ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ-ΙΑΤΡΙΚΗ ΣΧΟΛΗ
Α΄ ΚΛΙΝΙΚΗ ΑΦΡΟΔΙΣΙΩΝ ΚΑΙ ΔΕΡΜΑΤΙΚΩΝ ΝΟΣΩΝ
ΝΟΣΟΚΟΜΕΙΟ **«ΑΝΔΡΕΑΣ ΣΥΓΓΡΟΣ»**
ΔΙΕΥΘΥΝΤΗΣ : ΚΑΘΗΓΗΤΗΣ **ΑΛΕΞΑΝΔΡΟΣ Ι. ΣΤΡΑΤΗΓΟΣ**

The role of Artificial Intelligence in early Skin Cancer Diagnosis

Μελέτη της χρήσης της Τεχνητής Νοημοσύνης (Artificial Intelligence) στην έγκαιρη διάγνωση του καρκίνου δέρματος

Konstantinos Liopyris, MD

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΑΘΗΝΑ 2020

Memorial Sloan Kettering Cancer Center

PhD Thesis

The Role of Artificial Intelligence in early skin cancer diagnosis

Konstantinos Liopyris, MD

Memorial Sloan Kettering Cancer Center

Athens, 2020

Per Aspera Ad Astra*

Table of Contents

Acknowledgements

«Ἓν οἶδα ὅτι οὐδὲν οἶδα» - "I know that I know nothing", Socrates

Knowledge is a lifelong process, and medicine is a lifelong commitment. Realizing the amplitude of things that you do not know is a mesmerizing feeling, but at the same time, it is a constant motivation to always explore, always proceed, always inquire.

This thesis concludes an eight-year journey, from Athens to New York City and back again. A journey full of passion, full of knowledge, full of experiences; a journey that would have never been possible without the people that leaved their mark to it.

I would like to express my sincere gratitude to my supervisor, Dr. Alexandros Stratigos; his constant support, knowledge, faith in me, and motivation to always reach higher have all played an integral role in this thesis. More importantly, I would like to thank him because he was the reason why I chose to follow dermatology, and skin oncology in particular. As a medical student, seeing his approach to the patient with cancer, his personality and his attitude conformed both my career and subsequently, life path.

Colleagues, that turn into friends, that turn into family. Cristian Navarrete, Michael Marchetti and Dr. Ashfaq Marghoob, thank you could never be enough for what you have offered me. My personality, my stance towards science, towards medicine, my overall life philosophy has been shaped by the endless hours we spent together. You taught me everything I know, you made NYC a second home to me, and

our endless conversations have defined the way I am as a doctor and as a human being. You are a constant inspiration, and I can never repay that.

This journey, along with this research would never have been possible without the constant support and guidance of Dr. Alan Halpern. Not only did he show faith in me to accept me to Memorial Sloan Kettering Cancer Center, but throughout these years, he has been a role model, and a constant inspiration to always dream big, to always aim high, to always be the best that you can be.

Along this journey I met many brilliant and talented physicians; their approach to medicine and their overall way of thinking has shaped the way I practice medicine. Dr. Dimitrios Rigopoulos, Dr. Electra Nicolaidou, Dr. Stamatios Gregoriou, Dr. Aimilios Lallas, Dr. Zoe Apalla, and so many others, thank you for being the unique people and physicians that you are.

Finally, and most importantly, I would like to thank my family. Without them nothing would be possible; without them nothing would have meaning. Anastasia, Vasileios, Nancy, Orestis, Iraklis, you made me the human being I am, no words of gratitude can express the way I feel about you.

"An expert is a man who has made all the mistakes which can be made, in a narrow field" - Niels Bohr

I still have a lot of mistakes to make, but, judging by the endless list of mistakes I have already made, and keep on making, I think I am on the right track.

Abstract

Artificial Intelligence (AI) has been incorporated in a wide spectrum of our daily lives, and Medicine could be no exemption to it. A plurality of scientific articles has explored the application of AI in fields such as Ophthalmology and Radiology, while there are already FDA approved, AI applications, which are used in clinical practice. Through our research we explored: (1) Can AI be used in early skin cancer diagnosis? (2) Which are the pitfalls of AI algorithms in Dermatology and in which possible ways could they be improved upon? (3) How could AI be of use in Dermatology, and in which ways could it be used? (4) Which is the best approach to the research conducted with regards to AI algorithms in early skin cancer diagnosis, and how should these results be interpreted? And (5), how can we improve the diagnostic accuracy of clinicians and AI algorithms for early skin cancer diagnosis and more specifically, melanoma? Finally, we attempted to describe the overall framework, within which, AI algorithms could be proven useful in clinical practice, and more importantly, beneficial to the patients.

Περίληψη

Η Τεχνητή Νοημοσύνη έχει ενσωματωθεί σε πολλές πτυχές της καθημερινότητας μας και η Ιατρική δεν θα μπορούσε να αποτελέσει εξαίρεση. Πλειάδα επιστημονικών άρθρων έχουν εξερευνήσει την εφαρμογή της Τεχνητής Νοημοσύνης σε τομείς όπως η Οφθαλμολογία και η Ακτινολογία, ενώ υπάρχουν ήδη εγκεκριμένες εφαρμογές Τεχνητής Νοημοσύνης οι οποίες χρησιμοποιούνται στην καθημερινή κλινική πράξη. Κατά τη διάρκεια της διδακτορικής μας διατριβής επιχειρήσαμε να διερευνήσουμε (1) Το κατά πόσον μπορεί η Τεχνητή Νοημοσύνη να χρησιμοποιηθεί στην έγκαιρη διάγνωση του καρκίνου δέρματος; (2) Ποια είναι τα μειονεκτήματα των εφαρμογών της Τεχνητής Νοημοσύνης στη δερματολογία μέχρι στιγμής και πως μπορούν να βελτιωθούν; (3) Που θα μπορούσε δυνητικά να βρει εφαρμογή η Τεχνητή Νοημοσύνη στη δερματολογία και με ποιο τρόπο θα ήταν χρήσιμη; (4) Ποια είναι η βέλτιστη προσέγγιση και ερμηνεία των αποτελεσμάτων που παράγονται τόσο από την έρευνα, όσο και από την εφαρμογή της Τεχνητής Νοημοσύνης στη δερματολογία; Και (5) Πως μπορούμε να βελτιώσουμε τη διαγνωστική ακρίβεια τόσο των κλινικών ιατρών, όσο και των αλγορίθμων Τεχνητής Νοημοσύνης για την έγκαιρη διάγνωση του καρκίνου δέρματος και ειδικότερα του μελανώματος; Τέλος, επιχειρήσαμε να περιγράψουμε το πλαίσιο, εντός του οποίου, οι αλγόριθμοι Τεχνητής Νοημοσύνης θα μπορούσαν να φανούν χρήσιμοι στην κλινική πράξη, προς όφελος των ασθενών.

Εισαγωγή

Η Τεχνητή Νοημοσύνη (Artificial Intelligence - AI) είναι ένα ταχέως αναπτυσσόμενο πεδίο έρευνας το οποίο συμπεριλαμβάνει ένα ευρύ φάσμα τομέων της καθημερινότητας. Το φάσμα αυτό εκτείνεται από την αναγνώριση φωνητικών εντολών έως την αυτοματοποίηση της εκπαίδευσης, τη βελτιστοποίηση των μεταφορών και του τομέα των ανανεώσιμων πηγών ενέργειας.[1-3] Η ιατρική δεν θα μπορούσε να είναι εξαίρεση· μια πλειάδα πρόσφατων επιστημονικών άρθρων διερευνούν τις εφαρμογές που θα μπορούσε να έχει το AI σε τομείς όπως η ακτινολογία, η ιστοπαθολογία, η οφθαλμολογία και άλλοι.[4-6]

Ο όρος 'Τεχνητή Νοημοσύνη' (AI) αναφέρεται στην ανάπτυξη αυτοματοποιημένων υπολογιστικών συστημάτων τα όποια έχουν τη δυνατότητα να επιτελούν δράσεις οι οποίες κανονικά απαιτούν ανθρώπινη νοημοσύνη, όπως η οπτική αντίληψη, η αναγνώριση ομιλίας, η λήψη αποφάσεων και η μετάφραση μεταξύ διαφόρων γλωσσών.[7] Ένας τομέας του AI είναι το "machine learning" (ML), ο οποίος αφορά στην επιστημονική μελέτη αλγορίθμων και στατιστικών μοντέλων που χρησιμοποιούνται από υπολογιστές προκειμένου να επιτελέσουν μια συγκεκριμένη λειτουργία, χωρίς να χρησιμοποιήσουν συγκεκριμένες οδηγίες ως προς την επιτέλεση της λειτουργίας αυτής. Οι αλγόριθμοι ML χτίζουν ένα μαθηματικό μοντέλο βασισμένο σε δεδομένα τα οποία έχουν χρησιμοποιηθεί για την εκπαίδευση τους, προκειμένου να κάνουν προβλέψεις ή να λάβουν αποφάσεις, χωρίς να είναι συγκεκριμένα προγραμματισμένοι για αυτό το σκοπό.[8] Αυτοί οι αλγόριθμοι ML είναι εκείνοι που έχουν βρει και τη μεγαλύτερη εφαρμογή στην ιατρική, ενώ τα τελευταία χρόνια, ο Αμερικάνικος Οργανισμός Τροφίμων και Φαρμάκων (FDA) έχει δώσει την έγκριση του σε 64 αλγορίθμους τεχνητής νοημοσύνης προκειμένου να χρησιμοποιηθούν στην κλινική πράξη και οι 29 εξ αυτών είναι αλγόριθμοι ML (45%).[9] Η

συντριπτική πλειοψηφία των αλγορίθμων αυτών αφορούν σε εφαρμογές στην ακτινολογία, την οφθαλμολογία, την παθολογία και την επείγουσα ιατρική.

Σε αυτά τα πλαίσια, προσπαθήσαμε να διερευνήσουμε τις εφαρμογές που θα μπορούσαν να έχουν οι αλγόριθμοι AI στη διάγνωση του καρκίνου δέρματος καθώς και να διευκολύνουμε την έρευνα στον τομέα της δερματολογικής ογκολογίας. Ωστόσο, η έλλειψη μεγάλων, δημόσιων βάσεων δεδομένων έχει περιορίσει την πρόοδο των αλγορίθμων deep learning για τη διάγνωση καρκίνου δέρματος. Συνεπακόλουθα, μέχρι στιγμής, κανένας αλγόριθμος δεν έχει χρησιμοποιηθεί αποτελεσματικά στην καθημερινή κλινική πράξη. Προκειμένου να υπερκεράσουμε αυτά τα εμπόδια δημιουργήσαμε το International Skin Imaging Collaboration Archive (ISIC - www.isic-archive.com) το οποίο είναι ένα ανοιχτό αρχείο δερματολογικών εικόνων (open source) οι οποίες είναι διαθέσιμες τόσο για εκπαιδευτικούς, όσο και για ερευνητικούς σκοπούς.[10] Το ISIC Archive μπορεί να χρησιμοποιηθεί από κλινικούς ιατρούς και από ερευνητές τεχνητής νοημοσύνης και έχει χρησιμοποιηθεί μέχρι στιγμής σε περισσότερες από 5000 δημοσιεύσεις.[10]

Ο λόγος που ασχοληθήκαμε με τον καρκίνο δέρματος έγκειται στο γεγονός ότι ο καρκίνος δέρματος είναι η πλέον συχνή μορφή καρκίνου με περισσότερα από 5 εκατομμύρια περιστατικά το χρόνο στις ΗΠΑ μόνο, ενώ το μελάνωμα είναι η 5η συχνότερη αιτία θανάτου από καρκίνο, με περισσότερους από 9000 θανάτους ετησίως.[11-14] Τα τελευταία χρόνια έχει καταγραφεί σημαντική πρόοδος τόσο στην επιβίωση, όσο και στην ποιότητα ζωής των ανθρώπων με μεταστατικό μελάνωμα και τοπικά προχωρημένο ή μεταστατικό βασικοκυτταρικό καρκίνωμα (BCC) ή ακανθοκυτταρικό καρκίνωμα (SCC) χάρη στις νέες συστηματικές θεραπείες, οι οποίες περιλαμβάνουν τα immune checkpoint inhibitors και τους εκλεκτικούς αναστολείς μονοπατιών που εμπλέκονται στην καρκινογένεση, όπως οι αναστολείς του μονοπατιού “Hedgehog”.[15-19] Ωστόσο, παρά τη μεγάλη πρόοδο που έχει καταγραφεί στις

συστηματικές θεραπείες, το κυριότερο όπλο μας στην αντιμετώπιση της θνητότητας και της νοσηρότητας του καρκίνου δέρματος παραμένει η έγκαιρη διάγνωση και η χειρουργική εξαίρεση του όγκου.[19-21]

Σε αυτή την κατεύθυνση έχει συνδράμει σημαντικά τα τελευταία 20 χρόνια η δερματοσκόπηση. Η δερματοσκόπηση είναι μια ευρέως διαδεδομένη, μη επεμβατική, διαγνωστική μέθοδος, η οποία φαίνεται να υπερέχει έναντι της απλής κλινικής εξέτασης για την διάγνωση του καρκίνου δέρματος.[22,23] Η δερματοσκόπηση έχει γίνει αναπόσπαστο κομμάτι της δερματολογίας και θεωρείται πλέον «το στηθοσκόπιο των δερματολόγων», ενώ η χρήση της έχει επεκταθεί και σε άλλους τομείς της γενικής δερματολογίας, εκτός της δερματολογικής ογκολογίας.[24,25] Ένα επιπρόσθετο πλεονέκτημα της δερματοσκόπησης είναι πως οι δερματοσκοπικές εικόνες λαμβάνονται σε απευθείας επαφή με το δέρμα, με standardized μεγέθυνση 10x, καθιστώντας κατ' αυτό τον τρόπο μη αναγνωρίσιμες τις βλάβες δέρματος οι οποίες χρησιμοποιούνται στο ISIC Archive, και προστατεύοντας τα προσωπικά δεδομένα των ασθενών.

Μέσω της έρευνας μας επιχειρήσαμε να διαλευκάνουμε περισσότερο τα εξής: (1) Το κατά πόσον μπορεί η Τεχνητή Νοημοσύνη να χρησιμοποιηθεί στην έγκαιρη διάγνωση του καρκίνου δέρματος; (2) Ποια είναι τα μειονεκτήματα των εφαρμογών της Τεχνητής Νοημοσύνης στη δερματολογία μέχρι στιγμής και πως μπορούν να βελτιωθούν; (3) Που θα μπορούσε δυνητικά να βρει εφαρμογή η Τεχνητή Νοημοσύνη στη δερματολογία και με ποιο τρόπο θα ήταν δυνητικά χρήσιμη; (4) Ποια είναι η βέλτιστη προσέγγιση και ερμηνεία των αποτελεσμάτων που παράγονται τόσο από την έρευνα, όσο και από την εφαρμογή της Τεχνητής Νοημοσύνης στη δερματολογία; Και τέλος, (5) Πως μπορούμε να βελτιώσουμε τη διαγνωστική ακρίβεια τόσο των κλινικών ιατρών, όσο και των αλγορίθμων Τεχνητής Νοημοσύνης για την έγκαιρη διάγνωση του καρκίνου δέρματος και ειδικότερα του μελανώματος;

Προκειμένου να επιτύχουμε τους στόχους αυτούς, κατά τη διάρκεια της διατριβής, σε συνεργασία με μια ομάδα διακεκριμένων ερευνητών από την Αμερική, την Ελλάδα, τη Χιλή, την Ισπανία, την Αυστραλία και την Αυστρία προχωρήσαμε σε μια σειρά ερευνητικών προσπαθειών. Μέσω του ISIC Archive διοργανώσαμε και εξακολουθούμε να διοργανώνουμε ετήσιους διαγωνισμούς, τα ISIC Grand Challenges, όπου ερευνητές αλγορίθμων Τεχνητής Νοημοσύνης διαγωνίζονται σε διαρκώς μεγαλύτερα datasets για τη διάγνωση όγκων δέρματος, διερευνώντας τη διαγνωστική ακρίβεια (lesion classification) των αλγορίθμων AI και συγκρίνοντας παράλληλα την επίδοση τους με αυτή των κλινικών ιατρών. Ταυτόχρονα, διερευνήσαμε τη δυνατότητα των αλγορίθμων AI να διαχωρίζουν τις βλάβες δέρματος, από το περιβάλλον φυσιολογικό δέρμα (lesion segmentation), αλλά και τη δυνατότητα τους να εντοπίζουν προεπιλεγμένα δερματοσκοπικά κριτήρια (dermoscopic feature detection).[26-29] Επιπλέον, επιχειρήσαμε να προσδιορίσουμε σε ποιους κλινικούς ιατρούς θα μπορούσε να φανούν χρήσιμες οι εφαρμογές Τεχνητής Νοημοσύνης, εξετάζοντας τη διαγνωστική ακρίβεια ιατρών σε διαφορετικά στάδια της εκπαίδευσης τους, συγκριτικά με τους αλγορίθμους, ή και με τη βοήθεια αυτών.[30] Σε ένα άλλο σκέλος της έρευνας μας, διερευνήσαμε την επίδοση δημοσιευμένων αλγορίθμων τεχνητής νοημοσύνης σε ένα δημόσια διαθέσιμο, standardized dataset καρκίνων δέρματος, προκειμένου να ανακαλύψουμε τυχόν αδυναμίες τους, και να βοηθήσουμε την ερευνητική κοινότητα στη βελτίωση των επιδόσεων τους.[31,32] Κατά τη διάρκεια της έρευνας μας εφαρμόσαμε εναλλακτικές στατιστικές προσεγγίσεις και ερμηνευτικές προσπάθειες, οι οποίες δεν είχαν χρησιμοποιηθεί προηγουμένως στον τομέα, προκειμένου να διασταυρώσουμε και να ερμηνεύσουμε τα αποτελέσματα τόσο της δικής μας έρευνας, όσο και αποτελέσματα άλλων ερευνητών.[33] Τέλος, επιχειρήσαμε να βελτιώσουμε τη διαγνωστική ακρίβεια των κλινικών ιατρών στη διάγνωση καρκίνου δέρματος, τόσο περιγράφοντας καινούργια δερματοσκοπικά κριτήρια για τη διάγνωση του καρκίνου δέρματος, και ειδικότερα δύσκολων στη διάγνωση καρκίνων δέρματος και μιμητών τους, όσο και διοργανώνοντας την πρώτη Expert Agreement Study on Dermoscopy μελανοκυτταρικών βλαβών (EASY study).[34-37] Η εν λόγω

μελέτη εξέτασε την συμφωνία ειδικών στη δερματοσκόπηση στην εντόπιση δερματοσκοπικών κριτηρίων εντός μιας βλάβης, προκειμένου να ελέγξουμε την αξιοπιστία και την αναπαραγωγιμότητα των δερματοσκοπικών κριτηρίων για τη διάγνωση του μελανώματος.

ISIC Grand Challenges και Reader Studies

Εισαγωγή

Όπως εκθέσαμε ανωτέρω, προκειμένου να διερευνήσουμε τις πιθανές εφαρμογές που μπορεί να έχει η Τεχνητή Νοημοσύνη στην έγκαιρη διάγνωση του καρκίνου δέρματος, καθώς και για να αντιμετωπίσουμε το έλλειμμα δημόσια διαθέσιμων βάσεων δεδομένων δερματολογικών, και ειδικότερα δερματοσκοπικών εικόνων, δημιουργήσαμε το ISIC Archive.[10] Μέσω αυτού διοργανώνουμε από το 2016 ετήσια ISIC Grand Challenges, με διαρκώς αυξανόμενο επίπεδο δυσκολίας, αριθμό διαγνώσεων και προκλήσεων στις οποίες διαγωνίζονται ερευνητές Τεχνητής Νοημοσύνης από όλο τον κόσμο, καθιστώντας το ISIC Archive το σημείο αναφοράς για την έρευνα στο πεδίο. Τα ISIC Grand Challenges πραγματοποιούνται ετησίως, στα πλαίσια προβεβλημένων, διεθνών συνεδρίων· ειδικότερα, στο International Symposium on Biomedical Imaging (ISBI, 2016-2017), Medical Image Computing and Computer Assisted Intervention (MICCAI, 2018-2020), στο Conference on Computer Vision and Pattern Recognition (CVPR, 2019-2020) και στο Society for Imaging Informatics (SIIM, 2019-2020). Στα πλαίσια της διατριβής αυτής, ανέλαβα το ρόλο του clinical coordinator του ISIC Archive συμμετέχοντας ενεργά σε όλους τους διαγωνισμούς οι οποίοι διοργανώθηκαν από το 2017 έως το 2020, ενώ από το 2020 το Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, μέσω του Νοσοκομείου Δερματικών και Αφροδισίων Νοσημάτων «Ανδρέας Συγγρός» είναι επίσημα τμήμα της συνεργασίας που έχει οικοδομηθεί μεταξύ του Memorial Sloan Kettering Cancer Center, του Medical University of Vienna, της Barcelona Hospital Clinic, του Emory University, του University of Queensland και του Melanoma Institute Australia and Sydney Melanoma Diagnostic Center.

Τα ISIC Grand Challenges αποτελούν διαγωνισμούς όπου ερευνητές Τεχνητής Νοημοσύνης καταθέτουν τους αλγορίθμους τους, οι οποίοι διαγωνίζονται στη διάγνωση καρκίνου δέρματος και μιμητών του. Οι διαγωνισμοί αυτοί ακολουθούνται από reader studies οι οποίες διεξάγονται ανάμεσα σε κλινικούς ιατρούς διαφορετικών ειδικοτήτων και διαφορετικών βαθμίδων εκπαίδευσης (δερματολόγοι, ειδικευόμενοι δερματολογίας, γενικοί ιατροί, ειδικευόμενοι γενικής ιατρικής), προκειμένου να διερευνηθεί (1) η διαγνωστική ακρίβεια των αλγορίθμων Τεχνητής Νοημοσύνης έναντι των κλινικών ιατρών, και (2) πού θα μπορούσαν οι αλγόριθμοι αυτοί να φανούν χρήσιμοι. Οι διαγωνισμοί αυτοί εμπλουτίζονται κάθε χρόνο με μεγαλύτερο αριθμό βλαβών, περισσότερες διαγνώσεις, καθώς και ποικίλες ερευνητικές προκλήσεις οι οποίες καθορίζουν την εξέλιξη του πεδίου, στην προσπάθεια να μπορέσουν οι αλγόριθμοι αυτοί να καταστούν χρήσιμοι στην κλινική πράξη.[26-29]

Στον πρώτο διαγωνισμό ISIC Grand Challenge 2016 οι αλγόριθμοι Τεχνητής Νοημοσύνης διαγωνίστηκαν σε 1279 δερματοσκοπικές βλάβες (248 μελανώματα και 1031 σπίλους) οι οποίες χωρίστηκαν σε ένα training (n=900, 19.2% μελανώματα) και ένα test dataset (n=379, 19.8% μελανώματα).[27] Στη reader study η οποία ακολούθησε, περιλαμβάνοντας 8 expert readers βρήκαμε πως ο fusion algorithm των κορυφαίων συμμετεχόντων αλγορίθμων είχε καλύτερη επίδοση από τους ειδικούς (ειδικότητα 76% έναντι 59% των experts, p=0.02). [38]

Εδώ θα παρουσιάσουμε τα αποτελέσματα του ISIC Grand Challenge 2017, όπου επεκτείναμε τον αριθμό των βλαβών και τις κατηγορίες βλαβών, συμπεριλαμβάνοντας μελανώματα, σπίλους και σμηγματορροϊκές υπερκερατώσεις. Ταυτόχρονα επεκτείναμε και τις δοκιμασίες στις οποίες διαγωνίστηκαν οι διαφορετικοί αλγόριθμοι για να συμπεριλάβουμε 3 tasks: (1) lesion segmentation, (2) dermoscopic feature detection και (3) classification.[28] Επιπλέον θα παρουσιάσουμε, τη reader study η οποία ακολούθησε το ISIC Challenge 2017, όπου συγκρίναμε τη διαγνωστική ακρίβεια του αλγορίθμου

τεχνητής νοημοσύνης με την καλύτερη επίδοση στη διάγνωση μελανώματος, με τη διαγνωστική ακρίβεια έμπειρων δερματολόγων, ειδικών στη διάγνωση και θεραπεία καρκίνου δέρματος, καθώς και ειδικευόμενων δερματολογίας. Στα πλαίσια αυτής της μελέτης διερευνήσαμε τα αποτελέσματα που θα είχε η ενδεχόμενη χρήση των διαγνωστικών προβλέψεων των αλγορίθμων από τους κλινικούς ιατρούς στις περιπτώσεις όπου η αυτοπεποίθηση τους (confidence) στη διάγνωση ήταν χαμηλή.[30]

ISIC Grand Challenge 2017

Υλικά και Μέθοδοι, ISIC Grand Challenge 2017

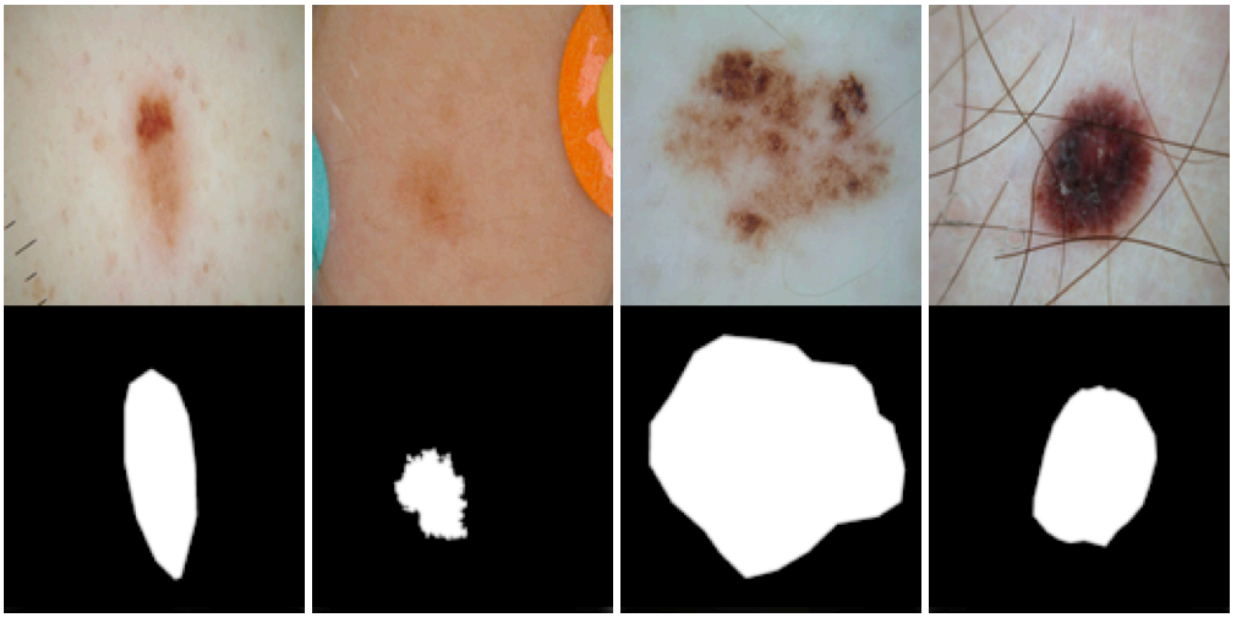
Το ISIC Grand Challenge 2017 αποτελούταν από 3 tasks: lesion segmentation, dermoscopic feature detection, και disease classification. Για κάθε ένα task, τα data αποτελούνταν από δερματοσκοπικές εικόνες και αντιστοιχούσαν σε ground truth annotations, χωρισμένα σε training set (n=2000), validation set (n=150) και test set (n=600) datasets. Οι προβλέψεις των διαγωνιζόμενων μπορούσαν να υποβληθούν τόσο στο validation, όσο και στο test dataset. Οι υποβολές στο validation dataset προμήθευαν τους διαγωνιζόμενους με άμεσο feedback στη μορφή αποτελέσματος επίδοσης, καθώς και πληροφορίες αναφορικά με την κατάταξη των διαγωνιζομένων, συγκριτικά με τους συνδιαγωνιζόμενους τους. Τα αποτελέσματα επιδόσεων στο test dataset έγιναν γνωστά στους συμμετέχοντες μόνο μετά τη λήξη της προθεσμίας του Challenge· τα datasets στα οποία εκπαιδεύτηκαν (training dataset) και διαγωνίστηκαν οι αλγόριθμοι (validation and test datasets) είναι διαθέσιμα στον ακόλουθο ιστότοπο: http://challenge2017.isic-archive.com/.

1ο σκέλος, Lesion segmentation task

Από τους διαγωνιζόμενους ζητήθηκε να υποβάλλουν αυτοματοποιημένες προβλέψεις των ορίων των βλαβών δέρματος (lesion segmentations) από δερματοσκοπικές εικόνες, σε μορφή binary masks (λευκό χρώμα η βλάβη - μαύρο χρώμα το περιβάλλον της εικόνας). Το training dataset για αυτό το σκέλος είχε δοθεί στους διαγωνιζόμενους με τη μορφή της πρωτότυπης εικόνας, καθώς και σε binary masks, μετά από μη-αυτοματοποιημένο καθορισμό των ορίων της βλάβης από 2 εκ των συγγραφέων (K.L., M.M.). Pixel values αξίας 255 θεωρούνταν εντός των ορίων της βλάβης, ενώ pixel values αξίας 0 θεωρούνταν εκτός. (**Figure 1)**

Figure 1.

Στο άνω σκέλος της εικόνας βλέπουμε τις πρωτότυπες δερματοσκοπικές εικόνες, και στο κάτω μέρος, τα binary segmentation masks, όπως δόθηκαν στους διαγωνιζόμενους στο training dataset. Με λευκό χρώμα απεικονίζονται τα όρια των βλαβών, ενώ με μαύρο, το περιβάλλον δέρμα, το οποίο δεν περιλαμβάνει κάποια βλάβη.

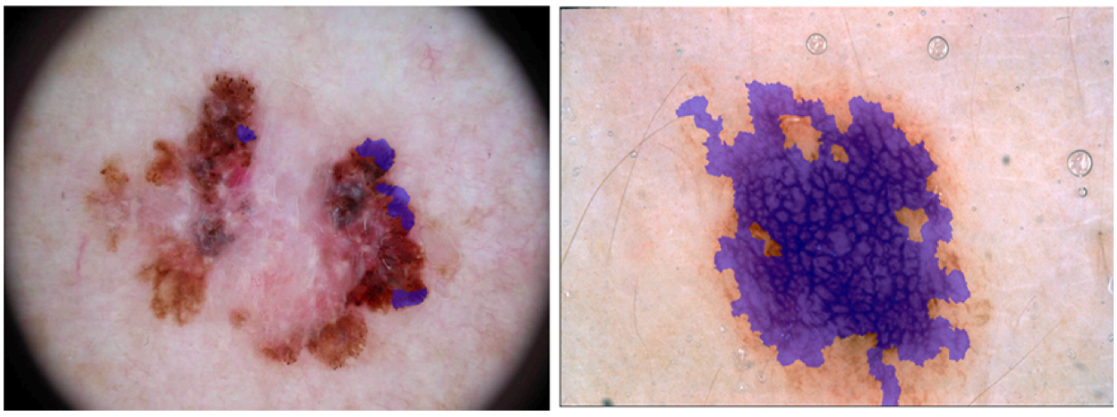


2ο Σκέλος, Dermoscopic Feature Classification Task

Από τους διαγωνιζόμενους ζητήθηκε να ταυτοποιήσουν αυτοματοποιημένα τα ακόλουθα τέσσερα δερματοσκοπικά κριτήρια στις δερματοσκοπικές εικόνες: “network”, “negative network”, “streaks” και “milia-like cysts”.[39-41] Η αναγνώριση των δερματοσκοπικών κριτηρίων αφορούσε τόσο στην ταξινόμηση τους, όσο και στην χωρική εντόπιση τους εντός της εικόνας. **(Figure 2)** Προκειμένου να μειώσουμε το variability και το dimensionality της χωρικής εντόπισης των δερματοσκοπικών κριτηρίων, οι δερματοσκοπικές εικόνες υποδιαιρέθηκαν σε ‘superpixels’, χρησιμοποιώντας τον αλγόριθμο SLIC.[42-44] Τα training data για αυτό το task συμπεριλάμβαναν την πρωτότυπη εικόνα, καθώς και τους συνοδούς αριθμούς μασκών - ‘superpixels’ όπου βρίσκονταν τα εν λόγω δερματοσκοπικά κριτήρια· ένας εκ των συγγραφέων (K.L.) είχε εκ των προτέρων υπογραμμίσει χωρικά ποια είναι και που βρίσκονται τα κριτήρια στις εικόνες. Τα validation και test sets περιελάμβαναν δερματοσκοπικές εικόνες και τα αντίστοιχα superpixels, χωρίς ωστόσο τα annotations.

Figure 2.

Α. Δερματοσκοπική εικόνα ενός μελανώματος με χωρικό καθορισμό superpixels όπου εντοπίζεται το δερματοσκοπικό κριτήριο “streaks”, το οποίο έχει βρεθεί να έχει αυξημένο Odds Ratio για τη διάγνωση μελανώματος, και Β. Δερματοσκοπική εικόνα ενός σπίλου με τα αντίστοιχα superpixels όπου εντοπίζεται το κριτήριο “network”

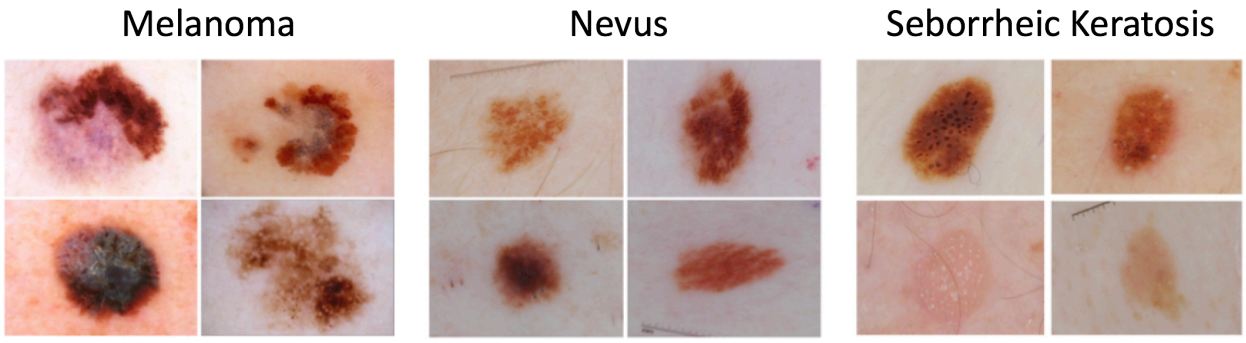


3ο Σκέλος, Disease Classification Task

Από τους διαγωνιζόμενους αιτήθηκε να ταξινομήσουν τις εικόνες σε μια εκ των 3 κατηγοριών συμπεριλαμβανομένων των: μελάνωμα (n=374 training, 30 validation, 117 test), σμηγματορροϊκή υπερκεράτωση (n=254, 42, and 90), και καλοήθης σπίλος (n=1372, 78, 393), με σκορ ταξινόμησης κυμαινόμενο μεταξύ 0.0 και 1.0 για κάθε κατηγορία, με το 0.5 να είναι το ορόσημο προκειμένου να θεωρηθεί ένα αποτέλεσμα θετικό (binary decision threshold). Τα training data για αυτό το σκέλος περιελάμβαναν τις δερματοσκοπικές εικόνες με το gold standard για κάθε διάγνωση (ιστοπαθολογική επιβεβαίωση για το μελάνωμα, ιστοπαθολογική επιβεβαίωση, ή επιβεβαίωση διάγνωσης από ειδικούς στη δερματοσκόπηση για τους σπίλους και τις σμηγματορροϊκές υπερκερατώσεις), καθώς και την ηλικία του ασθενούς κατά προσέγγιση 5ετίας, καθώς και το φύλο. **(Figure 3)**

Figure 3.

Αντιπροσωπευτικές εικόνες μελανωμάτων, σπίλων, και σμηγματορροϊκών υπερκερατώσεων που συμπεριλήφθηκαν στο dataset μας.



Αξιολόγηση αποτελεσμάτων

Οι παράμετροι που χρησιμοποιήσαμε για να αξιολογήσουμε την επίδοση των αλγορίθμων έχουν περιγραφεί αναλυτικά σε προηγούμενες μελέτες.[27,38] Για το σκέλος της ταξινόμησης των εικόνων σε διαφορετικές διαγνώσεις χρησιμοποιήσαμε το όριο του 0.5 ως θετικό για μια κατηγορία.

Για το 1ο σκέλος του lesion segmentation, pixel values >128 θεωρήθηκαν ως θετική απάντηση, και αξίες μικρότερες αυτού, θεωρήθηκαν αρνητικές. Προκειμένου να αξιολογήσουμε τα αποτελέσματα της ταξινόμησης ως προς τη διάγνωση, υπολογίσαμε την περιοχή κάτω από την καμπύλη (Area Under the Curve - AUC), από τα receiver operating characteristic curve (ROC).[27] Επιπρόσθετα, αναφορικά με την ταξινόμηση του μελανώματος, η ειδικότητα των αλγορίθμων υπολογίστηκε με βάση το ROC curve, όπου η ευαισθησία ήταν 82%, 89%, και 95%, αντιστοιχώντας στα επιθυμητά επίπεδα επίδοσης ενός δερματολόγου, εξειδικευμένου στον καρκίνο δέρματος.[38] Τέλος, οι επιδόσεις για το task 1, lesion segmentation, συγκρίθηκαν χρησιμοποιώντας το Jaccard Index, Dice coefficient και την ακρίβεια ως προς τα pixel.

Αποτελέσματα, ISIC Grand Challenge 2017

Στο ISIC Challenge 2017 είχαμε 593 εγγραφές, 81 πειραματικές προ-καταχωρήσεις αποτελεσμάτων, και 46 τελικές καταχωρήσεις αποτελεσμάτων (συμπεριλαμβανομένων και ισάριθμων δημοσιεύσεων στο arXiv.org) . Μέχρι εκείνη τη στιγμή, αυτή ήταν η μεγαλύτερη, προτυποποιημένη, συγκριτική μελέτη στον τομέα, περικλείοντας τον αριθμό των βλαβών που συμπεριλάβαμε, τον αριθμό των αλγορίθμων που εκτιμήσαμε και τον αριθμό των συμμετεχόντων. Ακολούθως παρατίθενται τα αποτελέσματα για κάθε ξεχωριστό σκέλος της έρευνας.

1ο Σκέλος, Lesion Segmentation

Εικοσιένα συμμετέχοντες κατέθεσαν τις προβλέψεις τους σε αυτό το σκέλος του διαγωνισμού για το test set ενώ 39 συμμετέχοντες συμμετείχαν στο validation set. Ο κορυφαίος των συμμετεχόντων πέτυχε μέσο Jaccard Index 0.765, ακρίβεια 93.4% και Dice coefficient 0.849, χρησιμοποιώντας μια παραλλαγή ενός fully convolutional network, ακολουθώντας μια προσέγγιση deep learning.[45]

Αντιπροσωπευτικά παραδείγματα καθορισμού ορίων (segmentations) ανευρίσκονται στο **Figure 4**, και ένα ιστόγραμμα του Jaccard index μιας μεμονωμένης εικόνας στο **Figure 5**. Segmentations τα οποία επιτυγχάνουν Jaccard Index >0.8 τείνουν να εμφανίζονται οπτικά ακριβή, αντίθετα, όταν πέφτει κάτω του 0.7, η ακρίβεια του καθορισμού των ορίων αμφισβητείται.[26] Στο εν λόγω task, το Jaccard Index του κορυφαίου αλγόριθμου αυτόματου segmentation έπεσε κάτω από το όριο του 0.7 σε 156 από τις 600 εικόνες, ενώ για 91 εικόνες, το Jaccard Index ήταν μικρότερο του 0.6, έχοντας ένα failure rate το οποίο κυμάνθηκε από 15% έως 26%.

Figure 4.

Στην πρώτη σειρά βλέπουμε πέντε από τις πρωτότυπες εικόνες στις οποίες έπρεπε να γίνει ο καθορισμός των ορίων, segmentation, στη 2η σειρά διακρίνουμε το ground truth των ορίων, όπως καθορίστηκαν από τους διοργανωτές του ISIC Challenge, και στην 3η σειρά διακρίνουμε τις προβλέψεις για τα όρια των εν λόγω βλαβών, όπως τα περιέγραψε ο κορυφαίος εκ των διαγωνιζόμενων αλγορίθμων.

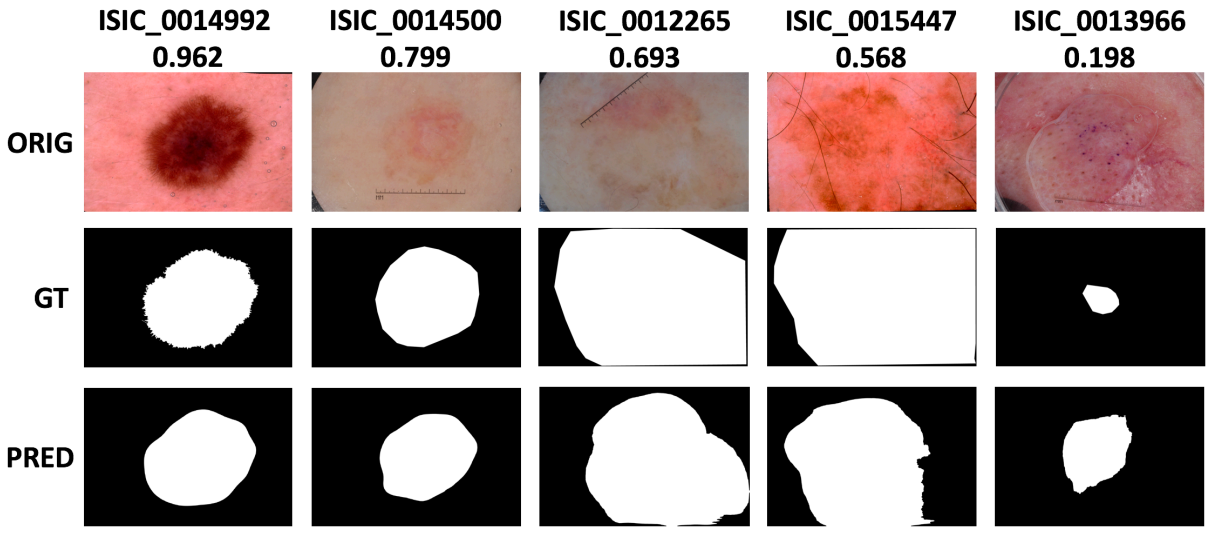
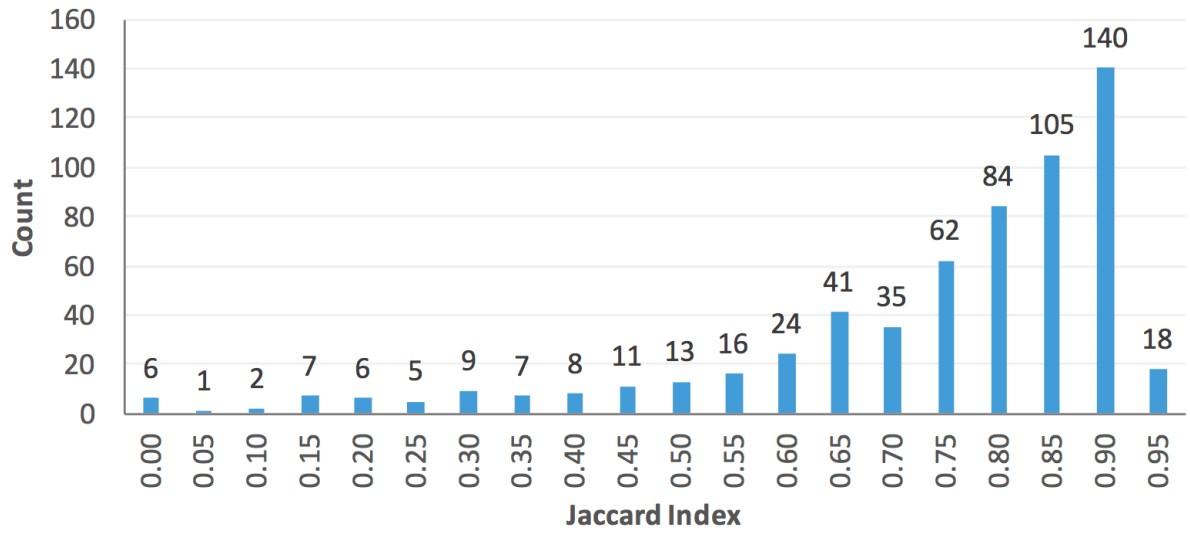


Figure 5

Παρατίθεται το ιστόγραμμα των αποτελεσμάτων βάσει του Jaccard Index για τον κορυφαίο αλγόριθμο.



2ο Σκέλος, Dermoscopic feature classification task

Για 2η συνεχή χρονιά, αυτό το task έλαβε τη μικρότερη συμμετοχή, συγκριτικά με τα υπόλοιπα tasks του ISIC Challenge· μόλις 3 τελικές υποβολές αποτελεσμάτων, από 2 ομάδες ελήφθησαν.[46,47] Το αν αυτή η μικρή συμμετοχή είναι αποτέλεσμα του τρόπου περιγραφής του task, ή αποτέλεσμα θεώρησης του ως μειωμένης σημασίας διερευνάται. Μολαταύτα, τα αποτελέσματα τα οποία λάβαμε καταδεικνύουν πως η χωρική εντόπιση δερματοσκοπικών κριτηρίων από αλγόριθμους ML δείχνουν πως είναι εφικτός στόχος, δείχνοντας μια μέση περιοχή κάτω από την καμπύλη (AUC) μεγαλύτερη από 0.75, και κατά σημεία να αγγίζει το 0.9. **(Table 1)**

Table 1.

Σε αυτό τον πίνακα βλέπουμε τις επιδόσεις των διαφορετικών μεθόδων που ακολούθησαν οι 2 ομάδες που διαγωνίστηκαν σε αυτό το σκέλος του ISIC Challenge. Με bold βλέπουμε τον νικητή για κάθε ξεχωριστό δερματοσκοπικό κριτήριο.[46,47]

Method / Rank	AVG	Network	Negative Network	Streaks	Milia-Like Cyst
1[29]	**0.895**	**0.945**	**0.869**	**0.960**	0.807
2[30]	0.833	0.835	0.762	0.896	**0.838**
3[30]	0.832	0.828	0.762	0.900	0.837

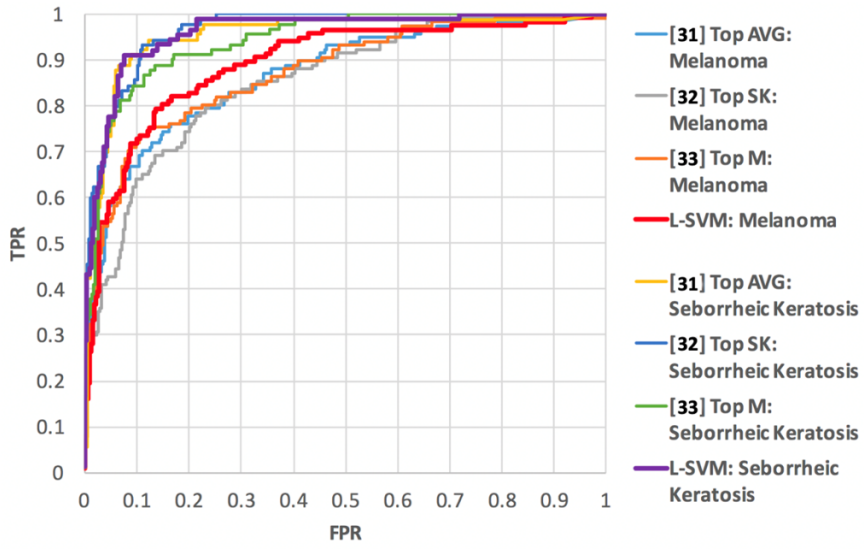
3ο Σκέλος, Ταξινόμηση Νόσου (Disease Classification)

Αυτό το σκέλος του διαγωνισμού έλαβε 23 τελικές υποβολές αποτελεσμάτων στο test dataset, και 39 υποβολές στο validation dataset. Τα χαρακτηριστικά της επίδοσης του τελικού νικητή αυτού του σκέλους του διαγωνισμού[48], καθώς και των νικητών των υπόλοιπων επιμέρους διαγνώσεων (σμηγματορροϊκές υπερκερατώσεις[49], μελανώματα[50]) βρίσκονται στον **Table 2.** Στον ίδιο Πίνακα βρίσκονται και οι 3 στρατηγικές fusion που ακολουθήσαμε, προκειμένου να εξάγουμε τα τελικά συμπεράσματα μας, τα score averaging (AVGSC), linear support vector machine (L-SVM) & non-linear support vector machine (NL-SVM).[38] Όλες οι στρατηγικές fusion που ακολουθήσαμε χρησιμοποίησαν ένα πιθανολογικό (probabilistic) μοντέλο, παράγοντας μια πιθανότητα η οποία κινούνταν μεταξύ του 0.0 και του 1.0 για κάθε ξεχωριστό νόσημα, όποια πρόβλεψη ξεπερνούσε το 0.5 θεωρούνταν θετική για την κατηγορία νόσου όπου προέκυψε. Οι ROC καμπύλες για τους νικητές αυτού του σκέλους του διαγωνισμού για τις 3 κατηγορίες νοσημάτων βρίσκονται στο **Figure 6.**

Τα κύρια ευρήματα μας περιλαμβάνουν τα εξής: 1) Όλοι οι κορυφαίοι συμμετέχοντες χρησιμοποίησαν κάποια παραλλαγή ενός deep learning network, ενώ χρησιμοποίησαν και εξωτερικά δεδομένα προκειμένου να εκπαιδεύσουν το μοντέλο τους, πλην του training dataset που τους παρείχαμε.[48-50] 2) Η ταξινόμηση των σμηγματορροϊκών υπερκερατώσεων φαίνεται να είναι ευκολότερο task συγκριτικά με τη διάγνωση μελανώματος, αυτό μπορεί να οφείλεται σε χαρακτηριστικά του νοσήματος, ή ενδεχομένως ενός bias στα dataset. 3) Ο συνολικός νικητής του διαγωνισμού δεν ήταν πρώτος σε επίδοση σε καμία από τις επιμέρους κατηγορίες, ωστόσο το μοντέλο τους παρήγαγε τα πιο αξιόπιστα αποτελέσματα. 4) Οι πιο περιπλεγμένες στρατηγικές fusion οδήγησαν σε πτώση της επίδοσης των αλγορίθμων, ενώ οι πιο απλές μέθοδοι παρήγαγαν τα πλέον αξιόπιστα αποτελέσματα.[38] Αυτή είναι η 2η μελέτη που εξερευνά τις δυνατότητες των αλγορίθμων τεχνητής νοημοσύνης στην ταξινόμηση δερματοσκοπικών εικόνων, και τα αποτελέσματα και των 2 μελετών συγκλίνουν πως η χρήση ενός fusion αλγόριθμου μεταξύ των επιμέρους κορυφαίων συμμετεχόντων έχει καλύτερη επίδοση συγκριτικά με κάθε μεμονωμένο αλγόριθμο.

Εικόνα 6.

ROC curves των κορυφαίων αλγορίθμων για κάθε κατηγορία,[48-50] και επίδοση του fusion αλγόριθμου.



Συζήτηση, ISIC Grand Challenge 2017

Το International Skin Imaging Collaboration (ISIC) Archive χρησιμοποιήθηκε προκειμένου να φιλοξενήσει τον 2ο δημόσιο διαγωνισμό για τη διάγνωση δερματολογικών νοσημάτων μέσω δερματοσκοπικών εικόνων στο International Symposium on Biomedical Imaging (ISBI) 2017. Ο διαγωνισμός περιλάμβανε 3 επιμέρους σκέλη: segmentation, feature selection (4 δερματοσκοπικά κριτήρια), και disease classification (3 νοσήματα). Ο διαγωνισμός περιέλαβε 2000 εικόνες για εκπαίδευση των αλγορίθμων (training dataset), 150 εικόνες για επιβεβαίωση αποτελεσμάτων (validation dataset), και 600 εικόνες στο test set, ενώ παραλάβαμε 593 εγγραφές, 81 πειραματικές προ-καταχωρήσεις αποτελεσμάτων, και 46 τελικές καταχωρήσεις αποτελεσμάτων από τους διαγωνιζόμενους, αποτελώντας έτσι την μεγαλύτερη, προτυποποιημένη, συγκριτική μελέτη στον τομέα.

Οι αλγόριθμοι τεχνητής νοημοσύνης δείχνουν να έχουν την ικανότητα τόσο να παράγουν αξιόπιστο καθορισμό των ορίων των βλαβών δέρματος σε δερματοσκοπικές εικόνες, όσο και να ανιχνεύουν επιμέρους δερματοσκοπικά κριτήρια, και σημαντικότερα, να ταξινομούν με ακρίβεια διαφορετικά νοσήματα, συμπεριλαμβανομένου του μελανώματος. Ακολούθως θα εκθέσουμε τη reader study την οποία διοργανώσαμε προκειμένου να συγκρίνουμε τα αποτελέσματα των αλγορίθμων τόσο με δερματολόγους έμπειρους στη διάγνωση του καρκίνου δέρματος, όσο και ειδικευόμενους δερματολογίας διερευνώντας το ρόλο που ενδεχομένως θα μπορούσαν να αποκτήσουν οι αλγόριθμοι Τεχνητής Νοημοσύνης στην έγκαιρη διάγνωση του καρκίνου δέρματος.[30]

ISIC Grand Challenge 2017, Reader Study

Εισαγωγή, ISIC Grand Challenge 2017, Reader Study

Στη μελέτη μας αυτή, συγκρίναμε τη διαγνωστική ακρίβεια του αλγορίθμου τεχνητής νοημοσύνης με την καλύτερη επίδοση στη διάγνωση μελανώματος από το ISIC Challenge 2017,[28] με αυτή έμπειρων δερματολόγων, ειδικών στη διάγνωση και θεραπεία καρκίνου δέρματος, καθώς και ειδικευόμενων δερματολογίας σε μια reader study. Ταυτόχρονα διερευνήσαμε τα αποτελέσματα που θα είχε η πιθανή χρήση του αλγορίθμου αυτού στη διαγνωστική ακρίβεια των κλινικών ιατρών στις περιπτώσεις όπου η αυτοπεποίθηση (confidence) τους στη διάγνωση ήταν χαμηλή.[30]

Υλικά και Μέθοδοι, ISIC Grand Challenge 2017, Reader Study

Λάβαμε έγκριση για τη μελέτη από το επιστημονικό συμβούλιο του Memorial Sloan Kettering Cancer Center, και η μελέτη πραγματοποιήθηκε σύμφωνα με τη διακήρυξη του Ελσίνκι. Τα αποτελέσματα και οι μέθοδοι του 2017 ISIC Challenge περιεγράφηκαν ανωτέρω.[28] Επιλέξαμε 2,750 δερματοσκοπικές εικόνες υψηλής ανάλυσης από το ISIC Archive (www.isic-archive.com): 521 (19%) μελανώματα, 1,843 (67%) σπίλους και 386 (14%) σμηγματορροϊκές υπερκερατώσεις. Οι εικόνες διαμοιράστηκαν τυχαία σε training dataset (n=2,000), validation dataset (n=150) και test dataset (n=600). Εικοσιτρείς (23) αλγόριθμοι κατατέθηκαν στη δοκιμασία για την ταξινόμηση μελανωμάτων (Task 3), και όλοι χρησιμοποιήσαν νευρωνικά δίκτυα (neural networks) και deep learning, μια μορφή machine learning που χρησιμοποιεί πολλαπλά υπολογιστικά επίπεδα προκειμένου να ταυτοποιήσει όλο και περισσότερο αφηρημένα σχήματα σε μια εικόνα.[51] Οι αλγόριθμοι αξιολογήθηκαν με βάση την

επίδοση τους κάτω από το receiver operating curve (ROC) και ακολούθως επιλέξαμε τον αλγόριθμο με την καλύτερη επίδοση για τις αναλύσεις μας.[28]

Η καμπύλη ROC είναι μια καμπύλη η οποία δημιουργείται αντιπαραθέτοντας την ευαισθησία μια διαγνωστικής διαδικασίας έναντι του αριθμού των ψευδώς αρνητικών αποτελεσμάτων (1-ειδικότητα) σε διάφορα όρια (thresholds). Κατ' αυτόν τον τρόπο, η καμπύλη ROC είναι ένα κοινώς αποδεκτό μέτρο της δυνατότητας ενός τεστ να ταξινομήσει εάν μια συγκεκριμένη συνθήκη είναι παρούσα ή απούσα. Μια καμπύλη ROC της τάξεως του 0.5 υποδηλώνει πως η διαγνωστική διαδικασία που ακολουθείται δεν έχει καμία δυνατότητα να κάνει αυτό το διαχωρισμό, ενώ μια ROC καμπύλη της τάξεως του 1.0 υποδηλώνει πως το εν λόγω τεστ έχει τέλεια ταξινομητική ιδιότητα. Οι καμπύλες ROC μπορούν να χρησιμοποιηθούν προκειμένου να καθορίσουν ένα αποδεκτό όριο για κάθε διαγνωστική διαδικασία, ωστόσο, η επιλογή του ορίου εξαρτάται από το σκοπό εκάστης διαγνωστικής διαδικασίας, και η ισορροπία μεταξύ ευαισθησίας και ειδικότητας εξαρτάται από το κλινικό σενάριο στο οποίο απευθύνεται κάθε συγκεκριμένη δοκιμασία.[52]

Για τη reader study μας χρησιμοποιήσαμε 150 δερματοσκοπικές εικόνες από 3 διαφορετικές διαγνώσεις [50 μελανώματα (15 διηθητικά, 20 in situ και 15 αταξινόμητα), 50 σπίλους και 50 σμηγματορροϊκές υπερκερατώσεις]. Το μέσο βάθος διείσδυσης Breslow για τα διηθητικά μελανώματα ήταν 0.3 (0.15-3.3) mm. Οχτώ δερματολόγοι με εξειδίκευση στη διάγνωση και αντιμετώπιση του καρκίνου δέρματος (experts), καθώς και 10 ειδικευόμενοι δερματολογίας συμφώνησαν να συμμετάσχουν στη μελέτη· ένας εκ των ειδικευομένων δεν ολοκλήρωσε τη μελέτη και αποκλείστηκε από κάθε περαιτέρω ανάλυση. Ο μέσος χρόνος κλινικής εμπειρίας και χρήσης δερματοσκόπησης των 8 experts που συμμετείχαν στη μελέτη ήταν 14 (4-32) και 14.5 (7-28) αντίστοιχα. Οι experts που συμμετείχαν προέρχονται από 4 χώρες (4 από τις ΗΠΑ, 2 από την Ισπανία, 1 από το Ισραήλ και 1 από

την Κολομβία), ενώ όλοι οι ειδικευόμενοι δερματολογίας προέρχονται από τις ΗΠΑ. Οι συμμετέχοντες στη μελέτη (readers), ταξινόμησαν τις βλάβες ως μελάνωμα, σπίλο, ή σμηγματορροϊκή υπερκεράτωση, πρότειναν τον χειρισμό που θα έκαναν για κάθε βλάβη (βιοψία ή παρακολούθηση) και ανέφεραν την αυτοπεποίθηση τους ως προς τη διάγνωση που παρείχαν σε κλίμακα Likert από 0 (εξαιρετικά μη βέβαιος) έως 6 (εξαιρετικά βέβαιος). Συνολικά προέκυψαν 1200 αξιολογήσεις εικόνων από τους ειδικούς και 1350 αξιολογήσεις από τους ειδικευόμενους. Οι αξιολογήσεις πραγματοποιήθηκαν μέσω της πλατφόρμας του ISIC archive, και οι readers ήταν blinded ως προς τη διάγνωση και τα κλινικά δεδομένα. Δεν υπήρχε κανένας χρονικός περιορισμός και οι συμμετέχοντες είχαν τη δυνατότητα να ολοκληρώσουν τις εκτιμήσεις τους μετά από όσες συνεδρίες επιθυμούσαν. Προκειμένου να υπάρξει άμεση σύγκριση με τους συμμετέχοντες ιατρούς, η επίδοση του κορυφαίου αλγόριθμου επανυπολογίστηκε συγκεκριμένα για τις 150 βλάβες που συμπεριλήφθηκαν στη μελέτη.

