

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



-Επεξηγούμενη τεχνητή νοημοσύνη (Explainable Artificial Intelligence - xAI) σε
προβλήματα στον τομέα της Βιοϊατρικής-

Μεταξάς Κωνσταντίνος

Επιβλέπων: -Αριστείδης Βραχάτης-

2024

Επιβλέπων καθηγητής

Αριστείδης Βραχάτης, *Επίκουρος Καθηγητής*

Ιόνιο Πανεπιστήμιο

Τριμελής επιτροπή

Αριστείδης Βραχάτης, *Επίκουρος Καθηγητής*

Ιόνιο Πανεπιστήμιο

Θεμιστοκλής Έξαρχος, *Αναπληρωτής Καθηγητής*

Ιόνιο Πανεπιστήμιο

Παναγιώτης Βλάμος, *Καθηγητής*

Ιόνιο Πανεπιστήμιο

Αυτή η προσπάθεια αφιερώνεται στην οικογένειά μου

Περίληψη

Σκεπός της παρούσας διπλωματικής εργασίας είναι η μελέτη των εφαρμογών της επεξηγούμενης τεχνητής νοημοσύνης σε βιοϊατρικά προβλήματα, η εφαρμογή αλγορίθμων βασισμένους σε δένδρα όπως ο XGBoost, CatBoost και Random Forest μέσα από τις αρχές της επεξηγούμενης τεχνητής νοημοσύνης, ο εντοπισμός σετ χαρακτηριστικών (features) που είναι κυρίαρχο σε ένα σετ δεδομένων, η εφαρμογή διαδικασιών Μηχανικής Μάθησης στο σετ χαρακτηριστικών (features) που είναι κυρίαρχο σε ένα σετ δεδομένων και τέλος η ερμηνεία των αποτελεσμάτων των μεθοδολογιών αυτών σε ένα πραγματικό βιοϊατρικό πρόβλημα.

Πίνακας περιεχομένων

Περίληψη	4
Εισαγωγή	7
ΚΕΦΑΛΑΙΟ 1. ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ	9
Ιστορική προέλευση της ανάγκης για εξηγήσιμη τεχνητή νοημοσύνη	11
ΚΕΦΑΛΑΙΟ 2. Η ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΒΙΟΪΑΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ (ΧΑΙ)	15
Κύριοι Βιοϊατρικοί στόχοι του ΧΑΙ	17
ΧΑΙ σχεδιασμένο για μη προγραμματιστές	18
Κύριοι Βιοϊατρικοί στόχοι του ΧΑΙ	20
Αξιοπιστία	21
ΠΕΡΑΙΤΕΡΩ ΚΑΙ ΣΥΝΑΦΕΙΣ ΣΤΟΧΟΙ ΤΗΣ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ	21
Διαφάνεια	23
Κατανοητότητα	24
Πληροφόρηση	24
Προσβασιμότητα	25
Μεταβιβασιμότητα	26
Συντονισμένη ερμηνεία ΑΙ μεταξύ εμπειρογνομόνων πληροφορικής και βιοϊατρικού τομέα	27

ΧΑΙ στις βιοϊατρικές εκδόσεις	28
Πρακτικές επιπτώσεις του ΧΑΙ στην υγειονομική περίθαλψη	30
Ανίχνευση και μετριάσμός μεροληψίας	30
Μελέτες περιπτώσεων και εμπειρικά στοιχεία	31
Παραδείγματα χρήσης ΧΑΙ στην ιατρική επιστήμη	31
Συμβατική απεικόνιση/ακτίνες	32
Υπολογιστική Τομογραφία (CT)	36
Απεικόνιση μαγνητικού συντονισμού (MRI)	39
Αρχή φόρμας	44
Οπτική απεικόνιση	44
Δερματολογία	44
Διαβητική αμφιβληστροειδοπάθεια	46
Καρκίνος	49
Κλινικά/παρατηρητικά δεδομένα	51
Ανίχνευση καρκίνου	52
Προκλήσεις και μελλοντικές κατευθύνσεις	54
Κατανόηση και βελτίωση της επεξηγηματικότητας	55
ΧΑΙ με επίκεντρο τα δεδομένα	56
Αρχή φόρμας	57
ΧΑΙ με επίκεντρο το μοντέλο	57
Human-in-the-loop	58
ΒΙΒΛΙΟΓΡΑΦΙΑ	60

Εισαγωγή

Η τεχνητή νοημοσύνη αναφέρεται σε έναν τομέα της επιστήμης των υπολογιστών που επιδιώκει την αυτοματοποίηση ανθρώπινων δραστηριοτήτων που συνήθως θεωρούνται δείγματα ευφυούς ανθρώπινης συμπεριφοράς (Luger, 2004). Οι μέθοδοι αυτές περιλαμβάνουν την κατανόηση της ανθρώπινης γλώσσας, την αναπαράσταση και τη χρήση γνώσεων, τη λογική σκέψη, τον προγραμματισμό, την επίλυση προβλημάτων και την εκτίμηση κινδύνων, συμπεριλαμβανομένης της εικασίας και της μάθησης από την εμπειρία. Η μηχανική μάθηση είναι σήμερα μακράν η πιο δημοφιλής μέθοδος που χρησιμοποιείται στην τεχνητή νοημοσύνη και μπορεί να αναφερθεί σε δύο διαφορετικές μορφές (Lötsch and Ultsch, 2017): πρώτον, προσεγγίσεις στις οποίες μια κατηγορία πολύ γενικών αλγορίθμων (τεχνητά νευρωνικά δίκτυα, ταξινομητές, προγνωστικοί μηχανισμοί, συσχετιστικές μνήμες κ.λπ.) συντονίζονται με βάση παραδείγματα για τη βελτιστοποίηση της πρόβλεψης ή της ταξινόμησης νέων, αθέατων περιπτώσεων. Σε αυτήν την περίπτωση, οι τεχνικές που ανιχνεύουν δομές στα δεδομένα και περιγράφουν αυτές τις δομές με έναν τρόπο που επιτρέπει στην πληροφορία να χρησιμοποιηθεί για τη σωστή ταξινόμηση νέων παραδειγμάτων και να είναι κατανοητή από ανθρώπους, αποτελούν τη διαδικασία εξαγωγής γνώσης από τα δεδομένα (Murphy, 2014; Dhar, 2013).

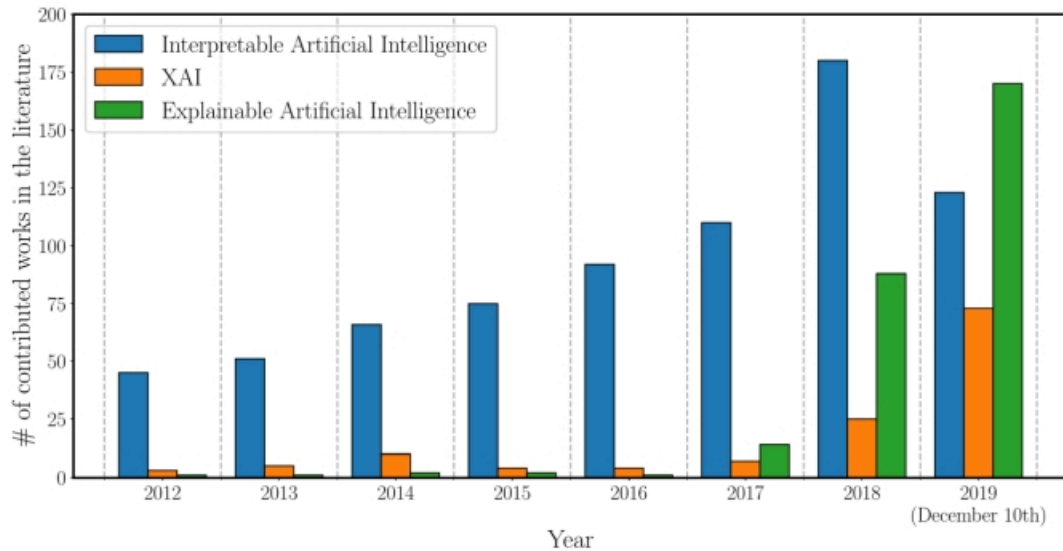
Η αυξανόμενη σημασία της τεχνητής νοημοσύνης και των αλγορίθμων μηχανικής μάθησης στη βιοϊατρική έρευνα επιφέρει αυξημένη επίδραση στις κλινικές διαδικασίες λήψης αποφάσεων. Αυτό έχει άμεσες επιπτώσεις στην ιατρική πρακτική και στην επικοινωνία μεταξύ ιατρών και ασθενών καθώς η αλληλεπίδραση μεταξύ γιατρού-ασθενούς έχει βάσεις στην προσωπική εμπιστοσύνη των ιατρικών συμβουλών και των θεραπευτικών αποφάσεων. Ωστόσο, εάν η διάγνωση ή η επιλογή της πιο υποσχόμενης θεραπείας δεν γίνεται πλέον αποκλειστικά από τον ιατρό, αλλά σε μεγάλο βαθμό από ένα σύστημα με αλγόριθμους μάθησης και τεχνητή νοημοσύνη, τότε οι αποφάσεις γίνονται λιγότερο διαφανείς. Οι γιατροί δεν μπορεί να θεωρηθεί ότι διαθέτουν τις απαραίτητες γνώσεις πληροφορικής για να κατανοήσουν τη διαδικασία λήψης αποφάσεων ενός αλγορίθμου και θα πρέπει να είναι σε θέση να επικοινωνούν τη διαδικασία αυτή με όλες τις σχετικές λεπτομέρειες στους ασθενείς τους με κατανοητό τρόπο.

Η Ευρωπαϊκή Ένωση (ΕΕ) έχει αναγνωρίσει το πρόβλημα που δημιουργεί η λήψη ιατρικών αποφάσεων βάσει αλγορίθμων για τα δικαιώματα ενημέρωσης των θιγόμενων ασθενών και έχει δημοσιεύσει ένα έγγραφο-ορόσημο στο οποίο επισημαίνεται η ανάγκη για επεξηγήσεις των αποφάσεων που λαμβάνονται μέσω υπολογιστή, ώστε να μπορούν να κοινοποιούνται στους θιγόμενους ασθενείς με κατανοητό τρόπο (Hamon et al., 2020). Η λύση βρίσκεται στην έννοια της εξηγήσιμης ΤΝ (XAI), η οποία προσελκύει ολοένα και μεγαλύτερο επιστημονικό ενδιαφέρον (Arrieta et al., 2020). Αυτό συνάδει με τις προσπάθειες του αμερικανικού στρατού να αποκτήσει εξηγήσιμα μοντέλα που καθιστούν διαφανείς τις αποφάσεις που λαμβάνονται από αυτόνομα συστήματα (<https://www.darpa.mil/program/explainable-artificial-intelligence> (Turek, 2016), πρόσβαση στις 15 Δεκεμβρίου 2021). Χωρίς την απαραίτητη κατανόηση, η μηχανική μάθηση βασίζεται στη διαδικασία δοκιμής και σφάλματος και έχει συγκριθεί με τους μεσαιωνικούς αλχημιστές (Hutson, 2018).

