 εικόνες για εκπαίδευση και 135 εικόνες για επικύρωση. Το test set περιλάμβανε 379 εικόνες με τις αντίστοιχες μάσκες. Όλες οι εικόνες και οι μάσκες μετατράπηκαν σε μέγεθος 224×224 pixels.

Σύνολο δεδομένων	Πλήθος εικόνων
Train	765
Validation	135
Test	379
Σύνολο	1.279

Πίνακας 4.40. Κατανομή δεδομένων στο πείραμα FAT-Net-inspired τμηματοποίησης στο ISIC 2016

Ο βασικός CNN κλάδος ήταν ένας ResNet34 encoder με προεκπαιδευμένα βάρη ImageNet. Από τον encoder εξάγονταν χαρακτηριστικά σε τέσσερις διαφορετικές χωρικές κλίμακες: $1/4$, $1/8$, $1/16$ και $1/32$ της αρχικής εικόνας. Παράλληλα, χρησιμοποιήθηκε ένας Transformer κλάδος, ο οποίος ξεκινούσε από το βαθύτερο χαρακτηριστικό επίπεδο του ResNet34 και εφάρμοζε μηχανισμό multi-head attention. Με αυτόν τον τρόπο, το μοντέλο μπορούσε να συνδυάσει τοπικά χαρακτηριστικά από το CNN με πιο συνολική πληροφορία συμφραζομένων από τον attention κλάδο. Για τη σύνδεση των CNN και Transformer χαρακτηριστικών χρησιμοποιήθηκαν Feature Adaptation Modules. Τα modules αυτά ευθυγράμμιζαν τα χαρακτηριστικά των δύο κλάδων και τα συνδυάζαν μέσω μηχανισμού στάθμισης. Στη συνέχεια, ο decoder έκανε σταδιακή ανάλυση της εικόνας με upsampling blocks και skip connections, μέχρι να παραχθεί τελική δυαδική μάσκα με μία έξοδο ανά pixel. Η αρχιτεκτονική αυτή είναι πιο σύνθετη από ένα απλό encoder-decoder μοντέλο, επειδή δεν βασίζεται μόνο σε συνεκτικά χαρακτηριστικά, αλλά ενσωματώνει και global context μέσω του attention μηχανισμού.

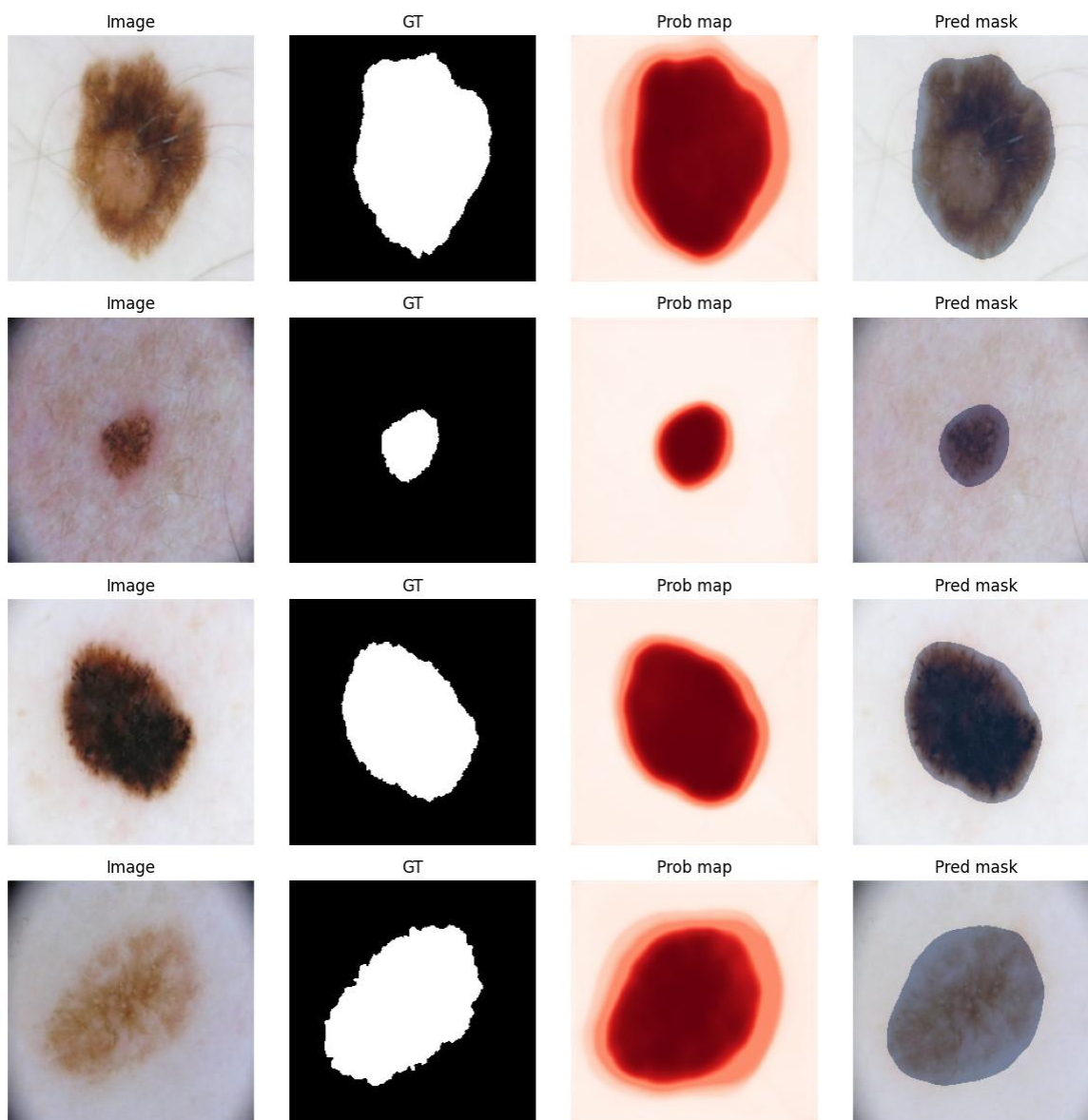
Η εκπαίδευση πραγματοποιήθηκε για μέγιστο αριθμό 20 εποχών, με batch size 8, learning rate $2e-4$ και weight decay $1e-4$. Ως συνάρτηση απώλειας χρησιμοποιήθηκε συνδυασμός BCEWithLogitsLoss και soft Dice loss, με βάρη 0.4 και 0.6 αντίστοιχα. Αυτή η επιλογή είναι κατάλληλη για segmentation, επειδή η BCE βοηθά στην pixel-wise ταξινόμηση, ενώ η Dice loss δίνει έμφαση στην επικάλυψη ανάμεσα στην προβλεπόμενη περιοχή και τη μάσκα αναφοράς. Το καλύτερο μοντέλο επιλέχθηκε με βάση το validation IoU. Η καλύτερη επίδοση εμφανίστηκε πολύ νωρίς, στη 2η εποχή, όπου το μοντέλο πέτυχε $val_loss = 0.3050$,

$val_iou = 80.64\%$ και $val_acc = 95.29\%$. Στην 1η εποχή το validation IoU ήταν ήδη υψηλό, 0.7875, ενώ στην 4η εποχή παρέμεινε κοντά, με $val_iou = 80\%$. Μετά την 4η εποχή η απόδοση άρχισε να μειώνεται, φτάνοντας σε $val_iou = 74\%$ στην 7η εποχή, όπου ενεργοποιήθηκε το early stopping. Αυτό δείχνει ότι το μοντέλο έμαθε γρήγορα τη βασική δομή του προβλήματος, αλλά στη συνέχεια δεν παρουσίασε σταθερή περαιτέρω βελτίωση στο validation set.

Epoch	Train loss	Train Dice*	Train IoU (%)	Train Acc (%)	Val loss	Val IoU (%)	Val Acc (%)
1	0.4387	1.2630	63.15%	88.43%	0.3516	78.75%	93.64%
2	0.3750	1.3322	66.61%	90.07%	0.3050	80.64%	95.29%
3	0.3290	1.3952	69.76%	91.30%	0.2866	75.86%	93.97%
4	0.3002	1.4160	70.80%	91.76%	0.2343	80.00%	94.86%
5	0.2785	1.4233	71.16%	91.91%	0.2578	78.32%	94.65%
6	0.2629	1.4246	71.23%	91.92%	0.2332	75.81%	93.64%
7	0.2366	1.4715	73.57%	92.71%	0.2211	74.00%	94.12%

Πίνακας 4.41. Πορεία εκπαίδευσης του FAT-Net-inspired μοντέλου στο ISIC 2016

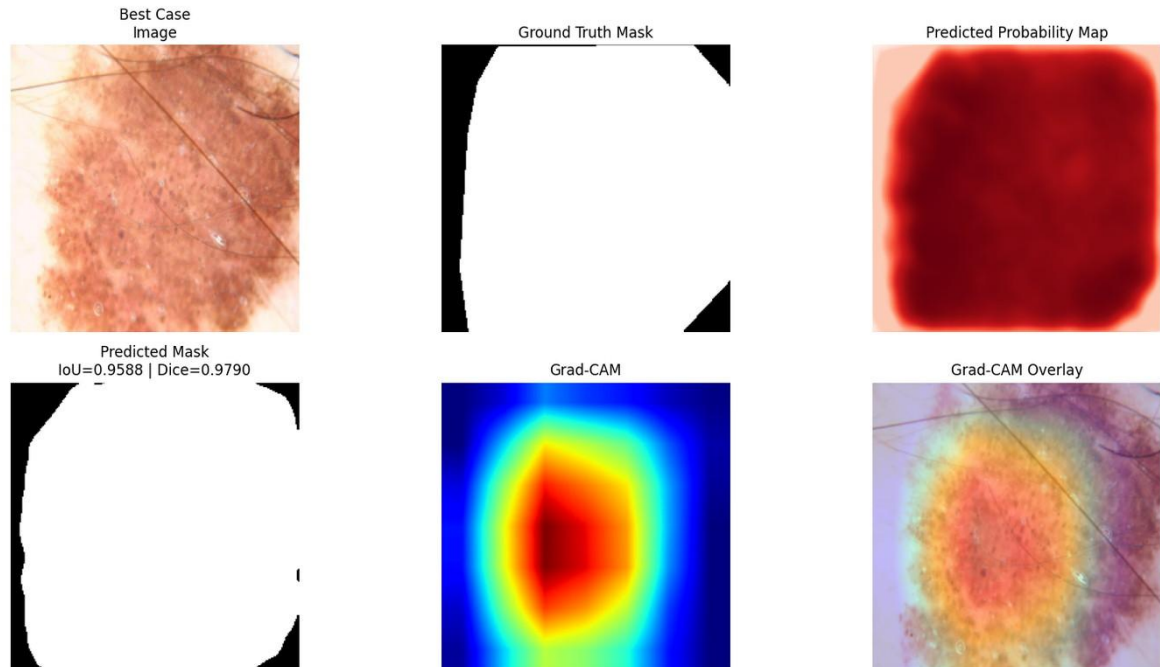
Στο test set, το καλύτερο checkpoint της 2ης εποχής πέτυχε Test Loss = 0.3022, Test Dice = 88.19%, Test IoU = 80.27% και Test Accuracy = 94.67%. Τα αποτελέσματα αυτά δείχνουν ότι το μοντέλο είχε αρκετά καλή ικανότητα τμηματοποίησης. Το IoU = 80.27% σημαίνει ότι η προβλεπόμενη μάσκα είχε υψηλή επικάλυψη με τη μάσκα αναφοράς, ενώ το Dice = 88.19% δείχνει ισχυρή συμφωνία ανάμεσα στις δύο περιοχές. Η υψηλή pixel accuracy πρέπει να διαβαστεί προσεκτικά, γιατί στις εικόνες segmentation το υπόβαθρο συνήθως καταλαμβάνει μεγάλο μέρος της εικόνας. Γι' αυτό οι πιο ουσιαστικές μετρικές εδώ είναι το Dice και το IoU. Στην ποιοτική αξιολόγηση παρουσιάστηκαν παραδείγματα από το test set, όπου συγκρίνονται η αρχική εικόνα, η μάσκα αναφοράς και η προβλεπόμενη μάσκα. Σε αρκετές περιπτώσεις, το μοντέλο κατάφερε να αποδώσει ικανοποιητικά το γενικό σχήμα της βλάβης και να ακολουθήσει τα βασικά όριά της. Αυτό είναι σημαντικό, επειδή στις δερματοσκοπικές εικόνες τα όρια της βλάβης δεν είναι πάντα καθαρά. Συχνά υπάρχουν περιοχές με σταδιακή μετάβαση χρώματος, τρίχες, φωτιστικές διαφορές ή ανομοιογενές υπόβαθρο, στοιχεία που δυσκολεύουν την ακριβή τμηματοποίηση.



Εικόνα 4.56. Ενδεικτικά αποτελέσματα τμηματοποίησης του FAT-Net-inspired μοντέλου στο test set

Για την ερμηνεία των προβλέψεων εφαρμόστηκε Grad-CAM πάνω στο segmentation μοντέλο. Επειδή το μοντέλο παράγει μάσκα και όχι μία απλή κλάση, χρησιμοποιήθηκε ειδικός wrapper που μετατρέπει την έξοδο segmentation σε scalar score, παίρνοντας τον μέσο όρο των logits στον χωρικό χάρτη. Ως target layer χρησιμοποιήθηκε ένα συνελκτικό επίπεδο στο κεντρικό τμήμα του μοντέλου, πριν από τον decoder. Η Grad-CAM ανάλυση, λοιπόν, δεν δείχνει μια δεύτερη μάσκα τμηματοποίησης, αλλά τις περιοχές που επηρέασαν περισσότερο την παραγωγή της μάσκας. Για πιο αναλυτική ποιοτική αξιολόγηση επιλέχθηκαν τρεις περιπτώσεις από το test set: η καλύτερη περίπτωση, μία ενδιάμεση περίπτωση και η δυσκολότερη περίπτωση, με βάση το ατομικό IoU. Η καλύτερη περίπτωση είχε δείκτη $idx = 211$ και εικόνα ISIC_0009980, με $IoU = 0.9588$ και $Dice = 0.9790$. Η τιμή αυτή δείχνει σχεδόν πλήρη συμφωνία ανάμεσα στην προβλεπόμενη μάσκα και τη μάσκα

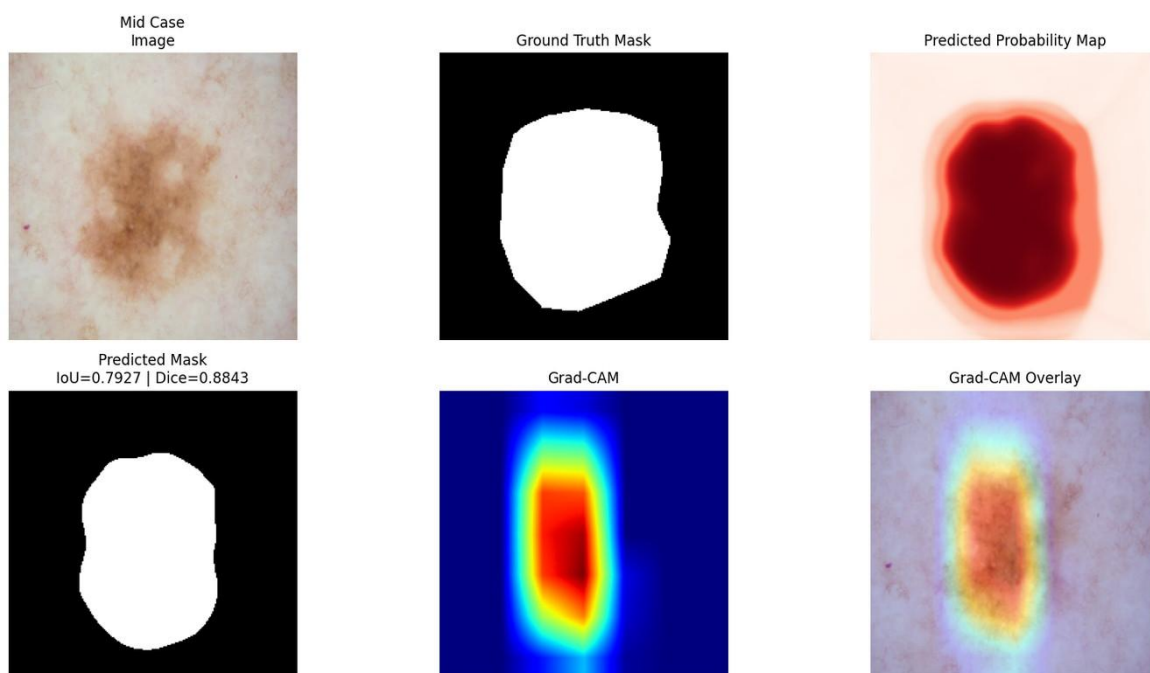
αναφοράς. Στις κορυφαίες περιπτώσεις του test set, τα IoU κυμάνθηκαν περίπου από 0.9496 έως 0.9588, κάτι που δείχνει ότι το μοντέλο μπορούσε να παράγει πολύ ακριβείς μάσκες σε ευνοϊκές εικόνες.



Εικόνα 4.57. Καλύτερη περίπτωση τμηματοποίησης στο test set, με IoU = 95.88% και Dice = 97.90%

Στην καλύτερη περίπτωση, η Grad-CAM ενεργοποίηση εμφανίζεται κυρίως γύρω από την περιοχή της βλάβης. Αυτό είναι θετικό, γιατί δείχνει ότι το μοντέλο δεν στηρίζεται κυρίως σε άσχετο φόντο. Η προβλεπόμενη μάσκα ακολουθεί πολύ καλά την περιοχή της βλάβης. Η διαφορά από τη μάσκα αναφοράς είναι μικρή και περιορίζεται κυρίως σε λεπτομέρειες των ορίων. Αυτό δείχνει ότι το μοντέλο μπορεί να αποδώσει με ακρίβεια τη βλάβη όταν υπάρχει σαφής αντίθεση ανάμεσα στη βλάβη και στο περιβάλλον δέρμα. Σε τέτοιες εικόνες, ο συνδυασμός CNN χαρακτηριστικών και attention-based global context φαίνεται να λειτουργεί θετικά.

Στην ενδιάμεση περίπτωση, η Grad-CAM μπορεί να δείχνει εστίαση σε σημαντικό μέρος της βλάβης, αλλά η σχέση με την τελική μάσκα είναι λιγότερο καθαρή. Αυτό είναι χρήσιμο, γιατί δείχνει ότι η ερμηνευτική εικόνα δεν είναι πάντα τόσο σταθερή όσο στις πολύ καλές περιπτώσεις. Το μοντέλο μπορεί να εντοπίζει τη σωστή περιοχή, αλλά να δυσκολεύεται στα ακριβή όρια ή στο μέγεθος της προβλεπόμενης μάσκας.



Εικόνα 4.58. Ενδιάμεση περίπτωση τμηματοποίησης στο test set

Συνολικά, το πείραμα δείχνει ότι η FAT-Net-inspired αρχιτεκτονική είχε καλή απόδοση στο ISIC 2016, με Test Dice = 88.19% και Test IoU = 80.27%. Η καλύτερη εποχή ήταν η 2η, γεγονός που δείχνει ότι το μοντέλο έμαθε γρήγορα χρήσιμες αναπαραστάσεις για την τμηματοποίηση. Παράλληλα, η πτώση του validation IoU μετά τις πρώτες εποχές δείχνει ότι η περαιτέρω εκπαίδευση δεν βελτίωσε απαραίτητα τη γενίκευση και πιθανόν οδήγησε σε πιο ασταθή συμπεριφορά στο validation set.

4.13 Πείραμα VGG-U-Net στο ISIC 2016 με Grad-CAM και SHAP

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο VGG-U-Net για την τμηματοποίηση δερματικών βλαβών στο σύνολο δεδομένων ISIC 2016. Σε αντίθεση με τα πειράματα ταξινόμησης, όπου το μοντέλο προβλέπει μία συνολική ετικέτα για κάθε εικόνα, εδώ ο στόχος ήταν η παραγωγή μιας δυαδικής μάσκας που διαχωρίζει τη βλάβη από το υπόλοιπο δέρμα. Άρα η αξιολόγηση δεν βασίστηκε κυρίως στην απλή ακρίβεια, αλλά σε μετρικές επικάλυψης, όπως το Dice και το IoU.

Το dataset περιλάμβανε 900 ζεύγη εικόνας-μάσκας από το training μέρος του ISIC 2016. Κάθε ζεύγος αποτελούνταν από μία δερματοσκοπική εικόνα και την αντίστοιχη ground truth μάσκα. Το σύνολο χωρίστηκε εσωτερικά σε training και validation set, με αναλογία περίπου 80% / 20%, δηλαδή περίπου 720 εικόνες για εκπαίδευση και 180 εικόνες για επικύρωση. Οι εικόνες και οι μάσκες μετατράπηκαν σε μέγεθος 256×256 pixels. Οι εικόνες κανονικοποιήθηκαν με τις τιμές του ImageNet, ενώ οι μάσκες μετατράπηκαν σε δυαδική μορφή.

Περιγραφή	Τιμή
Paired images	900
Train	≈ 720
Validation	≈ 180
Test	—
Image size	256×256

Πίνακας 4.42. Κατανομή δεδομένων στο πείραμα VGG-U-Net στο ISIC 2016

Η αρχιτεκτονική βασίστηκε σε έναν VGG16-BN encoder με προεκπαιδευμένα βάρη από το ImageNet. Από τον encoder αξιοποιήθηκαν τα διαδοχικά blocks του VGG, ώστε να εξαχθούν χαρακτηριστικά σε διαφορετικά επίπεδα ανάλυσης. Στη συνέχεια, χρησιμοποιήθηκε decoder τύπου U-Net, με upsampling και skip connections από τα αντίστοιχα επίπεδα του encoder. Η λογική αυτής της αρχιτεκτονικής είναι ότι τα βαθύτερα επίπεδα μαθαίνουν πιο γενικά χαρακτηριστικά της βλάβης, ενώ τα πρώτα επίπεδα διατηρούν χωρικές λεπτομέρειες που είναι χρήσιμες για την ακριβή οριοθέτηση.

Η εκπαίδευση πραγματοποιήθηκε για 10 εποχές, με optimizer Adam και learning rate $1e-4$. Η απόδοση μετρήθηκε με Dice και IoU, ενώ το καλύτερο μοντέλο επιλέχθηκε με βάση το validation Dice. Ήδη από την 1η εποχή το μοντέλο εμφάνισε υψηλή επίδοση, με val_dice = 0.8906 και val_iou = 0.8103. Αυτό δείχνει ότι ο προεκπαιδευμένος VGG encoder βοήθησε το μοντέλο να αποκτήσει γρήγορα χρήσιμες αναπαραστάσεις για τη βλάβη.

Η καλύτερη επίδοση καταγράφηκε στην 8η εποχή, όπου το μοντέλο πέτυχε val_dice = 90.96% και val_iou = 83.95%. Στη 10η εποχή οι τιμές παρέμειναν υψηλές, με val_dice = 90.52% και val_iou = 83.41%, αλλά δεν ξεπέρασαν την 8η εποχή. Η πορεία αυτή δείχνει ότι το μοντέλο γενίκευσε καλά, αλλά μετά από ένα σημείο δεν υπήρξε ουσιαστική περαιτέρω βελτίωση στο validation set.

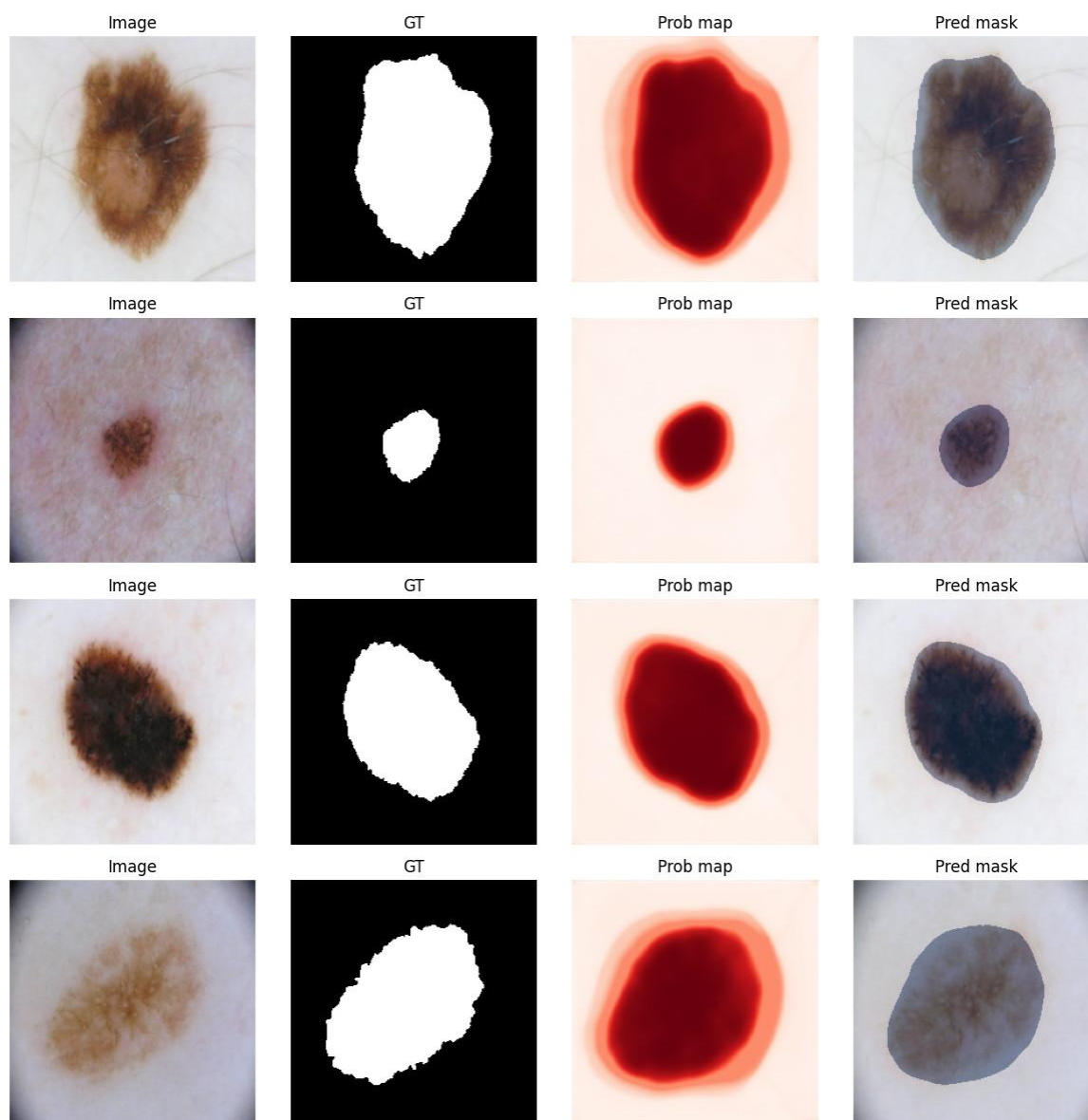
Epoch	Train loss	Train Dice (%)	Train IoU (%)	Val loss	Val Dice (%)	Val IoU (%)
1	0.6436	81.56%	71.83%	0.4727	89.06%	81.03%
3	0.3979	88.49%	80.73%	0.3746	90.10%	82.84%
4	0.3428	90.09%	83.10%	0.3420	90.14%	83.00%
5	0.3000	91.00%	84.29%	0.2896	90.51%	83.30%
8	0.2470	91.90%	85.64%	0.3418	90.96%	83.95%
10	0.1994	92.91%	87.28%	0.2685	90.52%	83.41%

Πίνακας 4.43. Πορεία εκπαίδευσης του VGG-U-Net στο ISIC 2016

Η βελτίωση του training Dice από 81.56% σε 92.91% δείχνει ότι το μοντέλο συνέχισε να προσαρμόζεται στα δεδομένα εκπαίδευσης. Το validation Dice, όμως, κινήθηκε από νωρίς

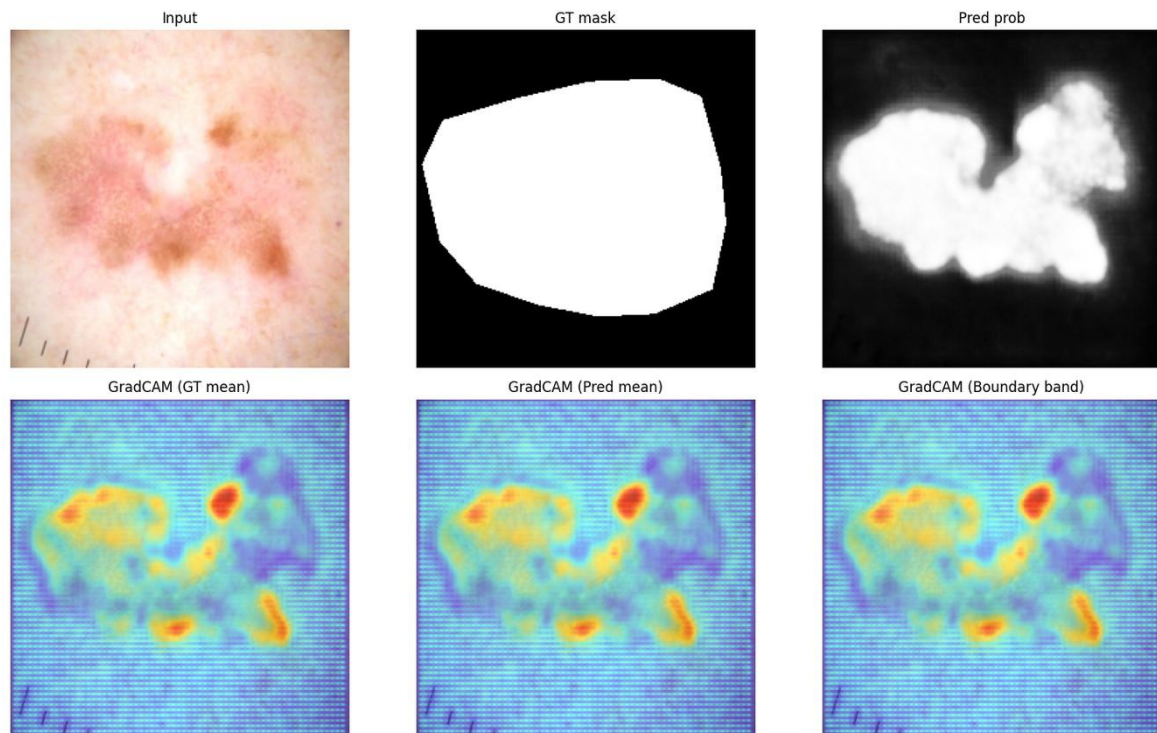
σε υψηλά επίπεδα και σταθεροποιήθηκε γύρω από το 0.90. Αυτό δείχνει ότι το μοντέλο πέτυχε γρήγορα καλή γενίκευση, αλλά οι μεταγενέστερες εποχές βελτίωσαν περισσότερο το training set παρά το validation set.

Στην ποιοτική αξιολόγηση παρουσιάστηκαν ενδεικτικά παραδείγματα με την αρχική εικόνα, τη μάσκα αναφοράς και την προβλεπόμενη μάσκα. Σε αρκετές περιπτώσεις, το VGG-U-Net κατάφερε να ακολουθήσει με ικανοποιητικό τρόπο το γενικό σχήμα της βλάβης. Οι βασικές αποκλίσεις εμφανίζονται κυρίως στα όρια, κάτι αναμενόμενο στις δερματοσκοπικές εικόνες, όπου η μετάβαση από τη βλάβη στο φυσιολογικό δέρμα δεν είναι πάντα καθαρή.



Εικόνα 4.59. Ενδεικτικά αποτελέσματα τμηματοποίησης του VGG-U-Net στο validation set

Για την ερμηνεία της πρόβλεψης εφαρμόστηκε Grad-CAM πάνω στο segmentation μοντέλο. Επειδή το VGG-U-Net δεν παράγει μία κλάση αλλά έναν χωρικό χάρτη μάσκας, η έξοδος μετατράπηκε σε scalar score, ώστε να μπορεί να εφαρμοστεί η Grad-CAM. Πρακτικά, η μέθοδος χρησιμοποιήθηκε για να δείξει ποιες περιοχές της εικόνας επηρέασαν περισσότερο την παραγωγή της μάσκας. Αυτό είναι σημαντικό, γιατί σε ένα segmentation μοντέλο δεν αρκεί να δούμε μόνο αν η μάσκα είναι σωστή, χρειάζεται να εξετάσουμε και αν το μοντέλο εστιάζει σε σχετικές περιοχές της εικόνας.

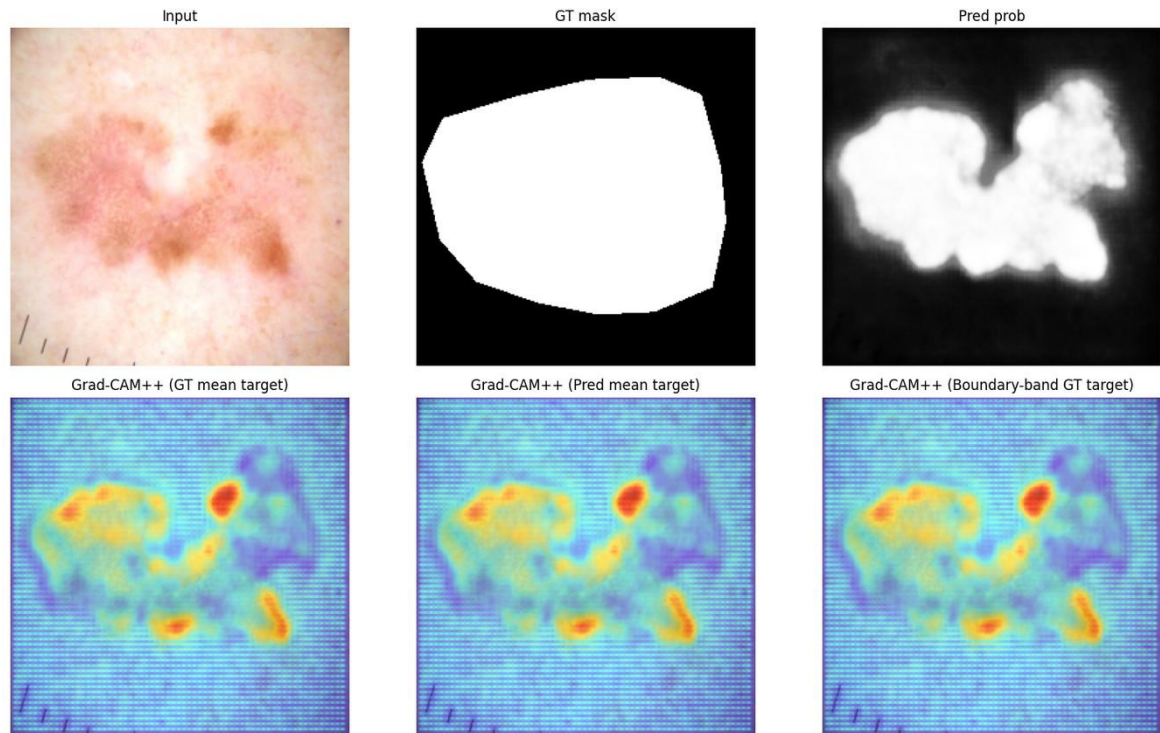


Εικόνα 4.60. Grad-CAM ανάλυση του VGG-U-Net για παράδειγμα τμηματοποίησης

Η Grad-CAM εικόνα δείχνει ότι οι περιοχές ενεργοποίησης βρίσκονται κυρίως γύρω από τη βλάβη. Αυτό είναι θετικό, γιατί σημαίνει ότι το μοντέλο δεν στηρίζεται κυρίως σε άσχετο φόντο. Παρ' όλα αυτά, η Grad-CAM δεν πρέπει να ερμηνεύεται ως δεύτερη μάσκα τμηματοποίησης. Δεν δείχνει τα ακριβή όρια της βλάβης, αλλά τις περιοχές που συνέβαλαν στην πρόβλεψη. Στο συγκεκριμένο πείραμα, οι CAM χάρτες που υπολογίστηκαν ως προς τη ground truth μάσκα, την predicted mask και την edge περιοχή είχαν σχεδόν ταυτόσημη μορφή. Οι συσχετίσεις ήταν πολύ υψηλές: $\text{corr}(\text{gt}, \text{pred}) = 0.99999$, $\text{corr}(\text{gt}, \text{edge}) = 0.99997$ και $\text{corr}(\text{pred}, \text{edge}) = 0.99997$. Αυτό δείχνει ότι οι τρεις απλές CAM εκδοχές έδιναν σχεδόν την ίδια πληροφορία.