Χρησιμοποιήσαμε περιγραφικά στατιστικά προκειμένου να εξερευνήσουμε την κατανομή των αποτελεσμάτων των readers και του αλγορίθμου για την ταξινόμηση των βλαβών και την αυτοπεποίθηση για την ταξινόμηση (confidence level). Επιπλέον, στην ανάλυση μας συμπεριλάβαμε τη διαγνωστική ακρίβεια των readers, τόσο για την ταξινόμηση, όσο και για τη διαχείριση των βλαβών που εξετάσαμε. Η διαγνωστική ακρίβεια του αλγόριθμου συμπεριλήφθηκε μόνο όσον αφορά στην ταξινόμηση των βλαβών, ενώ υπολογίσαμε τα ROC curves για τον αλγόριθμο και τους readers ξεχωριστά, καθώς και τις κατηγοριοποιημένες ομάδες των readers (experts vs. ειδικευόμενοι). Προκειμένου να συγκρίνουμε τις περιοχές της καμπύλης ROC μεταξύ του αλγορίθμου και των ιατρών χρησιμοποιήσαμε μη-παραμετρική προσέγγιση.[53]

Όταν η αυτοπεποίθηση των ιατρών που συμμετείχαν στη μελέτη για την ταξινόμηση των βλαβών ήταν χαμηλή (confidence 0-3), στα αποτελέσματα τους συνυπολογίσαμε την πρόβλεψη του αλγορίθμου.

Αυτό επιτεύχθηκε διχοτομώντας τα αποτελέσματα του αλγορίθμου με όριο μια ευαισθησία της τάξεως του 90%. Μετά τον συνυπολογισμό των αποτελεσμάτων των readers με τη βοήθεια του αλγορίθμου υπολογίσαμε εκ νέου την διαγνωστική ακρίβεια. Όλες οι αναλύσεις έγιναν χρησιμοποιώντας Stata v.14.2, Stata Corporation, College Station, TX.

Αποτελέσματα, ISIC Challenge 2017, Reader Study

Η συνολική ευαισθησία, ειδικότητα και η καμπύλη ROC για τους ειδικευμένους δερματολόγους που συμμετείχαν στη μελέτη ήταν 76.0% (95% CI:71.5–80.1), 72.6% (95% CI:69.4–75.7) και 0.74 (95% CI:0.72–0.77) αντίστοιχα. Η συνολική ευαισθησία, ειδικότητα και η καμπύλη ROC για τους ειδικευόμενους δερματολογίας που συμμετείχαν στη μελέτη για τη σωστή ταξινόμηση μελανώματος ήταν 56.0% (95% CI:51.3–60.6), 76.3% (95% CI:133 73.4–79.1) και 0.66 (95% CI:0.6–0.69), αντιστοίχως. Η καμπύλη ROC του αλγορίθμου με την καλύτερη επίδοση για την ταξινόμηση μελανώματος ήταν 0.8685 (**Figure 7),** η οποία ήταν καλύτερη από τη συνολική επίδοση (σε ROC curve) τόσο των ειδικών, όσο και των ειδικευομένων δερματολόγων ($p<0.01$ για όλες τις συγκρίσεις).

Λαμβάνοντας ως όριο σύγκρισης την συνολική ευαισθησία των ειδικών δερματολόγων για την ορθή ταξινόμηση των βλαβών, η οποία ήταν της τάξεως του 76.0%, ο αλγόριθμος τεχνητής νοημοσύνης είχε ειδικότητα για την ταξινόμηση 85.0%, υψηλότερη από αυτή των ειδικών που ήταν 72.6% ($p=0.001$). Αντίστοιχα, η συνολική ευαισθησία των ειδικών δερματολόγων για τη διαχείριση (management) των βλαβών ήταν 89.0% και με αυτό ως όριο σύγκρισης, η ειδικότητα του αλγορίθμου τεχνητής νοημοσύνης ήταν 61%, υψηλότερη από αυτή των δερματολόγων που ήταν 51.1% ($p=0.02$)

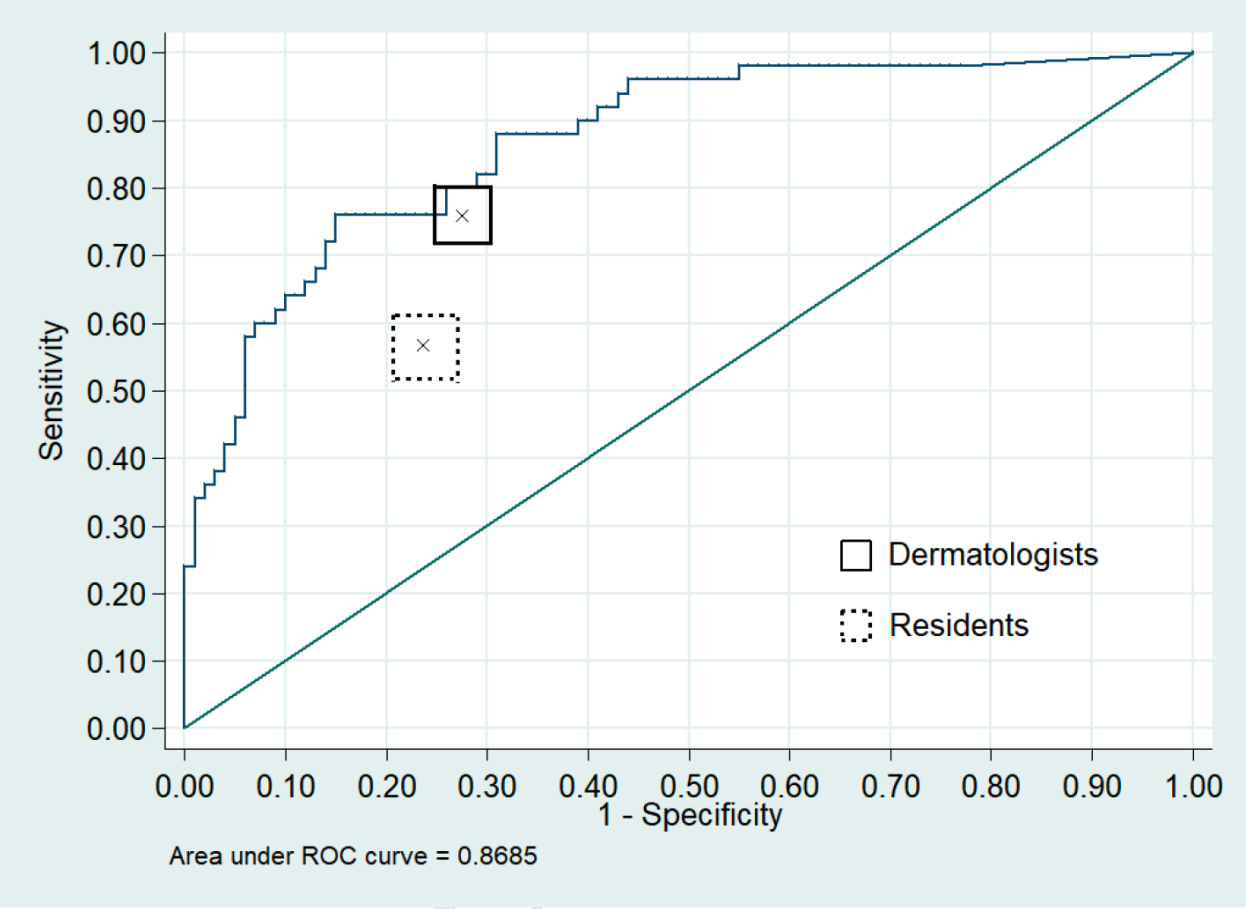
Προκειμένου να διερευνήσουμε τη δυνατότητα των αλγορίθμων τεχνητής νοημοσύνης να βοηθήσουν στη διάγνωση μελανώματος, χρησιμοποιήσαμε τις προβλέψεις του αλγορίθμου στις περιπτώσεις όπου η αυτοπεποίθηση των readers στην κλίμακα Likert ήταν χαμηλή (0-3). (**Table 2)** Αυτές οι περιπτώσεις αποτελούσαν το 51% του συνολικού αριθμού των εκτιμήσεων των ειδικευομένων, και το 26.6% των ειδικών δερματολόγων αντίστοιχα. Μετά τη χρήση των προβλέψεων του αλγορίθμου, η ευαισθησία των ειδικευομένων δερματολογίας αυξήθηκε σε 72.9% από 56.0%, με αντίστοιχη μείωση της ειδικότητας σε 72.6% από 76.3%. Η διαγνωστική ακρίβεια των ειδικευομένων αυξήθηκε σε 72.6% (n=981 από το συνολικό αριθμό βλαβών 1350) από 69.4% (n=939). Επιπλέον, αναφορικά με τους ειδικούς δερματολόγους, η χρήση του αλγορίθμου οδήγησε σε αύξηση της ευαισθησίας τους σε 80.8%, από 76.0% που ήταν, ενώ και η ειδικότητα τους για τη διάγνωση μελανώματος αυξήθηκε σε 72.8% από 72.6%.

Table 2. Μέτρηση της διαγνωστικής ακρίβειας για την ταξινόμηση βλαβών, με βάση την αναφερόμενη αυτοπεποίθηση στη διάγνωση, καθώς και η συνοδός ευαισθησία (sensitivity) και ειδικότητα (specificity) για τους ειδικευόμενους δερματολογίας (residents) και τους ειδικούς δερματολόγους (dermatologists). Οι readers ανέφεραν μέση αυτοπεποίθηση στη διάγνωση τους 3.7 (SD=1.51), με τους ειδικούς δερματολόγους να έχουν υψηλότερη αυτοπεποίθηση από τους ειδικευομένους (p<0.001).

Group and confidence level	n (%)	Sensitivity (95% CI)	P_{trend}	Specificity (95% CI)	P_{trend}
Residents					
0	7 (0.5)	100.0 (2.5-100.0)	.54	16.7 (0.4-64.1)	<.001
1	160 (11.8)	57.6 (44.1-70.4)		61.4 (51.2-70.9)	
2	238 (17.6)	48.6 (36.9-60.6)		73.2 (65.7-79.8)	
3	289 (21.4)	53.6 (43.2-63.8)		70.3 (63.3-76.7)	
4	397 (29.4)	51.8 (43.1-60.4)		81.9 (76.7-86.4)	
5	204 (15.1)	63.1 (50.2-74.7)		87.8 (81.1-92.7)	
6	55 (4.1)	100.0 (80.5-100.0)		89.5 (75.2-97.1)	
Dermatologists					
0	26 (2.2)	75.0 (34.9-96.8)	.002	61.1 (35.7-82.7)	<.001
1	65 (5.4)	62.5 (40.6-81.2)		68.3 (51.9-81.9)	
2	97 (8.1)	52.0 (31.3-69.8)		58.3 (46.1-69.8)	
3	131 (10.9)	67.3 (52.9-79.7)		63.3 (51.7-73.9)	
4	301 (25.1)	74.3 (64.8-82.3)		64.8 (57.7-71.5)	
5	342 (28.5)	79.5 (70.8-86.5)		76.1 (70.0-81.4)	
6	238 (19.8)	91.9 (83.2-97.0)		90.2 (84.6-94.3)	

Figure 7.

Διαγνωστική ακρίβεια του αλγορίθμου με την κορυφαία επίδοση από το ISIC Challenge 2017, καθώς και των ειδικευομένων δερματολογίας (residents) και των ειδικών (dermatologists) που συμμετείχαν στη μελέτη για τις 150 εικόνες που εξετάστηκαν. Η καμπύλη ROC δείχνει την ευαισθησία και την ειδικότητα του αλγορίθμου με την κορυφαία επίδοση (μπλε καμπύλη). Το συμπαγές μαύρο κουτί δείχνει την συνολική επίδοση των 8 ειδικών δερματολόγων ενώ το διακεκομμένο μαύρο κουτί δείχνει την επίδοση των ειδικευομένων δερματολογίας, καθώς και το 95% Confidence Interval τους.



Συζήτηση, ISIC Challenge 2017, Reader Study

Τα αποτελέσματα μας, σε συνδυασμό με άλλες έρευνες στο πεδίο,[54-56] καταδεικνύουν ότι τα νευρωνικά δίκτυα (deep neural networks) και οι αλγόριθμοι τεχνητής νοημοσύνης μπορούν να ταξινομήσουν εικόνες μελανώματος με μεγάλη ακρίβεια. Εν συγκρίσει με το ISIC Challenge 2016,[27,38] παρατηρήσαμε βελτίωση της διαγνωστικής ακρίβειας του αλγορίθμου με την κορυφαία επίδοση συγκρινόμενο με τους ιατρούς - readers της μελέτης. Αυτό το εύρημα οδηγεί στο πιθανό συμπέρασμα πως η επίδοση των υπαρχόντων αλγορίθμων ολοένα βελτιώνεται, ενδεχομένως λόγω της ύπαρξης μεγαλύτερων και ποιοτικότερων βάσεων δεδομένων για την ανάπτυξη αυτών.[10]

Μολονότι αρκετές μελέτες έχουν δείξει πως οι αλγόριθμοι τεχνητής νοημοσύνης έχουν διαγνωστική ακρίβεια εφάμιλλη, αν όχι ανώτερη, αυτής των κλινικών δερματολόγων σε reader studies, ακόμα και experts του τομέα, η δυνατότητα εφαρμογής τους στην κλινική πράξη παραμένει αμφίβολη. Προκειμένου να διερευνήσουμε αυτή τη δυνατότητα, υπολογίσαμε αν και κατά πόσο η χρήση των προβλέψεων των αλγορίθμων θα άλλαζε τη διάγνωση και τη διαχείριση, στις περιπτώσεις όπου ο ιατρός έχει χαμηλή αυτοπεποίθηση στη διάγνωση του. Σε αυτή την ανάλυση βρήκαμε ότι η ευαισθησία, και η συνολική διαγνωστική ακρίβεια των readers αυξήθηκε μετά την προσθήκη της πρόβλεψης των αλγορίθμων. Περαιτέρω μελέτες χρειάζονται, προκειμένου να διερευνήσουμε τα όρια (thresholds), τα οποία θα ήταν χρήσιμα στην κλινική πράξη και θα ωφελούσαν τόσο τους ιατρούς, όσο και τους ασθενείς, σε ένα ευρύ φάσμα κλινικών σεναρίων, μελέτες τις οποίες ήδη διεξάγουμε στο νοσοκομείο Ανδρέας Συγγρός.

Οι περιορισμοί της έρευνας μας είναι οι εξής: το test dataset που χρησιμοποιήσαμε για τη μελέτη αυτή δεν περιλάμβανε όλο το εύρος των αλλοιώσεων δέρματος, ειδικότερα, κλασσικές καλοήθεις βλάβες,

καθώς και εικόνες μελανωμάτων με σπανιότερες μορφές και εμφανίσεις. Επιπροσθέτως, όλη η έρευνα έγινε σε ένα τεχνητό περιβάλλον, ενώ οι ιατροί που συμμετείχαν στη μελέτη δεν είχαν πρόσβαση σε όλα τα δεδομένα τα οποία λαμβάνουν υπόψιν τους όταν εκτιμούν έναν ασθενή με φυσική παρουσία (ηλικία, ατομικό και οικογενειακό ιστορικό μελανώματος, το αν η βλάβη είναι συμπτωματική κ.α.). Τέλος, δεν πραγματοποιήσαμε εξωτερική, ανεξάρτητη επιβεβαίωση των αποτελεσμάτων του αλγορίθμου, η οποία είναι σημαντική προκειμένου να επιβεβαιωθεί το κατά πόσο οι προβλέψεις των αλγορίθμων τεχνητής νοημοσύνης μπορούν να γενικευτούν στον γενικό πληθυσμό.[31]

Επιπλέον, οι συγκρίσεις της διαγνωστικής ακρίβειας των δερματολόγων με τους αλγόριθμους τεχνητής νοημοσύνης για τη διάγνωση μελανώματος μέσω των reader studies θα πρέπει να ερμηνεύονται με προσοχή. Μια συσκευή αυτόματης διάγνωσης μελανώματος, η οποία είχε λάβει έγκριση από τον Food and Drug Administration (FDA) των ΗΠΑ είχε δείξει υψηλή ευαισθησία για τη διάγνωση μελανώματος και στο να αυξάνει τόσο την ευαισθησία, όσο και της ειδικότητα των ιατρών μετά από κλινική και δερματοσκοπική εκτίμηση του ασθενούς μέσω reader studies. [57] Μολαταύτα, η συσκευή αυτή αποσύρθηκε το 2017, καθώς αποδείχτηκε πως δεν είχε εφάμιλλα αποτελέσματα στην καθημερινή κλινική πράξη.

Αυτό που διαφοροποιεί αυτή τη μελέτη από άλλες αντίστοιχες,[54-56] είναι πως εξετάσαμε τη διαγνωστική ακρίβεια των αλγορίθμων τεχνητής νοημοσύνης χρησιμοποιώντας μια δημόσια προσβάσιμη βάση δεδομένων με δερματοσκοπικές εικόνες, οι οποίες μπορούν να χρησιμοποιηθούν από ερευνητές για περαιτέρω μελέτες και επιβεβαίωση των αποτελεσμάτων μας. Επιπλέον, συγκρίναμε τη διαγνωστική ακρίβεια των readers με αυτή του αλγορίθμου τεχνητής νοημοσύνης με την κορυφαία επίδοση από το ISIC Challenge 2017,[28] χρησιμοποιώντας έτσι την τελευταία λέξη της τεχνολογίας και της έρευνας στον τομέα προκειμένου να προχωρήσουμε στις συγκρίσεις μας. Τα

ετήσια ISIC Challenges για τη διάγνωση μελανώματος είναι οι μεγαλύτερες συνεργατικές και συγκριτικές μελέτες της αυτοματοποιημένης ανίχνευσης καρκίνου του δέρματος μέχρι στιγμής και έχουν καταστεί σημείο αναφοράς για την έρευνα στον τομέα αυτό. Καθώς το αρχείο μας μεγαλώνει, αναμένουμε να φιλοξενούμε ετήσια challenges με διαρκώς αυξανόμενη βάση δεδομένων, συμπεριλαμβάνοντας όλο και περισσότερες διαγνώσεις καθώς και κλινικά δεδομένα τα οποία είναι χρήσιμα στην καθημερινή πράξη. Οι μελέτες οι οποίες ακολούθησαν (2018-2020) συμπεριέλαβαν όλο το εύρος των καρκίνων δέρματος, όπως βασικοκυτταρικό καρκίνωμα, ακανθοκυτταρικό καρκίνωμα καθώς και επιπρόσθετους μιμητές τους.[58] Επιπρόσθετα, στα πλαίσια της διατριβής μας επιχειρήσαμε να επανεξετάσουμε το task 2 για την απομόνωση συγκεκριμένων δερματοσκοπικών χαρακτηριστικών μέσω της Expert Agreement Study on Dermoscopy μελανοκυτταρικών βλαβών (EASY Study). Ο εντοπισμός συγκεκριμένων δερματοσκοπικών χαρακτηριστικών που συμπεριελήφθη στα ISIC Challenges 2016 & 2017 επέδειξε χαμηλότερη επίδοση των αλγορίθμων συγκριτικά με τα άλλα δυο tasks και στα πλαίσια της EASY Study, την οποία θα εκθέσουμε στην πορεία, επιχειρήσαμε να επαναπροσδιορίσουμε τα δερματοσκοπικά χαρακτηριστικά για τη διάγνωση σπίλων και μελανώματος.[27,28]

Συμπέρασμα, ISIC Challenge 2017

Συμπερασματικά, ο αλγόριθμος με την κορυφαία επίδοση από το ISIC Challenge 2017 έδειξε καλύτερη διαγνωστική ακρίβεια τόσο συγκριτικά με τους ειδικευόμενους δερματολογίας, όσο και με experts, οι οποίοι εξειδικεύονται στη διάγνωση και τη θεραπεία του καρκίνου του δέρματος, στο τεχνητό περιβάλλον μιας reader study. Η ευαισθησία και η συνολική διαγνωστική ακρίβεια των ιατρών βελτιώθηκε, μετά την προσθήκη των προβλέψεων των αλγορίθμων, στις περιπτώσεις όπου οι κλινικοί ιατροί ανέφεραν μειωμένη αυτοπεποίθηση ως προς τη διάγνωση τους, δείχνοντας έτσι, πως οι

αλγόριθμοι τεχνητής νοημοσύνης θα μπορούσαν να έχουν ρόλο στην κλινική πράξη σε ένα τέτοιο σενάριο. Μελλοντικές μελέτες οι οποίες θα χρησιμοποιήσουν τους αλγορίθμους κατά τη διάρκεια της εξέτασης ενός ασθενούς είναι αναγκαίες προκειμένου να επιβεβαιωθούν αυτά τα αρχικά αποτελέσματα, ενώ ήδη διεξάγουμε μια τέτοια μελέτη στο νοσοκομείο Δερματικών και Αφροδισίων Νοσημάτων «Ανδρέας Συγγρός», με υποσχόμενα αποτελέσματα.

Διερεύνηση των μειονεκτημάτων των αλγορίθμων Τεχνητής Νοημοσύνης για τη διάγνωση δερματολογικών παθήσεων

Εισαγωγή

Στα πλαίσια της έρευνας μας προσπαθήσαμε να διερευνήσουμε τόσο τη διαγνωστική ακρίβεια, όσο και το generalizability υπαρχόντων αλγορίθμων τεχνητής νοημοσύνης για τη διάγνωση δερματολογικών παθήσεων, και ειδικότερα για τη διάγνωση του καρκίνου δέρματος. Για αυτό το λόγο διεξάγαμε δυο ερευνητικές προσπάθειες επαλήθευσης δημόσια διαθέσιμων, διαγνωστικών αλγόριθμων τεχνητής νοημοσύνης.[56,59]

Automated Dermatological Diagnosis: Hype, or Reality?[31]

Το 2018 οι Han et al. έκαναν μια εξαιρετικά σημαντική συνεισφορά στην εφαρμογή της Τεχνητής Νοημοσύνης στη διάγνωση δερματολογικών παθήσεων. Μολονότι προηγούμενες ερευνητικές προσπάθειες είχαν δείξει πως οι αλγόριθμοι τεχνητής νοημοσύνης μπορούν να διαγιγνώσκουν τον καρκίνο του δέρματος με ακρίβεια εφάμιλλη αυτής των κλινικών ιατρών, οι Han et al. ήταν οι πρώτοι οι οποίοι έκαναν τον αλγόριθμο τους δημόσια διαθέσιμο για εξωτερική επαλήθευση των ευρημάτων τους.[38,55,56,60] Ο εν λόγω αλγόριθμος είχε εκπαιδευτεί σε 12 διαγνώσεις οι οποίες συμπεριελάμβαναν το μελάνωμα, το BCC, το SCC, το SCC in situ, την ακτινική υπερκεράτωση, τη σμηγματορροϊκή υπερκεράτωση, τους μελανοκυτταρικούς σπίλους, την ηλιακή φακή, το αιμαγγείωμα, το πυογόνο κοκκίωμα, το δερματοΐνωμα και τις μυρμηκίες. Προκειμένου να διερευνήσουμε το generalizability του αλγορίθμου τους σε έναν ξεχωριστό πληθυσμό ασθενών, επιλέξαμε 100 καρκίνους δέρματος, οι οποίοι είχαν υποστεί βιοψία και οι οποίοι διέθεταν εικόνες υψηλής ποιότητας και ευκρίνειας. Οι εικόνες

αυτές αποτελούνταν από 37 εικόνες μελανώματος, 40 εικόνες BCC και 23 εικόνες SCC, οι οποίες είναι δημόσια προσβάσιμες στο ISIC-archive (https://isic-archive.com/#images, dataset name: 2018 JID Editorial Images).

Το δημόσιο αυτό dataset αποτελείται από 15 in situ μελανώματα και 22 διηθητικά μελανώματα, με μέσο όρο Breslow thickness τα 0.6mm (0.1-3.1mm), ενώ τα 14 από τα SCC ήταν SCC in situ και τα 9 ήταν διηθητικά. Οι βλάβες εντοπίζονταν στον τράχηλο/κεφάλι (n=26), στον κορμό (n=21), στα άνω άκρα (n=30) και στα κάτω άκρα (n=23), η μέση ηλικία των ασθενών ήταν τα 66.8 έτη, 67% των ασθενών ήταν άνδρες, ενώ όλες οι βλάβες που προέρχονταν από Καυκάσιους ασθενείς στις νότιες ΗΠΑ. Ακολούθως, ανεβάσαμε τις εικόνες αυτές στην διαδικτυακή εφαρμογή των Han et al. η οποία είναι διαθέσιμη στο http://dx.medicalphoto.org/.

Συνολικά ο αλγόριθμος των Han et al. βρήκε τη σωστή ιστοπαθολογική διάγνωση στο 29% των βλαβών (**Table 3)**. Λαμβάνοντας υπόψιν και τις 5 διαγνώσεις που προσέφερε ο αλγόριθμος ως αποτέλεσμα στη διαφορική του διάγνωση (ανεξαρτήτως πιθανότητας), η σωστή διάγνωση συμπεριλήφθηκε στο 58% των βλαβών. Δε διαπιστώθηκε καμία στατιστικά σημαντική διαφορά ανάμεσα στις βλάβες οι οποίες ταξινομήθηκαν σωστά και λανθασμένα (0.711 vs.0.715, P = 0.94, paired t-test). Αναφορικά με τις εικόνες μελανωμάτων οι οποίες συμπεριλήφθηκαν στη μελέτη μας, το μελάνωμα ήταν η 1^{η} διάγνωση στο 13.5% των περιπτώσεων (5 από 37), με μέσος (range) πιθανότητα 0.82 (0.42-0.99). Συμπεριλαμβάνοντας και τις 5 διαγνώσεις που ο αλγόριθμος προσέφερε ως διαφορική διάγνωση, το μελάνωμα συμπεριλαμβανόταν στο 35.1% των περιπτώσεων (13 από 37), με μέση (εύρος) πιθανότητα 0.43 (0.02-0.99). Ανάμεσα στις 8 εικόνες μελανώματος όπου η διάγνωση του μελανώματος ήταν 2^{η} ή 3^{η} πιθανολογικά διάγνωση το μέσο (range) score πιθανότητας ήταν 0.18 (0.02-0.037). Τα αποτελέσματα μας έδειξαν ότι η ευαισθησία του αλγορίθμου των Han et al. που δημοσιεύτηκαν το 2018 στο Journal of

Investigative Dermatology, και ειδικότερα η ευαισθησία του για τη διάγνωση μελανώματος ήταν σημαντικά μειωμένη όταν εφαρμόστηκε σε βλάβες προερχόμενες από διαφορετικό πληθυσμό από αυτόν στον οποίο εκπαιδεύτηκε το μοντέλο AI, περιορίζοντας έτσι το generalizability του. Τέλος διαπιστώσαμε πως χειρισμοί ως προς τη φωτεινότητα ή την αντίθεση της εικόνας μπορούν να επηρεάσουν το αποτέλεσμα του αλγορίθμου **(Figure 8).** [31]

Multiclass artificial intelligence in dermatology - progress but still room for improvement[32]

Το 2020, οι ίδιοι συγγραφείς επέκτειναν τον διαγνωστικό τους αλγόριθμο AI, προκειμένου να συμπεριλάβουν 174 διαγνώσεις, εκπαιδεύοντας τον σε 220.680 εικόνες, και επαληθεύοντας τα αποτελέσματα τους για 134 εκ των διαγνώσεων αυτών και με εξωτερικά dataset. [59] Προκειμένου να επαληθεύσουμε τα αποτελέσματα τους, και να εξετάσουμε αν βελτιώθηκε το generalizability του αλγορίθμου τους, χρησιμοποιήσαμε τις ίδιες 100 βλάβες που χρησιμοποιήθηκαν και στον αλγόριθμο του 2018. [31] Ανεβάσαμε τις βλάβες στη διαδικτυακή εφαρμογή του καινούργιου αλγορίθμου μεταξύ 7 και 9 Απριλίου 2020 (https://modelderm.com/). Η διαδικτυακή αυτή εφαρμογή επιτρέπει στους χρήστες να επιλέξουν μια περιοχή ενδιαφέροντος (region of interest - ROI), τετράγωνου σχήματος και με μέγεθος 250pixels-by-250pixels εντός της εικόνα που έχει κατατεθεί. Η επιλογή αυτού του ROI γίνεται μειώνοντας τη μεγέθυνση της εικόνα και τοποθετώντας στο κέντρο του πεδίου.

Πραγματοποιήσαμε μια σειρά από τέσσερα πειράματα. Στη σειρά uploads που περιγράφεται από τις οδηγίες χρήσης και καταγράψαμε ως 'intended use', επιλέξαμε ως ROI αυτό το οποίο (i) είχε τη βλάβη στο κέντρο και (ii) κάλυπτε ~80% της βλάβης. Έπειτα ανεβάσαμε τις ίδιες βλάβες και προσαρμόσαμε το ROI ώστε η βλάβη να είναι εκτός του κέντρου του πεδίου, 'intended use off-center'. Ταυτόχρονα,

χρησιμοποιήσαμε το ‘drag and drop’, όπου χρησιμοποιήσαμε τη μεγέθυνση που επέλεγε αυτόματα ο αλγόριθμος. Τέλος χρησιμοποιήσαμε και το ‘1x magnification’, όπου τοποθετήσαμε τη βλάβη στο κέντρο και προσαρμόσαμε τη μεγέθυνση ώστε να καλύπτει το μεγαλύτερο δυνατόν πεδίο (όσο εγγύτερα στη μεγέθυνση 1x ήταν δυνατόν). Για κάθε εικόνα, και για τα 4 πειράματα, καταγράψαμε τις τρείς πρώτες διαγνώσεις του αλγορίθμου καθώς και τις πιθανότητες που τους αντιστοιχούσαν. Επιπλέον καταγράψαμε την πιθανότητα για κακοήθεια που έδινε ως αποτέλεσμα ο αλγόριθμος, ‘malignancy probability’, το οποίο είναι το άθροισμα των πιθανοτήτων που παράγει ο αλγόριθμος για μελάνωμα, BCC και SCC. Υπολογίσαμε τις top-1 accuracy, top-3 accuracy, και το average και standard deviation (SD) σκορ κακοήθειας (malignancy score). Η ευαισθησία του αλγορίθμου υπολογίστηκε χρησιμοποιώντας το όριο του 10% για πιθανότητα κακοήθειας, με βάση τη δημοσίευση των Han et al.[59]

Η συνολική top-1 και top-3 διαγνωστική ακρίβεια του νέου μοντέλου ήταν βελτιωμένη συγκριτικά με αυτή του 2018 (top-1: 39% vs. 32%, p=0.38; top-3: 63% vs. 60%, p=0.77) (**Table 4**). Η διαγνωστική ακρίβεια για το μελάνωμα βελτιώθηκε σημαντικότερα, συγκριτικά με τους άλλους καρκίνους δέρματος (top-1: 29.7% vs. 13.5%, p=0.16; top-3: 56.7% vs. 35.1%, p=0.10), ωστόσο σε απόλυτα ποσοστά παρέμεινε χαμηλή (**Figure 9**). Η top-1 διαγνωστική ακρίβεια ήταν υψηλότερη για το BCC (55%), ακολουθούμενη από το μελάνωμα (29.7%) και το SCC (17.3%) (X^2=8.6, p=0.01). Ο μέσος όρος της πιθανότητας κακοήθειας (SD) των 100 εικόνων καρκίνου του δέρματος ήταν 41.1% (30.4), καταδεικνύοντας ότι το model calibration επιδέχεται βελτίωσης. (**Table 4**).

Συζήτηση

Μολονότι κάποιος θα μπορούσε να ερμηνεύσει τα αποτελέσματα μας ως ενδείξεις χαμηλής επίδοσης των εν λόγω αλγορίθμων, είναι σημαντικό να λάβουμε υπόψιν τους εγγενείς περιορισμούς και τις προκλήσεις που σχετίζονται με την αυτοματοποιημένη διάγνωση του καρκίνου δέρματος. Αναφορικά με τον αλγόριθμο του 2018, μόλις 20.000 εικόνες χρησιμοποιήθηκαν για την εκπαίδευση του αλγόριθμου τους, και εξ αυτών, μόλις οι 6.000 ήταν εικόνες κακοήθειας.

Σε αμφότερες τις έρευνες μας βρήκαμε περιορισμένο generalizability των αλγορίθμων σε ένα εξωτερικό dataset, ενώ η top-1 και top-3 διαγνωστική ακρίβεια για μελάνωμα, BCC και SCC ήταν χαμηλότερη από αυτή την οποία ανέφεραν οι συγγραφείς. Εντοπίσαμε ότι η διαγνωστική ακρίβεια του αλγόριθμου εξαρτάται σε μεγάλο βαθμό από τη μεγέθυνση με την οποία θα υποβληθεί μια εικόνα προς ανάλυση, το ανατομικό σημείο όπου βρίσκονταν οι βλάβες, καθώς και από το αν οι εικόνες είναι προσεκτικά τοποθετημένες στο κέντρο του ROI. Τα ευρήματα μας αυτά είναι συμβατά με αντίστοιχες έρευνες οι οποίες έχουν δείξει ότι οι αλγόριθμοι αυτόματης διάγνωσης δερματολογικών παθήσεων μπορούν να επηρεαστούν από πλειάδα παραγόντων, μεταξύ αυτών του ‘image noise’, της περιστροφής των εικόνων, των παρεμβάσεων ως προς τη φωτεινότητα και την αντίθεση των εικόνων, καθώς και από pen markings.[31,61-63] Επιπλέον, μια συχνή κριτική η οποία ασκείται στους ερευνητές των αλγορίθμων τεχνητής νοημοσύνης, όπως και στο ίδιο το ISIC-Archive, περιλαμβάνει τον περιορισμό ως προς επί το πλείστον Καυκάσιους ασθενείς, έχοντας έτσι περιορισμένη ποικιλία ως προς την εμφάνιση κάποιων νοσημάτων.[64]

Κάποιοι ενδεχόμενοι τρόποι βελτίωσης των διαγνωστικών αλγορίθμων για την αυτοματοποιημένη διάγνωση δερματολογικών νοσημάτων, και ειδικότερα, καρκίνου δέρματος συμπεριλαμβάνουν: 1. Τη χρήση προτυποποιημένων μεθόδων για την λήψη φωτογραφιών, ώστε να παράγονται standardized datasets, 2. Την εκπαίδευση των αλγορίθμων αυτών σε πολύ ευρεία datasets, τα οποία θα περιλαμβάνουν φωτογραφίες από ποικιλία φωτογραφικών εξοπλισμών αλλά και άτομα διαφορετικής φυλής, ηλικίας, φύλου ώστε τα training set να περιλαμβάνουν μια ευρύτερη ποικιλία παρουσίασης δερματολογικών νοσημάτων και συνακόλουθα να είναι ευκολότερο το generalizability των αποτελεσμάτων τους.

Table 3.

Συχνότητες για το cross-classification μεταξύ ιστοπαθολογικής διάγνωσης των εικόνων του dataset μας και της πιθανότερης διάγνωσης που παρείχε ως αποτέλεσμα ο αλγόριθμος, καθώς και ο μέσος όρος της πιθανότητας συσχέτισης με την ακριβή διάγνωση για τον αλγόριθμο του 2018. [56]

	Web app categorization										
Histopathologic diagnosis	**Melanoma**	**Basal cell carcinoma**	**Intraepithelial carcinoma**	**Squamous cell carcinoma**	**Hemangioma**	**Lentigo**	**Actinic keratosis**	**Nevus**	**Seborrheic keratosis**	**Wart**	**Total**
Melanoma	**5** **0.82**	2 0.96	6 0.70	3 0.59	1 0.96	12 0.82	1 0.94	5 0.65	0	2 0.82	37
Basal cell carcinoma	0	**19** **0.68**	10 0.78	1 0.64	3 0.83	1 0.81	1 0.98	2 0.74	1 0.37	2 0.57	40
Intraepithelial carcinoma	0	6 0.59	**4** **0.83**	1 0.51	0	1 0.52	1 0.46	0	0	1 0.85	14
Squamous cell carcinoma	1 0.17	1 0.46	4 0.87	**1** **0.30**	0	0	0	0	0	2 0.36	9
Total	6	28	24	6	4	14	3	7	1	7	100

The bold values represent the "correct" diagnosis.

Table 4: Διαγνωστική ακρίβεια ανάλογα με τη διάγνωση, τον αλγόριθμο που χρησιμοποιήθηκε και την πειραματική μέθοδο.[56,59]

	Han et al 2020 174-disease algorithm				**Han et al 2018 12-disease algorithm**
	Intended Use	*Intended Use, Off-Center*	*Drag and Drop*	*1x Magnification*	*Intended Use*
Overall Top-1 Accuracy, n=100	39%	37%	28%	24%	32%
Overall Top-3 Accuracy, n=100	63%	65%	65%	48%	60%
Overall Top-Any Accuracy, n=100	-	-	-	-	-
Melanoma Top-1 Accuracy, n=37	29.7%	29.7%	16.2%	2.7%	13.5%
Melanoma Top-3 Accuracy, n=37	56.7%	56.7%	59.4%	18.9%	35.1%

Melanoma Top-Any Accuracy, n=37	-	-	-	-	-
Malignancy Probability, mean (SD), n=100	41.1% (30.4)	40.9% (29)	41.3% (26.7)	32.3% (26.4)	-
Sensitivity (malignancy probability ≥10%)	77%	80%	83%	70%	-

Figure 8.

Μεταβολή της ταξινόμησης του αλγόριθμου του 2018 μέσω χειρισμού της εικόνας[31,56] **(a, b)** Βασικοκυτταρικό καρκίνωμα. Η αρχική εικόνα μεταβλήθηκε αυξάνοντας τη μεγέθυνση, κάνοντας τον αλγόριθμο να δώσει διαφορετική ταξινόμηση. Στην εικόνα (a) το αποτέλεσμα του αλγόριθμου ήταν εφηλίδα (99.2% confidence) ενώ στην εικόνα (b) βασικοκυτταρικό καρκίνωμα (96.9% confidence). **(c, d)** Μελάνωμα. Στην αρχική εικόνα (c) μεταβλήθηκε η φωτεινότητα και η αντίθεση (d). Ο αλγόριθμος έκανε διαφορετικό classification στις δύο εικόνες με την εικόνα (c) να ταξινομείται ως μελάνωμα (99% confidence) ενώ η εικόνα (d) ταξινομήθηκε ως αιμαγγείωμα (98% confidence). **(e, f)** Μελάνωμα. Η αρχική εικόνα (e) μεταβλήθηκε αλλάζοντας τον προσανατολισμό της βλάβης (f) και ο αλγόριθμος αντίστοιχα μετέβαλε την ταξινόμηση του. Η αρχική εικόνα ταξινομήθηκε ως εφηλίδα (74% confidence), μελάνωμα (12% confidence),

σπίλος (5% confidence), ενώ η εικόνα (f) ταξινομήθηκε ως μελάνωμα (40.5% confidence), εφηλίδα (32% confidence), σπίλος (24%confidence). Όλες οι εικόνες προέρχονται από το International Skin Imaging Collaboration Archive (https://isicarchive.com/#images, dataset name: 2018 JID Editorial Images)

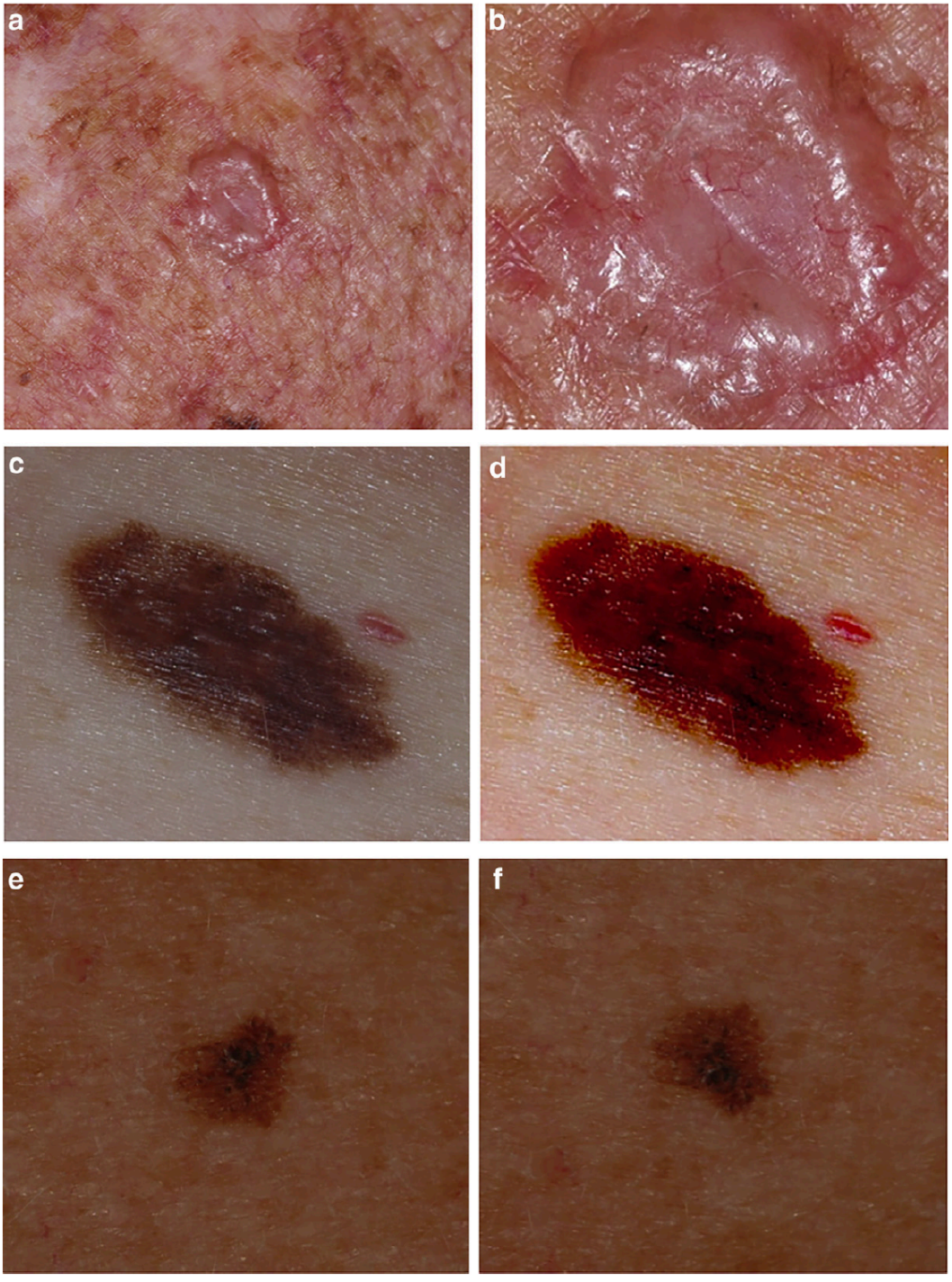
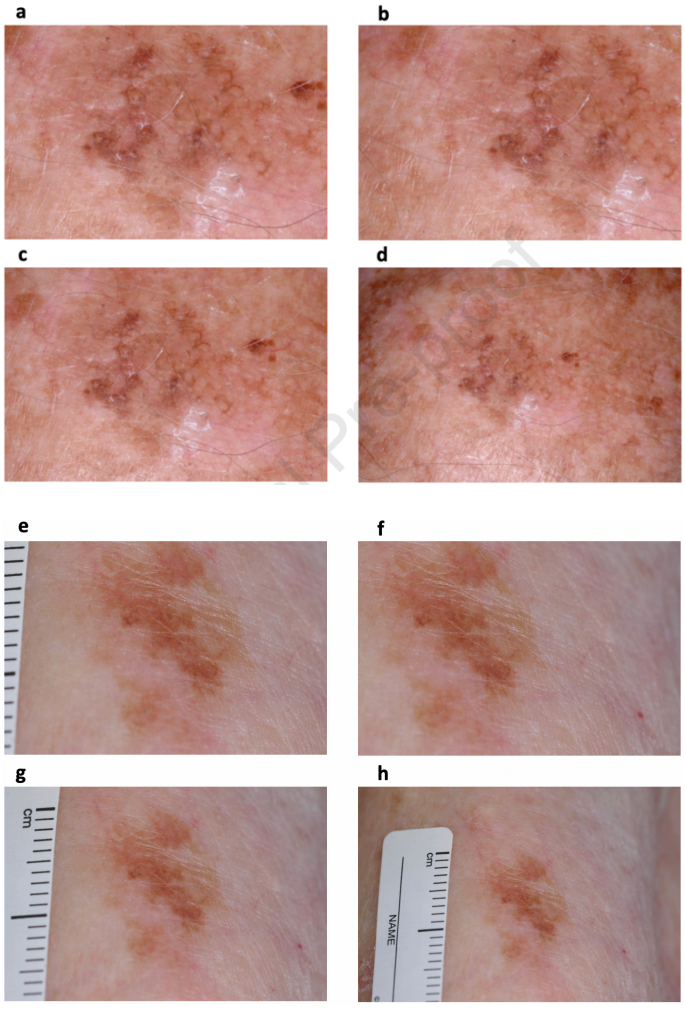


Figure 9.

Κλινικές εικόνες δύο μελανωμάτων και με τις τέσσερις πειραματικές συνθήκες που χρησιμοποιήσαμε, τα οποία διαγνώσθηκαν λανθασμένα από τον αλγόριθμο των Han et al.[59]

Melanoma in situ (A – D). A. 'Intended use', 1η διάγνωση πορορκεράτωση (πιθανότητα 27%) **B. 'Intended use, off centered'**, κύρια διάγνωση πορορκεράτωση (πιθανότητα 27%). **C. 'Drag and drop'**, κύρια διάγνωση lentigo (πιθανότητα 67%). **D. '1x magnification'**, κύρια διάγνωση lentigo (πιθανότητα 51%). **Melanoma 0.2 mm thickness (E-H). E.** 'Intended use', κύρια διάγνωση λεύκη (πιθανότητα 33%). **F. 'Intended use, off centered'**, κύρια διάγνωση λεύκη (πιθανότητα 27%). **G. 'Drag and drop'**, κύρια διάγνωση ουλή (πιθανότητα 35%). **H. '1x magnification'**, κύρια διάγνωση μη ειδική βλάβη, (πιθανότητα 52%).



Ερμηνευτική προσέγγιση της εφαρμογής των αλγορίθμων Τεχνητής Νοημοσύνης για τη διάγνωση του καρκίνου δέρματος στην κλινική πράξη.

Όπως εκθέσαμε και στην εισαγωγή, η έγκαιρη διάγνωση και η χειρουργική αφαίρεση του καρκίνου δέρματος, και ειδικότερα του μελανώματος, είναι ο κύριος τρόπος μείωσης της θνητότητας και της νοσηρότητας από αυτούς τους όγκους.[19-21] Υπάρχει μια πλειάδα εμπορικά διαθέσιμων, μη επεμβατικών συσκευών, οι οποίες μπορούν να χρησιμοποιηθούν κατά τη διάρκεια της κλινικής εξέτασης προκειμένου να διευκολύνουν τη διάγνωση του μελανώματος, συμπεριλαμβανομένων των Raman spectroscopy, multispectral instrumentation και εφαρμογών Τεχνητής νοημοσύνης, ενσωματωμένων σε συστήματα ψηφιακής δερματοσκόπησης.[57,65-68]

Σε μια σημαντική συνεισφορά στον τομέα της έγκαιρης διάγνωση του μελανώματος, οι Maclellan et al. διερεύνησαν την ευαισθησία και την ειδικότητα που προσφέρουν τρείς από αυτές τις συσκευές, και συγκεκριμένα τα FotoFinder®, Melafind®, Verisante Aura™, συγκρίνοντας τες με τις αντίστοιχες ενός κλινικού δερματολόγου, και ενός ειδικού στη δερματοσκόπηση σε 209 μελανοκυτταρικές βλάβες σε 184 ασθενείς οι οποίες υπεβλήθησαν σε βιοψία.[69] Στην έρευνα τους αυτή βρήκαν πως ο ειδικός στη δερματοσκόπηση, ο οποίος έκανε teleconsultation, είχε ευαισθησία 84.5% και ειδικότητα 82.6% για τη διάγνωση μελανώματος, ο κλινικός δερματολόγος είχε ευαισθησία 96.6% και ειδικότητα 32.2%, ενώ οι μη επεμβατικές συσκευές είχαν αντίστοιχα ευαισθησία και ειδικότητα: MelaFind® 82.5% και 52.4%, Verisante Aura™ 21.4% και 86.2%, ενώ το FotoFinder® Moleanalyzer Pro (88.1%, 78.8%). Στα συμπεράσματα αυτής τους της μελέτης καταλήγουν πως ο αλγόριθμος τεχνητής νοημοσύνης που είναι

ενσωματωμένος στο σύστημα ψηφιακής δερματοσκόπησης FotoFinder® θα μπορούσε να αποδειχθεί κλινικά χρήσιμος με βάση τις επιδόσεις του για τη διάγνωση μελανώματος.

Προκειμένου να διερευνήσουμε αυτή την υπόθεση, καθώς και για να μπορέσουμε να αξιολογήσουμε εν γένει την εφαρμογή της Τεχνητής Νοημοσύνης στη δερματολογία εν γένει, και ειδικότερα στον καρκίνο του δέρματος, αρχικά προσπαθήσαμε να επαληθεύσουμε τα ευρήματα της μελέτης αυτής και έπειτα επιχειρήσαμε μια διαφορετική προσέγγιση στη διαδικασία λήψης της απόφασης για βιοψία μιας ύποπτης βλάβης.[33] Θα ήταν βεβιασμένο να εξάγουμε συμπεράσματα για το generalizability των ευρημάτων των Maclellan et al. βασιζόμενοι σε αυτή τους τη μελέτη, δεδομένου ότι το dataset τους είχε μεγάλο bias ως προς τον επιπολασμό του μελανώματος (28.2% των βλαβών που συμπεριλήφθηκαν στη μελέτη ήταν μελάνωμα), και είναι μάλλον αμφίβολο ότι τα αποτελέσματα ως προς την ευαισθησία και την ειδικότητα των συσκευών που χρησιμοποιήθηκαν θα ήταν τα ίδια σε έναν γενικό πληθυσμό, με πολύ περισσότερες καλοήθεις βλάβες.

Επιπρόσθετα, ένα σημείο το οποίο δεν έχει τονιστεί ιδιαίτερα είναι ότι μπορεί να είναι δύσκολο να συγκρίνουμε την ενδεχόμενη χρησιμότητα ανταγωνιστικών διαγνωστικών στρατηγικών χρησιμοποιώντας παραδοσιακά μέτρα για την ταξινόμηση της διαγνωστικής ακρίβειας, όπως ευαισθησία, ειδικότητα και Area Under the Curve (AUC). Μια εναλλακτική προσέγγιση που προτείνουμε, και η οποία ενδεχομένως να μπορεί να βρει μια θέση στην αξιολόγηση της χρησιμότητας των εφαρμογών Τεχνητής Νοημοσύνης, είναι αυτή του Net Benefit και Decision Curve Analysis (DCA).[70,71] Αρχικά χρειάζεται να καθορίσουμε το σημείο ανταλλαγής μεταξύ δυο διαφορετικών αποτελεσμάτων, όπως τη βιοψία ενός μελανώματος έναντι της βιοψίας ενός σπίλου. Ως θεωρητικό παράδειγμα μπορούμε να εκλάβουμε ότι ένας δερματολόγος μπορεί να θεωρήσει ότι το ‘κακό’ που προξενεί η βιοψία 9 σπίλων, δηλαδή καλοηθών μελανοκυτταρικών βλαβών, ισούται με το ‘καλό’ -

benefit, της βιοψίας ενός μελανώματος. Αυτό παράγει ένα exchange rate 1:9 και αντίστοιχα ένα probability threshold της τάξεως του 10%. Κατά αντιστοιχία, ένας δερματολόγος θα ήταν πρόθυμος να προβεί σε βιοψία μιας μελανοκυτταρικής βλάβης εάν το ρίσκο η εν λόγω βλάβη να είναι μελάνωμα είναι της τάξεως του 10%, αλλά όχι αν είναι 9% ή μικρότερο. Το exchange rate, η ευαισθησία και η ειδικότητα μιας διαγνωστικής μεθόδου χρησιμοποιούνται προκειμένου να παραχθεί το Net Benefit (Net Benefit= (true positives/n) – [(false positives/n) × (weighting factor)]. weighting factor = threshold probability/(1 – threshold probability).