ΚΕΦΑΛΑΙΟ 1. ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Η Τεχνητή Νοημοσύνη (TN) αναδεικνύεται ως κεντρικός παράγοντας σε πολλούς τομείς δραστηριότητας που έχουν υιοθετήσει τις νέες τεχνολογίες πληροφοριών (Russell et al., 2016). Παρότι οι αρχές της Τεχνητής Νοημοσύνης έχουν αποκτήσει ρίζες εδώ και πολλές δεκαετίες, υπάρχει γενική συμφωνία όσον αφορά την ουσιαστική σημασία που έχουν σήμερα οι εξελιγμένες μηχανές που είναι ικανές για μάθηση, συλλογισμό και προσαρμογή. Με τις ικανότητές τους, οι αλγόριθμοι TN επιτυγχάνουν εκπληκτικές αποδόσεις στην επίλυση όλο και πιο σύνθετων υπολογιστικών προβλημάτων, γεγονός που καθιστά αυτήν την τεχνολογία ζωτικής σημασίας για τη μελλοντική πρόοδο της ανθρώπινης κοινωνίας (West, 2018). Η πολυπλοκότητα των συστημάτων που χρησιμοποιούν τεχνητή νοημοσύνη έχει αυξηθεί πρόσφατα σε τέτοιο βαθμό, ώστε σχεδόν καμία ανθρώπινη παρέμβαση δεν απαιτείται για τον σχεδιασμό και την ανάπτυξή τους. Έτσι υπάρχει μια αυξανόμενη ανάγκη για την κατανόηση του τρόπου με τον οποίο οι αποφάσεις αυτές ενισχύονται με μεθόδους τεχνητής νοημοσύνης (Goodman et al., 2017).

Ενώ τα πρώτα συστήματα τεχνητής νοημοσύνης ήταν εύκολα ερμηνεύσιμα, τα τελευταία χρόνια παρατηρείται η άνοδος αδιαφανών συστημάτων λήψης αποφάσεων, όπως τα Βαθιά Νευρωνικά Δίκτυα (DNN). Η επιτυχία των μοντέλων βαθιάς μάθησης (DL), όπως τα Δικτυακά Νευρωνικά Δίκτυα (DNNs), προέρχεται από τη συνδυαστική δράση αποτελεσματικών αλγορίθμων μάθησης και του παραμετρικού τους χώρου (Castelvecchi, 2016). Το αντίθετο του μαύρου κουτιού είναι η διαφάνεια, δηλαδή η προσπάθεια να κατανοήσουμε απευθείας τους μηχανισμούς λειτουργίας ενός μοντέλου (Lipton, 2018).



Εικόνα 1: Αυξητική τάση των συνεισφορών για το XAI και τις συναφείς έννοιες. Εξέλιξη του αριθμού των συνολικών δημοσιεύσεων των οποίων ο τίτλος, η περίληψη ή/και οι λέξεις-κλειδιά αναφέρονται στον τομέα του XAI κατά τα τελευταία έτη. Δεδομένα που ανακτήθηκαν από το Scopus® (10 Δεκεμβρίου 2019) με τη χρήση των όρων αναζήτησης που αναφέρονται στο υπόμνημα κατά την αναζήτηση σε αυτή τη βάση δεδομένων. Είναι ενδιαφέρον να σημειωθεί η λανθάνουσα ανάγκη για ερμηνεύσιμα μοντέλα TN με την πάροδο του χρόνου (η οποία ανταποκρίνεται στη διαίσθηση, καθώς η ερμηνεία αποτελεί απαίτηση σε πολλά σενάρια), ωστόσο μόλις το 2017 το ενδιαφέρον για τεχνικές επεξήγησης των μοντέλων TN διεισδύσε σε όλη την ερευνητική κοινότητα (Barredo Arrieta et al., 2010).

Καθώς τα μοντέλα μηχανικής μάθησης (ML) μαύρου κουτιού χρησιμοποιούνται όλο και περισσότερο για να κάνουν σημαντικές προβλέψεις σε κρίσιμα πλαίσια, η απαίτηση για διαφάνεια αυξάνεται από τους διάφορους ενδιαφερόμενους φορείς της TN (Preece et al., 2018). Μεγάλη επικυδινότητα υπάρχει ως προς το να χρησιμοποιηθούν αποφάσεις που δεν είναι δικαιολογημένες, νόμιμες ή που απλώς δεν επιτρέπουν την απόκτηση λεπτομερών εξηγήσεων για τη συμπεριφορά τους (Gunning, 2017). Σύμφωνα με τους Tjoa and Guan, (2019) «οι εξηγήσεις που υποστηρίζουν την έξοδο ενός μοντέλου είναι ζωτικής σημασίας, π.χ. στην ιατρική ακριβείας, όπου οι ειδικοί απαιτούν πολύ περισσότερες πληροφορίες από το μοντέλο από μια απλή δυαδική πρόβλεψη για την υποστήριξη της διάγνωσής τους». Άλλα παραδείγματα περιλαμβάνουν, μεταξύ άλλων, τα αυτόνομα οχήματα στις μεταφορές, την ασφάλεια και τα χρηματοοικονομικά.

Σε γενικές γραμμές, οι άνθρωποι είναι επιφυλακτικοί στην υιοθέτηση τεχνικών που δεν είναι άμεσα ερμηνεύσιμες, εφαρμόσιμες και αξιόπιστες (Zhu et al., 2018), δεδομένης της αυξανόμενης ζήτησης για ηθική τεχνητή νοημοσύνη (Goodman et al., 2017). Αυτό είναι αληθές υπό την έννοια ότι υπάρχει ένας συμβιβασμός μεταξύ της

απόδοσης ενός μοντέλου και της διαφάνειάς του (Došilović et al., 2018). Η βελτίωση της κατανόησης ενός συστήματος μπορεί να οδηγήσει στη διόρθωση των ελλείψεων του. Κατά την ανάπτυξη ενός μοντέλου ML, η εξέταση της ερμηνείας ως πρόσθετου οδηγού σχεδίασης μπορεί να βελτιώσει την υλοποίησή του για 3 λόγους:

- Η ερμηνεία συμβάλλει στη διασφάλιση της αμεροληψίας κατά τη λήψη αποφάσεων, δηλαδή στον εντοπισμό και, κατά συνέπεια, τη διόρθωση της μεροληψίας στο σύνολο δεδομένων εκπαίδευσης.
- Η ερμηνεία διευκολύνει την παροχή ευρωστίας με την επισήμανση πιθανών αντίξοων διαταραχών που θα μπορούσαν να αλλάξουν την πρόβλεψη.
- Η ερμηνεία μπορεί να λειτουργήσει ως ασφάλεια ότι μόνο οι μεταβλητές με νόημα συμπεραίνουν την έξοδο, δηλαδή εγγυάται ότι μια υποκείμενη αληθινή αιτιότητα υπάρχει στη συλλογιστική του μοντέλου.

Σύμφωνα με τον Hall (2018), « η ερμηνεία του συστήματος θα πρέπει, προκειμένου να θεωρείται πρακτική, να παρέχει είτε κατανόηση των μηχανισμών και των προβλέψεων του μοντέλου, είτε οπτικοποίηση των κανόνων διάκρισης του μοντέλου, είτε υποδείξεις σχετικά με το τι θα μπορούσε να διαταράξει το μοντέλο». Προκειμένου να αποφευχθεί ο περιορισμός της αποτελεσματικότητας της τρέχουσας γενιάς συστημάτων τεχνητής νοημοσύνης, η eXplainable AI (XAI) (Gunning, 2017) προτείνει τη δημιουργία μιας σειράς τεχνικών ML που 1) παράγουν πιο επεξηγήσιμα μοντέλα διατηρώντας παράλληλα υψηλό επίπεδο μαθησιακής απόδοσης (π.χ. ακρίβεια πρόβλεψης) και 2) επιτρέπουν στους ανθρώπους να κατανοούν, να εμπιστεύονται κατάλληλα και να διαχειρίζονται αποτελεσματικά την αναδυόμενη γενιά τεχνητά ευφυών συνεργατών. Στο XAI συμάλλουν οι κοινωνικές επιστήμες (Miller, 2019) και η ψυχολογία της εξήγησης.

Ιστορική προέλευση της ανάγκης για εξηγήσιμη τεχνητή νοημοσύνη

Η επεξηγήσιμη τεχνητή νοημοσύνη δεν είναι ένας νέος τομέας (Goebel et al., 2018). Τα συστήματα τεχνητής νοημοσύνης υποστηρίχθηκαν εκτενώς κατά τη δεκαετία του 1980 και του 1990. Αυτά τα συστήματα βασίζονταν σε μια ακριβή και τυπική αναπαράσταση της ανθρώπινης γνώσης, χρησιμοποιώντας κατηγορηματική λογική, γράφους όπως οι κατευθυνόμενοι ακυκλικοί γράφοι (DAG), και μια προσεγγιστική συλλογιστική, όπως η θεωρία Bayes (1763), η ασαφής συλλογιστική και η θεωρία

Dempster-Shafer (2001). Ένα παράδειγμα ενός επιτυχημένου συστήματος βασισμένου σε αυτές τις γνώσεις είναι η βάση γνώσης Gene Ontology (Ashburner et al., 2000).

Ωστόσο, ένα σοβαρό μειονέκτημα αυτών των συστημάτων είναι ότι η Τεχνητή Νοημοσύνη χρειάζεται μια αναπαράσταση γνώσης που έχει δημιουργηθεί χειροκίνητα πριν αρχίσει να εργάζεται. Τα περισσότερα από τα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται σήμερα δεν βασίζονται ούτε σε γνώση ούτε παράγουν γνώση. Ειδικότερα από τη σκοπιά της βασισμένης στη γνώση τεχνητής νοημοσύνης, τα συστήματα αυτά θυσιάζουν την κατανόηση και την επεξηγηματικότητα υπέρ της απόδοσης.