Για να εξεταστούν πιο ουσιαστικά τα όρια της βλάβης, δημιουργήθηκε και ένα boundary contrast CAM. Σε αυτή την περίπτωση, η ανάλυση δεν περιορίστηκε στη γενική περιοχή της μάσκας, αλλά συνέκρινε την περιοχή των ορίων με το εσωτερικό της βλάβης. Το αποτέλεσμα είχε πολύ χαμηλότερη συσχέτιση με το αρχικό CAM, με $\text{corr}(\text{gt}, \text{edge_contrast}) = 0.5278$ και $\text{mean} |\text{gt} - \text{edge_contrast}| = 0.1185$. Αυτό σημαίνει ότι το

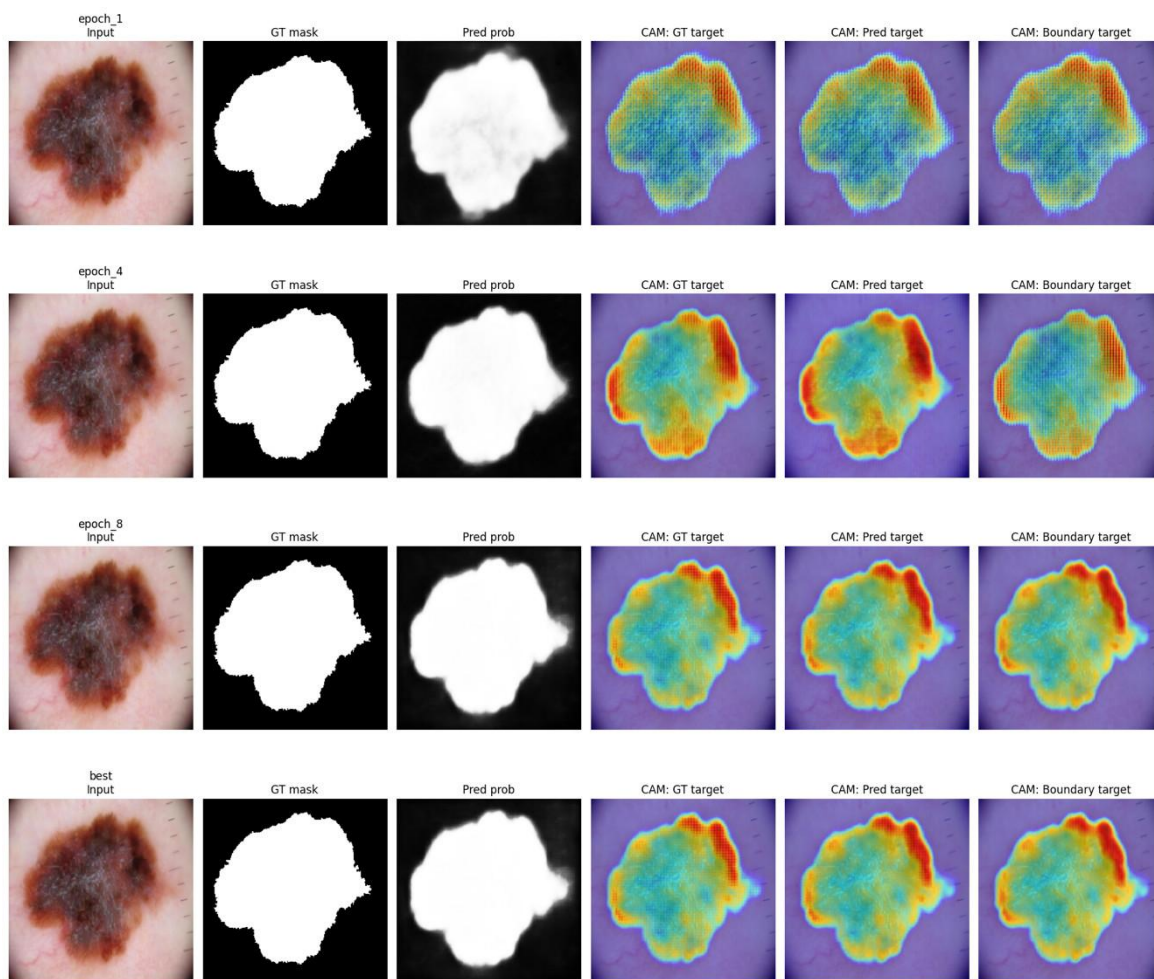
boundary contrast CAM ανέδειξε διαφορετική πληροφορία και είναι πιο χρήσιμο για να δούμε αν το μοντέλο επηρεάζεται από τα όρια και όχι μόνο από το εσωτερικό της βλάβης.



Εικόνα 4.61. Boundary contrast CAM για την ανάλυση των ορίων της βλάβης

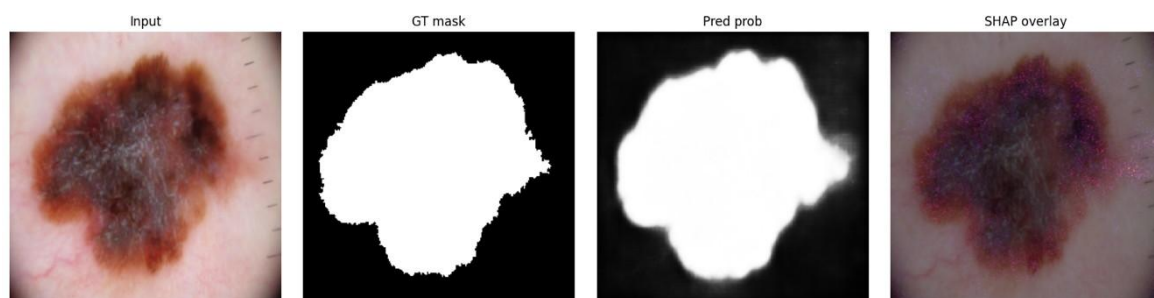
Η εικόνα του boundary contrast CAM είναι από τις πιο χρήσιμες αυτού του πειράματος. Σε προβλήματα segmentation, το μοντέλο μπορεί να βρίσκει σωστά το κέντρο της βλάβης, αλλά να κάνει λάθη στα περιγράμματα. Αυτό έχει σημασία, γιατί η ποιότητα της τμηματοποίησης δεν εξαρτάται μόνο από το αν εντοπίζεται η βλάβη, αλλά και από το πόσο σωστά οριοθετείται. Εδώ το boundary CAM προσθέτει κάτι πιο ειδικό από μια απλή Grad-CAM εικόνα.

Επιπλέον, πραγματοποιήθηκε across-epoch Grad-CAM++ ανάλυση με checkpoints από τις εποχές 1, 4, 8 και από το καλύτερο μοντέλο. Η ανάλυση αυτή είναι χρήσιμη, επειδή δείχνει πώς μεταβάλλεται η οπτική εστίαση του μοντέλου κατά τη διάρκεια της εκπαίδευσης. Στις αρχικές εποχές η ενεργοποίηση είναι συνήθως πιο αδρή ή λιγότερο σταθερή, ενώ στις μεταγενέστερες εποχές και ιδιαίτερα στο καλύτερο checkpoint η εστίαση εμφανίζεται πιο κοντά στην περιοχή της βλάβης. Με τον τρόπο αυτό, η ανάλυση δεν περιορίζεται μόνο στην τελική μάσκα, αλλά επιτρέπει να παρακολουθηθεί και η εξέλιξη της ερμηνείας του μοντέλου μέσα στον χρόνο εκπαίδευσης.



Εικόνα 4.62. Grad-CAM++ across-epoch ανάλυση του VGG-U-Net στο ISIC 2016

Τέλος, εφαρμόστηκε και SHAP στο VGG-U-Net. Επειδή το μοντέλο παράγει μάσκα, χρησιμοποιήθηκε wrapper που μετατρέπει την έξοδο σε scalar score μέσω της μέσης ενεργοποίησης των logits..



Εικόνα 4.63. SHAP ανάλυση για πρόβλεψη τμηματοποίησης του VGG-U-Net

Συνολικά, το πείραμα με VGG-U-Net στο ISIC 2016 έδειξε πολύ καλή επίδοση τμηματοποίησης, με καλύτερο validation Dice = 90.96% και validation IoU = 83.95% στην 8η εποχή. Η χρήση προεκπαιδευμένου VGG16-BN encoder φαίνεται να βοήθησε το

μοντέλο να αποκτήσει γρήγορα ισχυρές αναπαραστάσεις, αφού η επίδοση ήταν ήδη υψηλή από την πρώτη εποχή. Σε σχέση με ένα απλό U-Net, το VGG-U-Net έχει το πλεονέκτημα της μεταφοράς μάθησης, αλλά είναι βαρύτερο υπολογιστικά. Το πιο ουσιαστικό κομμάτι του πειράματος είναι ότι η αξιολόγηση δεν έμεινε μόνο στις μετρικές Dice και IoU. Οι Grad-CAM εικόνες έδειξαν ότι το μοντέλο εστιάζει κυρίως στην περιοχή της βλάβης, ενώ το boundary contrast CAM πρόσθεσε πιο ειδική πληροφορία για τα όρια. Η SHAP ανάλυση, παρότι περιορισμένη σε λίγα δείγματα, έδωσε μια δεύτερη, πιο τοπική ερμηνεία. Με αυτόν τον τρόπο, το πείραμα παρουσιάζει όχι μόνο το αν η τμηματοποίηση είναι αριθμητικά καλή, αλλά και το πώς το μοντέλο φαίνεται να αξιοποιεί την εικόνα για να παράγει τη μάσκα.

4.14 Πείραμα PAD-UFES-20 με τρία backbones και εξηγησιμότητα Grad-CAM++ / SHAP

Στο πείραμα αυτό αξιολογήθηκε ένα σχήμα δυαδικής ταξινόμησης στο σύνολο δεδομένων PAD-UFES-20. Το αρχικό dataset περιλάμβανε διαφορετικές διαγνωστικές κατηγορίες, οι οποίες ομαδοποιήθηκαν σε δύο ευρύτερες κλάσεις: benign και suspicious. Στην κλάση benign εντάχθηκαν οι κατηγορίες NEV και SEK, ενώ στην κλάση suspicious εντάχθηκαν οι ACK, BCC, MEL και SCC. Με αυτόν τον τρόπο, το πρόβλημα διαμορφώθηκε ως ένα σενάριο αρχικού ελέγχου, όπου το μοντέλο καλείται να ξεχωρίσει τις πιθανώς ύποπτες βλάβες από τις λιγότερο ανησυχητικές.

Το τελικό dataframe περιλάμβανε 2.298 εικόνες. Από αυτές, οι 1.819 ανήκαν στην κλάση suspicious και οι 479 στην κλάση benign. Η κατανομή των αρχικών κατηγοριών ήταν: BCC = 845, ACK = 730, NEV = 244, SEK = 235, SCC = 192 και MEL = 52. Η ανισορροπία είναι εμφανής, καθώς η suspicious κλάση είναι πολύ μεγαλύτερη από την benign. Αυτό επηρεάζει ιδιαίτερα την ερμηνεία της accuracy, επειδή ένα μοντέλο μπορεί να έχει υψηλή συνολική ακρίβεια κυρίως λόγω της καλής επίδοσης στη μεγαλύτερη κλάση.

Κατηγορία	Πλήθος
Suspicious	1819
Benign	479

Αρχική διάγνωση	Πλήθος
BCC	845
ACK	730
NEV	244
SEK	235
SCC	192
MEL	52

Πίνακας 4.44. Κατανομή των κατηγοριών στο PAD-UFES-20

Το dataset χωρίστηκε σε 1.608 εικόνες εκπαίδευσης, 345 εικόνες επικύρωσης και 345 εικόνες ελέγχου. Στο training set υπήρχαν 1.273 suspicious και 335 benign εικόνες. Στο validation set υπήρχαν 273 suspicious και 72 benign, ενώ η ίδια κατανομή υπήρχε και στο test set. Άρα, σε όλα τα splits διατηρήθηκε η βασική ανισορροπία του dataset.

Split	Σύνολο	Benign	Suspicious
Train	1608	335	1273
Validation	345	72	273
Test	345	72	273

Πίνακας 4.45. Κατανομή train, validation και test στο πείραμα PAD-UFES-20

Για τη σύγκριση χρησιμοποιήθηκαν τρία διαφορετικά backbones: ResNet50, EfficientNet-B0 και ConvNeXt-Tiny. Και τα τρία μοντέλα εκπαιδεύτηκαν για 8 εποχές και αξιολογήθηκαν με το ίδιο πρωτόκολλο. Σε κάθε εποχή έγινε επιλογή threshold στο validation set, ώστε να βελτιστοποιηθεί η δυαδική απόφαση. Αυτό ήταν απαραίτητο, επειδή το dataset είναι ανισόρροπο και το σταθερό threshold 0.50 δεν είναι πάντα το πιο κατάλληλο.

Το ResNet50 είχε την καλύτερη συνολική επίδοση στο test set. Το καλύτερο checkpoint ήταν στην 4η εποχή, με threshold 0.29. Στο test set πέτυχε accuracy = 93.62%, precision = 97.72%, recall = 94.14%, F1-score = 95.90% και AUC = 96.33% για την κλάση suspicious. Ο πίνακας σύγκρισης ήταν:

Κατηγορία	Τιμή
True Negatives	66
False Positives	6
False Negatives	16
True Positives	257

Αυτό σημαίνει ότι από τις 72 benign εικόνες, οι 66 ταξινομήθηκαν σωστά και 6 ταξινομήθηκαν λανθασμένα ως suspicious. Από τις 273 suspicious εικόνες, οι 257 εντοπίστηκαν σωστά, ενώ 16 ταξινομήθηκαν λανθασμένα ως benign. Η τελευταία κατηγορία λαθών είναι πιο κρίσιμη, γιατί αντιστοιχεί σε ύποπτες βλάβες που το μοντέλο δεν αναγνώρισε.

Το EfficientNet-B0 είχε καλύτερο checkpoint στην 8η εποχή, με threshold 0.06. Στο test set πέτυχε accuracy = 91.88%, precision = 96.23%, recall = 93.41%, F1-score = 94.80% και AUC = 94.15%. Ο πίνακας σύγκρισης ήταν:

Κατηγορία	Τιμή
True Negatives / Class 0 σωστά	62
False Positives	10

Κατηγορία	Τιμή
False Negatives	18
True Positives / Class 1 σωστά	255

Το μοντέλο είχε καλή συνολική συμπεριφορά, αλλά σε σχέση με το ResNet50 είχε περισσότερα false positives και false negatives. Συγκεκριμένα, ταξινομήσε 10 benign εικόνες ως suspicious και έχασε 18 suspicious εικόνες.

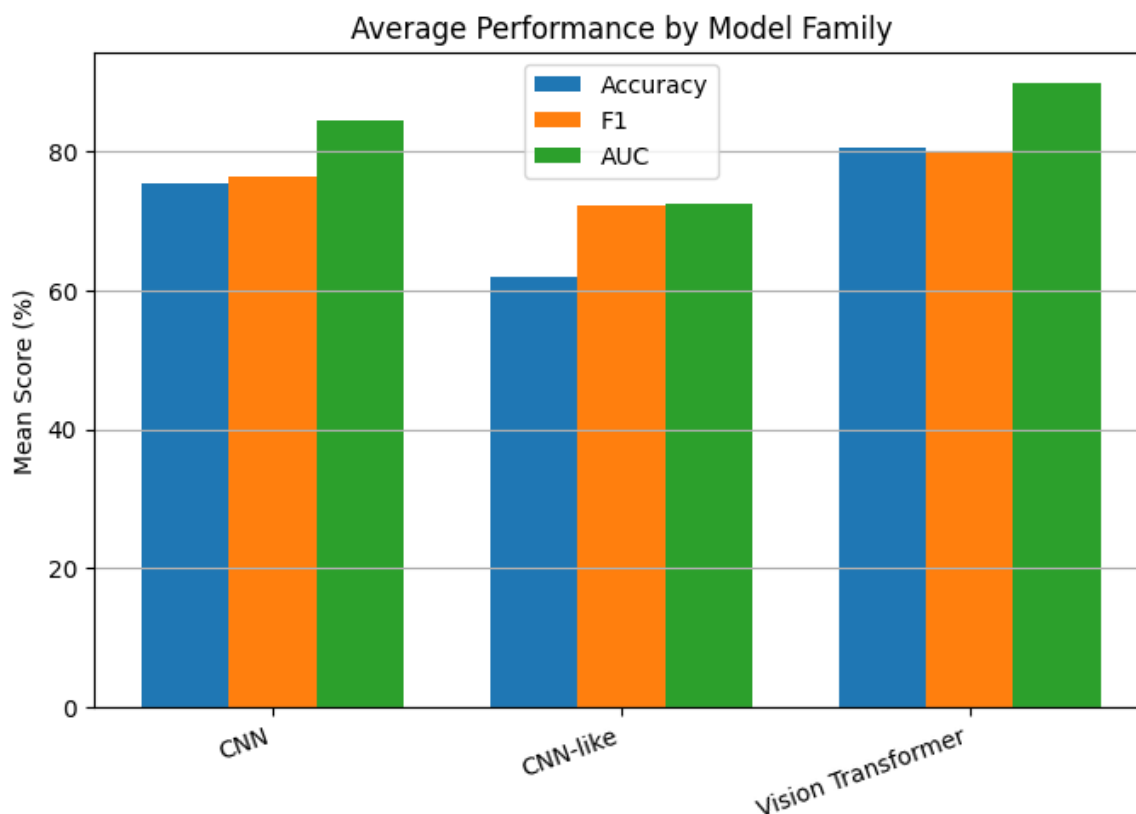
Το ConvNeXt-Tiny είχε καλύτερο checkpoint στην 7η εποχή, με threshold 0.07. Στο test set πέτυχε accuracy = 91.88%, precision = 94.22%, recall = 95.60%, F1-score = 94.91% και AUC = 95.88%. Ο πίνακας σύγκρισης ήταν:

Κατηγορία	Τιμή
True Negatives / Class 0 σωστά	56
False Positives	16
False Negatives	12
True Positives / Class 1 σωστά	261

Το ConvNeXt-Tiny είχε το υψηλότερο recall για την suspicious κλάση, καθώς εντόπισε 261 από τις 273 suspicious εικόνες. Αυτό σημαίνει ότι είχε λιγότερα false negatives από τα άλλα δύο μοντέλα. Από την άλλη πλευρά, είχε περισσότερα false positives, αφού 16 benign εικόνες ταξινομήθηκαν ως suspicious. Άρα, το ConvNeXt-Tiny ήταν πιο ευαίσθητο, αλλά λιγότερο αυστηρό.

Model	Best epoch	Threshold	Accuracy	Precision	Recall	F1	AUC
ResNet50	4	0.29	93.62%	97.72%	94.14%	95.90%	96.33%
EfficientNet-B0	8	0.06	91.88%	96.23%	93.41%	94.80%	94.15%
ConvNeXt-Tiny	7	0.07	91.88%	94.22%	95.60%	94.91%	95.88%

Πίνακας 4.46. Σύγκριση των τριών backbones στο PAD-UFES-20

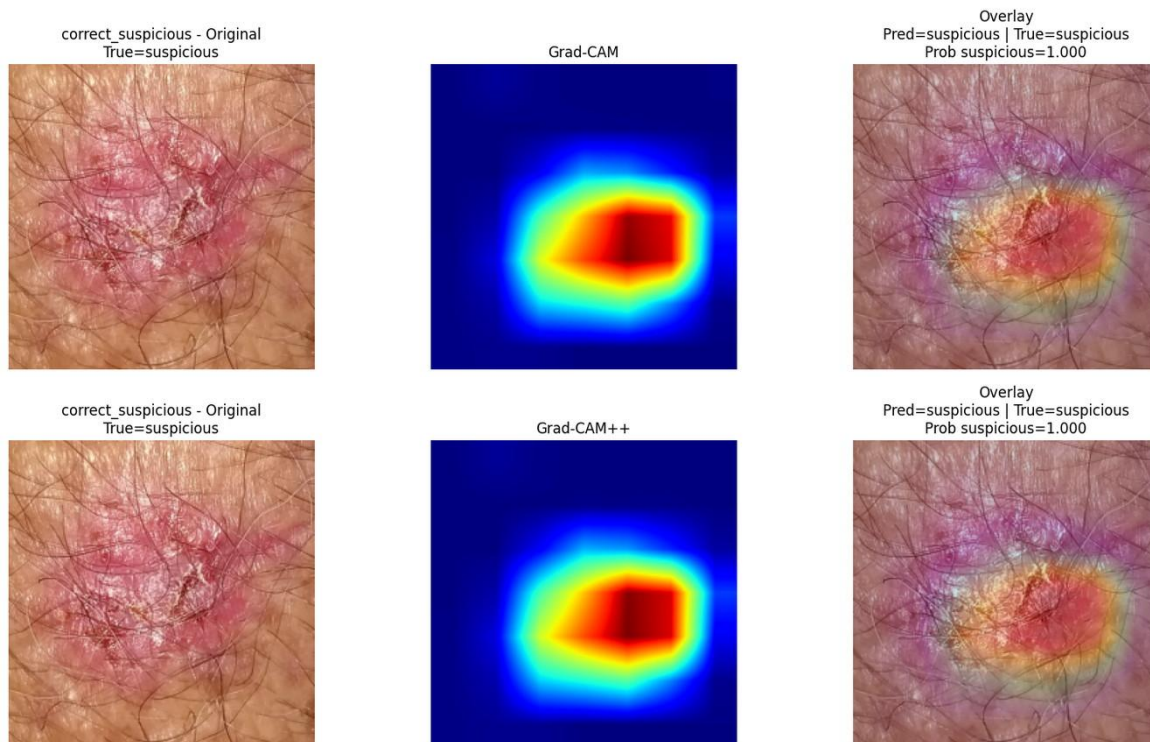


Εικόνα 4.64. Σύγκριση των backbones ως προς Accuracy, Precision, Recall, F1 και AUC

Η σύγκριση των τριών μοντέλων δείχνει ότι το ResNet50 ήταν συνολικά το πιο ισορροπημένο. Είχε την υψηλότερη accuracy, precision, F1-score και AUC. Το ConvNeXt-Tiny είχε ελαφρώς χαμηλότερη συνολική απόδοση, αλλά πέτυχε το καλύτερο recall για την suspicious κλάση, κάτι που έχει πρακτικό ενδιαφέρον σε ένα σενάριο αρχικού ελέγχου. Το EfficientNet-B0 κινήθηκε κοντά στα άλλα δύο μοντέλα, αλλά δεν ξεπέρασε το ResNet50 σε κάποια βασική μετρική. Ο πίνακας σύγκρισης του ResNet50 δείχνει καθαρά τη συμπεριφορά του καλύτερου μοντέλου. Η μεγαλύτερη μάζα των προβλέψεων βρίσκεται στη σωστή suspicious κλάση, με 257 σωστές προβλέψεις. Παρ' όλα αυτά, υπάρχουν 16 false negatives, δηλαδή suspicious εικόνες που ταξινομήθηκαν ως benign. Αυτά τα λάθη είναι σημαντικά, γιατί σε ένα ιατρικό πλαίσιο η αποτυχία εντοπισμού μιας ύποπτης βλάβης έχει μεγαλύτερη βαρύτητα από ένα false positive.

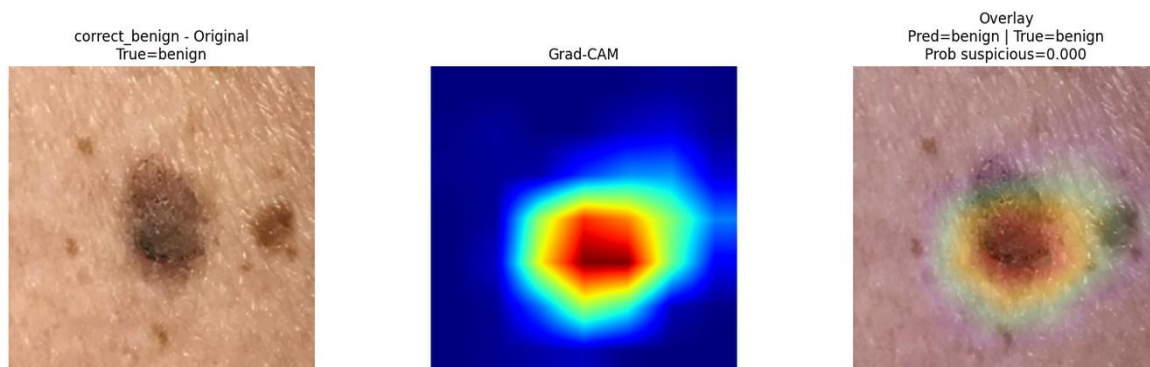
Για την ποιοτική αξιολόγηση επιλέχθηκαν τρεις χαρακτηριστικές περιπτώσεις: μία σωστή πρόβλεψη suspicious, μία σωστή πρόβλεψη benign και μία λανθασμένη πρόβλεψη. Στην ορθή suspicious περίπτωση, το μοντέλο προέβλεψε σωστά την εικόνα ως suspicious με πιθανότητα 1.000 στα CAM outputs και 0.848 στη SHAP ανάλυση. Στην ορθή benign περίπτωση, η πιθανότητα suspicious ήταν σχεδόν μηδενική στα CAM outputs και 0.154 στη SHAP ανάλυση. Στη λανθασμένη περίπτωση, η πραγματική κλάση ήταν suspicious, αλλά το μοντέλο προέβλεψε benign, με πιθανότητα suspicious μόλις 0.005. Αυτή η περίπτωση είναι ιδιαίτερα χρήσιμη για την ανάλυση, γιατί δείχνει πότε το μοντέλο αποτυγχάνει σε πιο κρίσιμο λάθος.

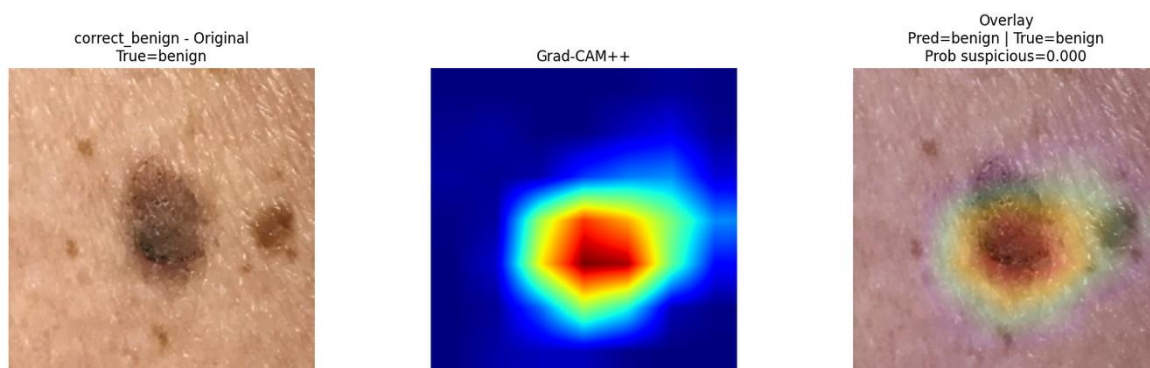
Για την εξηγησιμότητα χρησιμοποιήθηκαν Grad-CAM και Grad-CAM++. Οι δύο μέθοδοι εφαρμόστηκαν στο καλύτερο μοντέλο, δηλαδή στο ResNet50, και χρησιμοποιήθηκαν για να εξεταστεί αν η απόφαση βασίστηκε στην περιοχή της βλάβης ή σε στοιχεία του φόντου.



Εικόνα 4.65. Grad-CAM και Grad-CAM++ ανάλυση για ορθή πρόβλεψη suspicious στο PAD-UFES-20

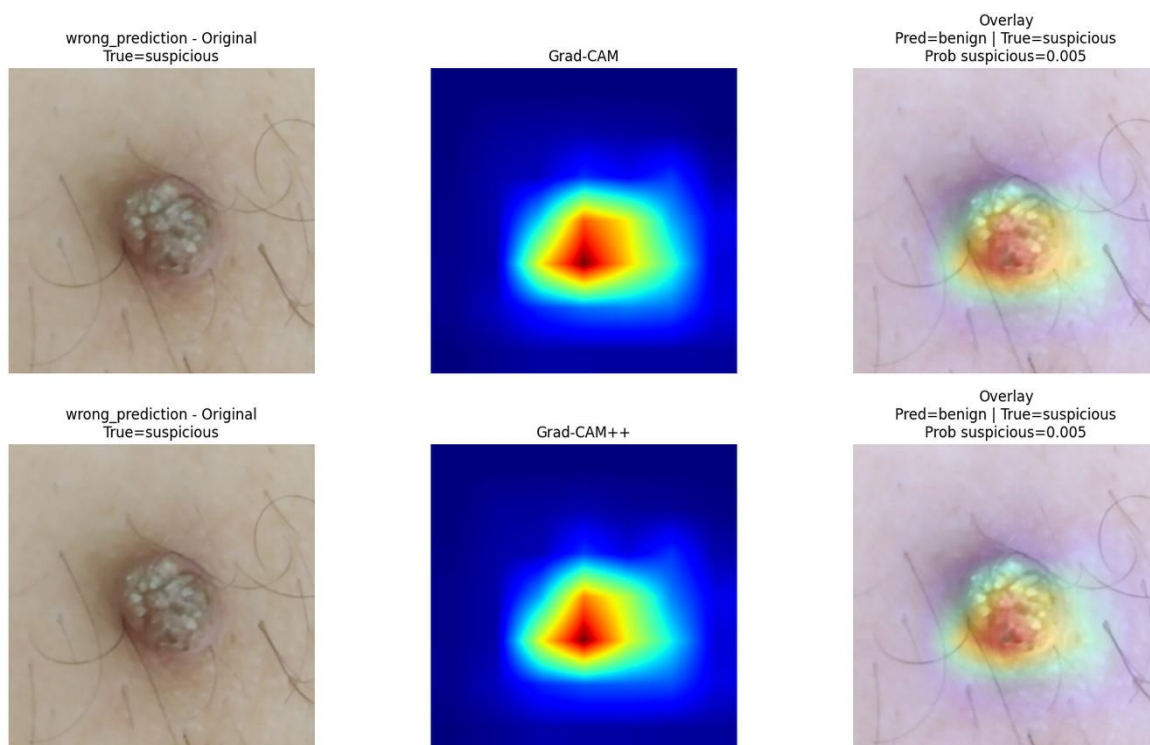
Στην ορθή suspicious περίπτωση, η Grad-CAM++ εστίαση βρίσκεται πάνω στην κεντρική περιοχή της βλάβης. Η πρόβλεψη είναι σωστή και η πιθανότητα suspicious είναι 1.000, άρα το μοντέλο εμφανίζεται πολύ βέβαιο. Η οπτική ερμηνεία είναι σχετικά πειστική, γιατί η ενεργοποίηση δεν φαίνεται να προέρχεται κυρίως από άσχετο φόντο, αλλά από την περιοχή της αλλοίωσης.





Εικόνα 4.66. Grad-CAM και Grad-CAM++ ανάλυση για ορθή πρόβλεψη benign στο PAD-UFES-20

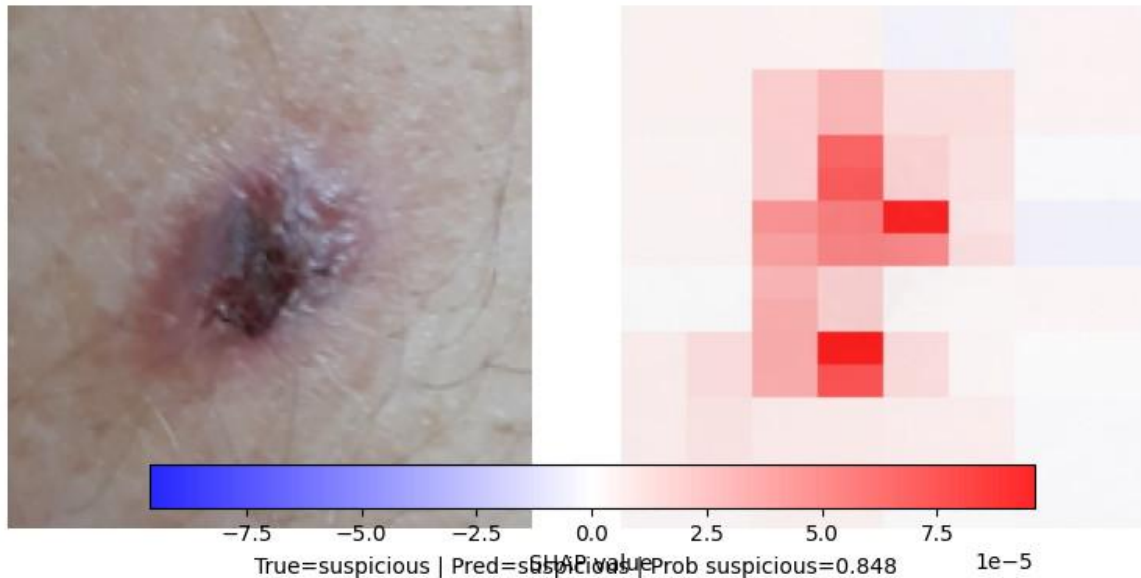
Στην ορθή benign περίπτωση, το μοντέλο απορρίπτει σωστά την suspicious κλάση. Η Grad-CAM++ εστιάζει στη βλάβη, αλλά η τελική πιθανότητα suspicious είναι 0.000, δηλαδή το μοντέλο δεν θεωρεί τα χαρακτηριστικά της εικόνας ύποπτα. Αυτό το παράδειγμα είναι χρήσιμο γιατί δείχνει ότι η εστίαση πάνω στη βλάβη δεν σημαίνει αυτόματα πρόβλεψη suspicious. Το μοντέλο μπορεί να κοιτάζει τη σωστή περιοχή, αλλά να τη χαρακτηρίζει ως benign.