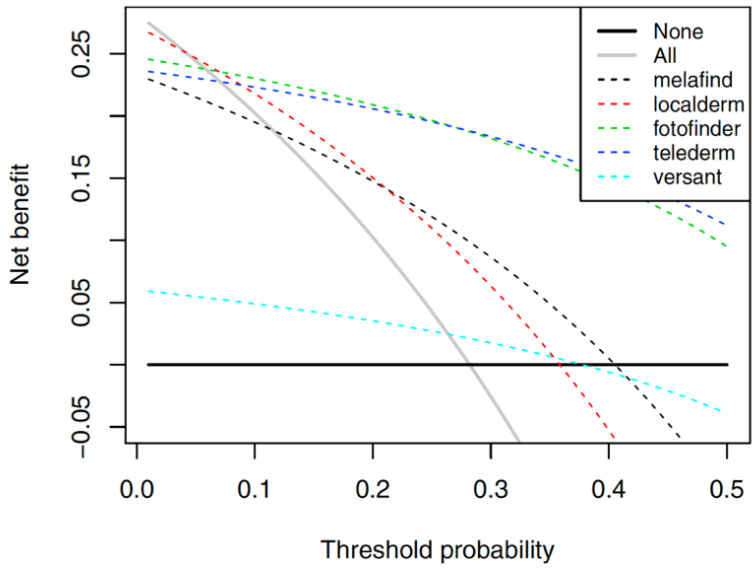
Επειδή υπό διαφορετικές συνθήκες πιθανά μεταβάλλεται το exchange rate, οι δερματολόγοι και οι ασθενείς επί παραδείγματι μπορεί να έχουν διαφορετικό exchange rate, είναι χρήσιμο να υπολογιστεί το Net Benefit σε ένα μεγάλο εύρος κλινικών πιθανοτήτων και σεναρίων, το οποίο παράγει το Decision Curve. Για τη μελέτη των Maclellan et al. υπολογίσαμε το εύρος αυτό μεταξύ 5% και 10% για την βιοψία ενός μελανώματος.[69] Χρησιμοποιώντας τα στοιχεία για τη διαγνωστική ακρίβεια που παρουσίασαν στη μελέτη τους υπολογίσαμε τα Decision Curves για τις 5 πειραματικές συνθήκες που παρουσιάστηκαν (ειδικός στη δερματοσκόπηση, κλινικός δερματολόγος και οι 3 εμπορικά διαθέσιμες συσκευές). Συγκεκριμένα, υπολογίσαμε τα Decision Curves για τις ανταγωνιστικές διαγνωστικές μεθόδους σε όλο το εύρος, από το να υφίστανται βιοψία όλες οι βλάβες (ευαισθησία 100%, ειδικότητα 0%) έως του να μην υφίσταται βιοψία καμία από τις βλάβες (ευαισθησία 0%, ειδικότητα 100%). **(Figure 10)**

Βρήκαμε ότι η βέλτιστη προσέγγιση εξαρτάται εν πολλοίς από το probability threshold. Εάν το όριο αυτό είναι της τάξεως του 5%-7% (exchange rates από 19:1 έως 13:1), ο κλινικός δερματολόγος, ο οποίος διαθέτει και την υψηλότερη ευαισθησία συγκρινόμενος με τις υπόλοιπες πειραματικές μεθόδους, έχει το υψηλότερο Net Benefit. Εάν αυξήσουμε το επιθυμητό probability threshold στο εύρος 8% έως 10% (exchange rates από 12:1 έως 9:1), το σύστημα Τεχνητής Νοημοσύνης του

FotoFinder έχει το υψηλότερο Net Benefit. Ωστόσο, οι απόλυτες διαφορές στο Net Benefit μεταξύ των στρατηγικών οι οποίες περιλαμβάνουν βιοψία όλων των βλαβών, βιοψία των βλαβών με βάση τη διαγνωστική ακρίβεια του κλινικού δερματολόγου, και βιοψία με βάση το αποτέλεσμα του αλγορίθμου του FotoFinder είναι μικρές σε αυτά τα thresholds, ειδικά με ένα dataset το οποίο περιλαμβάνει τόσο μεγάλο ποσοστό κακοηθειών (28.2%). Μολαταύτα, η χρήση του Net Benefit και των Decision Curve Analyses, θα μπορούσε να είναι ένας χρήσιμος δείκτης για την αξιολόγηση της διαγνωστικής ακρίβειας, και της χρησιμότητας που θα μπορούσαν να έχουν οι αλγόριθμοι Τεχνητής Νοημοσύνης στην κλινική πράξη.

Figure 10.

Decision Curves για τις διαγνωστικές προσεγγίσεις που χρησιμοποιήθηκαν από τους Maclellan et al. έχοντας το Net Benefit ως μεταβλητή του threshold probability.[69] Ανάμεσα στις διαφορετικές προσεγγίσεις, η βιοψία όλων των βλαβών έχει το υψηλότερο Net Benefit για threshold probabilities από 1% έως 4%. Βιοψία με βάση τον κλινικό δερματολόγο έχει το υψηλότερο Net Benefit για threshold probabilities από 5% έως 7%. Και βιοψία βλαβών με βάση τα αποτελέσματα του αλγορίθμου του FotoFinder έχει το υψηλότερο Net Benefit για threshold probabilities από 8% έως 10%.



Περιγραφή νέων δερματοσκοπικών κριτηρίων για τη διάγνωση του καρκίνου δέρματος και μιμητών του και η Expert Agreement Study on Dermoscopy μελανοκυτταρικών βλαβών.

Όπως περιγράψαμε και στην εισαγωγή, η δερματοσκόπηση είναι μια εύχρηστη, μη επεμβατική τεχνική, η οποία έχει αποδειχθεί ότι βελτιώνει τη διαγνωστική ακρίβεια των κλινικών ιατρών για τη διάγνωση του καρκίνου δέρματος.[23,72] Τα τελευταία 30 χρόνια έχουν καταβληθεί σημαντικές προσπάθειες για την περιγραφή δερματοσκοπικών διαγνωστικών κριτηρίων τα οποία βοηθούν στη διαφορική διάγνωση του καρκίνου δέρματος από καλοήθεις βλάβες που μπορεί να τον μιμούνται.[73-85] Η αναγνώριση αυτών των κριτηρίων έχει σημαίνοντα ρόλο κατά την κλινική εξέταση μιας ύποπτης βλάβης για καρκίνο δέρματος, ενώ μπορεί να διαδραματίσει σημαντικό ρόλο και για την εκπαίδευση των αλγορίθμων Τεχνητής Νοημοσύνης Machine Learning προκειμένου να αυξηθεί η διαγνωστική τους ακρίβεια.[30,38,39,86]

Κατά τη διάρκεια της διατριβής αυτής προβήκαμε στην περιγραφή καινούργιων δερματοσκοπικών κριτηρίων τόσο για τη διάγνωση του μελανώματος, όσο και του βασικοκυτταρικού καρκινώματος, καθώς και ενός συχνού μιμητή του καρκίνου δέρματος - του Lichen Planus-Like Keratosis (LPLK). Τέλος, προκειμένου τα δερματοσκοπικά κριτήρια να μπορούν να χρησιμοποιηθούν για τη διάγνωση του μελανώματος, χρειάζεται να είναι αξιόπιστα και αναπαράξιμα, για αυτό το λόγο προβήκαμε και στην πρώτη Expert Agreement Study on Dermoscopy μελανοκυτταρικών βλαβών, η οποία προσδιόρισε όχι μόνο τη συμφωνία ειδικών στη δερματοσκόπηση από όλο τον κόσμο για την παρουσία δερματοσκοπικών κριτηρίων, αλλά και την εντόπιση τους εντός μιας βλάβης. Τα αποτελέσματα μας αυτά μπορούν να χρησιμοποιηθούν προκειμένου να καθοδηγήσουν τους διαγνωστικούς αλγορίθμους

που απευθύνονται σε κλινικούς ιατρούς οι οποίοι ασχολούνται με τη διάγνωση του καρκίνου δέρματος, και για να βελτιωθούν οι αλγόριθμοι Τεχνητής Νοημοσύνης για τη διάγνωση του καρκίνου δέρματος.

Δερματοσκοπική παρουσίαση Αμελανωτικού Μελανώματος των άκρων σε σύγκριση με τα Αιμαγγειώματα των άκρων.[34]

Εισαγωγή

Το μελάνωμα των άκρων έχει φτωχότερη πρόγνωση συγκριτικά με άλλους τύπους μελανώματος, οφειλόμενα κυρίως στη δυσκολία της διάγνωσης του και συνεπακόλουθα, στη διάγνωση του σε πιο προχωρημένα στάδια της νόσου.[87] Μια αναδρομική μελέτη 53 μελανωμάτων των άκρων σε ένα κέντρο βρήκε ότι σε τουλάχιστον 34% (n=18) εξ αυτών, αρχικά, χάθηκε η διάγνωση του μελανώματος. Από αυτές τις περιπτώσεις όπου έγινε λανθασμένη διάγνωση, το 50% (n=9) ήταν αμελανωτικά μελανώματα.[88] Η αμελανωτική παρουσίαση του μελανώματος των άκρων περιγράφεται πολύ σπάνια και τα κλινικά και δερματοσκοπικά της χαρακτηριστικά είναι άγνωστα. Οι Özdemir et al. περιέγραψαν ένα δερματοσκοπικό πρότυπο το οποίο ανευρίσκεται στην περιφέρεια των μελανωμάτων των άκρων ως “vascularized parallel ridge pattern,” το οποίο περιγράφεται ως ερύθημα και dotted vessels τα οποία πληρούν τα ridges και εξαιρούν τα furrows.[89] Ωστόσο, τα αιμαγγειώματα των άκρων έχουν επίσης περιγραφεί να παρουσιάζουν ένα παρόμοιο πρότυπο με αγγεία διατεταγμένα σε parallel ridge pattern στη δερματοσκόπηση.[90,91] Σε αυτή μας τη μελέτη περιγράψαμε τη δερματοσκοπική παρουσίαση ενός αμελανωτικού μελανώματος των άκρων και συγκρίναμε τα δερματοσκοπικά ευρήματα μας με αυτά των αιμαγγειωμάτων των άκρων.[34]

Ευρήματα

Το περιστατικό που περιγράψαμε αφορά σε έναν ενήλικα άνδρα ο οποίος παρουσιάστηκε με μια καινούργια διάγνωση μελανώματος στο μεγάλο δάκτυλο του αριστερού ποδός. Ο ασθενής ανέφερε μια διαρκώς επιδεινούμενη δυστροφία του όνυχος κατά τη διάρκεια της τελευταίας πενταετίας. Η βιοψία όνυχος επιβεβαίωσε τη διάγνωση μελανώματος, βάθους τουλάχιστον 0.57 mm κατά Breslow. Η παρατήρηση της περιφέρειας του δακτύλου ανέδειξε μια κλινικά ύποπτη περιοχή μελάγχρωσης καθώς και μια προεξάρχουσα ερυθρή περιοχή όγκου στη μεσότητα της περιφέρειας. Η δερματοσκοπική εξέταση της εξέρυθρης περιοχής ανέδειξε πολλαπλά στικτά αγγεία, χαοτικά διανεμημένα στα ridges, χωρίς να περιλαμβάνονται τα furrows **(Figure 11A)**. Η punch βιοψία της ερυθρής πλάκας έδειξε μελάνωμα in situ, με συμμετοχή τόσο του εκκρινούς πόρου, όσο και επέκταση στους εκκρινείς αδένες **(Figure, 11C)**. Ο ασθενής υπεβλήθη σε εγχείρηση ακρωτηριασμού του δακτύλου στο επίπεδο της άπω φαλαγγο-φαλαγγικής άρθρωσης και το τελικό βάθος Breslow του όγκου ήταν 4.6 mm, ενώ η βιοψία φρουρού λεμφαδένα ήταν αρνητική για κακοήθεια. Επιπλέον συγκρίναμε τα δερματοσκοπικά ευρήματα μας στο αμελανωτικό μελάνωμα άκρων με τα δερματοσκοπικά ευρήματα σε 3 ασθενείς με αιμαγγειώματα άκρων. Όλα τα αιμαγγειώματα παρουσίαζαν parallel ridge pattern, αποτελούμενο από ερυθρές προς ερυθροϊώδεις κουκίδες. Σε αντίθεση με το αμελανωτικό μελάνωμα, τα στικτά αγγεία στα αιμαγγειώματα είχαν οργανωμένη κατανομή στην περιφέρεια των ridges, χωρίς να συμπεριλαμβάνουν τους εκκρινείς πόρους **(Figure, 11B)**. Ιστοπαθολογικά, η εξέταση ενός εκ των αιμαγγειωμάτων ανέδειξε εστιακούς σχηματισμούς αγγείων, τα οποία επεκτείνονταν στις χοριακές λάχνες, χωρίς να συμπεριλαμβάνουν τους εκκρινείς αδένες **(Figure, 11D)**.

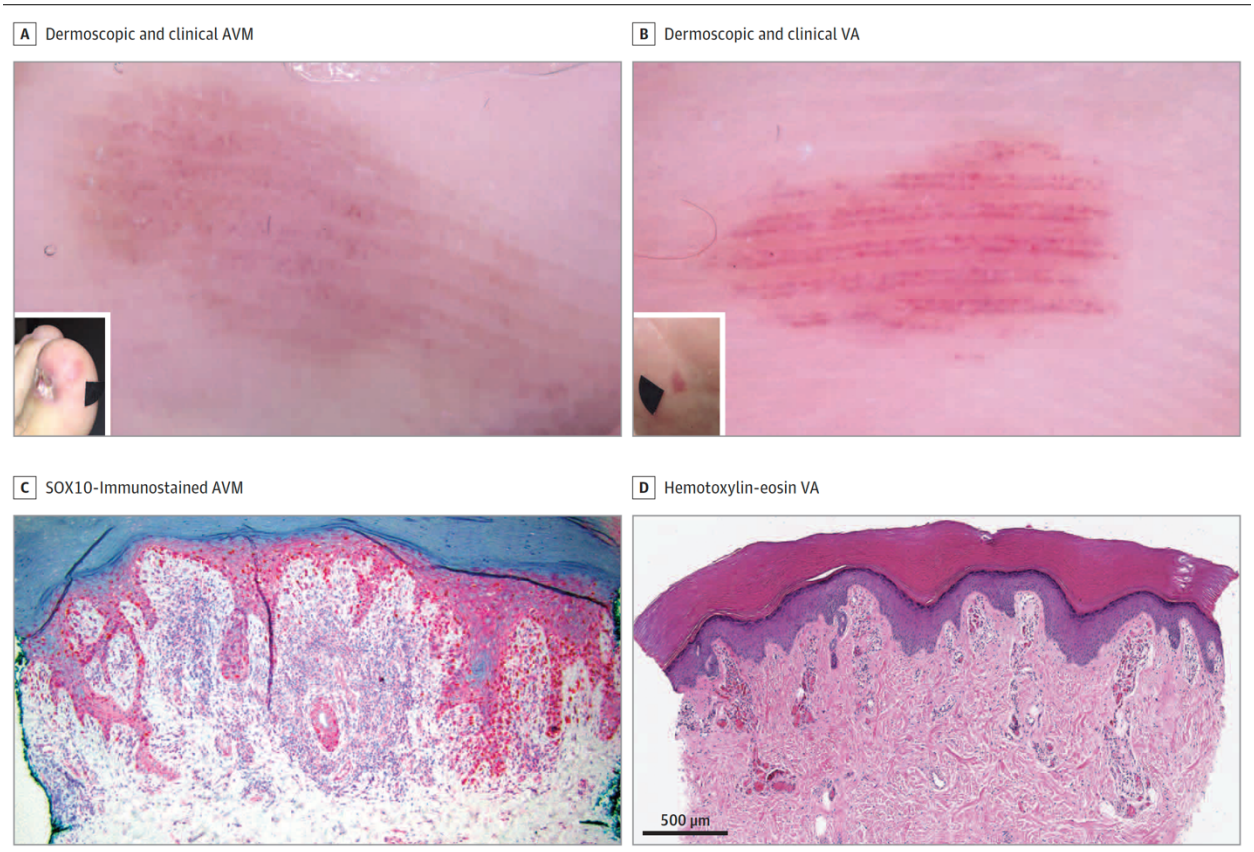
Συζήτηση

Η παρουσία του μελαγχρωματικού parallel ridge pattern έχει δειχθεί σε πλειάδα μελετών να σχετίζεται με τη διάγνωση του acral lentiginous melanoma και είναι ιδιαίτερα χρήσιμο για τη διάγνωση του μελανώματος in situ των άκρων.[92-94] Στη μελέτη μας αυτή περιγράψαμε μια περίπτωση προεξαρχόντως αμελανωτικού μελανώματος των άκρων το οποίο επεδείκνυε στη δερματοσκόπηση ένα αγγειακό parallel ridge pattern, αποτελούμενο από χαοτικά διανεμημένα στικτά αγγεία. Επιπροσθέτως, στη μελέτη μας συγκρίναμε την παρουσίαση των αμελανωτικών μελανωμάτων των άκρων με αυτή των αιμαγγειωμάτων των άκρων, τα οποία είναι η κύρια διαφορική διάγνωση, και βρήκαμε πως τα αιμαγγειώματα παρουσιάζονται επίσης με αγγειακό parallel ridge pattern. Το parallel ridge pattern των αιμαγγειωμάτων χαρακτηρίζεται ως linear,double-dotted ridge pattern και αποτελείται από ερυθρές προς ερυθροϊώδεις κουκίδες (dotted vessels), τα οποία διανέμονται στην περιφέρεια των ridges, ενώ σημαντικό διαφοροδιαγνωστικό στοιχείο είναι ότι δεν εμπλέκουν τους εκκρινείς πόρους.[90,91] Οι παρατηρήσεις μας αυτές θα χρειαστούν επιβεβαίωση στο μέλλον, ωστόσο μπορούν να αποδειχθούν δυνητικά χρήσιμες για τη διαφορική διάγνωση του αμελανωτικού μελανώματος των άκρων από τα αιμαγγειώματα των άκρων.

Figure 11.

A και **B,** Δερματοσκόπηση πολωμένου φωτός, μεγέθυνση x10. **A.** Δερματοσκοπική και κλινική (ένθετο) φωτογραφία αμελανωτικού μελανώματος άκρων με parallel ridge pattern και χαοτική διανομή ερυθρών στικτών αγγείων με συμμετοχή των εκκρινών πόρων. **B.** Δερματοσκοπική και κλινική (ένθετο) φωτογραφία αιμαγγειώματος άκρων με parallel ridge pattern αποτελούμενο από στικτά αγγεία τα οποία είναι κατανεμημένα σε σειρά στα άκρα των ridges και χωρίς συμμετοχή των εκκρινών πόρων. **C.** Ανοσοϊστοχημική ανάλυση του αμελανωτικού μελανώματος in situ το οποίο δείχνει άτυπα

μελανοκύτταρα στην επιδερμίδα και συμμετοχή των εκκρινών πόρων, με ήπια αύξηση στην πυκνότητα των αγγείων στο χόριο, συσχετιζόμενη πιθανώς με φλεγμονή. **D.** Ιστοπαθολογική εικόνα ενός εκ των αιμαγγειωμάτων που εξετάσαμε, η οποία αναδεικνύει αγγειακούς σχηματισμούς που φτάνουν μέχρι τις χοριακές λάχνες χωρίς να συμπεριλαμβάνουν τους εκκρινείς πόρους.



Συσχέτιση των Multiple Aggregated Yellow-White Globules με τη διάγνωση του μη-μελαγχρωματικού Βασικοκυτταρικού Καρκινώματος.[37]

Εισαγωγή

Το Βασικοκυτταρικό καρκίνωμα (BCC) είναι ο πιο συχνός καρκίνος στον κόσμο.[95] Η επίπτωση του BCC αυξάνει διαρκώς με περισσότερες από 2 εκατομμύρια περιπτώσεις να διαγιγνώσκονται ετήσια στις ΗΠΑ.[12] Τα δερματοσκοπικά κριτήρια για τη διάγνωση του βασικοκυτταρικού καρκινώματος περιεγράφηκαν το 2000 από τους Menzies et al. και περιλαμβάνουν τα blue-gray ovoid nests, multiple non aggregated blue-gray globules, ulceration, arborizing telangiectasia, spoke-wheel structures, και τις leaflike areas.[73] Πιο πρόσφατα, προστέθηκε το κριτήριο των shiny white structures, και ειδικότερα των shiny white blotches and strands, ως καινούργιο διαγνωστικό κριτήριο για το BCC.[96] Αυτά τα κριτήρια έχουν επιβεβαιωθεί ότι προσφέρουν υψηλή διαγνωστική ακρίβεια για τη διάγνωση του BCC με μια συνολική ευαισθησία της τάξεως του 91.2% και ειδικότητα της τάξεως του 95% σύμφωνα με μια πρόσφατη μετανάλυση.[97] Ωστόσο, η ευαισθησία και η ειδικότητα των διαγνωστικών αυτών κριτηρίων για τα μη-μελαγχρωματικά BCC είναι χαμηλότερες (84.3% ευαισθησία και 73.2% ειδικότητα).[97]

Με αυτά τα δεδομένα, η ανεύρεση νέων δερματοσκοπικών κριτηρίων για τη διάγνωση των μη-μελαγχρωματικών BCC είναι αναγκαία.[98] Βρήκαμε πως κάποια BCC επιδεικνύουν στη δερματοσκόπηση πολλαπλά aggregated yellow-white (MAY) globules. Αυτό το δερματοσκοπικό κριτήριο διαφέρει από προηγουμένως περιγεγραμμένα λευκά ή κιτρινωπά κριτήρια στα BCC.[99,100] Προκειμένου να διερευνήσουμε τη συχνότητα εμφάνισης αυτού του κριτηρίου στα BCC καθώς και τη διαγνωστική του ακρίβεια κάναμε μια αναδρομική μελέτη σε κλινικές και δερματοσκοπικές εικόνες βλαβών οι οποίες είχαν το μη-μελαγχρωματικό BCC ως πιθανή διαφορική διάγνωση.

Μέθοδοι

Αυτή ήταν μια αναδρομική, case control μελέτη, η οποία πραγματοποιήθηκε από την 1η Ιουλίου 2017, έως την 1η Ιουλίου 2019. Όλες οι εικόνες που συμπεριλήφθηκαν στη μελέτη προήλθαν από μια deidentified database συνεχόμενων βλαβών από μια δερματολογική κλινική στο Plantation, Florida. Δεδομένου του σχετικά χαμηλού αριθμού αμελανωτικών μελανωμάτων σε αυτό το dataset ελέγξαμε 2169 μελανώματα από τη βιβλιοθήκη του International Skin Imaging Collaboration (ISIC) archive, και 22 αμελανωτικά μελανώματα ανευρέθηκαν και συμπεριλήφθηκαν στη μελέτη μας.[10] Η μελέτη αυτή εγκρίθηκε από την επιτροπή ηθικής του Memorial Sloan Kettering Cancer Center. Οι εικόνες που αναλύθηκαν ήταν αποκλειστικά close-up, μεγεθυμένες κλινικές καθώς και δερματοσκοπικές εικόνες, ενώ κανένα αναγνωρίσιμο χαρακτηριστικό των ασθενών δεν ανευρέθηκε σε καμία από τις εικόνες που αναλύθηκαν. Τρείς από τους ερευνητές (C.N.-D., K.L, and A.R.) εξετάσαμε όλες τις εικόνες του dataset και συμπεριλάβαμε μόνο εκείνες τις βλάβες οι οποίες είχαν ιστοπαθολογική επιβεβαίωση και ήταν κλινικά μη-μελαγχρωματικές. Αποκλείσαμε εικόνες βλαβών οι οποίες ιστοπαθολογικά ήταν collision tumors, είχαν αμφίβολη ιστοπαθολογική διάγνωση, ή δεν είχαν πολωμένες και μη-πολωμένες δερματοσκοπικές εικόνες. Συμπεριλάβαμε όλους τους ιστολογικούς υπότυπους BCC ενώ ως controls συμπεριλάβαμε βλάβες οι οποίες συνήθως εισέρχονται στη διαφορική διάγνωση του μη-μελαγχρωματικού BCC: ακανθοκυτταρικό καρκίνωμα, (SCC), χοριακούς σπίλους - intradermal nevus (IDN), αμελανωτικό μελάνωμα, lichen planus–like keratosis (LPLK), δεσμοπλαστικό τριχοεπιθηλίωμα (DT), εξαρτηματικούς όγκους (όπως, fibrofolliculoma), και φλεγμονώδεις παθήσεις (όπως δερματίτιδα και ψωρίαση). Τα δημογραφικά χαρακτηριστικά των ασθενών, η ανατομική τοποθεσία των βλαβών και ο ιστολογικός υπότυπος των όγκων καταγράφηκαν.

Κλινικές και Δερματοσκοπικές εικόνες

Οι κλινικές και δερματοσκοπικές εικόνες ελήφθησαν με μια κάμερα Nikon 1 (Nikon USA Inc) και ένα Dermlite DL2 pro HR για τις δερματοσκοπικές εικόνες (3Gen). Συμπεριλάβαμε τόσο πολωμένες, όσο και μη-πολωμένες δερματοσκοπικές εικόνες. Οι δερματοσκοπικές εικόνες αναλύθηκαν από τρεις εκ των ερευνητών (C.N.-D., K.L., and A.R.), οι οποίοι ήταν blinded ως προς τη διάγνωση για consensus. Ένας τέταρτος αξιολογητής (A.A.M.) επέλυε την όποια διαφωνία προέκυπτε. Οι δερματοσκοπικές εικόνες αναλύθηκαν για τα δερματοσκοπικά κριτήρια όπως περιεγράφηκαν στο πιο πρόσφατο consensus για τη δερματοσκόπηση.[39] Αναλύσαμε ειδικότερα τα δερματοσκοπικά κριτήρια για BCC, ενώ το κύριο δερματοσκοπικό κριτήριο το οποίο αναζητούσαμε ήταν τα MAY globules, τα οποία περιεγράφηκαν από εμάς ως multiple, aggregated, white-to-yellowish globules arranged in clusters. Το εν λόγω δερματοσκοπικό κριτήριο είναι εμφανές τόσο με το πολωμένο όσο και με το μη-πολωμένο φως στη δερματοσκόπηση διαχωρίζοντας το από τα shiny white structures (blotches and strands) και από τα milia-like cysts αντίστοιχα **(Figure 12)**. Αξιολογήσαμε τις δερματοσκοπικές εικόνες για την παρουσία ή την απουσία των MAY globules σε consensus οι τρεις ερευνητές (C.N.-D., K.L., and A.R.) για την πλειονότητα των βλαβών που συμπεριλάβαμε. Προκειμένου να αξιολογήσουμε το interrater agreement για την παρουσία των MAY globules σε 150 συνεχόμενες βλάβες, οι ίδιοι ερευνητές πραγματοποιήσαμε ανεξάρτητη αξιολόγηση.

Reflectance Confocal Microscopy, Optical Coherence Microscopy και Histopathological correlation

Σε μια υποκατηγορία βλαβών οι οποίες διαγνώστηκαν προοπτικά και επεδείκνυαν τα MAY globules εφαρμόσαμε reflectance confocal microscopy (RCM) και optical coherence tomography(OCT) πριν την πραγματοποίηση βιοψίας. Οι εικόνες ελήφθησαν με ένα handheld RCM device (VivaScope 1500 and/or 3000; Caliber ID), ενώ για τις εικόνες OCT χρησιμοποιήσαμε μια συσκευή η οποία αναπτύχθηκε

πρόσφατα με τη συμμετοχή συνεργατών μας, καθώς και τα κριτήρια που περιγράφηκαν στην ίδια μελέτη.[101] Τα κριτήρια που χρησιμοποιήθηκαν για το RCM είναι αυτά που περιγράφηκαν σε μια πρόσφατη systematic review.[102] Σε αυτές τις βλάβες το histopathologic correlation πραγματοποιήθηκε με την τεχνική του precision biopsy.[103]

Κύρια Ευρήματα

Το κύριο εύρημα της μελέτης μας ήταν η κατανομή της παρουσίας ή της απουσίας των MAY globules για τη διάγνωση του BCC συγκρινόμενο με όλες τις άλλες διαγνώσεις οι οποίες είχαν τον ρόλο του control. Δευτερεύοντα αποτελέσματα ήταν η κατανομή των MAY globules κατά υπότυπο BCC, καθώς και κατά ανατομική τοποθεσία του όγκου. Προκειμένου να αναλύσουμε τους διαφορετικούς ιστολογικούς υπότυπους του BCC τους χωρίσαμε σε high risk (morpheaform και infiltrative) και σε low risk (superficial and nodular) BCCs.

Αποτελέσματα

Η αναδρομική εξέταση 2555 βλαβών ανέδειξε 643 βλάβες οι οποίες θα μπορούσαν να συμπεριληφθούν στη μελέτη μας σε 621 ασθενείς με κλινικά μη-μελαγχρωματικούς όγκους. Οι 9 βλάβες αποκλείσθηκαν λόγω: μη πραγματοποίησης βιοψίας (4 όγκοι), collision tumors (4 όγκοι) και μη διαθέσιμης, μη-πολωμένης δερματοσκοπικής εικόνας (1 όγκος). Η τελική μας ανάλυση συμπεριέλαβε 656 βλάβες (634 βλάβες από το database και 22 αμελανωτικά μελανώματα από το ISIC archive) σε 643 ασθενείς. Η μέση (SD) ηλικία του cohort μας ήταν τα 63.1 (14.9) χρόνια, και 381 (58.1%) εκ των ασθενών ήταν άνδρες. Οι 194 (29.6%) βλάβες ήταν στον τράχηλο και το κεφάλι. Ένα σύνολο 291 βλαβών σε 278 ασθενείς (44.4%; μέση [SD] ηλικία, 61.9 [14.9]; 190 [64.3%] άνδρες) ήταν BCCs (cases), και 365 βλάβες (55.6%) σε 365 ασθενείς (μέση [SD] ηλικία, 63.9 [14.9]; 191 [53.1%] άνδρες)

αντιστοιχούσαν σε άλλες διαγνώσεις (controls). Το μέσο (SD) μέγεθος όγκων σε ολόκληρο το cohort ήταν 7.6 (4.9) mm, με ένα μέσο (SD) μέγεθος όγκου 6.8 (4.8) mm για τα BCC και 8.2 (4.9) mm για τις υπόλοιπες διαγνώσεις (P < .001). Οι διαγνώσεις των ασθενών και οι υπότυποι των BCC του cohort μας παρουσιάζονται στον **Table 5**. Τα BCCs εντοπίζονταν στο κεφάλι και τον τράχηλο σε 124 (42.6%) των ασθενών και στον κορμό και στα άκρα σε 167 (57.4%) των ασθενών. Αναφορικά με τις άλλες διαγνώσεις, οι όγκοι εντοπίζονταν στο κεφάλι και τον τράχηλο σε 70 ασθενείς (19.5%) και στον κορμό και τα άκρα σε 289 ασθενείς (79.2%) (P<.001). Οι υπότυποι των BCC στη μελέτη μας ήταν: nodular για 224 βλάβες (76.7%), superficial για 27 (9.2%), infiltrative για 24 (8.2%), morpheaform for 8 (2.7%), sclerosing για 2 (0.7%), keratotic για 2 (0.7%), ινοεπιθηλίωμα του Pinkus για 1 (0.3%), και basosquamous για 1 (0.3%). Για 2 όγκους, δεν υπήρχαν διαθέσιμα δεδομένα για τον υπότυπο.

Διαγνωστικά κριτήρια

Τα MAY globules ανευρέθηκαν σε 64 από τις 656 βλάβες που συμπεριλήφθηκαν (9.8%; 95% CI, 7.6%-12.3%). Τα MAY globules παρατηρήθηκαν σε 61 από τα 291 BCC (21.0%; 95%CI, 16.4%-26.1%) και σε 3 από τις 365 περιπτώσεις με άλλες διαγνώσεις (0.8%; 95% CI, 0%-2.3%) (P < .001). Η παρουσία των MAY globules στα BCCs συσχετίσθηκε με μια ευαισθησία της τάξεως του 20.9% (95%CI, 16.4%-26.1%), και μια ειδικότητα της τάξης του 99.2% (95%CI, 97.6%-99.8%), ένα positive predictive value της τάξεως του 95.3% (95% CI, 86.6%-94.5%), και ένα negative predictive value της τάξεως του 61.0%(95% CI, 59.6%-62.4%). Το odds ratio για τη διάγνωση του BCC ήταν 32.0 (96% CI,9.9-103.2). Το positive likelihood ratio ήταν 25.4 (96% CI, 8.0-80.0), και το negative likelihood ratio ήταν 0.8 (96% CI, 0.7-0.8).

Ανατομική Εντόπιση

Αξιολογώντας την παρουσία των MAY globules στις βλάβες που εντοπίζονταν στο κεφάλι και τον τράχηλο(n = 194), 51 ασθενείς (26.3%) παρουσιάστηκαν με αυτό το κριτήριο. Συνολικά 124 εκ των 194

βλαβών (63.9%) που εντοπίζονταν στο κεφάλι και τον τράχηλο αντιστοιχούσαν σε BCC. Εκ των BCC που εντοπίζονταν στο κεφάλι και τον τράχηλο, 48 εκ των 124 (38.7%) επεδείκνυαν τα MAY globules, συγκριτικά με 3 εκ των 70 βλαβών (4.2%) με άλλες διαγνώσεις (P < .001). Το odds ratio για τη διάγνωση του BCC όταν τα MAY globules ήταν παρόντα σε μια βλάβη ήταν 14.1 (95%CI, 4.2-47.4) για τις βλάβες οι οποίες εντοπίζονταν στο κεφάλι και τον τράχηλο στο cohort που εξετάσαμε. Οι 3 βλάβες που παρουσιάστηκαν με το κριτήριο και δεν ήταν BCC ήταν ιστοπαθολογικά DT (n = 2) και SCC (n = 1).

Ανάλυση Υποτύπου BCC

Τα MAY globules παρατηρήθηκαν σε 18 εκ των 32 high-risk BCCs (56.2%) (ήτοι, infiltrative και morpheaform) και σε 41 εκ των 210 low-risk BCCs (19.5%) (P < .001) **(Figure 12, Figure 13A, and Figure 14A)**. Τα MAY globules ήταν 6.5 φορές πιο πιθανό να παρατηρηθούν σε BCC υψηλότερου ρίσκου συγκριτικά με αυτά χαμηλού ρίσκου (odds ratio, 6.5; 95%CI, 3.1-14.3). Το κριτήριο αυτό δεν παρατηρήθηκε σε κανένα εκ των 27 superficial BCCs **(Table 6).**

Interrater Agreement

Παρατηρήσαμε πολύ υψηλό interrater agreement για την παρουσία των MAY globules (κ = 0.89; 95% CI, 0.75-0.94). Το interrater agreements για τα υπόλοιπα δερματοσκοπικά κριτήρια BCC ήταν 0.94 (95% CI, 0.94-0.97) για τα arborizing vessels, 0.83 (95% CI, 0.81-0.90) για τα shiny white structures (blotches and strands), 0.78 (95% CI, 0.63-0.84) για τα in-focus dots, και 0.73 (95% CI, 0.66-0.77) για το ulceration **(Table 6).**

Κριτήρια RCM και OCT

Κατά την εξέταση υπό RCM, όλες από τις 4 περιπτώσεις που συμπεριλάβαμε προοπτικά είχαν hyperreflective amorphous areas (**Figure 13B & C** και **Figure 14B)** σε προσθήκη των κλασσικών

κριτηρίων για το BCC (tumor nests with palisading and clefting). Η εξέταση OCT ήταν διαθέσιμη για 2 περιπτώσεις και ανέδειξε hyperreflective areas, οι οποίες παρήγαγαν optical shadow **(Figure 13D** και **Figure 14C)**.

Histopathologic Correlation

Στις 4 βλάβες που εξετάσαμε με precision biopsy, τα aggregated yellow globules συσχετίστηκαν με μεμονωμένες, στρογγυλές περιοχές δυστροφικής ασβεστοποίησης, εντός, ή πέριξ των tumor nodules και με την παρουσία ασβεστοποιημένων κερατινικών κύστεων **(Figure 13E & F** και **Figure 14D)**. Επιπλέον, 2 εκ των περιπτώσεων που αναλύθηκαν ιστοπαθολογικά ανέδειξαν μικρές εναποθέσεις ασβεστίου στο superficial dermis, σε συσχέτιση με μικρές κερατινικές κύστεις.

Συζήτηση

Σε αυτή την αναδρομική, case-control μελέτη 656 βλαβών σε 643 ασθενείς με μη-μελαγχρωματικούς όγκους, βρήκαμε ότι η παρουσία του κριτηρίου των MAY globules συσχετίζεται με τη διάγνωση του BCC. Επιπλέον, βρήκαμε ότι η παρουσία των MAY globules συσχετίζεται με υψηλού ρίσκου υπότυπους BCC. Παρότι και άλλοι όγκοι εκτός των BCC μπορεί να επιδεικνύουν λευκωπά δερματοσκοπικά κριτήρια, όπως milia-like cysts ή και shiny white structures, παρατηρήσαμε ότι η παρουσία των MAY globules παρατηρήθηκε σχεδόν αποκλειστικά σε BCCs. Μολονότι το κριτήριο αυτό παρατηρήθηκε σε μόλις 21.0% των μη-μελαγχρωματικών BCC που εξετάστηκαν, η συχνότητα αυτή δε διαφέρει σημαντικά από τη συχνότητα άλλων BCC-specific δερματοσκοπικών κριτηρίων, όπως τα spoke-wheel structures, τα concentric structures, και τα leaf-like areas, τα οποία έχουν αναφερόμενη συχνότητα η οποία κυμαίνεται από 8% έως 20%.[73,104] Ωστόσο, όταν ήταν παρόντα, τα MAY globules συσχετίζονταν με τη διάγνωση του BCC και ειδικότερα με υπότυπους BCC υψηλού ρίσκου.

Η δερματοσκόπηση έχει βελτιώσει τη διαγνωστική ακρίβεια για τους ερυθρούς όγκους, απεικονίζοντας στοιχεία και δομές οι οποίες είναι αόρατες δια γυμνού οφθαλμού, βελτιώνοντας έτσι τη διαγνωστική ακρίβεια για τα SCC, τα αμελανωτικά μελανώματα και τα BCC μεταξύ άλλων.[98]

Παρόλα αυτά, οι υπομελανωτικές και αμελανωτικές βλάβες παραμένουν δύσκολες στη διάγνωση τους αποκλειστικά με τη χρήση δερματοσκόπησης.[36,98] Μια άλλη κατηγορία βλαβών η οποία μπορεί να είναι δύσκολη στη διάγνωση της με τη χρήση μόνο της δερματοσκόπησης είναι αυτή των BCC που παρουσιάζονται ως λευκές βλατίδες στο κεφάλι και τον τράχηλο ατόμων με έντονη ηλιακή έκθεση.[35] Η παρουσία των MAY globules μπορεί να είναι ένα σημαντικό στοιχείο για τη διάγνωση του BCC σε αυτές τις περιπτώσεις, όπως και στη διαφοροδιάγνωση του BCC από άλλες διαγνώσεις με τις οποίες μπορεί να υπάρξει δυσκολία διάγνωσης, όπως οι χοριακοί σπίλοι.[98] Εκτός από το να βελτιώνει τη διαγνωστική ακρίβεια για τα BCC, η δερματοσκόπηση έχει επιπροσθέτως βελτιώσει τη δυνατότητα μας να αναγνωρίζουμε τον υπότυπο των BCC (κυρίως, superficial vs nodular), δυνατότητα η οποία μπορεί με τη σειρά της να καθοδηγήσει την θεραπεία ακόμα και χωρίς τη λήψη βιοψίας, κατά την κλινική εξέταση.[80,105] Στην μελέτη μας, τα MAY globules δεν παρατηρήθηκαν σε κανένα από τα superficial BCC τα οποία εξετάστηκαν. Σημαντικά, η παρουσία των MAY globules συσχετίστηκε με πιο επιθετικούς υπότυπους BCC. Συνεπώς, η παρουσία των MAY globules θα μπορούσε να διαδραματίσει σημαντικό ρόλο στην απόφαση χειρισμού ενός περιστατικού κατά τη διάρκεια της κλινικής εξέτασης, αποφεύγοντας τοπικές καταστρεπτικές θεραπείες, όπως υγρό άζωτο, καθώς και να καθοδηγήσει τον τρόπο λήψης της βιοψίας, ώστε να αποφευχθεί ενδεχόμενη λανθασμένη ταξινόμηση ενός επιθετικού τύπου BCC.[106]

Προηγούμενες μελέτες έχουν βρει πως ασβεστοποιήσεις μπορεί να είναι παρούσες ιστοπαθολογικά σε ένα 11% έως 21% των BCCs.[107,108] Επιπλέον οι Slowdkoska et al βρήκε ότι αυτές οι ασβεστοποιήσεις είναι πιο συνήθεις σε υψηλού ρίσκου ιστολογικούς υπότυπος BCC (44% σε infiltrative ή morpheaform) συγκριτικά με χαμηλού ρίσκου ιστολογικούς υπότυπους (22%).[107] Τα δερματοσκοπικά ευρήματα της

μελέτης μας επικυρώνουν αυτά τα αποτελέσματα, με τα MAY globules να είναι παρόντα προεξαρχόντως σε υψηλού ρίσκου BCCs (56%) συγκριτικά με ιστολογικούς υπότυπους χαμηλότερου ρίσκου (superficial and nodular, 19%).

Περιορισμοί

Η μελέτη μας περιορίζεται από την αναδρομική της φύση, καθώς και από τον περιορισμένο αριθμό καλοήθων όγκων οι οποίοι παρουσιάζονται με λευκές δομές στη δερματοσκόπηση (sebaceous hyperplasias, molluscum contagiosum και pilomatrixomas). Αυτές οι βλάβες ωστόσο συνήθως είναι πολλαπλές και σχετικά εύκολες στη διάγνωση τους με την κλινική εξέταση σε συνδυασμό με τη δερματοσκόπηση, και σπάνια υπεισέρχονται στη διαφορική διάγνωση των BCC, ή υφίστανται βιοψία. Ένας επιπλέον περιορισμός της μελέτης μας έγκειται στο ότι προκειμένου να εμπλουτίσουμε τους μη μελαγχρωματικούς όγκους οι οποίοι δεν είναι BCC ανατρέξαμε στο ISIC Archive, έτσι ενδέχεται η αληθής συχνότητα εμφάνισης των MAY globules να μην μπορεί να υπολογιστεί με ακρίβεια. Τέλος ο πληθυσμός που συμπεριελήφθη στη μελέτη μας αποτελείται από Καυκάσιους ασθενείς, με ιστορικό έντονης ηλιακής έκθεσης (κάτοικοι της Florida), περιορίζοντας έτσι το generalizability των ευρημάτων μας.

Συμπέρασμα

Τα αποτελέσματα μας δείχνουν πως τα MAY globules θα μπορούσαν να έχουν χρησιμότητα ως ένα καινούργιο δερματοσκοπικό κριτήριο για τη διάγνωση των μη μελαγχρωματικών BCC, και ειδικότερα, των BCC υψηλού ρίσκου. Αυτές οι δομές ενδέχεται να συνδέονται με ασβεστοποιήσεις.

Επικύρωση των αποτελεσμάτων μας θα χρειαστούν με δεδομένα από άλλα κέντρα με διαφορετικό πληθυσμό ασθενών.

Tables

Table 5.

Κατηγορίες βλαβών που συμπεριλήφθηκαν στη μελέτη μας και υπότυποι BCC.

Characteristic	No. (%) of total cases (N = 656)
Diagnosis	
BCC	291 (44.4)
SCC	114 (17.4)
Actinic keratosis	42 (6.4)
LPLK	37 (5.6)
Amelanotic or hypomelanotic melanoma	31 (4.7)
Seborrheic keratosis	29 (4.4)
Bowen disease	13 (2.0)
Keratoacanthoma	12 (1.8)
Intradermal nevus	12 (1.8)
Dermatitis	11 (1.7)
Sebaceous hyperplasia	6 (0.9)
Dermatofibroma	5 (0.8)
Desmoplastic trichoepithelioma	4 (0.7)
Psoriasis	4 (0.7)
Molluscum contagiosum	2 (0.3)
Other[a]	43 (6.6)
BCC subtype[b]	
Nodular	224 (76.9)
Superficial	27 (9.3)
Infiltrative	24 (8.2)
Morpheaform or sclerosing	10 (3.4)
Keratotic	2 (0.7)
Basosquamous	1 (0.3)
Pinkus	1 (0.3)

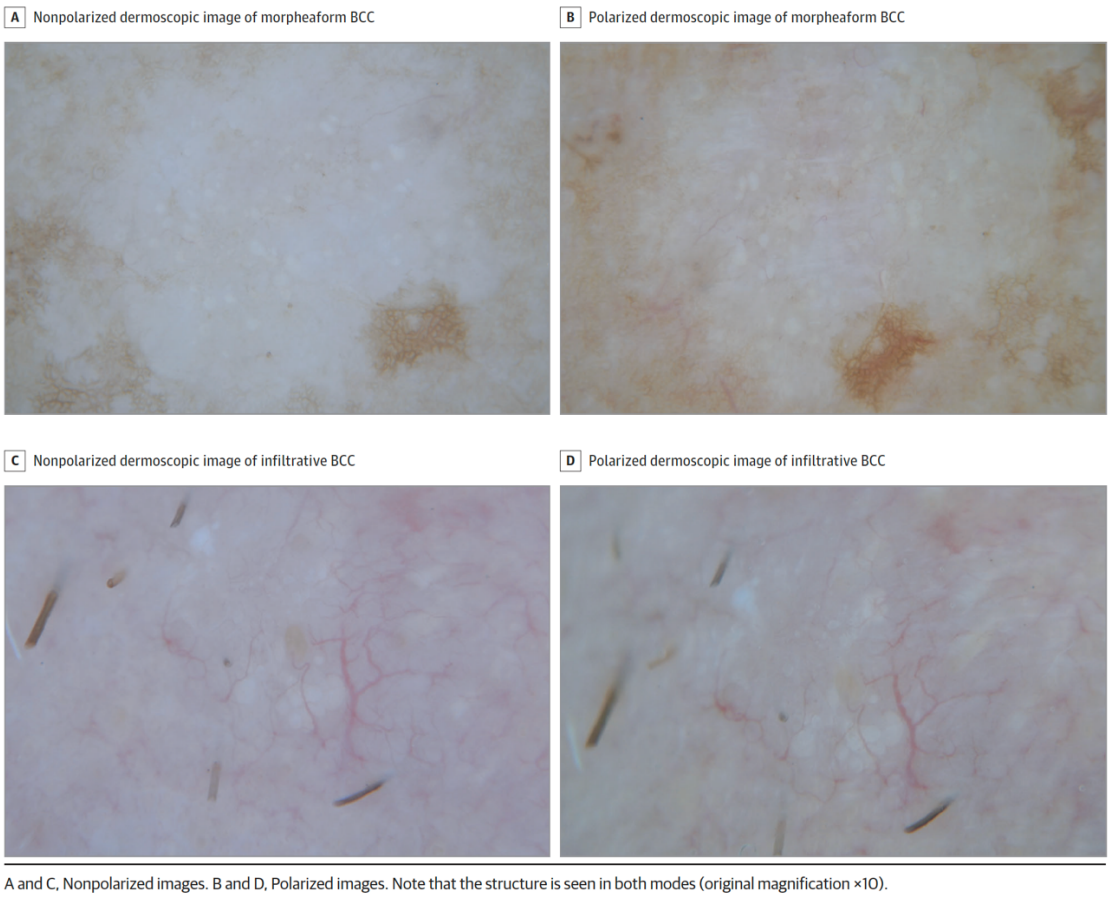
Table 6.

Ανάλυση των δερματοσκοπικών χαρακτηριστικών των βλαβών που συμπεριλήφθηκαν.

Characteristic	Cases, No. (%)			OR (95% CI)	κ (95% CI)
	BCC (n = 291)	Other diagnoses (n = 365)	Total (N = 656)		
Multiple aggregated yellow-white globules	61 (21.0)	3 (0.8)	64 (9.8)	32.0 (9.9 to 103.2)	0.895 (0.753 to 0.937)
Ulceration[a]	50 (17.2)	60 (16.4)	110 (16.8)	1.1 (0.7 to 1.6)	0.7261 (0.664 to 0.768)
Arborizing telangiectasia[a]	139 (47.8)	22 (6.0)	161 (24.5)	14.3 (8.7 to 23.2)	0.935 (0.935 to 0.966)
Ovoid nest[a]	16 (5.5)	0 (0)	16 (2.4)	NR	−0.008 (0.017 to 0.003)
Blue-gray globules[a]	24 (8.3)	0 (0)	24 (3.7)	NR	0.189 (0.126 to 0.494)
Blotches and strands[a]	214 (73.5)	70 (19.2)	284 (43.3)	11.7 (8.1 to 16.9)	0.827 (0.806 to 0.904)
Spoke-wheel structures[a]	4 (1.4)	0 (0)	4 (0.6)	NR	0.328 (0.189 to 0.497)
Leaflike areas[a]	36 (12.4)	1 (0.3)	37 (5.6)	51.4 (7.0 to 377.2)	0.747 (0.747 to 1.0)
Concentric structures[a]	15 (5.2)	0 (0)	15 (2.3)	NR	0.272 (−0.008 to 0.392)
Short-fine telangiectasia[a]	115 (39.5)	25 (6.9)	140 (21.3)	8.9 (5.6 to 14.2)	0.484 (0.398 to 0.592)
In-focus dots[a]	75 (25.8)	6 (1.6)	81 (12.4)	20.8 (8.9 to 48.5)	0.782 (0.629 to 0.835)
Multiple small erosions[a]	23 (7.9)	3 (0.8)	26 (4.0)	10.4 (3.1 to 34.8)	-
Serpentine vessels	1 (0.3)	16 (4.4)	17 (2.6)	0.1 (0 to 0.6)	−0.014 (−0.017 to −0.008)
Milialike cysts	15 (5.2)	16 (4.4)	31 (4.7)	1.2 (0.6 to 2.4)	0.601 (0.477 to 0.791)
Polymorphous vessels	11 (3.8)	46 (12.6)	57 (8.7)	0.3 (0.1 to 0.5)	0.669 (0.507 to 0.742)
Shiny white lines	3 (1.0)	28 (7.7)	31 (4.7)	0.1 (0 to 0.4)	0.851 (0.658 to 1.0)
Rosettes	17 (5.8)	51 (14.0)	68 (10.4)	0.4 (0.2 to 0.7)	0.741 (0.676 to 0.747)
Peppering	2 (0.7)	7 (1.9)	9 (1.4)	0.4 (0.1 to 1.7)	−0.005 (−0.005 to −0.003)
White circles	3 (1.0)	35 (9.6)	38 (5.8)	0.1 (0 to 0.3)	0.767 (0.719 to 0.82)
Scale	15 (5.2)	179 (49.0)	194 (29.6)	0.1 (0 to 0.1)	0.807 (0.782 to 0.824)
Glomerular vessels	11 (3.8)	115 (31.5)	126 (19.2)	0.1 (0 to 0.2)	0.650 (0.501 to 0.697)
Hairpin vessels	10 (3.4)	41 (11.2)	51 (7.8)	0.3 (0.1 to 0.6)	0.556 (0.332 to 0.63)
Orange color	5 (1.7)	50 (13.7)	55 (8.4)	0.1 (0 to 0.3)	0.647 (0.538 to 0.779)

Figures

Figure 12: Morpheaform και Infiltrative BCC τα οποία επιδεικνύουν τα MAY globules.

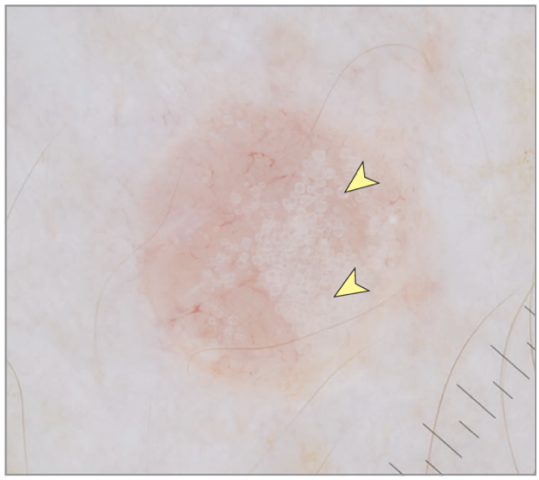


A and C, Nonpolarized images. B and D, Polarized images. Note that the structure is seen in both modes (original magnification ×10).

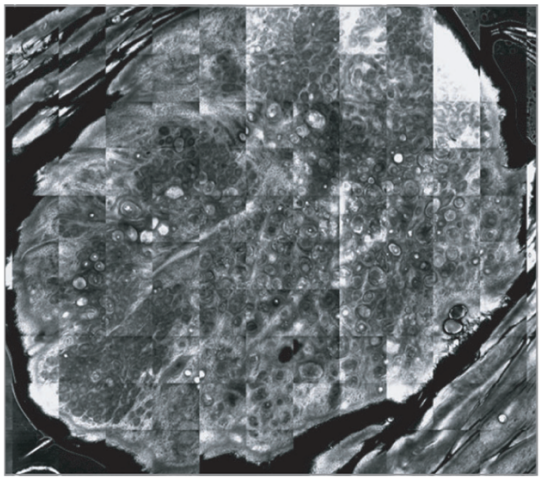
Figure 13.

A. Δερματοσκοπική εμφάνιση των MAY globules (κίτρινα βέλη). **B.** Πανοραμική όψη RCM δείχνει έναν καλά καθορισμένο όγκο με hyperreflective, amorphous areas. **C.** Το RCM δείχνει tumor nodules (μπλε βέλη) και hyperreflective, amorphous areas (κίτρινα βέλη). **D.** Η OCT δείχνει hyperreflective δομές με ακουστική σκιά (κίτρινα βέλη) και hyporeflective οζίδια (μπλε βέλη). **E.** Ιστοπαθολογική ανάλυση (en face) η οποία δείχνει tumor islands με clefting και palisading (μπλε βέλη) και εναποθέσεις ασβεστίου (κίτρινα βέλη). **F.** Ιστοπαθολογική ανάλυση (vertical) η οποία δείχνει tumor islands με clefting και palisading (μπλε βέλη) και εναποθέσεις ασβεστίου (κίτρινα βέλη) υποεπιδερμιδικά.

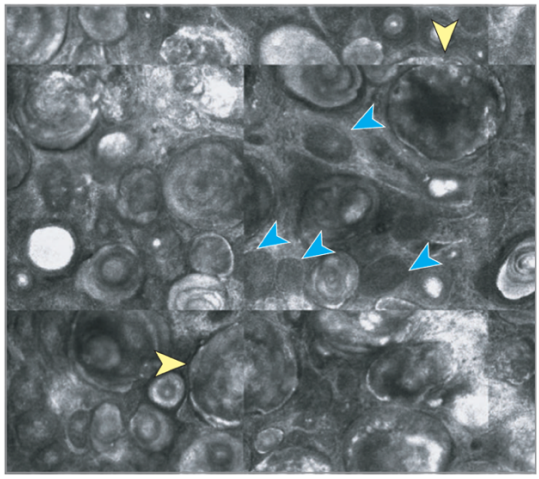
A Dermoscopic image



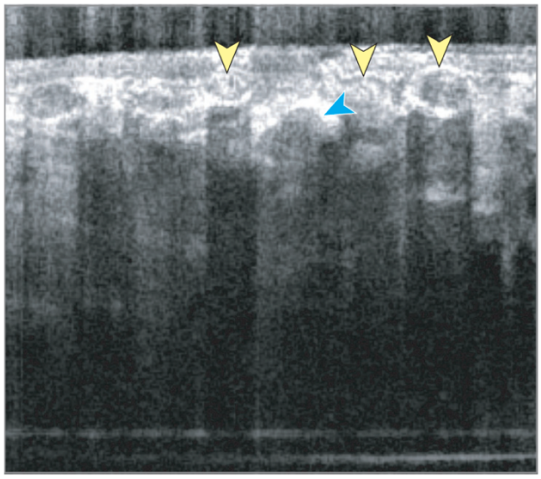
B Reflectance confocal microscopy image, panoramic view



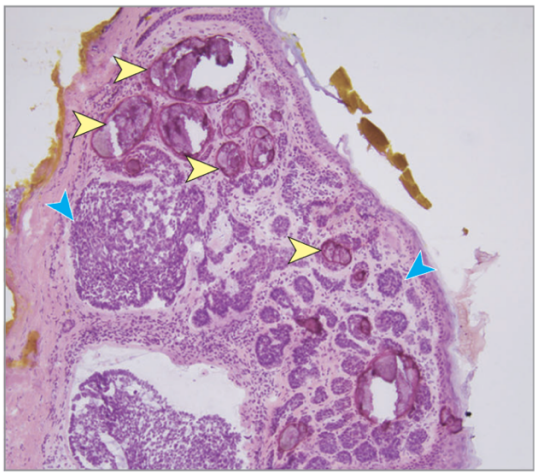
C Reflectance confocal microscopy image



D Optical coherence tomography



E Histopathologic analysis, en face view



F Histopathologic analysis, vertical view

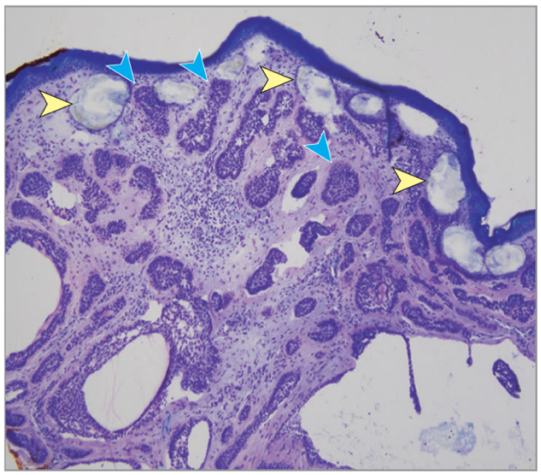
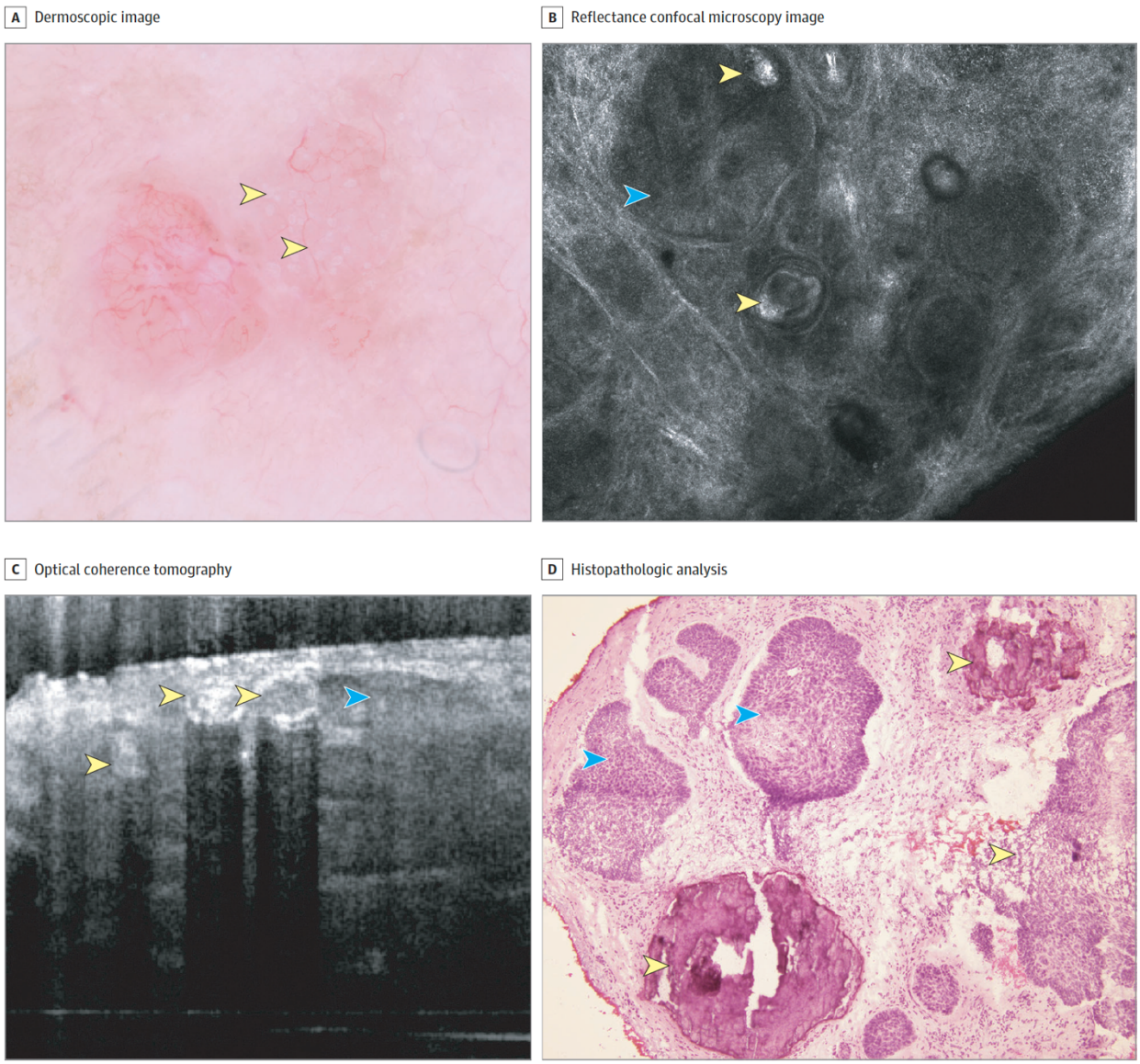


Figure 14.