Επιπρόσθετα μια διαφανής απόφαση με βάση την τεχνητή νοημοσύνη θα μπορούσε ιδανικά να επιτευχθεί όταν είναι διαθέσιμη μια υγιής επιστημονική θεωρία ως βάση για τον τρόπο λειτουργίας του υποκείμενου συστήματος ML. Στη συνέχεια, το αξιόπιστο σύστημα μπορεί να εξάγει λογικά συμπεράσματα με βάση αυτή τη θεωρία για να καταλήξει στα συμπεράσματά του. Στα συστήματα που βασίζονται σε υγιή επιστημονική θεωρία, μπορεί να εκτιμηθεί το εύρος και η ακρίβεια μιας πρόβλεψης. Επιπλέον, μπορεί να δοθεί ένα σκεπτικό (εξήγηση) για το αποτέλεσμα (Murray and Dermott, 2000). Στην αστρονομία, οι τρεις νόμοι του Κεπλέρ μπορούν να προκύψουν από τον νόμο της βαρύτητας του Νεύτωνα. Από αυτούς τους νόμους μπορεί να συναχθεί η θέση ενός πλανήτη.

Συχνά, τα συστήματα μηχανικής μάθησης χρησιμοποιούνται για εργασίες όπου δεν υπάρχει δεδομένη ή γνωστή μια επιστημονική θεωρία. Για παράδειγμα, αναπτύσσονται τεχνικές μηχανικής μάθησης για να ανιχνεύουν παθολογίες σε ασθενείς βάσει διαφόρων μετρήσεων της γονιδιακής έκφρασης, ακόμα και όταν οι ακριβείς μοριακές διαδικασίες που σχετίζονται με την ασθένεια είναι εν μέρει άγνωστες ή δεν έχουν πλήρη κατανόηση. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να εκπαιδευτούν να αναγνωρίζουν πρότυπα ή συσχετίσεις στα δεδομένα χωρίς την ανάγκη για προκαθορισμένους κανόνες ή μοντέλα. Αντίθετα, μπορούν να εξετάσουν τα δεδομένα και να εξάγουν αυτόματα πρότυπα ή να κάνουν προβλέψεις βάσει αυτών των προτύπων. Με αυτόν τον τρόπο, η μηχανική μάθηση μπορεί να είναι χρήσιμη για την ανακάλυψη νέων πληροφοριών ή πρότυπων σε δεδομένα όπου η επιστημονική κατανόηση είναι ελλιπής. Επιπλέον, μπορεί να παρέχει προβλέψεις ή εκτιμήσεις που

θα μπορούσαν να οδηγήσουν σε περαιτέρω έρευνα και κατανόηση των διαδικασιών που συμβαίνουν σε αυτά τα δεδομένα.

Σε τέτοιες περιπτώσεις, η βιβλιογραφία της μηχανικής μάθησης αρκείται στο να μετρά την "ποιότητα" ενός διαγνωστικού συστήματος μετρώντας την ακρίβεια των προβλέψεών του σε ένα περιορισμένο σύνολο δεδομένων που δεν χρησιμοποιήθηκε κατά την ανάπτυξη (εκπαίδευση, μάθηση, προσαρμογή και ρύθμιση) του συστήματος, δηλαδή τα λεγόμενα "δεδομένα δοκιμής" (Liaw and Wiener, 2002; Hryniewska et al., 2021).

Σε αυτήν την προσέγγιση, ο αλγόριθμος εκπαιδεύεται χρησιμοποιώντας προσεκτικά επιλεγμένα σύνολα δεδομένων εκπαίδευσης και δοκιμής, με στόχο την ανάπτυξη της ικανότητάς του να εκτελεί μια συγκεκριμένη εργασία, όπως η πραγματοποίηση μιας κλινικής διάγνωσης. Μέσω αυτής της διαδικασίας αξιολογείται η ικανότητά του να γενικεύει τις γνώσεις του σε νέες, άγνωστες περιπτώσεις. Στην ουσία, η εμπιστοσύνη προκύπτει από μια μέτρηση της απόδοσης σε αόρατα δεδομένα. Ωστόσο, συχνά, αυτά τα "αόρατα" δεδομένα ήταν ήδη διαθέσιμα κατά την ανάπτυξη του μοντέλου.

Οι περιορισμοί των συστημάτων ML που βασίζονται σε δεξιότητες είναι προφανείς: για δεδομένα που μοιάζουν πολύ στη δομή τους με τα δεδομένα εκπαίδευσης, το σύστημα θα έχει καλή απόδοση. Για δεδομένα που έχουν διαφορετική δομή, το σύστημα ML που βασίζεται σε δεξιότητες θα αποτύχει και μπορεί να μην αναγνωρίσει καν ότι τα δεδομένα αυτά δεν εμπίπτουν στον τομέα δεξιοτήτων του αλγορίθμου. Στην αστρονομία, αυτό μοιάζει με το μοντέλο του επίκυκλου της πλανητικής κίνησης. Μπορεί να θεωρηθεί ως μια εμπειρική ανάλυση Fourier της πλανητικής κίνησης, με μια σειρά από μεγαλύτερους και μικρότερους κύκλους που επικάθονται (Murschel, 1995). Για μικρές χρονικές περιόδους και υπό "κανονικές" συνθήκες, το μοντέλο του επίκυκλου μπορεί να προβλέψει τη θέση ενός πλανήτη με κάποια ακρίβεια (Hanson, 1960). Ωστόσο, δεν είναι γνωστό πότε η πρόβλεψη είναι σωστή ή λανθασμένη.

Για τα συστήματα ML που βασίζονται σε δεξιότητες, δεν έχει νόημα να ζητείται εξήγηση για μια απόφαση. Τα συστήματα τύπου "συσχετιστικής μνήμης", για παράδειγμα, αποθηκεύουν όλες τις περιπτώσεις και τις διαγνώσεις τους σε μια μνήμη (βάση δεδομένων). Σύμφωνα με τους Cover and Hart (1967), «η διάγνωση μιας νέας περίπτωσης καθορίζεται με την αναζήτηση των πιο όμοιων περιπτώσεων και την

ανάθεση της πλειοψηφίας των διαγνώσεων της πιο όμοιας περίπτωσης». Ένα παράδειγμα αυτού του τύπου αλγορίθμου είναι ο αλγόριθμος ταξινόμησης k-κοντινότερων γειτόνων. Οι προσπάθειες λεπτομερούς ανάλυσης αλγορίθμων που βασίζονται σε δεξιότητες οδηγούν μόνο στην κατανόηση του μαθηματικού μοντέλου που χρησιμοποιείται σε ένα τέτοιο σύστημα. Για παράδειγμα, η διάγνωση του ασθενούς A είναι D επειδή ο A είναι πιο όμοιος με τον ασθενή X, ο οποίος είχε διάγνωση D στο παρελθόν. Σε πρόσθετο επίπεδο, η εξασφάλιση της δικαιοσύνης και η αποφυγή διακρίσεων κατά μειονοτήτων, καθώς και άλλων ηθικών αρχών όπως η αποφυγή προκλήσεων βλάβης σε ανθρώπους, δεν μπορούν να εξασφαλιστούν ή να επιβληθούν στα συστήματα που βασίζονται σε δεξιότητες (βλ. <https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe>, πρόσβαση στις 15 Δεκεμβρίου 2021).

ΚΕΦΑΛΑΙΟ 2. Η ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΒΙΟΪΑΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ (ΧΑΙ)

Οι κύριοι τύποι ταξινομητών που χρησιμοποιούνται στη μηχανική μάθηση είναι συμβολικοί (Newell and Simon, 1976) ή υποσυμβολικοί (Smolensky, 2010) ταξινομητές. Για τους συμβολικούς ταξινομητές, η απόφαση για τον τρόπο με τον οποίο επιτυγχάνεται μια ταξινόμηση μπορεί να ερμηνευθεί από έναν ειδικό του τομέα ως συνδυασμός συνθηκών στα χαρακτηριστικά. Ένα παράδειγμα θα μπορούσε να είναι ο ταξινομητής που χρησιμοποιεί σύνολο κανόνων, οργανωμένων ιεραρχικά ή μη, για να καθορίσει σε ποια κατηγορία ανήκει ένα συγκεκριμένο αντικείμενο. Αυτή η δομή συμβαδίζει με τη σημερινή απαίτηση για Εξηγήσιμη Τεχνητή Νοημοσύνη. Αντίθετα, οι υποσυμβολικοί αλγόριθμοι δεν παρέχουν διαφάνεια σχετικά με τα συγκεκριμένα κριτήρια που χρησιμοποιούνται για την ανάθεση ενός αντικειμένου σε μια συγκεκριμένη κατηγορία, όπως για παράδειγμα υγιές ή άρρωστο. Ένα παράδειγμα υποσυμβολικών ταξινομητών είναι οι ταξινομητές Random Forest (Ho, 1995; Breiman, 2001), οι οποίοι βασίζονται σε έναν αριθμό διαφορετικών, ασυσχέτιστων και απλών δέντρων απόφασης.

Σύμφωνα με τη μελέτη των Lötsch και συνεργατών του (2021), παρατηρείται ότι είναι εφικτό να περιοριστεί ο αριθμός και ο τύπος των χαρακτηριστικών που χρησιμοποιούνται για τη λήψη αποφάσεων και να δοθεί ιεραρχία στη σημασία τους. Παρόλα αυτά, η συγκεκριμένη διαδικασία παραμένει αδιαφανής και δύσκολα κατανοητή, λειτουργώντας ως μαύρο κουτί. Για παράδειγμα, μια μέθοδος που έχει αναπτυχθεί για τα Random Forest είναι η ανάλυση αντιπροσωπευτικών δέντρων στο δάσος (Banerjee et al., 2012) (Εικόνα 2). Το αποτέλεσμα είναι μια μικρή επιλογή καλά λειτουργικών πρωτότυπων δέντρων από ένα σύνολο 1500 δέντρων στο Random Forest, στα οποία μπορούν να εντοπιστούν πιθανές διαδικασίες λήψης αποφάσεων. Μια εναλλακτική προσέγγιση για την ανάλυση μη ιεραρχικών κανόνων απόφασης σε μοντέλα όπως τα Random Forest και άλλους συμβολικούς ταξινομητές είναι η μέθοδος LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro κ.ά., 2016). Αυτή η μέθοδος δημιουργεί ένα ερμηνεύσιμο μοντέλο τοπικά γύρω από την ενιαία πρόβλεψη ενός εκπαιδευμένου μοντέλου τεχνητής νοημοσύνης, όπως ένα Random Forest. Αυτό επιτυγχάνεται με την προσαρμογή των κανόνων ανάθεσης για ένα μόνο παράδειγμα δεδομένων, όπως έναν μόνο ασθενή, αλλάζοντας τις τιμές των

χαρακτηριστικών και στη συνέχεια παρατηρώντας τον αντίκτυπο στην ταξινόμηση. Το αποτέλεσμα του LIME είναι ένα σύνολο κανόνων που αντιπροσωπεύουν τη συνεισφορά κάθε χαρακτηριστικού στην πρόβλεψη για μια μεμονωμένη περίπτωση δεδομένων, παρέχοντας μια τοπική ερμηνεία.