Εικόνα 4.67. Grad-CAM και Grad-CAM++ ανάλυση για λανθασμένη πρόβλεψη suspicious ως benign

Στη λανθασμένη περίπτωση, η πραγματική κλάση ήταν suspicious, αλλά το μοντέλο προέβλεψε benign με πολύ χαμηλή πιθανότητα suspicious, 0.005. Η Grad-CAM++ εστιάζει

στην περιοχή της βλάβης, άρα το λάθος δεν φαίνεται να οφείλεται απλώς στο ότι το μοντέλο κοιτάζει εντελώς άσχετη περιοχή. Πιο πιθανό είναι ότι τα οπτικά χαρακτηριστικά της συγκεκριμένης βλάβης δεν ερμηνεύτηκαν ως αρκετά ύποπτα. Αυτό είναι σημαντικό, γιατί δείχνει τη διαφορά ανάμεσα στη χωρική εστίαση και στη σωστή ταξινόμηση.



Εικόνα 4.68. SHAP ανάλυση για ορθή πρόβλεψη suspicious στο PAD-UFES-20

Στη SHAP εικόνα της ορθής suspicious περίπτωσης, οι θετικές συνεισφορές εμφανίζονται κυρίως πάνω στην περιοχή της βλάβης. Αυτό ενισχύει την ερμηνεία που δίνει και η Grad-CAM++: η σωστή πρόβλεψη δεν βασίστηκε τυχαία στο φόντο, αλλά σε οπτικά στοιχεία της αλλοίωσης. Η πιθανότητα suspicious ήταν 0.848, άρα πρόκειται για μια αρκετά βέβαιη και σωστή πρόβλεψη.

Συνολικά, το πείραμα στο PAD-UFES-20 έδειξε ότι τα τρία backbones πέτυχαν υψηλές επιδόσεις στη δυαδική ταξινόμηση benign / suspicious. Το ResNet50 ήταν το καλύτερο συνολικά, με Accuracy = 93.62%, F1 = 95.90% και AUC = 96.33%. Το ConvNeXt-Tiny είχε το υψηλότερο recall, 95.60%, άρα εντόπισε περισσότερες suspicious εικόνες, αλλά με περισσότερα false positives. Το EfficientNet-B0 είχε επίσης καλή απόδοση, αλλά βρέθηκε ελαφρώς χαμηλότερα από τα άλλα δύο στις βασικές μετρικές. Το βασικό συμπέρασμα είναι ότι η ταξινόμηση suspicious / benign στο PAD-UFES-20 φαίνεται εφικτή με ισχυρά CNN backbones, αλλά η ανισορροπία των κλάσεων και τα false negatives παραμένουν σημαντικά σημεία προσοχής. Η εξηγησιμότητα με Grad-CAM++ και SHAP πρόσθεσε χρήσιμη πληροφορία, επειδή έδειξε πότε το μοντέλο εστιάζει πάνω στη βλάβη και πότε, παρότι εστιάζει σωστά χωρικά, οδηγείται σε λάθος πρόβλεψη.

4.15 Πολυταξική ταξινόμηση στο HAM10000 με PAM-CNN-BiLSTM και εξηγησιμότητα Grad-CAM++ / SHAP

Στο πείραμα αυτό εξετάστηκε η πολυταξική ταξινόμηση του συνόλου δεδομένων HAM10000 με ένα μοντέλο που συνδυάζει ResNet50, Parallel Attention Module (PAM) και BiLSTM. Ο στόχος ήταν η αναγνώριση των επτά κατηγοριών του dataset, δηλαδή akiec, bcc, bkl, df, mel, nv και vasc. Το συγκεκριμένο πείραμα έχει ιδιαίτερο ενδιαφέρον, επειδή δεν περιορίζεται σε ένα απλό δυαδικό σχήμα, αλλά επιχειρεί να διαχωρίσει επτά διαφορετικές διαγνώσεις, κάτι που είναι αισθητά πιο απαιτητικό.

Το HAM10000 περιλάμβανε συνολικά 10.015 εικόνες. Η αρχική κατανομή των κλάσεων ήταν έντονα ανισόρροπη: nv = 6705, mel = 1113, bkl = 1099, bcc = 514, akiec = 327, vasc = 142 και df = 115. Η ανισορροπία αυτή είναι ένα από τα βασικά χαρακτηριστικά του dataset και επηρεάζει άμεσα τόσο την εκπαίδευση όσο και την ερμηνεία των αποτελεσμάτων. Η κλάση nv κυριαρχεί αριθμητικά, ενώ κλάσεις όπως df και vasc έχουν πολύ μικρότερη εκπροσώπηση.

Κλάση	Πλήθος εικόνων	Ποσοστό (%)
nv	6.705	66.95%
mel	1.113	11.11%
bkl	1.099	10.97%
bcc	514	5.13%
akiec	327	3.27%
vasc	142	1.42%
df	115	1.15%
Σύνολο	10.015	100.00%

Πίνακας 4.47. Κατανομή των κλάσεων στο HAM10000

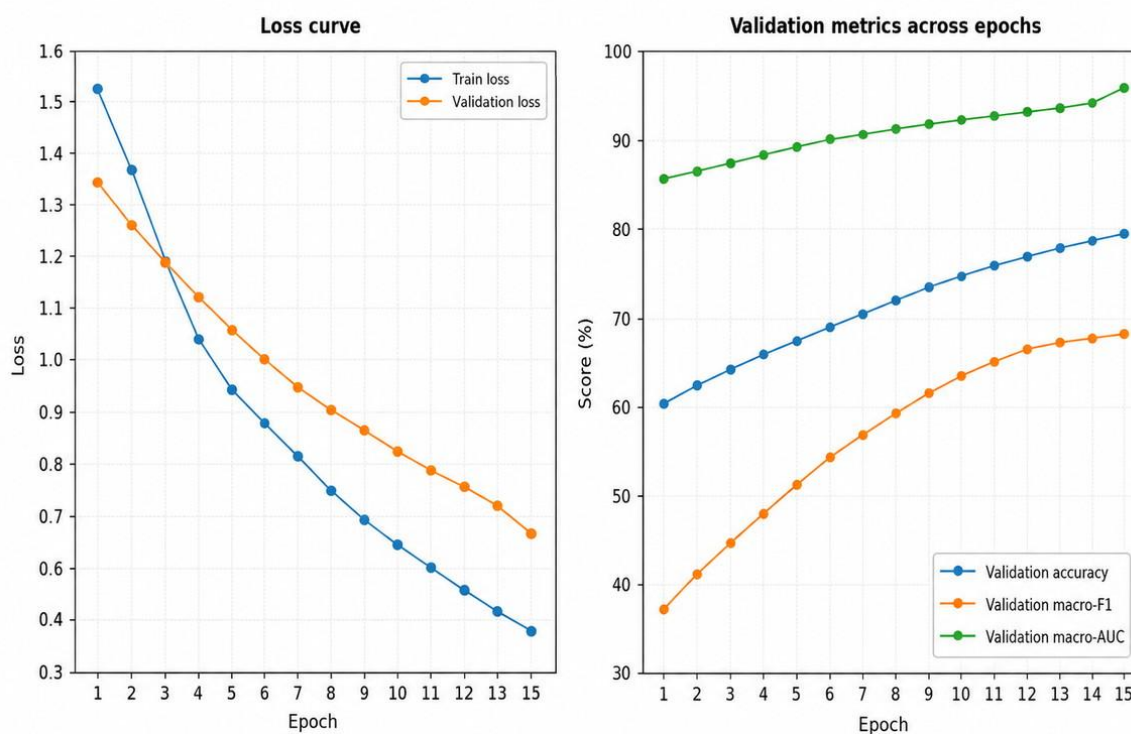
Μετά την προεπεξεργασία, το τελικό dataframe παρέμεινε στις 10.015 εικόνες, χωρίς ελλείψεις στις διαδρομές των αρχείων. Οι εικόνες χωρίστηκαν σε 7010 για εκπαίδευση, 1502 για επικύρωση και 1503 για έλεγχο. Η κατανομή των κλάσεων διατηρήθηκε και στα τρία splits, ώστε η αξιολόγηση να γίνει με τρόπο πιο αντιπροσωπευτικό της συνολικής δομής του dataset.

Στο training set υπήρχαν 4693 nv, 779 mel, 769 bkl, 360 bcc, 229 akiec, 99 vasc και 81 df εικόνες. Στο validation set υπήρχαν 1006 nv, 167 mel, 165 bkl, 77 bcc, 49 akiec, 21 vasc και 17 df. Αντίστοιχα, στο test set υπήρχαν 1006 nv, 167 mel, 165 bkl, 77 bcc, 49 akiec, 22 vasc και 17 df.

Σύνολο δεδομένων	nv	mel	bkl	bcc	akiec	vasc	df	Σύνολο
Train	4693	779	769	360	229	99	81	7010
Validation	1006	167	165	77	49	21	17	1502
Test	1006	167	165	77	49	22	17	1503
Σύνολο	6705	1113	1099	514	327	142	115	10015

Πίνακας 4.48. Κατανομή των κλάσεων σε train, validation και test σύνολο

Η εκπαίδευση πραγματοποιήθηκε για 15 εποχές. Εξαιτίας της ανισορροπίας των κλάσεων, χρησιμοποιήθηκαν class weights, ώστε το μοντέλο να μη μάθει σχεδόν αποκλειστικά την κυρίαρχη κλάση nv. Η διαδικασία της εκπαίδευσης έδειξε σταδιακή βελτίωση. Στην 1η εποχή, το μοντέλο είχε validation accuracy = 60,25%, validation macro-F1 = 37,62% και validation macro-AUC = 85,98%. Μέχρι την 15η εποχή, οι επιδόσεις βελτιώθηκαν σε validation accuracy = 79,56%, validation macro-F1 = 68,19% και validation macro-AUC = 95,88%. Το καλύτερο checkpoint αποθηκεύτηκε τελικά στην 15η εποχή, καθώς εκεί καταγράφηκε η υψηλότερη τιμή macro-F1 στο validation set.



Εικόνα 4.69. Εξέλιξη της απώλειας εκπαίδευσης/επικύρωσης και των validation μετρικών κατά τη διάρκεια της εκπαίδευσης

Από τις καμπύλες εκπαίδευσης φαίνεται ότι η training loss ακολουθεί σταθερά πτωτική πορεία, από 1.5174 στην πρώτη εποχή σε 0.6829 στην τελευταία. Η validation loss επίσης μειώνεται συνολικά, παρότι παρουσιάζει μικρές διακυμάνσεις στη μέση της εκπαίδευσης. Το στοιχείο αυτό είναι αναμενόμενο σε ένα πιο δύσκολο πολυταξικό πρόβλημα, ειδικά όταν

υπάρχουν αρκετές υποαντιπροσωπευόμενες κλάσεις. Το πιο ουσιαστικό σημείο είναι ότι η macro-F1 αυξάνεται από περίπου 0.38 σε περίπου 0.68, ενώ η macro-AUC φτάνει πολύ ψηλά, κοντά στο 0.96. Με απλά λόγια, το μοντέλο γίνεται προοδευτικά καλύτερο όχι μόνο στη συνολική ακρίβεια, αλλά και στην ισορροπημένη συμπεριφορά απέναντι στις διαφορετικές κλάσεις.

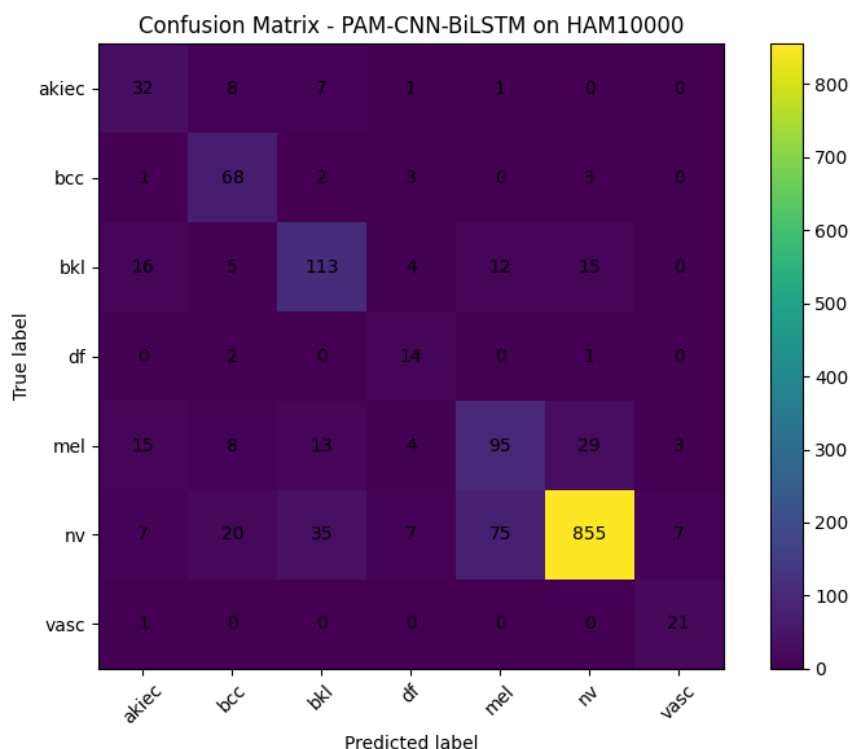
Στο test set, το τελικό μοντέλο πέτυχε Accuracy = 79,71%, Macro-F1 = 67,40%, Weighted-F1 = 80,62%, Macro-Precision = 61,28%, Macro-Recall = 77,40% και Macro-AUC = 95,18%. Οι τιμές αυτές δείχνουν ότι το μοντέλο έχει καλή συνολική απόδοση, αλλά και ότι το πρόβλημα δεν είναι εύκολο. Η διαφορά ανάμεσα στο weighted-F1 και στο macro-F1 δείχνει καθαρά την επίδραση της ανισορροπίας: το μοντέλο αποδίδει καλύτερα στις πιο συχνές κλάσεις, ενώ οι μικρότερες κλάσεις παραμένουν πιο απαιτητικές.

Μετρική	Τιμή (%)
Accuracy	79.71%
Macro-F1	67.40%
Weighted-F1	80.62%
Macro-Precision	61.28%
Macro-Recall	77.40%
Macro-AUC	95.18%

Πίνακας 4.49. Συνολικά αποτελέσματα του PAM-CNN-BiLSTM στο test set του HAM10000

Αν δούμε πιο αναλυτικά τα αποτελέσματα ανά κατηγορία, η καλύτερη επίδοση εμφανίζεται στην κλάση nv, με precision = 0.95, recall = 0.85 και F1-score = 0.90. Πολύ καλή συμπεριφορά εμφανίζεται και στις κλάσεις bcc και vasc, όπου το μοντέλο πέτυχε F1-score = 0.72 και 0.79 αντίστοιχα. Η κλάση df είχε μικρό αριθμό δειγμάτων, αλλά αξιοσημείωτα υψηλό recall = 0.82, παρ' όλο που το precision = 0.42 παραμένει χαμηλό. Οι πιο δύσκολες κλάσεις ήταν κυρίως η mel και η akiec, όπου το μοντέλο είχε χαμηλότερη ακρίβεια διάκρισης. Για την mel, οι τιμές ήταν precision = 0.52, recall = 0.57 και F1-score = 0.54, κάτι που δείχνει ότι το μελάνωμα εξακολουθεί να αποτελεί κλάση με ουσιαστική δυσκολία. Το σημείο αυτό είναι σημαντικό, γιατί η accuracy από μόνη της δεν αρκεί για να περιγράψει τι πραγματικά συμβαίνει. Ένα ποσοστό κοντά στο 80% ακούγεται αρκετά καλό, αλλά η macro-F1 δείχνει ότι η απόδοση δεν είναι το ίδιο ισχυρή σε όλες τις κλάσεις. Στην πράξη, η πολυταξική διάκριση ανάμεσα σε παρόμοιες δερματοσκοπικές εικόνες είναι ένα αρκετά πιο σύνθετο πρόβλημα από ό,τι φαίνεται με μια πρώτη ματιά.

Ο πίνακας σύγχυσης δείχνει με σαφήνεια πού δυσκολεύεται περισσότερο το μοντέλο. Η κλάση nv έχει τις περισσότερες σωστές ταξινομήσεις (855 σωστές προβλέψεις), κάτι αναμενόμενο αφού είναι και η μεγαλύτερη κλάση. Από την άλλη, υπάρχουν αξιοσημείωτες συγχύσεις ανάμεσα σε mel και nv, καθώς 29 περιπτώσεις melanoma ταξινομήθηκαν ως nv, ενώ και 75 περιπτώσεις nv ταξινομήθηκαν ως mel. Παρόμοια, η κλάση bkl συγχέεται με akiec, mel και nv, ενώ η akiec επίσης παρουσιάζει κάποια διασπορά προς γειτονικές κλάσεις.

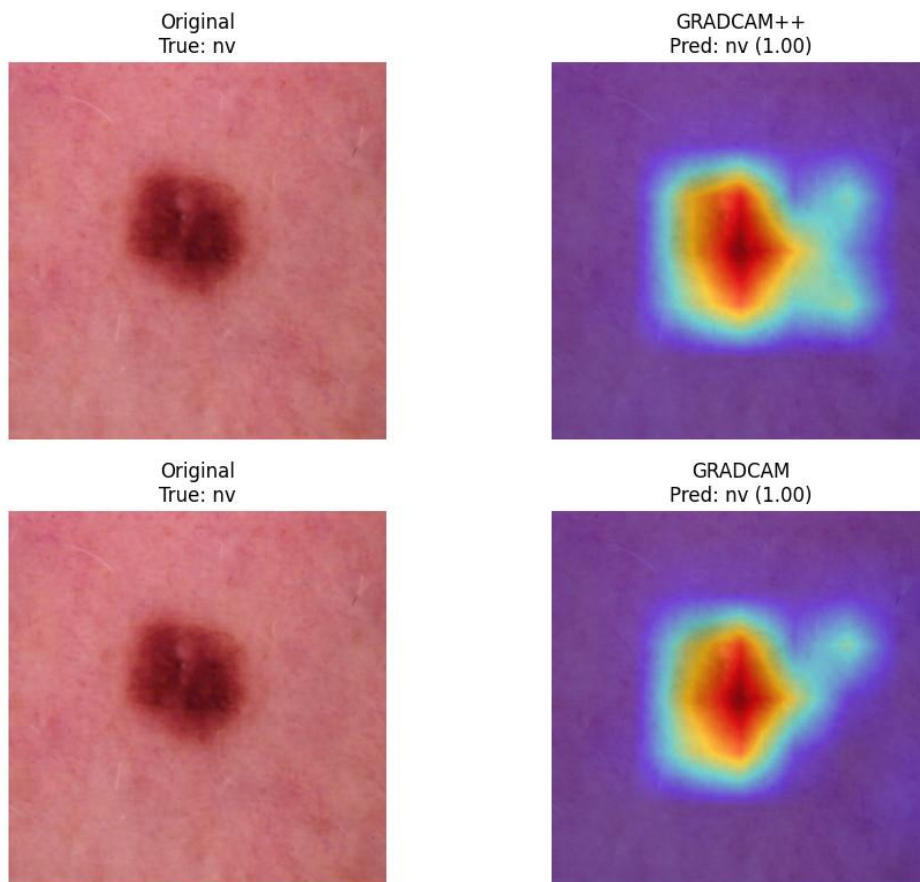


Εικόνα 4.70. Πίνακας σύγχυσης του PAM-CNN-BiLSTM στο test set του HAM10000

Αυτή η εικόνα είναι χρήσιμη, γιατί δείχνει ότι τα λάθη δεν κατανέμονται τυχαία. Το μοντέλο μπερδεύεται κυρίως σε ζεύγη κλάσεων που έχουν οπτικές ομοιότητες, και αυτό έχει ενδιαφέρον και κλινικά αλλά και τεχνικά. Δεν πρόκειται απλώς για “αποτυχία” του μοντέλου, αλλά για σημεία όπου η οπτική πληροφορία είναι πιο αμφίσημη.

Για την ποιοτική αξιολόγηση εφαρμόστηκαν Grad-CAM, Grad-CAM++ και SHAP. Από τα παραδείγματα που παρήχθησαν, δεν αξίζει να χρησιμοποιηθούν όλα στο κεφάλαιο. Ορισμένα heatmaps είναι αρκετά διάχυτα ή δίνουν έμφαση και σε περιοχές εκτός βλάβης, οπότε αν τα βάλουμε όλα, περισσότερο θα μπερδέψουν την παρουσίαση. Για αυτό, είναι προτιμότερο να επιλεγούν μόνο τα πιο καθαρά και ερμηνεύσιμα παραδείγματα. Από τα σωστά παραδείγματα, το πιο καθαρό είναι ένα παράδειγμα nevus (nv) που ταξινομήθηκε σωστά ως nv με πιθανότητα 1.00, όπου το Grad-CAM++ συγκεντρώνεται αρκετά καθαρά πάνω στην περιοχή της βλάβης. Πολύ καλό είναι και το δεύτερο σωστό παράδειγμα nv, στο οποίο το μοντέλο και πάλι προβλέπει nv (1.00) και η ενεργοποίηση ακολουθεί τη βασική μορφή της βλάβης. Ένα ακόμη χρήσιμο σωστό παράδειγμα είναι η περίπτωση bkl → bkl (0.63). Παρότι η πιθανότητα είναι χαμηλότερη και η ενεργοποίηση πιο

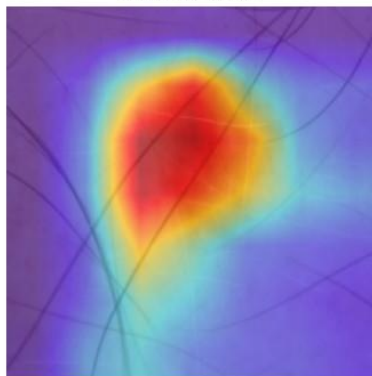
περιορισμένη, η εικόνα αυτή είναι χρήσιμη γιατί δείχνει μια σωστή πρόβλεψη και σε κλάση διαφορετική από την κυρίαρχη nv.



Original
True: nv



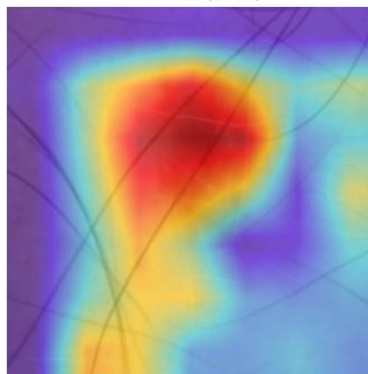
GRADCAM++
Pred: nv (1.00)

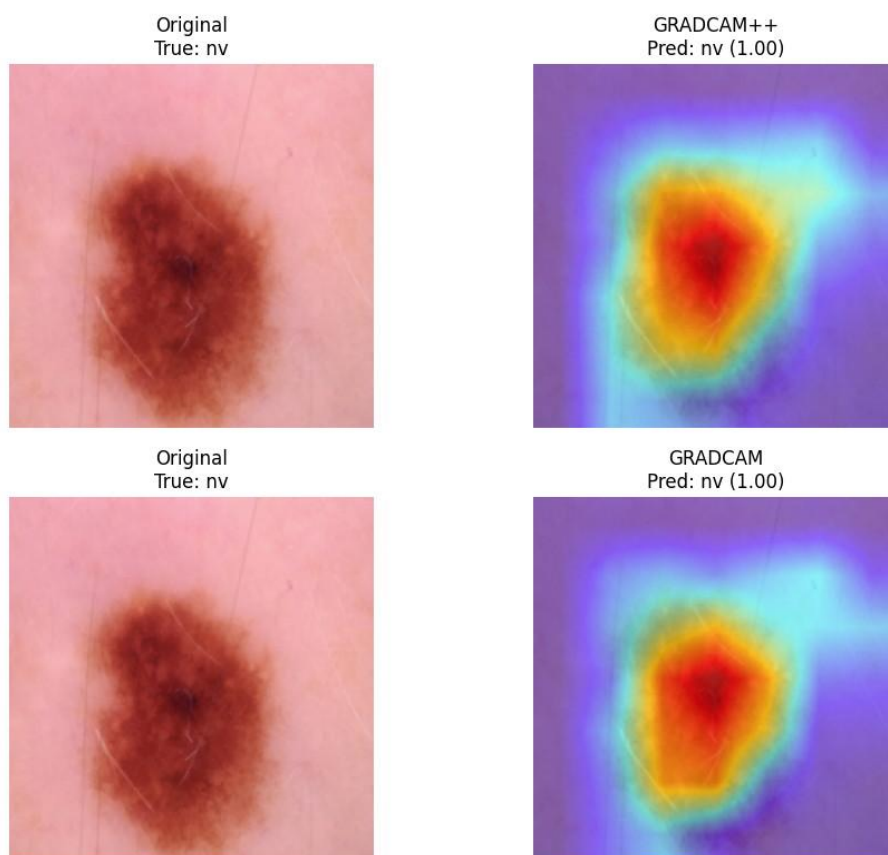


Original
True: nv



GRADCAM
Pred: nv (1.00)

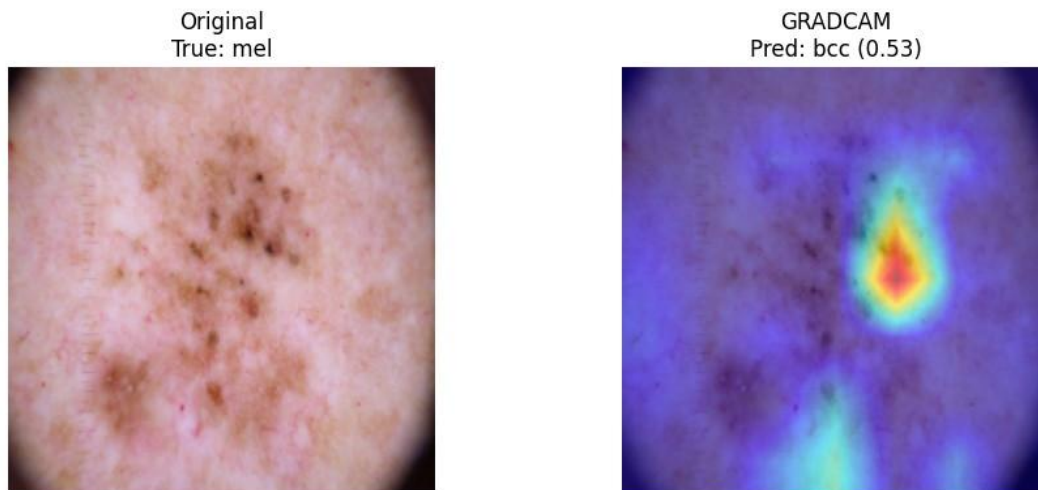




Εικόνα 4.71. Ενδεικτικά σωστά παραδείγματα και αντίστοιχοι Grad-CAM και Grad-CAM++ χάρτες

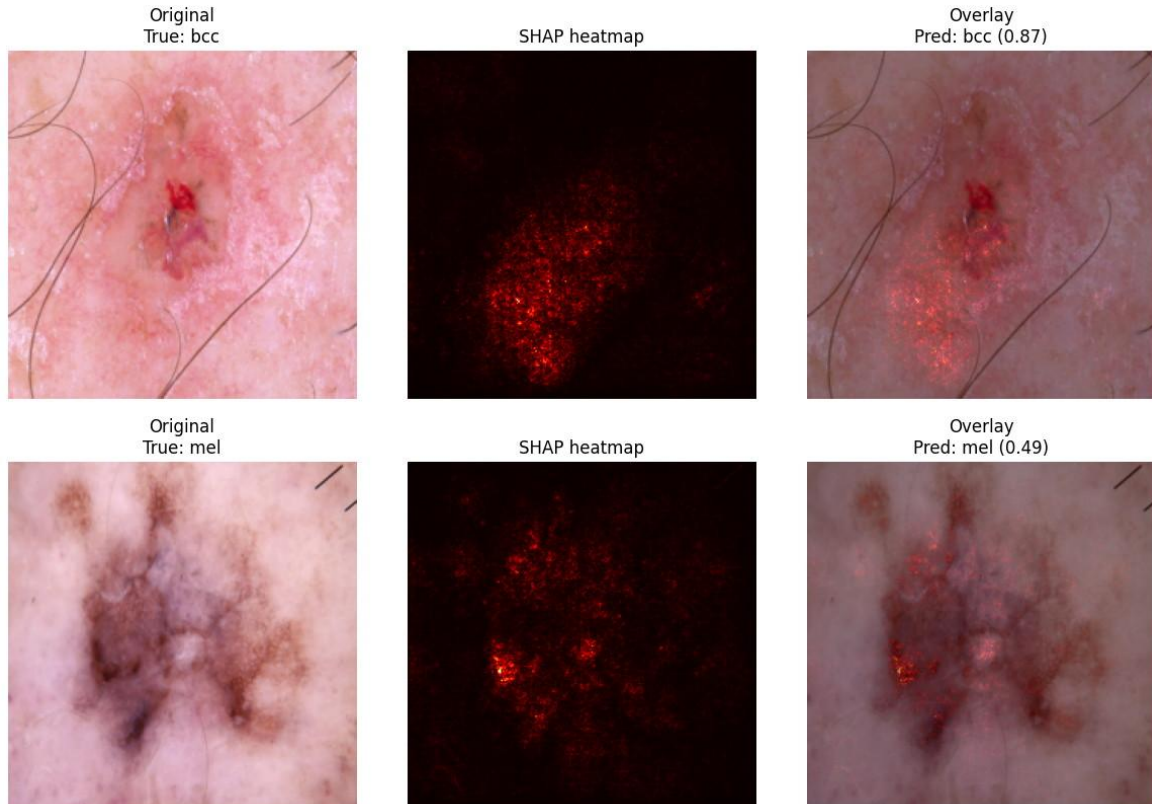
Στα σωστά παραδείγματα, το Grad-CAM++ φαίνεται γενικά πιο καθαρό από το απλό Grad-CAM. Το heatmap είναι πιο συγκεντρωμένο και, στις καλύτερες περιπτώσεις, καλύπτει τη βασική περιοχή της βλάβης χωρίς να απλώνεται υπερβολικά στο φόντο. Αυτό δεν σημαίνει ότι η εξήγηση είναι τέλεια σε όλες τις περιπτώσεις, αλλά είναι αρκετά πειστική ώστε να υποστηρίξει το ότι το μοντέλο “κοιτάζει” το σωστό μέρος της εικόνας όταν η πρόβλεψη είναι σωστή.

Από τα λανθασμένα παραδείγματα, το πιο σημαντικό είναι η περίπτωση $mel \rightarrow bcc$ (0.53). Η εικόνα αυτή έχει μεγαλύτερη ερμηνευτική αξία, γιατί αφορά ένα μελάνωμα που το μοντέλο δεν αναγνώρισε σωστά. Το Grad-CAM++ δείχνει ότι υπάρχει εστίαση σε συγκεκριμένο σημείο της βλάβης, αλλά αυτή η εστίαση δεν οδήγησε στη σωστή διάγνωση. Με άλλα λόγια, το πρόβλημα εδώ δεν είναι απλώς ότι το μοντέλο αγνόησε τη βλάβη, αλλά ότι αξιολόγησε λανθασμένα τα χαρακτηριστικά της.



Εικόνα 4.72. Παράδειγμα λανθασμένης πρόβλεψης και αντίστοιχο Grad-CAM++ χάρτης

Για το SHAP, η πρώτη εικόνα είναι το σωστό παράδειγμα $bcc \rightarrow bcc$ (prob = 0.87). Οι θετικές περιοχές εμφανίζονται σε τμήματα που σχετίζονται οπτικά με τη βλάβη. Η δεύτερη εικόνα είναι σωστό παράδειγμα $mel \rightarrow mel$ (prob = 0.49). Η πιθανότητα εδώ είναι χαμηλότερη, κάτι που κάνει το παράδειγμα ακόμη πιο ενδιαφέρον, το μοντέλο ταξινομεί σωστά, αλλά με σχετικά μικρότερη βεβαιότητα. Αυτό μας επιτρέπει να δείξουμε ότι το SHAP δεν χρησιμοποιείται μόνο σε εύκολες και πολύ βέβαιες περιπτώσεις.



Εικόνα 4.73. Ενδεικτικές SHAP απεικονίσεις για σωστές ταξινομήσεις στο HAM10000

Συνολικά, το πείραμα αυτό δείχνει ότι το PAM-CNN-BiLSTM μπορεί να επιτύχει αρκετά καλή απόδοση στην πολυταξική ταξινόμηση του HAM10000, παρότι πρόκειται για πρόβλημα με υψηλή ανισορροπία και ουσιαστικές οπτικές ομοιότητες ανάμεσα σε ορισμένες κλάσεις. Η τελική επίδοση του μοντέλου, με Accuracy = 79,71%, Macro-F1 = 67,40% και Macro-AUC = 95,18%, είναι ικανοποιητική, αλλά δεν πρέπει να ερμηνευτεί επιφανειακά. Η υψηλή τιμή του weighted-F1 σε σχέση με το macro-F1 θυμίζει ότι οι πιο μεγάλες κλάσεις ωφελούν περισσότερο το τελικό σκορ, ενώ οι μικρότερες και πιο δύσκολες κλάσεις συνεχίζουν να πιέζουν την επίδοση προς τα κάτω.

4.16 Πείραμα δυαδικής ταξινόμησης στο ISIC 2019 με EfficientNet-B0, Grad-CAM++ και t-SNE

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο EfficientNet-B0 για δυαδική ταξινόμηση δερματικών βλαβών στο σύνολο δεδομένων ISIC 2019. Όπως και στο προηγούμενο πείραμα με το EfficientNet-B4, οι αρχικές κατηγορίες δεν χρησιμοποιήθηκαν ως ξεχωριστές πολυταξικές ετικέτες, αλλά ομαδοποιήθηκαν σε δύο γενικότερες ομάδες: benign-like και suspicious. Στην ομάδα suspicious εντάχθηκαν οι κατηγορίες AK, BCC, MEL και SCC, ενώ στην ομάδα benign-like εντάχθηκαν οι NV, BKL, DF και VASC. Η αρχική βάση περιλάμβανε 25.329 εικόνες. Η κατανομή των αρχικών κλάσεων ήταν έντονα ανισορροπη, με την κλάση NV να έχει 12.875 εικόνες, ενώ μικρότερες κλάσεις όπως DF και VASC είχαν μόλις 239 και 253 εικόνες αντίστοιχα. Μετά τη δυαδική ομαδοποίηση, η κλάση benign-like περιλάμβανε 15.990 εικόνες, ενώ η κλάση suspicious περιλάμβανε 9.339 εικόνες. Για να περιοριστεί η επίδραση αυτής της ανισορροπίας, δημιουργήθηκε ισορροπημένο υποσύνολο με 3.000 benign-like και 3.000 suspicious εικόνες, δηλαδή 6.000 εικόνες συνολικά. Το ισορροπημένο υποσύνολο χωρίστηκε σε 4.199 εικόνες εκπαίδευσης, 901 εικόνες επικύρωσης και 900 εικόνες ελέγχου. Η κατανομή παρέμεινε σχεδόν απόλυτα ισορροπημένη σε όλα τα splits. Στο training set υπήρχαν 2.099 benign-like και 2.100 suspicious εικόνες. Στο validation set υπήρχαν 451 benign-like και 450 suspicious, ενώ στο test set υπήρχαν 450 εικόνες από κάθε κλάση.