A. Δερματοσκοπική εμφάνιση BCC η οποία απεικονίζει MAY globules (κίτρινα βέλη) **B.** RCM εικόνα η οποία απεικονίζει tumor nodules (μπλε βέλη) και ασβεστοποιήσεις (κίτρινα βέλη) **C.** OCT η οποία απεικονίζει hyperreflective δομές με ακουστική σκιά (κίτρινα βέλη) και hyporeflective οζίδια (μπλε βέλη). **D.** Ιστοπαθολογική ανάλυση (en face) η οποία δείχνει tumor islands με clefting και palisading (μπλε βέλη) και εναποθέσεις ασβεστίου (κίτρινα βέλη)



Expert Agreement Study on Dermoscopy Μελανοκυτταρικών βλαβών (EASY study)

Εισαγωγή

Όπως εκθέσαμε και στην εισαγωγή της διατριβής, η Δερματοσκόπηση είναι μια μη επεμβατική διαγνωστική τεχνική η οποία έχει αποδειχθεί ότι βοηθά στην έγκαιρη διάγνωση του καρκίνου του δέρματος και ειδικότερα του μελανώματος.[22,71,107] Η Δερματοσκόπηση χρησιμοποιεί πολωμένο ή μη πολωμένο φως και σταθερή μεγέθυνση 10x προκειμένου να απεικονίσει στοιχεία δερματολογικών βλαβών που βρίσκονται κάτω από την κεράτινη στιβάδα της επιδερμίδας.[40,41] Υπάρχει ένα μεγάλο εύρος ερευνών οι οποίες υπογραμμίζουν τα πλεονεκτήματα της δερματοσκόπησης, ωστόσο, ενώ η δερματοσκόπηση αδιαμφισβήτητα βελτιώνει τη διαγνωστική ακρίβεια των δερματολόγων για τη διάγνωση του καρκίνου δέρματος, η συμφωνία μεταξύ των παρατηρητών (interobserver agreement) για την παρουσία συγκεκριμένων δερματοσκοπικών κριτηρίων παραμένει κακή έως μέτρια, ακόμα και ανάμεσα σε experts της δερματοσκόπησης.[75,109]

Οι λόγοι για αυτή την κακή συμφωνία δεν έχουν διευκρινιστεί πλήρως αλλά θα μπορούσαν να περιλαμβάνουν: (1) διαφορές στην αντίληψη των διαφόρων κριτηρίων ανάμεσα στους παρατηρητές, (2) πιθανή αλληλοεπικάλυψη διαφόρων κριτηρίων, (3) έλλειψη προτυποποίησης της δερματοσκοπικής ορολογίας, (4) δερματοσκοπικά κριτήρια τα οποία είναι μη αναπαράξιμα, (5) χρήση ακατάλληλων στατιστικών μεθόδων για την μέτρηση της συμφωνίας μεταξύ των παρατηρητών. Πρότερες προσπάθειες στον τομέα έχουν επιχειρήσει την προτυποποίηση της δερματοσκοπικής ορολογίας, ωστόσο, αυτά τα κριτήρια δεν έχουν ελεγχθεί για την αναπαραγωγιμότητα τους, καθώς και την αξιοπιστία τους.[39]

Προκειμένου να επιτευχθεί αυξημένη χρήση της δερματοσκόπησης από τους κλινικούς ιατρούς καθώς και για να βελτιωθεί η διαγνωστική ακρίβεια των ιατρών η βελτίωση του interobserver agreement, όπως και η προτυποποίηση της δερματοσκοπικής ορολογίας είναι προαπαιτούμενα. Για την επίτευξη αυτού του στόχου σχεδιάσαμε την Expert Agreement Study on DermoscopY (EASY Study) για μελανοκυτταρικές βλάβες. Σε αυτή τη μελέτη επιχειρήσαμε να υπολογίσουμε την interobserver agreement των ειδικών στη δερματοσκόπηση σε 31 (τριανταένα) δερματοσκοπικά κριτήρια τα οποία είναι ειδικά για τη διάγνωση μελανοκυτταρικών βλαβών, όπως προέκυψαν από το τελευταίο consensus.[39]

Προς αυτό το σκοπό, ζητήσαμε από ειδικούς στη δερματοσκόπηση να υποβάλλουν 'αρχετυπικές' βλάβες, οι οποίες εμπεριείχαν τα 31 δερματοσκοπικά κριτήρια με παραδειγματικό τρόπο. Αυτές οι βλάβες χρησιμοποιήθηκαν προκειμένου να διερευνήσουμε τη συμφωνία των ειδικών ως προς την παρουσία, ή την απουσία των συγκεκριμένων κριτηρίων. Επιπροσθέτως, μέσω του www.isic-archive.com αναπτύξαμε μια πλατφόρμα για διεξοδική περιγραφή (annotation) των κριτηρίων αυτών από τους ειδικούς της δερματοσκόπησης. Μέσω αυτής της πλατφόρμας οι ειδικοί μπορούσαν να προσδιορίσουν που ακριβώς στη βλάβη βρίσκονται τα 31 κριτήρια, κάνοντας spatial annotation των superpixel (ομάδες pixel τα οποία ομοιάζουν μεταξύ τους) που τα περιείχαν. Η περιγραφή των κριτηρίων από τους ειδικούς στο επίπεδο της βλάβης επιτρέπει την ανάλυση του interobserver agreement για την παρουσία ή όχι των κριτηρίων, ενώ η ανάλυση στο επίπεδο των superpixels επιτρέπει την ανάλυση της αλληλοεπικάλυψης των εν λόγω κριτηρίων.

Υλικά και Μέθοδοι

Αυτή ήταν μια cross-sectional, observational μελέτη η οποία πραγματοποιήθηκε μεταξύ της 1ης Σεπτεμβρίου, 2017 και της 31ης Ιανουαρίου, 2020. Προσκλήσεις στάλθηκαν μέσω email σε 32 ειδικούς στη δερματοσκόπηση. Ως ειδικούς προσδιορίσαμε δερματολόγους με περισσότερα από 10 χρόνια εμπειρίας στη δερματοσκόπηση και οι οποίοι έχουν σημαντική συνεισφορά στην εξέλιξη του πεδίου τόσο μέσω έρευνας όσο και διδασκαλίας. Ζητήσαμε από τους ειδικούς να συνεισφέρουν 1 έως 3 εικόνες βλαβών για κάθε ένα από τα 31 μελανοκυτταρικά δερματοσκοπικά κριτήρια, οι οποίες να απεικονίζουν πρότυπα παραδείγματα των κριτηρίων αυτών.[39] **(Supplementary Table S1**: Ορισμοί κριτηρίων και οι συντμήσεις τους) Ως πρότυπη εικόνα ορίσαμε τις εικόνες οι οποίες είχαν εξαιρετική ποιότητα και απεικόνιζαν το κριτήριο για το οποίο κατατέθηκαν in-focus και με παραδειγματικό τρόπο. Συνολικά, 25 ειδικοί (81%) συνεισέφεραν 964 εικόνες μελανοκυτταρικών βλαβών. Εκ των 32 ειδικών οι οποίοι προσεκλήθησαν αρχικά να συμμετάσχουν στη μελέτη, 21 (66%) ολοκλήρωσαν το annotation όλων των βλαβών και πέντε επιπλέον ειδικοί προσεκλήθησαν να συμμετάσχουν. Όλες οι εικόνες είναι διαθέσιμες στο International Skin Imaging Collaboration Archive (ISIC Archive).[10]

Προκειμένου να καθορίσουμε τον απαραίτητο αριθμό readers και των αξιολογήσεων που έκαστος έπρεπε να πραγματοποιήσει χρησιμοποιήσαμε προσομοιώσεις Monte-Carlo της intraclass correlation coefficient. Υπολογίσαμε ότι με ένα confidence interval της τάξεως του 95% και interclass correlation coefficient της τάξεως του 0.5 ως το μέτρο της συμφωνίας, 5 readers ανά dataset θα αρκούσε προκειμένου να αξιολογήσουμε τη συμφωνία των readers για τα 31 δερματοσκοπικά κριτήρια.

Επιλογή Δερματοσκοπικών Κριτηρίων

Συμπεριλάβαμε και τα 31 μελανοκυτταρικά δερματοσκοπικά κριτήρια από το consensus του 2016. [39] Αυτή η λίστα συμπεριελάμβανε 14 κριτήρια τα οποία έχουν υψηλή ειδικότητα για τη διάγνωση μελανώματος.[41]

Δημιουργία των επιμέρους Dataset

Τρεις ειδικοί στη δερματοσκόπηση (K.L., C.N-D. and A.A.M.) επέλεξαν 310 από τις εικόνες που εστάλησαν από τους ειδικούς μέσω consensus. Επελέγησαν 10 εικόνες ανά δερματοσκοπικό κριτήριο με βάση την ποιότητα της εικόνας, την παραδειγματική παρουσίαση του κάθε κριτηρίου και την παρουσία όλων των metadata. Αυτές οι εικόνες χρησιμοποιήθηκαν προκειμένου να δημιουργηθούν 5 datasets, το καθένα αποτελούμενο από 62 δερματοσκοπικές εικόνες. Κάθε ένα από τα 5 dataset συμπεριελάμβανε 2 'exemplar' δερματοσκοπικές εικόνες για καθένα από τα 31 δερματοσκοπικά κριτήρια. Οι 25 ειδικοί στη δερματοσκόπηση κατανεμήθηκαν τυχαία σε ένα από τα 5 datasets, προκειμένου να προβούν στο annotation των εικόνων. Τα annotations δεν ολοκληρώθηκαν για ένα από τα 5 datasets και αποκλείσθηκαν από περαιτέρω ανάλυση.

Δημιουργία των Superpixels

Ακολουθώντας το upload των δερματοσκοπικών εικόνων στο ISIC Archive (www.isic-archive.com), ένας 'χάρτης' από superpixel δημιουργείται αυτόματα. Το segmentation εικόνων σε superpixels είναι μια τεχνική για την ομαδοποίηση γειτονικών pixel με βάση την ομοιότητα τους και την εγγύτητά τους στο χώρο. Χρησιμοποιήσαμε τον αλγόριθμο SLIC (Simple Linear Iterative Clustering) βασιζόμενοι στο χρώμα των pixel για να πραγματοποιήσουμε την ομαδοποίηση βασιζόμενοι στις χρωματικές αντιθέσεις μεταξύ γειτονικών pixel.[43,110] **(Figure 15)**.

Annotation των δερματοσκοπικών εικόνων

Αναπτύξαμε ένα online εργαλείο μέσω του ISIC Archive (ISIC annotation tool) που επιτρέπει στους annotators να επιλέγουν δερματοσκοπικά κριτήρια τόσο στο επίπεδο της βλάβης (lesion-level annotations), όσο και στο επίπεδο των superpixels (spatial annotations).[10] Τα annotations στο επίπεδο της βλάβης επιτρέπει την επιλογή σε δυαδικό επίπεδο - παρουσία ή απουσία ενός δερματοσκοπικού κριτηρίου - χωρίς να προσδιορίζεται σε ποιο σημείο εντός της βλάβης παρατηρείται το εν λόγω κριτήριο. Αντίθετα, τα superpixel annotations επιτρέπουν να προσδιορίζεται χωρικά που εντοπίζεται κάθε κριτήριο. Το ISIC annotation tool μπορεί να χρησιμοποιηθεί για την ταυτοποίηση πολλαπλών δερματοσκοπικών κριτηρίων εντός μιας βλάβης, ακόμα και αν αλληλεπικαλύπτονται. Οι ειδικοί οι οποίοι πραγματοποίησαν τα annotations δεν γνώριζαν τη διάγνωση των βλαβών και ήταν ελεύθεροι να επιλέξουν όποια κριτήρια διέκριναν στις βλάβες σύμφωνα με την αντίληψη τους. (Επίδειξη του πως λειτουργεί η πλατφόρμα μπορεί να βρεθεί εδώ https://youtu.be/jgJdCD3k3Es) Παράδειγμα μιας εκ των δερματοσκοπικών εικόνων με το διαχωρισμό σε superpixels καθώς και τα annotations από τους 5 ειδικούς μπορεί να βρεθεί στην **Figure 15.**

Ανάλυση του Agreement

Πραγματοποιήσαμε τρία επίπεδα ανάλυσης: (1) Agreement για την παρουσία του κριτηρίου για το οποίο κατατέθηκε η εικόνα (exemplar feature) από τους ειδικούς (2) Agreement για τα exemplar features στο επίπεδο των superpixel & (3) Agreement για τα non-exemplar features στο επίπεδο των superpixel.

Agreement στο επίπεδο της βλάβης

Συμφωνία στην παρουσία του exemplar feature:

Παραδειγματικές εικόνες για καθένα από τα 31 κριτήρια διαμοιράστηκαν σε καθένα από τα 4 dataset τα οποία συμπεριλήφθηκαν στη μελέτη μας (2 exemplars για κάθε κριτήριο). Ακολούθως τα δεδομένα συμπτύχθηκαν σε ένα dataset προκειμένου να αναλυθούν. Υπολογίσαμε την ποσοστιαία συμφωνία, το Fleiss kappa και το Gwet's AC1 για καθένα από τα 31 δερματοσκοπικά κριτήρια. [111] Ο υπολογισμός αμφότερων των Fleiss kappa και Gwet's AC1 έγινε προκειμένου να συνυπολογίσουμε την παράδοξη επίδοση της Fleiss kappa στα όρια της κατανομής της ποσοστιαίας συμφωνίας. Προκειμένου να διερευνήσουμε περαιτέρω πιθανή αλληλοεπικάλυψη μεταξύ των δερματοσκοπικών κριτηρίων τα συνδυάσαμε σε 7 'super-categories' κριτηρίων (Dots, Globules/Clods, Lines, Network, Regression Structures, Structureless and Vessels), διατηρώντας ξεχωριστά τα δομικά διαφορετικά κριτήρια Shiny White Structures, Angulated Lines/Polygons, και Negative Network με βάση συναίνεση μεταξύ 4 εκ των ερευνητών (K.L., C.N-D., M.A.M, & A.A.M.). Η μέτρηση της συμφωνίας πραγματοποιήθηκε και για τις 'super-categories', ενώ η ερμηνεία των Kappa & Gwet's AC1 πραγματοποιήθηκε όπως υπογραμμίστηκε από τους Landis and Koch: 0<0.4 (poor agreement), 0.4<0.75 (fair to good), και 0.75-1.0 (excellent agreement).[112]

Συμφωνία στο επίπεδο των Superpixel:

Προκειμένου να εκτιμήσουμε τη συμφωνία μεταξύ των readers στο ίδιο superpixel, καθώς και την αλληλοεπικάλυψη, ή τη σύγχυση μεταξύ των διαφορετικών κριτηρίων, υπολογίσαμε το Dice coefficient ανάμεσα σε κάθε πιθανό ζεύγος κριτηρίων για κάθε ένα από τα superpixels τα οποία επέλεξαν οι

readers.[113,114] Αυτός ο υπολογισμός γεννά έναν αριθμό μεταξύ μηδενικής συμφωνίας (0.0, 0%) σε περιπτώσεις πλήρους διαφωνίας και μέγιστης συμφωνίας (1.0, 100%) (δηλαδή επιλογή ακριβώς των ίδιων superpixel). Προκειμένου να απεικονίσουμε την ανάλυση αυτή μεταξύ των ζευγών κριτηρίων δημιουργήσαμε ένα confusion matrix. Ο κώδικας για αυτές τις αναλύσεις είναι διαθέσιμος online στο Github repository.[115]

Αποτελέσματα

Annotations στο επίπεδο της βλάβης:

Είκοσι ειδικοί προχώρησαν σε annotation 248 δερματοσκοπικών εικόνων (8 exemplar εικόνες ανά δερματοσκοπικό κριτήριο) σε ομάδες των 5 ειδικών, για ένα σύνολο 4507 παρατηρήσεων δερματοσκοπικών κριτηρίων. Μονήρεις παρατηρήσεις ενός κριτηρίου αποτέλεσε το 22.4% όλων των παρατηρήσεων, ενώ απόλυτη συμφωνία (μεταξύ όλων των readers) προέκυψε σε 65 εικόνες (26.2%) **(Supplementary Table S2)**

Agreement για το exemplar feature στο επίπεδο της βλάβης:

Τα αποτελέσματα του agreement παρουσιάζονται στο **Table 7**. Η ποσοστιαία συμφωνία για τα διάφορα κριτήρια ποίκιλλε. Τα υψηλότερα ποσοστά agreement παρατηρήθηκαν για τα κριτήρια: 'Peppering / Granularity' (92%), 'Shiny White Streaks' (90%), 'Typical Network' και 'Irregular Blotch' (86%), 'Negative Network' (84%), 'Irregular Globules' και 'Dotted Vessels' (82%) και 'Scar-like Depigmentation' και 'Blue-Whitish Veil' (80%). Τα υπόλοιπα 22 κριτήρια παρήγαγαν μικρότερα ποσοστά συμφωνίας **(Table 7).** Συνολικά το Fleiss kappa έδειξε χαμηλό agreement (<0.4) για όλα τα μελανοκυτταρικά δερματοσκοπικά κριτήρια, με την εξαίρεση των: 'Rim of brown globules' και 'Irregular

Blotch'; 0.44 & 0.42, αντίστοιχα. Αυτά τα αποτελέσματα οφείλονται στην παραδοξότητα της επίδοσης του Fleiss kappa σε εξαιρετικά υψηλές, ή χαμηλές τιμές του επιπολασμού ενός χαρακτηριστικού. Χρησιμοποιώντας το Gwet's AC, βρήκαμε πολύ υψηλή συμφωνία για τα κριτήρια 'Irregular Globules' (0.78), 'Typical Network' (0.83), 'Peppering / Granularity' (0.91), 'Shiny White Streaks' (0.89), 'Negative Network' (0.81), 'Irregular Blotch' (0.82), 'Blue-Whitish Veil' (0.76) και 'Dotted Vessels' (0.77). Τα υπόλοιπα κριτήρια επέδειξαν χαμηλή έως μέτρια συμφωνία.

Μετά την σύμπτυξη των επιμέρους κριτηρίων σε 'super-categories' βασιζόμενοι σε δομικές ομοιότητες, παρατηρήσαμε υψηλότερα επίπεδα συνολικού ποσοστιαίου agreement, όπως και τιμών Gwet's AC. Τα κριτήρια 'Globules/clods', 'Network', 'Regression Structures', 'Shiny White Structures', 'Negative Network' και 'Vessels' επέδειξαν εξαιρετικό agreement με τιμές Gwet's AC >0.81. Μέτρια συμφωνία εμφάνισαν τα κριτήρια 'Lines' και 'Structureless', ενώ τα κριτήρια 'Dots' και 'Angulated lines' εμφάνισαν χαμηλό agreement.

Superpixel level annotations:

Καθεμία από τις 248 δερματοσκοπικές εικόνες που αναλύθηκαν στη μελέτη μας χωρίστηκε σε περίπου 1,000 superpixel (mean=1001.4, SD=18.1). Ένα σύνολο 47,524 Superpixel έγινε annotate από τους ειδικούς σε αυτές τις εικόνες. Διαφωνία αναφορικά με τα δερματοσκοπικά κριτήρια προέκυψε στο 81.5% των superpixels που έγιναν annotate (N=38,732 superpixels).

Percent Superpixel Agreement on the exemplar feature:

Υπήρξαν 9 κριτήρια τα οποία παρουσιαζόμενα ως exemplar feature εμφάνισαν χωρική (spatial) agreement >10% ανάμεσα στους 5 ειδικούς οι οποίοι ήταν readers. Αυτά τα κριτήρια ήταν τα *'Typical Network'* με 36.2% απόλυτο (100%) agreement ανάμεσα στους readers, *'Cobblestone Pattern'* με

24.7%,*'Rim of Brown Globules'* με 19.6%, '*Blotch Regular'* με 17.6%, '*Blotch Irregular'* με 17.2%, *'Negative Network'* με 15.3%, '*Shiny White Streaks'* με 15.2%, *'Peppering / Granularity'* με 11.3% και '*Blue-Whitish Veil'* με 11.3%. **(Table 8).**

Δέκα εκ των 31 κριτηρίων δεν εμφάνισαν καθόλου spatial agreement ανάμεσα στους expert readers (0%), αυτά ήταν: 'Regular Dots', 'Milky Red Globules', 'Regular Globules', 'Angulated lines', 'Branched Streaks', 'Broadened Network', 'Delicate Network', 'Homogeneous Pattern', 'Milky Red Areas' και 'Corkscrew Vessels'.

Percent Superpixel Agreement on all features (exemplar and non-exemplar):

Τα κριτήρια με το υψηλότερο ποσοστό απόλυτου agreement (100%) ανάμεσα στους readers ήταν: το *'Cobblestone Pattern'* με 14.63% απόλυτου (100%) agreement, *'Typical Network'* με 11.88%, *'Rim of Brown Globules'* με 10.86%, *'Dotted Vessels'* με 5.83%, *'Negative Network'* με 5.27%, '*Shiny White Streaks'* με 4.99%, *'Peppering / Granularity'* με 3.82%, *'Starburst Pattern'* με 3.36%, '*Blotch Regular'* με 2.86%, '*Blotch Irregular'* με 2.37% και '*Polymorphous Vessels'* με 2.09% απόλυτου (100%) spatial agreement ανάμεσα στους readers **(Supplementary Table S2).** Τα ίδια δέκα κριτήρια που επέδειξαν μηδενική συμφωνία όταν παρουσιάστηκαν ως exemplar features είχαν επίσης μηδενικό agreement εξετάζοντας τα κριτήρια εν τω συνόλω τους.

Confusion matrix και DICE coefficient (exemplar and non-exemplar features):

Το ποσοστιαίο agreement ανάμεσα στους experts στο επίπεδο των superpixel ήταν συγκριτικά χαμηλό, ωστόσο μέσω της ανάλυσης του DICE coefficient εντοπίσαμε 50 ζεύγη κριτηρίων τα οποία επέδειξαν σταθερά υψηλή αλληλοεπικάλυψη με μέσο DICE coefficient ≥0.5 (**Table 9)**. Επί παραδείγματι, *'Atypical Network'* και *'Broadened Network'* έγιναν annotate 85 φορές από διαφορετικούς readers, σε

παρόμοιες περιοχές superpixel με DICE coefficient της τάξεως του 0.584· αντίστοιχα τα *'Delicate Network'* και *'Typical network'* έγιναν annotate 67 φορές από διαφορετικούς readers σε παρόμοιες περιοχές superpixel με DICE coefficient της τάξεως του 0.637, ενώ τα κριτήρια *'Broadened Network'* & *'Typical Network'* έγιναν annotate 33 φορές από διαφορετικούς readers σε παρόμοιες περιοχές superpixel με DICE coefficient της τάξεως του 0.658, δείχνοντας έτσι την πιθανή αλληλοεπικάλυψη των ορισμών ανάμεσα στα 4 αυτά μελανοκυτταρικά δερματοσκοπικά κριτήρια. Επιπροσθέτως, το κριτήριο '*Homogeneous Pattern'* – ένα γενικό pattern μοιάζει να είναι μη ειδικό και εμφανίστηκε να αλληλεπικαλύπτεται με σχεδόν όλα τα δερματοσκοπικά κριτήρια τα οποία αναλύθηκαν στη μελέτη μας (**Figure 16).** Παραδείγματα αμφότερων των αποτελεσμάτων (υψηλό overlap/agreement σε ένα κριτήριο και εξίσου υψηλή διαφωνία σε άλλα κριτήρια) μπορούν να ανευρεθούν στο **Supplementary Figure 1.**

Συζήτηση:

Σε αυτή τη μελέτη η οποία συμπεριέλαβε 20 διεθνείς experts στη δερματοσκόπηση αξιολογήσαμε το agreement για τα 31 αναγνωρισμένα, μελανοκυτταρικά, δερματοσκοπικά κριτήρια σε 248 εικόνες μελανοκυτταρικών βλαβών.[39] Το agreement στο επίπεδο της βλάβης ήταν μέτριο για πολλά από τα προτεινόμενα κριτήρια, με μονήρεις παρατηρήσεις να απαρτίζουν το 22.4% όλων των annotations από τους expert readers. Βρήκαμε σημαντική συμφωνία σε 14 από τα 31 κριτήρια (45.2%) ενώ εξαιρετικό agreement προέκυψε μόλις σε 8 από τα 31 κριτήρια (25.8%), ακόμα και όταν παρουσιάστηκαν ως exemplar features.[111,112] Ενδιαφέρουσα ήταν η παρατήρηση πως από τα 14 κριτήρια με υψηλό agreement τα 7 ήταν κριτήρια με υψηλή ειδικότητα για τη διάγνωση μελανώματος, καταδεικνύοντας πως αυτά τα κριτήρια είναι απαραίτητα για την κλινική πράξη. Επιπλέον, η σύμπτυξη των 31 κριτηρίων σε επιμέρους 'super-categories' βασιζόμενη σε δομικές ομοιότητες ανέδειξε καλύτερο agreement, με

την εξαίρεση των ‘Dots’. Αυτά τα ευρήματα υπονοούν πως ενδεχομένως να υπάρχει αλληλοεπικάλυψη τόσο στους ορισμούς, όσο και στην αντίληψη των διαφόρων κριτηρίων που εντάσσονται σε αυτές τις ‘super-categories’, τα οποία θα μπορούσαν να αναλυθούν περαιτέρω ή να επαναπροσδιορισθούν προκειμένου να δημιουργηθούν πιο αξιόπιστα και αναπαράξιμα διαγνωστικά κριτήρια.

Προηγηθείσες έρευνες στο πεδίο είχαν δείξει χαμηλή έως μέτρια συμφωνία για την πλειονότητα των δερματοσκοπικών κριτηρίων.[75,109] Ωστόσο, αυτό που διαχωρίζει την έρευνα μας είναι εν πρώτοις η χρήση των exemplar δερματοσκοπικών εικόνων, οι οποίες υποβλήθηκαν από τους ειδικούς οι οποίοι έχουν περιγράψει την πλειονότητα των υπό εξέταση κριτηρίων. Η χρήση αυτών των εικόνων επέφερε υψηλότερο agreement συγκρινόμενη με πρότερες έρευνες ενώ δημιούργησε ένα δημόσια διαθέσιμο αρχείο ‘exemplar’ εικόνων το οποίο μπορεί να χρησιμοποιηθεί από όλους για την πρόοδο της έρευνας του πεδίου.[10] Επιπροσθέτως, η μελέτη μας είναι η πρώτη που αξιοποίησε το spatial localization χρησιμοποιώντας τα superpixel ώστε να περιγράψει το agreement για τα δερματοσκοπικά κριτήρια. Αυτή η καινοτόμος προσέγγιση μας επέτρεψε να προσδιορίσουμε τα κριτήρια τα οποία αλληλεπικαλύπτονται, δίνοντας μας έτσι τη δυνατότητα να αναλύσουμε έτι περαιτέρω την οπτική αντίληψη των ειδικών στη δερματοσκόπηση. Τέλος, ο επιπρόσθετος υπολογισμός του Gwet’s AC μας προσέφερε τη δυνατότητα να υπερβούμε τους περιορισμούς της Fleiss kappa στις ακραίες τιμές της κατανομής των κριτηρίων.

Το συνολικό agreement, ποικίλλε για τα διάφορα δερματοσκοπικά κριτήρια. Ωστόσο, ένα σημαντικό τμήμα της μελέτης μας αφορούσε στη χρήση των superpixel για τον χωρικό προσδιορισμό εκάστου κριτηρίου στις δερματοσκοπικές εικόνες που συμπεριλήφθηκαν στη μελέτη μας. Τα superpixel annotations μας επιτρέπουν την επιπρόσθετη κατανόηση των αλληλοεπικαλύψεων μεταξύ διαφορετικών δερματοσκοπικών κριτηρίων, καθώς και νέες προσεγγίσεις για την ανάλυση του

agreement. Η συμφωνία στο επίπεδο των superpixel ήταν χαμηλή· μόλις 19.6% του συνόλου των superpixel τα οποία έγιναν annotate από τους ειδικούς επέδειξαν συμφωνία μεταξύ 2 ή περισσοτέρων readers, ενώ μόλις 11 κριτήρια έδειξαν spatial agreement >2% ανάμεσα στο σύνολο των αξιολογητών. Ωστόσο, στη μελέτη μας προσδιορίσαμε 50 ζεύγη κριτηρίων τα οποία επέδειξαν σταθερά υψηλό DICE overlap (>0.5). Κατ' αυτό τον τρόπο μπορέσαμε να εντοπίσουμε τα κριτήρια τα οποία μπορεί να προκαλέσουν σύγχυση, ή τα οποία έχουν αλληλεπικαλυπτόμενους ορισμούς, ακόμα και για τους ειδικούς στη δερματοσκόπηση (**Table 9, Figure 16).** Στη μελέτη μας συμπεριλάβαμε 4 διαφορετικά κριτήρια υπό την 'super-category' *'Network / Reticulation',* τα οποία παρουσιάζουν δομικές ομοιότητες, βασιζόμενοι στο τελευταίο consensus της δερματοσκοπικής ορολογίας.[39] Δύο εξ αυτών είναι ειδικά για τη διάγνωση μελανώματος (*Atypical Network και Broadened Network*), ένα συσχετίζεται με τη διάγνωση σπίλων (Typical Network), ενώ ένα είναι μη ειδικό (*Delicate Network*).[39,41] Ωστόσο, παρά τους διαφορετικούς ορισμούς τους και τη διαφορετική τους σημασία στην κλινική πράξη, τα κριτήρια αυτά επέδειξαν ένα σταθερά υψηλό DICE coefficient στη μελέτη μας. Επιπροσθέτως, το ίδιο ίσχυσε και για τα 'Irregular Globules' (OR για διάγνωση μελανώματος 1.7-4.8) και τα 'Regular Globules', ένα κριτήριο ειδικό για τη διάγνωση καλοηθών μελοκυτταρικών βλαβών. (**Table 9, Figure 16**).[39,41] Μελλοντικές μελέτες χρειάζονται προκειμένου να διευκρινιστεί αν αυτά τα ευρήματα οφείλονται σε διαφορετική αντίληψη των αλληλεπικαλυπτόμενων κριτηρίων από τους ειδικούς, λόγω έλλειψης επαρκούς κατανόησης των ορισμών, ή λόγω σύγχρονης παρουσίας των εν λόγω κριτηρίων.

Η προσέγγιση της ανάλυσης εικόνων μέσω superpixel είχε χρησιμοποιηθεί παλαιότερα προκειμένου να θέσει το ground truth για την εκμάθηση δερματοσκοπικών κριτηρίων από αλγόριθμους τεχνητής νοημοσύνης, ωστόσο αυτό το annotation είχε πραγματοποιηθεί από μόνο έναν expert (K.L) και η επίδοση των αλγορίθμων στην αναγνώριση των κριτηρίων ήταν χαμηλή.[28,29] Οι αλγόριθμοι τεχνητής νοημοσύνης έχουν επιδείξει υψηλές δυνατότητες για τη διάγνωση μελανώματος

και άλλων καρκίνων δέρματος και θα μπορούσαν να αποδειχθούν πολύτιμοι σύμμαχοι στην κλινική πράξη.[30,58,116] Στη μελέτη μας δείξαμε πως πολλαπλοί readers προσεκτικά επιλεγμένων δερματοσκοπικών εικόνων θα μπορούσαν να αποτελέσουν τη βάση για την καλύτερη κατανόηση των δερματοσκοπικών κριτηρίων και θα μπορούσαν να βελτιώσουν την επίδοση των αλγορίθμων τεχνητής νοημοσύνης μέσω επιβλεπόμενης εκμάθησης.

Ο πρωταρχικός στόχος της μελέτης μας αφορούσε στη διερεύνηση του agreement μεταξύ των ειδικών στη δερματοσκόπηση για τα 31 μελανοκυτταρικά δερματοσκοπικά κριτήρια. Τα αποτελέσματα μας έδειξαν τα κριτήρια τα οποία έχουν υψηλότερα ποσοστά συμφωνίας, καθώς και τα κριτήρια τα οποία αλληλεπικαλύπτονται με άλλα. Αυτού του τύπου οι μελέτες μπορούν ενδεχόμενα να αποτελέσουν τη βάση για βελτιωμένους διαγνωστικούς αλγόριθμους, όπως και για προτυποποιημένη ορολογία στη δερματοσκόπηση, οδηγώντας έτσι σε υψηλότερη διαγνωστική ακρίβεια, τόσο από τους κλινικούς δερματολόγους, όσο και από τους αλγορίθμους τεχνητής νοημοσύνης. Ένας σημαντικός στόχος της μελέτης αυτής αφορούσε στη δημιουργία ενός πρότυπου αρχείου εικόνων, οι οποίες θα επιδείκνυαν τα 31 μελανοκυτταρικά δερματοσκοπικά κριτήρια με παραδειγματικό τρόπο για αποτελεσματική επιστημονική επικοινωνία, για διδαχή και για ασκήσεις αλγορίθμων τεχνητής νοημοσύνης. Αυτές οι 248 εικόνες, με τα συνοδά τους annotations και τις περιοχές συμφωνίας ανάμεσα στους παγκόσμιους ειδικούς της δερματοσκόπησης βρίσκονται στο ISIC archive και είναι διαθέσιμες στο κοινό.[10]

Περιορισμοί:

Η χρήση των superpixels είναι μια καινοτόμος μέθοδος για την αξιολόγηση της συμφωνίας μεταξύ διαφορετικών readers. Ωστόσο, αυτές οι προκαθορισμένες περιοχές ενδέχεται να μην είναι ιδεατές για τον προσδιορισμό δερματοσκοπικών κριτηρίων που εκτείνονται πέρα από τα όρια αυτά· ειδικά για κριτήρια τα οποία εμπεριέχουν υψηλό επίπεδο αντίθεσης, όπως το Network. Ένας άλλος σημαντικός περιορισμός της μελέτης μας είναι η αξιοποίηση των expert readers· η γενίκευση των αποτελεσμάτων που παράχθηκαν από αυτούς τους εξειδικευμένους readers ενδεχομένως να είναι αδύνατη στο γενικότερο πληθυσμό των δερματολόγων. Επιπροσθέτως, παρά τις προσπάθειες μας να συμπεριλάβουμε ένα μεγάλο εύρος ειδικών από όλο τον κόσμο είχαμε μόλις 2 readers από τη Λατινική Αμερική και την Αυστραλία, και κανέναν από την Αφρική και την Ασία. Τέλος, αυτή η μελέτη δεν θα έπρεπε σε καμία περίπτωση να εκληφθεί ως επιδημιολογική μελέτη, μιας και εμπλουτίσαμε το dataset μας με πληθώρα μελανωμάτων προκειμένου να έχουμε επαρκή αντιπροσώπευση των κριτηρίων που είναι ειδικά για τη διάγνωση μελανώματος. Η επιδημιολογία των κριτηρίων αυτών έχει μελετηθεί επαρκώς σε προηγούμενες έρευνες και ο σκοπός της μελέτης μας ήταν να εξετάσουμε τη συμφωνία μεταξύ ειδικών στην παρουσία και την εντόπιση των μελανοκυτταρικών δερματοσκοπικών κριτηρίων.

Συμπέρασμα:

Η συμφωνία στα δερματοσκοπικά κριτήρια παραμένει ποικίλη ακόμα και όταν χρησιμοποιούμε εξαιρετικά επιλεγμένες δερματοσκοπικές εικόνες, με readers ειδικούς στη δερματοσκόπηση. Η χρήση των exemplar images διευκολύνει το agreement, ενώ η χρήση των superpixels παρέχει σημαντικές πληροφορίες για τα κριτήρια τα οποία αλληλεπικαλύπτονται, ή των οποίων οι ορισμοί συγχέονται. Στη μελέτη μας βρήκαμε ότι τα μισά από τα δερματοσκοπικά κριτήρια τα οποία είναι ειδικά για τη διάγνωση μελανώματος έχουν σημαντικό agreement ενώ άλλα όπως τα 'Broadened Network' ,'Delicate Network', 'Tan Peripheral Brown Areas' έχουν πολύ χαμηλό agreement, ακόμα και στο ιδεατό σενάριο των exemplar images και ενδεχομένως να χρειάζεται είτε να αποκλειστούν από τα κριτήρια διάγνωσης μελανώματος, είτε να συμπτυχθούν μεταξύ τους. Τέλος η ανάλυση της αλληλοεπικάλυψης των δερματοσκοπικών κριτηρίων στο επίπεδο των superpixel ανέδειξε σημαντικό overlap μεταξύ διαφόρων κριτηρίων, υπογραμμίζοντας έτσι την σύγχυση μεταξύ των ορισμών διαφόρων κριτηρίων. Αυτές οι παρατηρήσεις είναι απαιτούμενες προϋποθέσεις προκειμένου να τεθούν στέρεες βάσεις για την οικοδόμηση consensus, όπως και για τον βέλτιστο ορισμό των κριτηρίων, ώστε να είναι δυνατή η γενίκευση τους και η καθημέρα χρήση τους. Τα αποτελέσματα μας μπορούν να αποτελέσουν έναν οδικό χάρτη προκειμένου να επιτευχθεί καλύτερη προτυποποίηση των δερματοσκοπικών κριτηρίων και για την καθοδήγηση των δερματοσκοπικών διαγνωστικών αλγορίθμων προκειμένου να επιτευχθεί υψηλότερη χρήση και αποδοχή της δερματοσκόπησης ανάμεσα στους κλινικούς δερματολόγους.

Tables:

Table 7. Μέτρηση του agreement για μεμονωμένα δερματοσκοπικά κριτήρια, καθώς και για τις 'super-categories' στο επίπεδο της βλάβης.

Individual Dermoscopic Features					**Combined Dermoscopic Super-feature categories**			
Variable	**% agreement**	**Fleiss' kappa**	**Gwet's AC**		**Variable**	**% agreement**	**kappa**	**Gwet's AC**
Dots : Irregular	60.00%	0.0954	0.2829		Dots	61.00%	0.1371	0.2884*
Dots : Regular	60.00%	0.2000	0.2000					
Globules / Clods : Cobblestone pattern	60.00%	.0.0809	0.2918		Globules / Clods	91.00%	0.3706*	0.8950*
Globules / Clods : Irregular	82.00%	-0.0989	0.7847*					
Globules / Clods : Regular	46.00%	-0.1315	-0.0301					
Globules / Clods : Rim of brown globule	78.00%	0.4419	0.6368*					

Lines : Branched streaks	76.00%	0.0033	0.6839*		Lines	70.77%	0.3542*	0.4659*
Lines : Pseudopods	54.00%	-0.1170	0.2180					
Lines : Radial streaming	65.93%	0.0143	0.4792					
Lines : Starburst (pseudopods/radial)	54.00%	0.0417	0.1154					
Network : Atypical pigment network / Re	70.00%	-0.0631	0.5821*		Network	100%	1.0*	1.0*
Network : Broadened pigment network / R	38.00%	-0.2917	-0.1923					
Network : Delicate Pigment Network / Re	48.00%	-0.0552	-0.0252					
Network : Typical pigment network / Ret	86.00%	0.2222	0.8293*					

Regression structures : Peppering / Gra	92.00%	-0.0417	0.9133*		Regression structures	96.00%	-0.0204	0.9584*
Regression structures : Scarlike depigm	80.00%	0.3225	0.7162*					
Shiny white structures : Shiny white st	90.00%	-0.0526	0.8895*		Shiny white structures	90.00%	-0.0526	0.8895*
Independent: Angulated lines / Polygons / Zig-zag	60.00%	0.1883	0.2114		Angulated lines/ polygons	60.00%	0.1883	0.2114
Independent : Negative pigment network (independent)	84.00%	-0.0870	0.8124		Negative network	84.00%	-0.0870	0.8124
Structureless : Blotch irregular	86.00%	0.4186	0.8156*		Structureless	75.06	0.1699*	0.6434*
Structureless : Blotch regular	66.00%	0.2052	0.4058					

Structureless : Blue-whitish veil	80.00%	-0.1111	0.7561*					
Structureless : Milky red areas	58.00%	0.0349	0.2564					
Structureless : Structureless brown (ta	42.67%	-0.1485	-0.1448					
Structureless: Homogeneous : NOS	46.57%	-0.0876	-0.0876					
Vessels : Comma	76.00%	0.0588	0.6779*		Vessels	89.49%	0.3462*	0.8748*
Vessels : Corkscrew	58.00%	-0.1858	0.3496					
Vessels : Dotted	82.00%	0.1477	0.7718*					
Vessels : Linear irregular	68.89%	0.2742	0.4555*					
Vessels : Polymorphous	54.00%	-0.0403	0.1753					

Vessels: Milky red globules	67.00%	-0.0124	0.5104*				

Table 8.

Superpixel agreement on the exemplar feature.

Ποσοστιαία συμφωνία ανάμεσα στους ειδικούς στο spatial localization (superpixel agreement) του exemplar features. Ο Μέσος Αριθμός Readers (**Mean Number of Readers)** δείχνει το μέσο αριθμό των ειδικών που έκαναν annotate τις εικόνες για έκαστο κριτήριο. Το ποσοστό της συμφωνίας ανάμεσα στην πλειονότητα των readers (τουλάχιστον 3 από τους 5) ≥60% βρίσκεται στο **3RA %**, ενώ κάτω από το **Full - 5RA%** μπορούμε να δούμε το ποσοστό της απόλυτης συμφωνίας για κάθε κριτήριο (και οι 5 readers έκαναν annotate την ίδια περιοχή / superpixels για το εν λόγω κριτήριο).

Feature	**Mean Number of Readers**	**3RA - %**	**5RA- %**
Dots : Irregular	3.875	0.192648	0.003571
Dots : Regular	2.5	0.250801	0
Globules / Clods : Cobblestone pattern	3.625	0.629329	0.247115
Globules / Clods : Irregular	4.875	0.349432	0.043504

Globules / Clods : Milky red	1.25	0.016079	0
Globules / Clods : Regular	2.875	0.299281	0
Globules / Clods : Rim of brown globules	3.625	0.408223	0.195766
Lines : Angulated lines / Polygons / Zig-zag pattern	3.5	0.156645	0
Lines : Branched streaks	0.875	0	0
Lines : Pseudopods	3.5	0.300592	0.003906
Lines : Radial streaming	4	0.20898	0.035716
Network : Atypical pigment network / Reticulation	4.25	0.30728	0.031322
Network : Broadened pigment network / Reticulation	1.875	0.092118	0
Network : Delicate Pigment Network / Reticulation	2.125	0.01616	0
Network : Negative pigment network	4.5	0.644936	0.152862
Network : Typical pigment network / Reticulation	4.375	0.756599	0.362415
Pattern : Homogeneous : NOS	2.125	0.180564	0

Pattern : Starburst	3.125	0.330246	0.041268
Regression structures : Peppering / Granularity	4.625	0.504966	0.112971
Regression structures : Scarlike depigmentation	3.875	0.312567	0.050974
Shiny white structures : Shiny white streaks	5	0.532319	0.151495
Structureless : Blotch irregular	4.125	0.417691	0.172234
Structureless : Blotch regular	3.375	0.508637	0.175758
Structureless : Blue-whitish veil	4.375	0.4768	0.112656
Structureless : Milky red areas	3	0.184145	0
Structureless : Structureless brown (tan)	2.5	0.157379	0.033333
Vessels : Comma	4.375	0.176428	0.027243
Vessels : Corkscrew	1.125	0	0
Vessels : Dotted	4.5	0.541738	0.082328
Vessels : Linear irregular	3.875	0.247682	0.036885
Vessels : Polymorphous	3.25	0.315864	0.002083

Table 9.

Ζεύγη κριτηρίων με υψηλό (>0.5) DICE spatial overlap στα superpixel annotations. **N of pairs** δείχνει πόσες φορές προέκυψαν τα ζεύγη με το υψηλό overlap στη μελέτη μας, ενώ **% overlap** δείχνει το ποσοστό του spatial overlap.

N of pairs	% overlap	Feature 1	Feature 2
92	52.80%	Network : Atypical pigment network / Reticulation	Network : Typical pigment network / Reticulation
85	58.40%	Network : Atypical pigment network / Reticulation	etwork : Broadened pigment network / Reticulation
68	55.60%	Vessels : Linear irregular	Vessels : Polymorphous
67	63.70%	Network : Delicate Pigment Network / Reticulation	Network : Typical pigment network / Reticulation
64	67.00%	Lines : Radial streaming	Pattern : Starburst
58	86.50%	Globules / Clods : Cobblestone pattern	Globules / Clods : Regular
42	90.10%	Pattern : Homogeneous : NOS	Structureless : Blotch regular
33	65.80%	Network : Broadened pigment network / Reticulation	Network : Atypical pigment network / Reticulation
31	71.50%	Globules / Clods : Irregular	Globules / Clods : Regular

29	66.30%	Pattern : Homogeneous : NOS	ıctureless : Structureless brown (tan peripheral area)
29	50.70%	Network : Negative pigment network / Reticulation	Shiny white structures : Shiny white streaks
25	62.20%	Dots : Regular	Globules / Clods : Regular
24	67.90%	Globules / Clods : Regular	Globules / Clods : Rim of brown globules
24	56.20%	Pattern : Homogeneous : NOS	Structureless : Blotch irregular
22	50.30%	Dots : Irregular	Dots : Regular
18	70.10%	Lines : Pseudopods	Pattern : Starburst
16	54.90%	Pattern : Homogeneous : NOS	Structureless : Blue-whitish veil
14	72.30%	Structureless : Blotch regular	Structureless : Blue-whitish veil
14	67.00%	Globules / Clods : Rim of brown globules	Lines : Pseudopods
13	50.30%	Structureless : Blotch regular	Structureless : Structureless brown (tan
12	57.50%	Dots : Regular	Network : Atypical pigment network / Reticulation
9	56.80%	Globules / Clods : Milky red	Vessels : Polymorphous
9	52.40%	Globules / Clods : Regular	Network : Negative pigment network / Reticulation

7	92.90%	Structureless : Blotch irregular	Structureless : Blotch regular
5	71.60%	Lines : Branched streaks	Pattern : Starburst
5	61.40%	Globules / Clods : Cobblestone pattern	Pattern : Homogeneous : NOS
5	50.00%	Vessels : Corkscrew	Vessels : Polymorphous
4	89.40%	Dots : Regular	Globules / Clods : Cobblestone pattern
4	74.10%	Globules / Clods : Cobblestone pattern	Network : Negative pigment network
4	72.30%	Globules / Clods : Rim of brown globules	Pattern : Starburst
3	50.30%	ression structures : Peppering / Granularity	Structureless : Blotch regular

Figure 15.

A. Παράδειγμα δερματοσκοπικής εικόνας (ISIC_0022328) ενός μελανώματος που συμπεριλήφθηκε στη μελέτη μας για το 'exemplar' feature 'Structureless: Blue Whitish Veil'; **B-F.** Η δερματοσκοπική εικόνα επικαλυπτόμενη από τα superpixel outlines και τα annotation εκάστου εκ των 5 ειδικών για τα κριτήρια τα οποία επέλεξαν. **G.** Χρωματική απεικόνιση εκάστου εκ των κριτηρίων που επελέγησαν για αυτή την εικόνα. **H.** Συμφωνία μεταξύ όλων των ειδικών.

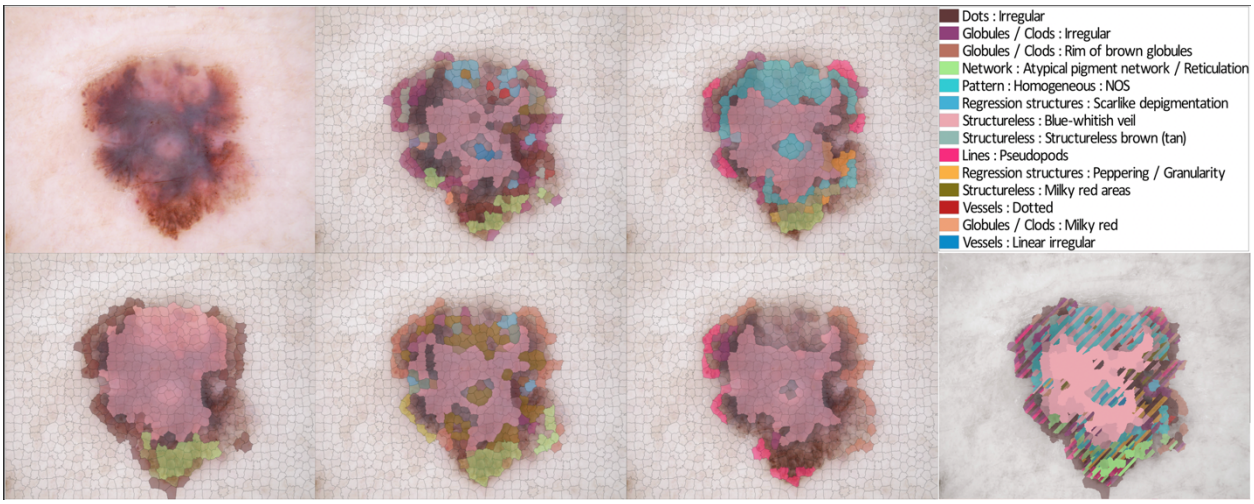
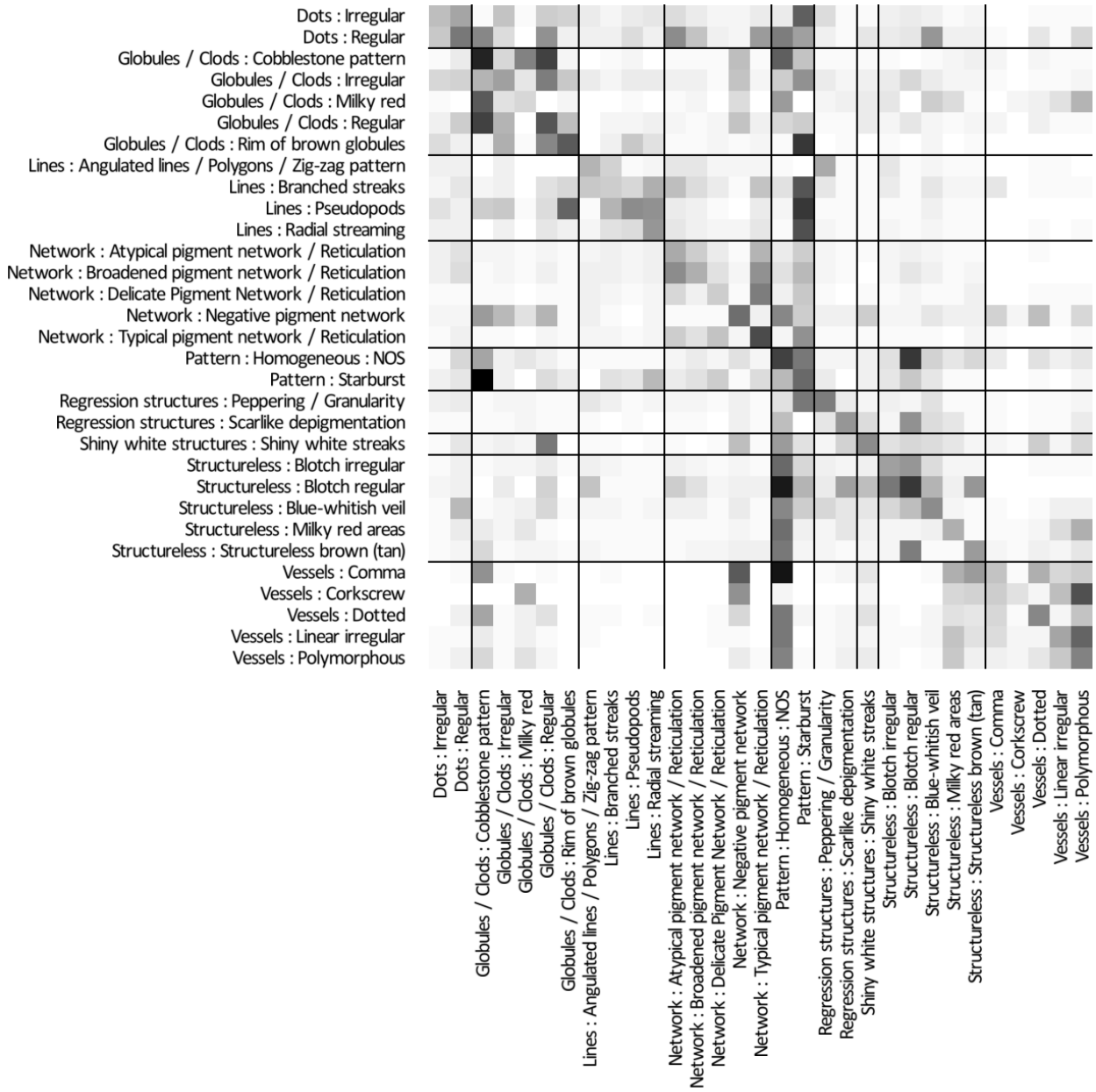


Figure 16.