Εικόνα 2: Παράδειγμα μιας προσπάθειας να καταστούν οι υποσυμβολικοί ταξινομητές διαφανείς όσον αφορά τη δομή απόφασης κατά μήκος της οποίας γίνεται η ανάθεση κλάσης με την εξαγωγή καλά συμπεριφερόμενων και αντιπροσωπευτικών δέντρων από έναν ταξινομητή Random Forest. Δημιουργήθηκε ένας υποσυμβολικός ταξινομητής Random Forest με μέγεθος 1500 δέντρων που περιείχε έως και $d = 3$ μεταβλητές που σχετίζονται με τον πόνο με τον καθορισμό υπερπαραμέτρων, χρησιμοποιώντας τη βιβλιοθήκη R "randomForest" (<https://cran.r-project.org/package=randomForest> (Liaw and Wiener, 2002), πρόσβαση στις 15 Δεκεμβρίου 2021). Οι μεταβλητές που σχετίζονται με τον πόνο προέρχονται από το (Lötsch et al., 2021) και αποτελούνται από κατώφλια για διάφορα ερεθίσματα που καταγράφηκαν σε ποσοτικές αισθητηριακές δοκιμασίες σε μια μελέτη πειραματικά προκαλούμενου πόνου σε ανθρώπους, δηλαδή κατώφλια πόνου για βλαβερή θερμότητα και κρύο. Οι μεταβλητές που χρησιμοποιήθηκαν περιλάμβαναν τα κατώφλια πόνου θερμότητας (HPT) και τα κατώφλια πόνου ψύχους (CPT). Τα δεδομένα πόνου περιελάμβαναν τα μετασχηματισμένα κατά z κατώφλια πόνου για θερμικά ή ψυχρά ερεθίσματα που καταγράφηκαν σε συνθήκες ελέγχου και μετά από ακτινοβολία UV-B, καθώς και τις επιδράσεις UV-B που καταγράφηκαν ως η διαφορά μεταξύ των μετασχηματισμένων κατά z κατωφλίων (zHPTbaseline, zHPTUVB, zCPTbaseline, zCPTUVB, UVBEffHeat και UVBEffcold) που αποκτήθηκαν από 84 υγιή άτομα. Η διαδικασία ανάλυσης των αντιπροσωπευτικών δέντρων στο δάσος οδήγησε στην εντοπίση τεσσάρων δέντρων που παρουσιάζονται στον πίνακα C. Αυτή η ανάλυση βασίστηκε στο εκπαιδευμένο μοντέλο Random Forest, το οποίο χρησιμοποιήθηκε για προβλέψεις, ενώ παράλληλα αναγνώρισε τα αντιπροσωπευτικά δέντρα με βάση τη μετρική d_2 , με τη χρήση ευκλείδειας απόστασης. Το αποτέλεσμα ήταν τα δέντρα με αριθμό 732, 905, 913 και 1070 από τα 1500 δέντρα του δάσους. Αυτοί οι υπολογισμοί και τα διαγράμματα πραγματοποιήθηκαν με τη χρήση των βιβλιοθηκών R "reprtree" (<https://github.com/araastat/reprtree/blob/master/R/ReprTree.R> (Dasgupta and Reprtre, 2014), πρόσβαση στις 15 Δεκεμβρίου 2021) και "gggraph"

(<https://cran.r-project.org/package=ggraph>, πρόσβαση στις 15 Δεκεμβρίου 2021). Στο σχήμα παρουσιάζονται τα αντιπροσωπευτικά δέντρα, με τις αναθέσεις κλάσεων ως χρωματιστά φύλλα στα αντίστοιχα κατώτατα σημεία. Το σχήμα δημιουργήθηκε χρησιμοποιώντας το πακέτο λογισμικού R και το πακέτο R "ggplot2" (Knutson, 2020) (Πηγή: Löttsch et al., 2022).

Κύριοι Βιοϊατρικοί στόχοι του XAI

Οι προσεγγίσεις για την επεξήγηση των αποφάσεων των αλγορίθμων βαθιάς μάθησης για βιοϊατρικές εργασίες επικεντρώνονται κυρίως στην οπτικοποίηση των στοιχείων που συνέβαλαν σε κάθε απόφαση (Holzinger, 2020). Για παράδειγμα, μια από αυτές τις μεθόδους είναι οι διαδραστικοί χάρτες θερμότητας (Bach et al., 2016). Σύμφωνα με τους Holzinger (2020), Montavon (2019), « υπάρχουν διάφοροι τρόποι με τους οποίους τέτοιες μηχανικές εξηγήσεις μπορούν να αναδείξουν ποια είσοδος είναι σχετική με μια ληφθείσα έξοδος, χρησιμοποιώντας τις κλίσεις ως μια πολυμεταβλητή γενίκευση του υποπροϊόντος, όπου το νευρωνικό δίκτυο θεωρείται ως συνάρτηση και η εξήγηση βασίζεται στην κλίση της συνάρτησης που είναι διαθέσιμη από τον αλγόριθμο backpropagation». Ο όγκος των μελετών σχετικά με τις μεθόδους ερμηνευσιμότητας της μηχανικής μάθησης τα τελευταία χρόνια καταδεικνύει το αυξανόμενο ενδιαφέρον για αυτόν τον ερευνητικό τομέα. Ωστόσο, παρά τη ραγδαία ανάπτυξη, ο στόχος της κατανοητότητας για τους ειδικούς (στατιστικούς) θυσιάζεται υπέρ της κατανοητότητας για τους επαγγελματίες (Linardatos et al., 2020).

Έγινε μια προσπάθεια να καθοριστεί η θέση του XAI στο βιοϊατρικό πλαίσιο γενικότερα, ενώ το πρωταρχικό μέλημα της επίτευξης επεξηγησιμότητας και ιχνηλασιμότητας των αποφάσεων που βασίζονται στη μηχανική μάθηση είναι εγγενής. Προτάθηκε μια περαιτέρω ανάλυση με βάση τη μελέτη των όρων που χρησιμοποιούνται συχνά στο πλαίσιο του XAI (Arrieta et al., 2020). Κατά συνέπεια, το XAI θα πρέπει να εξυπηρετεί τους ακόλουθους στόχους, ήτοι (i) αξιοπιστία, (ii) αιτιότητα, (iii) μεταβιβάσιμότητα, (iv) πληροφοριακότητα, (v) εμπιστοσύνη, (vi) δικαιοσύνη, (vii) προσβασιμότητα, (viii) διαδραστικότητα και (ix) επίγνωση της ιδιωτικής ζωής.

Οι μέθοδοι μηχανικής μάθησης (ML) για εργασίες ταξινόμησης αποφασίζουν ποια κλάση (π.χ. μια κλινική διάγνωση) είναι κατάλληλη για μια δεδομένη περίπτωση.

Όταν η μοίρα ενός ατόμου εξαρτάται από το αποτέλεσμα μιας τέτοιας απόφασης που λαμβάνεται από έναν αλγόριθμο, η αξιοπιστία του συστήματος ML έχει ιδιαίτερη σημασία.

Ο όρος αξιόπιστη TN χρησιμοποιείται όλο και περισσότερο ως εναλλακτική του όρου XAI στην κλινική έρευνα και στη λήψη αποφάσεων με τη βοήθεια της TN, όταν η έννοια της XAI χρησιμοποιείται στο πλαίσιο της αλληλεπίδρασης ασθενούς-ιατρού. Η θεμελιώδης ιδέα που διατρέχει την αξιοπιστία μιας Τεχνητής Νοημοσύνης (TN) είναι να αποκτηθεί η εμπιστοσύνη των ατόμων ή των οργανισμών στο μοντέλο TN μέσω της εξήγησης των χαρακτηριστικών και των αιτιών που οδηγούν στην έξοδο της TN. Για παράδειγμα, εάν ούτε οι γιατροί ούτε οι ασθενείς εμπιστεύονται μια πρόβλεψη βασισμένη σε TN για μια κλινική διάγνωση, είναι απίθανο κάποιος από αυτούς να ακολουθήσει την πρόβλεψη (Thiebes et al., 2021). Μια λύση παρέχεται από τη μέθοδο LIME που προαναφέρθηκε. Ένα πρόσφατο παράδειγμα χρήσης αυτής της μεθόδου ήταν η διαφανής ανάθεση σε συγκεκριμένες ομάδες φαινοτύπων πόνου με βάση ένα τυχαίο δάσος (Lötsch and Malkusch, 2021).

XAI σχεδιασμένο για μη προγραμματιστές

Η εμβάθυνση στις μαθηματικές λεπτομέρειες ενός αλγορίθμου τεχνητής νοημοσύνης μπορεί να αποτελέσει μια διαδικασία κατανόησης για ειδικούς στη στατιστική ή την επιστήμη των υπολογιστών. Ωστόσο, όταν πρόκειται για το μέλλον και την ευημερία των ανθρώπων, αυτή η "εξήγηση του προγραμματιστή" είναι ανεπαρκής. Για παράδειγμα, η Παγκόσμια Τράπεζα απαιτεί από τα συστήματα τεχνητής νοημοσύνης που χρησιμοποιούνται για τη βαθμολόγηση πιστώσεων "την ικανότητα των ανθρώπων να ερμηνεύουν, να κατανοούν, να εξηγούν και να δικαιολογούν τις αποφάσεις που λαμβάνονται με μεθόδους που χρησιμοποιούν μεγάλο αριθμό μεταβλητών" (Knutson, 2020). Τελικά, ένας άνθρωπος πρέπει να είναι σε θέση να αναλάβει την ευθύνη για τις συνέπειες ενός συστήματος τεχνητής νοημοσύνης.

Τα γραμμικά μοντέλα, ειδικά τα δομικά απλά, είναι κατανοητά από μαθηματικούς, στατιστικούς ή επιστήμονες πληροφορικής (δηλαδή, τους προγραμματιστές). Ωστόσο, αυτά τα μοντέλα έχουν τους περιορισμούς τους. Η ανάπτυξη των "παράλληλων κατανεμημένων επεξεργασιών" έχει προσπαθήσει να ξεπεράσει αυτούς τους περιορισμούς των γραμμικών συστημάτων. Τέτοια μοντέλα αποτελούνται από

έναν πολύ μεγάλο αριθμό μη γραμμικών συναρτήσεων. Ρυθμίζοντας πολλές παραμέτρους, ένα τέτοιο μοντέλο μπορεί να "μάθει" να αναπαράγει δεδομένες καταστάσεις εισόδου-εξόδου. Λόγω της μεγάλης αριθμού αλληλεπιδρώντων στοιχειωδών διεργασιών (νευρώνων) σημαντική είναι η κατανόηση των λεπτομερειών των ενός τέτοιου συστήματος σε πεπερασμένο χρόνο δεν είναι ούτε σκοπούμενη ούτε δυνατή. Τέτοια συστήματα αναφέρονται ως "μαύρα κουτιά".