Split	Σύνολο	Benign-like	Suspicious
Train	4199	2099	2100
Validation	901	451	450
Test	900	450	450

Πίνακας 4.50. Κατανομή του balanced subset στο πείραμα EfficientNet-B0

Το μοντέλο εκπαιδεύτηκε για 15 εποχές. Σε κάθε εποχή έγινε αξιολόγηση στο validation set και επιλογή threshold, ώστε η τελική δυαδική απόφαση να μην εξαρτάται υποχρεωτικά από το τυπικό όριο 0.50. Αυτή η διαδικασία είναι χρήσιμη σε ιατρικά προβλήματα, επειδή η μετακίνηση του threshold μπορεί να επηρεάσει σημαντικά τη σχέση ανάμεσα σε precision και recall.

Στην 1η εποχή, το μοντέλο είχε $\text{val_acc} = 72,70\%$, $\text{val_precision} = 71,70\%$, $\text{val_recall} = 74,89\%$, $\text{val_f1} = 73,26\%$ και $\text{val_auc} = 79,93\%$. Η απόδοση βελτιώθηκε σταδιακά κατά την εκπαίδευση. Στην 5η εποχή, το validation F1 ανέβηκε σε $80,74\%$, ενώ στην 10η εποχή το validation AUC έφτασε στο $89,19\%$. Το καλύτερο checkpoint αποθηκεύτηκε στην 14η εποχή, όπου το μοντέλο πέτυχε $\text{val_acc} = 81,91\%$, $\text{val_precision} = 81,26\%$, $\text{val_recall} = 82,89\%$, $\text{val_f1} = 82,07\%$ και $\text{val_auc} = 89,42\%$, με threshold 0.43.

Epoch	Train loss	Val accuracy (%)	Val precision (%)	Val recall (%)	Val F1 (%)	Val AUC (%)	Threshold
1	1.2713	72.70%	71.70%	74.89%	73.26%	79.93%	0.19
3	0.5618	76.58%	72.42%	85.78%	78.54%	85.31%	0.10
5	0.2918	81.47%	83.93%	77.78%	80.74%	87.61%	0.75
10	0.1333	81.35%	82.94%	78.89%	80.87%	89.19%	0.49
14	0.0939	81.91%	81.26%	82.89%	82.07%	89.42%	0.43
15	0.0899	80.47%	76.97%	86.89%	81.63%	89.17%	0.23

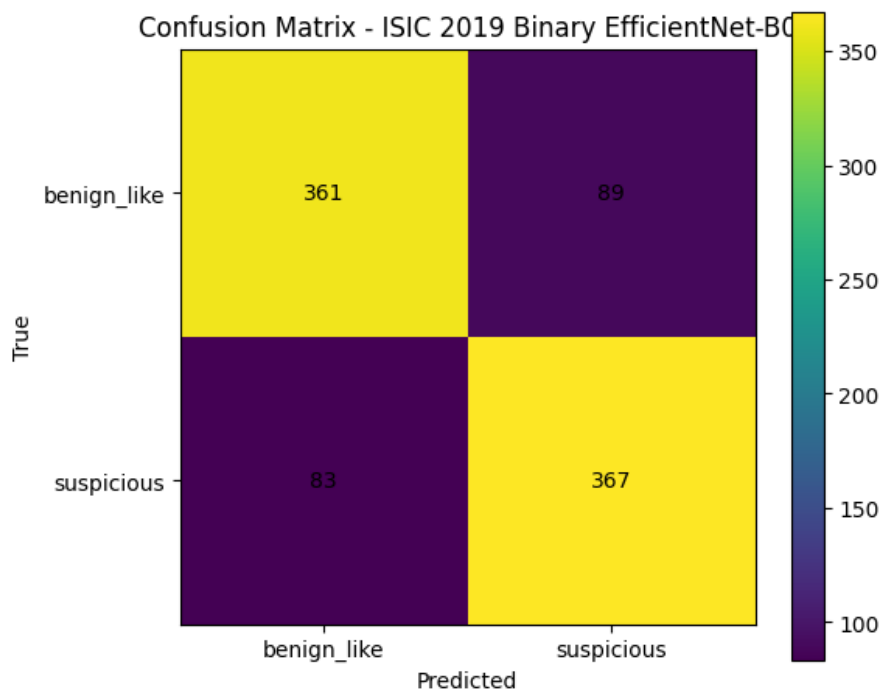
Πίνακας 4.51. Πορεία εκπαίδευσης του EfficientNet-B0 στο ISIC 2019

Στο test set, το καλύτερο μοντέλο της 14ης εποχής αξιολογήθηκε με threshold 0.43. Τα αποτελέσματα ήταν Test Accuracy = $80,89\%$, Test Precision = $80,48\%$, Test Recall = $81,56\%$, Test F1 = $81,02\%$ και Test AUC = $88,81\%$. Σε αντίθεση με το προηγούμενο EfficientNet-B4 πείραμα, εδώ οι δύο κλάσεις εμφανίζουν πιο ισορροπημένη συμπεριφορά, καθώς το test set έχει ακριβώς ίδιο αριθμό εικόνων ανά κατηγορία.

Metric	Value
Best epoch	14
Threshold	0.43
Test Accuracy	80.89%
Test Precision	80.48%
Test Recall	81.56%
Test F1	81.02%
Test AUC	88.81%

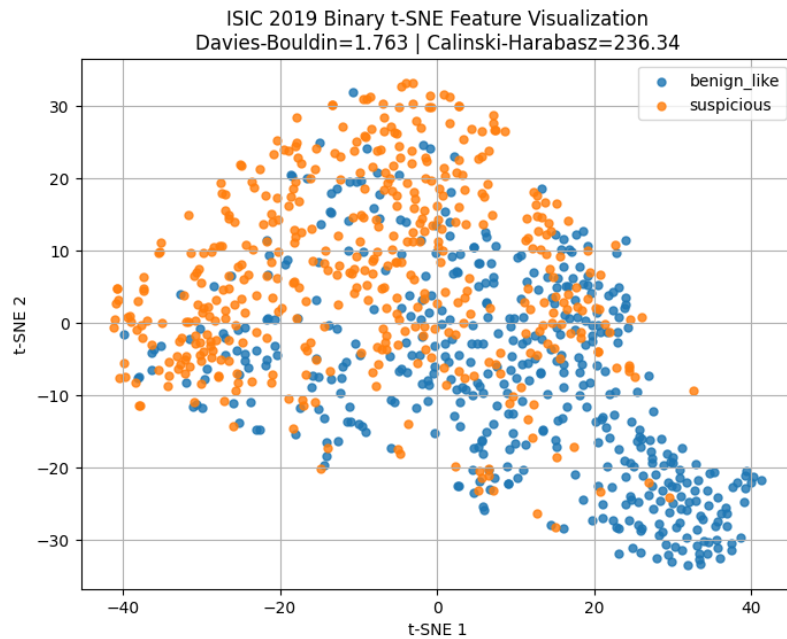
Πίνακας 4.52. Test αποτελέσματα του EfficientNet-B0 στο ISIC 2019

Ο πίνακας σύγχυσης ήταν:



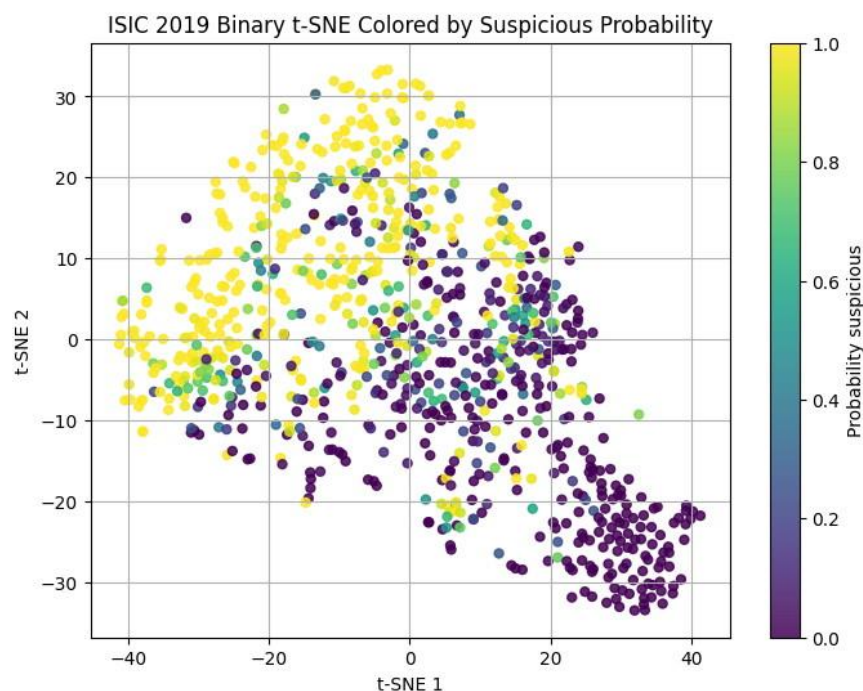
Εικόνα 4.74. Πίνακας σύγχυσης του EfficientNet-B0 στο test set του ISIC 2019

Αυτό σημαίνει ότι από τις 450 benign-like εικόνες, οι 361 ταξινομήθηκαν σωστά και 89 ταξινομήθηκαν λανθασμένα ως suspicious. Από τις 450 suspicious εικόνες, οι 367 εντοπίστηκαν σωστά, ενώ 83 ταξινομήθηκαν λανθασμένα ως benign-like. Τα δύο είδη λαθών είναι σχετικά κοντά αριθμητικά, κάτι που δείχνει ότι το μοντέλο δεν είναι υπερβολικά στραμμένο προς τη μία από τις δύο κλάσεις. Παρ' όλα αυτά, τα 83 false negatives για την suspicious κλάση παραμένουν κρίσιμα, γιατί αντιστοιχούν σε ύποπτες βλάβες που το μοντέλο δεν αναγνώρισε. Από το classification report φαίνεται ότι η κλάση benign-like είχε precision = 0.81, recall = 0.80 και F1-score = 0.81. Η κλάση suspicious είχε precision = 0.80, recall = 0.82 και F1-score = 0.81. Η σχεδόν ίδια τιμή F1 και για τις δύο κλάσεις δείχνει ότι η χρήση balanced subset βοήθησε το μοντέλο να μην ευνοεί υπερβολικά κάποια κατηγορία. Στη συνέχεια έγινε οπτικοποίηση των εσωτερικών χαρακτηριστικών του μοντέλου με t-SNE. Η πρώτη t-SNE εικόνα παρουσιάζει τα δείγματα χρωματισμένα με βάση την πραγματική δυαδική κλάση. Η οπτικοποίηση συνοδεύεται από Davies-Bouldin = 1.763 και Calinski-Harabasz = 236.34. Η εικόνα δείχνει ότι υπάρχει κάποια ομαδοποίηση των χαρακτηριστικών, αλλά οι δύο κλάσεις δεν χωρίζονται πλήρως. Υπάρχουν περιοχές όπου τα benign-like και suspicious δείγματα αναμειγνύονται, κάτι που συμφωνεί με το ότι η τελική accuracy δεν ξεπερνά το 81%.



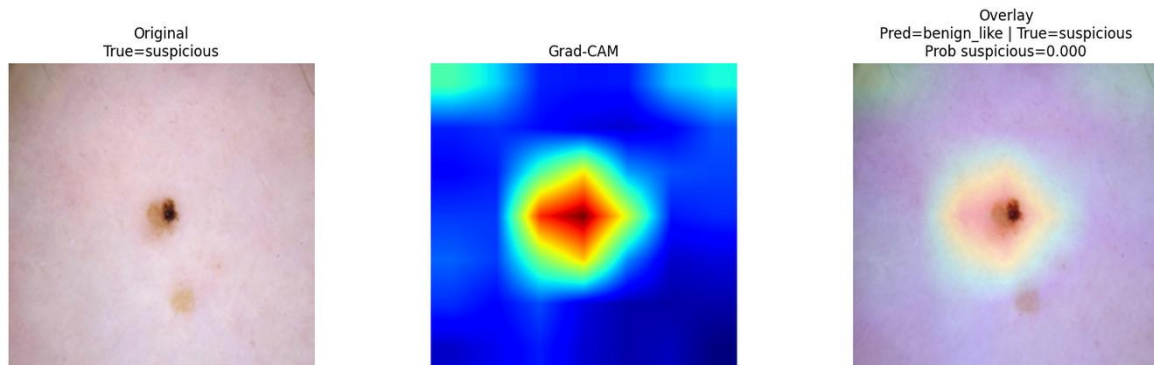
Εικόνα 4.75. t-SNE απεικόνιση των χαρακτηριστικών του EfficientNet-B0 ανά δυαδική κλάση

Η δεύτερη t-SNE εικόνα χρωματίζει τα ίδια χαρακτηριστικά με βάση την πιθανότητα suspicious που αποδίδει το μοντέλο. Αυτή η εικόνα είναι πιο ενδιαφέρουσα ερμηνευτικά, γιατί δεν δείχνει μόνο την πραγματική ετικέτα, αλλά και το πόσο ύποπτα θεωρεί τα δείγματα το μοντέλο. Τα σημεία με υψηλή πιθανότητα suspicious συγκεντρώνονται περισσότερο σε συγκεκριμένες περιοχές του feature space, ενώ άλλες περιοχές έχουν χαμηλότερες πιθανότητες. Η μετάβαση όμως δεν είναι απόλυτα καθαρή, γεγονός που δείχνει ότι το μοντέλο διατηρεί αβεβαιότητα για αρκετές εικόνες.



Εικόνα 4.76. t-SNE απεικόνιση χρωματισμένη με βάση την πιθανότητα suspicious

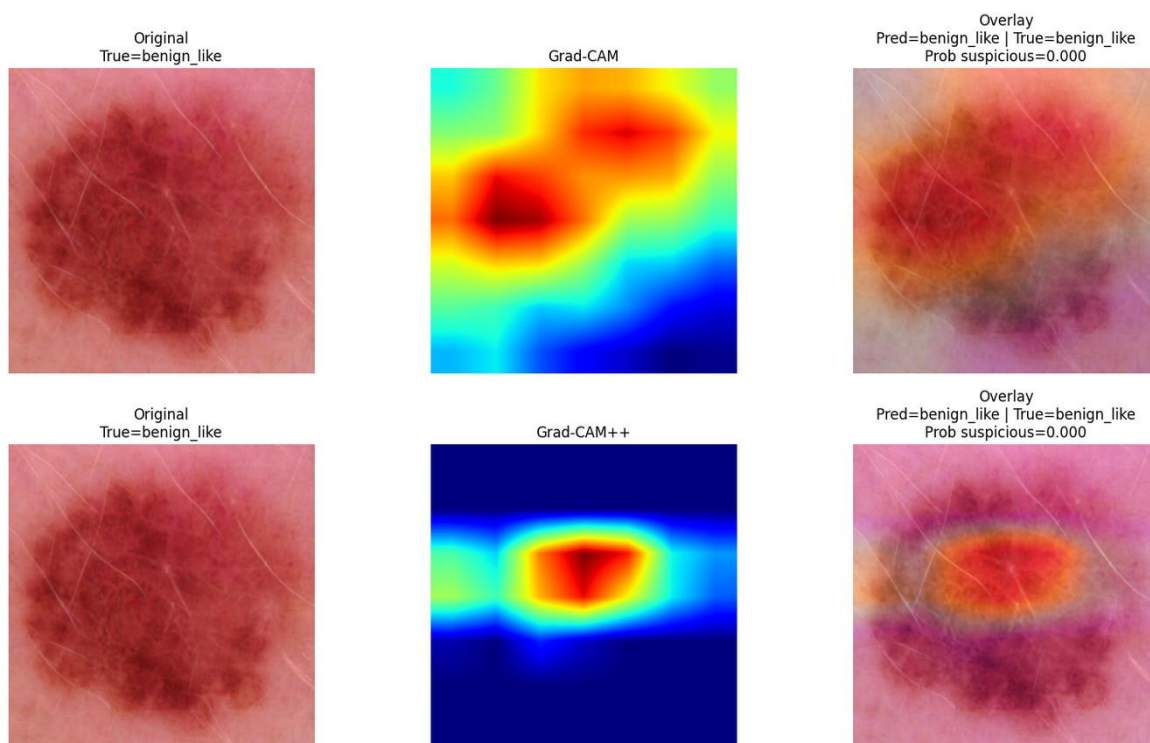
Για την ποιοτική ερμηνεία των προβλέψεων χρησιμοποιήθηκαν Grad-CAM και Grad-CAM++. Η λανθασμένη suspicious εικόνα προβλέφθηκε ως benign-like, με probability suspicious = 0.000. Στο Grad-CAM, μια λανθασμένη suspicious εικόνα προβλέφθηκε ως benign-like, με probability suspicious = 0.000, η ενεργοποίηση βρίσκεται καθαρά πάνω στη μικρή βλάβη,



Εικόνα 4.77. Grad-CAM ανάλυση λανθασμένης suspicious περίπτωσης που ταξινομήθηκε ως benign-like

Η περίπτωση αυτή είναι σημαντική, γιατί δείχνει ότι το μοντέλο μπορεί να εντοπίζει τη βλάβη χωρικά, αλλά να την ερμηνεύει λανθασμένα. Δηλαδή το λάθος δεν είναι απαραίτητα ότι το μοντέλο κοιτάζει αλλού. Το πρόβλημα είναι ότι τα οπτικά χαρακτηριστικά που εντοπίζει δεν οδηγούν στη σωστή απόφαση. Αυτό είναι ένα από τα βασικά σημεία που πρέπει να αναδειχθεί στην εξηγησιμότητα.

Η δεύτερη περίπτωση είναι μια σωστή benign-like πρόβλεψη, όπου το μοντέλο προβλέπει benign-like με probability suspicious = 0.000. Εδώ η Grad-CAM++ εστιάζει αρκετά καθαρά πάνω στην κεντρική περιοχή της βλάβης. Η εικόνα αυτή είναι χρήσιμη γιατί δείχνει ότι το μοντέλο μπορεί να κοιτάζει τη σωστή περιοχή και να απορρίπτει σωστά την suspicious κλάση.



Εικόνα 4.78. Grad-CAM και Grad-CAM++ ανάλυση για ορθή benign-like πρόβλεψη

Συνολικά, το πείραμα με EfficientNet-B0 στο ISIC 2019 έδειξε μια αρκετά ισορροπημένη απόδοση ανάμεσα στις δύο δυαδικές κλάσεις. Το μοντέλο πέτυχε Test Accuracy = 80,89%, Test F1 = 81,02% και Test AUC = 88,81%. Σε σχέση με το προηγούμενο EfficientNet-B4 πείραμα, η απόδοση είναι ελαφρώς καλύτερη ως προς accuracy, F1 και AUC, αλλά το σημαντικότερο είναι ότι οι δύο κλάσεις έχουν πιο ισορροπημένη συμπεριφορά στο classification report. Η t-SNE ανάλυση έδειξε ότι το feature space του EfficientNet-B0 οργανώνει μερικώς τις δύο ομάδες, αλλά δεν τις χωρίζει απόλυτα. Αυτό εξηγεί γιατί το μοντέλο πετυχαίνει ικανοποιητική αλλά όχι άριστη απόδοση. Οι Grad-CAM και Grad-CAM++ εικόνες πρόσθεσαν ένα δεύτερο επίπεδο ερμηνείας. Σε ορισμένες περιπτώσεις το μοντέλο εστιάζει καθαρά πάνω στη βλάβη, ενώ σε άλλες η ενεργοποίηση είναι λιγότερο αξιόπιστη.

4.17 Συγκριτική αξιολόγηση Swin-Tiny + SwiGLU και EfficientNet-B4 στο ISIC 2019

Στο παρόν στάδιο συγκρίνονται τα δύο βασικά πειράματα δυαδικής ταξινόμησης στο ISIC 2019. Το πρώτο πείραμα βασίστηκε στο Swin-Tiny με κεφαλή SwiGLU, ενώ το δεύτερο στο EfficientNet-B4. Και στις δύο περιπτώσεις, οι αρχικές κατηγορίες του ISIC 2019 ομαδοποιήθηκαν σε δύο κλάσεις: benign-like και suspicious. Πρέπει να διευκρινιστεί ότι η σύγκριση δεν είναι απολύτως ισοδύναμη, επειδή τα δύο μοντέλα εκπαιδεύτηκαν σε διαφορετικό μέγεθος ισορροπημένου υποσυνόλου. Το Swin-Tiny + SwiGLU εκπαιδεύτηκε σε σύνολο 6000 εικόνων, ενώ το EfficientNet-B4 σε σύνολο 3000 εικόνων. Για τον λόγο

αυτό, τα αποτελέσματα δεν πρέπει να διαβαστούν ως απόλυτη απόδειξη υπεροχής μιας αρχιτεκτονικής, αλλά ως συγκριτική ένδειξη μέσα στο συγκεκριμένο πειραματικό πλαίσιο.

Παράμετρος	Swin-Tiny + SwiGLU	EfficientNet-B4
Dataset	ISIC 2019	ISIC 2019
Πρόβλημα	Binary classification	Binary classification
Κλάσεις	benign-like / suspicious	benign-like / suspicious
Balanced subset	6000 εικόνες	3000 εικόνες
Εικόνες ανά κλάση	3000 / 3000	1500 / 1500
Train set	4199	2099
Validation set	901	451
Test set	900	450
Test ανά κλάση	450 / 450	225 / 225
Best epoch	10	10

Μετρική	Swin-Tiny + SwiGLU	EfficientNet-B4
Accuracy	81.44%	79.11%
Precision	76.55%	75.49%
Recall	90.67%	86.22%
F1-score	83.01%	80.50%
AUC	90.69%	87.91%
Best epoch	10	10
Threshold	0.29	0.31
Test samples	900	450

Πίνακας 4.53. Συγκριτικός πίνακας και μετρικές σε Swin-Tiny + SwiGLU και EfficientNet-B4

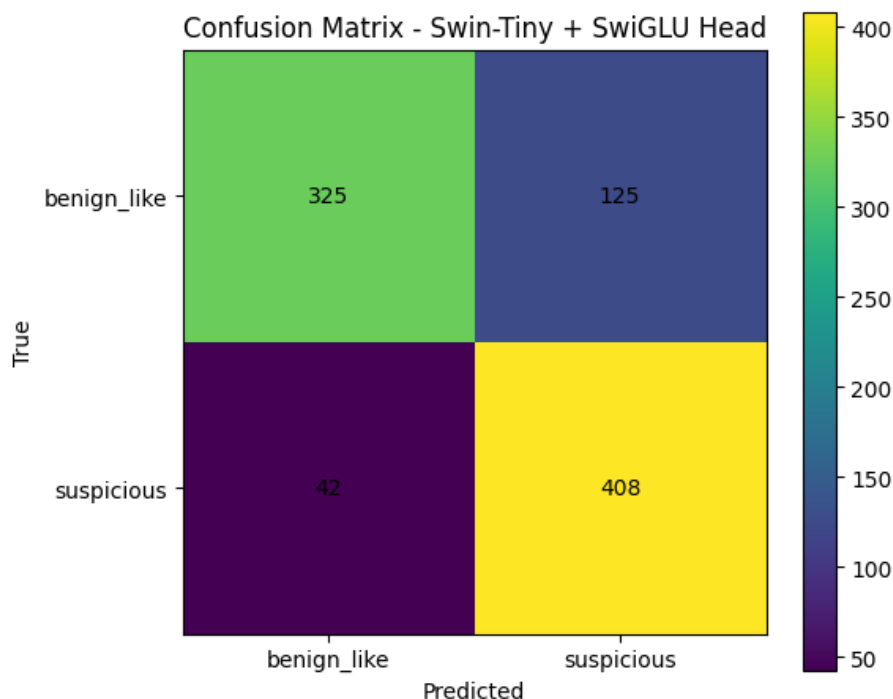
Με βάση τις συνολικές μετρικές, το Swin-Tiny + SwiGLU εμφανίζει καλύτερη απόδοση από το EfficientNet-B4. Η διαφορά δεν είναι πολύ μεγάλη, αλλά είναι σταθερή σε όλες τις βασικές μετρικές. Το Swin-Tiny + SwiGLU έχει υψηλότερο accuracy, recall, F1-score και AUC. Η πιο ουσιαστική διαφορά βρίσκεται στο recall της suspicious κλάσης. Το Swin-Tiny + SwiGLU πέτυχε recall 90.67%, ενώ το EfficientNet-B4 πέτυχε 86.22%. Σε ένα πρόβλημα ταξινόμησης δερματικών βλαβών, αυτή η διαφορά έχει ιδιαίτερη βαρύτητα, επειδή δείχνει πόσες ύποπτες βλάβες εντοπίζονται σωστά από το μοντέλο.

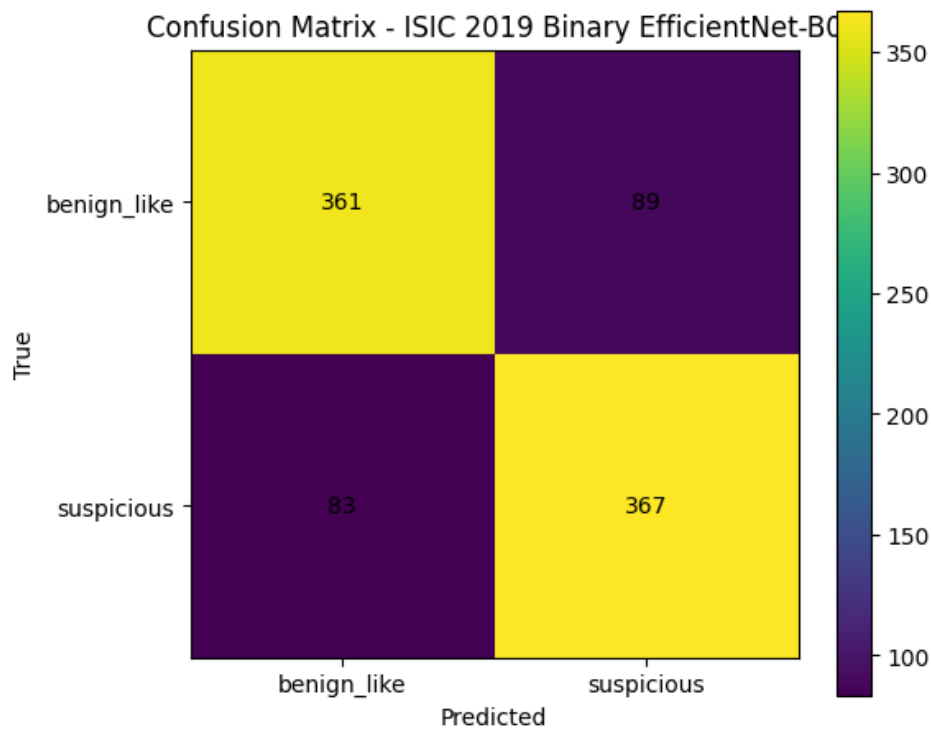
Η confusion matrix του Swin-Tiny + SwiGLU δείχνει ότι το μοντέλο αναγνώρισε σωστά 408 από τις 450 suspicious εικόνες, ενώ ταξινόμησε λανθασμένα 42 suspicious δείγματα ως benign-like. Αντίστοιχα, το EfficientNet-B4 αναγνώρισε σωστά 194 από τις 225 suspicious

εικόνες και ταξινομήσε λανθασμένα 31 suspicious δείγματα ως benign-like. Αν και τα δύο μοντέλα εμφανίζουν καλή ευαισθησία στην suspicious κλάση, το Swin-Tiny + SwiGLU φαίνεται να διατηρεί υψηλότερη ικανότητα εντοπισμού των ύποπτων περιπτώσεων. Το ενδιαφέρον εύρημα είναι ότι τα δύο μοντέλα έχουν σχεδόν ίδια συμπεριφορά στην κλάση benign-like. Το Swin-Tiny + SwiGLU ταξινομεί σωστά το 72.22% των benign-like εικόνων, ενώ το EfficientNet-B4 το 72.00%. Άρα η βασική διαφορά ανάμεσα στα δύο μοντέλα δεν βρίσκεται στην αναγνώριση των benign-like περιπτώσεων. Η διαφορά εντοπίζεται κυρίως στην κλάση suspicious. Το Swin-Tiny + SwiGLU αναγνωρίζει σωστά το 90.67% των suspicious εικόνων, ενώ το EfficientNet-B4 το 86.22%. Παράλληλα, το Swin-Tiny + SwiGLU έχει μικρότερο ποσοστό false negatives στην suspicious κλάση, δηλαδή 9.33%, σε σχέση με το 13.78% του EfficientNet-B4.

Μοντέλο	Benign-like σωστά	Benign-like ως suspicious	Suspicious σωστά	Suspicious ως benign-like
Swin-Tiny + SwiGLU	325/450 = 72.22%	125/450 = 27.78%	408/450 = 90.67%	42/450 = 9.33%
EfficientNet-B4	162/225 = 72.00%	63/225 = 28.00%	194/225 = 86.22%	31/225 = 13.78%

Πίνακας 4.54. Σύγκριση των confusion matrices με βάση τα ποσοστά ανά κλάση

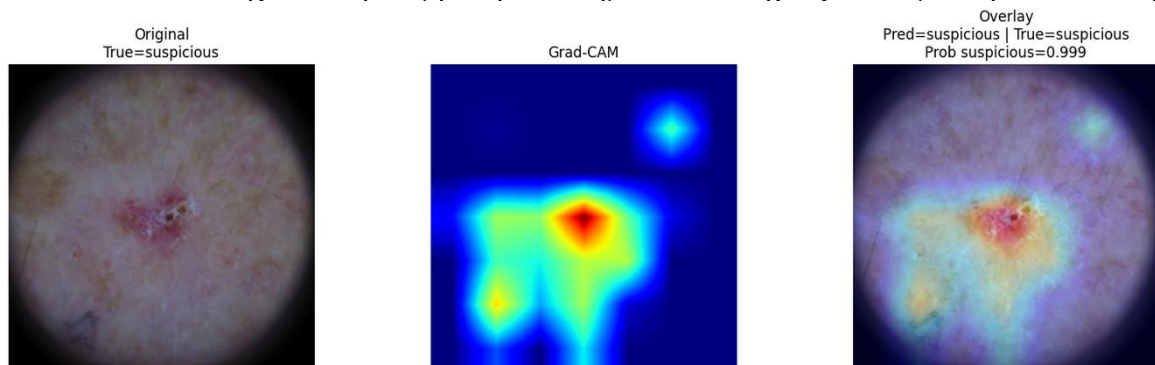




Εικόνα 4.79. Confusion matrix του μοντέλου Swin-Tiny + SwiGLU και του μοντέλου EfficientNet-B4 στο test set.

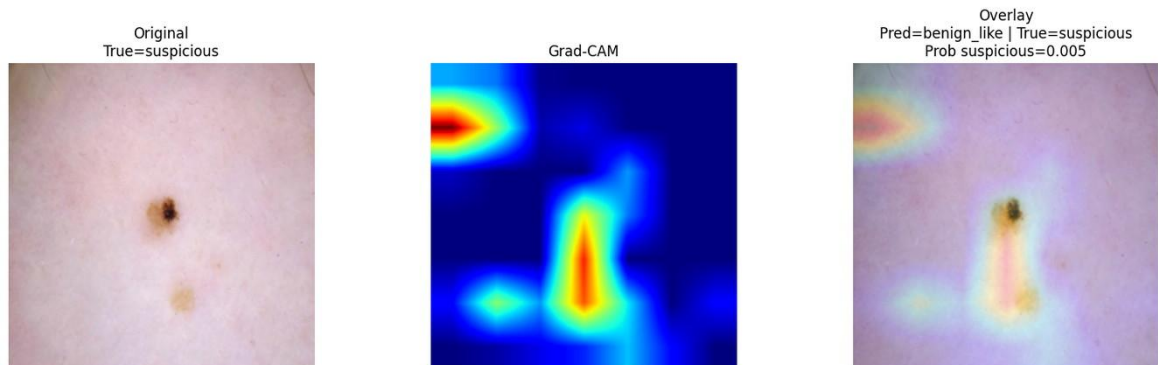
Σε επίπεδο αρχιτεκτονικής, η διαφορά αυτή μπορεί να συνδεθεί με τον τρόπο που τα δύο μοντέλα επεξεργάζονται την εικόνα. Το EfficientNet-B4, ως CNN μοντέλο, βασίζεται κυρίως σε τοπικά χαρακτηριστικά, όπως υφή, χρώμα και περιγράμματα. Το Swin-Tiny, από την άλλη, αξιοποιεί μηχανισμό προσοχής μέσα σε τοπικά παράθυρα και μπορεί να συνδυάσει πληροφορία από διαφορετικές περιοχές της εικόνας. Στην πράξη, αυτό φαίνεται να το βοήθησε να πετύχει ελαφρώς καλύτερη διάκριση ανάμεσα στις benign-like και suspicious βλάβες.

Στις εικόνες του Swin-Tiny + SwiGLU παρατηρείται ότι οι σωστές suspicious προβλέψεις έχουν συχνά πιο εμφανή εστίαση στην περιοχή της βλάβης ή γύρω από αυτή. Αυτό φαίνεται καθαρά στην περίπτωση όπου το μοντέλο προβλέπει suspicious με πιθανότητα 0.999. Στο αντίστοιχο Grad-CAM, η ενεργοποίηση καλύπτει την κεντρική περιοχή της βλάβης, ενώ το SHAP δίνει επίσης θετική συμβολή σε σημεία που σχετίζονται με την αλλοίωση.

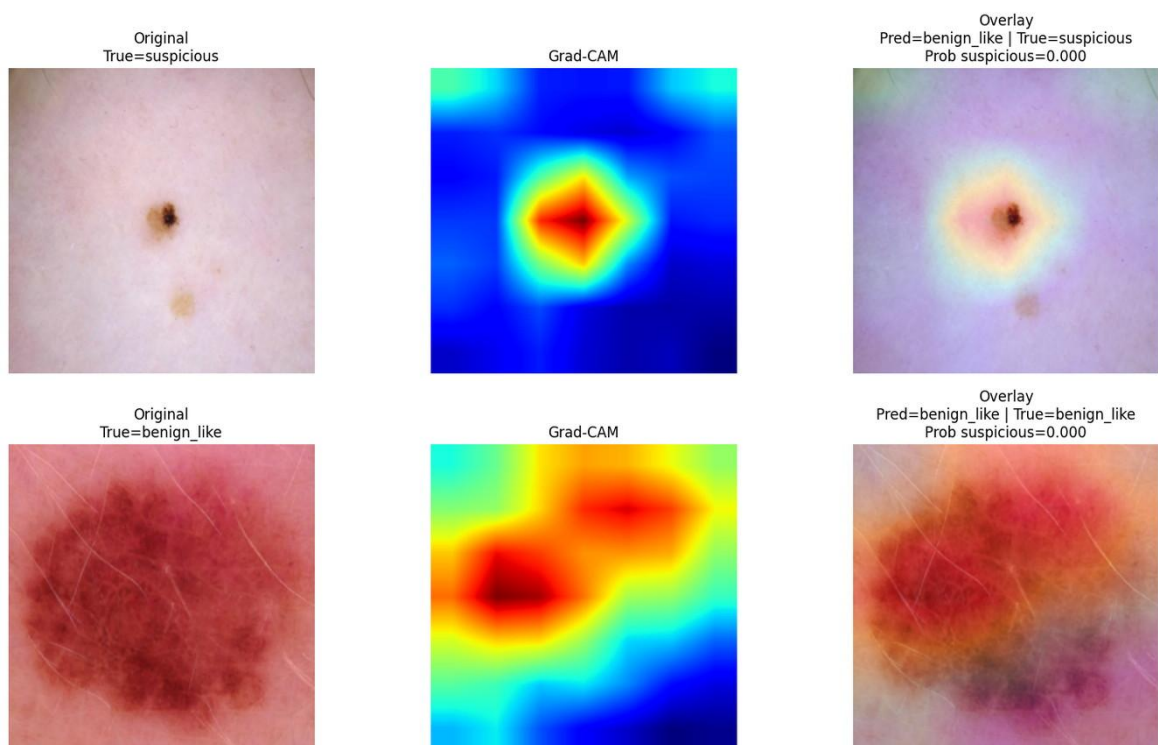


Εικόνα 4.80. Σωστή suspicious πρόβλεψη του Swin-Tiny + SwiGLU με Grad-CAM.