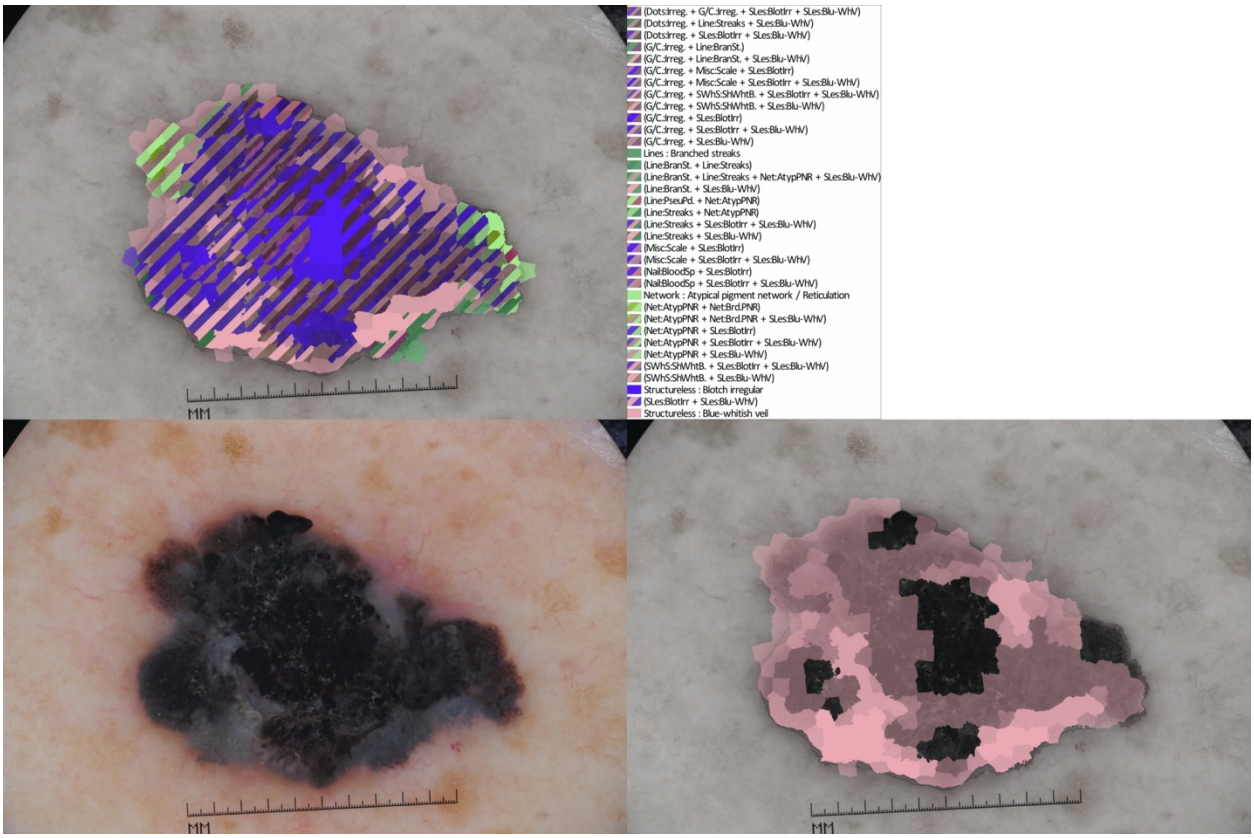
Confusability Matrix: Κάθε στοιχείο αυτού του confusion matrix είναι η διάμεσος της DICE coefficient για όλα τα ζεύγη κριτηρίων τα οποία προέκυψαν από τη μελέτη μας. Στις περιπτώσεις όπου τα διαγώνια στοιχεία είναι κατά το δυνατόν εγγύτερα στο 1.0 (σκούρο χρώμα), φαίνεται ότι τα ζεύγη των readers (κάθε reader συγκρινόμενος με καθένα από τους υπόλοιπους 4 οι οποίοι έκαναν annotate το ίδιο dataset) επέλεξαν παρόμοια superpixel για τα ίδια κριτήρια στην ίδια εικόνα. Όταν όλα τα στοιχεία της διαγώνιου πλησιάζουν το 0.0 (ανοιχτό χρώμα), φαίνεται ότι οι readers δεν επέλεξαν τα ίδια superpixel για τα εν λόγω κριτήρια. Από την εικόνα αυτή μπορούμε να εξάγουμε το συμπέρασμα ότι

επί παραδείγματι, το μη ειδικό κριτήριο ‘Pattern : Homogeneous’, το οποίο μπορεί να παρατηρηθεί σε οποιοδήποτε σημείο μιας βλάβης παρουσιάζει overlap με μια πλειάδα άλλων δερματοσκοπικών κριτηρίων, συνήθως με τα ‘Blotch : Regular’ and ‘Vessels : Comma’, ωστόσο, η ποσοστιαία συμφωνία ανάμεσα στους readers για το εν λόγω κριτήριο ήταν 0%. Επιπλέον, τα κριτήρια ‘Dots : Regular’ & ‘Dots : Irregular’ έχουν συχνό overlap μεταξύ τους καθώς και με τα ‘Globules / Clods’, ενώ τόσο το agreement στο επίπεδο των superpixel όπως και το Gwet’s AC ήταν χαμηλό για αυτά τα κριτήρια.



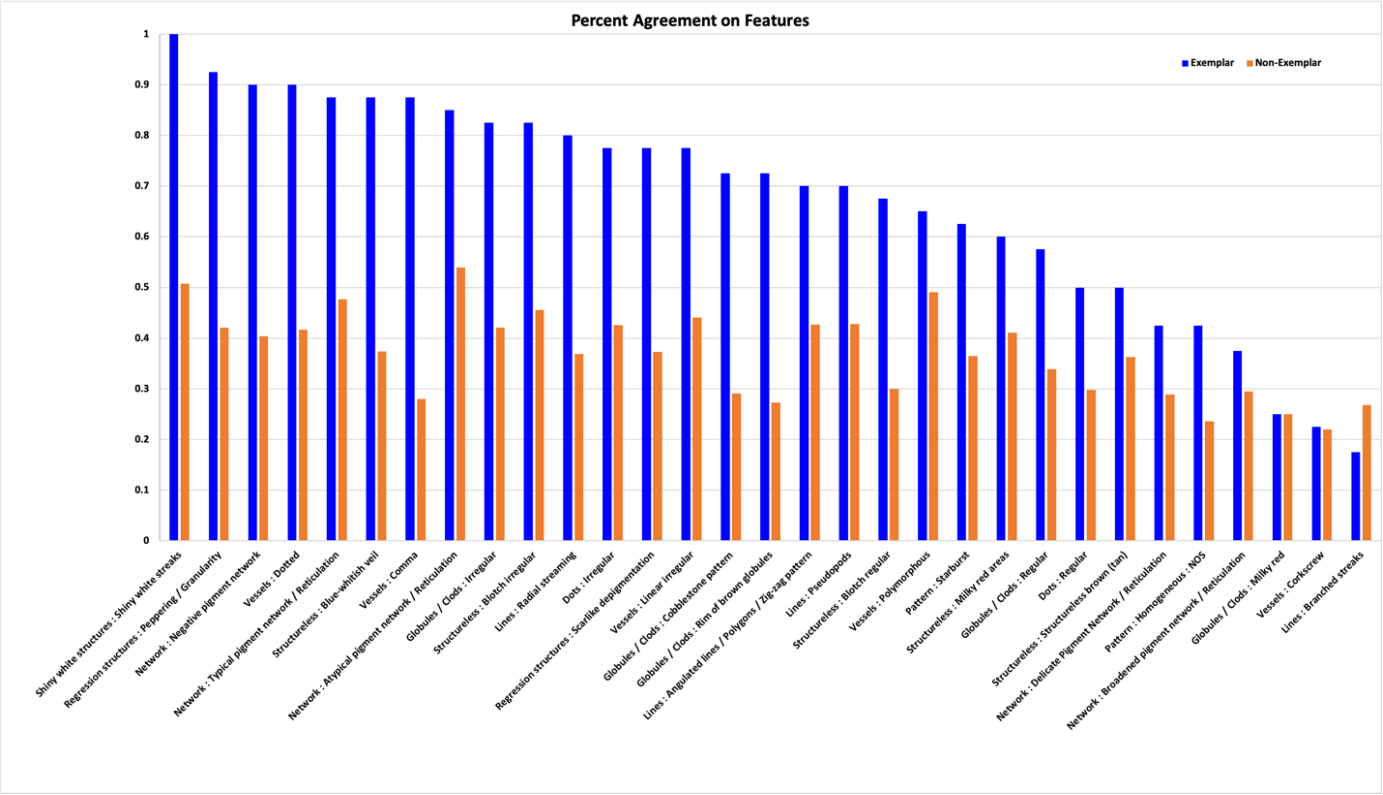
Supplementary Figure 1.

Παράδειγμα μιας δερματοσκοπικής εικόνας από τη μελέτη μας (ISIC_0016128), η οποία επιδεικνύει τόσο υψηλή συμφωνία για το exemplar feature (‘*Structureless : Blue-whitish veil’* - ροζ χρώμα), και υψηλό DICE overlap ανάμεσα στους readers για τα κριτήρια *‘Network : Atypical pigment network / Reticulation’* & *‘Network : Broadened pigment network / Reticulation’*.



Supplementary Figure 2.

Ποσοστιαία συμφωνία για όλα τα κριτήρια στο επίπεδο της βλάβης και για τις 248 εικόνες που συμπεριλήφθηκαν στη μελέτη μας. Μια συμφωνία της τάξεως του 100% (1.0) αντιπροσωπεύει θετική αναγνώριση του κριτηρίου από όλους τους ειδικούς, κάθε φορά που το κριτήριο αναγνωρίστηκε από τουλάχιστον έναν ειδικό σε μια βλάβη. Με μπλε χρώμα βλέπουμε τη συμφωνία όταν το κριτήριο παρουσιάστηκε ως exemplar feature ενώ με πορτοκαλί χρώμα είναι το agreement για όταν το κριτήριο αναγνωρίστηκε και ήταν το non-exemplar feature.



Tables

Supplementary Table S1:

Terminology Glossary: Η Δερματοσκόπηση έχει μια εκτεταμένη ορολογία και 2 διαφορετικές προσεγγίσεις στην ονοματολογία, μια περιγραφική και μια μεταφορική. Εδώ μπορούμε να δούμε τα ονόματα των κριτηρίων που συμπεριλάβαμε στη μελέτη μας, καθώς και την ανταπόκριση μεταξύ περιγραφικής και μεταφορικής ορολογίας, με τους συνοδούς ορισμούς και τις συντμήσεις που χρησιμοποιήσαμε στο κείμενο μας.[39]

Metaphoric Terminology	**Definition**	**Descriptive terminology**	**Abbreviation**
Dots: Regular	Dots clustered at the center of the lesion, or located on the network lines or in the hole of the network (also called target network)	Target: Dots, brown, central (in the center of hypopigmented spaces between reticular lines)	Regular Dots
Dots: Irregular	Any distribution of dots other than dots as described for regular dots		Irregular Dots
Globules / Clods: Cobblestone Pattern	Polygonal globules symmetrically distributed throughout lesion	Clods, brown or skin colored, large and polygonal	Cobblestone Pattern

Globules / Clods: Irregular	Globules with variability in color, size, shape, or spacing and distributed in an asymmetric/disorganized fashion		Irregular Globules
Globules / Clods: Milky-red		Clods, pink and small	Milky-red Globules
Globules / Clods: Regular	Globules with minimal variability in their color, size, and shape located in the center of a lesion with surrounding network, or along the perimeter or throughout the entire lesion	Clods, small, round or oval	Regular Globules
Globules / Clods: Rim of brown globules	Globules distributed at the periphery of lesion	Clods, brown, circumferential	Rim of Brown Globules
Lines: Angulated lines / Polygons / Zig-Zag	Gray-brown lines that are connected at an angle or coalescing to form polygons	Lines, angulated or polygonal (non-facial skin)	Angulated Lines

Lines: Branched streaks	Atypical network with broken/interrupted lines and incomplete connections		Branched Streaks
Lines: Pseudopods	Bulbous and often kinked projections seen at the lesion edge, either directly associated with a network or solid tumor border		Pseudopods
Lines: Radial streaming	Radial linear extensions at the lesion edge	Lines, radial and segmental	Radial streaming
Network: Atypical pigment network / Reticulation	Network with increased variability in line color, thickness and spacing. Gray color to lines or disorganized distribution	Lines, reticular and thick or reticular lines that vary in color	Atypical Network
Network: Broadened pigment network / Reticulation	Widening of the network lines	Lines, reticular and thick	Broadened Network
Network: Delicate pigment network / Reticulation	Fine thin network	Lines, reticular and thin	Delicate Network
Network: Typical pigment network / Reticulation	Network with minimal variability in the color, thickness, and spacing of the lines; symmetrically distributed	Lines, reticular	Typical Network

Network: Negative pigment network / Reticulation	Serpiginous interconnecting broadened hypopigmented lines that surround elongated and curvilinear brown structures	Lines, reticular, hypopigmented, around brown clods	Negative Network
Patterns: Starburst pattern	This pattern consists of tiered peripheral globules, pseudopods, or streaks (or a combination of them), located around the entire perimeter of the lesion	Pseudopods, circumferential or lines, radial, circumferential	Starburst Pattern
Patterns: Homogeneous pattern	A pattern lacking any definable pigment structures, also know as structureless pattern	Structureless, any color	Homogeneous Pattern
Regression structures: Peppering/granularity	Consists of fine dots with a blue-gray color	Dots, gray	Peppering / Granularity
Regression structures: Scar-like depigmentation	Area of white that is whiter than surrounding normal-appearing skin (true scarring); it should not be confused with hypopigmentation or depigmentation caused by simple loss of melanin; shiny white structures and blood vessels are not seen in areas of regression	Structureless zone, white	Scar-like Depigmentation

Shiny white structures: Shiny white streaks	Short discrete white lines oriented parallel and orthogonal (perpendicular) to each other seen only under polarized dermoscopy	Lines, white, perpendicular	Shiny White Streaks
Structureless: Blue whitish veil	A raised/palpable blotch of blue hue with an overlying whitish groundglass haze	Structureless zone, blue	Blue whitish veil
Structureless: Blotch Regular	One blotch within center of lesion and surrounded by network		Regular Blotch
Structureless: Blotch Irregular	More than one blotch or a blotch that is located off center		Irregular Blotch
Structureless: Tan (Brown) Peripheral Structureless areas		Structureless, brown (tan), eccentric	Tan Peripheral Structureless Areas
Structureless: Milky-red areas	Milky-white appearance or pinkish structureless areas (strawberry and ice cream-like), consisting a red vascular blush with no specific distinguishable vessels		Milky-red Areas
Vessel morphology, monomorphous: Dots	Tiny pinpoint vessels		Dotted Vessels
Vessel morphology, monomorphous: Comma	Linear, curved, short vessels	Curved	Comma Vessels

Vessel morphology, monomorphous: Corkscrew	Twisted looped vessels with bends twisted along a central axis	Helical	Corkscrew Vessels
Vessel morphology, monomorphous: Linear irregular / Serpentine	Linear, curved or serpentine vessels,	Serpentine	Linear Irregular Vessels
Vessel morphology: Polymorphous			Polymorphous Vessels

Supplementary Table S2.

Δερματοσκοπικά κριτήρια ειδικά για τη διάγνωση μελανοκυτταρικών βλαβών (σπίλοι και μελανώματα) τα οποία συμπεριλήφθηκαν στη μελέτη μας, ο συνολικός αριθμός παρατηρήσεων (lesion level - Total N of observations), ο αριθμός των εικόνων στις οποίες παρατηρήθηκαν (in images) και η συμφωνία η οποία παρατηρήθηκε για τα κριτήρια αυτά κατ' αντιστοιχία. Παρατηρήσεις από μεμονωμένους readers - Orphan observations, ≥ 40% - 2-RA, συμφωνία μεταξύ τουλάχιστον 2 readers, ≥ 60% - 3-RA, συμφωνία μεταξύ τουλάχιστον 3 readers, ≥ 80% - 4-RA, συμφωνία μεταξύ τουλάχιστον 4 readers και 100% agreement, συμφωνία μεταξύ όλων των readers οι οποίοι αξιολόγησαν τις εικόνες.

Dermoscopic Feature	**Total N of observations**	**In images**	**Orphan observations**	**Lesions ≥ 40% agreement (≥2RA)**	**Lesions ≥ 60% agreement (≥3RA)**	**Lesions ≥ 80% Agreement (≥4RA)**	**Lesions with 100% agreement**

Dots : Irregular	**268**	**124**	**48**	**76**	**44**	**18**	**6**
Dots : Regular	**102**	**65**	**37**	**28**	**7**	**2**	**0**
Globules / Clods : Cobblestone pattern	**60**	**29**	**14**	**15**	**9**	**5**	**2**
Globules / Clods : Irregular	**319**	**150**	**67**	**83**	**49**	**27**	**10**
Globules / Clods : Milky red	**40**	**31**	**24**	**7**	**2**	**0**	**0**
Globules / Clods : Regular	**103**	**57**	**28**	**29**	**14**	**3**	**0**
Globules / Clods : Rim of brown globules	**58**	**29**	**18**	**11**	**8**	**6**	**4**

Lines : Angulated lines/Polygons/Zig-zag pattern	**72**	**30**	**9**	**21**	**16**	**5**	**0**
Lines : Branched streaks	**82**	**61**	**43**	**18**	**3**	**0**	**0**
Lines : Pseudopods	**116**	**50**	**19**	**31**	**23**	**11**	**1**
Lines : Radial streaming	**135**	**66**	**33**	**33**	**21**	**12**	**3**
Network : Atypical pigment network / Reticulation	**344**	**128**	**32**	**96**	**67**	**39**	**14**
Network : Broadened pigment	**130**	**86**	**51**	**35**	**9**	**0**	**0**

network / Reticulation							
Network : Delicate Pigment Network / Reticulation	**124**	**81**	**55**	**26**	**12**	**5**	**0**
Network : Negative pigment network	**124**	**54**	**22**	**32**	**19**	**13**	**6**
Network : Typical pigment network / Reticulation	**245**	**101**	**34**	**67**	**39**	**26**	**12**
Pattern : Homogeneous : NOS	**88**	**71**	**57**	**14**	**2**	**1**	**0**

Pattern : Starburst	**53**	**24**	**10**	**14**	**8**	**5**	**2**
Regression structures : Peppering / Granularity	**249**	**114**	**48**	**66**	**35**	**23**	**11**
Regression structures : Scarlike depigmentation	**161**	**81**	**42**	**39**	**24**	**11**	**6**
Shiny white structures : Shiny white streaks	**218**	**82**	**24**	**58**	**40**	**25**	**13**
Structureless : Blotch irregular	**227**	**98**	**44**	**54**	**37**	**27**	**11**
Structureless : Blotch regular	**71**	**39**	**24**	**15**	**9**	**6**	**2**

Structureless : Blue-whitish veil	**183**	**93**	**39**	**54**	**27**	**6**	**3**
Structureless : Milky red areas	**169**	**80**	**32**	**48**	**25**	**15**	**1**
Structureles: Structureless brown (tan)	**250**	**138**	**62**	**76**	**28**	**7**	**1**
Vessels : Comma	**63**	**28**	**16**	**12**	**10**	**8**	**5**
Vessels : Corkscrew	**20**	**17**	**14**	**3**	**0**	**0**	**0**
Vessels : Dotted	**147**	**63**	**26**	**37**	**25**	**15**	**7**
Vessels : Linear irregular	**143**	**61**	**22**	**39**	**25**	**16**	**2**
Vessels : Polymorphous	**143**	**54**	**14**	**40**	**28**	**16**	**5**

Συζήτηση

Ο καρκίνος του δέρματος είναι ο πιο συχνός καρκίνος στον κόσμο.[95] Υπολογίζεται πως μόνο στις ΗΠΑ η επίπτωση του μη μελανοκυτταρικού καρκίνου δέρματος (βασικοκυτταρικό και ακανθοκυτταρικό καρκίνωμα) ενδέχεται να ξεπερνάει τα πέντε εκατομμύρια περιπτώσεις το χρόνο.[11,12] Ταυτόχρονα, ετησίως, καταγράφονται 220,000 περιπτώσεις μελανώματος, της πιο θανατηφόρας μορφής καρκίνου του δέρματος, καθώς και 37,000 θάνατοι εξαιτίας του μελανώματος σε Ευρώπη και ΗΠΑ.[117] Τα τελευταία χρόνια έχει καταγραφεί σημαντική πρόοδος στην επιβίωση και την ποιότητα ζωής των ασθενών με προχωρημένο καρκίνο του δέρματος.[15-19] Παρόλα αυτά, ο σημαντικότερος τρόπος για την μείωση τόσο της θνητότητας, όσο και της νοσηρότητας εξαιτίας του καρκίνου δέρματος παραμένει η έγκαιρη διάγνωση και χειρουργική εξαίρεση του όγκου.[19-21]

Στη διατριβή μας εξετάσαμε τον πιθανό ρόλο που θα μπορούσαν να αποκτήσουν οι αλγόριθμοι Τεχνητής Νοημοσύνης στην έγκαιρη διάγνωση του καρκίνου του δέρματος. Οι αλγόριθμοι οι οποίοι βασίζονται στο Machine Learning έχουν δείξει πολλά υποσχόμενα αποτελέσματα σε διάφορους τομείς της ιατρικής όπως η οφθαλμολογία και η ακτινολογία ενώ ήδη υπάρχουν FDA εγκεκριμένοι αλγόριθμοι οι οποίοι χρησιμοποιούνται επικουρικά στις ανωτέρω ειδικότητες.[9] Μέσω του International Skin Imaging Collaboration (ISIC) Archive (www.isic-archive.com) επιχειρήσαμε να διερευνήσουμε τόσο τη διαγνωστική ακρίβεια που έχουν οι αλγόριθμοι Τεχνητής Νοημοσύνης για τη διάγνωση του καρκίνου του δέρματος, όσο και ενδεχόμενες εφαρμογές τους στην κλινική πράξη. Το ISIC Archive είναι μια συνεργασία μεταξύ της ακαδημαϊκής κοινότητας και της βιομηχανίας, η οποία έχει πρωταρχικό στόχο τη μείωση της θνησιμότητας του μελανώματος μέσω της εφαρμογή νέων απεικονιστικών τεχνολογιών. Προς αυτό το στόχο, το ISIC Archive φιλοξενεί ετήσιους διαγωνισμούς (2016-2020) χρησιμοποιώντας το

διαρκώς αυξανόμενο αρχείο δερματοσκοπικών του εικόνων. Εδώ εκθέσαμε τα αποτελέσματα του ISIC Challenge 2017, καθώς και τα ευρήματα μας από την reader study η οποία επακολούθησε και η οποία συμπεριελάμβανε τόσο ειδικούς στη διάγνωση του καρκίνου δέρματος, όσο και ειδικευόμενους δερματολόγους.[28,30] Στην έρευνα μας αυτή, βρήκαμε ότι ο κορυφαίος αλγόριθμος, σε ένα δεδομένο όριο ευαισθησίας, είχε καλύτερη ειδικότητα για τη διάγνωση μελανώματος και από τους ειδικούς και από τους ειδικευόμενους. Το πιο εντυπωσιακό από τα ευρήματα μας ωστόσο ήταν το γεγονός ότι σε περιπτώσεις όπου οι κλινικοί ιατροί, και ειδικά εκείνοι με τη μικρότερη εμπειρία (ειδικευόμενοι), είχαν χαμηλή εμπιστοσύνη στη διάγνωση τους, η επικουρική χρήση του αλγορίθμου θα μπορούσε να βελτιώσει τη διαγνωστική τους ακρίβεια, ευρήματα που επιβεβαιώθηκαν και σε επακόλουθες έρευνες.[59,116]

Ακολούθως, επιδιώξαμε να ελέγξουμε την αναπαραγωγιμότητα δημοσιευμένων αλγορίθμων Τεχνητής Νοημοσύνης, ενδεχόμενους παράγοντες που μπορεί να επηρεάζουν το διαγνωστικό τους αποτέλεσμα, καθώς και τη δυνατότητα τους να εφαρμοστούν σε διαφορετικό πληθυσμό από αυτόν στον οποίον εκπαιδεύτηκαν.[56,59] Χρησιμοποιήσαμε ένα standardized dataset καρκίνων δέρματος, το οποίο είναι δημόσιο,[10] προκειμένου να ελέγξουμε δυο δημόσια διαθέσιμους αλγορίθμους και βρήκαμε πως η διαγνωστική ακρίβεια των αλγορίθμων Τεχνητής Νοημοσύνης μειώνεται σημαντικά όταν χρησιμοποιούνται βλάβες από ασθενείς διαφορετικού τύπου δέρματος.[31,32] Επιπρόσθετα βρήκαμε πως οι αλγόριθμοι Τεχνητής Νοημοσύνης οι οποίοι χρησιμοποιούν εικόνες για τη διάγνωση δερματολογικών παθήσεων μπορεί να επηρεαστούν από διάφορους παράγοντες, όπως από την περιστροφή της εικόνας, ή από παρεμβάσεις ως προς τη φωτεινότητα ή την αντίθεση της εικόνας.[31] Τέλος διαπιστώσαμε πως η διαγνωστική ακρίβεια των αλγορίθμων εξαρτάται σε μεγάλο βαθμό από τη μεγέθυνση με την οποία θα υποβληθεί μια εικόνα προς ανάλυση, το ανατομικό σημείο όπου βρίσκονταν οι βλάβες, καθώς και από το αν οι εικόνες είναι προσεκτικά τοποθετημένες στο κέντρο του

ROI.[32] Τα ευρήματα μας αυτά συμβαδίζουν με αντίστοιχες έρευνες στο πεδίο, καταδεικνύοντας έτσι κάποιες από τις εγγενείς αδυναμίες τους, οι οποίες θα πρέπει να υπερκεραστούν προκειμένου να βρουν εφαρμογή και χρησιμότητα στην καθημερινή κλινική πρακτική.[61-63]

Στα πλαίσια της διατριβής, προτείναμε νέες ερμηνευτικές και στατιστικές προσεγγίσεις, οι οποίες θα μπορούσαν να είναι πιο κατάλληλες για την αξιολόγηση της εφαρμογής αυτών των νέων τεχνολογιών στη διάγνωση του καρκίνου δέρματος, καταδεικνύοντας ταυτόχρονα την ανάγκη να δοκιμαστούν σε έναν γενικό πληθυσμό, και όχι μόνο στα πλαίσια ερευνών με αυστηρά επιλεγμένα dataset στα οποία υπερεκπροσωπούνται οι καρκίνοι δέρματος, και έτι περαιτέρω, το μελάνωμα. Σε αυτά τα πλαίσια, θα απαιτηθούν και διπλά τυφλές, τυχαιοποιημένες μελέτες, καθώς και προοπτικές μελέτες αξιολόγησης των αλγορίθμων Τεχνητής Νοημοσύνης, μελέτες τις οποίες ήδη τρέχουμε στο νοσοκομείο «Ανδρέας Συγγρός», επιχειρώντας να προσδιορίσουμε με μεγαλύτερη ακρίβεια τη χρησιμότητα των αλγορίθμων αυτών.

Όπως εκθέσαμε προηγουμένως, η δερματοσκόπηση είναι μια εδραιωμένη, μη επεμβατική τεχνική η οποία έχει βοηθήσει σημαντικά στην έγκαιρη διάγνωση του καρκίνου του δέρματος, τόσο των μη-μελανοκυτταρικών καρκίνων, όσο και του μελανώματος. [22,23] Ωστόσο, η συμφωνία πάνω στα επιμέρους δερματοσκοπικά κριτήρια παραμένει από χαμηλή έως μέτρια, ακόμα και ανάμεσα σε ειδικούς της δερματοσκόπησης. [39,75,109] Μέσω της έρευνας μας περιγράψαμε νέα δερματοσκοπικά κριτήρια i. για τη διάγνωση δύσκολων στη διάγνωση μελανωμάτων, όπως το αμελανωτικό μελάνωμα των άκρων, ii. Βασικοκυτταρικών καρκινωμάτων υψηλού ρίσκου, όπως τα morpheaform και infiltrative και τέλος, iii. συχνών μιμητών τους.[34,36,37,118] Επιπλέον, διεξάγαμε την Expert Agreement Study on Dermoscopy of melanocytic lesions μέσω της οποίας προσδιορίσαμε την αξιοπιστία, την αναπαραγωγιμότητα και την αλληλοεπικάλυψη των δερματοσκοπικών κριτηρίων που

χρησιμοποιούνται για τη διαφοροδιάγνωση καλοήθων από κακοήθεις μελανοκυτταρικές βλάβες. Τα αποτελέσματα μας μπορούν να χρησιμοποιηθούν ώστε 1. Να βελτιωθούν οι διαγνωστικοί αλγόριθμοι που χρησιμοποιούν οι κλινικοί ιατροί για τη διάγνωση μελανώματος και 2. να χρησιμοποιηθούν από ερευνητές Τεχνητής Νοημοσύνης για πιο στοχευμένη εκπαίδευση των αλγορίθμων AI.[27,28]

Ένα ερώτημα το οποίο τίθεται συχνά αφορά στο τι εξυπηρετεί η μελέτη των αλγορίθμων Τεχνητής Νοημοσύνης για τη διάγνωση του καρκίνου δέρματος. Κατά πόσο χρειαζόμαστε τους αλγορίθμους Τεχνητής Νοημοσύνης για τη διάγνωση του καρκίνου δέρματος, και εν τέλει, αν η χρήση τους θα αποδειχθεί χρήσιμη για τον ασθενή και αν θα εξυπηρετήσει τον τελικό στόχο, ήτοι την καλύτερη παροχή φροντίδας στον ασθενή.

Σε αυτό το σημείο χρειάζεται να προβούμε σε κάποιες παραδοχές και να αναγνωρίσουμε κάποιες αληθείς προκλήσεις που αντιμετωπίζουμε ως κλινικοί ιατροί αναφορικά με τη διάγνωση του καρκίνου δέρματος. Η πλειονότητα των μελανωμάτων, περί το 75%, δεν διαγιγνώσκεται αρχικά από τους δερματολόγους, αλλά εντοπίζεται από τον ίδιο τον ασθενή ή κάποιο μέλος της οικογένειας του.[119-124] Επιπλέον, ένα μεγάλο ποσοστό ασθενών δεν θα λάβει ποτέ εξειδικευμένη δερματολογική φροντίδα, λόγω περιορισμένης πρόσβασης.[125]. Ενώ, τέλος, η διαγνωστική μας ακρίβεια ως δερματολόγοι για τη διάγνωση του μελανώματος δεν είναι εξαιρετικά υψηλή.[126] Η εφαρμογή των αλγορίθμων Τεχνητής Νοημοσύνης θα μπορούσε ενδεχομένως να άρει κάποιους από αυτούς τους περιορισμούς, ειδικά εάν μπορέσουν να γίνουν ευρέως διαθέσιμες, ενσωματωμένες σε εφαρμογές έξυπνων κινητών τηλεφώνων επί παραδείγματι.

Ωστόσο, σε αυτό το σημείο υπεισέρχεται ένα εξαιρετικά σημαντικό ερώτημα, το οποίο κατά την άποψη μας θα πρέπει να απαντηθεί από την επιστημονική κοινότητα. Έστω ότι σε λίγα χρόνια δημιουργούνται

αλγόριθμοι Τεχνητής Νοημοσύνης με σχεδόν τέλεια διαγνωστική ακρίβεια, σε ποιους θα απευθύνεται ένα τόσο ισχυρό διαγνωστικό εργαλείο; Άραγε θα πρέπει να είναι διαθέσιμο σε όλους, ακόμα και στο ευρύ κοινό; Και σημαντικότερα, κατά την Ιπποκρατική θεώρηση, η χρήση του θα επιφέρει περισσότερο όφελος από ότι ζημία; Τα τελευταία χρόνια παρατηρείται μια διαρκής αύξηση του αριθμού μελανωμάτων που διαγιγνώσκονται ετησίως, ενώ η θνητότητα από το μελάνωμα παραμένει σταθερή, ενώ εσχάτως, για πρώτη φορά σε διάρκεια δεκαετιών, παρατηρείται μείωση της.[127] Αντίστοιχη είναι η τάση και για τη διάγνωση των μη-μελανοκυτταρικών καρκίνων δέρματος.[12,95] Κατά την άποψη μας, η μείωση της θνητότητας του μελανώματος και των άλλων καρκίνων δέρματος οφείλεται σε μεγάλο βαθμό στις νέες, στοχευμένες, συστηματικές θεραπείες,[17,19] ενώ η αύξηση στον επιπολασμό του καρκίνου δέρματος οφείλεται σε 3 σημαντικούς παράγοντες: 1. Το προσδόκιμο ζωής, ειδικά στις δυτικές κοινωνίες αυξάνεται διαρκώς και αναμένεται να διατηρήσει την ίδια τάση καθ' όλο τον 21ο αιώνα,[128] 2. Η ευαισθησία και η ειδικότητα μας, ως κλινικών ιατρών έχει βελτιωθεί σημαντικά με την προσθήκη νέων διαγνωστικών εργαλείων, όπως η δερματοσκόπηση και το Reflectance Confocal Microscopy,[22,23,72] και 3. Σημαντικό ρόλο στην προαναφερθείσα αύξηση της διάγνωσης παίζει το diagnostic drive, η διαρκής ώθηση για τη διάγνωση όλο και μικρότερων μελανωμάτων, όλο και πιο πρώιμων καρκίνων δέρματος.[129-131]

Το ερώτημα το οποίο εγείρεται ωστόσο είναι αν διαγιγνώσκοντας καρκίνους δέρματος σε εξαιρετικά πρώιμο στάδιο, ή καρκίνους οι οποίοι θα είχαν αμφίβολη επίπτωση στο προσδόκιμο ή στην ποιότητα ζωής ενός ασθενούς κάνουμε περισσότερο καλό παρά κακό.[131-133] Σε αυτό το πλαίσιο έγκειται και το αν η υπερβολική χρήση της τεχνολογίας και των πλέον εξελιγμένων διαγνωστικών εργαλείων μπορεί να οδηγεί σε αντίθετα από τα επιθυμητά αποτελέσματα. Δηλαδή, εάν και κατά πόσον η χρήση της τεχνολογίας μπορεί να οδηγεί στη διάγνωση «ακίνδυνων» καρκίνων δέρματος, οι οποίοι ενδεχομένως δεν θα προκαλούσαν σοβαρές, ή και καμία επίπτωση στον ασθενή εάν δεν ήμασταν τόσο ικανοί στην

ανίχνευση τους, ερωτήματα αντίστοιχα με αυτά που εγείρονται και στο screening για τον καρκίνο του μαστού και του προστάτη.[131,134] Τι μπορεί να συμβεί αντίστοιχα εάν ένα τόσο ισχυρό διαγνωστικό εργαλείο, με τη δυνατότητα να ανιχνεύει την παραμικρή μεταβολή σε μέγεθος, σχήμα, ή χρώμα μιας βλάβης δέρματος, βρεθεί στα χέρια δισεκατομμυρίων ασθενών. Η πίεση στα συστήματα υγείας θα μπορούσε να είναι τρομακτική, για την διάγνωση καρκίνων με αμφίβολη κακοήθη πρόγνωση (uncertain malignant potential) ενώ ταυτόχρονα το όφελος που θα αποκόμιζαν οι ασθενείς περιορισμένο.

Τέλος, σε όλα τα προαναφερθέντα θα πρέπει να ληφθεί υπόψιν και ο ρόλος των ιατρών, ως κάτι περισσότερο από ανθρώπους οι οποίοι απλά διαγιγνώσκουν και θεραπεύουν μια βλάβη δέρματος, ειδικά όταν αυτή η βλάβη είναι καρκίνος δέρματος. Η σχέση εμπιστοσύνης ιατρού-ασθενούς είναι ενδεχομένως ένα από τα σημαντικότερα βήματα στην παροχή φροντίδας του ασθενούς, πολλώ δε μάλλον, του ασθενούς με καρκίνο. Η χρήση των νέων τεχνολογιών, ειδικά εάν μας κάνουν καλύτερους κλινικούς ιατρούς, είναι παραπάνω από ευπρόσδεκτη, μόνο όμως ενταγμένες σε ένα πλαίσιο στο οποίο θα μπορέσουμε να παρέχουμε καλύτερη φροντίδα στους ασθενείς μας.[135]

Καταλήγοντας, η πρόοδος της τεχνολογίας έχει αναμφισβήτητα προσφέρει ανεκτίμητες υπηρεσίες στο ανθρώπινο γένος και το ίδιο αναμένεται να γίνει και με την Τέταρτη Βιομηχανική Επανάσταση, όπου κυρίαρχο ρόλο θα διαδραματίζουν οι αλγόριθμοι Τεχνητής Νοημοσύνης. Αναμένουμε πως η χρήση αυτής της νέας τεχνολογίας θα διαδραματίσει εξίσου σημαντικό ρόλο και στην Ιατρική, και στη δερματολογία συγκεκριμένα. Κατά τη γνώμη μας ωστόσο, θα πρέπει να είναι η επιστημονική κοινότητα εν τω συνόλω η οποία θα καθορίσει τα πλαίσια χρήσης των αλγορίθμων Τεχνητής Νοημοσύνης, ώστε εν τέλει, οι κλινικοί ιατροί, υποβοηθούμενοι από εφαρμογές Τεχνητής Νοημοσύνης, θα μπορούν διαρκώς βελτιούμενα, να «Ωφελέειν, η μη βλάπτειν».

Συμπέρασμα

Συμπερασματικά, στα πλαίσια της διατριβής μας, (1) βρήκαμε ότι οι αλγόριθμοι Τεχνητής Νοημοσύνης είναι πολλά υποσχόμενοι και θα μπορούσαν μελλοντικά να συμβάλλουν σημαντικά στη βελτίωση της διαγνωστικής ακρίβειας των κλινικών ιατρών για τη διάγνωση του καρκίνου δέρματος. (2) Εντοπίσαμε πως ιατροί με μικρότερη κλινική εμπειρία ενδεχομένως να ωφεληθούν περισσότερο από τη χρήση τους. (3) Καταδείξαμε κάποιες σημαντικές εγγενείς αδυναμίες των αλγορίθμων Τεχνητής Νοημοσύνης οι οποίες οφείλουν να υπερκεραστούν προτού ενταχθούν στην κλινική πράξη. (4) Προτείναμε νέους τρόπους προσέγγισης και στατιστικής ανάλυσης για την ερμηνεία των αποτελεσμάτων που παράγουν οι έρευνες με αλγορίθμους Τεχνητής Νοημοσύνης. Και, (5) περιγράψαμε καινούργια διαγνωστικά κριτήρια για τη διάγνωση του καρκίνου του δέρματος, αλλά και εντοπίσαμε αδυναμίες των δερματοσκοπικών, διαγνωστικών κριτηρίων για τη διάγνωση του μελανώματος, τα οποία θα μπορούσαν να χρησιμοποιηθούν για τη βελτίωση των αλγορίθμων Τεχνητής Νοημοσύνης. Τέλος, περιγράψαμε τα πλεονεκτήματα, αλλά και τα ενδεχόμενα μειονεκτήματα που θα μπορούσαν να προκύψουν από τη χρήση της Τεχνητής Νοημοσύνης για τη διάγνωση του καρκίνου δέρματος, καθώς και το πλαίσιο στο οποίο μπορούν να ενταχθούν, πάντα υπό την καθοδήγηση και την εποπτεία της επιστημονικής κοινότητας.

Introduction

Artificial Intelligence (AI) is a rapidly developing research field which involves a wide spectrum of daily life. This spectrum extends from voice recognition to automation of education, transport and renewable energy sources.[1-3] Medicine could be no exemption; a plurality of recent scientific publications has explored AI application in fields such as ophthalmology, radiology and pathology among others.[4-6]

The term AI refers to the development of automated computing systems that are capable of performing actions which, under normal circumstances, would require human intelligence, such as visual recognition, speech recognition, decision making, and translation.[7] A subfield of AI is "machine learning" (ML), which entails the scientific research of algorithms and statistical models used by computers in order to perform a function, without being specifically instructed how to perform that action. ML algorithms build a mathematical model, based on their training data in order to make forecasts, or perform a specific function, without being specifically instructed how to perform that action.[8] These ML algorithms are the ones that are mainly used in Medicine, while, over the recent years, the American Food and Drug Administration has approved 64 AI algorithms for use in clinical practice, with 29 of them being ML algorithms (45%).[9] The vast majority of these algorithms concerns radiology, ophthalmology, internal medicine and emergency medicine.[9]

Herein we attempted to explore the potential that AI algorithms could have in early skin cancer diagnosis, as well as to facilitate research in the field of skin oncology. However, the lack of large, public databases has so far limited the progress of deep learning algorithms for the diagnosis of skin cancer

and accordingly, no AI algorithm has, this far, been applied in daily clinical practice efficiently. In order to overcome these barriers we created International Skin Imaging Collaboration Archive (ISIC - www.isic-archive.com) which is an open source, public archive of skin images, available to the public for both educational and research purposes.[10] ISIC Archive can be used by both physicians and AI researchers and has been utilized to date to fuel more than 5000 publications.[10]

The reason we elected to deal with skin cancer lays on the fact that skin cancer is the most common cancer, with more than 5 million cases in the United States alone, while melanoma is the 5th cause of cancer death, with more than 9000 deaths each year.[11-14] Recently, there has been significant progress both in survival and quality of life of patients with metastatic melanoma, or regionally advanced and metastatic Basal Cell Carcinoma (BCC) or Squamous Cell Carcinoma (SCC). This progress has been attributed to the new, targeted, systemic therapies for these diseases, including immune check point inhibitors and selective inhibitors of pathways implicated in carcinogenesis, such as 'Hedgehog' pathway inhibitors.[15-19] However, despite progress in systemic treatments, the mainstay for skin cancer treatment remains early diagnosis and surgical excision of the tumor.[19-21]

Dermoscopy has played a critical role towards this direction over the past 20 years. Dermoscopy is a widely available, non-invasive, diagnostic technique, which has been proven to aid in early skin cancer diagnosis compared to naked eye examination alone.[22,23] Dermoscopy has become an intrinsic piece of dermatology and is now considered "dermatologist's stethoscope", while its use has expanded beyond the realms of skin oncology to include general dermatology as well.[24,25] An additional advantage of dermoscopy is that dermoscopic pictures are captured with the dermatoscope directly attached to the skin of the patient, with a standardized 10x magnification, thus preventing dermoscopic images to render a patient recognizable while protecting sensitive patient information.

Through our research we attempted to shed some light to the following: (1) Can AI be used in early skin cancer diagnosis? (2) Which are the pitfalls of AI algorithms in Dermatology and in which possible ways could they be improved upon? (3) How could AI be of use in Dermatology, and in which ways could it be used? (4) Which is the best approach to the research conducted with regards to AI algorithms in early skin cancer diagnosis, and how should these results be interpreted? And (5), how can we improve the diagnostic accuracy of clinicians and AI algorithms for early skin cancer diagnosis, and more specifically, melanoma? Finally, we attempted to describe the overall framework, within which, AI algorithms could be proven useful in clinical practice, and more importantly, beneficial to the patients.

To achieve these goals, we collaborated during our thesis with distinguished scientists from the USA, Greece, Chile, Spain, Australia, and Austria. Via ISIC Archive we have been organizing (2016-2020) annual Grand Challenges, where AI researchers compete in constantly expanding datasets for skin cancer diagnosis. During these challenges we explored the diagnostic accuracy of AI algorithms and compared them to the diagnostic accuracy of clinicians. We also examined the potential of AI algorithms to segment lesions from background skin, as well as their capabilities in detecting preselected dermoscopic features.[26-29] Additionally, we attempted to detect under which conditions these AI algorithms could be useful in the clinical setting by examining the diagnostic accuracy of clinicians at different stages of their training, compared to AI algorithms, or in adjunct with them.[30] We also explored the generalizability and reproducibility of published, publicly available AI algorithms on a standardized, public skin cancer dataset in order to reveal potential pitfalls of these algorithms and aid the scientific community in improving these pitfalls.[31,32] During our research we applied novel statistical approaches and ways of interpretation to better comprehend the results of both our research and the research of other scientists.[33] Finally, we tried to improve the diagnostic accuracy of both clinicians and AI algorithms by describing new dermoscopic criteria for the diagnosis of skin cancer, especially difficult to diagnose skin

cancers and its mimickers, and by conducting the first Expert Agreement Study on Dermoscopy of melanocytic lesions (EASY study).[34-37] EASY study examined the agreement of experts in dermoscopy for the reliability and reproducibility of established dermoscopic features for differentiating nevi and melanomas, not only on the presence of the feature, but also on its spatial localization.

ISIC Grand Challenges and Reader Studies

Introduction

As discussed prior, in order to examine the potential application of AI in early skin cancer diagnosis, as well as to face the lack of publicly available databases of skin lesion images, and particularly, dermoscopic images, we created ISIC Archive.[10] Through ISIC Archive we have been organizing since 2016 annual ISIC Grand challenges with continuously growing level of difficulty and complexity, along with number of images, diagnoses as well as challenges, in which AI investigators from around the world compete, making ISIC Archive the reference standard for research in the field.

ISIC Grand Challenges take place annually during well-established, international conferences. In particular, International Symposium on Biomedical Imaging (ISBI, 2016-2017), Medical Image Computing and Computer Assisted Intervention (MICCAI, 2018-2020), Conference on Computer Vision and Pattern Recognition (CVPR, 2019-2020) and Society for Imaging Informatics (SIIM, 2019-2020). During the conduct of this thesis, I served as a clinical coordinator of ISIC Archive, actively participating in all the competitions that took place from 2017 till now, while, since 2020, National and Kapodistrian University of Athens, via 'Andreas Sygros' Hospital for Skin and Venereal diseases is officially a part of the established collaboration between Memorial Sloan Kettering Cancer Center, Medical University of Vienna, Barcelona Hospital Clinic, Emory University, University of Queensland and Sydney Melanoma Diagnostic Center. ISIC Grand Challenges are competitions where AI investigators compete with their algorithms in the accurate diagnosis of skin cancer and its benign mimickers. These competitions are followed by reader studies, conducted among clinicians of different subspecialties and training stage

(dermatologists, dermatology residents, general practitioners and residents of general medicine) in order to examine (1) the diagnostic accuracy of these algorithms compared to clinicians and (2) where these algorithms could be useful. These competitions are enriched each year with a larger number of lesions, diagnoses and research challenges which determine the evolution of the field, in an effort to identify the potential utility of AI algorithms in skin cancer diagnosis.[26-29]

In the first ISIC Grand Challenge 2016, AI algorithms competed on 1279 dermoscopic images (248 melanomas and 1031 nevi) which were partitioned in a training (n=900l 19.2% melanomas) and a test dataset (n=379, 19.8% melanomas).[27] In the reader study that followed, including 8 expert readers, we found that the fusion algorithm of the best performing algorithms had a better performance compared to the experts (sensitivity 76% vs 59% for the experts, p=0.02).[38]

Herein we are going to present the results of ISIC Grand Challenge 2017, where we expanded the number of diagnoses to include melanomas, nevi and seborrheic keratoses. At the same time, we expanded the challenges where the algorithms competed to include 3 tasks: (1) lesion segmentation, (2) dermoscopic feature detection και (3) classification.[28] Additionally, we will present the reader study that followed the challenge, where we compared the diagnostic accuracy of the best performing algorithm with the performance of expert dermatologists and dermatology residents. We also examined what would happen if we were to substitute the diagnosis of clinicians by the algorithm prediction, in cases where the clinical confidence in diagnosis was low.[30]

ISIC Grand Challenge 2017

Materials and Methods, ISIC Grand Challenge 2017

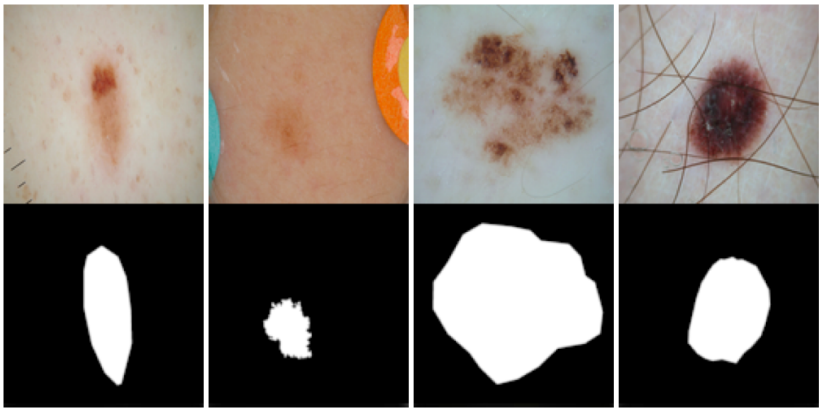
The 2017 challenge consisted of 3 tasks: lesion segmentation, dermoscopic feature detection, and disease classification. For each, data consisted of images and corresponding ground truth annotations, split into training (n=2000), validation (n=150), and holdout test (n=600) datasets. Predictions could be submitted on validation and test datasets. The validation submissions provided instantaneous feedback in the form of performance evaluations, as well as ranking in comparison to other participants. Test submissions only provided feedback after the submission deadline. The training, validation, and test datasets continue to be available for download from the following address: http://challenge2017.isic-archive.com/

Part 1: Lesion Segmentation

Task Participants were asked to submit automated predictions of lesion segmentations from dermoscopic images in the form of binary masks **(Figure 1)**. Lesion segmentation training data included the original image, paired with the expert manual tracing of the lesion boundaries in the form of a binary mask, pre-annotated by 2 of the experts participating (K.L. and M.M.) where pixel values of 255 are considered inside the area of the lesion, and pixel values of 0 are outside

Figure 1.

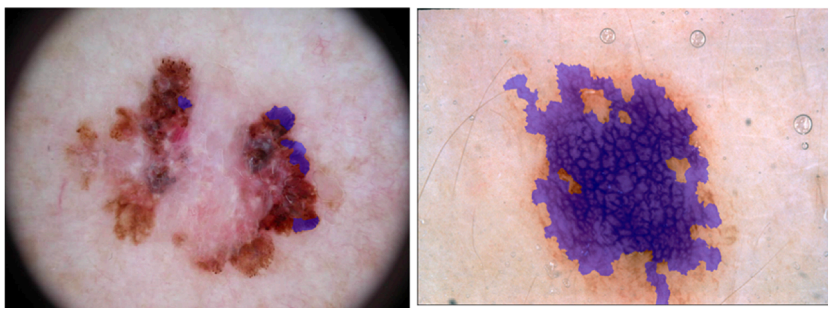
Example lesion segmentation. Left: original dermoscopic image. Right: binary segmentation mask.



Part 2: Dermoscopic Feature Classification

Participants were asked to automatically detect the following four clinically defined dermoscopic features: “network,” “negative network,” “streaks,” and “milia-like cysts,”.[39-41] Pattern detection involved both localization and classification **(Figure 2)**. To reduce the variability and dimensionality of spatial feature annotations, the lesion images were subdivided into superpixels using the SLIC algorithm.[42-44] Lesion dermoscopic feature data included the original lesion image and a corresponding set of superpixel masks, paired with superpixel-wise expert annotations for the presence or absence of the dermoscopic features (ground truth was set by one of the participating experts, K.L.). Validation and test sets included images and superpixels without annotation.

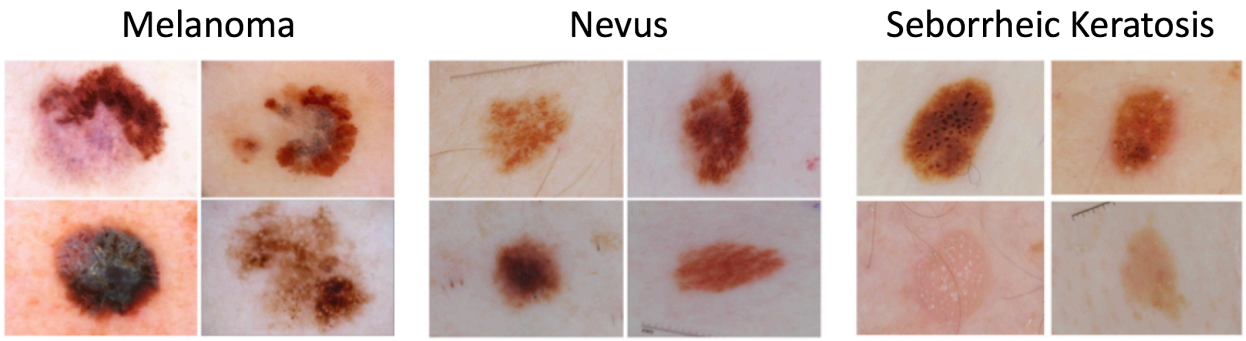
Figure 2. Images from “Part 2: Dermoscopic Feature Classification”. Ground truth labels highlighted in purple. Left: Streaks. Right: Pigment Network.



Part 3: Disease Classification Task

Participants were asked to classify images as belonging to one of 3 categories **(Figure 3)**, including "melanoma" (374 training, 30 validation, 117 test), "seborrheic keratosis" (254, 42, and 90), and "benign nevi" (1372, 78, 393), with classification scores normalized between 0.0 to 1.0 for each category (and 0.5 as binary decision threshold). Lesion classification data included the original image paired with the gold standard diagnosis, as well as approximate age (5-year intervals) and gender when available.

Figure 3. Example images from "Part 3: Disease Classification." Ground truth labels written above



EVALUATION METRICS

Details of evaluation metrics have been previously described.[27,38] For classification decisions, any confidence above 0.5 was considered positive for a category. For segmentation tasks, pixel values above 128 were considered positive, and pixel values below were considered negative. For evaluation of classification decisions, the area under curve (AUC) measurement from the receiver operating characteristic (ROC) curve was computed.[27] Additionally, for classification of melanoma, specificity was measured on the operating curve where sensitivity was equal to 82%, 89%, and 95%, corresponding to dermatologist classification and management performance levels, and theoretically desired sensitivity

levels, respectively.[38] Segmentation submissions were compared using the Jaccard Index, Dice coefficient, and pixel-wise accuracy.[27] Participant ranking used Jaccard.

Results, ISIC Grand Challenge 2017

The 2017 challenge saw 593 registrations, 81 pre-submissions, and 46 finalized submissions (including a 4-page arXiv paper with each). The associated workshop at ISBI 2017 saw approximately 50 attendees. To date, this has been the largest standardized and comparative study in this field, accounting for the size of the dataset, the number of algorithms evaluated, and the number of participants. In the following, the results for each challenge part are investigated.

Part 1: Lesion Segmentation

21 sets of prediction scores on the final test set were submitted for the segmentation task, and 39 were submitted to the validation set. The top ranked participant achieved an average Jaccard Index of 0.765, accuracy of 93.4%, and Dice coefficient of 0.849, using a variation of a fully convolutional network ensemble (a deep learning approach).[45] Example segmentations are shown in **Figure 4**, and a histogram of individual image Jaccard Index measurements is shown in **Figure 5**. Subjectively assessing the quality of the segmentations, one can observe that segmentations of Jaccard Index 0.8 or above tend to appear visually "correct." This observation is consistent with prior reports that measured an inter-observer agreement of 0.786 on a subset of 100 images from the ISIC 2016 Challenge. When Jaccard falls to 0.7 or below, the "correctness" of the segmentation can be debated. 156 out of 600 images (26%) fell at or below a Jaccard of 0.7. 91 images (15.2%) fell at or below Jaccard of 0.6. This suggests a failure rate of 15% to 26%, which is higher than the pixel-wise failure rate of 6.6%.

Figure 4.

Part 1 example segmentations from top ranked participant submission. Top Row: Original images. Middle Row: Ground truth segmentations. Bottom Row: Participant predictions. ISIC identifiers and Jaccard Index values are listed at each column head.

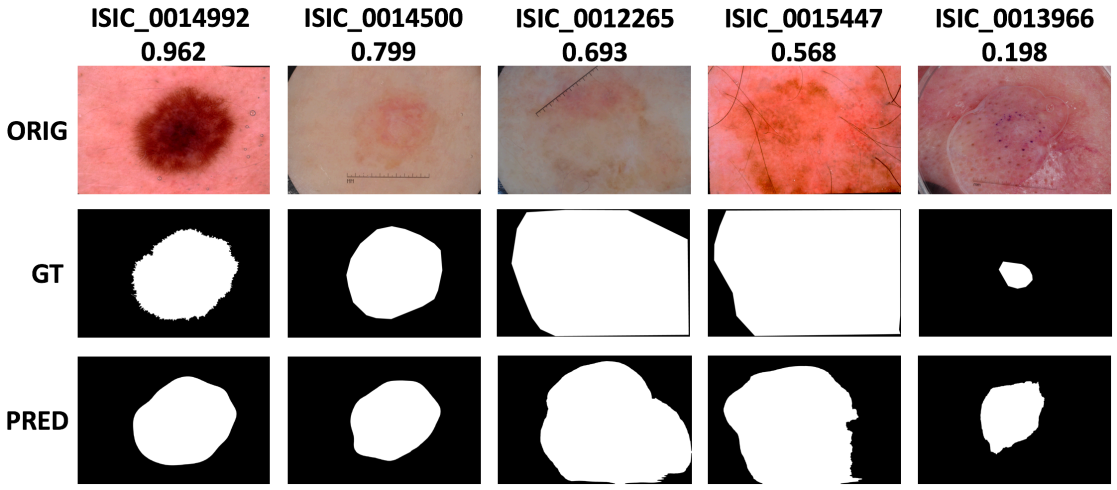
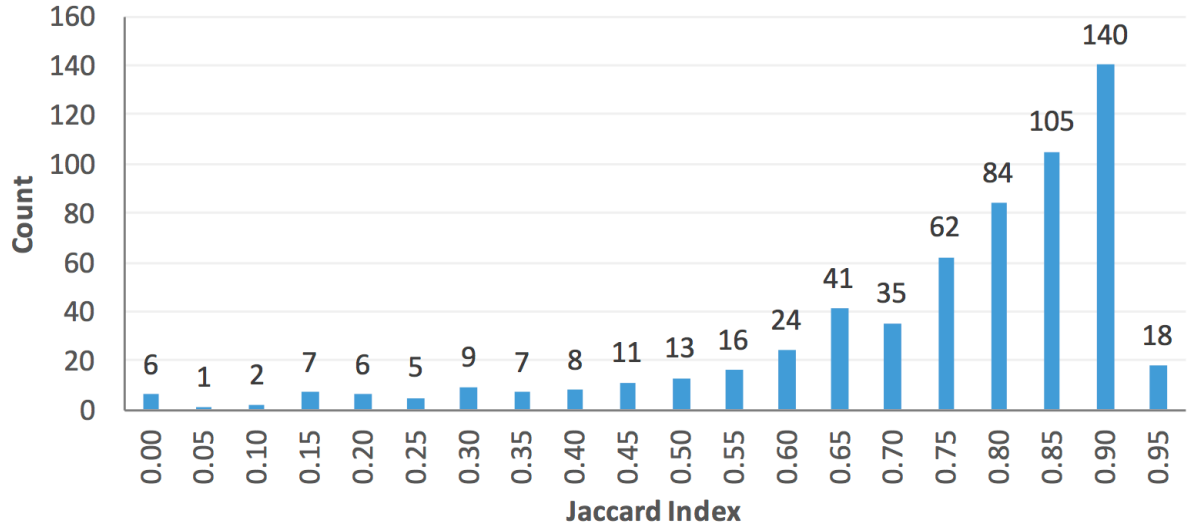


Figure 5.