Το XAI απαιτεί μεθόδους ανακάλυψης γνώσης που να μπορούν να χρησιμοποιηθούν από τη μηχανή και να εξηγηθούν σε έναν εμπειρογνώμονα του τομέα ή ακόμη και σε έναν απλό άνθρωπο. Οι πιο ακριβείς ορισμοί του XAI είναι παλαιότεροι. Στην έρευνα για την τεχνητή νοημοσύνη βασισμένη στη γνώση (Michalski, Craven and Shavlik, 1996). Αυτή η κατανόηση και αποδοχή του μοντέλου εξαγωγής συμπερασμάτων της TN πρέπει να είναι τέτοια ώστε ο εμπειρογνώμονας να είναι πρόθυμος τελικά να αναλάβει τη νομική ευθύνη για τις αποφάσεις της TN.

Τα συστήματα XAI δεν μπορούν να βασίζονται αποκλειστικά στις δεξιότητες και τις εμπειρίες. Αντίθετα, πρέπει να λαμβάνουν τις αποφάσεις τους χρησιμοποιώντας επιστημονική λογική συλλογιστική, βασιζόμενα στην αναγνωρισμένη εμπειρογνωμοσύνη. Τα συστήματα XAI πρέπει να είναι σε θέση να εξηγούν κάθε απόφαση και τα συμπεράσματά τους με έναν τρόπο που να είναι κατανοητός από τον ειδικό του τομέα (τομέας κατανόησης). Είναι σημαντικό, τα συστήματα XAI να στηρίζονται σε γνώσεις που είναι επεξεργάσιμες από μηχανές και προσανατολισμένες στην ανθρώπινη γλώσσα.

Τα λεγόμενα "έμπειρα" ή "βασισμένα στη γνώση" συστήματα (Yanase and Triantaphyllou, 2019) πληρούν αυτή την απαίτηση. Συνήθως βασίζονται σε μια αναπαράσταση των εννοιών, των γεγονότων, των κανόνων, των σχέσεων και των θεωριών σε έναν συγκεκριμένο τομέα (Ultsch et al., 2006). Η βάση γνώσης GeneOntology (Ashburner et al., 2020) παρέχει ένα παράδειγμα αναπαράστασης γνώσης στον τομέα της κυτταρικής βιολογίας και της γενετικής. Ένα σύστημα XAI καταλήγει σε συμπεράσματα (αποφάσεις/διαγνώσεις) εφαρμόζοντας τυπικές μεθόδους επιστημονικής συλλογιστικής, όπως ο λογισμός προθέσεων (Hodges, 2001). Υπάρχουν συστήματα TN με μηχανική μάθηση που χρησιμοποιούν αυτό το είδος συλλογισμού, δηλαδή τα επιμέρους βήματα απόφασης παρέχονται απευθείας από τις λεγόμενες "συμβολικές" μεθόδους μηχανικής μάθησης που βασίζονται στην ανάθεση κλάσης μιας περίπτωσης σε ένα σύνολο ιεραρχικά ή μη ιεραρχικά

οργανωμένων κανόνων (Loh, 2011). Παραδείγματα αποτελούν τα ιεραρχικά δέντρα ταξινόμησης και παλινδρόμησης (CART) ή η μη ιεραρχική επαναλαμβανόμενη αυξητική αποκοπή για τη μείωση των σφαλμάτων (RIPPER (Cohen, 1995). Σύμφωνα με τους Marewski and Gigerenzer (2021), « μεταξύ των αλγορίθμων που βασίζονται σε συμβολικά δέντρα, τα λεγόμενα "Fast and Frugal Trees" (FFTs) παρέχουν ιδιαίτερα απλά δέντρα απόφασης, τα οποία συνήθως αποτελούνται από 1 έως 5 πληροφορίες». Αυτό τα καθιστά ιδιαίτερα κατάλληλα για βιοϊατρικά προβλήματα, καθώς μιμούνται τις διαδικασίες για τη λήψη μιας κλινικής διάγνωσης.

Ωστόσο, μια σημαντική απαίτηση για τα συστήματα τεχνητής νοημοσύνης με μηχανική μάθηση συχνά παραβλέπεται: η κατανόηση της γνώσης που χρησιμοποιείται στο σύστημα από έναν ειδικό του τομέα. Τα συμβολικά συστήματα με μηχανική μάθηση συχνά στερούνται αυτής της ιδιότητας. Για παράδειγμα, τα δέντρα αποφάσεων μπορεί να αποτελούνται από εκατοντάδες συνθήκες. Πανομοιότυπα υποδέντρα μπορεί να χρησιμοποιούνται επανειλημμένα σε διαφορετικούς κλάδους ενός τέτοιου δέντρου αποφάσεων. Είναι αποδεκτό για έναν αλγόριθμο υπολογιστή να βασίζει μια απόφαση σε εκατοντάδες συνθήκες. Αντίθετα οι άνθρωποι, παρουσιάζουν περιορισμένη ικανότητα όσον αφορά την πολυπλοκότητα και τον πλεονασμό των μοντέλων ή των εξηγήσεων. Σύμφωνα με το νόμο του Miller, «το τυπικό όριο της ανθρώπινης ικανότητας επεξεργασίας πληροφοριών είναι 7 ± 2 στοιχεία (Miller, 1956)». Οι εξηγήσεις των ΧΑΙ πρέπει, επομένως, να είναι όσο το δυνατόν απλούστερες (ξυράφι του Όκαμ) και να χρησιμοποιούν αφαιρέσεις (γενικεύσεις) από παραδείγματα καταστάσεων.

Κύριοι Βιοϊατρικοί στόχοι του ΧΑΙ

Οι προσεγγίσεις για την επεξήγηση των αποφάσεων των αλγορίθμων βαθιάς μάθησης για βιοϊατρικές εργασίες έχουν ως κύριο μέλημά τους την οπτικοποίηση των στοιχείων που συνέβαλαν σε κάθε απόφαση (Holzinger, 2020). Χαρακτηριστικό παράδειγμα είναι οι διαδραστικοί χάρτες θερμότητας (Bach, 2016). Σύμφωνα με τους Holzinger (2020) και Montavon (2019), « υπάρχουν διάφοροι τρόποι με τους οποίους τέτοιες μηχανικές εξηγήσεις μπορούν να αναδείξουν ποια είσοδος είναι σχετική με μια ληφθείσα έξοδος, χρησιμοποιώντας τις κλίσεις ως μια πολυμεταβλητή γενίκευση του υποπροϊόντος, όπου το νευρωνικό δίκτυο θεωρείται ως συνάρτηση και η εξήγηση

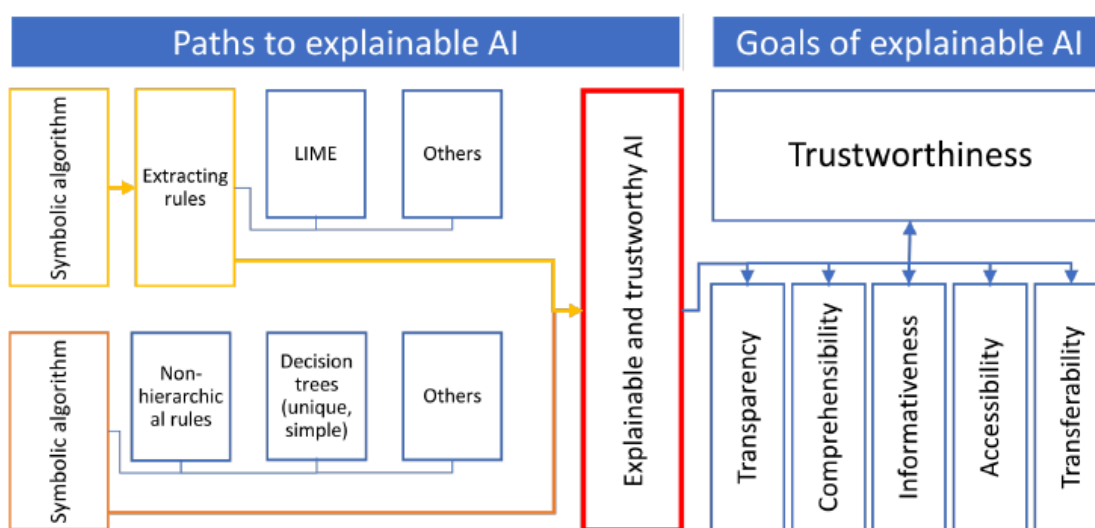
βασίζεται στην κλίση της συνάρτησης που είναι διαθέσιμη από τον αλγόριθμο backpropagation». Ο όγκος των μελετών σχετικά με τις μεθόδους ερμηνευσιμότητας της μηχανικής μάθησης τα τελευταία χρόνια καταδεικνύει το αυξανόμενο ενδιαφέρον για αυτόν τον ερευνητικό τομέα. Παρόλη την ανάπτυξη και την εξέλιξη, ο στόχος της κατανόησης από ειδικούς όπως οι στατιστικολόγοι υποβαθμίζεται προς όφελος της κατανόησης από επαγγελματίες (Linardatos et al., 2020). Προσπαθείται να διατυπωθεί η θέση της Εξηγημένης Τεχνητής Νοημοσύνης (ETN) στον τομέα της ιατρικής γενικότερα. Ο κύριος στόχος είναι η επίτευξη επεξηγηματικότητας και ιχνηλασιμότητας των αποφάσεων που βασίζονται στη μηχανική μάθηση, μελετώντας περαιτέρω τους όρους που συχνά χρησιμοποιούνται στο πλαίσιο της ETN. Συνεπώς, η ETN πρέπει να εξυπηρετεί τους ακόλουθους στόχους: (i) αξιοπιστία, (ii) αιτιότητα, (iii) μεταβιβασιμότητα, (iv) πληροφοριακότητα, (v) εμπιστοσύνη, (vi) δικαιοσύνη, (vii) προσβασιμότητα, (viii) διαδραστικότητα και (ix) επίγνωση της ιδιωτικής ζωής.