Αντίθετα, στην περίπτωση όπου η πραγματική κλάση είναι suspicious, αλλά το μοντέλο προβλέπει benign-like με πολύ χαμηλή πιθανότητα suspicious, οι χάρτες ερμηνευσιμότητας δεν εστιάζουν με καθαρό τρόπο στην ύποπτη περιοχή. Η ενεργοποίηση εμφανίζεται πιο διάχυτη ή μετατοπισμένη, κάτι που βοηθά στην ερμηνεία του σφάλματος. Με άλλα λόγια, η λανθασμένη πρόβλεψη δεν φαίνεται μόνο στις μετρικές, αλλά και στην οπτική συμπεριφορά του μοντέλου.



Εικόνα 4.81. Λανθασμένη suspicious πρόβλεψη του Swin-Tiny + SwiGLU



Εικόνα 4.82. Λανθασμένη suspicious και benign_like πρόβλεψη του EfficientNet-B4.

Η εικόνα παρουσιάζει περίπτωση όπου το μοντέλο δεν αναγνώρισε σωστά την suspicious κλάση, γεγονός που συνδέεται με τα false negatives της confusion matrix. Η συγκριτική αξιολόγηση των δύο πειραμάτων έδειξε ότι και τα δύο μοντέλα ανταποκρίθηκαν

ικανοποιητικά στο πρόβλημα της δυαδικής ταξινόμησης δερματικών βλαβών στο ISIC 2019. Το EfficientNet-B4 πέτυχε Accuracy 79,11%, F1-score 80,50% και AUC 87,91%, ενώ το Swin-Tiny + SwiGLU παρουσίασε καλύτερη συνολική απόδοση, με Accuracy 81,44%, F1-score 83,01% και AUC 90,69%. Η πιο ουσιαστική διαφορά εντοπίζεται στο Recall της suspicious κλάσης, όπου το Swin-Tiny + SwiGLU έφτασε το 90,67%, ενώ το EfficientNet-B4 το 86,22%. Αυτό δείχνει ότι το Swin-Tiny + SwiGLU αναγνώρισε μεγαλύτερο ποσοστό ύποπτων βλαβών και είχε λιγότερα false negatives, στοιχείο ιδιαίτερα σημαντικό σε ένα πρόβλημα ιατρικής ταξινόμησης, όπου η λανθασμένη ταξινόμηση μιας suspicious βλάβης ως benign-like έχει μεγαλύτερη βαρύτητα. Από τις confusion matrices φάνηκε επίσης ότι τα δύο μοντέλα είχαν σχεδόν παρόμοια συμπεριφορά στην benign-like κλάση, ενώ η κύρια διαφοροποίηση εμφανίστηκε στην ικανότητα αναγνώρισης των suspicious εικόνων.

Η ανάλυση των εικόνων ερμηνευσιμότητας έδειξε ότι η καλύτερη αριθμητική απόδοση δεν σημαίνει απαραίτητα και καθαρότερη οπτική εξήγηση. Το Swin-Tiny + SwiGLU φάνηκε ισχυρότερο ως προς τις μετρικές, αλλά οι χάρτες Grad-CAM, Grad-CAM++ και SHAP δεν ήταν πάντα εύκολοι στην ερμηνεία, καθώς σε ορισμένες περιπτώσεις η ενεργοποίηση εμφανιζόταν διάχυτη ή μετατοπισμένη σε περιοχές εκτός της κύριας βλάβης. Αντίθετα, το EfficientNet-B4, ως CNN μοντέλο, μπορεί να προσφέρει πιο άμεσα αναγνώσιμες οπτικές εξηγήσεις, επειδή οι ενεργοποιήσεις του συνδέονται πιο καθαρά με χωρικά χαρακτηριστικά της εικόνας, όπως η υφή, το περίγραμμα και το κέντρο της βλάβης. Με βάση αυτά, το Swin-Tiny + SwiGLU φαίνεται πιο αποδοτικό στο συγκεκριμένο πειραματικό πλαίσιο, κυρίως λόγω της καλύτερης επίδοσης στη suspicious κλάση, ενώ το EfficientNet-B4 διατηρεί ενδιαφέρον ως πιο ερμηνεύσιμη CNN προσέγγιση.

4.18 Πείραμα δυαδικής ταξινόμησης στο ISIC 2019 με EfficientNet-B3 και Test-Time Augmentation

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο EfficientNet-B3 Noisy Student για δυαδική ταξινόμηση δερματικών βλαβών στο σύνολο δεδομένων ISIC 2019. Όπως και στα προηγούμενα δυαδικά πειράματα, οι αρχικές κατηγορίες του dataset ομαδοποιήθηκαν σε δύο ευρύτερες κλάσεις: benign-like και suspicious. Στην κατηγορία suspicious εντάχθηκαν οι βλάβες που θεωρούνται πιο ύποπτες, όπως AK, BCC, MEL και SCC, ενώ στην κατηγορία benign-like εντάχθηκαν οι NV, BKL, DF και VASC. Το αρχικό σύνολο περιλάμβανε 25.329 εικόνες. Η κατανομή των αρχικών κατηγοριών ήταν έντονα ανισόρροπη, με την κλάση NV να έχει τον μεγαλύτερο αριθμό εικόνων και τις κλάσεις DF και VASC να είναι πολύ μικρότερες. Μετά τη δυαδική ομαδοποίηση, η κλάση benign-like περιλάμβανε 15.990 εικόνες, ενώ η κλάση suspicious περιλάμβανε 9.339 εικόνες. Για να μειωθεί η επίδραση αυτής της ανισορροπίας, δημιουργήθηκε ισορροπημένο υποσύνολο με 2.000 εικόνες ανά κλάση, δηλαδή 4.000 εικόνες συνολικά. Το balanced subset χωρίστηκε σε 2.799 εικόνες εκπαίδευσης, 601 εικόνες επικύρωσης και 600 εικόνες ελέγχου. Στο training set υπήρχαν 1.400 suspicious και 1.399 benign-like εικόνες. Στο validation set υπήρχαν 300 suspicious και 301 benign-like, ενώ στο test set υπήρχαν 300 εικόνες από κάθε κλάση. Με αυτόν τον τρόπο, το test set ήταν πλήρως ισορροπημένο, κάτι που κάνει πιο καθαρή την ερμηνεία των τελικών μετρικών.

Split	Σύνολο	Benign-like	Suspicious
Train	2799	1399	1400
Validation	601	301	300
Test	600	300	300

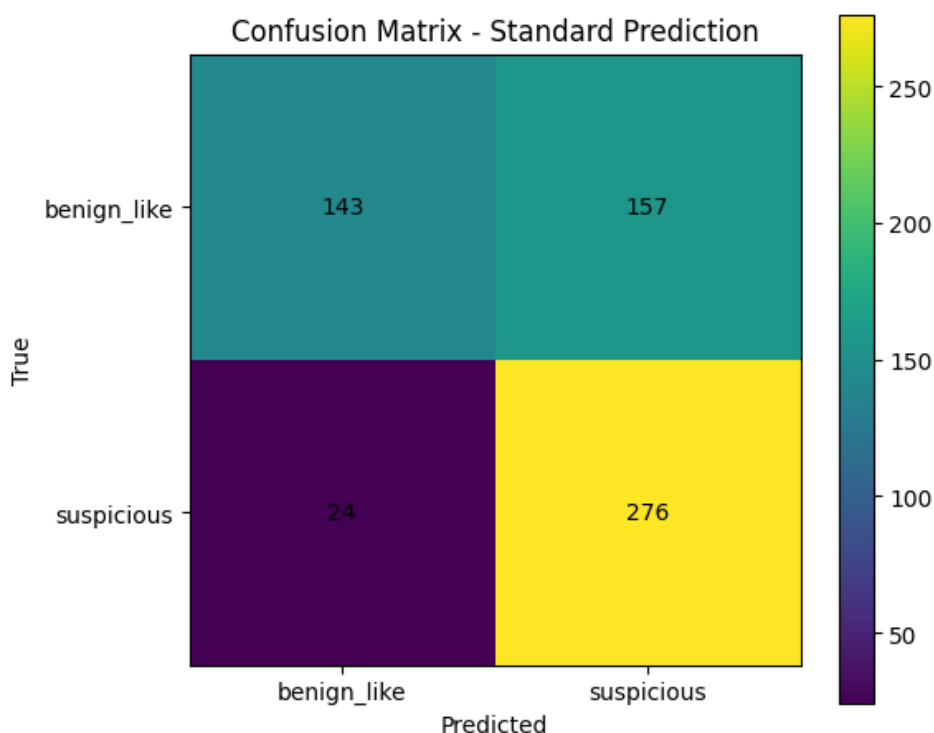
Πίνακας 4.55. Κατανομή του balanced subset στο πείραμα EfficientNet-B3

Το μοντέλο εκπαιδεύτηκε για 6 εποχές με εικόνες μεγέθους 300×300 pixels. Σε κάθε εποχή έγινε επιλογή threshold στο validation set, ώστε η τελική δυαδική απόφαση να μην εξαρτάται αποκλειστικά από το σταθερό όριο 0.50. Το καλύτερο checkpoint επιλέχθηκε στην 5η εποχή, όπου το validation AUC έφτασε στο 0.8308 και το validation F1 στο 0.7661, με threshold 0.12. Στην 6η εποχή το validation accuracy ανέβηκε σε 0.7421, αλλά το AUC μειώθηκε ελαφρά σε 0.8272, οπότε το καλύτερο μοντέλο παρέμεινε αυτό της 5ης εποχής.

Epoch	train_loss	val_acc	val_precision	val_recall	val_f1	val_auc	threshold
1	0.8719	0.7072	0.6685	0.8200	0.7365	0.7654	0.11
2	0.5609	0.7537	0.7171	0.8367	0.7723	0.7971	0.31
3	0.4439	0.7288	0.6789	0.8667	0.7613	0.8080	0.24
4	0.3480	0.7022	0.6384	0.9300	0.7571	0.8071	0.07
5	0.3153	0.7155	0.6497	0.9333	0.7661	0.8308	0.12
6	0.3124	0.7421	0.6954	0.8600	0.7690	0.8272	0.23

Πίνακας 4.56. Πορεία εκπαίδευσης του EfficientNet-B3 στο ISIC 2019

Στο πρώτο στάδιο αξιολόγησης έγινε standard inference, δηλαδή η κάθε εικόνα του test set πέρασε μία φορά από το μοντέλο. Με threshold 0.12, το μοντέλο πέτυχε Accuracy = 69,83%, Precision = 63,74%, Recall = 92,00%, F1 = 75,31% και AUC = 83,47%. Η συμπεριφορά αυτή δείχνει ότι το μοντέλο ήταν αρκετά ευαίσθητο στην κλάση suspicious, καθώς εντόπισε τα περισσότερα suspicious δείγματα. Από την άλλη πλευρά, το χαμηλότερο precision δείχνει ότι αρκετές benign-like εικόνες ταξινομήθηκαν λανθασμένα ως suspicious. Ο πίνακας σύγχυσης στο standard inference ήταν:



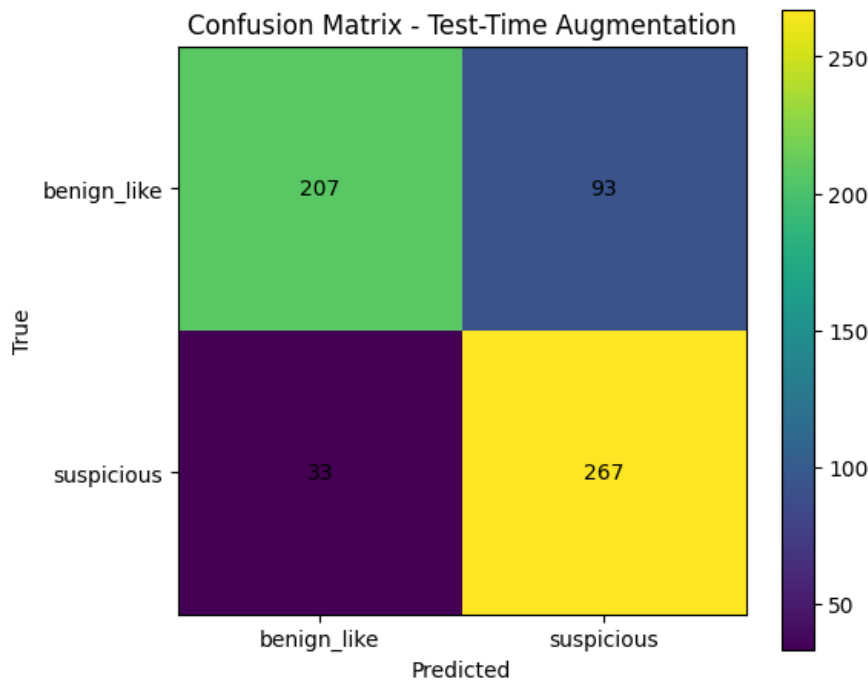
Εικόνα 4.87. Πίνακας σύγχυσης standard prediction

Αυτό σημαίνει ότι από τις 300 benign-like εικόνες, μόνο οι 143 ταξινομήθηκαν σωστά, ενώ 157 ταξινομήθηκαν λανθασμένα ως suspicious. Από τις 300 suspicious εικόνες, οι 276 εντοπίστηκαν σωστά και μόνο 24 ταξινομήθηκαν ως benign-like. Άρα το standard μοντέλο είχε πολύ υψηλό recall για suspicious, αλλά πλήρωσε αυτό το αποτέλεσμα με πολλά false positives.

Στο δεύτερο στάδιο εφαρμόστηκε Test-Time Augmentation (TTA). Με αυτή τη διαδικασία, κάθε test εικόνα δεν αξιολογείται μόνο στην αρχική της μορφή, αλλά και σε διαφορετικές επαυξημένες εκδοχές της. Οι προβλέψεις από αυτές τις εκδοχές συνδυάζονται, ώστε να προκύψει πιο σταθερή τελική πιθανότητα. Η λογική του TTA είναι ότι ένα αξιόπιστο μοντέλο δεν πρέπει να αλλάζει δραματικά πρόβλεψη λόγω μικρών μετασχηματισμών της εικόνας.

Για το TTA έγινε ξεχωριστή επιλογή threshold στο validation set. Το καλύτερο threshold ήταν 0.39, με Validation TTA F1 = 79.63%. Στο test set, η χρήση TTA βελτίωσε καθαρά την απόδοση: Accuracy = 79,00%, Precision = 74,17%, Recall = 89,00%, F1 = 80,91% και AUC = 88,02%.

Ο πίνακας σύγχυσης με TTA ήταν:

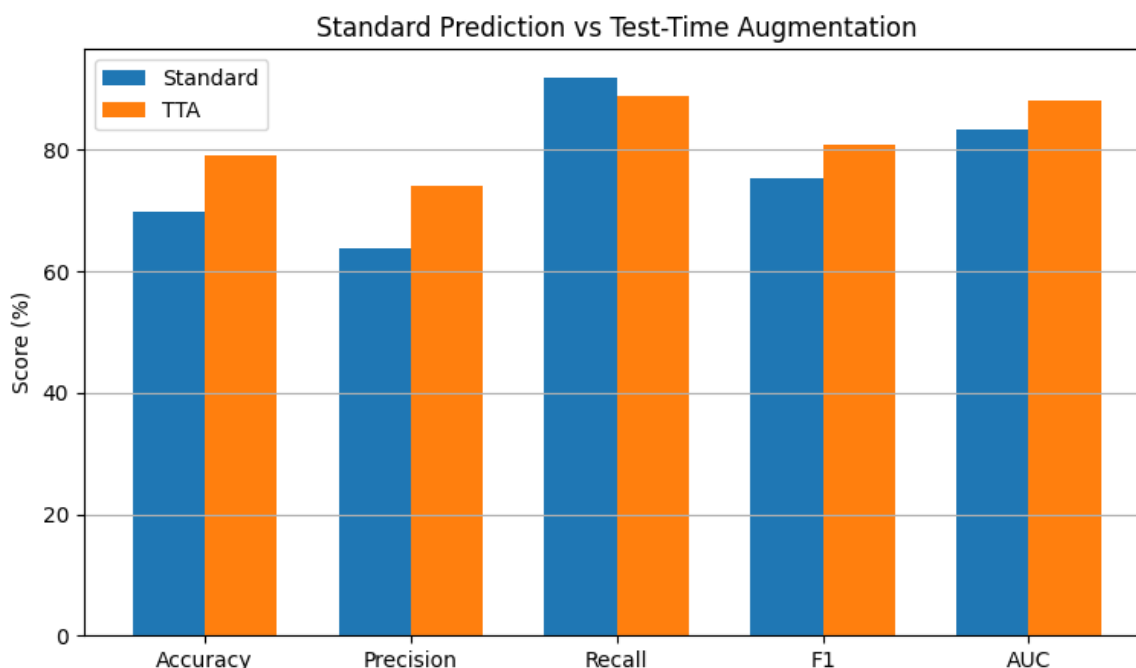


Εικόνα 4.83. Πίνακας σύγχυσης TTA

Με το TTA, από τις 300 benign-like εικόνες, οι 207 ταξινομήθηκαν σωστά και 93 ταξινομήθηκαν λανθασμένα ως suspicious. Από τις 300 suspicious εικόνες, οι 267 εντοπίστηκαν σωστά και 33 χάθηκαν ως benign-like. Σε σχέση με το standard inference, το TTA μείωσε σημαντικά τα false positives, από 157 σε 93, ενώ τα false negatives αυξήθηκαν ελαφρά, από 24 σε 33. Αυτό δείχνει ότι το TTA έκανε το μοντέλο πιο ισορροπημένο και λιγότερο υπερβολικά ευαίσθητο στην suspicious κλάση.

Η σύγκριση standard inference και TTA δείχνει ότι η τεχνική TTA βελτίωσε σχεδόν όλες τις βασικές μετρικές. Η accuracy αυξήθηκε από 69,83% σε 79,00%, το precision από 63,74% σε 74,17%, το F1-score από 75,31% σε 80,91% και το AUC από 83,47% σε 88,02%. Μόνο το recall μειώθηκε ελαφρά, από 92,00% σε 89,00%. Αυτή η μικρή μείωση είναι αποδεκτή, επειδή συνοδεύτηκε από μεγάλη μείωση των false positives και καλύτερη συνολική ισορροπία.

Setup	Threshold	Accuracy	Precision	Recall	F1	AUC
Standard	0.12	0.6983	0.6374	0.9200	0.7531	0.8347
TTA	0.39	0.7900	0.7417	0.8900	0.8091	0.8802



Εικόνα 4.84. Σύγκριση standard inference και Test-Time Augmentation ως προς τις βασικές μετρικές

Το πείραμα αυτό έχει ιδιαίτερη αξία γιατί δείχνει ότι η βελτίωση της απόδοσης δεν εξαρτάται μόνο από την αλλαγή της αρχιτεκτονικής. Εδώ το backbone παρέμεινε το ίδιο, αλλά άλλαξε ο τρόπος με τον οποίο γινόταν η πρόβλεψη στο test set. Με το TTA, το μοντέλο αξιοποίησε πολλαπλές εκδοχές της ίδιας εικόνας, κάτι που οδήγησε σε πιο σταθερές προβλέψεις και καλύτερη γενίκευση.

Συνολικά, το EfficientNet-B3 με TTA πέτυχε καλύτερη απόδοση από το standard inference, με $F1 = 80,91\%$ και $AUC = 88,02\%$. Το βασικό του πλεονέκτημα ήταν ότι μείωσε αρκετά τα false positives, χωρίς να χάσει υπερβολικά πολλά suspicious δείγματα. Παρ' όλα αυτά, τα 33 false negatives παραμένουν σημαντικά, επειδή αντιστοιχούν σε suspicious περιπτώσεις που ταξινομήθηκαν ως benign-like. Σε ένα ιατρικό πλαίσιο, τέτοιου είδους λάθη χρειάζονται ιδιαίτερη προσοχή.

4.19 Σύγκριση CNN και Vision Transformer μοντέλων στο ISIC 2019

Στο πείραμα αυτό πραγματοποιήθηκε συγκριτική αξιολόγηση διαφορετικών αρχιτεκτονικών για τη δυαδική ταξινόμηση δερματικών βλαβών στο σύνολο δεδομένων ISIC 2019. Όπως και στα προηγούμενα δυαδικά πειράματα, οι αρχικές κατηγορίες ομαδοποιήθηκαν σε δύο ευρύτερες ομάδες: benign-like και suspicious. Στην ομάδα suspicious εντάχθηκαν οι κατηγορίες AK, BCC, MEL και SCC, ενώ στην ομάδα benign-like εντάχθηκαν οι NV, BKL, DF και VASC. Η λογική του πειράματος ήταν να εξεταστεί αν οι αρχιτεκτονικές τύπου Vision Transformer μπορούν να αποδώσουν καλύτερα από κλασικά ή νεότερα CNN backbones στο ίδιο ακριβώς πρόβλημα. Το αρχικό σύνολο

περιλάμβανε 25.329 εικόνες. Η αρχική κατανομή ήταν έντονα ανισόρροπη, με την κλάση NV να έχει τον μεγαλύτερο αριθμό εικόνων, ενώ κλάσεις όπως DF και VASC είχαν πολύ μικρότερη παρουσία. Μετά τη δυαδική ομαδοποίηση, η κλάση benign-like περιλάμβανε 15.990 εικόνες, ενώ η κλάση suspicious περιλάμβανε 9.339 εικόνες. Για να υπάρχει πιο καθαρή σύγκριση ανάμεσα στα μοντέλα, δημιουργήθηκε ισορροπημένο υποσύνολο με 1.500 εικόνες ανά κλάση, δηλαδή 3.000 εικόνες συνολικά. Το balanced subset χωρίστηκε σε 2.099 εικόνες εκπαίδευσης, 451 εικόνες επικύρωσης και 450 εικόνες ελέγχου. Στο training set υπήρχαν 1.050 suspicious και 1.049 benign-like εικόνες. Στο validation set υπήρχαν 225 suspicious και 226 benign-like, ενώ στο test set υπήρχαν 225 εικόνες από κάθε κλάση. Με αυτόν τον τρόπο, το test set ήταν πλήρως ισορροπημένο, άρα η accuracy, το F1 και το AUC μπορούν να διαβαστούν πιο άμεσα.

Split	Σύνολο	Benign-like	Suspicious
Train	2099	1049	1050
Validation	451	226	225
Test	450	225	225

Πίνακας 4.57. Κατανομή των δεδομένων στα train, validation και test σύνολα

Συγκρίθηκαν έξι μοντέλα. Από την πλευρά των CNN χρησιμοποιήθηκαν τα EfficientNet-B0 και ReXNet-100, ενώ το ConvNeXt-Tiny τοποθετήθηκε ως νεότερη CNN-like αρχιτεκτονική. Από την πλευρά των Vision Transformers χρησιμοποιήθηκαν τα Swin-Tiny, DeiT-Small και ViT-Tiny. Όλα τα μοντέλα εκπαιδεύτηκαν για 5 εποχές, με κοινό πρωτόκολλο αξιολόγησης και επιλογή threshold στο validation set. Η επιλογή threshold ανά μοντέλο ήταν σημαντική, γιατί τα μοντέλα δεν παράγουν απαραίτητα πιθανότητες με την ίδια κατανομή.

Model	Family
EfficientNet-B0	CNN
ReXNet-100	CNN
ConvNeXt-Tiny	CNN-like
Swin-Tiny	Vision Transformer
DeiT-Small	Vision Transformer
ViT-Tiny	Vision Transformer

Πίνακας 4.58. Μοντέλα που συγκρίθηκαν στο πείραμα CNN vs Vision Transformer

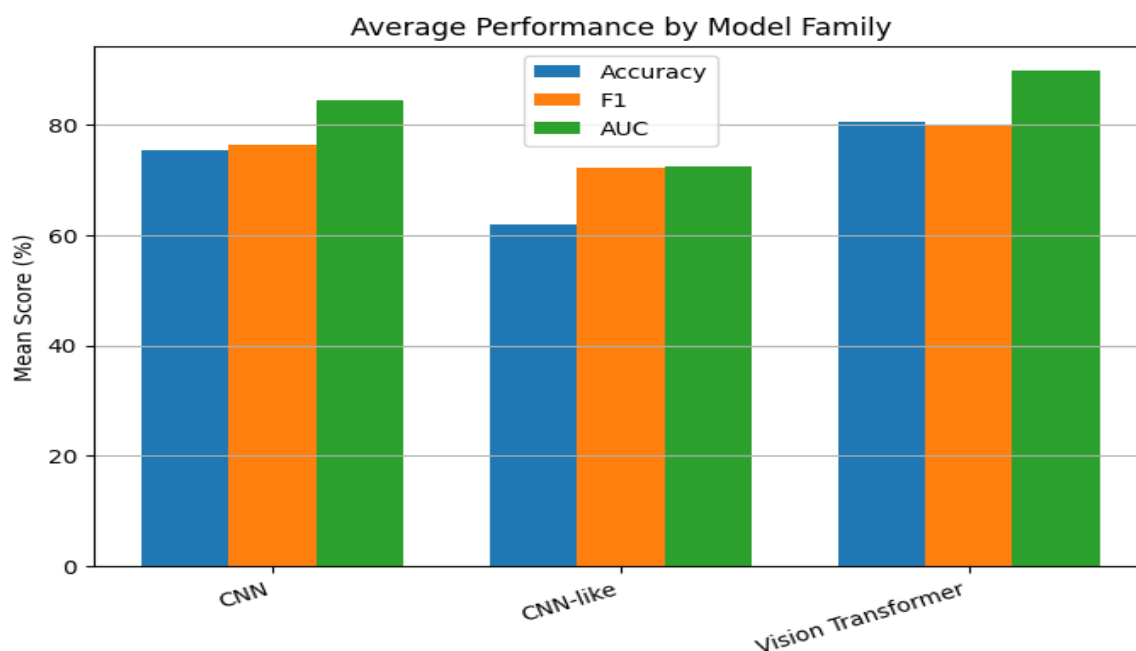
Στα τελικά αποτελέσματα, τα Vision Transformer μοντέλα είχαν συνολικά καλύτερη επίδοση από τα CNN μοντέλα. Το καλύτερο μοντέλο ως προς AUC ήταν το ViT-Tiny, με Test AUC = 90.74%, ενώ το καλύτερο ως προς F1-score και accuracy ήταν το Swin-Tiny, με Accuracy = 81.33% και F1 = 81.82%. Το DeiT-Small κινήθηκε επίσης σε υψηλό επίπεδο, με Accuracy = 79.78%, F1 = 79.73% και AUC = 88.81%. Από τα CNN, το ReXNet-100 είχε την καλύτερη συνολική ισορροπία, με Accuracy = 76.67%, F1 = 76.82% και AUC = 83.90%. Το EfficientNet-B0 είχε Accuracy = 74.22%, F1 = 76.13% και AUC = 85.13%. Το

ConvNeXt-Tiny είχε ιδιαίτερα υψηλό recall, 98.22%, αλλά πολύ χαμηλότερη precision, 56.96%, κάτι που δείχνει ότι ταξινομήσε υπερβολικά πολλές εικόνες ως suspicious.

Model	Family	Best epoch	Threshold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
EfficientNet-B0	CNN	3	0.27	74.22%	70.88%	82.22%	76.13%	85.13%
ReXNet-100	CNN	5	0.32	76.67%	76.32%	77.33%	76.82%	83.90%
ConvNeXt-Tiny	CNN-like	5	0.34	62.00%	56.96%	98.22%	72.10%	72.43%
Swin-Tiny	Vision Transformer	5	0.43	81.33%	79.75%	84.00%	81.82%	89.79%
DeiT-Small	Vision Transformer	5	0.46	79.78%	79.91%	79.56%	79.73%	88.81%
ViT-Tiny	Vision Transformer	5	0.55	80.22%	84.34%	74.22%	78.96%	90.74%

Πίνακας 4.59. Τελικά test αποτελέσματα των CNN και Vision Transformer μοντέλων

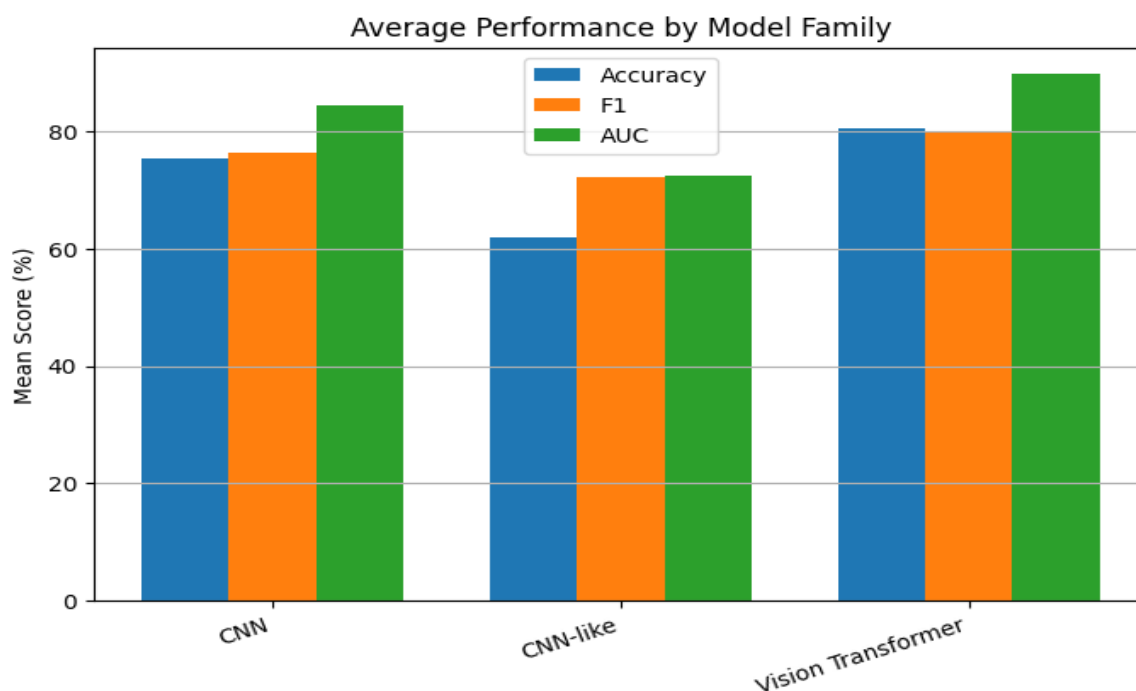
Η κατάταξη με βάση το AUC έφερε πρώτο το ViT-Tiny, δεύτερο το Swin-Tiny και τρίτο το DeiT-Small. Αυτό είναι σημαντικό, γιατί δείχνει ότι τα transformer-based μοντέλα είχαν καλύτερη συνολική διαχωριστική ικανότητα στο συγκεκριμένο δυαδικό πρόβλημα. Παρ' όλα αυτά, αν δοθεί μεγαλύτερη έμφαση στο F1-score, το Swin-Tiny είναι η πιο ισορροπημένη επιλογή, επειδή συνδυάζει υψηλό recall και αρκετά καλό precision.



Εικόνα 4.85. Σύγκριση CNN και Vision Transformer μοντέλων ως προς Accuracy, F1 και AUC

Η ανάλυση ανά οικογένεια επιβεβαιώνει την ίδια εικόνα. Τα δύο κλασικά CNN μοντέλα είχαν μέση Accuracy = 0.7544, μέσο F1 = 0.7648 και μέσο AUC = 0.8452. Τα Vision Transformer μοντέλα είχαν υψηλότερη μέση Accuracy = 0.8044, μέσο F1 = 0.8017 και μέσο AUC = 0.8978. Άρα, στο συγκεκριμένο πείραμα, οι transformer αρχιτεκτονικές είχαν πιο σταθερή και υψηλότερη απόδοση συνολικά.

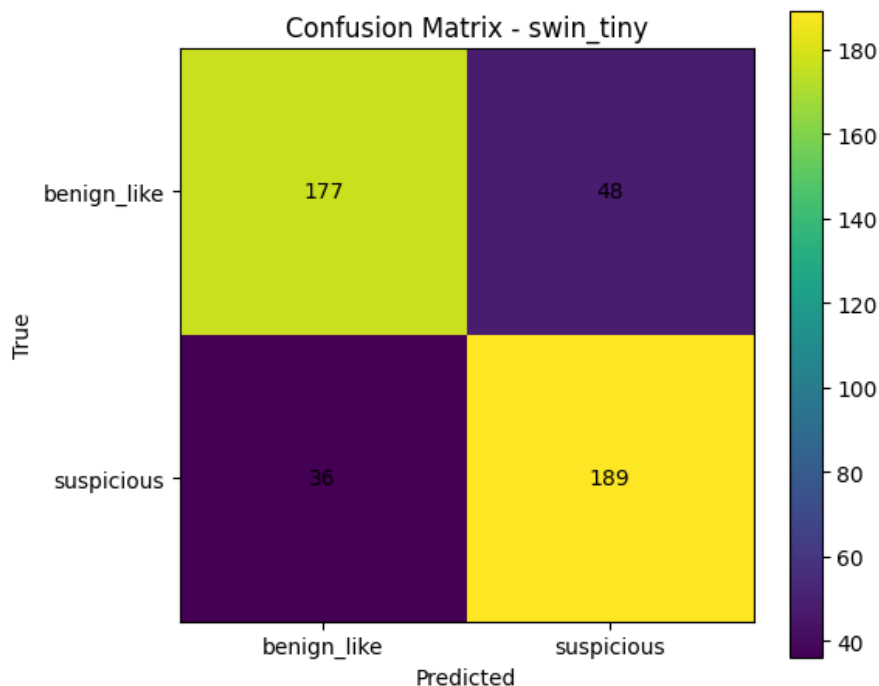
Family	Mean Accuracy (%)	Mean Precision (%)	Mean Recall (%)	Mean F1 (%)	Mean AUC (%)
CNN	75.44%	73.60%	79.78%	76.48%	84.52%
CNN-like	62.00%	56.96%	98.22%	72.10%	72.43%
Vision Transformer	80.44%	81.33%	79.26%	80.17%	89.78%



Εικόνα 4.86. Μέση απόδοση ανά οικογένεια μοντέλων

Αξίζει όμως να σχολιαστεί ξεχωριστά το ConvNeXt-Tiny. Παρότι θεωρείται σύγχρονη CNN-like αρχιτεκτονική, στο συγκεκριμένο πείραμα εμφάνισε πολύ ασταθή συμπεριφορά. Το recall ήταν πολύ υψηλό, 98,22%, δηλαδή το μοντέλο εντόπισε σχεδόν όλες τις suspicious εικόνες. Αυτό όμως έγινε με πολύ χαμηλή precision, 56,96%, καθώς ταξινομήσε μεγάλο αριθμό benign-like εικόνων ως suspicious. Ο πίνακας σύγχυσης το δείχνει καθαρά: από τις 225 benign-like εικόνες, μόνο 58 ταξινομήθηκαν σωστά, ενώ 167 προβλέφθηκαν λανθασμένα ως suspicious. Άρα η υψηλή ευαισθησία του ConvNeXt-Tiny είχε μεγάλο κόστος σε false positives.