Histogram of Jaccard Index values for individual images from top segmentation task participant submission.



Part 2: Dermoscopic Feature Classification Task

For the second year in a row, dermoscopic feature classification has received far less participation than other tasks. Only 3 submissions on the test set were received from 2 parties.[46,47] Whether this is due to the technical framing of the task (how well it maps to existing frameworks), or the perceived importance of the task, is a matter of current investigation. Regardless, performance levels of those submissions that were received demonstrated that localization of dermoscopic features is a tractable task for computer vision approaches. Top performance levels are shown in Table 1. AUC was above 0.75 ubiquitously, with an average close to 0.9

Table 1.

Part 2: Dermoscopic Feature Classification AUC Measurements. AVG = Average across all categories.

Method / Rank	AVG	Network	Negative Network	Streaks	Milia-Like Cyst
1[29]	**0.895**	**0.945**	**0.869**	**0.960**	0.807
2[30]	0.833	0.835	0.762	0.896	**0.838**
3[30]	0.832	0.828	0.762	0.900	0.837

Part 3: Disease Classification Task

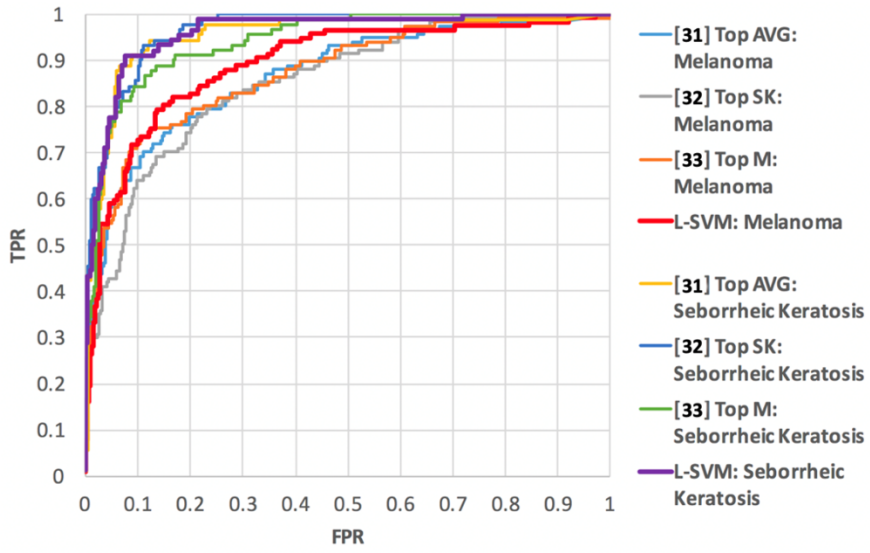
The disease classification task received 23 final test set submissions, and 39 validation set submissions. Performance characteristics of the average (AVG) classification winner,[48] seborrheic keratosis (SK) classification winner,[49] and melanoma (M) classification winner,[50] respectively, are shown in as well as 3 fusion strategies: score averaging (AVGSC), linear SVM (L-SVM), and non-linear SVM (NL-SVM) using a histogram intersection kernel.[38] Fusion strategies utilize all submissions on the final test set and are carried out via 3-fold cross-validation. SVM input feature vectors included all disease category predictions. Both SVM methods used probabilistic SVM score normalization, producing an output

confidence between 0.0 and 1.0 (with 0.5 as binary threshold), correlating with the probability of disease on a balanced dataset.[38] ROC curves for the 3 submissions and the best fusion strategy (Linear SVM) are shown in **Fig. 6**.

The 5 major trends observed involve the following: 1) All top submissions implemented various ensembles of deep learning networks. All used additional data sources to train, either from ISIC, in-house annotations, or external sources.[48-50] 2) Classification of seborrheic keratosis appears to be an easier task in this dataset, compared to melanoma classification. This may reflect aspects of the disease, or bias in the dataset.[49] The best performance came from the team that added additional weakly labelled pattern annotations to their training data. 3) The top average performer was not the best in any single classification category. 4) The most complex fusion approach (NL-SVM) led to a decrease in performance, whereas simpler methods led to overall improvements in performance, consistent with previous findings.[38] This challenge is the second benchmark to demonstrate that a collaborative among all participants outperforms any single method alone. 5) Not all thresholds balanced sensitivity and specificity. Probabilistic score normalization in fusions is effective at balancing sensitivity and specificity.[27,38]

Figure 6.

ROC curves for top 3 submissions to “Part 3: Disease Classification”, as well as linear SVM fusion.



Discussion, ISIC Grand Challenge 2017

The International Skin Imaging Collaboration (ISIC) archive was used to host the second public challenge on Skin Lesion Analysis Toward Melanoma Detection at the International Symposium on Biomedical Imaging (ISBI) 2017. The challenge was divided into 3 tasks: segmentation, feature detection (4 classes), and disease classification (3 classes). 2000 images were available for training, 150 for validation, and 600 for testing. The challenge involved 593 registrations, 81 pre-submissions, and 46 finalized submissions, making it the largest standardized and comparative study in this field.

AI algorithms show the potential to provide reliable lesion segmentations, detect specific dermoscopic features and more importantly, accurately classify skin lesions, including melanoma. Following we will discuss the reader study which we organized in order to compare the results of the AI algorithms from ISIC Grand Challenge 2017 with both expert dermatologists and dermatology residents. This study explored the potential role that AI algorithms could have in clinical practice, aiding in early skin cancer diagnosis. [30]

ISIC Grand Challenge 2017, Reader Study

Introduction, ISIC Grand Challenge 2017, Reader Study

In this study we compared the diagnostic accuracy of the best performing AI algorithm from ISIC Grand Challenge 2017[28] with that of expert dermatologists in diagnosing and treating skin cancer, as well as dermatology residents. We also examined how the potential use of this algorithm in cases where physicians' confidence in diagnosis was low would affect clinical outcomes.[30]

Materials and Methods, ISIC Grand Challenge 2017, Reader Study

Institutional review board approval was obtained at Memorial Sloan Kettering and the study was conducted in accordance with the Helsinki Declaration. Details of the challenge tasks, evaluation criteria, timeline, and participation were described above.[28] We selected 2750 high-quality dermoscopy images from the ISIC archive: 521 (19%) were melanomas, 1843 (67%) melanocytic nevi, and 386 (14%) SKs. Images were randomly allocated to training (n = 2000), validation (n = 150), and test (n = 600) data sets. Twenty-three algorithms were submitted to the melanoma classification challenge, and all used neural networks and deep learning, a form of machine learning that uses multiple processing layers to automatically identify increasingly abstract concepts present in data.[51]

Algorithms were ranked by area under the receiver operating characteristic (ROC) curve, and we chose the top-ranked algorithm for analyses.[28] A ROC curve is a graphical plot created by plotting sensitivity against the false positive rate (i.e., 1-specificity) at various threshold settings. The area under the ROC curve is therefore a global measure of the ability of a test to classify whether a specific condition is present or not present; an area under the ROC curve of 0.5 represents a test with no discriminating

ability (i.e., no better than chance alone), and an area under the ROC curve of 1.0 represents a test with perfect classification. A ROC curve can be used to determine an appropriate test cut-off, but the selection of a test threshold depends on the purpose of the test and the trade-off between sensitivity and specificity in the intended clinical scenario.[52]

A reader study was performed with 150 images (50 melanomas [15 invasive, 20 in situ, 15 not otherwise specified], 50 nevi, and 50 SKs) randomly selected from the test set. The median (range) Breslow depth for the invasive melanomas was 0.3 (0.15-3.3) mm. Eight dermatologists who specialize in skin cancer diagnosis and management and 10 dermatology residents agreed to participate in the study; after beginning evaluations, 1 resident did not complete the study and was removed. The dermatologists' mean number of years of clinical experience post-residency was 14 (range 4-32), and they had used dermoscopy for a mean of 14.5 (range 7- 28) years. The dermatologists originated from 4 countries (United States [n = 4], Spain [n = 2], Israel [n = 1], and Colombia [n = 1]), and all the dermatology residents were from the United States. Readers classified the lesions as melanoma, nevus, or SK; indicated a management decision (biopsy or observation); and reported diagnostic confidence on a Likert scale from 0 (extremely unconfident) to 6 (extremely confident). There were 1200 total image evaluations performed by dermatologists and 1350 by residents. Readers were blinded to diagnosis, clinical images, and metadata. There were no time restrictions and participants could complete evaluations over multiple sittings. For comparisons with human readers, algorithm performance metrics were calculated on the same 150 lesions from the reader study.

Descriptive statistics were used to explore the distributions of reader and algorithm results by lesion diagnostic classification and reader confidence. For readers, summary measures of diagnostic accuracy were estimated for lesion classification and management. Two sample tests for proportions were used

to assess differences in diagnostic accuracy measures between sample subgroups. Where applicable, variance estimates were inflated to address clustering of responses within readers. Algorithm diagnostic accuracy was assessed for lesion classification. ROC curves were calculated for algorithms, reader, and reader subgroups. To compare the ROC area between algorithms and human readers, we used a nonparametric approach.[53]

Reader results were imputed with algorithm responses when reader confidence in classification of the lesion was low (confidence classification 0-3). These imputations were accomplished by dichotomizing the algorithm with a predetermined sensitivity threshold of 90%. After imputation, diagnostic accuracy measures were recalculated. The alpha level for analyses was 5%, and tests were 2-sided. Analyses were performed using Stata version 14.2 (Stata Corporation, College Station, TX).

Results, ISIC Challenge 2017, Reader Study

The overall sensitivity, specificity, and ROC area of dermatologists for melanoma classification were 76.0% (95% confidence interval [CI] 71.5%-80.1%), 72.6% (95% CI 69.4%-75.7%), and 0.74 (95% CI 0.72- 0.77), respectively. The overall sensitivity, specificity, and ROC area of residents for melanoma classification were 56.0% (95% CI 51.3%-60.6%), 76.3% (95% CI 73.4%-79.1%), and 0.66 (95% CI 0.6-0.69), respectively. The ROC area of the top-ranked algorithm for melanoma classification was 0.8685 **(Figure 7)**, which was greater than the overall ROC areas for classification and management by dermatologists (0.74 and 0.70, respectively) and residents (0.66 and 0.67, respectively; P<.001 for all comparisons).

The specificities and sensitivities of dermatologists and residents for melanoma classification are provided in **Table 2**. At the dermatologists' overall sensitivity in classification of 76.0%, the computer

algorithm had a specificity of 85.0%, which was higher than the dermatologists' specificity of 72.6% (P = .001). At the dermatologists' overall sensitivity in management of 89.0%, the algorithm specificity was 61%, which was higher than the dermatologists' specificity of 51.1% (P = .02). To explore the feasibility of algorithms aiding lesion classification, we imputed algorithm classifications for reader evaluations with low confidence scores (range 0-3), constituting 51% of resident and 26.6% of dermatologist evaluations, respectively. After imputation, sensitivity of resident evaluations increased from 56.0% to 72.9%, with a decrease in specificity from 76.3% to 72.6%. The percentage of the 1350 evaluations correctly classified by residents increased from 69.4% (n = 939) to 72.6% (n = 981). The sensitivity of dermatologist classifications increased from 76.0% to 80.8%, and the specificity increased from 72.6% to 72.8%. The percentage of evaluations correctly classified by dermatologists increased from 73.8% (n = 885) to 75.4% (n = 905).

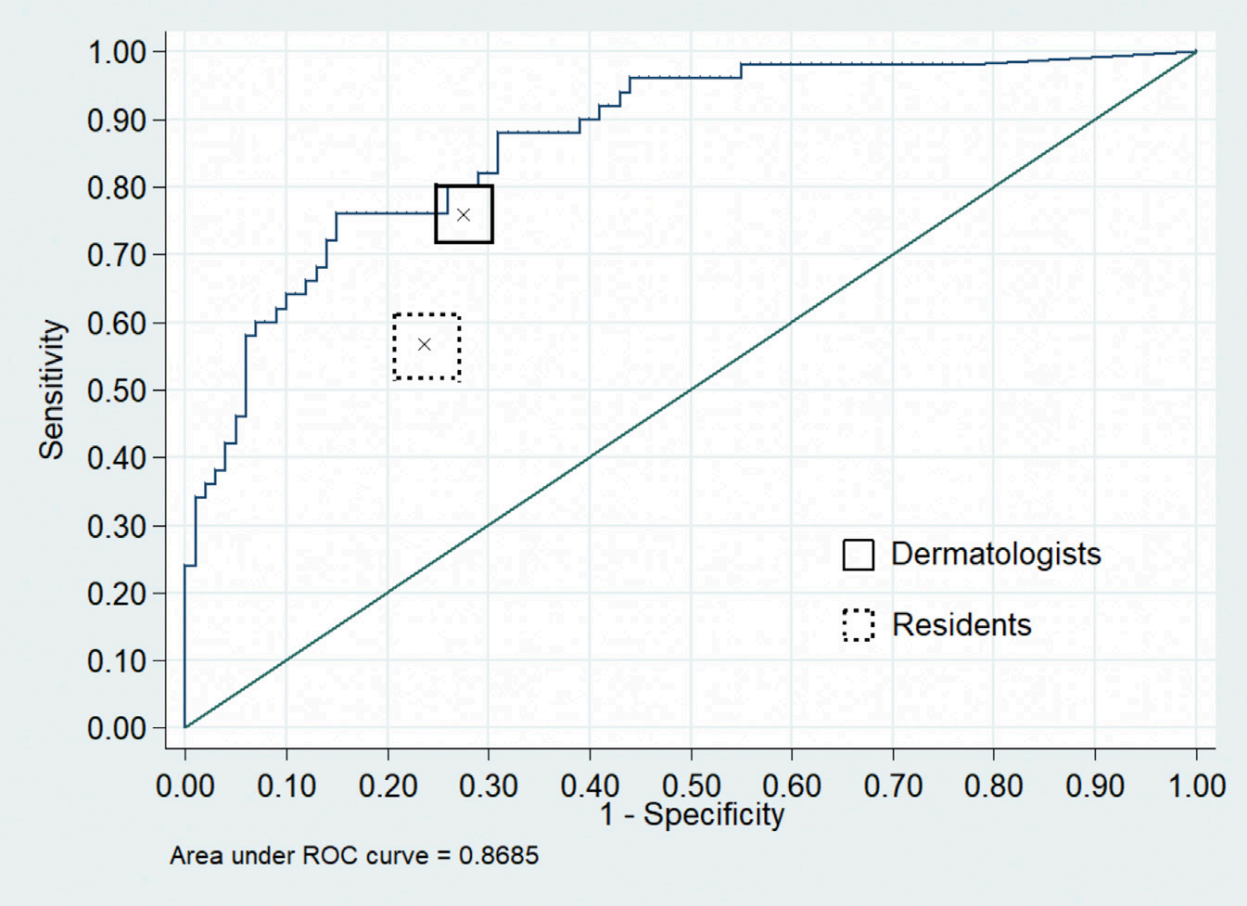
Table 2.

Measures of diagnostic accuracy for lesion classification by reader's reported confidence in the diagnosis

Group and confidence level	n (%)	Sensitivity (95% CI)	P_{trend}	Specificity (95% CI)	P_{trend}
Residents					
0	7 (0.5)	100.0 (2.5-100.0)	.54	16.7 (0.4-64.1)	<.001
1	160 (11.8)	57.6 (44.1-70.4)		61.4 (51.2-70.9)	
2	238 (17.6)	48.6 (36.9-60.6)		73.2 (65.7-79.8)	
3	289 (21.4)	53.6 (43.2-63.8)		70.3 (63.3-76.7)	
4	397 (29.4)	51.8 (43.1-60.4)		81.9 (76.7-86.4)	
5	204 (15.1)	63.1 (50.2-74.7)		87.8 (81.1-92.7)	
6	55 (4.1)	100.0 (80.5-100.0)		89.5 (75.2-97.1)	
Dermatologists					
0	26 (2.2)	75.0 (34.9-96.8)	.002	61.1 (35.7-82.7)	<.001
1	65 (5.4)	62.5 (40.6-81.2)		68.3 (51.9-81.9)	
2	97 (8.1)	52.0 (31.3-69.8)		58.3 (46.1-69.8)	
3	131 (10.9)	67.3 (52.9-79.7)		63.3 (51.7-73.9)	
4	301 (25.1)	74.3 (64.8-82.3)		64.8 (57.7-71.5)	
5	342 (28.5)	79.5 (70.8-86.5)		76.1 (70.0-81.4)	
6	238 (19.8)	91.9 (83.2-97.0)		90.2 (84.6-94.3)	

Figure 7.

Accuracy of the top-ranked algorithm, dermatologists, and residents for diagnosing melanoma using a 150-image data set. ROC curve (blue curve) demonstrates sensitivity and specificity of melanoma classification of the top-ranked algorithm from the 2017 International Skin Imaging Collaboration melanoma detection challenge. The x in the solid black box indicates the mean overall performance of 8 participating dermatologists, with the box indicating the 95% confidence intervals. The x in the dashed gray box indicates the mean overall performance of 9 participating residents, with the box indicating the 95% confidence intervals. ROC, Receiver operating characteristic.



Discussion, ISIC Grand Challenge 2017, Reader Study

These results and others demonstrate that deep neural networks can classify skin images of melanoma with high accuracy.[54-56] Compared with our 2016 challenge, we observed an increase in the relative diagnostic performance of the top-ranking algorithm compared with the same 8 dermatologist readers.[27,38] This finding suggests that the performance of algorithms is improving, possibly because of the availability of larger training data sets or advances in algorithm development.[10]

Although studies have demonstrated that algorithms can identify melanoma with diagnostic accuracy superior to dermatologists in reader studies, their clinical applicability remains uncertain. To examine the feasibility of an algorithm augmenting physician performance, we imputed algorithm classifications for lesions in which the physician reported low diagnostic confidence. We hypothesized that this would represent the most likely circumstance in which a physician would seek and use diagnostic help in a clinical setting. In this analysis, we found that the sensitivity and overall percentage of correct responses by readers increased by imputing algorithm classifications. Further studies are required to determine the optimal algorithm thresholds that would benefit physicians in a range of clinical settings and scenarios, and we are currently running similar studies in 'Andreas Sygros' Hospital for Cutaneous and Venereal Diseases.

There are notable limitations to our study. 1 Our test data set did not include the full spectrum of skin lesions, particularly banal lesions and less common presentations of melanoma, and the setting was artificial, considering that physicians did not have access to data typically used when evaluating patients (e.g., age, personal or family history of melanoma, lesion symptoms). We did not perform external validity analyses, which are important for demonstrating algorithm generalizability.[31]

Comparisons of skin cancer diagnostic accuracy of dermatologists with computer algorithms through reader studies should be cautiously interpreted. One device approved by the US Food and Drug Administration that used multispectral digital skin lesion analysis had been shown to have high melanoma sensitivity and to improve both the sensitivity and specificity of dermatologists after clinical and dermoscopic examination of suspicious skin lesions via reader studies; despite these apparent strengths, the device was discontinued in 2017.[57]

Unlike other studies examining the diagnostic accuracy of automated systems for skin cancer diagnosis,[54-56] our study used a data set that is publicly available for external use and future benchmarking. We further compared dermatologist accuracy to the top-ranked algorithm from a computer vision challenge, suggesting that the performance of the classifier is reflective of current state-of-the-art technology in machine learning.[28] Our annual ISIC melanoma detection challenges are the largest comparative studies of computerized skin cancer diagnosis to date and have attracted global participation. As our ISIC image archive expands, we anticipate hosting continuous public challenges with larger and more varied data sets with clinically relevant metadata. In the ISIC Grand Challenges that followed ours (2018-2020), a broader spectrum of skin cancer was included, i.e. Basal Cell Carcinoma, Squamous Cell Carcinoma and additional mimickers.[58] Additionally, during our thesis we attempted to reapproach task 2 for dermoscopic feature identification via our Expert Agreement Study on Dermoscopy of melanocytic lesions (EASY). Task 2 of ISIC Grand Challenges 2016 and 2017 received lower participation compared to the other 2 tasks, while the algorithms performance was also lower.[27,28] With EASY study we attempted a novel approach for the dermoscopic criteria used to differentiate nevi from melanomas.

Conclusion, ISIC Challenge 2017, Reader Study

In conclusion, the top-ranked algorithm from an international melanoma detection challenge exceeded the diagnostic accuracy of both dermatologists and residents in an artificial study setting. The sensitivity and overall percentage of correct evaluations by readers improved when imputing algorithm classifications for lesions in which the physician reported low diagnostic confidence, suggesting that augmented human classification is feasible. Future studies demonstrating clinical utility in a real-world setting are needed and we are already conducting similar studies in “Andreas Sygros’ Hospital for Cutaneous and Venereal Diseases.

Exploring the Pitfalls of Artificial Intelligence Algorithms for the Diagnosis of Skin Cancer

Introduction

During our thesis we explored both the diagnostic accuracy, as well as the generalizability of publicly available AI algorithms for the diagnosis of skin conditions, and specifically, skin cancer. Accordingly, we conducted two research projects to validate published, publicly available diagnostic AI algorithms.[56,59]

Automated Dermatological Diagnosis: Hype, or Reality? [31]

In 2018 Han et al. made a landmark contribution to the application of artificial intelligence (AI) in dermatologic diagnosis. Although previous studies have reported that computer algorithms can successfully diagnose skin cancer from medical images with human equivalency,[38,55,56,60] Han et al. have made their computer algorithm publicly available for external testing.[56]

To explore the generalizability of their computer classifier in a unique patient population, we selected 100 sequentially biopsied cutaneous melanomas (n= 37), basal cell carcinomas (n= 40), and squamous cell carcinomas (n= 23) with high-quality clinical images from the International Skin Imaging Collaboration Archive (https://isic-archive.com/#images, dataset name: 2018 JID Editorial Images). We uploaded them to the Han et al. web application (http://dx.medicalphoto.org/) on 7 March 2018. All lesions originated from Caucasian patients in the southern United States.

Our public dataset of clinical images composed of sequentially biopsied cutaneous melanomas (n=37), basal cell carcinomas (BCC, n=40), and squamous cell carcinomas (SCC) (14 in situ; 9 invasive). Fifteen melanomas were in situ. The mean (min-max) thickness of invasive melanomas was 0.6 mm (0.1-3.1 mm). All lesions originated from Caucasian patients in the United States and were located on the head/neck (n=26), trunk (n=21), upper extremities (n=30), and lower extremities (n=23). The mean age of the cohort was 66.8 years (min-max: 25-85) and 67% were male.

Overall, the Han et al. algorithm's first classification (i.e., highest probability output) matched the histopathological diagnosis in 29 of the 100 lesions (29%) **(Table 3)**. Considering any of the up to five classifications rendered per image by the web app algorithm (irrespective of the probability), the concordant or matching diagnosis was included for 58% of lesions (58 of 100). We found no difference in the probability output of the first classification among correctly and incorrectly diagnosed lesions (0.711 vs. 0.715, P= 0.94, paired t-test). Of the melanomas, melanoma was the first classification in 13.5% (5 of 37) lesions with a mean (range) probability score of 0.82 (0.42-0.99); considering any of the up to five classifications, melanoma was included in 35.1% (13 of 37) lesions with a mean (range) probability score of 0.43 (0.02-0.99). Among the eight melanomas with melanoma listed as the second or third classification, the mean (range) probability score was 0.18 (0.02-0.37). Our results suggest that the sensitivity of the Han et al. algorithm for skin cancer, and particularly melanoma, is considerably lower when applied to a different patient population, limiting its generalizability. Finally, we found that image manipulations as for brightness and/or contrast can affect the algorithm output. **(Figure 8)** [31]

Multiclass artificial intelligence in dermatology - progress but still room for improvement[32]

In 2020, the same authors expanded their diagnostic algorithm to improve their AI diagnostic algorithm to include 174 diagnoses, trained in 220.680 εικόνες and validating their results for 134 of these diagnoses with external datasets.[59]

We used the same public image dataset that we described above (https://isic-archive.com/#images, dataset name: 2018 JID Editorial Images) to test the new algorithm by Han et al. We uploaded images to https://modelderm.com/ on April 7-9, 2020. The web app allows users to select a 250px-by- 250px square region of interest (ROI). The ROI selection is performed by scaling down the image (between 0-1 times magnification of image) and panning within the image view window. This process changes the resolution of the images that are uploaded to the system. We conducted four sets of upload experiments. In the 'intended use' upload, we selected an ROI that (i) centered the lesion and (ii) was ~80% covered by the lesion. In the 'intended use off-center' upload, we adjusted the ROI to off-center the lesion. In the 'drag and drop' upload, we submitted images at the magnification selected by the web app. In the '1x magnification' upload, we centered the lesion and covered as much of the image as possible (i.e., as close to 1x magnification). For all uploads, we recorded the first three diagnoses and respective probabilities. We also recorded the sum 'malignancy probability' (sum of the probability outputs of melanoma, BCC, and SCC).

We calculated the top-1 and top-3 accuracies, and the average and standard deviation (SD) malignancy score. Algorithm sensitivity was calculated using a malignancy probability cut-off of 10% (per the website instructions). Unless specified, data refers to 'intended use' upload. Chi-squared and Fisher's

exact tests were used for comparisons of proportions in contingency tables. The overall top-1 and top-3 accuracies of the 174-disease algorithm showed improvement compared to the 12-disease algorithm (top-1: 39% vs. 32%, p=0.38; top-3: 63% vs. 60%, p=0.77) **(Table 4)**. The melanoma top-1 and top-3 accuracies increased more appreciably (top-1: 29.7% vs. 13.5%, p=0.16; top-3: 56.7% vs. 35.1%, p=0.10**) (Figure 9)**. The top-1 accuracy was highest for BCC (55%), followed by melanoma (29.7%) and SCC (17.3%) (p=0.01). **(Table 4)**

There was variability in overall top-1 accuracy (p=0.07) and melanoma top-1 accuracy (p=0.008) under the four upload conditions. In general, 'intended use' uploads performed similar to 'intended use off-center' and better than 'drag and drop' and '1x magnification'. However, 'drag and drop' performed best in overall and melanoma top-3 accuracy. The worst performance in top-1 and top-3 accuracies was consistently observed with '1x magnification'. Anatomic site was associated with top-1 accuracy for 'intended use off-center' uploads (p=0.006) but not with other experimental conditions. Combining all uploads (n=400), the top-1 accuracy was inferior for the lower extremities compared to all other sites combined (18% v. 36%, p<0.001). No associations between rulers, clothing, and hair were identified with accuracy.

Discussion

Although one may take these results to signify a poor performance of these computer classifiers, it is important to consider the inherent limitations and challenges associated with automated skin cancer diagnosis when interpreting these data. For the 2018 algorithm, although the authors collected more than 500,000 images, only approximately 20,000 (approximately 6,000 malignancies) were used in training their algorithm. Notably, these images were not collected in a standardized fashion and were

associated with limited clinical metadata. In both our studies we found limited generalizability for the AI algorithms in an external dataset, while top-1 and top-3 diagnostic accuracy for BCC, SCC and melanoma was lower than that reported by the authors. Classification performance was sensitive to perturbations in image magnification, implying that users must be careful to follow the intended use instructions, while performance varied by anatomic site. Additional variables reported to impact the performance of CNNs include adversarial 'noise', image rotation, brightness/contrast manipulation, rulers, and ink markings.[31,61-63] Additionally, a frequent critique to both artificial intelligence researchers, and ISIC Archive highlights that the training datasets for AI algorithms is mainly consisted of Caucasian patients, thus limiting the representation of possible variation in disease presentation.[64] Possible strategies to overcome these weaknesses include use of standards for acquiring images in dermatology, which could improve their quality, usability, and generalizability, or the development of more robust algorithms by training with a greater diversity of image types and settings.

Table 3.

Cross-classification frequencies of histopathological diagnosis and web app leading category (categorization with highest probability), along with the average probability associated with the rendered decision for 2018 algorithm.[56]

	Web app categorization										
Histopathologic diagnosis	**Melanoma**	**Basal cell carcinoma**	**Intraepithelial carcinoma**	**Squamous cell carcinoma**	**Hemangioma**	**Lentigo**	**Actinic keratosis**	**Nevus**	**Seborrheic keratosis**	**Wart**	**Total**
Melanoma	**5** **0.82**	2 0.96	6 0.70	3 0.59	1 0.96	12 0.82	1 0.94	5 0.65	0	2 0.82	37
Basal cell carcinoma	0	**19** **0.68**	10 0.78	1 0.64	3 0.83	1 0.81	1 0.98	2 0.74	1 0.37	2 0.57	40
Intraepithelial carcinoma	0	6 0.59	**4** **0.83**	1 0.51	0	1 0.52	1 0.46	0	0	1 0.85	14
Squamous cell carcinoma	1 0.17	1 0.46	4 0.87	**1** **0.30**	0	0	0	0	0	2 0.36	9
Total	6	28	24	6	4	14	3	7	1	7	100

The bold values represent the "correct" diagnosis.

Table 4.

Diagnostic accuracy depending on histopathological diagnosis for the two algorithms and each upload mode. [56,59]

	Han et al 2020 174-disease algorithm				**Han et al 2018 12-disease algorithm**
	Intended Use	*Intended Use, Off-Center*	*Drag and Drop*	*1x Magnification*	*Intended Use*
Overall Top-1 Accuracy, n=100	39%	37%	28%	24%	32%
Overall Top-3 Accuracy, n=100	63%	65%	65%	48%	60%
Overall Top-Any Accuracy, n=100	-	-	-	-	-
Melanoma Top-1 Accuracy, n=37	29.7%	29.7%	16.2%	2.7%	13.5%
Melanoma Top-3 Accuracy, n=37	56.7%	56.7%	59.4%	18.9%	35.1%

Melanoma Top-Any Accuracy, n=37	-	-	-	-	-
Malignancy Probability, mean (SD), n=100	41.1% (30.4)	40.9% (29)	41.3% (26.7)	32.3% (26.4)	-
Sensitivity (malignancy probability ≥10%)	77%	80%	83%	70%	-

Figure 8.

Modification of the web app classification output by image manipulation. (a, b) Basal cell carcinoma. The original image (a) was modified by zooming in (b). The two images gave different classifications: (a) lentigo (99.2% confidence); (b) intraepithelial carcinoma (96.9% confidence). **(c, d)** Melanoma. The original image (c) was modified by changing the contrast and brightness settings (d). The two images gave different classifications: (c) melanoma (99% confidence); (d) hemangioma (98% confidence). **(e, f)** Melanoma. The original image (e) was modified by flipping the image vertically (f). The two images gave different classifications: (e) lentigo (74% confidence), melanoma (12% confidence), nevus (5% confidence); (f) melanoma (40.5% confidence), lentigo (32% confidence), nevus (24% confidence). All images come from the International Skin Imaging Collaboration Archive (https://isicarchive.com/#images, dataset name: 2018 JID Editorial Images).

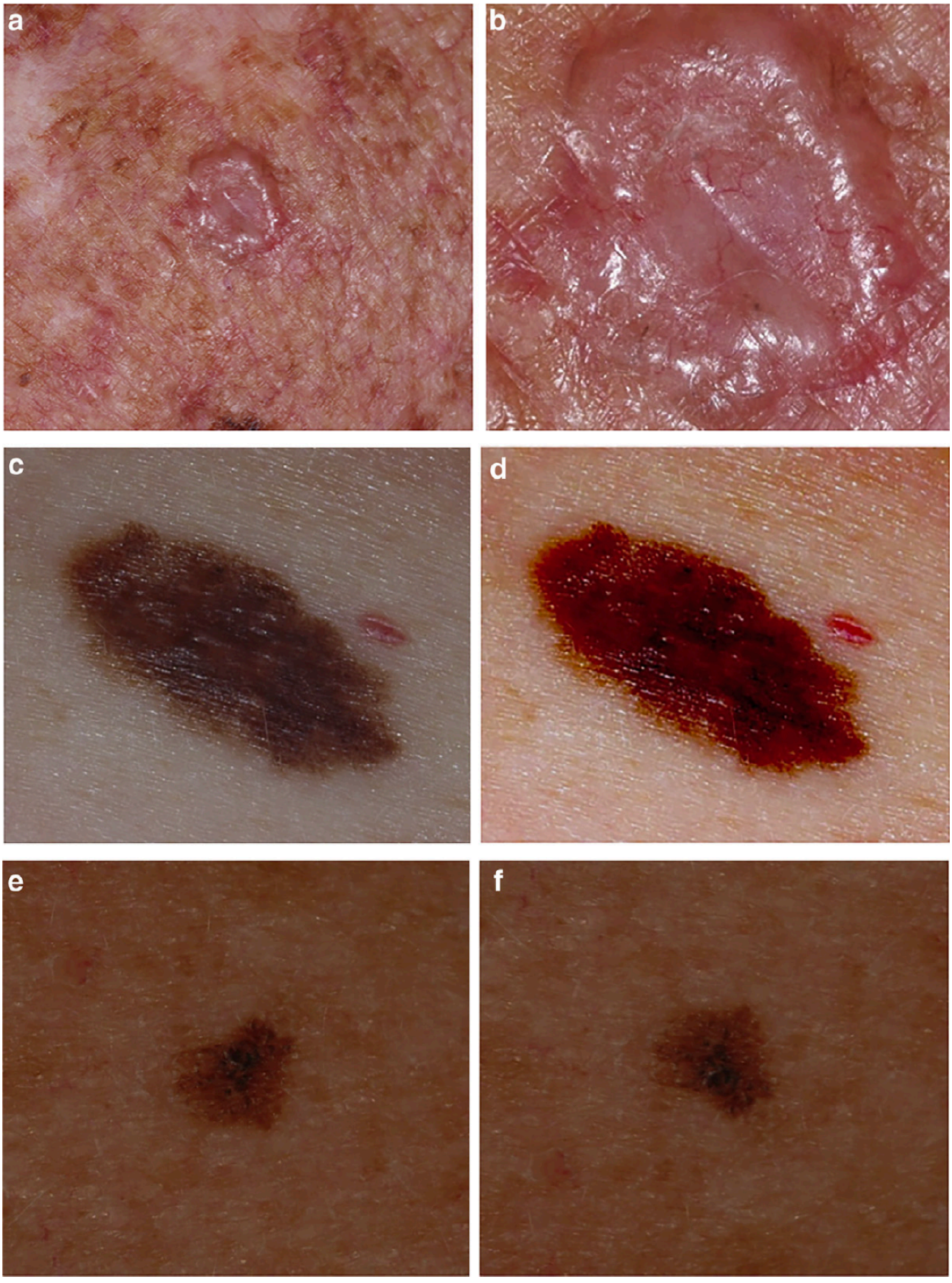
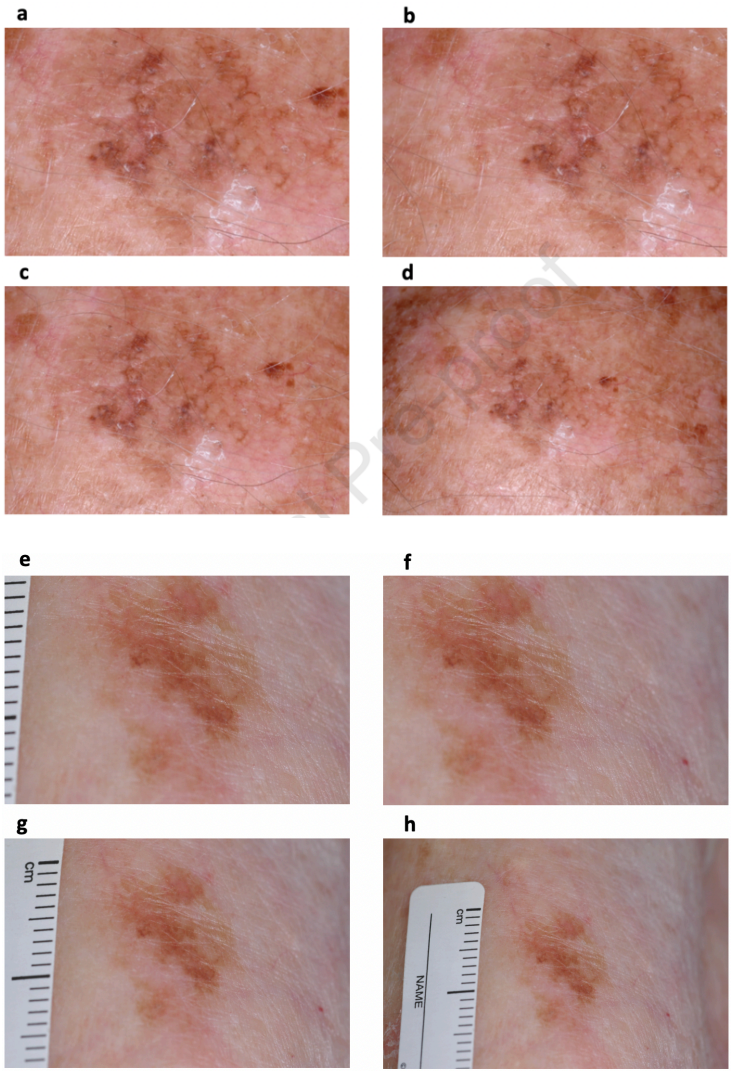

a
b
c
d
e
f

Figure 9.

Clinical images of the four upload experiments for two melanomas that were incorrectly diagnosed by the Han et al 2020 174-disease algorithm. Melanoma in situ (A –D). A. 'Intended use', leading diagnosis porokeratosis (probability 27%) **B. 'Intended use, off-centered',** leading diagnosis porokeratosis (probability 27%). **C. 'Drag and drop', l**eading diagnosis lentigo (probability 67%). **D. '1x magnification'**, leading diagnosis lentigo (probability 51%). **Melanoma 0.2 mm thickness (E-H). E**. 'Intended use', leading diagnosis vitiligo (probability 33%)**. F. 'Intended use, off centered',** leading diagnosis vitiligo (probability 27%). **G. 'Drag and drop',** leading diagnosis scar (probability 35%). **H. '1x magnification',** leading diagnosis non-specific lesion (probability 52%).



Interpretation of AI algorithms used in clinical practice for skin cancer diagnosis.

As discussed in the introduction, the mainstay for the treatment of skin cancer, and especially melanoma, is early skin cancer diagnosis and surgical excision.[19-21] There is an abundance of commercially available, non-invasive devices that can be used at bedside to aid in the diagnosis of skin cancer, including Raman spectroscopy, multispectral instrumentation and AI algorithms embedded to digital dermoscopy devices.[57,65-68]

Recently, MacLellan et al made an invaluable contribution through their investigator-initiated prospective diagnostic accuracy study comparing the classification performance of the FotoFinder (FotoFinder Systems GmbH, Germany), Melafind (STRATA Skin Sciences Inc, United States of America), and Verisante Aura (Verisante Inc, Canada) devices with a teledermoscopist and an onsite dermatologist on a consecutive series of 209 suspicious skin lesions referred to a specialty clinic.[69] These data provide important and much-needed insights into the accuracy of emerging diagnostic aids for cutaneous melanoma. It is speculative, however, to conclude that the addition of FotoFinder to the dermatologist's workflow would improve diagnostic accuracy, as the authors stated.

To test this hypothesis and to evaluate the addition of AI algorithms in clinical practice we initially attempted to validate their results and subsequently we suggested a different approach in the evaluation of the decision-making process of biopsying a lesion.[33] It would be hasty to conclude that the addition of AI algorithms would aid the clinicians perform better based on this heavily selected data set with a melanoma prevalence of 28.2%. It is uncertain if they would maintain accuracy, particularly

specificity, when applied to clinically less-concerning lesions. Additionally, an important point that has not been highlighted, is that it can be difficult to compare the clinical utility of competing diagnostic strategies by using traditional measures of classification accuracy (i.e., sensitivity, specificity, area under the curve). An alternative is to use net benefit and decision curve analysis.[70,71]

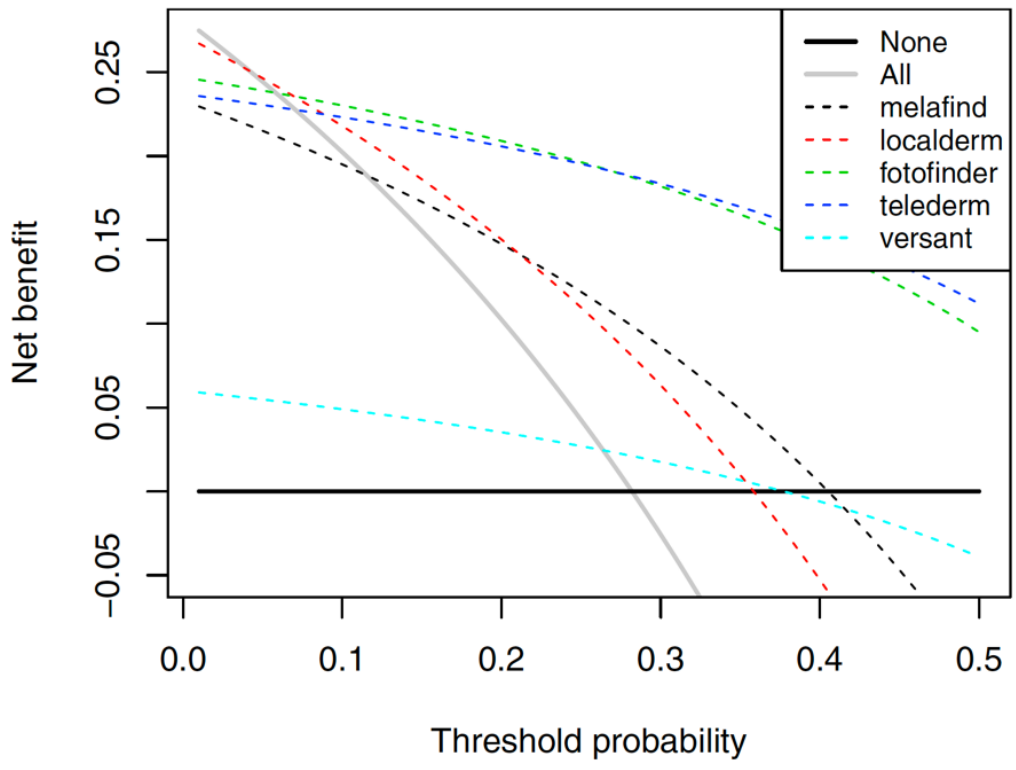
First, one must define the exchange rate between different endpoints, such as biopsying a melanoma versus biopsying a nevus. As a theoretical example, a dermatologist might consider the harm of 9 nevus biopsies equivalent to the benefit of 1 melanoma biopsy. This is an exchange rate of 1:9 and a probability threshold of 10%. Thus, the dermatologist would be willing to biopsy if the risk of melanoma was at least 10% but not if the risk was 9% or lower. The exchange rate and the sensitivity and specificity measures of a test are used to calculate the net benefit (Net Benefit= (true positives/n) – [(false positives/n) × (weighting factor)]. weighting factor = threshold probability/ (1 – threshold probability). Because individual dermatologists and patients likely have different exchange rates, it is beneficial to plot the net benefit across a range of clinically relevant probability thresholds, referred to as a decision curve. Here, we theoretically define that range as 5% to 10% for a melanoma biopsy. We used the diagnostic accuracy results reported by MacLellan et al and plotted decision curves for the 5 study strategies, along with decision curves for the competing strategies of biopsying all (sensitivity, 100%; specificity, 0%) and biopsying none (sensitivity, 0%; specificity, 100%)[69] **(Figure 10)**.

We found that the optimal approach depends on the probability threshold. From 5% to 7% (exchange rates of 19:1 to 13:1), the local dermatologist has the highest net benefit. From 8% to 10% (exchange rates of 12:1 to 9:1), FotoFinder has the highest net benefit. However, the absolute differences in net benefit between the strategies of biopsying all, biopsying based on the local dermatologist, and

biopsying based on FotoFinder are small at these thresholds. A more traditional decision analysis or cost-effectiveness analysis may shed further light on the optimal strategy.

Figure 10.

Decision curves for each studied diagnostic modality showing net benefit as a function of threshold probability. This analysis is based on the dataset as published by MacLellan et al, which had 209 lesions, 59 of which were melanoma (prevalence of 28.2%).[69] Among the treatment options, biopsying all has the highest net benefit for threshold probabilities of 1% to 4%, biopsying based on the decision of the local dermatologist has the highest net benefit for threshold probabilities of 5% to 7%, and biopsying based on the FotoFinder results has the highest net benefit for threshold probabilities of 8% to 10%. net benefit.



Description of novel dermoscopic features for skin cancer diagnosis and the Expert Agreement study on Dermoscopy of pigmented lesions.

As described in the introduction, dermoscopy is a widely available, non-invasive diagnostic technique which has been shown to aid in skin cancer diagnosis.[23,72] Over the past 30 years, important research efforts have described a wide variety of dermoscopic features that aid in the differential diagnosis of skin cancer from its benign mimickers.[73-85] Recognizing these features plays an important role during the clinical examination of a skin lesion suspicious for skin cancer, while it could also play an important role in training AI algorithms for skin cancer diagnosis.[30,38,39,86] During our thesis we described novel dermoscopic criteria for the diagnosis of melanoma and Basal Cell Carcinoma, as well as for Lichen Planus-Like Keratosis, a frequent benign mimicker.[34-37] Finally, for dermoscopic criteria to be useful in melanoma diagnosis, they should be reliable and reproducible; on this ground we performed the first Expert Agreement Study in Dermoscopy of melanocytic lesions, which explored expert agreement not only on the presence of a given feature within a lesion, but also its spatial localization. Our results could be used to guide the diagnostic algorithms used by clinicians for diagnosing skin cancer and to improve AI algorithms for skin cancer detection.

Dermoscopic Appearance of Amelanotic Volar Melanoma Compared with Volar Angioma[34]

Introduction

The survival rate of acral lentiginous melanoma is poorer than that of other cutaneous melanoma types, largely owing to difficulty in diagnosis and more advanced stages at presentation.[87] A single-center retrospective study of 53 acral melanomas found that at least 34% (n = 18) were initially misdiagnosed; of the misdiagnosed cases, 50% (n = 9) were amelanotic.[88] The amelanotic variant of acral volar melanoma is scarcely reported, and its clinical and dermoscopic characteristics are unknown. Özdemir et al described a dermoscopic feature on the periphery of pigmented acral lentiginous melanomas as a "vascularized parallel ridge pattern," defined as erythema and dotted vessels filling the ridges and sparing the furrows.[89] However, volar angiomas have similarly been observed to harbor a vascularized parallel ridge pattern on dermoscopy.[90,91] Herein, we describe the dermoscopic features of a subungual melanoma with an amelanotic volar component and compare these findings with the dermoscopic features of volar hemangiomas.

Findings

An adult man presented with a new diagnosis of melanoma of the left great toe. He reported a gradual worsening dystrophy of the left great toenail over the past 5 to 6 years. A biopsy of the nail bed confirmed a diagnosis of melanoma, at least 0.57 mm in Breslow depth. Clinical inspection of the left hallux revealed irregular brown pigmentation on the distal aspect of the dorsum of the hallux and a prominent pink tumor on the medial aspect. Dermoscopic examination of the pink plaque revealed chaotically distributed red dotted vessels on the ridges, sparing the furrows **(Figure 11A)**. A punch

biopsy of the red plaque was interpreted as melanoma in situ, with both superficial acrosyringial/ eccrine duct and deep eccrine gland involvement **(Figure 11C)**. The patient underwent amputation of the left great toe at the distal interphalangeal join, and the final Breslow depth was 4.6 mm. Findings of a sentinel lymph node biopsy of the left groin were negative.

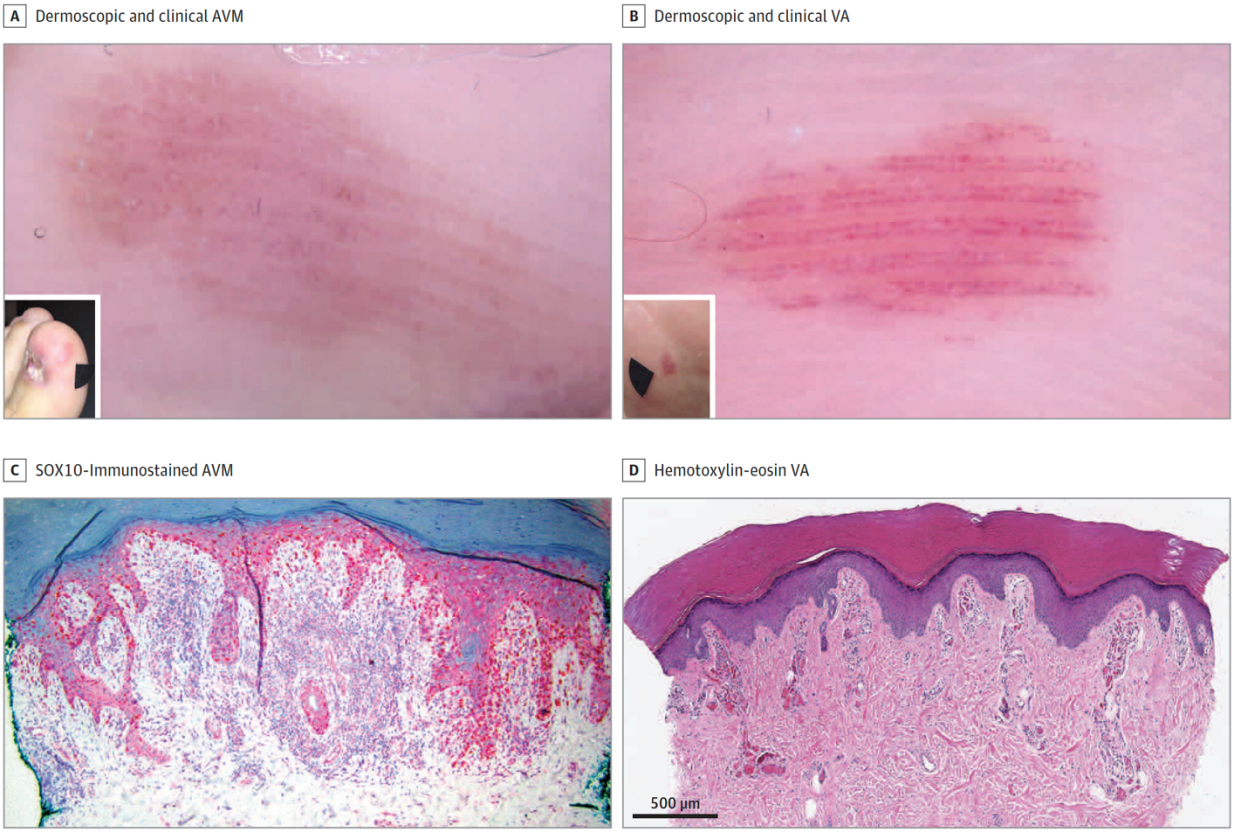
We also examined 3 patients with volar hemangiomas. They all presented dermoscopically with a parallel ridge pattern composed of red-to-purple dots. In contrast to the amelanotic volar melanoma, the dots were regularly aligned at the edges of the ridges, sparing the eccrine pores **(Figure 11B)**. Histopathological examination of a biopsy specimen of a volar hemangioma revealed capillary vascular proliferations that extended up into the dermal papillae, sparing the eccrine structures **(Figure 11D)**.

Discussion

The presence of a pigmented parallel ridge pattern has been shown in numerous studies to be associated with acral lentiginous melanoma and is particularly helpful for recognizing acral melanoma in situ.[92-94] Herein we describe a case of acral lentiginous melanoma with prominent amelanotic volar involvement that displayed a vascular parallel ridge pattern composed of chaotically distributed red dots. Additionally, we and others have found that volar hemangioma, which is an important differential diagnosis for amelanotic acral melanoma, has a distinct dermoscopic presentation of a parallel ridge vascular pattern composed of red or purple dots aligned at the edges of the ridges and sparing the eccrine pores (the linear, double-dotted ridge pattern).[90,91] Our observations require validation in future studies but may aid in distinguishing volar amelanotic melanoma from volar angioma.

Figure 11.

A and B Polarized light dermoscopy, original magnification ×10. **A**, Dermoscopic and clinical (inset) appearance of AVM with parallel ridge presentation of chaotically distributed red dots. **B**, Dermoscopic and clinical (inset) appearance of VA with parallel ridge presentation of red dots aligned at the edges of the ridges with sparing of the eccrine pores. **C**, Immunohistochemical analysis of AVM in situ using SOX10 immunostain, showing atypical melanocytes in the epidermis and involving the eccrine ducts, with a slight increase in the density of small vascular channels in the superficial dermis associated with inflammation (original magnification ×10). **D**, Histopathologic analysis of VA showing capillary vascular proliferations that extend up into the dermal papillae that surround the crista profunda intermedia and spare the eccrine structures (original magnification ×4).



Association of Multiple Aggregated Yellow-White Globules With Nonpigmented Basal Cell Carcinoma[37]

Introduction

Basal cell carcinoma (BCC) is the most common skin cancer worldwide.[95] BCC incidence is increasing, with more than 2 million cases of BCC diagnosed annually in the United States.[11,12] Dermoscopic features of BCC were first described by Menzies et al. in 2000 and included large blue-gray ovoid nests, multiple non aggregated blue-gray globules, ulceration, arborizing telangiectasia, spoke-wheel structures, and leaflike areas.[73] More recently, shiny white structures, specifically blotches and strands, were added as a new BCC dermoscopic criterion.[96] All these criteria have been confirmed to afford a high diagnostic accuracy for the diagnosis of BCC, with an overall sensitivity of 91.2% and a specificity of 95.0%, according to a recent meta-analysis.[97] However, the sensitivity and specificity of these dermoscopic features for nonpigmented BCC are lower (84.3% sensitivity and 73.2% specificity).[97] Given that most of the BCC criteria were defined for pigmented BCCs and that nonpigmented BCCs may be difficult to differentiate from other nonpigmented tumors, evaluation of new dermoscopic clues that may aid in the diagnosis of nonpigmented BCC and its differential diagnosis is needed.[98] Some BCCs display multiple aggregated yellow-white (MAY) globules. This dermoscopic feature differs from previously described milia-like cysts and from shiny white structures based on morphologic features and polarized vs nonpolarized light patterns of visualization. In 2014, Bellucci et al described the presence of yellow structures in 10% of BCCs; however, they were regarded mainly as milia-like cysts. Yellow orange structures were also described by Bañuls et al. but they were not characterized in extent.[99,100] To explore the prevalence and diagnostic accuracy of this newly characterized structure for the diagnosis of

BCC, we performed a retrospective assessment of clinical and dermoscopic images of lesions that could be included in the differential diagnosis of nonpigmented BCCs.

Materials and Methods

This retrospective case-control study was performed from July 1, 2017, to July 1, 2019. All images originated from a deidentified database of lesions consecutively seen in a single dermatology practice from January 1, 2009, to December 31, 2015, in Plantation, Florida. Given the relatively low frequency of amelanotic melanomas in the data set, we screened 2169 melanomas from the International Skin Imaging Collaboration (ISIC) archive, a publicly available image database. Twenty-two amelanotic melanomas were found and included. This study was approved by the institutional review board of Memorial Sloan Kettering Cancer Center. The images analyzed were only of close-up magnified dermoscopy images, and patient identifiers did not appear on the images. Therefore, the Memorial Sloan Kettering Cancer Center Institutional Review Board deemed that consent was not required for these magnified images of individual lesions. Three of us (C.N.-D., K.L, and A.R.) screened the clinical images of consecutive cases and included only those that had histopathologic results and were clinically nonpigmented. We excluded recurrent tumors, collision tumors, and cases with poor-quality images or lacking both polarized and nonpolarized modes. Cases included BCC of any subtype. Controls included lesions that are typically included in the differential diagnosis of nonpigmented BCC: squamous cell carcinoma (SCC), intradermal nevus (IDN), amelanotic melanoma, lichen planus–like keratosis, desmoplastic trichoepithelioma (DT), adnexal tumors (eg, fibrofolliculoma), and inflammatory diseases (eg, dermatitis and psoriasis). Patient age, sex, location of lesion, diagnosis, and predominant tumor subtype (if available) were recorded. Location was documented as specific anatomical locations and dichotomized into head and neck vs non–head and neck.