Αξιοπιστία

Οι μέθοδοι μηχανικής μάθησης για εργασίες ταξινόμησης αποφασίζουν το είδος της κλάσης (π.χ. μια κλινική διάγνωση) που είναι κατάλληλη για μια συγκεκριμένη περίπτωση. Ο όρος αξιόπιστη TN χρησιμοποιείται όλο και περισσότερο ως εναλλακτική λύση του όρου XAI σε κλινική έρευνα και τη λήψη αποφάσεων με τη βοήθεια της TN, όταν η έννοια του XAI χρησιμοποιείται σε πλαίσιο της αλληλεπίδρασης ασθενούς-ιατρού. Η βασική έννοια πίσω από την αξιοπιστία ενός τεχνητού νευρωνικού δικτύου είναι να εμπνεύσει εμπιστοσύνη στους ανθρώπους ή τους οργανισμούς σχετικά με το πώς λειτουργεί το μοντέλο του νευρωνικού δικτύου, με την εξήγηση των χαρακτηριστικών του και των λόγων πίσω από τα αποτελέσματά του. Αυτό συμβάλλει στην πλήρη αποδοτικότητα του νευρωνικού δικτύου. Για παράδειγμα, εάν ούτε οι γιατροί ούτε οι ασθενείς εμπιστεύονται μια συστάση βασισμένη σε αυτό, είναι απίθανο κάποιος από αυτούς να ακολουθήσει τη συστάση (Thiebes et al., 2021). Μια λύση που προσφέρεται είναι η μέθοδος LIME.

ΠΕΡΑΙΤΕΡΩ ΚΑΙ ΣΥΝΑΦΕΙΣ ΣΤΟΧΟΙ ΤΗΣ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΥΝΗΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

Οι ακόλουθες επεξηγήσεις και παραδείγματα επιχειρούν να αποτυπώσουν περισσότερους από τους κύριους στόχους που προσδιορίστηκαν στο (Arrieta et al., 2020), με έμφαση στη βιοϊατρική έρευνα και, ειδικότερα, στην κλινική λήψη αποφάσεων λήψη αποφάσεων με βάση την ΤΝ ή αλγορίθμους μηχανικής μάθησης (Σχήμα 3).



Εικόνα 3: Η γραφική αναπαράσταση της επίτευξης αξιόπιστων τεχνητών νευρωνικών δικτύων στη βιοϊατρική περιλαμβάνει την κλινική λήψη αποφάσεων και τις απαιτήσεις για μια τέτοια τεχνητή νοημοσύνη σε αυτό το περιβάλλον. Στο αριστερό μέρος φαίνονται οι διαδρομές προς την αξιόπιστη τεχνητή νοημοσύνη. Οι αποφάσεις που βασίζονται στην τεχνητή νοημοσύνη μπορούν να υλοποιηθούν ως συμβολικοί αλγόριθμοι, οι οποίοι συχνά χρησιμοποιούν κανόνες ή μικρά σύνολα κανόνων για την ταξινόμηση, ή ένας υποσυμβολικός και συχνά πιο ισχυρός τύπος αλγόριθμου μηχανικής μάθησης χρησιμοποιείται. Στη συνέχεια, εφαρμόζονται περαιτέρω μέθοδοι, όπως οι τοπικές ερμηνεύσιμες επεξηγήσεις με βάση το μοντέλο (local interpretable model-agnostic explanations, LIME), για την εξαγωγή κατανοητών κανόνων για την ανάθεση κλάσεων. Στο δεξιό μέρος παρουσιάζονται οι κύριοι στόχοι που ανατίθενται σε ένα XAI στη βιοϊατρική και κλινικό πλαίσιο, όπως προτείνεται στο (Arrieta et al., 2020), με κύριο στόχο τη λήψη κλινικών αποφάσεων βασισμένων στην ΤΝ. να είναι αξιόπιστες με το να είναι κατανοητές τόσο από τον ιατρό όσο και από τον ασθενή. Το σχήμα δημιουργήθηκε με τη χρήση του Microsoft PowerPoint.

365 (Redmond, WA, ΗΠΑ) σε Microsoft Windows 11 που εκτελείται σε ένα εικονική μηχανή που τροφοδοτείται από το VirtualBox 6.1 (Oracle Corporation, Austin, TX, USA) σε έναν υπολογιστή με Ubuntu Linux 20.04.03 LTS 64-bit.

Διαφάνεια

Η διαφάνεια ομαδοποιείται επίσης με τους όρους black box to white box, εγγενείς εξηγήσεις, κατανοητότητα ή κατανοητότητα. Όλοι αυτοί οι όροι αναφέρονται σε μια ακριβή περιγραφή των μαθηματικών/στατιστικών/αλγοριθμικών λεπτομερειών του τρόπου με τον οποίο λειτουργεί εσωτερικά το μοντέλο AI. Αυτό το είδος εξήγησης μπορεί να είναι κατανοητό από στατιστικούς, μαθηματικούς, και/ή στους επιστήμονες υπολογιστών. Σε έναν γιατρό ή ασθενή δεν έχει μεγάλη σημασία. Στον τραπεζικό τομέα, για παράδειγμα, η "διαφάνεια" μιας τεχνητής νοημοσύνης που αποφασίζει αν ένας πελάτης είναι επιλέξιμος για δάνειο βασίζεται στην εξίσωση $\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1$ [64]. Η επιτυχία της σημερινής "υποσυμβολικής" τεχνητής νοημοσύνης βασίζεται στην ανταλλαγή της κατανοητότητας του μοντέλου με την απόδοση του όσον αφορά την ακρίβεια των προβλέψεων της τεχνητής νοημοσύνης ή την ανάθεση τάξεων, όπως περιγράφεται λεπτομερέστερα παρακάτω στην ενότητα Ιστορία. Τέτοια μοντέλα, όπως τα τεχνητά νευρωνικά δίκτυα (ANN) (Rosenblatt, 1958) ή τα τυχαία δάση, χρησιμοποιούν έναν μεγάλο αριθμό μη γραμμικών συναρτήσεων ως νευρώνες ή δέντρα απόφασης που συνδέονται με χιλιάδες παράγοντες σύζευξης, όπως συνάψεις και βάρη, αντίστοιχα (Ho, 1995; Breiman, 2001). Αυτά τα μοντέλα, στη βαθιά μάθηση, μπορούν να οργανωθούν σε πολλά επίπεδα, επιτρέποντας την αναγνώριση πολύπλοκων προτύπων και σχέσεων στα δεδομένα.

Χάρη στην ισχύ των σύγχρονων υπολογιστών, έως και τους σημερινούς H/Y, οι πολλές χιλιάδες ή εκατομμύρια εσωτερικές παράμετροι μπορούν να βελτιστοποιηθούν για αυτά τα μοντέλα, έτσι ώστε να μπορούν να προσεγγιστούν εκείνες οι συναρτήσεις που αντιστοιχούν τα πολυδιάστατα πολυμεταβλητά δεδομένα εισόδου σε πολυτιμημένες εξόδους, όπως οι διάφορες διαγνωστικές κλάσεις. Με αυτόν τον τρόπο, τα επιμέρους στοιχεία (νευρώνες, δέντρα) μπορούν να περιγραφούν με ακρίβεια, αλλά η προκύπτουσα συλλογική συμπεριφορά του συστήματος δεν μπορεί να γίνει κατανοητή. Αυτό μπορεί να συγκριθεί με την αδυναμία εξήγησης των σκέψεων ή των ιδεών με την πυροδότηση των νευρώνων στον εγκέφαλο. Σύμφωνα

με τον Ultsch (2007), «στη θεωρία συστημάτων, αυτό θεωρείται μια από τις κεντρικές ιδιότητες των αναδυόμενων συστημάτων».

Κατανοητότητα

Αναφέρεται επίσης ως ερμηνευσιμότητα, η κατανοησιμότητα μιας εξήγησης σημαίνει την παροχή αιτιώδους και λογικής συναγωγής των αποτελεσμάτων (αποφάσεων) από τα δεδομένα (δεδομένα εισόδου) με τη χρήση των όρων, των διατυπώσεων και των μεθόδων λήψης αποφάσεων στην αντίστοιχη θεματική περιοχή. Οι κατανοητές εξηγήσεις διατυπώνονται με τη χρήση επαγωγικής λογικής, λαμβάνοντας υπόψη τις προσεγγίσεις (ασάφεια) και τους κινδύνους. Αυτή η έννοια είναι η ίδια με την έννοια της επεξηγηματικότητας της τεχνητής νοημοσύνης (AI), και χρησιμοποιείται ως επεξηγήσιμη τεχνητή νοημοσύνη (XAI) στο υπόλοιπο της παρούσας εργασίας. Το XAI (Explainable Artificial Intelligence) αναφέρεται στην τεχνητή νοημοσύνη όπου τα αποτελέσματα και οι λύσεις μπορούν να εξηγηθούν με έναν τρόπο που είναι κατανοητός από τον άνθρωπο. Ο όρος "εξηγήσιμος" χρησιμοποιείται και ορίζεται συχνά διαφορετικά από τους ερευνητές, καθώς δεν υπάρχει συγκεκριμένος μαθηματικός ορισμός (Gilpin et al., 2018). Το πιο σημαντικό είναι ότι ο άνθρωπος που πρέπει να κατανοήσει την απόφαση της AI είναι αμφιλεγόμενος: είναι ο ασθενής, ο γιατρός ή ένας επιστήμονας υπολογιστών ή στατιστικολόγος; Στον τομέα της εξηγήσιμης τεχνητής νοημοσύνης (XAI), ένα πρόβλημα που αντιμετωπίζεται είναι η ανάγκη για σαφής καθορισμός των όρων ερμηνευσιμότητας και εξηγησιμότητας, καθώς χρησιμοποιούνται με διαφορετικούς τρόπους από διαφορετικούς ερευνητές. Παρά την έλλειψη ενός συγκεκριμένου μαθηματικού ορισμού, ο όρος "κατανοητότητα" εμφανίζεται ως η πιο ουσιώδης έννοια στο XAI. Ωστόσο, υπάρχει συναίνεση ότι το XAI πρέπει να διασφαλίζει ότι οι υπολογιστικές αποφάσεις είναι διαφανείς, ώστε να μπορούν να κοινοποιούνται στους επηρεαζόμενους ασθενείς με κατανοητό τρόπο, όταν πρόκειται για βιοϊατρικές και ιδίως κλινικές αποφάσεις.

Πληροφόρηση

Η πληροφοριακότητα της AI είναι απαραίτητη σε συστήματα υποστήριξης κλινικών αποφάσεων με στόχο να βοηθούν τους γιατρούς σε διαγνωστικά ή θεραπευτικά καθήκοντα. Σύμφωνα με τον Lee et al., (2019), «τέτοια συστήματα βασίζονται στην AI και χρησιμοποιούνται όλο και περισσότερο στην κλινική πρακτική, με επίκαιρη

εστίαση στην ιατρική απεικόνιση. Η νόσος Αλτσχάιμερ μπορεί να διαγνωστεί από εικόνες μαγνητικού συντονισμού με την εκπαίδευση βαθιών νευρωνικών δικτύων για τον εντοπισμό μη φυσιολογικών περιοχών του εγκεφάλου». Ομοίως, τα βαθιά νευρωνικά δίκτυα εισήχθησαν στην κλινική απεικόνιση για να διευκολύνουν τη λήψη αποφάσεων με τη χρήση αυτών των δεδομένων (Papadimitroulas et al., 2021). Ωστόσο, ενώ τα αποτελέσματα αυτών των αναλύσεων φαίνονται λογικά σε έναν ιατρικό εμπειρογνώμονα επειδή συνάδουν με την ιατρική γνώση και, ως εκ τούτου, θα μπορούσαν να κοινοποιηθούν στον ασθενή, οι ακριβείς μηχανισμοί του τρόπου με τον οποίο γίνεται η διάγνωση για έναν συγκεκριμένο ασθενή παραμένουν ασαφείς. Τα βαθιά νευρωνικά δίκτυα είναι υποσυμβολικοί ταξινομητές με την παραπάνω έννοια, όπως και τα τυχαία δάση, των οποίων τα εκατοντάδες ή χιλιάδες δέντρα αποφάσεων επίσης δεν μπορούν να κατανοηθούν με όλες τις λεπτομέρειες από τον ιατρό και να κοινοποιηθούν με κατανοητό τρόπο στον ασθενή.