Συγκριτικά, το Swin-Tiny είχε πιο ισορροπημένη συμπεριφορά. Στον πίνακα σύγχυσης είχε 177 σωστές benign-like προβλέψεις και 189 σωστές suspicious προβλέψεις, με 48 false positives και 36 false negatives. Το ViT-Tiny, από την άλλη, είχε το υψηλότερο AUC, αλλά χαμηλότερο recall για suspicious, καθώς εντόπισε 167 από τις 225 suspicious εικόνες. Αυτό δείχνει ότι το ViT-Tiny είχε ισχυρή διαχωριστική ικανότητα συνολικά, αλλά με το επιλεγμένο threshold έγινε πιο αυστηρό στην απόδοση της suspicious κλάσης.



Εικόνα 4.87. Πίνακας σύγχυσης του καλύτερου Vision Transformer μοντέλου

Το βασικό συμπέρασμα από το πείραμα είναι ότι οι Vision Transformer αρχιτεκτονικές υπερέχουν συνολικά των CNN backbones στο συγκεκριμένο balanced subset του ISIC 2019. Η υπεροχή αυτή δεν είναι τεράστια, αλλά είναι σταθερή σε όλες τις βασικές μετρικές. Το Swin-Tiny ήταν το πιο ισορροπημένο μοντέλο, ενώ το ViT-Tiny είχε την καλύτερη συνολική διαχωριστική ικανότητα με βάση το AUC. Τα CNN μοντέλα είχαν αξιοπρεπή συμπεριφορά, αλλά δεν έφτασαν την επίδοση των transformer-based μοντέλων.

Παράλληλα, το πείραμα δείχνει ότι η επιλογή μοντέλου δεν πρέπει να βασίζεται σε μία μόνο μετρική. Αν ο στόχος είναι το υψηλότερο AUC, τότε το ViT-Tiny είναι πρώτο. Αν ο στόχος είναι η καλύτερη ισορροπία ανάμεσα σε precision και recall, τότε το Swin-Tiny είναι πιο κατάλληλο. Αν ο στόχος είναι να χαθούν όσο το δυνατόν λιγότερες suspicious εικόνες, τότε το ConvNeXt-Tiny έχει πολύ υψηλό recall, αλλά με τόσο πολλά false positives που δύσκολα θα ήταν πρακτικά αποδεκτό χωρίς περαιτέρω ρύθμιση.

4.20 Πείραμα δυαδικής ταξινόμησης στο ISIC 2019 με EfficientNet-B0, ανάλυση feature space και εξηγησιμότητα

Στο πείραμα αυτό αξιολογήθηκε ένα μοντέλο EfficientNet-B0 για τη δυαδική ταξινόμηση δερματικών βλαβών στο σύνολο δεδομένων ISIC 2019. Οι αρχικές διαγνωστικές κατηγορίες του dataset ομαδοποιήθηκαν σε δύο ευρύτερες κλάσεις: benign-like και suspicious. Στην κλάση suspicious εντάχθηκαν οι κατηγορίες AK, BCC, MEL και SCC, ενώ στην κλάση benign-like εντάχθηκαν οι NV, BKL, DF και VASC. Με αυτή τη διαμόρφωση, το πρόβλημα δεν αντιμετωπίστηκε ως πλήρης πολυταξική διάγνωση, αλλά ως ένα σενάριο αρχικού ελέγχου, όπου το μοντέλο καλείται να διαχωρίσει πιθανώς ύποπτες βλάβες από πιο καλοήθειες ή λιγότερο ανησυχητικές περιπτώσεις.

Το αρχικό σύνολο ISIC 2019 περιλάμβανε 25.331 εικόνες. Μετά τη δυαδική ομαδοποίηση, η κλάση benign-like περιλάμβανε 15.991 εικόνες, ενώ η κλάση suspicious περιλάμβανε 9.340 εικόνες. Η αρχική αυτή κατανομή ήταν ανισόρροπη, επομένως δημιουργήθηκε ένα ισορροπημένο υποσύνολο με 3.000 εικόνες ανά κλάση, δηλαδή 6.000 εικόνες συνολικά. Με αυτόν τον τρόπο περιορίστηκε η επίδραση της αριθμητικής υπεροχής της benign-like κλάσης και η αξιολόγηση έγινε σε πιο ελεγχόμενο πλαίσιο.

Το ισορροπημένο υποσύνολο χωρίστηκε σε 4.199 εικόνες εκπαίδευσης, 901 εικόνες επικύρωσης και 900 εικόνες ελέγχου. Στο training set υπήρχαν 2.099 benign-like και 2.100 suspicious εικόνες. Στο validation set υπήρχαν 451 benign-like και 450 suspicious, ενώ στο test set υπήρχαν 450 εικόνες από κάθε κλάση. Το test set, δηλαδή, ήταν πλήρως ισορροπημένο, κάτι που βοηθά στην πιο άμεση ερμηνεία της accuracy, του F1-score και του AUC.

Split	Σύνολο	Benign-like	Suspicious
Train	4199	2099	2100
Validation	901	451	450
Test	900	450	450

Πίνακας 4.60. Κατανομή των δεδομένων στα train, validation και test σύνολα

Η εκπαίδευση του EfficientNet-B0 πραγματοποιήθηκε για 12 εποχές. Σε κάθε εποχή υπολογίστηκαν οι βασικές validation μετρικές και έγινε επιλογή threshold, ώστε η τελική δυαδική απόφαση να μην εξαρτάται υποχρεωτικά από το τυπικό όριο 0.50. Αυτή η διαδικασία έχει ιδιαίτερη σημασία σε ιατρικά προβλήματα, επειδή η μεταβολή του threshold επηρεάζει άμεσα τη σχέση ανάμεσα σε precision και recall.

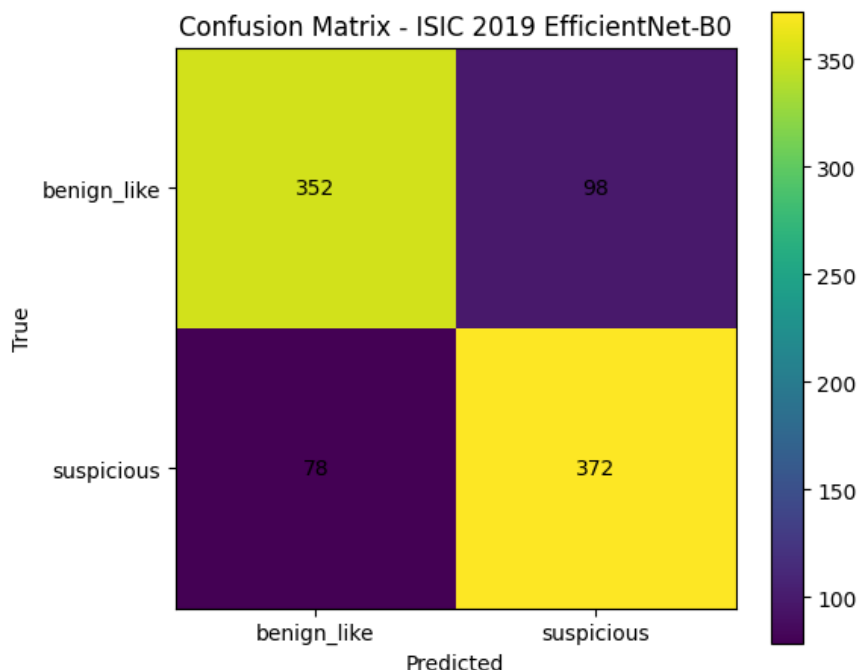
Στην 1η εποχή, το μοντέλο πέτυχε $\text{val_acc} = 71,14\%$, $\text{val_precision} = 67,72\%$, $\text{val_recall} = 80,67\%$, $\text{val_f1} = 73,63\%$ και $\text{val_auc} = 79,55\%$. Η απόδοση βελτιώθηκε σταδιακά στις επόμενες εποχές. Στην 5η εποχή, το validation F1 έφτασε στο 78,75%, ενώ στην 8η εποχή καταγράφηκε η καλύτερη συνολική επίδοση με $\text{val_acc} = 80,80\%$, $\text{val_precision} = 79,04\%$, $\text{val_recall} = 83,78\%$, $\text{val_f1} = 81,34\%$ και $\text{val_auc} = 88,66\%$. Για τον λόγο αυτό, η 8η εποχή επιλέχθηκε ως το καλύτερο checkpoint.

Epoch	Train loss	Val accuracy (%)	Val precision (%)	Val recall (%)	Val F1 (%)	Val AUC (%)	Threshold
1	1.3013	71.14%	67.72%	80.67%	73.63%	79.55%	0.09
2	0.6758	75.14%	72.16%	81.78%	76.67%	83.95%	0.28
5	0.2928	75.92%	70.40%	89.33%	78.75%	86.00%	0.11
7	0.2134	79.02%	75.14%	86.67%	80.50%	87.55%	0.25
8	0.1756	80.80%	79.04%	83.78%	81.34%	88.66%	0.47
10	0.1430	77.91%	72.69%	89.33%	80.16%	87.95%	0.13
12	0.1216	79.69%	76.86%	84.89%	80.68%	88.15%	0.28

Πίνακας 4.61. Πορεία εκπαίδευσης του EfficientNet-B0 στο ISIC 2019

Στο test set, το καλύτερο checkpoint της 8ης εποχής αξιολογήθηκε με threshold περίπου 0.47. Το μοντέλο πέτυχε Test Accuracy = 80,44%, Test Precision = 79,15%, Test Recall = 82,67%, Test F1 = 80,87% και Test AUC = 88,56%. Οι τιμές αυτές δείχνουν μια σχετικά ισορροπημένη συμπεριφορά ανάμεσα στις δύο κλάσεις. Το recall είναι ελαφρώς υψηλότερο από το precision, κάτι που σημαίνει ότι το μοντέλο εντοπίζει αρκετές suspicious περιπτώσεις, χωρίς όμως να μετακινείται υπερβολικά προς την ύποπτη κλάση.

Ο πίνακας σύγχυσης ήταν:



Εικόνα 4.88. Πίνακας σύγχυσης του EfficientNet-B0 στο test set του ISIC 2019

Αυτό σημαίνει ότι από τις 450 benign-like εικόνες, οι 352 ταξινομήθηκαν σωστά και οι 98 ταξινομήθηκαν λανθασμένα ως suspicious. Από τις 450 suspicious εικόνες, οι 372 εντοπίστηκαν σωστά, ενώ οι 78 ταξινομήθηκαν λανθασμένα ως benign-like. Τα 78 false negatives είναι το πιο κρίσιμο σημείο, επειδή αντιστοιχούν σε ύποπτες βλάβες που το μοντέλο δεν αναγνώρισε. Από την άλλη, τα 98 false positives δείχνουν ότι το μοντέλο εξακολουθεί να χαρακτηρίζει αρκετές benign-like εικόνες ως suspicious.

Metric	Value
Best epoch	8
Test Accuracy	0.8044
Test Precision	0.7915
Test Recall	0.8267
Test F1	0.8087
Test AUC	0.8856

Πίνακας 4.62. Test αποτελέσματα του EfficientNet-B0 στο ISIC 2019

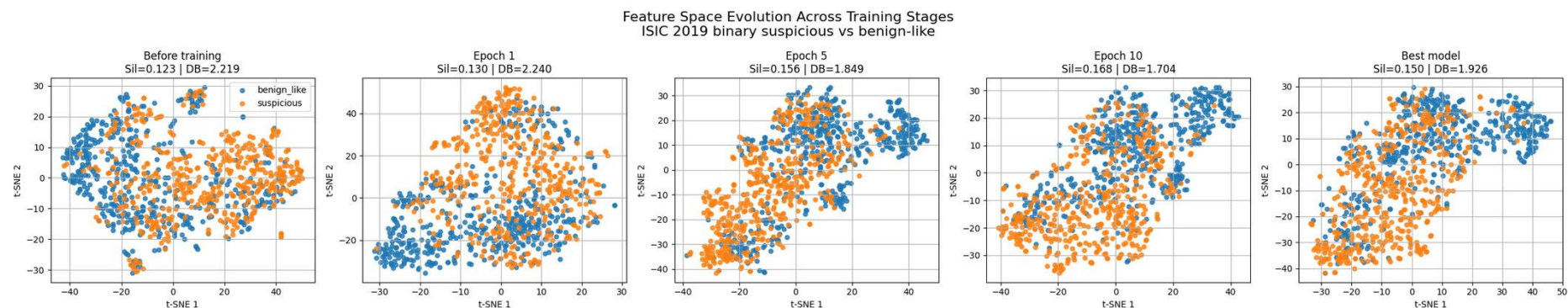
Ένα σημαντικό μέρος του πειράματος ήταν η ανάλυση του feature space του μοντέλου. Η ανάλυση αυτή δεν εξετάζει μεμονωμένα pixels ή περιοχές μιας εικόνας, αλλά το πώς οργανώνονται εσωτερικά οι αναπαραστάσεις που μαθαίνει το EfficientNet-B0. Για τον σκοπό αυτό εξήχθησαν χαρακτηριστικά από διαφορετικά στάδια εκπαίδευσης: before training, epoch 1, epoch 5, epoch 10 και best model. Στη συνέχεια εφαρμόστηκε t-SNE σε ισορροπημένο υποσύνολο 900 εικόνων, με 450 benign-like και 450 suspicious δείγματα. Η ανάλυση συνοδεύτηκε από ποσοτικές μετρικές που περιγράφουν τη χωριστικότητα και τη συνοχή των δύο ομάδων. Συγκεκριμένα, χρησιμοποιήθηκαν οι Davies–Bouldin index, Calinski–Harabasz score, Silhouette score, centroid distance, mean intra-class spread και separation ratio.

Πριν από την εκπαίδευση, το feature space ήταν σχετικά αδιαμόρφωτο. Ο Davies–Bouldin index ήταν 2.219, ο Calinski–Harabasz score ήταν 150.79, ο Silhouette score ήταν 0.123, ενώ το separation ratio ήταν 0.901. Η τιμή του separation ratio κάτω από τη μονάδα δείχνει ότι η απόσταση ανάμεσα στα κέντρα των δύο κλάσεων ήταν μικρότερη από τη μέση ενδοκλασική διασπορά. Με απλά λόγια, οι δύο κλάσεις δεν είχαν ακόμη καθαρή γεωμετρική οργάνωση στον χώρο των χαρακτηριστικών. Στην 1η εποχή, η βελτίωση ήταν περιορισμένη. Ο Davies–Bouldin αυξήθηκε ελαφρά σε 2.240, ενώ ο Calinski–Harabasz μειώθηκε σε 145.11. Παρότι ο Silhouette score ανέβηκε λίγο στο 0.130, η συνολική εικόνα δείχνει ότι το μοντέλο βρισκόταν ακόμη σε πρώιμο στάδιο προσαρμογής. Στην 5η εποχή, η οργάνωση των χαρακτηριστικών άρχισε να βελτιώνεται πιο καθαρά. Ο Davies–Bouldin μειώθηκε σε 1.849, ο Calinski–Harabasz ανέβηκε σε 207.20, ενώ το separation ratio έφτασε στο 1.082. Αυτό σημαίνει ότι τα κέντρα των δύο κλάσεων άρχισαν να απομακρύνονται περισσότερο σε σχέση με τη διασπορά των δειγμάτων μέσα στην κάθε κλάση. Η καλύτερη εικόνα ως προς τη γεωμετρική οργάνωση του feature space εμφανίστηκε στην 10η εποχή. Εκεί ο Davies–Bouldin μειώθηκε σε 1.704, ο Calinski–Harabasz ανέβηκε σε 250.31, ο Silhouette score έφτασε στο 0.168 και το separation ratio ανέβηκε σε 1.173. Αυτό δείχνει ότι, παρότι το καλύτερο μοντέλο με βάση το validation AUC ήταν αυτό της 8ης εποχής, η

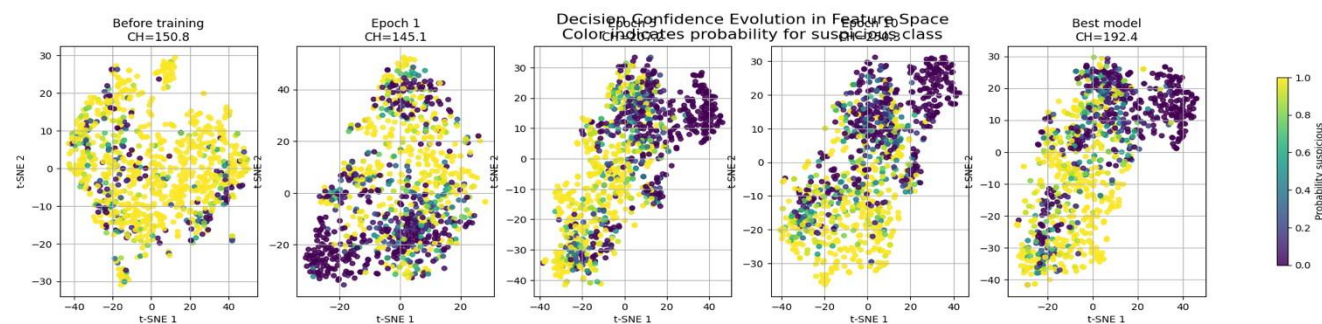
χωρική οργάνωση των χαρακτηριστικών συνέχισε να βελτιώνεται και μετά. Στο best model, δηλαδή στο checkpoint της 8ης εποχής, οι μετρικές παρέμειναν βελτιωμένες σε σχέση με την αρχική κατάσταση, αλλά δεν ήταν οι καλύτερες συνολικά. Ο Davies–Bouldin ήταν 1.926, ο Calinski–Harabasz 192.42, ο Silhouette score 0.150 και το separation ratio 1.038. Αυτό δείχνει ότι η καλύτερη εποχή για ταξινόμηση δεν ταυτίζεται απαραίτητα με την καλύτερη εποχή ως προς τη γεωμετρική δομή του feature space.

Stage	Epoch	Davies– Bouldin ↓	Calinski– Harabasz ↑	Silhouette ↑	Centroid distance ↑	Mean intra- class spread ↓	Separation ratio ↑
Before training	0	2.219	150.79	0.123	20.93	23.23	0.901
Epoch 1	1	2.240	145.11	0.130	19.78	22.15	0.893
Epoch 5	5	1.849	207.20	0.156	23.82	22.03	1.082
Epoch 10	10	1.704	250.31	0.168	24.33	20.73	1.173
Best model	8	1.926	192.42	0.150	22.69	21.85	1.038

Πίνακας 4.63. Σύγκριση του feature space σε διαφορετικά στάδια της εκπαίδευσης



Εικόνα 4.89. Εξέλιξη του feature space του EfficientNet-B0 κατά τη διάρκεια της εκπαίδευσης

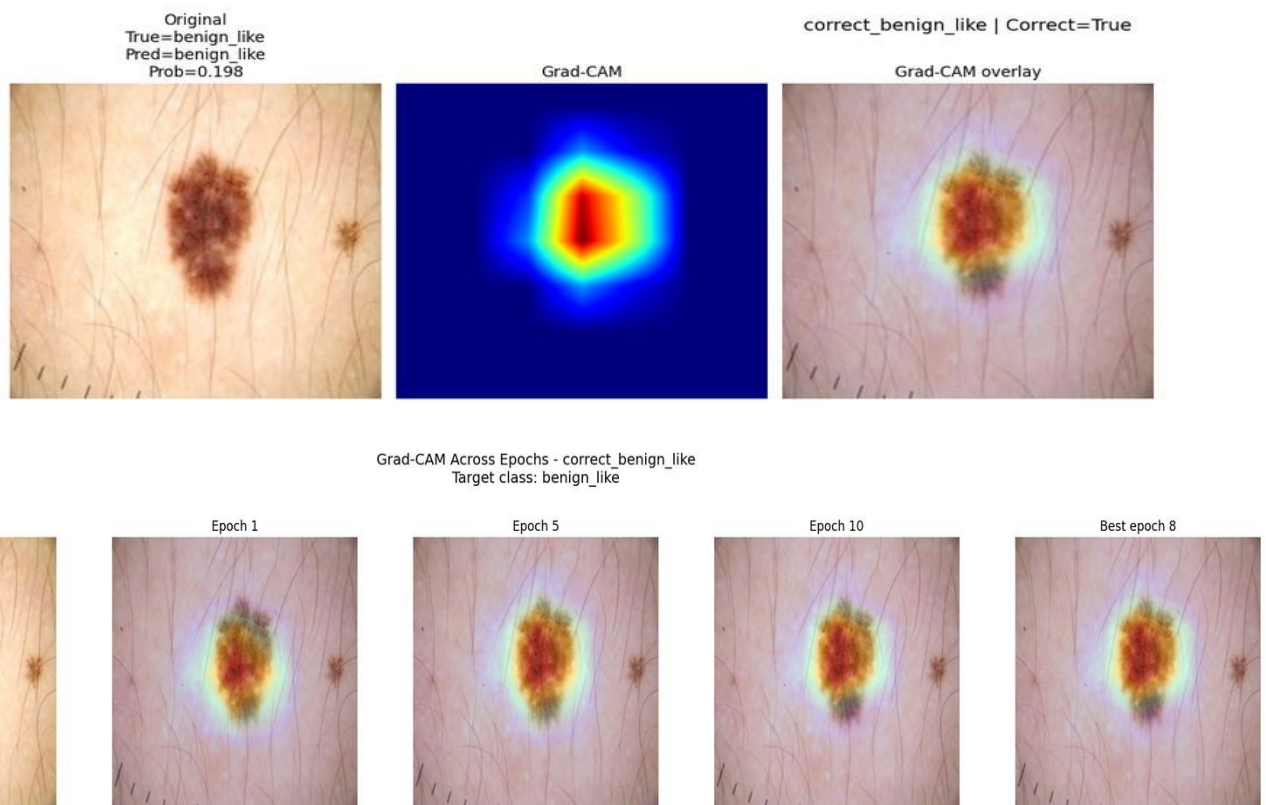


Εικόνα 4.90. Εξέλιξη της βεβαιότητας του μοντέλου στο feature space. Ο χρωματισμός αντιστοιχεί στην πιθανότητα που αποδίδει το EfficientNet-B0 στην κλάση suspicious.

Η ανάλυση αυτή δείχνει ότι η εκπαίδευση δεν βελτίωσε μόνο τις τελικές μετρικές ταξινόμησης, αλλά και την εσωτερική οργάνωση των χαρακτηριστικών. Οι δύο κλάσεις άρχισαν σταδιακά να σχηματίζουν πιο διακριτές ομάδες, παρότι εξακολουθεί να υπάρχει επικάλυψη. Αυτό είναι αναμενόμενο, επειδή οπτικά αρκετές benign-like και suspicious βλάβες μπορεί να μοιάζουν μεταξύ τους. Στην πράξη, το feature space δεν γίνεται απόλυτα καθαρό, αλλά γίνεται σαφώς πιο οργανωμένο από την αρχική κατάσταση.

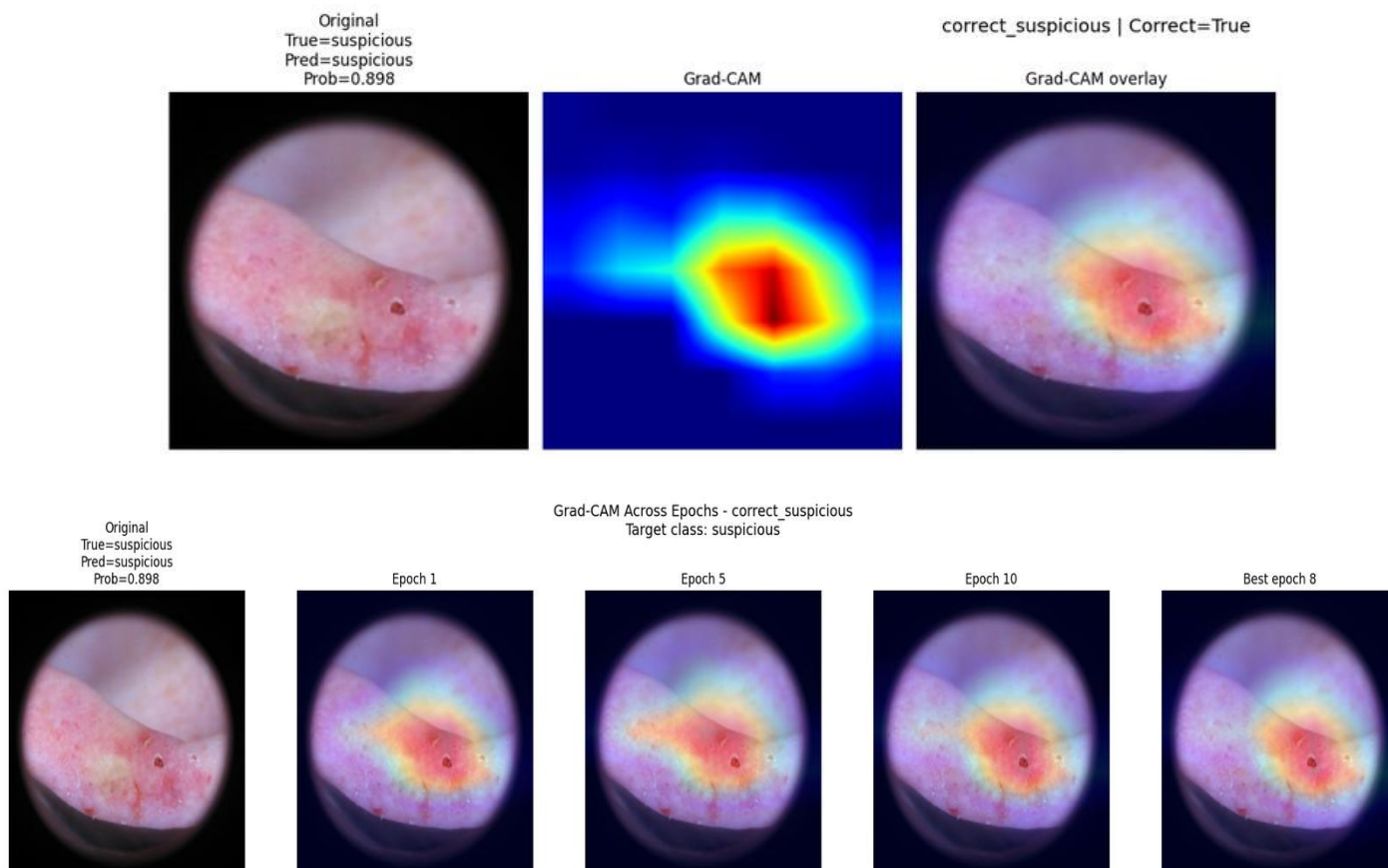
Πέρα από την ανάλυση του feature space, εφαρμόστηκαν και τοπικές τεχνικές εξηγησιμότητας, ώστε να εξεταστεί σε ποιες περιοχές της εικόνας βασίζεται το μοντέλο για να πάρει την τελική απόφαση. Συγκεκριμένα, χρησιμοποιήθηκαν Grad-CAM, Grad-CAM++, SHAP και Integrated Gradients / SmoothGrad. Από τις διαθέσιμες εικόνες, επιλέχθηκαν μόνο οι πιο αντιπροσωπευτικές, ώστε η ενότητα να μη φορτωθεί με χάρτες που δεν προσθέτουν ουσιαστική πληροφορία.

Σε αυτό το παράδειγμα, η πραγματική κλάση είναι benign-like, η πρόβλεψη του μοντέλου είναι επίσης benign-like, και η πιθανότητα για suspicious είναι 0.198. Στις πρώτες εποχές, η ενεργοποίηση είναι πιο γενική και λιγότερο σταθερή. Στις επόμενες εποχές, και ιδιαίτερα στο best checkpoint, η Grad-CAM εστίαση συγκεντρώνεται περισσότερο στην περιοχή της βλάβης. Αυτό δείχνει ότι το μοντέλο μαθαίνει σταδιακά να βασίζει την απόφασή του σε πιο σχετική οπτική πληροφορία.



Εικόνα 4.91. Grad-CAM overlay και Grad-CAM across-epoch ανάλυση για σωστή benign-like πρόβλεψη

Η δεύτερη χρήσιμη εικόνα είναι η Grad-CAM across-epoch ανάλυση για σωστή suspicious πρόβλεψη. Στο παράδειγμα αυτό, η πραγματική κλάση είναι suspicious, η πρόβλεψη είναι επίσης suspicious, και η πιθανότητα suspicious είναι 0.898. Η ενεργοποίηση καλύπτει τμήμα της παθολογικής περιοχής, αν και δεν είναι απόλυτα τέλεια. Παρ' όλα αυτά, είναι χρήσιμη εικόνα, γιατί δείχνει ότι το μοντέλο μπορεί να αναγνωρίζει οπτικά στοιχεία που συνδέονται με την suspicious κλάση. Σε σχέση με τη σωστή benign-like περίπτωση, η εστίαση είναι λιγότερο καθαρή, κάτι που δείχνει και τη δυσκολία της suspicious κατηγορίας.



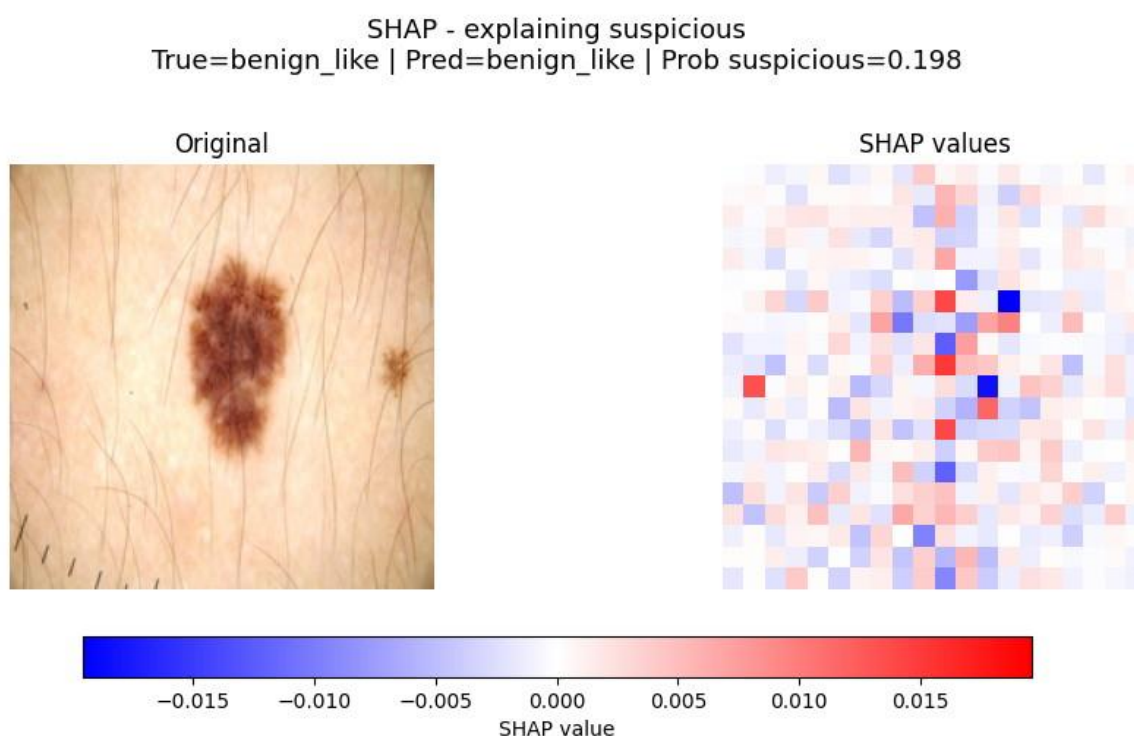
Εικόνα 4.92. Grad-CAM overlay και Grad-CAM across-epoch ανάλυση για σωστή suspicious πρόβλεψη

Ιδιαίτερο ενδιαφέρον έχει η εικόνα Grad-CAM across-epoch για λανθασμένη πρόβλεψη. Σε αυτή την περίπτωση, η πραγματική κλάση είναι benign-like, αλλά το μοντέλο την ταξινομήσε ως suspicious με πιθανότητα 1.000. Η εικόνα αυτή είναι πολύ χρήσιμη, γιατί δείχνει ένα κρίσιμο είδος λάθους: το μοντέλο υπερεκτιμά την επικινδυνότητα της βλάβης. Οι ενεργοποιήσεις εστιάζουν σε έντονα τοπικά χαρακτηριστικά της βλάβης, τα οποία πιθανότατα οδήγησαν το μοντέλο να τη θεωρήσει suspicious. Αυτό δείχνει ότι η λανθασμένη πρόβλεψη δεν προέκυψε απαραίτητα επειδή το μοντέλο “κοίταξε” τελείως άσχετη περιοχή, αλλά επειδή ερμήνευσε ορισμένα οπτικά χαρακτηριστικά ως πιο ύποπτα από όσο έπρεπε.



Εικόνα 4.93. Grad-CAM across-epoch ανάλυση σε λανθασμένη πρόβλεψη benign-like ως suspicious

Η SHAP ανάλυση εφαρμόστηκε συμπληρωματικά για να αποδοθεί η τοπική συνεισφορά περιοχών της εικόνας στην έξοδο suspicious. Από τις διαθέσιμες SHAP εικόνες, η πιο αντιπροσωπευτική είναι η περίπτωση όπου η πραγματική κλάση είναι benign-like, αλλά η πρόβλεψη είναι suspicious με πιθανότητα 1.000. Η SHAP εικόνα δείχνει ότι οι σημαντικότερες θετικές και αρνητικές συνεισφορές συγκεντρώνονται κοντά στην περιοχή της βλάβης, αν και η μορφή της είναι πιο χονδρική και λιγότερο άμεση από τη Grad-CAM.