Clinical and Dermoscopic Images

Clinical and dermoscopic images were captured with a Nikon 1 camera (Nikon USA Inc) using Dermlite DL2 pro HR for polarized images and Dermlite fluid for nonpolarized images at 10 × magnification (3Gen). Images were taken in nonpolarized and polarized modes; the 2 modes were sequentially analyzed for each patient. Dermoscopic images were evaluated by 3 of us (C.N.-D., K.L., and A.R.), who were blinded to diagnosis, for consensus. A fourth reviewer (A.A.M.) resolved disagreement when consensus was not achieved. Dermoscopic images were analyzed for criteria based on the latest dermoscopic consensus by Kittler et al.[39] We specifically analyzed BCC-specific criteria, including blotches and strands.[96,97] The main dermoscopic structure analyzed consisted of MAY globules, defined by us as multiple, aggregated, white-to-yellowish globules arranged in clusters. This structure is visible in polarized and nonpolarized light, differentiating it from shiny white structures (blotches and strands) and from milia-like cysts, respectively **(Figure 12)**. We evaluated dermoscopic images for the presence or absence of MAY globules. To evaluate for interrater agreement in classifying MAY globules in a subset of 150 consecutive lesions, we evaluated images independently blinded to the final diagnosis. This analysis was performed by the same 3 of us (C.N.-D., K.L., and A.R.).

Reflectance Confocal Microscopy, Optical Coherence Tomography, and Histopathologic Correlation

In a subgroup of prospectively diagnosed cases displaying MAY globules seen at a single dermatologic practice in Hauppauge, New York, reflectance confocal microscopy (RCM) and optical coherence tomography (OCT) images were obtained before biopsy. Images were obtained with an arm-mounted RCM and/or a handheld RCM device (VivaScope 1500 and/or 3000; Caliber ID).For OCT images, we used a recently designed RCM/OCT probe.[101] RCM criteria used were those described in a recent systematic

review.[102] OCT criteria were those used in a recent study.[101] In these cases, a histopathologic correlation was performed using the precision biopsy technique, as previously described.[103] In brief, precision biopsy enables 1-to-1 correlation with en face histopathologic images were evaluated by a Mohs micrographic surgeon (C.-C.J.C.) on frozen sections. Formalin-fixed, paraffin-embedded samples were evaluated by a dermatopathologist (K.J.B.).

Main Outcomes

Our primary outcome was the distribution of the presence or absence of clustered yellow globules for the diagnosis of BCC compared with all other diagnoses combined. Secondary outcomes were the distribution of MAY globules by BCC subtype and the distribution of MAY globules by anatomical location of the BCC. To analyze for different histologic BCC subtypes, we divided them into high risk (morpheaform and infiltrative) and low risk (superficial and nodular) BCCs.

Statistical Analysis

Distribution of participant and lesion characteristics was evaluated by histologic diagnosis of the study lesions. Descriptive statistics and graphical methods were used to describe the study participants and the characteristics of the individual lesions. The relative proportion of dermoscopic characteristics along with exact binomial 95% CIs were estimated. Interrater agreement was estimated using multirater κ along with binomial interpolations of the 95% CI. The κ estimates were interpreted per the guidelines of Landis and Koch.18 To assess the association between dermoscopic criteria and BCC type, logistic regression was performed with the dichotomous dependent variable being BCC vs other diagnosis and the independent variables being dermoscopic criteria. All analyses were 2-sided with an α level of 5%. $P < .05$ was considered to be statistically significant. Data analyses were conducted from July 1 to September 31, 2019. All analyses were performed with Stata, version 14.0 (StataCorp).

Results

A database review of 2555 lesions revealed 643 potential study lesions in 621 patients with clinically nonpigmented tumors; 9 cases were excluded (no biopsies performed in 4, collision tumors in 4, and polarized images not available in 1). Final analysis included 656 lesions (634 lesions from the database plus 22 amelanotic melanomas from the ISIC archive) in 643 patients. The mean (SD) age of the total cohort was 63.1 (14.9) years, and 381 (58.1%) were male. Of all 656 lesions, 194 (29.6%) were located on the head and neck. A total of 278 lesions in 291 patients (44.4%; mean [SD] age, 61.9 [14.9] years; 190 [64.3%] male) were BCCs (cases), and 365 lesions (55.6%) in 365 patients (mean [SD] patient age, 63.9 [14.9] years; 191 [53.1%] male) corresponded to other diagnoses (controls). The mean (SD) tumor size in the whole cohort was 7.6 (4.9) mm, with a mean (SD) tumor size of 6.8 (4.8) mm in the BCC group and 8.2 (4.9) mm in the other diagnosis group ($P < .001$). Patient diagnoses and BCC subtypes are given in **Table 5**. Basal cell carcinomas were located in the head and neck in 124 patients (42.6%) and the trunk and extremities in 167 (57.4%). For other diagnoses, tumors were located in the head and neck in 70 patients (19.5%) and the trunk and extremities in 289 patients (79.2%) ($P < .001$). Basal cell carcinoma subtype distribution was nodular for 224 lesions (76.7%), superficial for 27 (9.2%), infiltrative for 24 (8.2%), morpheaform for 8 (2.7%), sclerosing for 2 (0.7%), keratotic for 2 (0.7%), Pinkus for 1 (0.3%), and basosquamous for 1 (0.3%). In 2 cases, subtyping data were not available.

Diagnostic Features

MAY globules were found in 64 of 656 cases (9.8%; 95% CI, 7.6%-12.3%). The structure was seen in 61 of 291 BCC cases (21.0%; 95%CI, 16.4%-26.1%) and in 3 of 365 cases with other diagnoses (0.8%; 95% CI, 0%-2.3%) ($P < .001$). The presence of MAY globules in BCCs was associated with a sensitivity of 20.9% (95%CI, 16.4%-26.1%), a specificity of 99.2% (95%CI, 97.6%-99.8%), a positive predictive value of 95.3% (95% CI, 86.6%-94.5%), and a negative predictive value of 61.0%(95% CI, 59.6%-62.4%). The odds ratio

for diagnosis of BCC was 32.0 (96%CI,9.9-103.2). The positive likelihood ratio was 25.4 (96% CI, 8.0-80.0), and the negative likelihood ratio was 0.8 (96% CI, 0.7-0.8).

Anatomical Location

When evaluating the presence of MAY globules restricted to head and neck lesions (n = 194), 51 patients (26.3%) presented with this structure. A total of 124 of 194 head and neck lesions (63.9%) corresponded to BCCs. Of the BCCs located on the head and neck, 48 of 124 (38.7%) manifested with MAY globules compared with 3 of 70 cases (4.2%) with other diagnosis (P < .001). The odds ratio for diagnosis of BCC when the structure was present was 14.1 (95%CI, 4.2-47.4) for head and neck lesions. The 3 lesions that presented with MAY globules other than BCC corresponded to DT (n = 2) and SCC (n = 1).

Subtype Analysis

MAY globules were observed in 18 of 32 high-risk BCCs (56.2%) (i.e., infiltrative and morpheaform) and 41 of 210 low-risk BCCs (19.5%) (P < .001) **(Figure 12, Figure 13A, and Figure 14A)**. MAY globules were 6.5 times more likely to be observed in higher risk than lower-risk BCCs (odds ratio, 6.5; 95%CI, 3.1-14.3). The structure was not seen in any of the 27 superficial BCCs **(Table 6)**.

Interrater Agreement

We observed almost perfect interrater agreement for the presence of MAY globules (κ = 0.89; 95% CI, 0.75-0.94). The interrater agreements were 0.94 (95% CI, 0.94-0.97) for arborizing vessels, 0.83 (95% CI, 0.81-0.90) for shiny white structure, 0.78 (95% CI, 0.63-0.84) for in-focus dots, and 0.73 (95% CI, 0.66-0.77) for ulceration.

RCM and OCT Features

Under RCM, all 4 examined cases had hyperreflective amorphous areas (**Figure 13B & C** and **Figure 14B)** in addition to classic BCC-specific features (tumor nests with palisading and clefting). OCT was available for 2 cases. Lesions with these dermoscopic structures had hyperreflective areas, producing an optical shadow **(Figure 13D** and **Figure 14C)**.

Histopathologic Correlation

In the 4 cases examined with a precision biopsy, aggregated yellow globules correlated with isolated, round areas of dystrophic calcification in or around tumor nodules and with the presence of calcified keratocysts **(Figure 13E & F** and **Figure 14D)**. In addition, 2 cases were analyzed in formalin-fixed, paraffin-embedded tissue, showing small calcific deposits in the superficial dermis in association with small keratocysts.

Discussion

In this retrospective case-control study of 656 lesions in 643 patients with nonpigmented tumors, we found that the presence of MAY globules was associated with the diagnosis of BCC. In addition, the presence of MAY globules was associated with high-risk histologic subtypes. Although tumors other than BCC may display milia-like cysts and/or shiny white structures, we observed that the presence of MAY globules was almost exclusively seen in BCCs. Although this dermoscopic feature was seen in only 21.0% of the nonpigmented BCCs evaluated, its frequency is within the range of other BCC-specific criteria, such as spoke-wheel structures, concentric structures, and leaf-like areas, with reported prevalence ranging from 8% to 20%.[73,104] When present, however, MAY globules were highly associated with BCC, specifically with high-risk histologic subtypes.

Dermoscopy has improved the diagnostic accuracy of pink lesions by providing visualization of structures and clues not visible to the naked eye, thereby improving the diagnostic accuracy for SCC, amelanotic melanoma, and BCC, among others.[98] However, hypomelanotic and amelanotic lesions are still challenging to diagnose with dermoscopy alone.[36,98] Other subgroups of lesions that are difficult to diagnose on clinical and dermoscopic grounds alone are the recently described BCCs that present as white papules on chronic sun-damaged skin.[118] The presence of MAY globules may be an important clue toward the diagnosis of BCC in this group of lesions by narrowing the differential diagnosis toward BCC. Another scenario in which the presence of MAY globules might emerge as an important dermoscopic feature is in the differentiation of BCC from IDN on the face, which can be challenging.[98] If MAY globules are seen, one can rule out an IDN, prompting a biopsy. Besides improving the diagnostic accuracy of BCC, dermoscopy has also improved the ability to identify BCC subtypes (i.e., superficial vs nodular), which in turn can assist in real-time, bedside, management decisions.[80,105] In the present study, a relevant finding was that MAY globules were not seen in superficial BCC subtypes. More importantly, if MAY globules were present, there was a higher odd of a high-risk BCC. A histologic high-risk BCC could be missed because of (1) sampling errors at the time of partial biopsy, (2) superficial shave biopsies, and (3) a deeper-seated or mixed-type BCC.[106] If a pathology report indicates a superficial BCC in a lesion displaying MAY globules, one might request additional sectioning, or it may be prudent to avoid topical therapies and resort to more aggressive forms of treatment, such as surgical excision. In addition, the findings of MAY globules in BCCs can also guide the physician in selecting the most appropriate biopsy type (e.g., tangential vs punch vs incisional). On precision biopsy, we found that 4 cases of MAY globules were correlated with dystrophic calcification on histopathologic analysis. Two additional cases evaluated in formalin- fixed, paraffin-embedded samples confirmed that MAY globules were correlated with calcifications. Previous studies have found that calcifications are seen in 11% to 21% of all BCCs on histopathologic analysis.[107,108] Furthermore, Slowdkoska et al. found that these calcifications were

more common in high-risk histologic subtypes (44% in infiltrative or morpheaform) than in low-risk histologic subtypes (22%).[107] The dermoscopic findings in our study mirror these histopathologic findings, with 56% of histologic high-risk subtypes revealing MAY globules vs 19% in low-risk subtypes (superficial and nodular). In addition, Slowdkoska et al. reported that BCC with calcifications may display keratocyst formation, which we observed in the proximity of calcified nodules as well.[107]

Limitations

Limitations of this study include its retrospective, single center design and the low number of cases of benign lesions known to present with yellowish structures as controls (eg, sebaceus hyperplasias, molluscum contagiosum, and pilomatrixomas). These lesions are usually multiple and/or easy to diagnose on clinical and dermoscopic grounds alone and are not typically part of the differential diagnosis of equivocal lesions and are infrequently biopsied. In addition, to enrich the non-BCC tumor data set, we obtained a subset of lesions from the ISIC archive; thus, the real-life frequency of MAY globules in non-BCC tumors could not be quantified with confidence. An additional limitation is that the population was mainly composed of white patients with intense sun exposure (residents of Florida). Findings might vary based on different sun-exposure backgrounds and skin types, and these results should be validated in other populations.

Conclusion

The findings suggest that MAY globules may have utility as a new BCC dermoscopic criterion to aid in diagnosis and help in the identification of high-risk BCC histologic subtypes. These structures may be associated with calcifications. Validation of our results is needed with data sets from other centers that include lesions not routinely biopsied, such as sebaceous hyperplasia and molluscum contagiosum.

Tables

Table 5: Diagnoses and BCC subtypes

Characteristic	No. (%) of total cases (N = 656)
Diagnosis	
BCC	291 (44.4)
SCC	114 (17.4)
Actinic keratosis	42 (6.4)
LPLK	37 (5.6)
Amelanotic or hypomelanotic melanoma	31 (4.7)
Seborrheic keratosis	29 (4.4)
Bowen disease	13 (2.0)
Keratoacanthoma	12 (1.8)
Intradermal nevus	12 (1.8)
Dermatitis	11 (1.7)
Sebaceous hyperplasia	6 (0.9)
Dermatofibroma	5 (0.8)
Desmoplastic trichoepithelioma	4 (0.7)
Psoriasis	4 (0.7)
Molluscum contagiosum	2 (0.3)
Other[a]	43 (6.6)
BCC subtype[b]	
Nodular	224 (76.9)
Superficial	27 (9.3)
Infiltrative	24 (8.2)
Morpheaform or sclerosing	10 (3.4)
Keratotic	2 (0.7)
Basosquamous	1 (0.3)
Pinkus	1 (0.3)

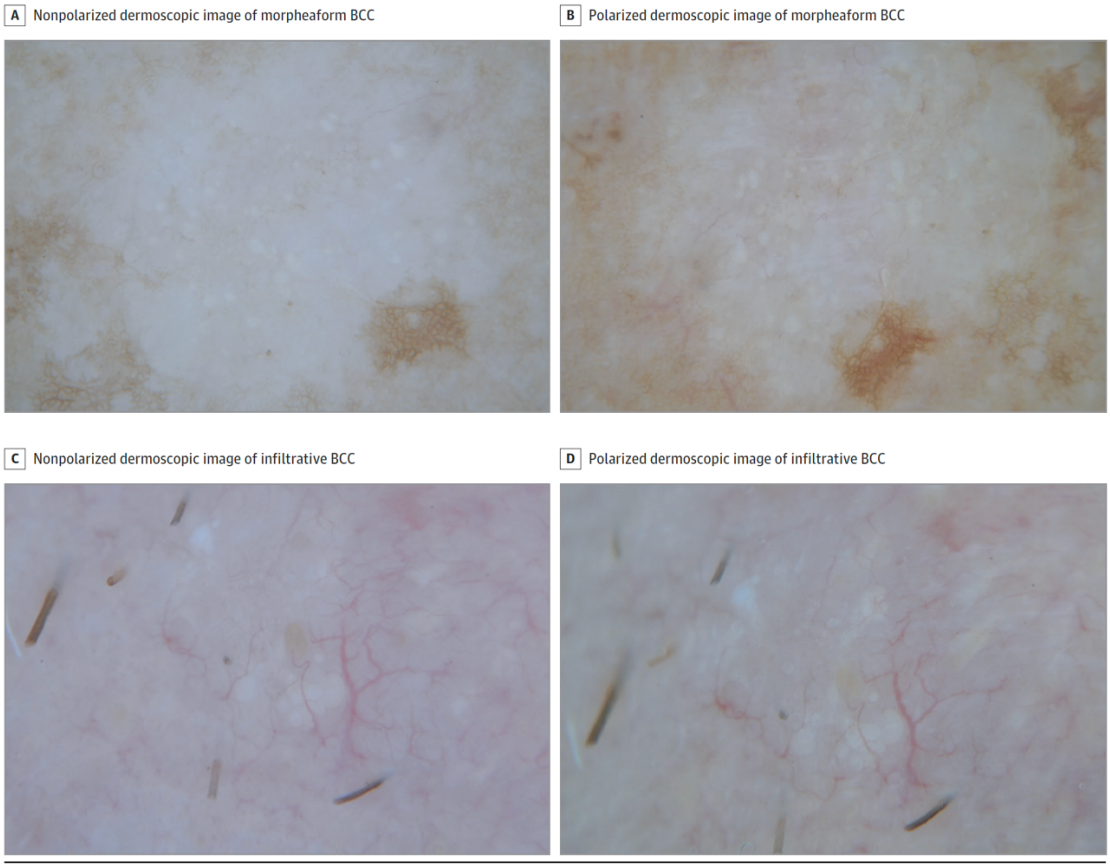
Table 6: Dermoscopic Characteristics

Characteristic	Cases, No. (%)			OR (95% CI)	κ (95% CI)
	BCC (n = 291)	Other diagnoses (n = 365)	Total (N = 656)		
Multiple aggregated yellow-white globules	61 (21.0)	3 (0.8)	64 (9.8)	32.0 (9.9 to 103.2)	0.895 (0.753 to 0.937)
Ulceration[a]	50 (17.2)	60 (16.4)	110 (16.8)	1.1 (0.7 to 1.6)	0.7261 (0.664 to 0.768)
Arborizing telangiectasia[a]	139 (47.8)	22 (6.0)	161 (24.5)	14.3 (8.7 to 23.2)	0.935 (0.935 to 0.966)
Ovoid nest[a]	16 (5.5)	0 (0)	16 (2.4)	NR	−0.008 (0.017 to 0.003)
Blue-gray globules[a]	24 (8.3)	0 (0)	24 (3.7)	NR	0.189 (0.126 to 0.494)
Blotches and strands[a]	214 (73.5)	70 (19.2)	284 (43.3)	11.7 (8.1 to 16.9)	0.827 (0.806 to 0.904)
Spoke-wheel structures[a]	4 (1.4)	0 (0)	4 (0.6)	NR	0.328 (0.189 to 0.497)
Leaflike areas[a]	36 (12.4)	1 (0.3)	37 (5.6)	51.4 (7.0 to 377.2)	0.747 (0.747 to 1.0)
Concentric structures[a]	15 (5.2)	0 (0)	15 (2.3)	NR	0.272 (−0.008 to 0.392)
Short-fine telangiectasia[a]	115 (39.5)	25 (6.9)	140 (21.3)	8.9 (5.6 to 14.2)	0.484 (0.398 to 0.592)
In-focus dots[a]	75 (25.8)	6 (1.6)	81 (12.4)	20.8 (8.9 to 48.5)	0.782 (0.629 to 0.835)
Multiple small erosions[a]	23 (7.9)	3 (0.8)	26 (4.0)	10.4 (3.1 to 34.8)	-
Serpentine vessels	1 (0.3)	16 (4.4)	17 (2.6)	0.1 (0 to 0.6)	−0.014 (−0.017 to −0.008)
Milialike cysts	15 (5.2)	16 (4.4)	31 (4.7)	1.2 (0.6 to 2.4)	0.601 (0.477 to 0.791)
Polymorphous vessels	11 (3.8)	46 (12.6)	57 (8.7)	0.3 (0.1 to 0.5)	0.669 (0.507 to 0.742)
Shiny white lines	3 (1.0)	28 (7.7)	31 (4.7)	0.1 (0 to 0.4)	0.851 (0.658 to 1.0)
Rosettes	17 (5.8)	51 (14.0)	68 (10.4)	0.4 (0.2 to 0.7)	0.741 (0.676 to 0.747)
Peppering	2 (0.7)	7 (1.9)	9 (1.4)	0.4 (0.1 to 1.7)	−0.005 (−0.005 to −0.003)
White circles	3 (1.0)	35 (9.6)	38 (5.8)	0.1 (0 to 0.3)	0.767 (0.719 to 0.82)
Scale	15 (5.2)	179 (49.0)	194 (29.6)	0.1 (0 to 0.1)	0.807 (0.782 to 0.824)
Glomerular vessels	11 (3.8)	115 (31.5)	126 (19.2)	0.1 (0 to 0.2)	0.650 (0.501 to 0.697)
Hairpin vessels	10 (3.4)	41 (11.2)	51 (7.8)	0.3 (0.1 to 0.6)	0.556 (0.332 to 0.63)
Orange color	5 (1.7)	50 (13.7)	55 (8.4)	0.1 (0 to 0.3)	0.647 (0.538 to 0.779)

Figures

Figure 12.

Morpheaform Basal Cell Carcinoma (BCC) and Infiltrative BCC Showing Multiple Aggregated Yellow-White Globules



A and C, Nonpolarized images. B and D, Polarized images. Note that the structure is seen in both modes (original magnification ×10).

Figure 13.

A. Dermoscopic appearance of MAY globules (yellow arrowheads). **B.** Reflectance confocal microscopy (RCM) panoramic view shows a well-defined tumor with hyperreflective amorphous areas (8 × 8 mm). **C**, RCM shows tumor nodules (blue arrowheads) and hyperreflective amorphous areas (yellow arrowheads) (750 × 750 µm). **D,** Optical coherence tomography shows hyperreflective structures with acoustic shadow (yellow arrowheads) and hyporeflective nodules (blue arrowheads). **E,** Histopathologic

analysis, en face view, shows tumor islands with palisading and clefting (blue arrowheads) and calcium deposits (yellow arrowheads) (hematoxylin-eosin, original magnification ×10). **F,** Histopathologic analysis, vertical view, shows tumor islands with palisading and clefting (blue arrows) and subepidermal calcium deposits (yellow arrowheads) (hematoxylin-eosin, original magnification ×10).

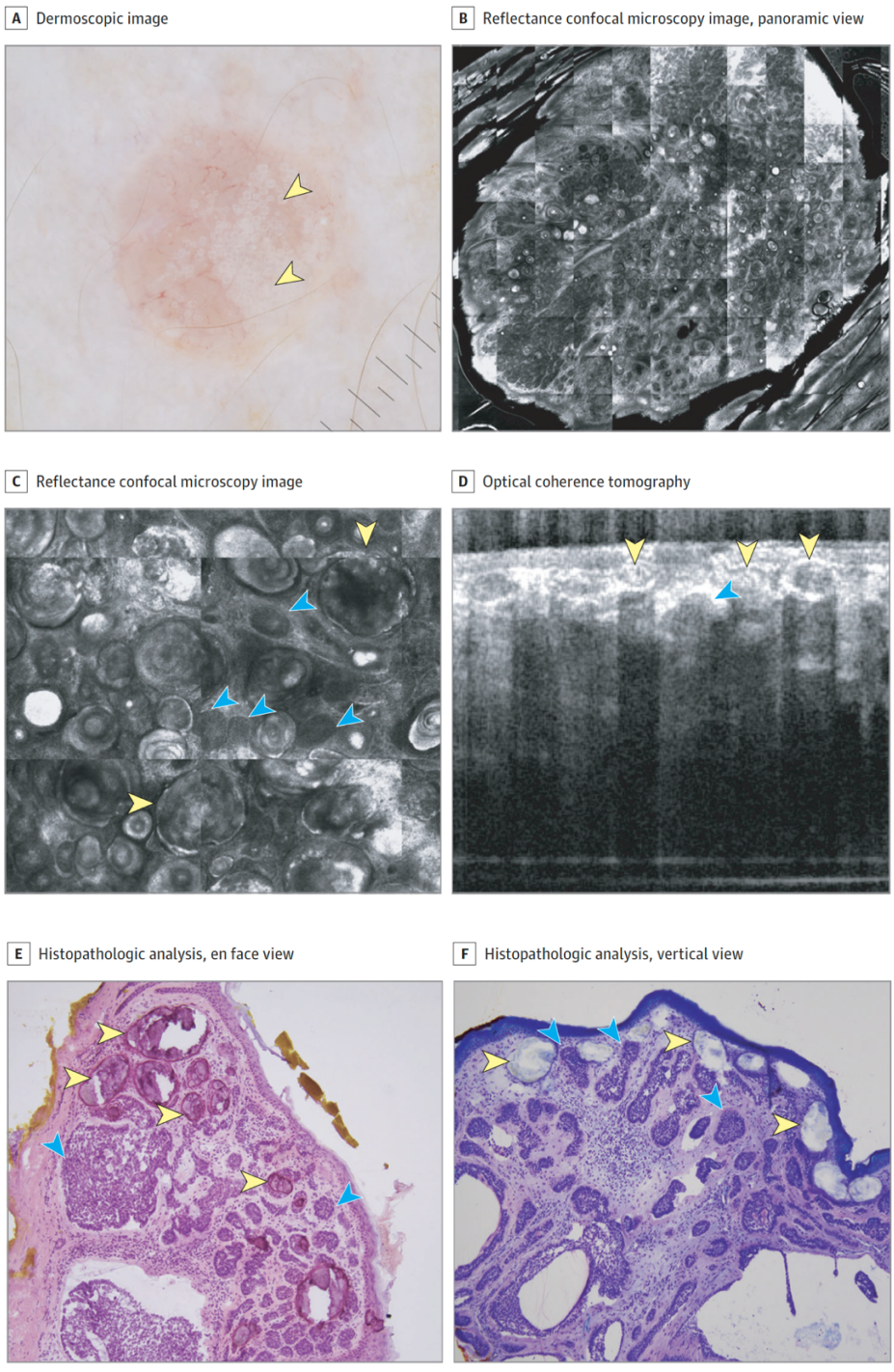
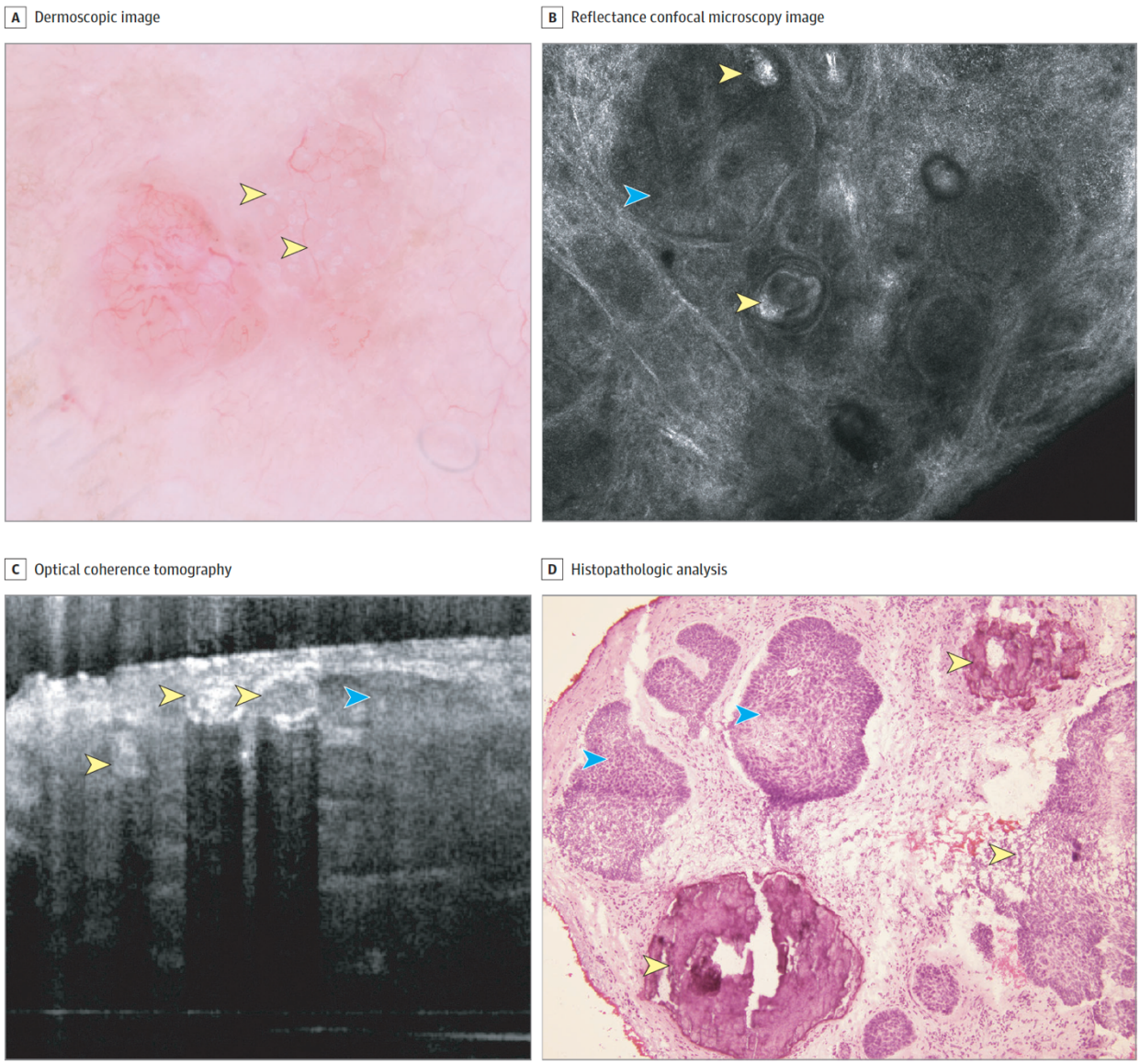


Figure 14.

A. Dermoscopic appearance showing MAY globules (yellow arrowheads). **B,** Reflectance confocal microscopy shows tumor nodules (blue arrowheads) and hyperreflective amorphous areas (yellow arrowheads). **C,** Optical coherence tomography shows hyperreflective structures with acoustic shadow (yellow arrowheads) and hyporeflective nodules (blue arrowheads). **D,** Histopathologic analysis shows tumor islands with palisading and clefting (blue arrowheads) and areas of calcium deposits (yellow arrowheads).



Expert Agreement Study on Dermoscopy of Melanocytic Lesions (EASY study)

Introduction

Dermoscopy is a non-invasive, diagnostic technique that aids in the early diagnosis of melanoma.[22,136,137] There is extensive literature highlighting the benefits of dermoscopy; however, while dermoscopy improves diagnostic accuracy, interobserver agreement for the presence of specific dermoscopic structures has been reported to be poor to moderate, even among experts.[75,109] Reasons for subpar agreement are poorly characterized but could include: (1) differences in the perception of feature definitions among experts, (2) potential overlap between features, (3) lack of standardization for dermoscopy terminology, (4) features that are intrinsically not reproducible, or (5) use of inappropriate statistical measures of agreement. Prior efforts in the field have attempted standardization of nomenclature, however these terms have not been formally tested for reproducibility and reliability.[39]

To achieve increased diagnostic accuracy and use among clinicians, agreement on dermoscopic structures and terminology standardization is necessary. To this objective, we designed the Expert Agreement Study on dermoscopY (EASY) of melanocytic lesions. In this study, we attempted to estimate the interobserver agreement of experts in dermoscopy on thirty-one melanocytic-specific dermoscopic features derived from the latest consensus agreement.[39] Towards this goal, dermoscopy experts were asked to submit 'exemplar' lesions displaying any of the previously-defined thirty-one melanocytic-specific structures.[39] These highly selective 'exemplar' images were used to construct datasets to investigate the agreement of experts on the presence or absence (lesion-level agreement) of dermoscopic features. In addition, a novel 'superpixel' (group of contextually similar pixels) annotation platform was created for readers to spatially annotate lesions by highlighting the superpixel containing

specific structures. While the lesion level annotation allows for the analysis of interobserver agreement for the presence or absence of specific structures, the superpixel annotation platform permits analysis of interobserver spatial overlap and/or agreement.

Methods

This cross-sectional, observational study was performed between September 1, 2017 and January 31, 2020. Invites were sent via email to 32 experts in dermoscopy. Experts were defined as clinicians with more than 10 years of experience in dermoscopy who have made significant contributions to research or teaching dermoscopy of pigmented lesions. We requested contributions of 1 to 3 exemplar-quality images for each of the thirty-one dermoscopic features.[39] **(Supplementary Table 1**: feature definitions and abbreviations) An exemplar-quality image was defined as one with excellent quality and in-focus depiction of a feature in which the experts had absolute confidence regarding its presence in the lesion. In total, 25 experts (81%) contributed 964 images of melanocytic lesions for the thirty-one features. Of the 32 experts initially invited to participate in the study, 21 (66%) completed the annotation phase. Five additional experts were subsequently invited to participate in image annotation, and all agreed to participate. The images were uploaded to the International Skin Imaging Collaboration Archive (ISIC Archive) and are available online.[10]

To determine the required number of readers and evaluations per reader, we used Monte-Carlo simulations of the intraclass correlation coefficient. We estimated that, with a confidence interval of 95% and an interclass correlation coefficient of 0.5 as the measure of agreement, 5 readers per study dataset would be efficient for evaluating agreement for the thirty-one dermoscopic features.

Feature Selection

All thirty-one melanocytic-specific dermoscopic features from the 2016 *International Dermoscopy Society* terminology consensus were selected.[39] This list included 14 criteria previously reported to be significantly associated with melanoma diagnosis.[41]

Dataset creation

Three experts (K.L., C.N-D. and A.A.M.) selected 310 of the contributed images via consensus; 10 images per exemplar feature were selected, based on image quality and presence of all assorted metadata. These images were used to generate 5 splits each containing 62 images. Each split contained 2 exemplar images for each of the thirty-one dermoscopic features. The 25 expert readers were randomly allocated into groups of 5 and assigned to one of the 5 datasets on which to perform the annotation process. Image annotations were not completed for one of the splits due to lack of participation (23 out of 25 experts completed annotations, 92%) and it was therefore excluded from further analysis.

Superpixel generation

Following image upload to the ISIC Archive (www.isic-archive.com), a superpixel map was automatically generated. The superpixel segmentation is a technique of grouping together the pixels of a given image into patches, according to their contextual similarity and spatial proximity. We applied SLIC (Simple Linear Iterative Clustering) algorithm on the pixel color values to create grouping distinctions based on contrast thresholds. **(Figure 1)**.[43,110]

Image annotations

We developed an annotation tool that allows for selection of features within images both on the lesion level (lesion-level annotations) and on the superpixel level (spatial annotations).[10] Lesion-level annotations allow for annotation of descriptive features in a binary fashion - present, or not present – without indicating their exact location within the lesion. In contrast, superpixel level annotation permits for spatial localization of features within the lesion. The annotation tool can be used to identify multiple overlapping features for any given lesion image. Annotators determined whether a feature was present, or absent in a given image and in a given superpixel, according to their perception. (Video demonstration of annotations can be found at https://youtu.be/jgJdCD3k3Es. Expert readers were blinded to the diagnosis and were invited to exhaustively annotate all thirty-one melanocytic-specific dermoscopic features in images. Readers were not instructed on the definitions of the features and annotated the features at their discretion. An example image displaying the superpixel division of an image and the assorted annotations performed by all five readers, along with their agreement is displayed in **Figure 15.**

Agreement Analysis

We performed three levels of analysis: (1) Agreement for the exemplar feature on the lesion level (2) Agreement for the exemplar features on the superpixel level, and (3) Agreement for the non-exemplar features on the superpixel level.

Lesion-level agreement

Agreement on the presence of exemplar feature:

Exemplars of each of the thirty-one features were allocated to each of the 4 splits (2 exemplar images per feature, per dataset). Data were combined into a single dataset for analysis. Percent agreement,

Fleiss kappa and Gwet's AC1 were estimated for each of the thirty-one dermoscopic features.[111] Both the Fleiss kappa and Gwet's AC1 were estimated due to the paradoxical performance of the Fleiss kappa at the extremes of the distribution of percent agreement. In order to further explore potential overlap among terms regarding similar features, these were combined into 7 'super-categories' of features (i.e. Dots, Globules/Clods, Lines, Network, Regression Structures, Structureless and Vessels), maintaining separate the structurally distinct features Shiny White Structures, Angulated Lines/Polygons, and Negative Network based on consensus among 4 investigators (K.L., C.N-D., M.A.M, and A.A.M.). Measures of agreement were estimated for the 'super-categories'. Kappa and Gwet's AC1 were interpreted as outlined by Landis and Koch: 0<0.4 (poor agreement), 0.4<0.75 (fair to good), and 0.75-1.0 (excellent agreement).[112]

Superpixel-level Agreement:

To assess inter-reader agreement on the same superpixel, as well as the confusability or overlap with other features, we computed both the percent agreement for each feature as well as the Dice coefficient for each possible cross-rater pair of selected superpixel sets. (i.e., for both the same feature as well as different features marked by two readers).[113,114]For the percent agreement we took the total number of superpixels annotated within an image for a specific feature as the denominator (100%), and subsequently we calculated the agreement for each feature on the annotated images. For Dice coefficient our calculations yielded a number between a minimum of 0.0 (0%), in cases of full disagreement (i.e., two mutually exclusive sets of superpixels), and a maximum of 1.0 (100%) in cases of complete agreement (i.e., exactly the same set of superpixels). To visually represent the full matrix of feature pair annotations for all study participants, we created a confusion matrix. Code in the python (version 3.7) language for these analyses is available at Github repository.[115]

Results

Lesion-level annotations

Twenty experts annotated 248 images (8 images per exemplar feature), in groups of 5 annotators for a total of 4507 feature markups. Single reader observations of a feature accounted for 22.4% of all observations, while agreement of all readers evaluating an image for a feature occurred in 65 images (26.2%). **Supplementary Table S2**

Agreement on the exemplar feature on the lesion level

Measures of agreement are presented in **Table 7**. Percent agreement varied by feature. Highest levels of percent agreement were observed for 'Peppering / Granularity' (92%); 'Shiny White Streaks' (90%); 'Typical Network' and 'Irregular Blotch' (86%); 'Negative Network' (84%); 'Irregular Globules' and 'Dotted Vessels' (82%) and 'Scar-like Depigmentation' and 'Blue-Whitish Veil' (80%). The remaining 22 features yielded lower levels of percent agreement **(Table 7).** Overall, Fleiss kappa showed poor agreement (<0.4) for all melanocytic specific features with the exception of 'Rim of brown globules' and 'Irregular Blotch'; 0.44 and 0.42, respectively. These results are due to the paradoxical nature of the statistic's performance at extremely low or high levels of feature prevalence. When using Gwet's AC, excellent agreement was observed for 'Irregular Globules' (0.78); 'Typical Network' (0.83), 'Peppering / Granularity' (0.91), 'Shiny White Streaks' (0.89), 'Negative Network' (0.81), 'Irregular Blotch' (0.82), 'Blue-Whitish Veil' (0.76), and 'Dotted Vessels' (0.77). The remaining structures showed poor to moderate agreement.

After collapsing the individual features into 'super-categories' based on structural similarities, higher levels of overall percent agreement and Gwet's AC were observed. 'Globules/clods', 'Network',

'Regression Structures', 'Shiny White Structures', 'Negative Network', and 'Vessels' showed excellent agreement with Gwet's AC values >0.81. Moderate agreement was observed for 'Lines' and 'Structureless'; 'Dots' and 'Angulated lines' yielded poor agreement.

Superpixel level annotations

Each of the 248 images in our set was delineated into approximately 1,000 (mean=1001.4, SD=18.1) superpixel patches. A total of 47,524 Superpixels were annotated by the expert readers in these images. Disagreement among readers occurred in 81.5% of the superpixels annotated (N=38,732 superpixels)

Percent Superpixel Agreement on the exemplar feature

There were 9 features that yielded 5 reader spatial agreement exceeding 10% of the superpixels where they were the exemplar feature (i.e. *'Typical Network'* with 36.2% absolute (100%) agreement among readers; *'Cobblestone Pattern'* with 24.7%; *'Rim of Brown Globules'* with 19.6%; '*Blotch Regular'* with 17.6%; '*Blotch Irregular'* with 17.2%; *'Negative Network'* with 15.3%; '*Shiny White Streaks',* 15.2%; *'Peppering / Granularity'* with 11.3%; and '*Blue-Whitish Veil'* with 11.3%) **(Table 8).**

Ten features yielded no spatial agreement among the expert readers (0%), those included: 'Regular Dots'; 'Milky Red Globules'; 'Regular Globules'; 'Angulated lines'; 'Branched Streaks'; 'Broadened Network'; 'Delicate Network'; 'Homogeneous Pattern'; 'Milky Red Areas' and 'Corkscrew Vessels'.

Percent Superpixel Agreement on all features (exemplar and non-exemplar)

The features with the highest level of absolute (100%) spatial agreement among readers were: *'Cobblestone Pattern'* with 14.63% absolute (100%) agreement among readers; *'Typical Network'* with 11.88%; *'Rim of Brown Globules'* with 10.86%; *'Dotted Vessels'* with 5.83%; *'Negative Network'* with 5.27%; '*Shiny White Streaks'* with 4.99%; *'Peppering / Granularity'* with 3.82%; *'Starburst Pattern'* with

3.36%; '*Blotch Regular'* with 2.86%; '*Blotch Irregular'* with 2.37% and '*Polymorphous Vessels'* with 2.09% absolute (100%) spatial agreement among readers **(Supplementary Table S2).** The same ten features that had no agreement in the 'exemplar' group did not yield any spatial agreement among 5 readers when evaluating both 'exemplar' and 'non exemplar'.

Confusion matrix and Dice coefficient (exemplar and non-exemplar features)

Percent agreement among experts on the superpixel level was comparatively low; however, there were thirty-one pairs of features displaying consistently high overlap with an average Dice coefficient ≥0.5 (**Table 9)**. *'Atypical Network'* and *'Broadened Network'* were annotated 85 times by different readers, in superpixel regions with a Dice coefficient of 0.584; *'Delicate Network'* and *'Typical network'* were annotated 67 times by different readers, in superpixel regions with a Dice coefficient of 0.637; *'Broadened Network'* and *'Typical Network'* were annotated 33 times by different readers, in superpixel regions with a Dice coefficient of 0.658, showing the potential definition overlap among the 4 different features. Furthermore, '*Homogeneous Pattern'* – a global pattern seems to be non-specific and overlapping with almost all other features annotated (**Figure 16).** Examples of both outcomes (high overlap/agreement in one feature, but equally strong disagreement in other features) can be found in **Supplementary Figure 1.**

Discussion

In this study including 20 international dermoscopy experts, we assessed the agreement on 248 dermoscopic images for thirty-one established melanocytic-specific criteria.[39] Lesion-level agreement was moderate for many of the suggested features, with single-reader observations accounting for 22.4% of all feature selections. We found significant agreement for 14 of the 31 (45.2%) features examined, while excellent agreement was achieved on only 8 of the 31 features (25.8%), even when they were presented as exemplars. These features were: 'Peppering / Granularity' (0.91); 'Shiny White Streaks' (0.89); 'Typical Pigment network' (0.83); 'Blotch Irregular' (0.82); 'Negative Network' (0.81); 'Irregular Globules' (0.78); 'Dotted Vessels' (0.77) and 'Blue Whitish Veil' (0.76). **(Table 7)** Interestingly, those with high agreement included 7 of the 14 (50%) melanoma-specific criteria examined, suggesting that they are pertinent to clinical practice.[41] However, collapsing the features into 'super-categories of features' based on structural similarity yielded higher levels of agreement with the exemption of 'Dots'. These findings suggest that there could be an overlap among definitions and the perception of these features among experts that could be further elucidated in order to create more reliable diagnostic criteria. As expected, positive identification by the expert readers of the thirty-one melanocytic specific features was higher when they were presented as exemplars compared to when non-exemplar features were identified in images by the experts (**Supplementary Figure 2).**

Prior efforts in the field have shown poor to moderate agreement for most dermoscopic features.[39,75,109] However, what distinguishes our study is the use of 'exemplar' images, submitted by the experts who initially described the majority of these features. The use of exemplar images yielded a better agreement compared to the prior efforts and has created a publicly-available exemplar dataset.[10] Additionally, this study also utilizes spatial localization based on the superpixel concept for specific

dermoscopic features. This novel approach allowed for the identification of overlapping features which may offer important insights on the visual perception of features by readers.

Overall agreement and chance-corrected agreement were highly variable for specific dermoscopic features. However, an important aspect of our study was the use of exhaustive (superpixel) annotations for the feature localization on a dermoscopic image. Superpixel annotations allow for additional refinement of our understanding of dermoscopic features and new approaches to feature agreement analysis **(Figure 15).** Agreement on the superpixel level was low; only 19.6% of all superpixels annotated showed any agreement between at least 2 readers and only 11 features achieved an agreement of >2% on the superpixel level among 5 readers. However, in our study we identified 50 pairs of features that displayed constantly a high (>0.5) Dice overlap, highlighting the features that may perceptually be confused with each other, or having definitions that may overlap with each other suggesting that differing terminology may in fact be referring to the same feature, or different features that occur concomitantly (**Table 9, Figure 16).** We included 4 different features termed *'Network / Reticulation',* with structural similarities based on the latest dermoscopy terminology consensus;[6] two of them are melanoma-specific features, that have been correlated with the diagnosis of melanoma on dermoscopy (i.e. *'Atypical Pigment Network',* OR for melanoma 2.0–9.0 and 'Broadened Pigment Network'), one is correlated with the diagnosis of nevi on dermoscopy (i.e. *' Typical pigment network'*), and one non-specific (i.e. *'Delicate Pigment Network*).[39,41] However, despite their different definitions and significance in clinical practice, these features showed consistently a high Dice overlap in our study. Furthermore, this was also the case for features *'Globules / Clods : Irregular'* (melanoma-specific feature, OR for melanoma 1.7-4.8) and *'Globules / Clods : Regular'*, which is a feature specific for benign melanocytic lesions (**Table 9; Figure 16**).[39,41] Future studies are needed to explore whether these

findings can be attributed to a different perception of overlapping features due to ambiguous definitions of these features, different cognitive perception of features or frequent coexistence of these features.

The superpixel approach has been previously used for the training and testing of machine learning algorithms (MLA); however, the 'ground truth' for this task was set by a single annotator and the performance of MLA was poor.[28,29] MLA have shown potential for diagnosing melanoma and other skin cancers on dermoscopic images and could potentially prove to be an adjunct to clinical practice. [30,58,116] Our results and the lack of agreement between annotators for the localization of features suggests that a single annotation for dermoscopic features may not be sufficient to determine ground truth annotations due to concerns related to reproducibility. We demonstrated that multiple reader annotation of curated datasets of dermoscopic images can help improve our understanding of image features, inform and improve MLA performance through supervised or active learning.

The primary goal of our study was to explore expert agreement for the thirty-one melanocytic-specific dermoscopic features; our results show the features that are more likely to be agreed upon, and which features may overlap with others. This type of study can be used as the basis to evaluate which features are reliable, reproducible, and can provide a critical basis for improved diagnostic algorithms by creating a standardized terminology for international consensus, potentially leading to higher diagnostic accuracy both by clinicians and MLAs. An important goal of our study was to create a 'gold standard' dataset of exemplar images displaying the melanocytic-specific criteria for medical teaching, effective scientific communication, and machine learning experiments. There are now 248 dermoscopic images publicly available with their assorted annotations and areas of agreement for the thirty-one melanocytic-specific criteria, generated by experts in dermoscopy.

Limitations

The use of superpixels is a novel method for the evaluation of agreement. However, these pre-determined regions might not be optimal for attribute identification as dermoscopic features may not be bounded by superpixels or contrast thresholds, especially features, such as network, that inherently contain high contrast. Another important limitation of this study is the use of expert readers; extrapolating the results of this study to the general dermatology community must be made with caution. Furthermore, despite our efforts to include a diverse group of readers, we only had two readers from Latin America, two readers from Australia, and none from Africa and Asia. Finally, this should not be interpreted as an epidemiologic study, since we have intentionally enriched the dataset for melanomas and their exemplar features; the prevalence of these features has been previously studied and the scope of this paper was to examine the expert agreement on the presence and localization of melanoma-specific features.

Conclusion

Agreement on dermoscopic features remains variable, even when using highly selected exemplar images among experts. The use of exemplar images facilitates the agreement process while the use of superpixels provides insights on commonly overlapping features and the intrinsic intricacies of dermoscopic features. We found that half of the melanoma-specific criteria have good reproducibility, yielding high agreement, while others, such as 'Broadened Pigment Network', 'Delicate Pigment Network', 'Tan Peripheral Brown Areas' have very low agreement even on the best-case scenario of exemplar images and may need elimination from the lexicon of dermoscopy. Finally, analysis of the superpixel spatial overlap revealed that only 19.6% of superpixel overlaps occurred between different features, indicating that there may be redundancy or confusion in feature terminology. These observations are the necessary prerequisites for enhanced consensus building and nomenclature standardization. Our results will be used as a roadmap to achieve better terminology standardization and to guide melanoma diagnostic algorithms that ultimately would increase worldwide use and acceptance of dermoscopy.

Tables

Table 7.

Measures of agreement for individual dermoscopic features and combined 'super-feature' categories on a lesion level.

Individual Dermoscopic Features					**Combined Dermoscopic Super-feature categories**			
Variable	**% agreement**	**Fleiss' kappa**	**Gwet's AC**		**Variable**	**% agreement**	**Kappa**	**Gwet's AC**
Dots : Irregular	60.00%	0.0954	0.2829		Dots	61.00%	0.1371	0.2884*
Dots : Regular	60.00%	0.2000	0.2000					
Globules / Clods : Cobblestone pattern	60.00%	.0.0809	0.2918		Globules / Clods	91.00%	0.3706*	0.8950*
Globules / Clods : Irregular	82.00%	-0.0989	0.7847*					
Globules / Clods : Regular	46.00%	-0.1315	-0.0301					
Globules / Clods : Rim of brown globule	78.00%	0.4419	0.6368*					

Lines : Branched streaks	76.00%	0.0033	0.6839 *		Lines	70.77%	0.3542*	0.4659*
Lines : Pseudopods	54.00%	-0.1170	0.2180					
Lines : Radial streaming	65.93%	0.0143	0.4792					
Lines : Starburst (pseudopods/radial)	54.00%	0.0417	0.1154					
Network : Atypical pigment network / Re	70.00%	-0.0631	0.5821 *		Network	100%	1.0*	1.0*
Network : Broadened pigment network / R	38.00%	-0.2917	-0.1923					
Network : Delicate Pigment Network / Re	48.00%	-0.0552	-0.0252					
Network : Typical pigment network / Ret	86.00%	0.2222	0.8293 *					

Regression structures : Peppering / Gra	92.00%	-0.0417	0.9133 *		Regression structures	96.00%	-0.0204	0.9584*
Regression structures : Scarlike depigm	80.00%	0.3225	0.7162 *					
Shiny white structures : Shiny white st	90.00%	-0.0526	0.8895 *		Shiny white structures	90.00%	-0.0526	0.8895*
Independent: Angulated lines / Polygons / Zig-zag	60.00%	0.1883	0.2114		Angulated lines/ polygons	60.00%	0.1883	0.2114
Independent : Negative pigment network (independent)	84.00%	-0.0870	0.8124		Negative network	84.00%	-0.0870	0.8124
Structureless : Blotch irregular	86.00%	0.4186	0.8156 *		Structureless	75.06	0.1699*	0.6434*

Structureless : Blotch regular	66.00%	0.2052	0.4058					
Structureless : Blue-whitish veil	80.00%	-0.1111	0.7561 *					
Structureless : Milky red areas	58.00%	0.0349	0.2564					
Structureless : Structureless brown (ta	42.67%	-0.1485	-0.1448					
Structureless: Homogeneous : NOS	46.57%	-0.0876	-0.0876					
Vessels : Comma	76.00%	0.0588	0.6779 *		Vessels	89.49%	0.3462*	0.8748*
Vessels : Corkscrew	58.00%	-0.1858	0.3496					
Vessels : Dotted	82.00%	0.1477	0.7718 *					
Vessels : Linear irregular	68.89%	0.2742	0.4555 *					
Vessels : Polymorphous	54.00%	-0.0403	0.1753					
Vessels: Milky red globules	67.00%	-0.0124	0.5104 *					

Table 8.

Superpixel agreement on the exemplar feature.

Percent agreement among the experts on the spatial localization (superpixel agreement) of the exemplar features. The **Mean Number of Readers** shows the average N of readers annotating the images for this feature. The percentage of agreement of ≥60% (at least 3 out of 5 readers annotating the same region / superpixels for the same feature) can be seen under **3RA %**, while under **Full - 5RA%** we can see the percentage of absolute agreement for each feature (5 out of 5 readers annotating the same region / superpixels for the same feature).

Feature	Mean Number of Readers	3RA - %	5RA- %
Dots : Irregular	3.875	0.192648	0.003571
Dots : Regular	2.5	0.250801	0
Globules / Clods : Cobblestone pattern	3.625	0.629329	0.247115
Globules / Clods : Irregular	4.875	0.349432	0.043504
Globules / Clods : Milky red	1.25	0.016079	0
Globules / Clods : Regular	2.875	0.299281	0
Globules / Clods : Rim of brown globules	3.625	0.408223	0.195766

Lines : Angulated lines / Polygons / Zig-zag pattern	3.5	0.156645	0
Lines : Branched streaks	0.875	0	0
Lines : Pseudopods	3.5	0.300592	0.003906
Lines : Radial streaming	4	0.20898	0.035716
Network : Atypical pigment network / Reticulation	4.25	0.30728	0.031322
Network : Broadened pigment network / Reticulation	1.875	0.092118	0
Network : Delicate Pigment Network / Reticulation	2.125	0.01616	0
Network : Negative pigment network	4.5	0.644936	0.152862
Network : Typical pigment network / Reticulation	4.375	0.756599	0.362415
Pattern : Homogeneous : NOS	2.125	0.180564	0
Pattern : Starburst	3.125	0.330246	0.041268
Regression structures : Peppering / Granularity	4.625	0.504966	0.112971

Regression structures : Scarlike depigmentation	3.875	0.312567	0.050974
Shiny white structures : Shiny white streaks	5	0.532319	0.151495
Structureless : Blotch irregular	4.125	0.417691	0.172234
Structureless : Blotch regular	3.375	0.508637	0.175758
Structureless : Blue-whitish veil	4.375	0.4768	0.112656
Structureless : Milky red areas	3	0.184145	0
Structureless : Structureless brown (tan)	2.5	0.157379	0.033333
Vessels : Comma	4.375	0.176428	0.027243
Vessels : Corkscrew	1.125	0	0
Vessels : Dotted	4.5	0.541738	0.082328
Vessels : Linear irregular	3.875	0.247682	0.036885
Vessels : Polymorphous	3.25	0.315864	0.002083

Table 9.

Pairs of features with high (>0.5) Dice spatial overlap on the superpixel annotations. **N of pairs** indicates how many times those pairs occurred with high overlap in our study, while **% overlap** shows the percentage of their spatial overlap.

N of pairs	% overlap	Feature 1	Feature 2
92	52.80%	Network : Atypical pigment network / Reticulation	Network : Typical pigment network / Reticulation
85	58.40%	Network : Atypical pigment network / Reticulation	Network : Broadened pigment network / Reticulation
68	55.60%	Vessels : Linear irregular	Vessels : Polymorphous
67	63.70%	Network : Delicate Pigment Network / Reticulation	Network : Typical pigment network / Reticulation
64	67.00%	Lines : Radial streaming	Pattern : Starburst
58	86.50%	Globules / Clods : Cobblestone pattern	Globules / Clods : Regular
42	90.10%	Pattern : Homogeneous : NOS	Structureless : Blotch regular
33	65.80%	Network : Broadened pigment network / Reticulation	Network : Atypical pigment network / Reticulation
31	71.50%	Globules / Clods : Irregular	Globules / Clods : Regular

29	66.30%	Pattern : Homogeneous : NOS	Structureless : Structureless brown (tan peripheral area)
29	50.70%	Network : Negative pigment network / Reticulation	Shiny white structures : Shiny white streaks
25	62.20%	Dots : Regular	Globules / Clods : Regular
24	67.90%	Globules / Clods : Regular	Globules / Clods : Rim of brown globules
24	56.20%	Pattern : Homogeneous : NOS	Structureless : Blotch irregular
22	50.30%	Dots : Irregular	Dots : Regular
18	70.10%	Lines : Pseudopods	Pattern : Starburst
16	54.90%	Pattern : Homogeneous : NOS	Structureless : Blue-whitish veil
14	72.30%	Structureless : Blotch regular	Structureless : Blue-whitish veil
14	67.00%	Globules / Clods : Rim of brown globules	Lines : Pseudopods
13	50.30%	Structureless : Blotch regular	Structureless : Structureless brown (tan
12	57.50%	Dots : Regular	Network : Atypical pigment network / Reticulation
9	56.80%	Globules / Clods : Milky red	Vessels : Polymorphous
9	52.40%	Globules / Clods : Regular	Network : Negative pigment network / Reticulation

7	92.90%	Structureless : Blotch irregular	Structureless : Blotch regular
5	71.60%	Lines : Branched streaks	Pattern : Starburst
5	61.40%	Globules / Clods : Cobblestone pattern	Pattern : Homogeneous : NOS
5	50.00%	Vessels : Corkscrew	Vessels : Polymorphous
4	89.40%	Dots : Regular	Globules / Clods : Cobblestone pattern
4	74.10%	Globules / Clods : Cobblestone pattern	Network : Negative pigment network
4	72.30%	Globules / Clods : Rim of brown globules	Pattern : Starburst
3	50.30%	Regression structures : Peppering / Granularity	Structureless : Blotch regular

Figures

Figure 15.

A. Sample image (ISIC_0022328) of a melanoma included in the study for 'exemplar' feature 'Structureless: Blue Whitish Veil'; **B-F.** Image overlaid with the superpixel outlines and annotation markups selected by each of five expert readers. **G.** Color representation of each feature selected by the readers. **H.** Agreement overlap among all readers.