Προσβασιμότητα

Η προσβασιμότητα μπορεί να ερμηνευτεί με τη συμμετοχή των τελικών χρηστών στη διαδικασία βελτίωσης και ανάπτυξης ενός αλγορίθμου ΑΙ, όπως έχει προταθεί προηγουμένως. Επιπρόσθετα η προσβασιμότητα μπορεί να θεωρηθεί η ικανότητα λήψης αποφάσεων βασισμένων στη μηχανική μάθηση χωρίς βαθύ προγραμματισμό και γνώσεις ΑΙ. Αυτό μπορεί να γίνει μέσω διαδραστικού προ-συσκευασμένου λογισμικού, όπως το πακέτο R "rattle" που χρησιμοποιείται στο δεύτερο κεφάλαιο της παρούσας έκθεσης ή άλλα διαδραστικά εργαλεία, όπως οι βιβλιοθήκες R "AdaptGauss" <https://cran.r-project.org/package=AdaptGauss>, που παρέχει διαδραστική προσαρμογή μοντέλων μίξης Gauss (Ultsch et al., 2015), "pguIMP" για την προεπεξεργασία συνόλων δεδομένων βιοϊατρικών εργαστηρίων, συμπεριλαμβανομένης της προσθήκης ελλειπόντων τιμών με μηχανική μάθηση (<https://cran.r-project.org/package=pguIMP> (Malkusch et al., 2021), όλα προσπελάστηκαν στις 15 Δεκεμβρίου 2021), και πολλά άλλα που θα απαιτούσαν ξεχωριστή εξέταση που υπερβαίνει το παρόν αντικείμενο.

Ωστόσο, η προσβασιμότητα μπορεί επίσης να γίνει κατανοητή από την άποψη της πνευματικής προσβασιμότητας. Ο στόχος υλοποιείται σε συμβολικές μορφές ΑΙ, συχνά σε απλούς αλγορίθμους μηχανικής μάθησης που έχουν τη μορφή ιεραρχικών ή

μη ιεραρχικών κανόνων. Ένας συμβολικός ταξινομητής βασισμένος σε κανόνες που περιλάμβανε 21 ατομικές ή συγκεντρωτικές παραμέτρους, συμπεριλαμβανομένων δημογραφικών χαρακτηριστικών, ψυχολογικών και σχετικών με τον πόνο παραμέτρων που καταγράφηκαν νωρίς μετά τη χειρουργική επέμβαση για τον καρκίνο του μαστού, προέβλεψε την επακόλουθη ανάπτυξη επίμονου πόνου με διασταυρωμένα επικυρωμένη ακρίβεια 86% και αρνητική προγνωστική αξία περίπου 95% όταν χρησιμοποιήθηκαν οι περισσότεροι μη ιεραρχικοί κανόνες (Lötsch et al., 2018). Ένα άλλο παράδειγμα από την έρευνα για τον πόνο είναι το πώς η υποσυμβολική AI μπορεί να υποβληθεί σε περαιτέρω ανάλυση για την εξαγωγή της ατομικής διαδικασίας λήψης αποφάσεων για κάθε περίπτωση ή πώς μπορεί να συμπληρωθεί από συμβολική AI που παρέχει κατανοητές εξηγήσεις για την ανάθεση σε ομάδες, παρόλο που η ακριβής διαδικασία λήψης αποφάσεων μπορεί να διαφέρει από εκείνη της υποσυμβολικής AI. Αυτό έχει αναπτυχθεί λεπτομερέστερα σε ένα σύστημα οπτικής ανάλυσης για τη σύγκριση πολλαπλών μοντέλων προβλέψεων για κλινικά δεδομένα (Li et al., 2020). Το σύστημα δίνει τη δυνατότητα σύγκρισης και αξιολόγησης διαφορετικών μοντέλων AI. Αυτό βασίζεται σε πληροφορίες, με στόχο να βοηθήσει τους κλινικούς ιατρούς στη λήψη αποφάσεων. Τα διαφορετικά μοντέλα συγκρίνονται ως προς τις κριτήρια πρόβλεψης που χρησιμοποιούνται και αξιολογείται η συνέπεια της εφαρμογής τους.

Μεταβιβασιμότητα

Η δυνατότητα μεταφοράς αναφέρθηκε ως ο δεύτερος συνηθέστερος λόγος για τη χρήση του XAI στην έρευνα (Liao et al., 2020). Η δυνατότητα μεταφοράς εξηγεί τον τρόπο λειτουργίας ενός μοντέλου AI και χρησιμεύει στην καλύτερη κατανόηση του υποκείμενου προβλήματος, ώστε η λύση να μπορεί να εφαρμοστεί ευκολότερα σε ένα διαφορετικό εφαρμογή ή πρόβλημα. Στην προαναφερθείσα κοόρτης για τον καρκίνο του μαστού, το αρχικό σύνολο κανόνων για την πρόβλεψη της επιμονής του πόνου περιλάμβανε πληροφορίες που συλλέχθηκαν από ασθενείς μέσω επαναλαμβανόμενης χρήσης ολοκληρωμένων ψυχολογικών ερωτηματολογίων. Μόλις αυτό αποδείχθηκε κατατοπιστική, η επιβλεπόμενη μηχανική μάθηση θα μπορούσε να χρησιμοποιηθεί για να μειώσει τα ερωτηματολόγια σε στοιχεία που σχετίζονται με το πλαίσιο του πόνου. Αυτό επιτεύχθηκε με τη δημιουργία μιας συντομότερης μορφής του ερωτηματολογίων που περιείχε μόνο επτά στοιχεία, τα οποία αντιπροσώπευαν το 10% του αρχικού ψυχολογικού ερωτήσεων, αλλά απέδωσε την ίδια προβλεπτική

απόδοση για την επιμονή του πόνου με το πλήρη ερωτηματολόγιο (Lötsch et al., 2018). Σίγουρα, αυτό το σύντομο ερωτηματολόγιο είναι πολύ πιο εύκολο για τους κλινικούς ιατρούς και ασθενείς να κατανοήσουν από τα πιο γενικά πλήρη ερωτηματολόγια.

Συντονισμένη ερμηνεία ΑΙ μεταξύ εμπειρογνομόνων πληροφορικής και βιοϊατρικού τομέα

Η επιστήμη της πληροφορικής περιγράφεται ως ένας ταχέως αναπτυσσόμενος διεπιστημονικός τομέας που χρησιμοποιεί προηγμένες υπολογιστικές δυνατότητες για την κατανόηση και την επίλυση πολύπλοκων προβλημάτων (Rao et al., 2020). Απαιτεί εγγενώς συνεργασία που περιλαμβάνει την ανταλλαγή και τη συνεργασία σε πληροφορίες και μεθόδους μεταξύ επαγγελματιών διαφορετικών τομέων, όπως για παράδειγμα γιατρών και εμπειρογνομόνων μηχανικής μάθησης (Benioff, 2005). Η συντονισμένη δημιουργία μοντέλων, με τη συμμετοχή διαφόρων εμπειρογνομόνων από διαφορετικά πεδία, είναι απαραίτητη για τον εντοπισμό και την εξάλειψη των παγίδων της μηχανικής μάθησης, όπως οι συγχυτικές μεταβλητές και οι υποκατάστατοι δείκτες, που συνήθως αναφέρονται ως συνδιακυμάνσεις που διαρρέουν δεδομένα. Ένα παράδειγμα είναι ο εντοπισμός προστατευτικής επίδρασης του ανταγωνιστή του υποδοχέα σεροτονίνης 5-HT₃ ondansetron, ενός αντιεμετικού που χρησιμοποιείται συνήθως για τη θεραπεία της ναυτίας και του εμέτου, έναντι της φλεβικής θρομβοεμβολής που σχετίζεται με τη νοσηλεία (Datta et al., 2021). Ενώ ένας αρχικός ταξινομητής πέτυχε ROC-AUC για την πρόβλεψη κινδύνου 0,92, μετά από μια συντονισμένη προσπάθεια των επιστημόνων βιοϊατρικής και πληροφορικής να αποκλείσουν συνδιακυμάνσεις που διαρρέουν δεδομένα, όπως η ειδική φαρμακολογική προφύλαξη ή η θεραπεία της θρομβοεμβολής, η ROC-AUC μειώθηκε σε 0,87, η οποία φαίνεται να αποτυπώνει πιο ρεαλιστικά το όφελος της ονδανσετρόνης σε αυτό το πλαίσιο. Τονίζεται η σημασία της συνεργασίας με ειδικούς στον τομέα της βιοϊατρικής για την αξιολόγηση των μοντέλων TN και την ερμηνεία των αποτελεσμάτων. Η επαγγελματική εμπειρογνομοσύνη στη βιοϊατρική μπορεί να προσφέρει πολύτιμες γνώσεις για την ερμηνεία των αποτελεσμάτων και τη διάκριση μεταξύ διαφορετικών μεταβλητών που μπορεί να επηρεάζουν τα αποτελέσματα της ανάλυσης. Η συνεργασία με βιοϊατρικούς επιτρέπει επίσης την ενσωμάτωση κλινικής

εμπειρογνωμοσύνης και πρακτικής στην ανάπτυξη και τη βελτίωση των μοντέλων TN, βοηθώντας έτσι στην ανάπτυξη πιο αξιόπιστων και ερμηνεύσιμων αλγορίθμων για τη λήψη κλινικών αποφάσεων. (Bhattacharya and Vogt, 2012).