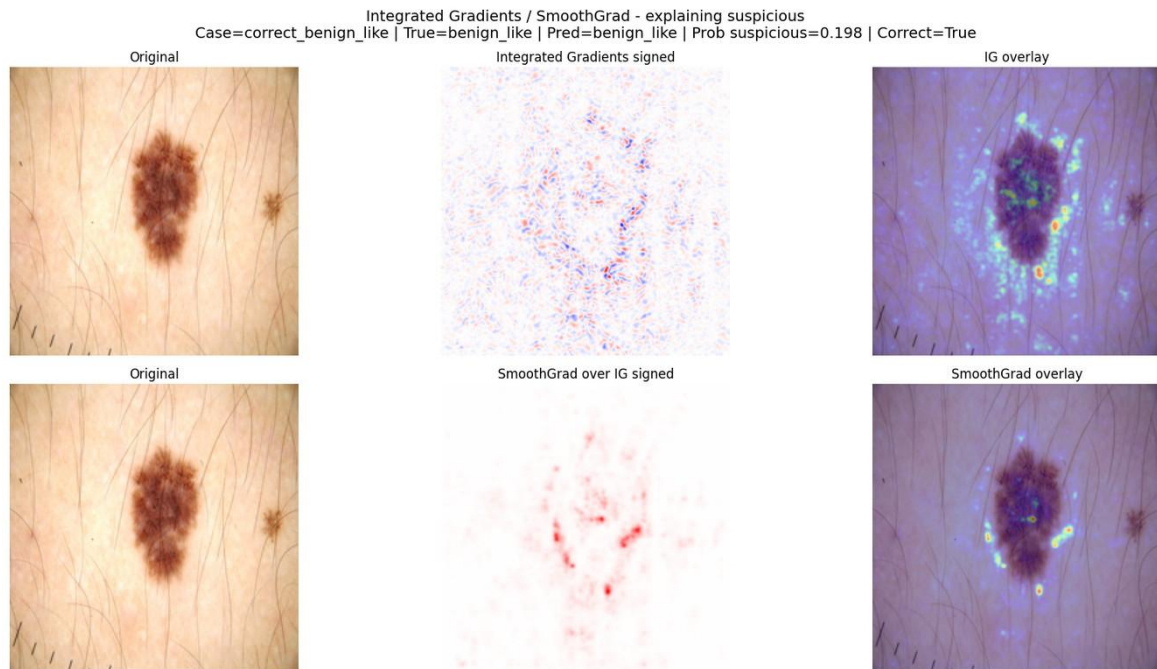


Εικόνα 4.94. SHAP ανάλυση για λανθασμένη πρόβλεψη benign-like ως suspicious

Η συγκεκριμένη SHAP εικόνα είναι χρήσιμη επειδή δίνει μια διαφορετική ανάγνωση του ίδιου λάθους. Δεν δείχνει απλώς το πού εστιάζει το μοντέλο, αλλά ποιες περιοχές συμβάλλουν στην πρόβλεψη της suspicious κλάσης. Με αυτόν τον τρόπο επιβεβαιώνει ότι

το λάθος συνδέεται με τοπικά χαρακτηριστικά της βλάβης και όχι αποκλειστικά με το φόντο.

Τέλος, χρησιμοποιήθηκαν Integrated Gradients / SmoothGrad ως συμπληρωματική μέθοδος.



Εικόνα 4.95. Ερμηνεία σωστής benign-like πρόβλεψης με Integrated Gradients και SmoothGrad

Συνολικά, η εξηγησιμότητα έδειξε ότι το μοντέλο στις σωστές προβλέψεις τείνει να εστιάζει σε περιοχές που σχετίζονται με τη βλάβη. Στις λανθασμένες προβλέψεις, όμως, η εικόνα είναι πιο σύνθετη. Το μοντέλο μπορεί να κοιτάζει τη βλάβη, αλλά να αποδίδει υπερβολική σημασία σε συγκεκριμένα τοπικά χαρακτηριστικά, οδηγώντας σε λανθασμένη ταξινόμηση. Αυτό είναι ίσως το πιο χρήσιμο συμπέρασμα της εξηγησιμότητας: η σωστή χωρική εστίαση δεν εγγυάται πάντα σωστή ταξινόμηση.

Συνολικά, το πείραμα με EfficientNet-B0 στο ISIC 2019 έδειξε ικανοποιητική απόδοση στη δυαδική ταξινόμηση benign-like / suspicious, με Accuracy = 80,44%, F1-score = 80,87% και AUC = 88,56% στο test set. Το μοντέλο εμφάνισε σχετικά ισορροπημένη συμπεριφορά ανάμεσα στις δύο κλάσεις, αν και παρέμειναν αρκετά false positives και false negatives.

5. Τεχνικές εξηγησιμότητας και οπτική ερμηνεία των αποτελεσμάτων

Στο προηγούμενο κεφάλαιο παρουσιάστηκαν τα πειραματικά αποτελέσματα των μοντέλων που εκπαιδεύτηκαν για την αναγνώριση και κατηγοριοποίηση δερματικών βλαβών. Οι μετρικές, όπως το accuracy, το precision, το recall, το F1-score και το AUC, έδωσαν μια πρώτη εικόνα για την απόδοση κάθε μοντέλου. Οι αριθμοί αυτοί είναι απαραίτητοι, γιατί επιτρέπουν τη σύγκριση ανάμεσα σε διαφορετικά μοντέλα, datasets και πειραματικές ρυθμίσεις. Στην περίπτωση της ιατρικής εικόνας, όμως, η αριθμητική επίδοση δεν αρκεί από μόνη της. Ένα μοντέλο μπορεί να εμφανίζει υψηλή ακρίβεια, χωρίς αυτό να σημαίνει απαραίτητα ότι εστιάζει στα σωστά σημεία της εικόνας. Μπορεί, για παράδειγμα, να επηρεάζεται από το φόντο, τον φωτισμό, τις σκιές, τα όρια της φωτογραφίας ή άλλα οπτικά στοιχεία που δεν σχετίζονται άμεσα με τη δερματική βλάβη. Αυτό είναι ένα κρίσιμο ζήτημα, ειδικά όταν η εργασία αφορά τη διάκριση μελανώματος και άλλων δερματικών αλλοιώσεων. Για τον λόγο αυτό, το παρόν κεφάλαιο εστιάζει στις τεχνικές εξηγησιμότητας που συνόδευαν τα αποτελέσματα του τέταρτου κεφαλαίου. Στόχος δεν είναι μόνο να παρουσιαστεί αν ένα μοντέλο ταξινομήσε σωστά ή λανθασμένα μια εικόνα, αλλά και να εξεταστεί πού φαίνεται να βασίζει την απόφασή του. Η ανάλυση αυτή δίνει μια δεύτερη, πιο οπτική ανάγνωση των πειραμάτων. Στην εργασία αξιοποιήθηκαν τεχνικές όπως Grad-CAM, Grad-CAM++, Score-CAM και SHAP. Παράλληλα, εξετάστηκε και η εξέλιξη της οπτικής εστίασης των μοντέλων σε διαφορετικές εποχές εκπαίδευσης, μέσα από across-epoch οπτικοποιήσεις. Με αυτόν τον τρόπο, η αξιολόγηση των μοντέλων δεν περιορίστηκε στις τελικές μετρικές, αλλά συνδέθηκε με την οπτική συμπεριφορά τους.

5.1 Η ανάγκη για εξηγησιμότητα σε προβλήματα ιατρικής εικόνας

Τα μοντέλα βαθιάς μάθησης μπορούν να αναγνωρίζουν σύνθετα μοτίβα σε εικόνες. Αυτό τα καθιστά χρήσιμα σε προβλήματα ταξινόμησης δερματικών βλαβών, όπου οι οπτικές διαφορές ανάμεσα στις κατηγορίες μπορεί να είναι λεπτές. Την ίδια στιγμή, όμως, η λειτουργία τους συχνά παραμένει δύσκολη στην ερμηνεία. Το μοντέλο δίνει μια πρόβλεψη, αλλά δεν είναι πάντα φανερό ποια στοιχεία της εικόνας οδήγησαν σε αυτήν. Στην παρούσα εργασία, η εξηγησιμότητα είχε ιδιαίτερη σημασία για δύο λόγους. Ο πρώτος λόγος είναι ότι τα datasets δερματικών βλαβών παρουσιάζουν συχνά ανισορροπία κατηγοριών. Σε τέτοιες περιπτώσεις, ένα μοντέλο μπορεί να δείχνει καλή γενική απόδοση, ενώ δυσκολεύεται σε λιγότερο συχνές αλλά κλινικά κρίσιμες κατηγορίες. Ο δεύτερος λόγος είναι ότι η σωστή πρόβλεψη δεν σημαίνει πάντα σωστή ερμηνευτική βάση. Ένα μοντέλο μπορεί να προβλέψει σωστά μια εικόνα, αλλά να έχει εστιάσει σε περιοχή που δεν αντιστοιχεί στη βλάβη.

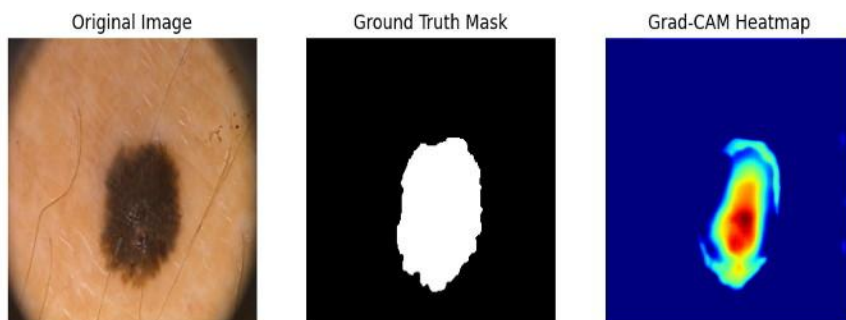
Οι τεχνικές εξηγησιμότητας λειτουργούν ως ένα πρόσθετο επίπεδο ελέγχου. Μέσα από θερμικούς χάρτες και οπτικοποιήσεις συμβολής χαρακτηριστικών, γίνεται δυνατή η παρατήρηση της περιοχής που επηρεάζει περισσότερο την πρόβλεψη. Αυτό δεν σημαίνει ότι οι τεχνικές αυτές αποκαλύπτουν πλήρως τον τρόπο «σκέψης» του μοντέλου. Προσφέρουν, όμως, χρήσιμες ενδείξεις για το αν η συμπεριφορά του είναι συμβατή με το αντικείμενο της ταξινόμησης. Στο πλαίσιο της εργασίας, η εξηγησιμότητα δεν αντιμετωπίστηκε ως διακοσμητικό στοιχείο. Αντίθετα, συνδέθηκε με τα αποτελέσματα του τέταρτου κεφαλαίου. Για κάθε βασικό πείραμα, οι οπτικοποιήσεις βοήθησαν να εξεταστεί

αν οι αριθμητικές επιδόσεις συνοδεύονται από λογική οπτική εστίαση. Αυτό είναι σημαντικό, γιατί σε ένα πρόβλημα ιατρικής εικόνας η αξιολόγηση χρειάζεται να είναι πιο προσεκτική από μια απλή αναφορά ποσοστών.

5.2 Grad-CAM και οπτική ερμηνεία προβλέψεων

Η πρώτη βασική τεχνική εξηγησιμότητας που εφαρμόστηκε στα πειράματα ήταν το Grad-CAM. Η μέθοδος αυτή παράγει έναν θερμικό χάρτη πάνω στην αρχική εικόνα, δείχνοντας ποιες περιοχές συνέβαλαν περισσότερο στην πρόβλεψη του μοντέλου. Στην πράξη, το Grad-CAM ήταν χρήσιμο επειδή έδινε μια άμεση οπτική ένδειξη της συμπεριφοράς του μοντέλου. Όταν η πιο έντονη περιοχή του heatmap βρισκόταν πάνω στη δερματική βλάβη, υπήρχε μια πρώτη ένδειξη ότι το μοντέλο βασίστηκε σε σχετική πληροφορία. Αν η ενεργοποίηση εμφανιζόταν κυρίως στο φόντο ή σε περιφερειακά σημεία, τότε η πρόβλεψη χρειαζόταν πιο προσεκτική ερμηνεία.

Στα πειράματα του τέταρτου κεφαλαίου, το Grad-CAM αξιοποιήθηκε τόσο σε σωστές όσο και σε λανθασμένες προβλέψεις. Αυτή η σύγκριση ήταν χρήσιμη, γιατί έδειξε ότι οι σωστές προβλέψεις δεν έχουν πάντα καθαρή οπτική εξήγηση. Αντίστοιχα, σε ορισμένες λανθασμένες προβλέψεις, το μοντέλο φαινόταν να εστιάζει πράγματι στη βλάβη, αλλά να μην καταφέρνει να διακρίνει σωστά την κατηγορία.



Εικόνα 5.1. Παράδειγμα Grad-CAM σε σωστά ταξινομημένη δερματική βλάβη

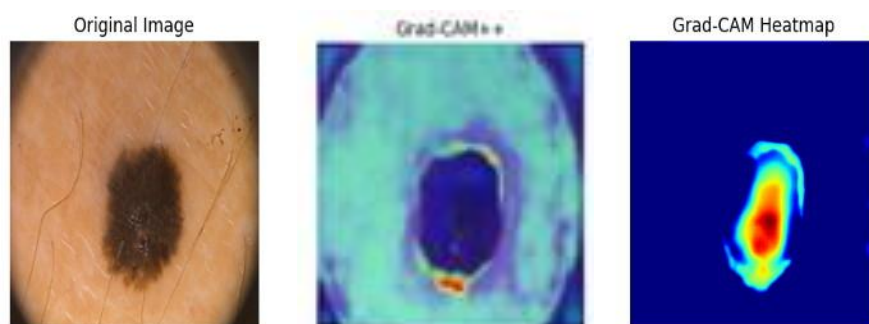
Η χρήση του Grad-CAM βοήθησε να γίνει πιο προσεκτική η ανάγνωση των αποτελεσμάτων. Δεν αρκούσε, δηλαδή, να αναφερθεί ότι ένα μοντέλο πέτυχε υψηλό AUC ή υψηλό F1-score. Έπρεπε να εξεταστεί και αν η οπτική του εστίαση βρίσκεται σε περιοχές που έχουν νόημα για το συγκεκριμένο πρόβλημα.

5.3 Grad-CAM++ και πιο λεπτομερής εστίαση

Δίπλα στο Grad-CAM, εφαρμόστηκε και το Grad-CAM++. Η τεχνική αυτή κινείται στην ίδια γενική λογική, αλλά μπορεί να δώσει πιο λεπτομερείς ή πιο συγκεντρωμένες ενεργοποιήσεις σε ορισμένες περιπτώσεις [13]. Αυτό έχει ενδιαφέρον στις δερματοσκοπικές εικόνες, επειδή οι βλάβες δεν έχουν πάντα απλή μορφή. Μπορεί να παρουσιάζουν ακανόνιστα όρια, ετερογένεια χρώματος, μικρές περιοχές διαφορετικής υφής ή σημεία που δεν είναι εξίσου εμφανή σε ολόκληρη την εικόνα. Σε τέτοιες περιπτώσεις, μια

πιο λεπτομερής οπτική ερμηνεία μπορεί να βοηθήσει στην καλύτερη παρατήρηση της συμπεριφοράς του μοντέλου.

Στα πειράματα, το Grad-CAM++ χρησιμοποιήθηκε για να εξεταστεί αν η εστίαση παραμένει κοντά στη βλάβη ή αν διαφοροποιείται σε σχέση με το απλό Grad-CAM. Σε κάποιες εικόνες, οι δύο τεχνικές έδιναν παρόμοια αποτελέσματα. Αυτό έκανε την ερμηνεία πιο σταθερή. Σε άλλες εικόνες, το Grad-CAM++ παρήγαγε πιο περιορισμένη ενεργοποίηση, κάτι που βοήθησε να εντοπιστούν συγκεκριμένα τμήματα της εικόνας που φαίνεται να επηρέασαν την πρόβλεψη.

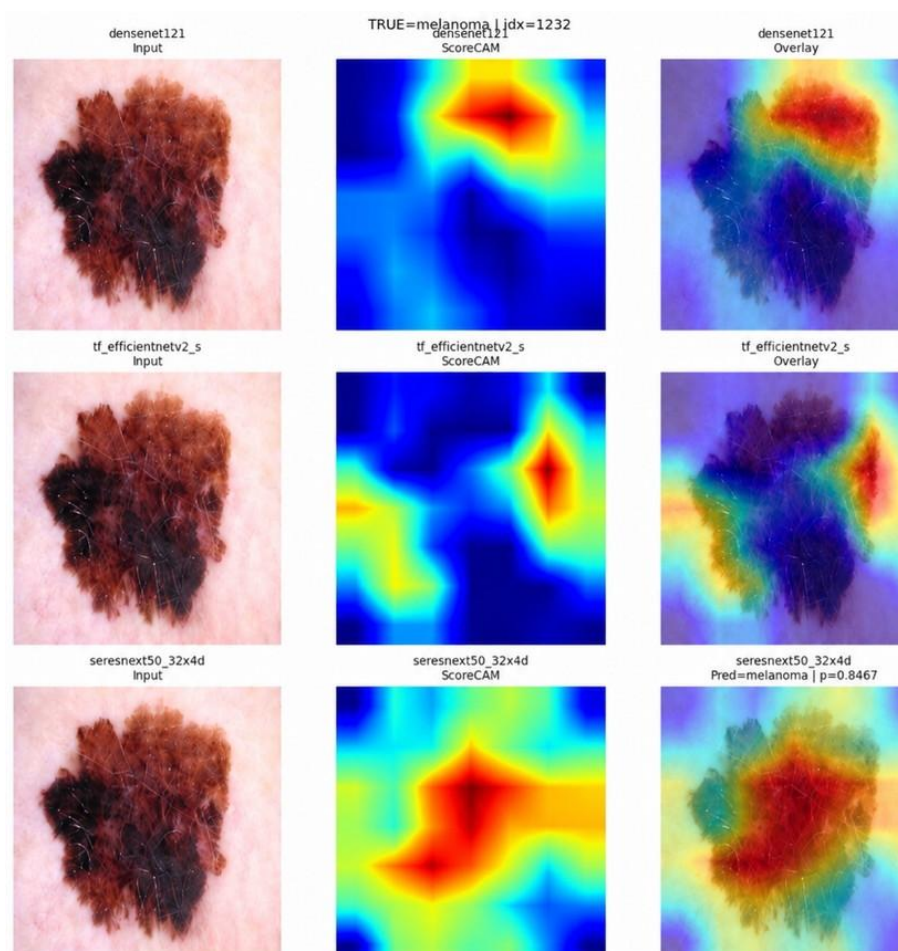


Εικόνα 5.2. Σύγκριση Grad-CAM και Grad-CAM++ στην ίδια δερματοσκοπική εικόνα

Η σύγκριση των δύο μεθόδων δείχνει ότι η εξηγησιμότητα δεν πρέπει να βασίζεται μηχανικά σε μία μόνο τεχνική. Ένα heatmap προσφέρει μια ένδειξη, όχι πλήρη απόδειξη. Για αυτόν τον λόγο, η εργασία αξιοποίησε περισσότερες από μία μεθόδους, ώστε η ερμηνεία να μην εξαρτάται αποκλειστικά από ένα μόνο είδος οπτικοποίησης.

5.4 Score-CAM και σύγκριση με τις gradient-based τεχνικές

Μια ακόμη τεχνική που εφαρμόστηκε ήταν το Score-CAM. Σε αντίθεση με τις μεθόδους που βασίζονται άμεσα στα gradients, το Score-CAM αξιολογεί την επίδραση περιοχών της εικόνας στην τελική απόκριση του μοντέλου [14]. Αυτό το καθιστά χρήσιμο ως συμπληρωματική τεχνική, επειδή δίνει μια διαφορετική οπτική ερμηνείας. Στην εργασία, το Score-CAM βοήθησε κυρίως στις περιπτώσεις όπου τα Grad-CAM heatmaps ήταν πιο διάχυτα ή όχι τόσο καθαρά. Όταν το Score-CAM ανέδειχνε παρόμοια περιοχή με το Grad-CAM και το Grad-CAM++, η ερμηνεία γινόταν πιο πειστική. Όταν οι τεχνικές έδιναν διαφορετικές περιοχές ενδιαφέροντος, αυτό έδειχνε ότι η συγκεκριμένη πρόβλεψη δεν πρέπει να ερμηνευθεί βιαστικά. Το Score-CAM δεν αντικατέστησε τις υπόλοιπες τεχνικές. Λειτουργήσε ως πρόσθετη μορφή ελέγχου. Η παρουσία πολλών τεχνικών εξηγησιμότητας βοήθησε να αναδειχθούν τόσο οι περιπτώσεις συμφωνίας όσο και οι περιπτώσεις απόκλισης ανάμεσα στις μεθόδους.



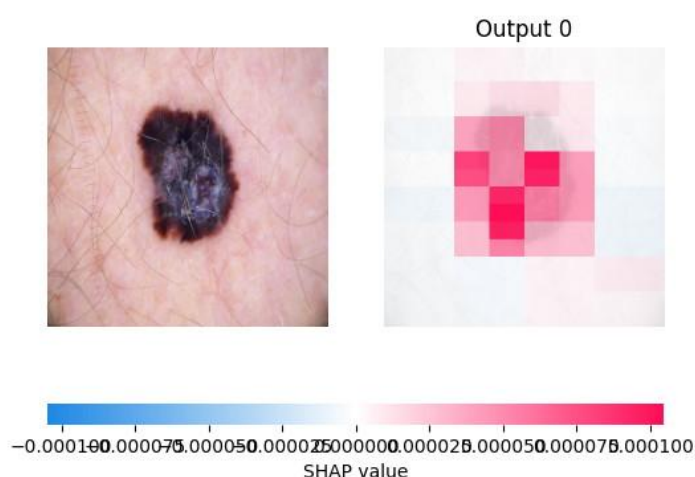
Εικόνα 5.3. Score-CAM σε 3 backbone για σωστή πρόβλεψη melanoma

5.5 SHAP και ανάλυση συμβολής χαρακτηριστικών

Μια κεντρική τεχνική εξηγησιμότητας που εφαρμόστηκε σε όλα τα πειράματα ήταν το SHAP. Σε αντίθεση με τις τεχνικές τύπου Grad-CAM, οι οποίες δείχνουν κυρίως περιοχές ενεργοποίησης, το SHAP βασίζεται στη λογική της συμβολής των χαρακτηριστικών στην τελική πρόβλεψη [15]. Με πιο απλά λόγια, προσπαθεί να δείξει ποια τμήματα της εικόνας ωθούν το μοντέλο προς μια συγκεκριμένη κατηγορία και ποια λειτουργούν αντίθετα. Η παρουσία του SHAP σε όλα τα πειράματα ήταν σημαντική, γιατί έδωσε μια πιο συστηματική μορφή εξηγησιμότητας. Δεν εφαρμόστηκε μόνο σε μεμονωμένες εικόνες, αλλά συνόδευσε συνολικά την πειραματική διαδικασία. Αυτό επέτρεψε να εξεταστεί αν η συμπεριφορά των μοντέλων παραμένει σταθερή σε διαφορετικά datasets, αρχιτεκτονικές και είδη ταξινόμησης.

Στο πλαίσιο της παρούσας εργασίας, οι οπτικοποιήσεις SHAP αξιοποιήθηκαν για να ελεγχθεί αν τα σημαντικά τμήματα της εικόνας συμπίπτουν με τη δερματική βλάβη ή αν η πρόβλεψη επηρεάζεται και από γειτονικές περιοχές. Αυτό ήταν χρήσιμο κυρίως σε εικόνες όπου τα heatmaps των Grad-CAM τεχνικών δεν ήταν απόλυτα καθαρά. Το SHAP πρόσφερε μια διαφορετική ανάγνωση, καθώς δεν έδειχνε απλώς «πού κοιτάζει» το μοντέλο, αλλά ποια

οπτικά στοιχεία φαίνεται να συμβάλλουν στην πρόβλεψη. Ιδιαίτερο ενδιαφέρον είχε η σύγκριση του SHAP με τις υπόλοιπες τεχνικές εξηγησιμότητας. Όταν οι περιοχές που αναδείκνυε το SHAP συμφωνούσαν με τις περιοχές υψηλής ενεργοποίησης από Grad-CAM, Grad-CAM++ ή Score-CAM, η ερμηνεία της πρόβλεψης γινόταν πιο σταθερή. Όταν υπήρχαν διαφορές, αυτό αποτελούσε ένδειξη ότι η συγκεκριμένη εικόνα χρειαζόταν πιο προσεκτική ανάλυση. Στις σωστές προβλέψεις, το SHAP βοήθησε να φανεί αν το μοντέλο βασίστηκε σε περιοχές που σχετίζονται με τη βλάβη. Σε αρκετές περιπτώσεις, οι πιο σημαντικές συνεισφορές εμφανίζονταν μέσα ή κοντά στην αλλοίωση. Αυτό ενίσχυε την εικόνα ότι η πρόβλεψη είχε λογική οπτική βάση. Σε πιο δύσκολες ή λανθασμένες προβλέψεις, το SHAP ανέδειξε ορισμένες φορές περιοχές που δεν ταυτίζονταν καθαρά με τη βλάβη. Αυτό έδειξε ότι το μοντέλο μπορεί να επηρεάζεται από λιγότερο σχετικές πληροφορίες.



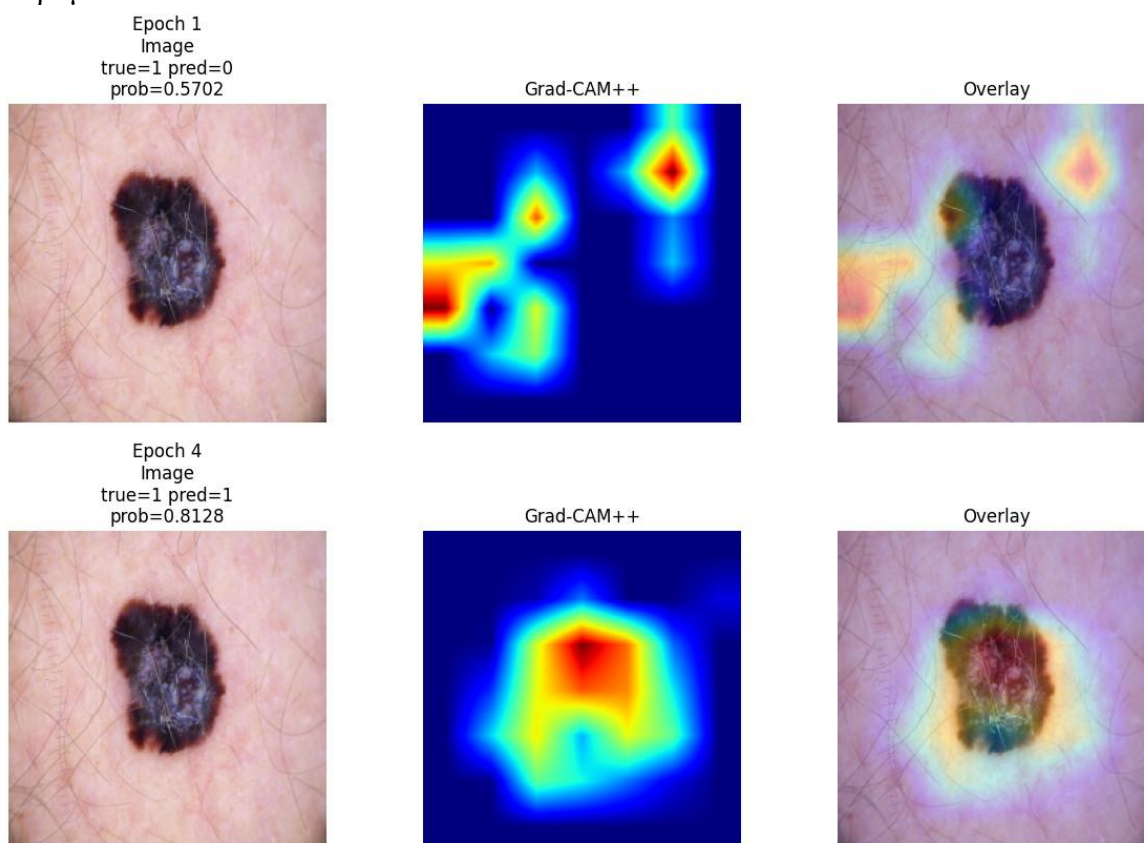
Εικόνα 5.4. Παράδειγμα μελανώματος με εξηγησιμότητα SHAP

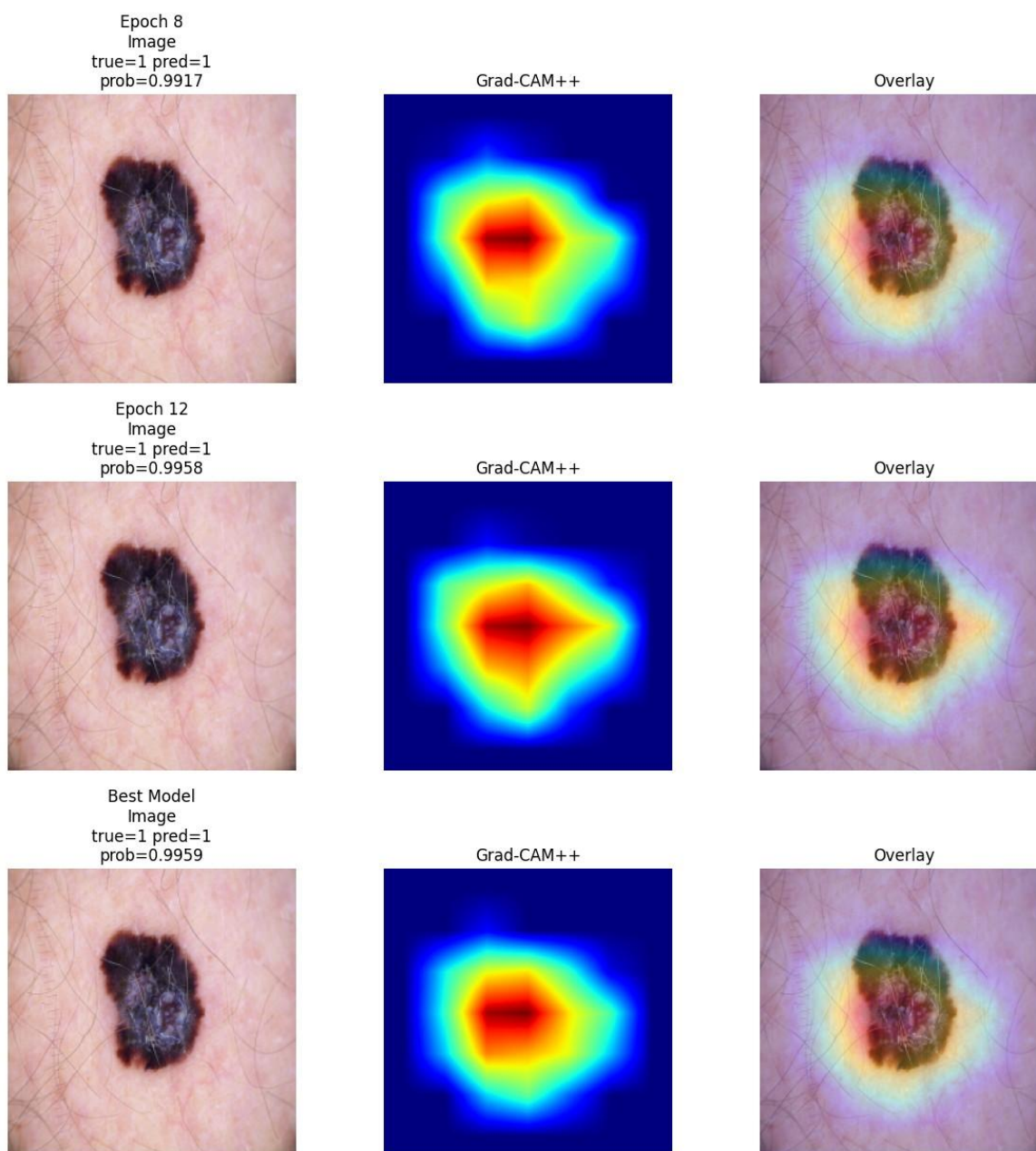
Η ερμηνεία του SHAP σε εικόνες χρειάζεται προσοχή. Οι δερματοσκοπικές εικόνες είναι δεδομένα υψηλής διάστασης και η τελική οπτικοποίηση μπορεί να επηρεαστεί από τον τρόπο με τον οποίο αναλύεται η εικόνα σε περιοχές ή χαρακτηριστικά. Για τον λόγο αυτό, το SHAP δεν αντιμετωπίστηκε ως μοναδική απάντηση στο ζήτημα της εξηγησιμότητας. Λειτουργήσε μαζί με τις υπόλοιπες τεχνικές, δίνοντας μια πιο πλήρη εικόνα της συμπεριφοράς των μοντέλων. Συνολικά, το SHAP αποτέλεσε βασικό μέρος της ανάλυσης. Η εφαρμογή του σε όλα τα πειράματα βοήθησε να υπάρχει κοινή γραμμή ερμηνείας, ανεξάρτητα από το dataset ή το μοντέλο που εξεταζόταν κάθε φορά.

5.6 Across-epoch explainability: εξέλιξη της προσοχής κατά την εκπαίδευση

Ένα από τα πιο ενδιαφέροντα στοιχεία της εργασίας ήταν η ανάλυση εξηγησιμότητας σε διαφορετικές εποχές εκπαίδευσης. Η λογική αυτής της προσέγγισης ήταν να μη μελετηθεί μόνο το τελικό heatmap του μοντέλου, αλλά και η πορεία της οπτικής εστίασης κατά τη διάρκεια της εκπαίδευσης. Στις πρώτες εποχές, ένα μοντέλο μπορεί να εστιάζει σε πιο γενικά ή ακόμη και άσχετα σημεία της εικόνας. Καθώς η εκπαίδευση προχωρά, είναι πιθανό η προσοχή του να μετακινείται πιο κοντά στη βλάβη. Αν αυτή η μεταβολή συνοδεύεται από

βελτίωση των validation metrics, τότε η εκπαίδευση αποκτά πιο πειστική ερμηνεία. Στα πειράματα του τέταρτου κεφαλαίου, η across-epoch ανάλυση συνδέθηκε με τις τιμές ανά εποχή, όπως validation loss, validation accuracy, F1-score και AUC. Αυτό επέτρεψε να εξεταστεί αν η αριθμητική βελτίωση συνοδευόταν και από πιο καθαρή οπτική εστίαση. Σε αρκετές περιπτώσεις, τα πρώτα heatmaps ήταν πιο διάχυτα, ενώ στις επόμενες εποχές η ενεργοποίηση γινόταν πιο συγκεντρωμένη στη βλάβη. Αυτό έδινε την εικόνα ότι το μοντέλο σταδιακά μαθαίνει πιο σχετική οπτική πληροφορία. Σε άλλες περιπτώσεις, οι μετρικές βελτιώνονταν χωρίς να υπάρχει αντίστοιχα καθαρή βελτίωση στο heatmap. Αυτή η παρατήρηση είναι σημαντική, γιατί δείχνει ότι οι αριθμοί και η οπτική εξήγηση δεν κινούνται πάντα με τον ίδιο τρόπο. Η across-epoch εξηγησιμότητα βοήθησε να παρουσιαστεί η εκπαίδευση όχι μόνο ως μια σειρά από τιμές loss και accuracy, αλλά ως διαδικασία σταδιακής μεταβολής της προσοχής του μοντέλου. Με ό,τι αυτό συνεπάγεται, η συγκεκριμένη ανάλυση πρόσθεσε ένα πιο δυναμικό στοιχείο στην ερμηνεία των πειραμάτων.