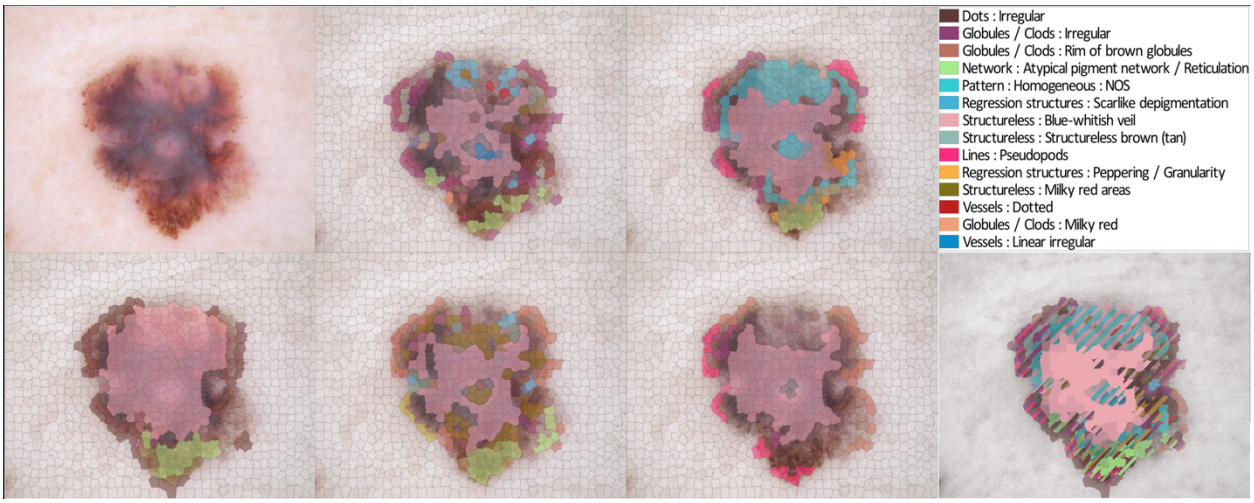
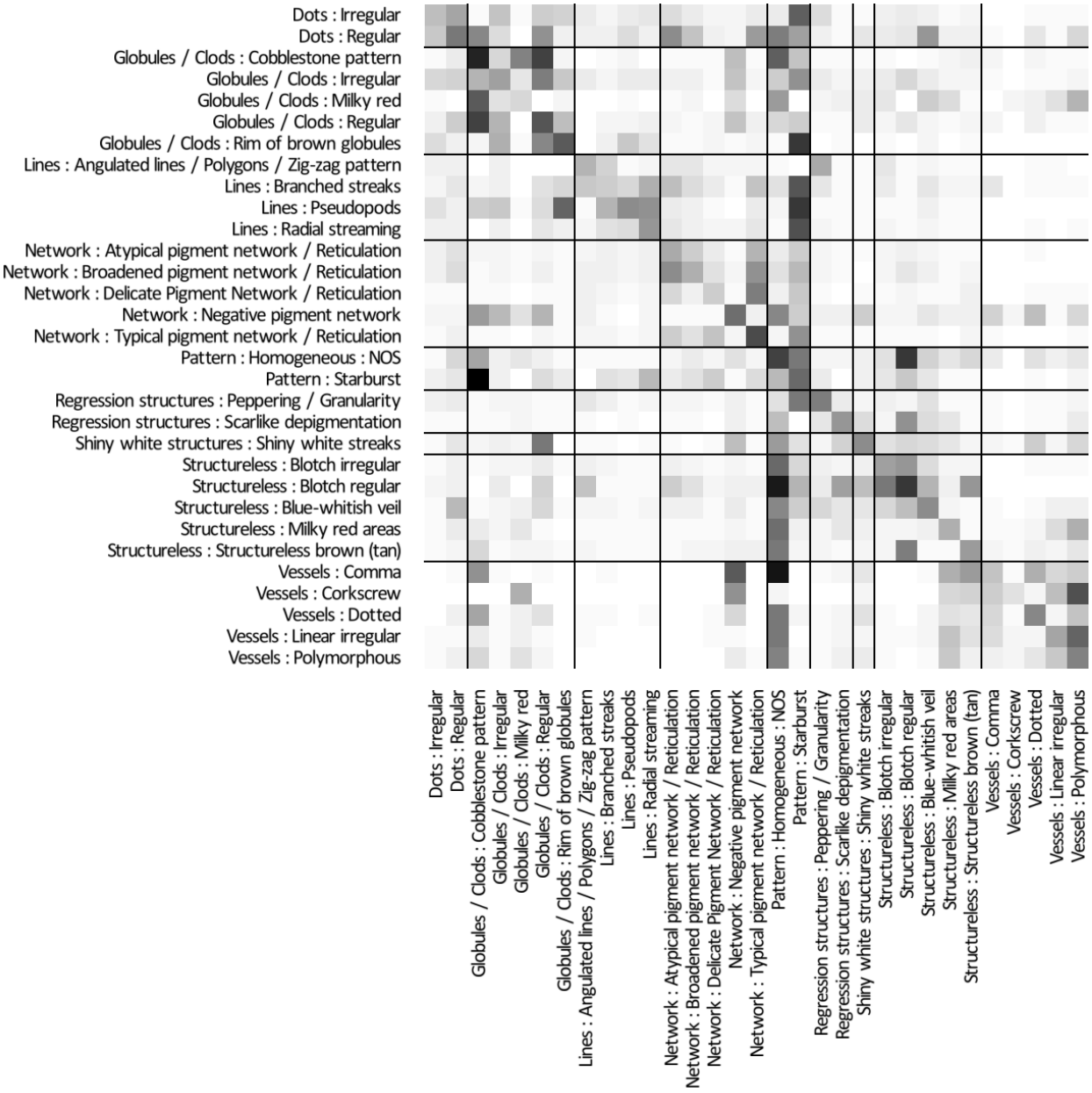


Figure 16.

Confusability Matrix: Each element of this confusion matrix is the median D coefficient across all pairs of features which occurred in the dataset. In the cases where the diagonal elements are as close to 1.0 as possible (dark color), this suggests that all pairs of readers (each individual reader directly compared to each of the other four -across all studies) selected close to identical sets of superpixels for the same feature and image. When all off-diagonal elements approach 0.0 (light color), this suggests that for different features, pairs of readers rarely selected overlapping superpixels. Feature 'Pattern : Homogeneous', a non-specific feature which can be observed in any region of the lesion overlaps with a multitude of features, commonly with 'Blotch : Regular' and 'Vessels : Comma', however the agreement

on the superpixel level for this feature among experts for this feature was 0%. Additionally, features ‘Dots : Regular’ and ‘Dots : Irregular’ overlap among them and with ‘Globules / Clods’, while both the superpixel level agreement and the Gwet’s AC was low for these features.



SUPPLEMENTARY MATERIAL

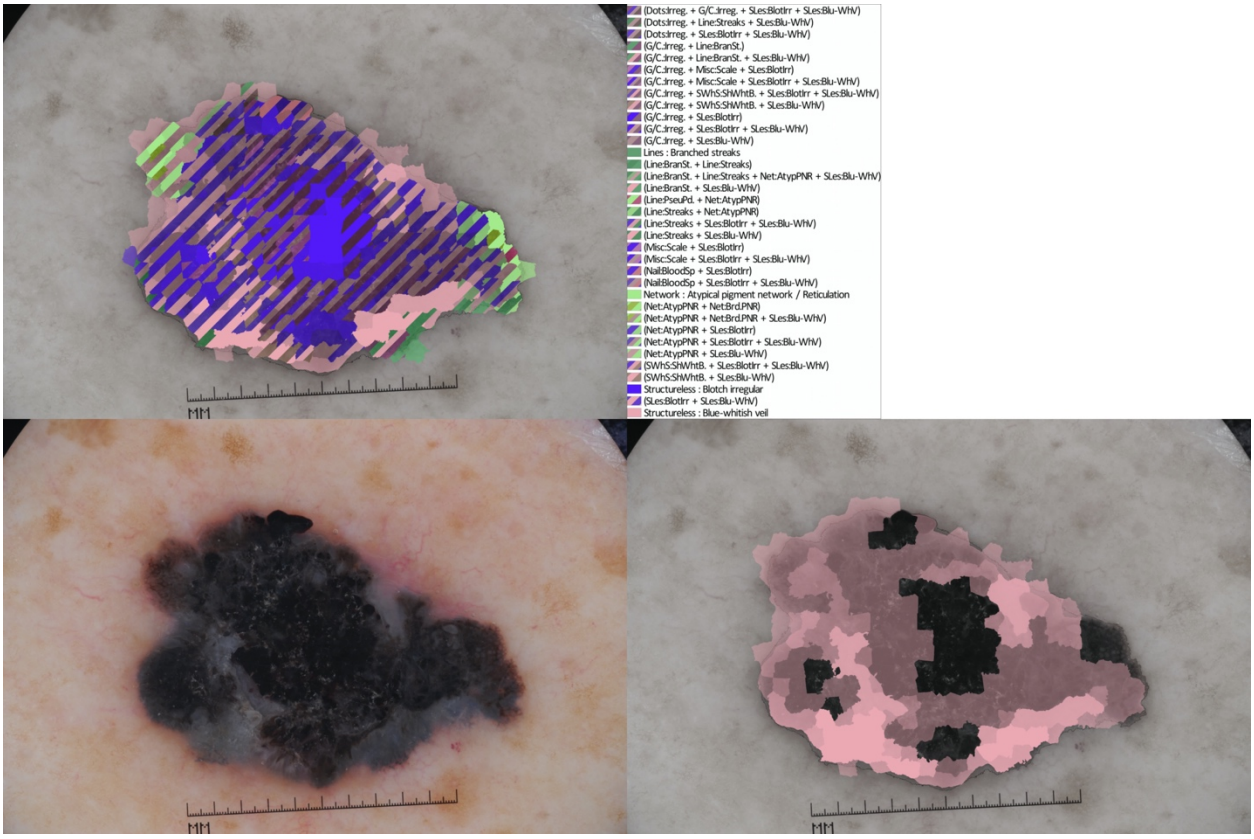
Supplementary Results Section:

Experts included in this study were: five experts from the USA; three from Spain; two from Austria; two from Italy and one expert from Australia, Chile, Colombia, France, Germany, Greece, Israel and Switzerland. Thirty percent of experts were female (n=6); all readers had more than 10 years of dermoscopy experience.

The median annotation time per image was 2:37 minutes (IQR 1:25 – 5:11 minutes). Each reader annotated an average of 3.8 (SD=2.4 per annotation, SD=1.53 across images) features per lesion for a total of 4507 feature markups. The average number of features annotated per image by the experts varied per diagnosis, from 2.86 (SD=1.80) for nevi up to 4.63 (SD=2.55) for melanomas ($p<0.001$).

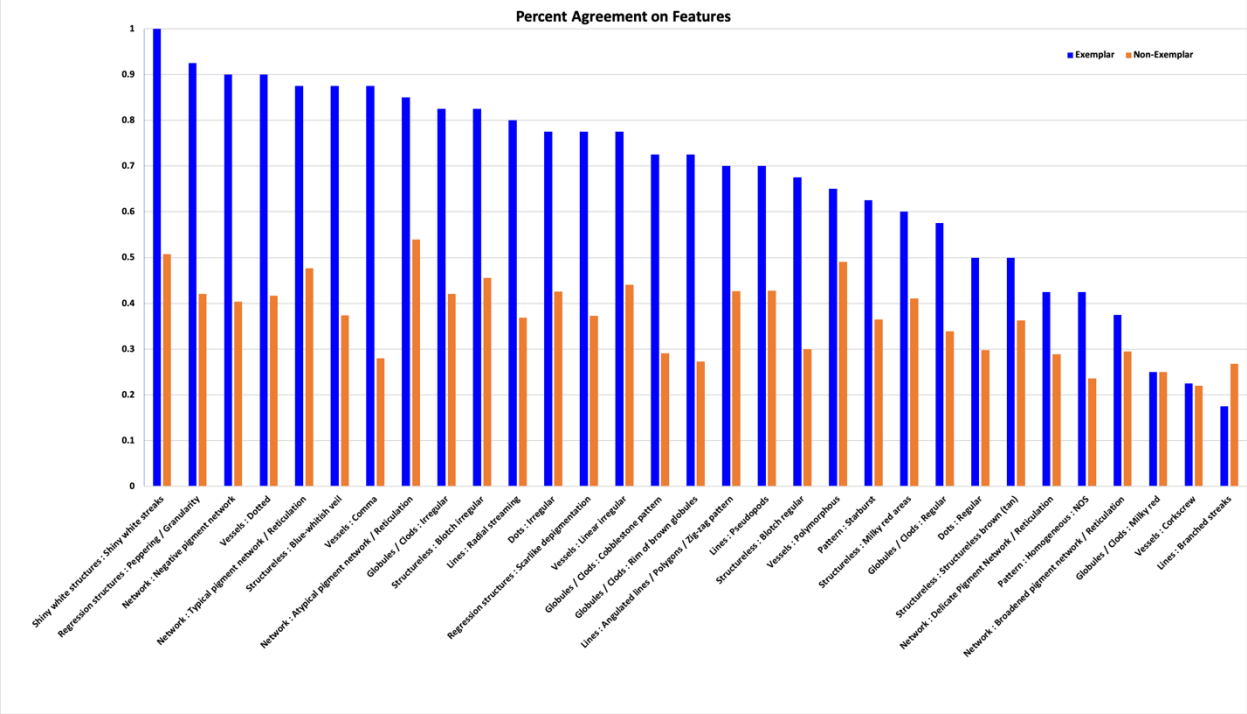
Supplementary Figure 1:

Example of an image included in the study (ISIC_0016128), displaying both examples of high superpixel agreement among readers for the exemplar feature ('*Structureless : Blue-whitish veil*' - highlighted in pink in these annotations), and high Dice overlap among annotations between *'Network : Atypical pigment network / Reticulation'* and *'Network : Broadened pigment network / Reticulation'.*


(Dots:Irreg. + G/C.:Irreg. + SLes:Blotlrr + SLes:Blu-Whv)
(Dots:Irreg. + Line:Streaks + SLes:Blu-Whv)
(Dots:Irreg. + SLes:Blotlrr + SLes:Blu-Whv)
(G/C.:Irreg. + Line:BranSt.)
(G/C.:Irreg. + Line:BranSt. + SLes:Blu-Whv)
(G/C.:Irreg. + Misc:Scale + SLes:Blotlrr)
(G/C.:Irreg. + Misc:Scale + SLes:Blotlrr + SLes:Blu-Whv)
(G/C.:Irreg. + SWhS:SHWhtB. + SLes:Blotlrr + SLes:Blu-Whv)
(G/C.:Irreg. + SWhS:SHWhtB. + SLes:Blu-Whv)
(G/C.:Irreg. + SLes:Blotlrr)
(G/C.:Irreg. + SLes:Blotlrr + SLes:Blu-Whv)
(G/C.:Irreg. + SLes:Blu-Whv)
Lines : Branched streaks
(Line:BranSt. + Line:Streaks)
(Line:BranSt. + Line:Streaks + Net:AtypPNR + SLes:Blu-Whv)
(Line:BranSt. + SLes:Blu-Whv)
(Line:PsePd. + Net:AtypPNR)
(Line:Streaks + Net:AtypPNR)
(Line:Streaks + SLes:Blotlrr + SLes:Blu-Whv)
(Line:Streaks + SLes:Blu-Whv)
(Misc:Scale + SLes:Blotlrr)
(Misc:Scale + SLes:Blotlrr + SLes:Blu-Whv)
(NailBloodSp + SLes:Blotlrr)
(NailBloodSp + SLes:Blotlrr + SLes:Blu-Whv)
Network : Atypical pigment network / Reticulation
(Net:AtypPNR + Net:Brd.PNR)
(Net:AtypPNR + Net:Brd.PNR + SLes:Blu-Whv)
(Net:AtypPNR + SLes:Blotlrr)
(Net:AtypPNR + SLes:Blotlrr + SLes:Blu-Whv)
(Net:AtypPNR + SLes:Blu-Whv)
(SWhS:SHWhtB. + SLes:Blotlrr + SLes:Blu-Whv)
(SWhS:SHWhtB. + SLes:Blu-Whv)
Structureless : Blotch irregular
(SLes:Blotlrr + SLes:Blu-Whv)
Structureless : Blue-whitish veil
MM
MM
MM

Supplementary Figure 2.

Percent agreement on the presence of features across all 248 images on the lesion level. An agreement of 100% represents positive identification of the feature by all 20 experts for all times the feature was identified in an image; with blue we see the agreement when the features occurred in exemplar images and with orange is the agreement for the same features when they were identified in non-exemplar images.



Supplementary Tables:

Supplementary Table S1.

Terminology Glossary: Dermoscopy has an extended terminology and two different approaches to the names of the features, metaphoric and descriptive. Below we provide the feature names and their correspondence between metaphoric and descriptive terminology, their definitions and abbreviations

which are used in the manuscript. Table adapted and modified with permission from Harald Kittler et al.

[39]

Metaphoric Terminology	**Definition**	**Descriptive terminology**	**Abbreviation**
Dots: Regular	Dots clustered at the center of the lesion, or located on the network lines or in the hole of the network (also called target network)	Target: Dots, brown, central (in the center of hypopigmented spaces between reticular lines)	Regular Dots
Dots: Irregular	Any distribution of dots other than dots as described for regular dots		Irregular Dots
Globules / Clods: Cobblestone Pattern	Polygonal globules symmetrically distributed throughout lesion	Clods, brown or skin colored, large and polygonal	Cobblestone Pattern
Globules / Clods: Irregular	Globules with variability in color, size, shape, or spacing and distributed in an asymmetric/disorganized fashion		Irregular Globules
Globules / Clods: Milky-red		Clods, pink and small	Milky-red Globules
Globules / Clods: Regular	Globules with minimal variability in their color, size, and shape located in the center of a lesion with surrounding network, or along the perimeter or throughout the entire lesion	Clods, small, round or oval	Regular Globules
Globules / Clods: Rim of brown globules	Globules distributed at the periphery of lesion	Clods, brown, circumferential	Rim of Brown Globules
Lines: Angulated lines / Polygons / Zig-Zag	Gray-brown lines that are connected at an angle or coalescing to form polygons	Lines, angulated or polygonal (non-facial skin)	Angulated Lines

Lines: Branched streaks	Atypical network with broken/interrupted lines and incomplete connections		Branched Streaks
Lines: Pseudopods	Bulbous and often kinked projections seen at the lesion edge, either directly associated with a network or solid tumor border		Pseudopods
Lines: Radial streaming	Radial linear extensions at the lesion edge	Lines, radial and segmental	Radial streaming
Network: Atypical pigment network / Reticulation	Network with increased variability in line color, thickness and spacing. Gray color to lines or disorganized distribution	Lines, reticular and thick or reticular lines that vary in color	Atypical Network
Network: Broadened pigment network / Reticulation	Widening of the network lines	Lines, reticular and thick	Broadened Network
Network: Delicate pigment network / Reticulation	Fine thin network	Lines, reticular and thin	Delicate Network
Network: Typical pigment network / Reticulation	Network with minimal variability in the color, thickness, and spacing of the lines; symmetrically distributed	Lines, reticular	Typical Network
Network: Negative pigment network / Reticulation	Serpiginous interconnecting broadened hypopigmented lines that surround elongated and curvilinear brown structures	Lines, reticular, hypopigmented, around brown clods	Negative Network
Patterns: Starburst pattern	This pattern consists of tiered peripheral globules, pseudopods, or streaks (or a combination of them),	Pseudopods, circumferential or lines, radial, circumferential	Starburst Pattern

	located around the entire perimeter of the lesion		
Patterns: Homogeneous pattern	A pattern lacking any definable pigment structures, also know as structureless pattern	Structureless, any color	Homogeneous Pattern
Regression structures: Peppering/granularity	Consists of fine dots with a blue-gray color	Dots, gray	Peppering / Granularity
Regression structures: Scar-like depigmentation	Area of white that is whiter than surrounding normal-appearing skin (true scarring); it should not be confused with hypopigmentation or depigmentation caused by simple loss of melanin; shiny white structures and blood vessels are not seen in areas of regression	Structureless zone, white	Scar-like Depigmentation
Shiny white structures: Shiny white streaks	Short discrete white lines oriented parallel and orthogonal (perpendicular) to each other seen only under polarized dermoscopy	Lines, white, perpendicular	Shiny White Streaks
Structureless: Blue whitish veil	A raised/palpable blotch of blue hue with an overlying whitish groundglass haze	Structureless zone, blue	Blue whitish veil
Structureless: Blotch Regular	One blotch within center of lesion and surrounded by network		Regular Blotch
Structureless: Blotch Irregular	More than one blotch or a blotch that is located off center		Irregular Blotch

Structureless: Tan (Brown) Peripheral Structureless areas		Structureless, brown (tan), eccentric	Tan Peripheral Structureless Areas
Structureless: Milky-red areas	Milky-white appearance or pinkish structureless areas (strawberry and ice cream-like), consisting a red vascular blush with no specific distinguishable vessels		Milky-red Areas
Vessel morphology, monomorphous: Dots	Tiny pinpoint vessels		Dotted Vessels
Vessel morphology, monomorphous: Comma	Linear, curved, short vessels	Curved	Comma Vessels
Vessel morphology, monomorphous: Corkscrew	Twisted looped vessels with bends twisted along a central axis	Helical	Corkscrew Vessels
Vessel morphology, monomorphous: Linear irregular / Serpentine	Linear, curved or serpentine vessels,	Serpentine	Linear Irregular Vessels
Vessel morphology: Polymorphous			Polymorphous Vessels

Supplementary Table S2.

Dermoscopic features specific for melanocytic lesions (nevi and melanomas) included in our study, the total Number of observations **(on the lesion-level**) of these features (**Total N of observations**), the number of images in which they were observed **(In images),** and the respected agreement they yielded. **Orphan observations** - observations by single readers; **2-RA** - agreement among 2 readers; **3-RA** - agreement among 3 readers, threshold of ≥60% for the presence of a feature in an image; **4-RA** -

agreement among 4 readers (80% agreement), and **5-RA (FA)** - full agreement among 5 readers, 100% agreement for the presence of a feature in an image.

Dermoscopic Feature	**Total N of observations**	**In images**	**Orphan observations**	**Lesions ≥ 40% agreement (≥2RA)**	**Lesions ≥ 60% agreement (≥3RA)**	**Lesions ≥ 80% Agreement (≥4RA)**	**Lesions with 100% agreement**
Dots : Irregular	**268**	**124**	**48**	**76**	**44**	**18**	**6**
Dots : Regular	**102**	**65**	**37**	**28**	**7**	**2**	**0**
Globules / Clods : Cobblestone pattern	**60**	**29**	**14**	**15**	**9**	**5**	**2**
Globules / Clods : Irregular	**319**	**150**	**67**	**83**	**49**	**27**	**10**

Globules / Clods : Milky red	**40**	**31**	**24**	**7**	**2**	**0**	**0**
Globules / Clods : Regular	**103**	**57**	**28**	**29**	**14**	**3**	**0**
Globules / Clods : Rim of brown globules	**58**	**29**	**18**	**11**	**8**	**6**	**4**
Lines : Angulated lines / Polygons / Zig-zag pattern	**72**	**30**	**9**	**21**	**16**	**5**	**0**
Lines : Branched streaks	**82**	**61**	**43**	**18**	**3**	**0**	**0**
Lines : Pseudopods	**116**	**50**	**19**	**31**	**23**	**11**	**1**

Lines : Radial streaming	**135**	**66**	**33**	**33**	**21**	**12**	**3**
Network : Atypical pigment network / Reticulation	**344**	**128**	**32**	**96**	**67**	**39**	**14**
Network : Broadened pigment network / Reticulation	**130**	**86**	**51**	**35**	**9**	**0**	**0**
Network : Delicate Pigment Network / Reticulation	**124**	**81**	**55**	**26**	**12**	**5**	**0**
Network : Negative pigment network	**124**	**54**	**22**	**32**	**19**	**13**	**6**

Network : Typical pigment network / Reticulation	**245**	**101**	**34**	**67**	**39**	**26**	**12**
Pattern : Homogeneous : NOS	**88**	**71**	**57**	**14**	**2**	**1**	**0**
Pattern : Starburst	**53**	**24**	**10**	**14**	**8**	**5**	**2**
Regression structures : Peppering / Granularity	**249**	**114**	**48**	**66**	**35**	**23**	**11**
Regression structures : Scarlike depigmentation	**161**	**81**	**42**	**39**	**24**	**11**	**6**

Shiny white structures : Shiny white streaks	**218**	**82**	**24**	**58**	**40**	**25**	**13**
Structureless : Blotch irregular	**227**	**98**	**44**	**54**	**37**	**27**	**11**
Structureless : Blotch regular	**71**	**39**	**24**	**15**	**9**	**6**	**2**
Structureless : Blue-whitish veil	**183**	**93**	**39**	**54**	**27**	**6**	**3**
Structureless : Milky red areas	**169**	**80**	**32**	**48**	**25**	**15**	**1**
Structureless : Structureless brown (tan)	**250**	**138**	**62**	**76**	**28**	**7**	**1**

Vessels : Comma	**63**	**28**	**16**	**12**	**10**	**8**	**5**
Vessels : Corkscrew	**20**	**17**	**14**	**3**	**0**	**0**	**0**
Vessels : Dotted	**147**	**63**	**26**	**37**	**25**	**15**	**7**
Vessels : Linear irregular	**143**	**61**	**22**	**39**	**25**	**16**	**2**
Vessels : Polymorphous	**143**	**54**	**14**	**40**	**28**	**16**	**5**

Discussion

Skin cancer is the most common cancer worldwide.[95] It is estimated that in the United States alone, non-melanoma skin cancer (BCC and SCC) could be exceeding 5 million cases per year.[11,12] At the same time, approximately 220.000 cases of melanoma, the deadliest form of skin cancer, are recorded each year, as well as, 37.000 deaths in Europe and the US.[117] In recent years there has been a significant improvement in the survival and quality of life of patients with advanced skin cancer.[15-19] Nevertheless, the mainstay for decreasing skin cancer morbidity and mortality remains early skin cancer diagnosis and surgical excision of the tumor.[19-21]

In our thesis we examined the potential role that AI algorithms could play in early skin cancer diagnosis. AI algorithms, and Machine Learning algorithms in particular, have displayed promising results in various fields of medicine, including ophthalmology and radiology, while there are already FDA approved AI algorithms that have been embedded into clinical practice.[9] Through ISIC Archive (www.isic-archive.com) we explored the diagnostic accuracy of AI algorithms for skin cancer diagnosis, as well as their possible implementations in everyday clinical practice. ISIC Archive is an academia - industry partnership whose primary goal is to decrease melanoma mortality with the use of new imaging technologies. Towards this goal, ISIC Archive has been hosting annual AI competitions (2016-2020) utilizing its constantly expanding dermoscopic images archive. Herein we presented the results of ISIC Grand Challenge 2017, as well as our findings from the reader study that followed.[28,30] We found that the best performing algorithm, at a given sensitivity threshold, had higher specificity for melanoma diagnosis, compared with both expert dermatologists and dermatology residents. Our most impressive finding was that in cases where clinicians had low confidence in their diagnosis, especially less

experienced clinicians, the adjunct use of the algorithm could potentially improve their diagnostic accuracy. These findings have been confirmed by subsequent research.[59,116]

Following, we explored the reproducibility of the results of published, publicly available AI algorithms, potential factors that could affect their diagnostic outcomes, as well as their generalizability in a population different than the one they were trained.[56,59] We selected and published a standardized skin cancer image dataset[10] to validate two publicly available AI algorithms and found that the diagnostic accuracy of these algorithms is significantly reduced when lesions from patients with a different skin type are used.[31,32] Additionally, we found that image based AI algorithm for skin cancer diagnosis can be affected by various factors, such as image rotation, or manipulations regarding image brightness or contrast.[31] We finally found that the diagnostic accuracy of AI algorithms is heavily dependent on whether the image is in focus and well centered.[32] Our findings are in accordance with similar research in the field, highlighting the intrinsic weaknesses of AI algorithms for skin cancer diagnosis.[61-63] These weaknesses need to be overcome before these algorithms are incorporated in everyday clinical practice.

During our thesis we suggested new statistical and interpretation approaches which could be more suitable when evaluating the application of AI algorithms in skin cancer diagnosis. At the same time, we highlighted the need for these algorithms to be tested in the general population, outside the biased settings of research, where skin cancer, and especially melanoma is heavily represented. On this ground, prospective, randomized clinical trials will be needed, and we are already running some of these studies in “Andreas Sygros” Hospital for Cutaneous and Venereal Diseases, trying to better define the potential utility of these AI algorithms.

As discussed prior, dermoscopy is a widely used, non-invasive, diagnostic technique that has significantly aided in early skin cancer diagnosis, both for melanoma and non-melanoma skin cancers.[22,23] However, agreement on specific dermoscopic criteria remains moderate to poor, even among dermoscopy experts.[39,75,109] Through our research we described novel dermoscopic criteria i. for diagnosing hard-to-diagnose melanomas, such as acral amelanotic melanoma; ii. For high-risk Basal Cell Carcinomas, and finally, iii. Their benign mimickers.[34,36,37,118] Furthermore, we conducted the Expert Agreement Study on Dermoscopy of Melanocytic Lesions, defining the reliability and reproducibility, along with the overlap of dermoscopic criteria used to differentiate nevi from melanomas. Our results could be utilized in order to: 1. Improve the diagnostic algorithms used by clinicians for skin cancer diagnosis, and 2. Improve the diagnostic accuracy of AI algorithms through targeted training.[27,28]

A question that is often posed, is what purposes the study of AI algorithms for diagnosing skin cancer serve. Why do we need AI algorithms for skin cancer diagnosis and finally, will their incorporation in clinical practice prove beneficial to the patient?

At this point we need to face some of the challenges that we, as physicians, face with regards to early skin cancer diagnosis. The majority of melanomas, approximately 75%, is not initially diagnosed by dermatologists, but by the patients themselves, or a family member.[119-124] Additionally, a high percentage of patients will not have access to specialized care.[125] Finally, our diagnostic accuracy for melanoma diagnosis, as dermatologists, is not exceedingly high.[126] Incorporating AI algorithms in clinical practice could overcome some of these limitations, especially if these algorithms were to be widely available via smartphone applications for example.

However, a critical question arises here, which according to our opinion, should be addressed by the scientific community. Hypothesizing that in the near future an AI algorithm with almost perfect sensitivity and specificity is generated. To whom should such a powerful diagnostic tool be addressed? Should it be available to the general public? And more importantly, from a Hippocratic standpoint, would its use produce more good than harm? In recent years a constant rise in the number of melanomas diagnosed is observed, while melanoma mortality remains stable, and recently, declining.[127] The trend is similar for non-melanoma skin cancers as well.[12,95] From our point of view, the decline in melanoma and non-melanoma skin cancer morbidity and mortality could be highly attributed to the novel, targeted, systemic treatments,[17,19] while the increase in skin cancer incidence is caused by the following factors: 1. The life expectancy, especially in western societies, is constantly rising and it is portrayed to maintain the same trend throughout the 21st century.[128] 2. Our sensitivity and specificity, as dermatologists, for skin cancer diagnosis has substantially improved drastically with the addition of novel diagnostic tools, such as dermoscopy and Reflectance Confocal Microscopy.[22,23,72] 3. The diagnostic drive and pressure to diagnose melanoma, and skin cancer in general, at an earlier stage has been a constant push towards that direction.[129-131]

However, we are not aware whether the diagnosis of skin cancer at an extremely early stage, or cancers that would have doubtful consequences at the patients' quality of life, or life expectancy, is doing more good than harm.[131-133] Within that spectrum, we are not aware whether the excessive use of technology and the most advanced diagnostic tools can lead to opposite than the desired results. Could the excessive use of technology lead to diagnosing "harmless" skin cancers, which would cause minimal, or no consequences to the patients if we were not so capable at diagnosing them? These are questions that are also being raised for the screening policies for breast and prostate cancer.[131,134] What could be

the outcome if such a powerful diagnostic tool, with the ability to detect minimal alterations in the size, shape and color of a skin lesion, finds its way to the hands of billions of people? The resulting pressure to health systems around the world could be tremendous, for diagnosing cancers with uncertain malignant potential, while the benefit to the patients could be limited.

Finally, an important thing to highlight to all the aforementioned is the role of physicians as something more than people that merely diagnose and treat a skin lesion, especially when the lesion in question is cancer. The relationship of trust between doctor and patient is one of the quintessential steps in the care of the patient, and furthermore the cancer patient. The use of new technologies, especially if they aid us in becoming better physicians, is more than welcome, yet, only within a framework where we will be able to provide better care for our patients.[135]

Concluding, technological advancements have offered invaluable services to human, and the same is anticipated for the Fourth Industrial Revolution, where AI algorithms are expected to play a major role. We anticipate that the use of this new technology will have an equally important role in Medicine and dermatology in particular. However, from our point of view, it has to be the scientific community as a wholesome that will define the framework within which AI will be used, in order for clinicians, aided by AI algorithms, to always “do good, or at least, do no harm”.

Conclusion

In conclusion, during our thesis (1) we identified that AI algorithms have a great potential to aid clinicians in early skin cancer diagnosis. (2) We found that clinicians with limited clinical experience could benefit more from their use. (3) We showed some important intrinsic weaknesses of AI algorithms; weaknesses that need to be overcome before they are embedded in clinical practice. (4) We suggested novel statistical and interpretation approaches for evaluating the results of the research conducted in AI algorithms and early skin cancer diagnosis. And (5), we described new dermoscopic criteria for diagnosing skin cancer and we identified some of the pitfalls of existing dermoscopic criteria for melanoma diagnosis. These results could be used to improve upon diagnostic algorithms used by physicians for skin cancer diagnosis and for better training of AI algorithms for the same purpose. Finally, we described the advantages and disadvantages that the use of AI algorithms for early skin cancer diagnosis could generate, as well as the framework within which they could be incorporated in clinical practice, always under the guidance and supervision of the scientific community.

Reference

1 Pedro F, Subosa M, Rivas A, Valverde P. Artificial intelligence in education : challenges and opportunities for sustainable development. *MINISTERIO DE EDUCACIÓN* 2019.

2 Liao L, Patterson DJ, Fox D, Kautz H. Learning and inferring transportation routines. *Artificial Intelligence* 2007; **171**: 311–31.

3 Jha SK, Bilalovic J, Jha A, Patel N, Zhang H. Renewable energy: Present research and future scope of Artificial Intelligence. *Renewable and Sustainable Energy Reviews* 2017; **77**: 297–317.

4 Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nature Reviews Cancer* 2018; **18**: 500–10.

5 Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019; **16**: 703–15.

6 Ting DSW, Pasquale LR, Peng L, *et al.* Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019; **103**: 167.

7 Artificial Intelligence. Oxford Dictionary. https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095426960 (accessed Feb 9, 2020).

8 Bishop CM. Pattern recognition and machine learning. Springer, 2006.

9 Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine* 2020; : 1–8.

10 International Skin Imaging Collaboration (ISIC) Archive. www.isic-archive.com.

11 Rogers HW, Weinstock MA, Harris AR, *et al.* Incidence estimate of nonmelanoma skin cancer in the United States, 2006. *Arch Dermatol* 2010; **146**: 283–7.

12 Rogers HW, Weinstock MA, Feldman SR, Coldiron BM. Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the US Population, 2012. *JAMA Dermatol* 2015; **151**: 1081–6.

13 Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. *CA Cancer J Clin* 2018; **68**: 7–30.

14 Society AC. Cancer Facts & Figures 2016. httpswww.cancer.orgcontentdamcancer-orgresearchcancer-facts-and-statisticsannual-cancer-facts-and-figurescancer-facts-and-figures-.pdf.

15 Pasquali S, Hadjinicolaou AV, Chiarion-Sileni V, Rossi CR, Mocellin S. Systemic treatments for metastatic cutaneous melanoma. *Cochrane Database of Systematic Reviews* 2018; **85**: 1979–338.

16 Ugurel S, Röhmel J, Ascierto PA, *et al.* Survival of patients with advanced metastatic melanoma: the impact of novel therapies-update 2017. *European Journal of Cancer* 2017; **83**: 247–57.

17 Xie P, Lefrançois P. Efficacy, safety, and comparison of sonic hedgehog inhibitors in basal cell carcinomas: A systematic review and meta-analysis. *Journal of the American Academy of Dermatology* 2018; **79**: 1089–1100.e17.

18 Lear JT, Basset-Seguin N, Kaatz M, *et al.* Treatment patterns and outcomes for patients with locally advanced basal cell carcinoma before availability of Hedgehog pathway inhibitors: a retrospective chart review. *Eur J Dermatol* 2017; **27**: 386–92.

19 Coit DG, Thompson JA, Albertini MR, *et al.* Cutaneous Melanoma, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2019; **17**: 367–402.

20 MD SMS, PhD HTM, MD CKB, *et al.* Guidelines of care for the management of primary cutaneous melanoma. *Journal of the American Academy of Dermatology* 2019; **80**: 208–50.

21 Gershenwald JE, Scolyer RA, Hess KR, *et al.* Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin* 2017; **67**: 472–92.

22 Dinnes J, Deeks JJ, Chuchu N, *et al.* Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. *Cochrane Database of Systematic Reviews* 2018.

23 Dinnes J, Deeks JJ, Chuchu N, *et al.* Visual inspection and dermoscopy, alone or in combination, for diagnosing keratinocyte skin cancers in adults. *Cochrane Database Syst Rev* 2018; **12**: CD011901.

24 Zalaudek I, Lallas A, Moscarella E, Longo C, Soyer HP, Argenziano G. The dermatologist's stethoscope-traditional and new applications of dermoscopy. *Dermatol Pract Concept* 2013; **3**: 67–71.

25 Errichetti E, Zalaudek I, Kittler H, *et al.* Standardization of dermoscopic terminology and basic dermoscopic parameters to evaluate in general dermatology (non-neoplastic dermatoses): an expert consensus on behalf of the International Dermoscopy Society. *British Journal of Dermatology* 2019; **19**: 56–14.

26 Codella N, Nguyen Q-B, Pankanti S, *et al.* Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. arXiv. 2016; **cs.CV**.

27 Gutman D, Codella NCF, Celebi E, *et al.* Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). 2016; published online May 3.

28 Codella NCF, Gutman D, Celebi ME, *et al.* Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). IEEE, 2018: 168–72.

29 Codella N, Rotemberg V, Tschandl P, *et al.* Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). 2019; published online Feb 9.

30 Michael A Marchetti MD, Konstantinos Liopyris MD, Stephen W Dusza D, *et al.* Computer Algorithms Show Potential for Improving Dermatologists' Accuracy to Diagnose Cutaneous Melanoma; Results of ISIC 2017. *Journal of the American Academy of Dermatology* 2019; : 1–18.

31 Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated Dermatological Diagnosis: Hype or Reality? *J Invest Dermatol* 2018; **138**: 2277–9.

32 Cristian Navarrete-Dechent MD, Konstantinos Liopyris MD, Michael A Marchetti MD. Multiclass artificial intelligence in dermatology - progress but still room for improvement. *J Invest Dermatol* 2020; : 1–15.

33 MD MAM, MD KL, MD CN-D. Net benefit and decision curve analysis of competing diagnostic strategies for cutaneous melanoma. *Journal of the American Academy of Dermatology* 2020; : 1–2.

34 Liopyris K, Navarrete-Dechent C, Mancebo SE, *et al.* Dermoscopic Appearance of Amelanotic Volar Melanoma Compared With Volar Angioma. *JAMA Dermatol* 2019; **155**: 500–2.

35 Liopyris K, Navarrete-Dechent C, Yélamos O, Marchetti MA, Rabinovitz H, Marghoob AA. Clinical, dermoscopic and reflectance confocal microscopy characterization of facial basal cell carcinomas presenting as small white lesions on sun-damaged skin. *British Journal of Dermatology* 2019; **180**: 229–30.

36 Liopyris K, Navarrete-Dechent C, Dusza SW, *et al.* Clinical and dermoscopic features associated with lichen planus-like keratoses that undergo skin biopsy: A single-center, observational study. *Australas J Dermatol* 2019; **60**: e119–26.

37 Navarrete-Dechent C, Liopyris K, Rishpon A, *et al.* Association of Multiple Aggregated Yellow-White Globules With Nonpigmented Basal Cell Carcinoma. *JAMA Dermatol* 2020; **156**: 882–9.

38 Marchetti MA, Codella NCF, Dusza SW, *et al.* Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *Journal of the American Academy of Dermatology* 2018; **78**: 270–1.

39 Kittler H, Marghoob AA, Argenziano G, *et al.* Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the International Society of Dermoscopy. *Journal of the American Academy of Dermatology* 2016; **74**: 1093–106.

40 Wolner Z, Yélamos O, Liopyris K, Rogers T, Marchetti M, A Marghoob A. Enhancing Skin Cancer Diagnosis with Dermoscopy.

41 Yélamos O, Braun RP, Liopyris K, *et al.* Dermoscopy and dermatopathology correlates of cutaneous neoplasms. *Journal of the American Academy of Dermatology* 2019; **80**: 341–63.

42 SLIC Superpixels. 2010; : 1–15.

43 Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans Pattern Anal Mach Intell* 2012; **34**: 2274–82.

44 Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans Pattern Anal Mach Intell* 2012; **34**: 2274–82.

45 1703.05165 YYAPA, 2017. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. *arxivorg*

.

46 Kawahara J, Hamarneh G. Fully Convolutional Neural Networks to Detect Clinical Dermoscopic Features. *IEEE Journal of Biomedical and Health Informatics* 2019; **23**: 578–85.

47 li Y, Shen L. Skin Lesion Analysis Towards Melanoma Detection Using Deep Learning Network. arXiv. 2017; **cs.CV**.

48 Matsunaga K, Hamada A, Minagawa A, Koga H. Image Classification of Melanoma, Nevus and Seborrheic Keratosis by Deep Neural Network Ensemble. arXiv. 2017; **cs.CV**.

49 Díaz IG. Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for the Diagnosis of Skin Lesions. 2017; published online March 6.

50 Menegola A, Tavares J, Fornaciali M, Li LT, Avila S, Valle E. RECOD Titans at ISIC Challenge 2017. arXiv. 2017; **cs.CV**.

51 Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol* 2017; **10**: 257–73.

52 Hoo ZH, Candlish J, Teare D. What is an ROC curve? *Emerg Med J* 2017; **34**: 357–9.

53 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 1988; **44**: 837–45.

54 Haenssle HA, Fink C, Schneiderbauer R, *et al.* Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 2018; **29**: 1836–42.

55 Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542**: 115–8.

56 Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *J Invest Dermatol* 2018; **138**: 1529–38.

57 Monheit G, Cognetta AB, Ferris L, *et al.* The performance of MelaFind: a prospective multicenter study. *Arch Dermatol* 2011; **147**: 188–94.

58 Tschandl P, Codella N, Akay BN, *et al.* Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol* 2019; **20**: 938–47.

59 Han SS, Park I, Chang SE, *et al.* Augmented Intelligence Dermatology: Deep Neural Networks Empower Medical Professionals in Diagnosing Skin Cancer and Predicting Treatment Options for 134 Skin Disorders. *J Invest Dermatol* 2020; : 1–9.

60 PhD LKFM, MS JAH, MS BG, *et al.* Computer-aided classification of melanocytic lesions using dermoscopic images. *Journal of the American Academy of Dermatology* 2015; **73**: 769–76.

61 Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science* 2019; **363**: 1287–9.

62 Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated Classification of Skin Lesions: From Pixels to Practice. *J Invest Dermatol* 2018; **138**: 2108–10.

63 Winkler JK, Fink C, Toberer F, *et al.* Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatol* 2019; **155**: 1135–41.

64 Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatol* 2018; **154**: 1247–2.

65 Lui H, Zhao J, McLean D, Zeng H. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. *Cancer Research* 2012; **72**: 2491–500.

66 Cukras AR. On the Comparison of Diagnosis and Management of Melanoma Between Dermatologists and MelaFind. *JAMA Dermatol* 2013; **149**: 622–3.

67 Wells R, Gutkowicz-Krusin D, Veledar E, Toledano A, Chen SC. Comparison of diagnostic and management sensitivity to melanoma between dermatologists and MelaFind: a pilot study. *Arch Dermatol* 2012; **148**: 1083–4.

68 Winkelmann RR, Farberg AS, Glazer AM, Rigel DS. Noninvasive Technologies for the Diagnosis of Cutaneous Melanoma. *Dermatol Clin* 2017; **35**: 453–6.

69 A Nikolas MacLellan MD, Emma L Price M, Pamela Publicover-Brouwer RN, *et al.* The Use of Non-Invasive Imaging Techniques in the Diagnosis of Melanoma: A Prospective Diagnostic Accuracy Study. *Journal of the American Academy of Dermatology* 2020; : 1–26.

70 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**: i6.

71 Vickers AJ, Ben Calster, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. 2019; : 1–8.

72 Dinnes J, Deeks JJ, Saleh D, *et al.* Reflectance confocal microscopy for diagnosing cutaneous melanoma in adults. *Cochrane Database of Systematic Reviews* 2018; **170**: 802–5.

73 Menzies SW, Westerhoff K, Rabinovitz H, Kopf AW, McCarthy WH, Katz B. Surface Microscopy of Pigmented Basal Cell Carcinoma. *Arch Dermatol* 2000; **136**: 1012–6.

74 Menzies SW, Kreusch J, Byth K, *et al.* Dermoscopic evaluation of amelanotic and hypomelanotic melanoma. *Arch Dermatol* 2008; **144**: 1120–7.

75 Argenziano G, Soyer HP, Chimenti S, *et al.* Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. *Journal of American Dermatology* 2003; **48**: 679–93.

76 Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol* 2002; **3**: 159–65.

77 Zalaudek I, Giacomel J, Schmid K, *et al.* Dermatoscopy of facial actinic keratosis, intraepidermal carcinoma, and invasive squamous cell carcinoma: A progression model. *Journal of the American Academy of Dermatology* 2012; **66**: 589–97.

78 Zalaudek I, Giacomel J, Argenziano G, *et al.* Dermoscopy of facial nonpigmented actinic keratosis. *British Journal of Dermatology* 2006; **155**: 951–6.

79 Zalaudek I, Schmid K, Marghoob AA, *et al.* Frequency of dermoscopic nevus subtypes by age and body site: a cross-sectional study. *Arch Dermatol* 2011; **147**: 663–70.

80 Lallas A, Tzellos T, Kyrgidis A, *et al.* Accuracy of dermoscopic criteria for discriminating superficial from other subtypes of basal cell carcinoma. *Journal of the American Academy of Dermatology* 2014; **70**: 303–11.

81 Lallas A, Tschandl P, Kyrgidis A, *et al.* Dermoscopic clues to differentiate facial lentigo maligna from pigmented actinic keratosis. *British Journal of Dermatology* 2016; **174**: 1079–85.

82 Zaballos P, Martí E, Cuéllar F, Puig S, Malvehy J. Dermoscopy of Lichenoid Regressing Seborrheic Keratosis. *Arch Dermatol* 2006; **142**: 1–1.

83 Zaballos P, Puig S, Llambrich A, Malvehy J. Dermoscopy of dermatofibromas: a prospective morphological study of 412 cases. *Arch Dermatol* 2008; **144**: 75–83.

84 Zaballos P, Ara M, Puig S, Malvehy J. Clinical and dermoscopic image of an intermediate stage of regressing seborrheic keratosis in a lichenoid keratosis. *Dermatol Surg* 2005; **31**: 102–3.

85 Rosendahl C, Cameron A, Argenziano G, Zalaudek I, Tschandl P, Kittler H. Dermoscopy of squamous cell carcinoma and keratoacanthoma. *Arch Dermatol* 2012; **148**: 1386–92.

86 Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of Computer-Aided Diagnosis of Melanoma. *JAMA Dermatol* 2019; **155**: 1291–9.

87 Bradford PT, Goldstein AM, McMaster ML, Tucker MA. Acral lentiginous melanoma: incidence and survival patterns in the United States, 1986-2005. *Arch Dermatol* 2009; **145**: 427–34.

88 Soon SL, Solomon ARJ, Papadopoulos D, Murray DR, McAlpine B, Washington CV. Acral lentiginous melanoma mimicking benign disease: the Emory experience. *Journal of the American Academy of Dermatology* 2003; **48**: 183–8.

89 Özdemir F, Errico MA, Yaman B, Karaarslan I. Acral lentiginous melanoma in the Turkish population and a new dermoscopic clue for the diagnosis. 2018; **8**: 140–8.

90 Phan A, Dalle S, Marcilly M-C, Bergues J-P, Thomas L. Benign dermoscopic parallel ridge pattern variants. *Arch Dermatol* 2011; **147**: 634.

91 Freites-Martinez A, Moreno-Torres A, Núñez AH, Martinez-Sanchez D, Huerta-Brogeras M, Borbujo J. Angioma serpiginosum: report of an unusual acral case and review of the literature. *An Bras Dermatol* 2015; **90**: 26–8.

92 Saida T, Miyazaki A, Oguchi S, *et al.* Significance of dermoscopic patterns in detecting malignant melanoma on acral volar skin: results of a multicenter study in Japan. *Arch Dermatol* 2004; **140**: 1233–8.

93 Braun RP, Thomas L, Dusza SW, *et al.* Dermoscopy of Acral Melanoma: A Multicenter Study on Behalf of the International Dermoscopy Society. 2013; **227**: 373–80.

94 Saida T, Oguchi S, Miyazaki A. Dermoscopy for acral pigmented skin lesions. *Clin Dermatol* 2002; **20**: 279–85.

95 Lomas A, Leonardi-Bee J, Bath-Hextall F. A systematic review of worldwide incidence of nonmelanoma skin cancer. *British Journal of Dermatology* 2012; **166**: 1069–80.

96 Navarrete-Dechent C, Bajaj S, Marchetti MA, Rabinovitz H, Dusza SW, Marghoob AA. Association of Shiny White Blotches and Strands With Nonpigmented Basal Cell Carcinoma. *JAMA Dermatol* 2016; **152**: 546–7.

97 Reiter O, Mimouni I, Gdalevich M, *et al.* The diagnostic accuracy of dermoscopy for basal cell carcinoma: A systematic review and meta-analysis. *Journal of the American Academy of Dermatology* 2019; **80**: 1380–8.

98 Sinz C, Tschandl P, Rosendahl C, *et al.* Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. *Journal of the American Academy of Dermatology* 2017; **77**: 1100–9.

99 Banuls J, Arribas P, Berbegal L, DeLeon FJ, Frances L, Zaballos P. Yellow and orange in cutaneous lesions: clinical and dermoscopic data. *Journal of the European Academy of Dermatology and Venereology* 2015; **29**: 2317–25.

100 Bellucci C, Arginelli F, Bassoli S, Magnoni C, Seidenari S. Dermoscopic yellow structures in basal cell carcinoma. *Journal of the European Academy of Dermatology and Venereology* 2014; **28**: 651–4.

101 Sahu A, Yélamos O, Iftimia N, *et al.* Evaluation of a Combined Reflectance Confocal Microscopy–Optical Coherence Tomography Device for Detection and Depth Assessment of Basal Cell Carcinoma. *JAMA Dermatol* 2018; **154**: 1175–10.

102 Navarrete-Dechent C, DeRosa AP, Longo C, *et al.* Reflectance confocal microscopy terminology glossary for nonmelanocytic skin lesions: A systematic review. *Journal of the American Academy of Dermatology* 2019; **80**: 1414–1427.e3.

103 Navarrete-Dechent C, Liopyris K, Cordova M, Busam KJ, Marghoob AA, Chen C-SJ. Reflectance Confocal Microscopic and En Face Histopathologic Correlation of the Dermoscopic 'Circle Within a Circle' in Lentigo Maligna. *JAMA Dermatol* 2018; **154**: 1092–4.

104 Altamura D, Menzies SW, Argenziano G, *et al.* Dermatoscopy of basal cell carcinoma: Morphologic variability of global and local features and accuracy of diagnosis. *Journal of American Dermatology* 2010; **62**: 67–75.

105 Longo C, Lallas A, Kyrgidis A, *et al.* Classifying distinct basal cell carcinoma subtype by means of dermatoscopy and reflectance confocal microscopy. *Journal of American Dermatology* 2014; **71**: 716–724.e1.

106 Haws AL, Rojano R, Tahan SR, Phung TL. Accuracy of biopsy sampling for subtyping basal cell carcinoma. *Journal of the American Academy of Dermatology* 2012; **66**: 106–11.

107 Slodkowska EA, Cribier B, Peltre B, Jones DM, Carlson JA. Calcifications Associated With Basal Cell Carcinoma: Prevalence, Characteristics, and Correlations. *Am J Dermatopathol* 2010; **32**.

108 Walsh JS, Perniciaro C, Randle HW. Calcifying Basal Cell Carcinomas. *Dermatol Surg* 1999; **25**.

109 Carrera C, Marchetti MA, Dusza SW, *et al.* Validity and Reliability of Dermoscopic Criteria Used to Differentiate Nevi From Melanoma. *JAMA Dermatol* 2016; **152**: 798–20.

110 Achanta R, Shaji A, Smith K, Lucchi A, Fua P. Slic superpixels (No. EPFL-REPORT-149300). 2010.

111 Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Med Res Methodol* 2013; **13**: 61.

112 Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 1977; **33**: 159–74.

113 Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology* 1945; **26**: 297–302.

114 Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Transactions on Medical Imaging*; **13**: 716–24.

115 neuroelf/isicarchive. https://github.com/neuroelf/isicarchive/blob/master/EASY - MARk-SUP - analyses.ipynb.

116 Tschandl P, Rinner C, Apalla Z, *et al.* Human–computer collaboration for skin cancer recognition. *Nat Med* 2020; : 1–13.

117 International Agency for Research on Cancer. Cancer Today: Population fact sheets. 2018.

118 Liopyris K, Navarrete-Dechent C, Yélamos O, Marchetti MA, Rabinovitz H, Marghoob AA. Clinical, dermoscopic and reflectance confocal microscopy characterization of facial basal cell carcinomas presenting as small white lesions on sun-damaged skin. *British Journal of Dermatology* 2019; **180**: 229–30.

119 Brady MS, Oliveria SA, Christos PJ, *et al.* Patterns of detection in patients with cutaneous melanoma. *Cancer* 2000; **89**: 342–7.

120 Carli P, De Giorgi V, Palli D, *et al.* Dermatologist detection and skin self-examination are associated with thinner melanomas: results from a survey of the Italian Multidisciplinary Group on Melanoma. *Arch Dermatol* 2003; **139**: 607–12.

121 Epstein DS, Lange JR, Gruber SB, Mofid M, Koch SE. Is physician detection associated with thinner melanomas? *JAMA* 1999; **281**: 640–3.

122 Schwartz JL, Wang TS, Hamilton TA, Lowe L, Sondak VK, Johnson TM. Thin primary cutaneous melanomas: associated detection patterns, lesion characteristics, and patient characteristics. *Cancer* 2002; **95**: 1562–8.

123 Körner A, Coroiu A, Martins C, Wang B. Predictors of skin self-examination before and after a melanoma diagnosis: the role of medical advice and patient's level of education. *Int Arch Med* 2013; **6**: 8.

124 Avilés-Izquierdo JA, Molina-López I, Rodríguez-Lomba E, Marquez-Rodas I, Suarez-Fernandez R, Lazaro-Ochaita P. Who detects melanoma? Impact of detection patterns on characteristics and prognosis of patients with melanoma. *Journal of the American Academy of Dermatology* 2016; **75**: 967–74.

125 Lakhani NA, Saraiya M, Thompson TD, King SC, Guy GPJ. Total body skin examination for skin cancer screening among U.S. adults from 2000 to 2010. *Prev Med* 2014; **61**: 75–80.

126 Argenziano G, Cerroni L, Zalaudek I, *et al.* Accuracy in melanoma detection: a 10-year multicenter survey. *Journal of the American Academy of Dermatology* 2012; **67**: 54–9.

127 Howlader N, Noone AM, Krapcho M, *et al.* SEER Cancer Statistics Review, 1975-2016, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2016/, based on November 2018 SEER data submission, posted to the SEER web site, April 2019.

128 PhD VK, PhD JEB, PhD CDM, PhD GL, PhD KF, FMedSci PME. Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble. *The Lancet* 2017; **389**: 1323–35.

129 Weinstock MA, Lott JP, Wang Q, *et al.* Skin biopsy utilization and melanoma incidence among Medicare beneficiaries. *British Journal of Dermatology* 2017; **176**: 949–54.

130 Welch HG, Kramer BS, Black WC. Epidemiologic Signatures in Cancer. *New England Journal of Medicine* 2019; **381**: 1378–86.

131 Welch HG, Black WC. Overdiagnosis in cancer. *JNCI Journal of the National Cancer Institute* 2010; **102**: 605–13.

132 Linos E, Chren MM. Is screening for basal cell carcinoma worthwhile? Too soon to tell. *British Journal of Dermatology* 2016; **174**: 1181–2.

133 Linos E, Parvataneni R, Stuart SE, Boscardin WJ, Landefeld CS, Chren M-M. Treatment of Nonfatal Conditions at the End of Life. *JAMA Intern Med* 2013; **173**: 1006–12.

134 Hofmann BM. Too much technology. *BMJ : British Medical Journal* 2015; **350**: h705.

135 Lallas A, Argenziano G. Artificial intelligence and melanoma diagnosis: ignoring human nature may lead to false predictions. *Dermatol Pract Concept* 2018; : 249–51.

136 Argenziano G, Soyer HP. Dermoscopy of pigmented skin lesions--a valuable tool for early diagnosis of melanoma. *Lancet Oncol* 2001; **2**: 443–9.

137 Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. 2002; **3**: 159–65.