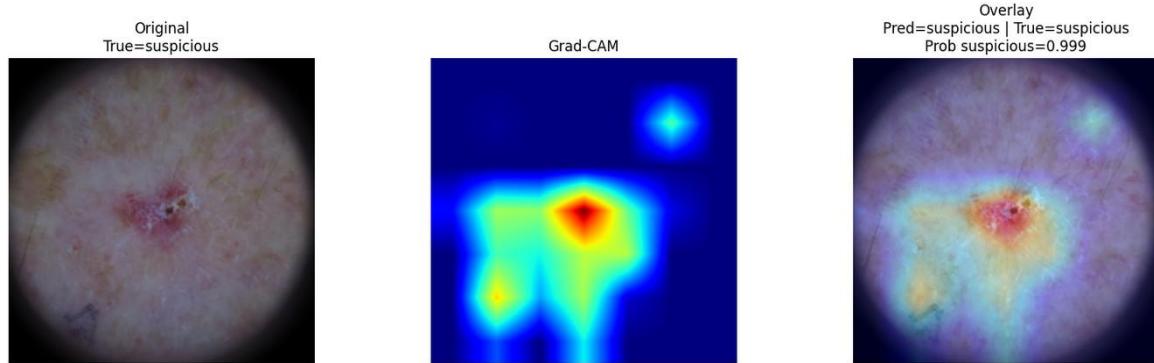


Εικόνα 5.5. Grad-CAM++ across-epoch ανάλυση για melanoma περίπτωση στο HAM10000

5.7 Ανάλυση σωστών προβλέψεων

Στις σωστές προβλέψεις, οι τεχνικές εξηγησιμότητας αξιοποιήθηκαν για να εξεταστεί αν το μοντέλο εστίαζε στη δερματική βλάβη. Αυτό είναι βασικό, γιατί μια σωστή ταξινόμηση δεν σημαίνει πάντα ότι το μοντέλο βασίστηκε στα κατάλληλα οπτικά στοιχεία. Σε αρκετές σωστές περιπτώσεις, τα Grad-CAM, Grad-CAM++ και Score-CAM heatmaps έδειξαν ενεργοποίηση πάνω ή κοντά στη βλάβη. Παρόμοια, οι SHAP οπτικοποιήσεις ανέδειξαν περιοχές που σχετίζονταν με την αλλοίωση. Όταν οι τεχνικές συμφωνούσαν μεταξύ τους, η πρόβλεψη μπορούσε να θεωρηθεί πιο σταθερή ερμηνευτικά. Υπήρχαν, βέβαια, και σωστές

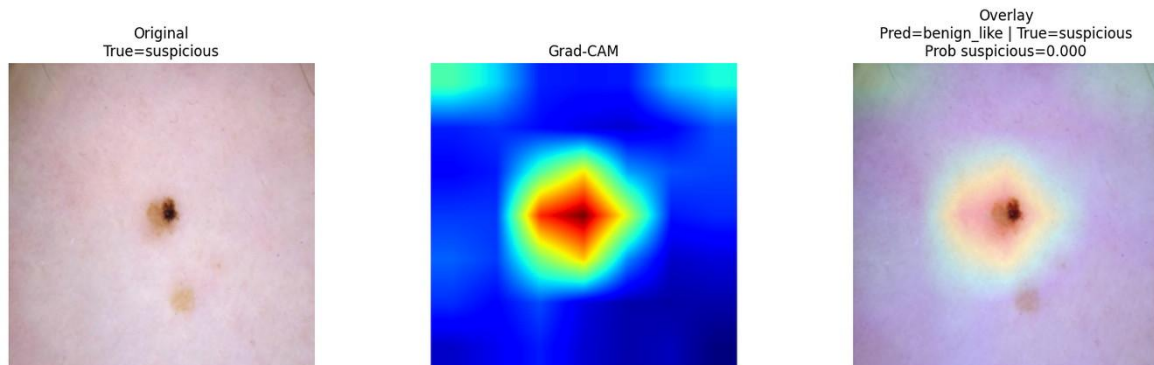
προβλέψεις όπου η εικόνα δεν ήταν τόσο καθαρή. Σε ορισμένες περιπτώσεις, το heatmap ήταν διάχυτο ή το SHAP ανέδειξε και περιοχές εκτός της βλάβης. Αυτό δεν ακυρώνει τη σωστή πρόβλεψη, αλλά δείχνει ότι το μοντέλο μπορεί να αξιοποιεί και ευρύτερα οπτικά στοιχεία της εικόνας. Η ανάλυση σωστών προβλέψεων έδειξε ότι η πιο πειστική περίπτωση είναι εκείνη όπου συνδυάζονται τρία στοιχεία: σωστή ετικέτα, καλές μετρικές και λογική οπτική εστίαση. Όταν αυτά τα στοιχεία συμφωνούν, το αποτέλεσμα αποκτά μεγαλύτερη ερμηνευτική σταθερότητα.



Εικόνα 5.6. Σωστή suspicious πρόβλεψη του Swin-Tiny + SwiGLU με Grad-CAM

5.8 Ανάλυση λανθασμένων ή δύσκολων προβλέψεων

Οι λανθασμένες και δύσκολες προβλέψεις είχαν ιδιαίτερη σημασία, γιατί ανέδειξαν τα όρια των μοντέλων. Μέσα από τις τεχνικές εξηγησιμότητας μπορούσε να εξεταστεί αν το λάθος συνδέεται με λανθασμένη εστίαση ή αν η εικόνα είναι από μόνη της δύσκολη για ταξινόμηση. Σε ορισμένες λανθασμένες προβλέψεις, τα heatmaps εμφάνιζαν ενεργοποίηση έξω από τη βλάβη. Παράλληλα, οι SHAP οπτικοποιήσεις ανέδειξαν περιοχές που δεν σχετίζονταν καθαρά με το δερματικό εύρημα. Σε τέτοιες περιπτώσεις, υπάρχει ένδειξη ότι το μοντέλο επηρεάστηκε από μη ουσιαστικά οπτικά στοιχεία, όπως φωτισμός, φόντο ή περιφερειακά σημεία της εικόνας. Υπήρχαν όμως και πιο σύνθετες περιπτώσεις. Σε ορισμένες λανθασμένες προβλέψεις, το μοντέλο φαινόταν να εστιάζει πράγματι στη βλάβη, αλλά η ταξινόμηση παρέμενε λανθασμένη. Αυτό δείχνει ότι η σωστή χωρική εστίαση δεν αρκεί πάντα. Το μοντέλο μπορεί να εντοπίζει τη σωστή περιοχή, αλλά να μην έχει μάθει επαρκώς τα χαρακτηριστικά που διαχωρίζουν τις κατηγορίες. Αυτή η παρατήρηση είναι σημαντική για την εργασία. Δείχνει ότι οι τεχνικές εξηγησιμότητας δεν πρέπει να ερμηνεύονται απλουστευτικά. Ένα καλό heatmap ή μια φαινομενικά σωστή SHAP οπτικοποίηση δεν εγγυάται σωστή ταξινόμηση. Αντίστοιχα, μια σωστή πρόβλεψη δεν σημαίνει πάντα ότι το μοντέλο έχει εστιάσει με τον επιθυμητό τρόπο.

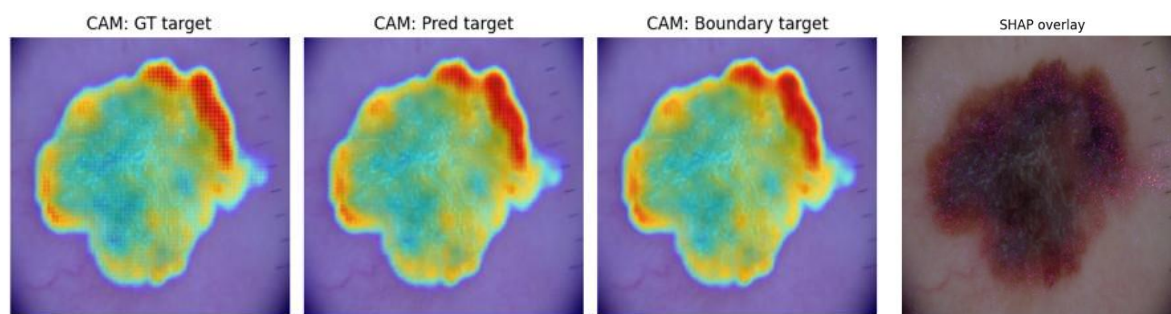


Εικόνα 5.7. Grad-CAM ανάλυση λανθασμένης suspicious περίπτωσης που ταξινομήθηκε ως benign-like

5.9 Σύγκριση heatmaps και SHAP με segmentation masks

Σε πειράματα όπου υπήρχαν διαθέσιμες μάσκες βλάβης, οι οπτικοποιήσεις εξηγησιμότητας συγκρίθηκαν με την πραγματική περιοχή της αλλοίωσης. Η σύγκριση αυτή είχε στόχο να εξετάσει αν η περιοχή που αναδεικνύεται από το μοντέλο συμπίπτει με τη βλάβη. Τα heatmaps από Grad-CAM, Grad-CAM++ και Score-CAM μπορούν να συγκριθούν οπτικά με τη μάσκα. Αν η ενεργοποίηση βρίσκεται μέσα ή κοντά στα όρια της βλάβης, τότε η ερμηνεία είναι πιο συμβατή με το αντικείμενο της ταξινόμησης. Αν η ενεργοποίηση εμφανίζεται κυρίως εκτός μάσκας, τότε η πρόβλεψη χρειάζεται προσοχή. Παρόμοια λογική εφαρμόστηκε και στις SHAP οπτικοποιήσεις. Το ζητούμενο ήταν να εξεταστεί αν οι περιοχές με υψηλή συμβολή βρίσκονται κοντά στη βλάβη ή αν απλώνονται σε άσχετα σημεία της εικόνας. Η ύπαρξη μάσκας πρόσφερε ένα πιο σαφές σημείο αναφοράς. Χωρίς μάσκα, η ερμηνεία βασίζεται κυρίως στην οπτική παρατήρηση. Με τη μάσκα, μπορεί να εξεταστεί πιο καθαρά αν η εστίαση του μοντέλου βρίσκεται στην περιοχή που αντιστοιχεί στη βλάβη. Πρέπει, βέβαια, να σημειωθεί ότι τα heatmaps και το SHAP δεν είναι segmentation μέθοδοι. Δεν έχουν ως στόχο να παράγουν ακριβές περίγραμμα της βλάβης. Δείχνουν περιοχές που επηρεάζουν την πρόβλεψη. Για αυτόν τον λόγο, η σύγκριση με μάσκες πρέπει να αντιμετωπίζεται ως συμπληρωματικός έλεγχος και όχι ως απόλυτη μέτρηση τμηματοποίησης.





Εικόνα 5.8. Σύγκριση segmentation mask με CAM heatmaps και SHAP overlay σε δερματοσκοπική εικόνα. Η εικόνα δείχνει κατά πόσο η οπτική εστίαση του μοντέλου αντιστοιχεί στην περιοχή της βλάβης.

5.10 Τεκμηρίωση GitHub

Κατά τη διάρκεια της εργασίας δημιουργήθηκε αποθετήριο στο GitHub, ώστε να οργανωθεί καλύτερα το τεχνικό υλικό των πειραμάτων. Εκεί συγκεντρώθηκαν τα βασικά notebooks και οι κώδικες που χρησιμοποιήθηκαν για τη φόρτωση και προετοιμασία των datasets, την εκπαίδευση των μοντέλων, την αξιολόγηση των αποτελεσμάτων και την παραγωγή των εικόνων εξηγησιμότητας. Η ανάγκη αυτή προέκυψε κυρίως επειδή τα πειράματα δεν βασίστηκαν σε ένα μόνο dataset ή σε ένα μόνο μοντέλο, αλλά περιλάμβαναν διαφορετικές δοκιμές, διαφορετικές αρχιτεκτονικές και διαφορετικές διαδικασίες προεπεξεργασίας. Στο αποθετήριο οργανώθηκαν στοιχεία που σχετίζονται με τα datasets που αξιοποιήθηκαν, τα προεκπαιδευμένα μοντέλα που προσαρμόστηκαν στα δεδομένα της εργασίας, τα scripts εκπαίδευσης και αξιολόγησης, καθώς και τα παραγόμενα αποτελέσματα, όπως μετρικές, πίνακες, confusion matrices και οπτικοποιήσεις. Ιδιαίτερη θέση είχαν και οι τεχνικές εξηγησιμότητας, όπως Grad-CAM, Grad-CAM++, Score-CAM και SHAP, καθώς οι εικόνες που προέκυψαν από αυτές συνδέθηκαν με την ερμηνεία των αποτελεσμάτων στα προηγούμενα κεφάλαια. Το GitHub δεν χρησιμοποιείται εδώ ως ανεξάρτητο ερευνητικό αποτέλεσμα, αλλά ως τρόπος τεκμηρίωσης της διαδικασίας. Μέσα από το αποθετήριο μπορεί να φανεί πιο καθαρά πώς οργανώθηκαν τα πειράματα, ποια βήματα ακολουθήθηκαν και πώς συνδέονται οι πίνακες και οι εικόνες της εργασίας με τον αντίστοιχο κώδικα. Με αυτή τη λογική, λειτουργεί υποστηρικτικά ως μέσο αναπαραγωγιμότητας, διαφάνειας και καλύτερης παρουσίασης της τεχνικής πλευράς της πτυχιακής. Τα πλήρη datasets δεν περιλαμβάνονται στο αποθετήριο λόγω μεγέθους και περιορισμών πρόσβασης, αλλά αναφέρονται στη σχετική τεκμηρίωση.

Διαθέσιμο στο: <https://github.com/Alekosbis/melanoma-classification-explainability>

6. Συμπεράσματα

Η παρούσα πτυχιακή εργασία είχε ως βασικό στόχο τη μελέτη της αναγνώρισης και κατηγοριοποίησης δερματικών βλαβών, με έμφαση στο μελάνωμα, μέσα από τεχνικές βαθιάς μάθησης και μοντέλα που μπορούν να υποστηρίξουν τη διάγνωση σε δερματοσκοπικές εικόνες. Το μελάνωμα αποτελεί μία από τις πιο σοβαρές μορφές καρκίνου του δέρματος, κυρίως επειδή η έγκαιρη αναγνώρισή του μπορεί να επηρεάσει σε μεγάλο βαθμό την πρόγνωση του ασθενούς. Μέσα σε αυτό το πλαίσιο, η εργασία δεν αντιμετώπισε την τεχνητή νοημοσύνη ως αντικατάσταση της ιατρικής κρίσης, αλλά ως ένα πιθανό βοηθητικό εργαλείο, το οποίο μπορεί να συμβάλει στην καλύτερη αξιολόγηση εικόνων και στη μείωση της αβεβαιότητας κατά την ταξινόμηση. Στο θεωρητικό μέρος εξετάστηκε αρχικά το ιατρικό και τεχνολογικό υπόβαθρο του προβλήματος. Παρουσιάστηκαν βασικές έννοιες που αφορούν τις δερματικές βλάβες, τα χαρακτηριστικά του μελανώματος, τον ρόλο της δερματοσκόπησης και τη σημασία των συνόλων δεδομένων στην ανάπτυξη μοντέλων μηχανικής μάθησης. Στη συνέχεια αναλύθηκαν τα κυριότερα datasets που συναντώνται στη βιβλιογραφία, όπως τα HAM10000, ISIC 2019, ISIC 2020, PH2 και Derm7pt, καθώς και οι βασικές μετρικές αξιολόγησης που χρειάζονται για την ερμηνεία των πειραματικών αποτελεσμάτων. Η αναφορά στις μετρικές κρίθηκε απαραίτητη, γιατί σε τέτοια προβλήματα η απλή ακρίβεια δεν είναι πάντα αρκετή. Σε σύνολα δεδομένων με ανισορροπία ανάμεσα στις κατηγορίες, μετρικές όπως το F1-score, η ευαισθησία, η ειδικότητα και το AUC δίνουν πιο ουσιαστική εικόνα για τη συμπεριφορά του μοντέλου. Στο πειραματικό μέρος της εργασίας δοκιμάστηκαν διαφορετικές προσεγγίσεις βαθιάς μάθησης για την ταξινόμηση δερματοσκοπικών εικόνων. Η διαδικασία περιλάμβανε την προετοιμασία των δεδομένων, τον διαχωρισμό τους σε σύνολα εκπαίδευσης και αξιολόγησης, την αντιμετώπιση της ανισορροπίας των κλάσεων και την εκπαίδευση μοντέλων με διαφορετικές αρχιτεκτονικές. Δόθηκε έμφαση τόσο σε CNN μοντέλα, όπως EfficientNet και ResNet, όσο και σε πιο σύγχρονες λογικές που σχετίζονται με μετασχηματιστές ή υβριδικές αρχιτεκτονικές. Στόχος δεν ήταν μόνο να επιτευχθούν καλές αριθμητικές επιδόσεις, αλλά να παρατηρηθεί και ο τρόπος με τον οποίο κάθε μοντέλο «αντιλαμβάνεται» τα οπτικά χαρακτηριστικά των εικόνων. Ένα σημαντικό μέρος της εργασίας ήταν η χρήση τεχνικών εξηγησιμότητας. Σε προβλήματα ιατρικής εικόνας, ένα μοντέλο που δίνει μια σωστή πρόβλεψη χωρίς καμία ένδειξη για το πού βασίστηκε παραμένει δύσκολο να αξιολογηθεί ουσιαστικά. Για τον λόγο αυτό, εντάχθηκαν τεχνικές όπως Grad-CAM, Grad-CAM++, Score-CAM, LIME, SHAP, Occlusion, UMAP και t-SNE. Οι τεχνικές αυτές βοήθησαν να εξεταστεί αν τα μοντέλα εστιάζουν σε περιοχές που σχετίζονται με τη βλάβη ή αν επηρεάζονται από άσχετα στοιχεία της εικόνας, όπως το φόντο, τα όρια της εικόνας ή τεχνικά χαρακτηριστικά της λήψης. Αυτό το σημείο είναι ιδιαίτερα κρίσιμο, γιατί ένα μοντέλο μπορεί να εμφανίζει καλές μετρικές, αλλά οι χάρτες εξηγησιμότητας να δείχνουν ότι η απόφασή του δεν στηρίζεται πάντα σε κλινικά χρήσιμες περιοχές. Από τα πειράματα φάνηκε ότι τα μοντέλα βαθιάς μάθησης μπορούν να πετύχουν ικανοποιητική απόδοση στην ταξινόμηση δερματικών βλαβών, ειδικά όταν τα δεδομένα έχουν οργανωθεί σωστά και όταν εφαρμόζονται κατάλληλες τεχνικές προεπεξεργασίας. Παράλληλα, η σύγκριση διαφορετικών αρχιτεκτονικών έδειξε ότι η υψηλή απόδοση δεν εξαρτάται μόνο από το μέγεθος ή την πολυπλοκότητα του μοντέλου. Στην πράξη, σημαντικό ρόλο παίζουν η ποιότητα του dataset, η ισορροπία των κατηγοριών, η στρατηγική εκπαίδευσης, η επιλογή των υπερπαραμέτρων και η σωστή αξιολόγηση σε δεδομένα που δεν έχει ξαναδεί το μοντέλο. Με άλλα λόγια, ένα πιο σύνθετο μοντέλο δεν οδηγεί πάντα αυτόματα σε πιο αξιόπιστο αποτέλεσμα. Ιδιαίτερη σημασία είχε και η ανάλυση

των οπτικών αποτελεσμάτων από τις τεχνικές εξηγησιμότητας. Σε ορισμένες περιπτώσεις οι χάρτες ενεργοποίησης ανέδειξαν περιοχές που σχετίζονταν με τη βλάβη, κάτι που ενισχύει την εμπιστοσύνη προς το μοντέλο. Σε άλλες περιπτώσεις, όμως, οι εικόνες εξηγησιμότητας δεν ήταν τόσο καθαρές ή έδειχναν διάχυτη προσοχή σε περιοχές που δεν είχαν άμεση σχέση με τη δερματική αλλοίωση. Αυτό δείχνει ότι η εξηγησιμότητα δεν πρέπει να χρησιμοποιείται απλώς ως οπτικό συμπλήρωμα στο τέλος της πειραματικής διαδικασίας. Αντίθετα, μπορεί να λειτουργήσει ως εργαλείο ελέγχου, βοηθώντας τον ερευνητή να καταλάβει πότε ένα μοντέλο μαθαίνει ουσιαστικά χαρακτηριστικά και πότε πιθανώς βασίζεται σε επιφανειακά μοτίβα.

Συνολικά, η εργασία ανέδειξε ότι η ταξινόμηση δερματοσκοπικών εικόνων είναι ένα σύνθετο πρόβλημα, όπου οι καλές μετρικές δεν αρκούν από μόνες τους. Η σύνδεση της απόδοσης με την εξηγησιμότητα δίνει πιο ολοκληρωμένη εικόνα για την ποιότητα ενός μοντέλου. Μέσα από αυτή τη διαδικασία έγινε φανερό ότι η τεχνητή νοημοσύνη μπορεί να έχει ουσιαστικό ρόλο στην υποστήριξη της δερματολογικής αξιολόγησης, αρκεί να συνοδεύεται από προσεκτική πειραματική τεκμηρίωση, σωστό έλεγχο των δεδομένων και ερμηνεία των αποτελεσμάτων με κριτική στάση.

6.1 Συνεισφορά της εργασίας

Η βασική συνεισφορά της εργασίας βρίσκεται στον συνδυασμό τριών στοιχείων: θεωρητικής μελέτης, πειραματικής αξιολόγησης και εξηγησιμότητας. Η εργασία δεν περιορίστηκε μόνο στην παρουσίαση μοντέλων από τη βιβλιογραφία, αλλά προχώρησε και σε πρακτικές δοκιμές με διαφορετικές αρχιτεκτονικές και datasets. Μέσα από αυτή τη διαδικασία έγινε προσπάθεια να φανεί πώς μεταβάλλεται η απόδοση των μοντέλων ανάλογα με τα δεδομένα, την αρχιτεκτονική και τον τρόπο αξιολόγησης. Παράλληλα, η ενσωμάτωση τεχνικών εξηγησιμότητας πρόσθεσε ένα πιο ουσιαστικό επίπεδο ανάλυσης. Αντί να παρουσιαστούν μόνο αριθμητικά αποτελέσματα, εξετάστηκε και το κατά πόσο τα μοντέλα εστιάζουν σε περιοχές που σχετίζονται με τη βλάβη. Αυτή η διάσταση είναι ιδιαίτερα χρήσιμη σε εφαρμογές ιατρικής εικόνας, όπου η αξιοπιστία ενός συστήματος δεν μπορεί να κριθεί μόνο από ένα ποσοστό επιτυχίας. Στην πράξη, η εργασία δείχνει ότι η εξηγησιμότητα μπορεί να βοηθήσει στην καλύτερη κατανόηση των αποφάσεων του μοντέλου και να αναδείξει πιθανά προβλήματα που δεν φαίνονται άμεσα από τις μετρικές.

6.2 Μελλοντικές κατευθύνσεις

Με βάση τα αποτελέσματα και τους περιορισμούς της παρούσας εργασίας, υπάρχουν αρκετές κατευθύνσεις για μελλοντική έρευνα. Μια πρώτη κατεύθυνση θα ήταν η εκπαίδευση και αξιολόγηση των μοντέλων σε μεγαλύτερα και πιο ετερογενή datasets, ώστε να ελεγχθεί καλύτερα η ικανότητα γενίκευσης. Θα είχε ενδιαφέρον, για παράδειγμα, να δοκιμαστούν τα ίδια μοντέλα σε δεδομένα από διαφορετικές πηγές και να εξεταστεί αν η απόδοσή τους παραμένει σταθερή. Μια δεύτερη κατεύθυνση αφορά τη βαθύτερη διερεύνηση υβριδικών αρχιτεκτονικών, οι οποίες συνδυάζουν τα πλεονεκτήματα των συνελκτικών νευρωνικών δικτύων με τη δυνατότητα των μετασχηματιστών να μοντελοποιούν σχέσεις σε μεγαλύτερη κλίμακα μέσα στην εικόνα. Τέτοιες προσεγγίσεις μπορεί να είναι ιδιαίτερα χρήσιμες σε δερματοσκοπικές εικόνες, όπου η διάγνωση δεν εξαρτάται μόνο από μικρές τοπικές λεπτομέρειες, αλλά και από τη συνολική μορφή, την

κατανομή του χρώματος και τα όρια της βλάβης. Μια τρίτη πιθανή κατεύθυνση είναι η πιο συστηματική αξιολόγηση των τεχνικών εξηγησιμότητας. Οι χάρτες Grad-CAM, SHAP, LIME και Occlusion μπορούν να προσφέρουν χρήσιμες ενδείξεις, αλλά χρειάζεται προσοχή στην ερμηνεία τους. Σε μελλοντική εργασία θα μπορούσαν να συγκριθούν οι εξηγήσεις του μοντέλου με πραγματικές μάσκες βλαβών ή με αξιολογήσεις ειδικών, ώστε να φανεί αν οι περιοχές που εντοπίζει το μοντέλο έχουν και κλινική σημασία. Τέλος, θα μπορούσε να εξεταστεί η ανάπτυξη ενός πιο ολοκληρωμένου συστήματος υποστήριξης απόφασης, το οποίο θα συνδυάζει την πρόβλεψη του μοντέλου, τις μετρικές εμπιστοσύνης και τις οπτικές εξηγήσεις σε ένα ενιαίο περιβάλλον. Ένα τέτοιο σύστημα δεν θα είχε ως στόχο να αντικαταστήσει τον γιατρό, αλλά να λειτουργήσει ως πρόσθετο εργαλείο πληροφόρησης.

Βιβλιογραφία

- [1] International Agency for Research on Cancer. (2022). Global Cancer Observatory: Cancer today. World Health Organization. <https://gco.iarc.fr/today>
- [2] Arnold, M., Singh, D., Laversanne, M., Vignat, J., Vaccarella, S., Meheus, F., Cust, A. E., de Vries, E., Whiteman, D. C., & Bray, F. (2022). Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatology*, 158(5), 495–503. <https://doi.org/10.1001/jamadermatol.2022.0160>
- [3] Guy, G. P., Jr., Machlin, S. R., Ekwueme, D. U., & Yabroff, K. R. (2015). Vital signs: Melanoma incidence and mortality trends and projections — United States, 1982–2030. *Morbidity and Mortality Weekly Report*, 64(21), 591–596.
- [4] American Cancer Society. (2026). Survival rates for melanoma skin cancer. American Cancer Society. <https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stage.html>
- [5] Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. (2002). Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3), 159–165. [https://doi.org/10.1016/S1470-2045\(02\)00679-4](https://doi.org/10.1016/S1470-2045(02)00679-4)
- [6] Abbasi, N. R., Shaw, H. M., Rigel, D. S., Friedman, R. J., McCarthy, W. H., Osman, I., Kopf, A. W., & Polsky, D. (2004). Early diagnosis of cutaneous melanoma: Revisiting the ABCD criteria. *JAMA*, 292(22), 2771–2776. <https://doi.org/10.1001/jama.292.22.2771>
- [7] Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), Article e0118432. <https://doi.org/10.1371/journal.pone.0118432>
- [8] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
- [9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*.
- [10] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28
- [11] Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., & Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv*. <https://arxiv.org/abs/1902.03368>
- [12] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). <https://doi.org/10.1109/ICCV.2017.74>
- [13] Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 839–847). <https://doi.org/10.1109/WACV.2018.00097>
- [14] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings*

- of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 24–25). <https://doi.org/10.1109/CVPRW50498.2020.00020>
- [15] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30.
- [16] Azhari, A. A., Yudistira, N., Widodo, A. W., & Yagi, Y. (2025). Two-stage CNN with weakly supervised segmentation for skin lesion classification. *Multimedia Tools and Applications*, 84(41), 49769–49800. <https://doi.org/10.1007/s11042-025-21091-8>
- [17] K, M. R., Krishna, G. S., Supriya, K., & Sorgile, M. (2025). LesionAid: Vision transformers-based skin lesion generation and classification – A practical review. *Multimedia Tools and Applications*, 84, 41405–41426. <https://doi.org/10.1007/s11042-025-20797-z>
- [18] Shah, S. A. H., Shah, S. T. H., Khaled, R., Buccoliero, A., Shah, S. B. H., Di Terlizzi, A., Di Benedetto, G., & Deriu, M. A. (2024). Explainable AI-Based Skin Cancer Detection Using CNN, Particle Swarm Optimization and Machine Learning. *Journal of Imaging*, 10(12), 332. <https://doi.org/10.3390/jimaging10120332>
- [19] Bie, Y., Luo, L., & Chen, H. (2024). MICA: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2), 837–845. <https://doi.org/10.1609/aaai.v38i2.27842>
- [20] ThangaPurni, J. S., & Braveen, M. (2025). Unified ARP-ViT-CNN system: Hybrid deep learning approach for segmenting and classifying multiple skin cancer lesions. *Array*, 28, 100515. <https://doi.org/10.1016/j.array.2025.100515>
- [21] Pandala, M. L., & Periyayagi, S. (2026). Optimal explainable vision transformer framework for skin cancer diagnosis with neural architecture search feature learning. *Biomedical Signal Processing and Control*, 112, 108723. <https://doi.org/10.1016/j.bspc.2025.108723>
- [22] Mahir, S. H., Tashrif, M. T. A., Shathe, A. S., Karim, M. A., Rahman, A., Kundu, D., Debnath, T., & Muhammad, G. (2026). PriFL-XAI: Hybrid privacy-preserving federated learning models for

- monkeypox detection through GAN augmentation and explainable AI. *Biomedical Signal Processing and Control*, 112, 108426. <https://doi.org/10.1016/j.bspc.2025.108426>
- [23] Babu, R. R., & Philip, F. M. (2024). Optimized deep learning for skin lesion segmentation and skin cancer detection. *Biomedical Signal Processing and Control*, 95, 106292. <https://doi.org/10.1016/j.bspc.2024.106292>
- [24] Ahmad, I., Amin, J., Lali, M. I. U., Abbas, F., & Sharif, M. I. (2024). A novel Deeplabv3+ and vision-based transformer model for segmentation and classification of skin lesions. *Biomedical Signal Processing and Control*, 92, 106084. <https://doi.org/10.1016/j.bspc.2024.106084>
- [25] Nakai, K., Chen, Y.-W., & Han, X.-H. (2022). Enhanced deep bottleneck transformer model for skin lesion classification. *Biomedical Signal Processing and Control*, 78, 103997. <https://doi.org/10.1016/j.bspc.2022.103997>
- [26] Reis, H. C., & Turk, V. (2024). Fusion of transformer attention and CNN features for skin cancer detection. *Applied Soft Computing*, 164, 112013. <https://doi.org/10.1016/j.asoc.2024.112013>
- [27] Aruk, I., Pacal, I., & Toprak, A. N. (2026). A comprehensive comparison of convolutional neural network and visual transformer models on skin cancer classification. *Computational Biology and Chemistry*, 120(Pt 2), 108713. <https://doi.org/10.1016/j.compbiolchem.2025.108713>
- [28] Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., & Wen, Z. (2022). FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis*, 76, 102327. <https://doi.org/10.1016/j.media.2021.102327>
- [29] Aruk, I., Pacal, I., & Toprak, A. N. (2025). A novel hybrid ConvNeXt-based approach for enhanced skin lesion classification. *Expert Systems with Applications*, 283, 127721. <https://doi.org/10.1016/j.eswa.2025.127721>
- [30] Ruga, T., Caroprese, L., Vocaturo, E., & Zumpano, E. (2026). MultiExCam: A multi approach and explainable artificial intelligence architecture for skin lesion classification. *Computer Methods and Programs in Biomedicine*, 273, 109081. <https://doi.org/10.1016/j.cmpb.2025.109081>
- [31] He, X., Tan, E.-L., Bi, H., Zhang, X., Zhao, S., & Lei, B. (2022). Fully transformer network for skin lesion analysis. *Medical Image Analysis*, 77, 102357. <https://doi.org/10.1016/j.media.2022.102357>
- [32] Ieracitano, C., Morabito, F. C., Hussain, A., Suffian, M., & Mammone, N. (2025). TIXAI: A Trustworthiness Index for eXplainable AI in skin lesions classification. *Neurocomputing*, 630, Article 129701. <https://doi.org/10.1016/j.neucom.2025.129701>
- [33] Batista, L. G., Bugatti, P. H., & Saito, P. T. M. (2022). Classification of skin lesion through active learning strategies. *Computer Methods and Programs in Biomedicine*, 226, 107122. <https://doi.org/10.1016/j.cmpb.2022.107122>
- [34] Xin, C., Liu, Z., Ma, Y., Wang, D., Zhang, J., Li, L., Zhou, Q., Xu, S., & Zhang, Y. (2024). Transformer guided self-adaptive network for multi-scale skin lesion image segmentation. *Computers in Biology and Medicine*, 169, 107846. <https://doi.org/10.1016/j.compbiomed.2023.107846>
- [35] Zhang, Y., Xie, F., & Chen, J. (2023). TFormer: A throughout fusion transformer for multi-modal skin lesion diagnosis. *Computers in Biology and Medicine*, 157, 106712. <https://doi.org/10.1016/j.compbiomed.2023.106712>
- [36] Xin, C., Liu, Z., Zhao, K., Miao, L., Ma, Y., Zhu, X., Zhou, Q., Wang, S., Li, L., Yang, F., Xu, S., & Chen, H. (2022). An improved transformer network for skin cancer classification. *Computers in Biology and Medicine*, 149, 105939. <https://doi.org/10.1016/j.compbiomed.2022.105939>
- [37] Rahman, Z., Hossain, M. S., Islam, M. R., Hasan, M. M., & Hridhee, R. A. (2021). An approach for multiclass skin lesion classification based on ensemble learning. *Informatics in Medicine Unlocked*, 25, 100659. <https://doi.org/10.1016/j.imu.2021.100659>
- [38] Ul Haq, I., AlZubi, A. A., Khan, A. A., Shi, X., Alsayaydeh, J. A. J., Yusof, M. F., Ibrahim, M. bin M., Almoghrabi, H., & Joarder, H. A. (2026). XAAI-ledger: An explainable CNN-transformer-based multi-modal deep learning framework for early detection of melanoma and non-melanoma

- skin cancers using dermoscopic and clinical data. *Biomedical Signal Processing and Control*, 123, Article 110410. <https://doi.org/10.1016/j.bspc.2026.110410>
- [39] Thaljaoui, A., Yousafzai, S. N., Nasir, I. M., Saidani, O., Fadhal, E., & Saidani, T. (2026). Explainable skin cancer diagnosis with parallel attention mechanism for segmentation and classification. *Biomedical Signal Processing and Control*, 113, Article 109159. <https://doi.org/10.1016/j.bspc.2025.109159>
- [40] Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., & Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). *arXiv*. <https://arxiv.org/abs/1902.03368>
- [41] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, Article 180161. <https://doi.org/10.1038/sdata.2018.161>
- [42] Mendonça, T., Ferreira, P. M., Marques, J. S., Marçal, A. R. S., & Rozeira, J. (2013). PH²: A dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5437–5440). <https://doi.org/10.1109/EMBC.2013.6610779>
- [43] Kawahara, J., Daneshvar, S., Argenziano, G., & Hamarneh, G. (2019). Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2), 538–546. <https://doi.org/10.1109/JBHI.2018.2824327>
- [44] International Skin Imaging Collaboration. (2019). ISIC 2019 challenge: Skin lesion analysis towards melanoma detection. ISIC Archive. <https://challenge.isic-archive.com/landing/2019/>
- [45] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Liopyris, K., Malvey, J., Musthaq, S., Nanda, J., Reiter, O., Shih, G., Stratigos, A., Tschandl, P., Weber, J., & Soyer, H. P. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data*, 8, Article 34. <https://doi.org/10.1038/s41597-021-00815-z>
- [46] Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. A. C., Okata-Karigane, U., Zou, J., & Chiou, A. S. (2022). SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained model debugging and analysis. *Advances in Neural Information Processing Systems*, 35, 18157–18167.
- [47] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ‘Why should I trust you?’ Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
- [48] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014* (pp. 818–833). Springer. https://doi.org/10.1007/978-3-319-10590-1_53
- [49] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- [50] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*. <https://arxiv.org/abs/1802.03426>
- [51] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [52] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1143).