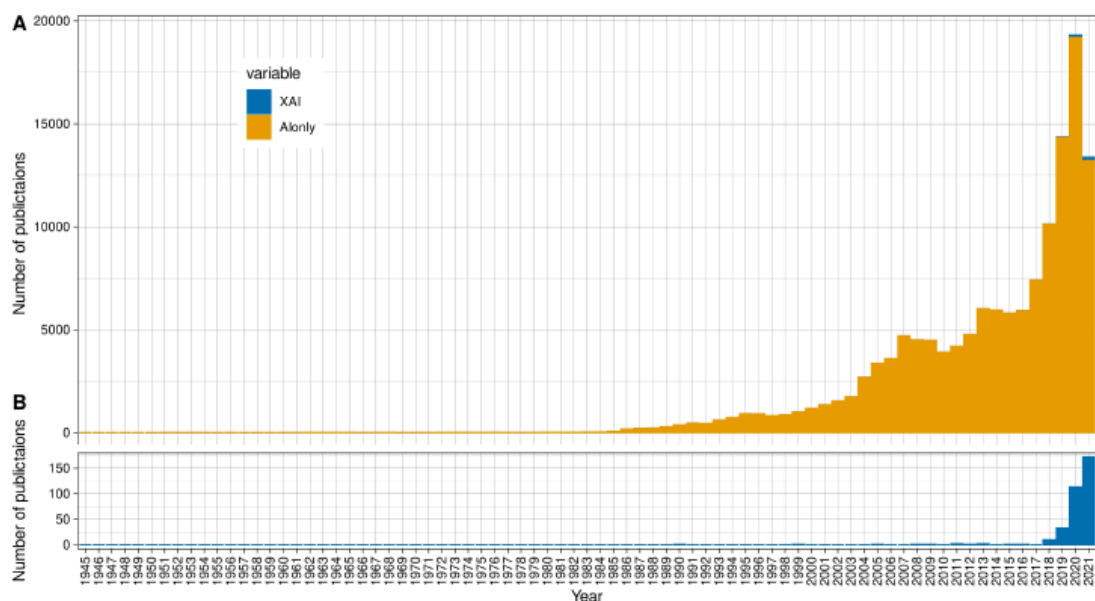
Φαίνεται ότι απαιτείται εμπειρογνωμοσύνη για την αντιμετώπιση των συγχυτικών μεταβλητών, έτσι ώστε ένας εμπειρογνώμονας να μπορεί να αποφασίσει ποια μεταβλητή μπορεί ενδεχομένως να θεωρηθεί συγχυτική μεταβλητή (και όχι μεσολαβητική μεταβλητή) ή υποκατάστατος δείκτης (VanderWeele, 2019). Όπως αναφέρεται αλλού, η πρόβλεψη απαιτεί μόνο συσχέτιση, αλλά η κατανόηση απαιτεί σημαντική γνώση που διέπει τους αιτιώδεις μηχανισμούς (Datta et al., 2021). Άλλα προβλήματα στη δημιουργία μοντέλων μηχανικής μάθησης περιλαμβάνουν την έλλειψη σημείων δεδομένων σε σχέση με τον αριθμό των διαθέσιμων μεταβλητών για επιλογή και τα αραιά σύνολα δεδομένων όπου πολλές από τις ετικέτες λείπουν. Και πάλι, ένας τρόπος αντιμετώπισης αυτών των ελλείψεων προτείνεται να συνίσταται στη συμμετοχή ανθρώπινων εμπειρογνομόνων του τομέα σε διάφορα στάδια της κατασκευής των συνόλων δεδομένων, της εκπαίδευσης και της αξιολόγησης των μοντέλων και, ιδίως, στην ενσωμάτωση προηγούμενων ιατρικών γνώσεων (Holzinger, 2021). Ένα παράδειγμα για τα αποτελέσματα αυτά είναι ο προσδιορισμός των οσφρητικών επιδράσεων διαφόρων φαρμάκων από ένα σύνολο δεδομένων με πολλά υποψήφια φάρμακα που εφαρμόστηκαν σε περιορισμένο αριθμό ασθενών, το οποίο έχει επίσης αξιολογηθεί τόσο από ειδικούς της βιοϊατρικής όσο και από ειδικούς της επιστήμης των υπολογιστών (Lötsch et al., 2015).

ΧΑΙ στις βιοϊατρικές εκδόσεις

Η τεχνητή νοημοσύνη και ιδιαίτερα η μηχανική μάθηση έχουν επεκταθεί σε πολλούς τομείς της καθημερινής ζωής και της επιστήμης, συμπεριλαμβανομένης και της βιοϊατρικής έρευνας. Στη βιοϊατρική έρευνα, η μηχανική μάθηση χρησιμοποιείται για την ανάλυση των δεδομένων υγείας, την πρόβλεψη διαγνώσεων και την ανάπτυξη νέων θεραπειών. Τα αλγόριθματα μηχανικής μάθησης μπορούν να ανακαλύψουν πρότυπα και συσχετίσεις σε μεγάλα σύνολα δεδομένων που θα μπορούσαν να ξεφύγουν από την ανθρώπινη ανάλυση. Για παράδειγμα, μπορούν να βοηθήσουν στην πρόβλεψη του κινδύνου ανάπτυξης μιας ασθένειας ή στην εξατομίκευση της θεραπείας βάσει των γονιδιακών προφίλ του ασθενούς.

Μια αυτοματοποιημένη αναζήτηση στη βάση δεδομένων PubMed στις 25 Σεπτεμβρίου 2021, με τη χρήση της βιβλιοθήκης R "RISmed" (<https://cran.r-project.org/package=RISmed> (Kovalchik, 2021), πρόσβαση στις 15 Δεκεμβρίου 2021) απέδωσε 166.938 επιτυχίες με τους όρους αναζήτησης ("machine-learning OR artificial intelligence OR explainable artificial intelligence") και 138.556 επιτυχίες με τους όρους αναζήτησης ("artificial intelligence OR explainable artificial intelligence"). Όταν αποκλείονται οι κριτικές με την προσθήκη "NOT review [PT]" στους όρους αναζήτησης, προέκυψαν 153.868 και 127.438 επιτυχίες, αντίστοιχα. Η παλαιότερη επιτυχία με τον όρο MeSH "τεχνητή νοημοσύνη" ήταν μια έκθεση του 1951 για μια νευρολογική έρευνα ρομπότ (Fletcher, 1951). Οι δημοσιεύσεις ανά έτος ήταν σπάνιες μέχρι τη δεκαετία του 1980 και δεν ξεπερνούσαν τις 100 ανά ετησίως μέχρι το 1986 (Σχήμα 4Α).

Η ένταξη του XAI στη βιοϊατρική έρευνα αποτελεί σχετικά πρόσφατη ανάπτυξη. Η αυτοματοποιημένη αναζήτηση στη βάση δεδομένων PubMed με τη χρήση του όρου "εξηγήσιμη τεχνητή νοημοσύνη" ανέδειξε 340 εγγραφές μέχρι τις 25 Σεπτεμβρίου 2021, με την πρώτη έρευνα να χρονολογείται από το 1990. Ωστόσο, παρατηρείται αύξηση του ενδιαφέροντος για το XAI στη βιοϊατρική, με τον αριθμό των δημοσιεύσεων να αυξάνεται σημαντικά τα τελευταία χρόνια. Συγκεκριμένα, το 2020 καταγράφηκαν 113 άρθρα, ενώ μόνο το 2021 ο αριθμός αυτός ανήλθε ήδη σε 172 άρθρα. Αυτό ανταποκρίνεται στις ημερομηνίες δημοσίευσης των βασικών άρθρων που αναφέρονται στις προηγούμενες ενότητες σχετικά με τις έννοιες της τεχνητής νοημοσύνης και του εξηγήσιμου τεχνητού συστήματος νοημοσύνης.



Εικόνα 4: Στοιβαγμένο ραβδόγραμμα των δημοσιεύσεων που παρατίθενται στο PubMed ανά έτος, με ιδιαίτερη έμφαση στις δημοσιεύσεις που βρέθηκαν με τον όρο αναζήτησης ("machine-learning OR artificial intelligence OR explainedable artificial intelligence") NOT (review[PT]), με το ποσοστό των δημοσιεύσεων που βρέθηκαν με τον όρο αναζήτησης ("explainedable artificial intelligence" NOT (review[PT])) να διαχωρίζεται ως μπλε τμήματα κάθε ράβδου. (Α) Οι δημοσιεύσεις ανά έτος ήταν σπάνιες μέχρι τη δεκαετία του 1980 και δεν ξεπέρασαν τις 100 ανά έτος μέχρι το 1986. (Β) Διευρυμένη προβολή της τελευταίας αναζήτησης, δηλαδή μόνο για το "XAI". Το σχήμα δημιουργήθηκε με τη χρήση του πακέτου λογισμικού R (έκδοση 4.1.2 για Linux-<http://CRAN.R-project.org/>) και της βιβλιοθήκης "ggplot2" (<https://cran.r-project.org/package=ggplot2>, όλα προσπελάστηκαν στις 15 Δεκεμβρίου 2021).

Πρακτικές επιπτώσεις του XAI στην υγειονομική περίθαλψη

Η επεξηγήσιμη τεχνητή νοημοσύνη (Explainable Artificial Intelligence - XAI) έχει κερδίσει όλο και μεγαλύτερη προσοχή στον τομέα της υγειονομικής περίθαλψης λόγω της δυνατότητάς της να βελτιώσει τα αποτελέσματα των ασθενών και να ενισχύσει τη λήψη ιατρικών αποφάσεων.

Μία από τις βασικές πρακτικές επιπτώσεις του XAI στην υγειονομική περίθαλψη είναι η δυνατότητα βελτίωσης των αποτελεσμάτων των ασθενών παρέχοντας πληροφορίες για τη διαδικασία λήψης αποφάσεων των μοντέλων AI. Οι τεχνικές XAI, όπως οι LIME, SHAP και Grad-CAM, έχουν επιτρέψει στους επαγγελματίες του τομέα της υγειονομικής περίθαλψης να αποκτήσουν μια βαθύτερη κατανόηση των χαρακτηριστικών που συμβάλλουν στις προβλέψεις του μοντέλου. Αυτή η ερμηνευσιμότητα δίνει τη δυνατότητα στους κλινικούς γιατρούς να κάνουν πιο τεκμηριωμένες και σίγουρες αποφάσεις, οδηγώντας σε ακριβείς διαγνώσεις, βελτιστοποιημένη θεραπεία σχέδια θεραπείας και καλύτερη φροντίδα των ασθενών.

Ανίχνευση και μετριασμός μεροληψίας

Ένας ακόμα πρακτικός ρόλος του εξηγήσιμου τεχνητού συστήματος νοημοσύνης (XAI) στην υγειονομική περίθαλψη είναι η συμβολή του στον εντοπισμό πιθανών προκαταλήψεων στα μοντέλα τεχνητής νοημοσύνης. Αυτό μπορεί να συμβάλει στην αποφυγή ανεπιθύμητων συνεπειών και στην αντιμετώπιση τυχόν προκαταλήψεων που

μπορεί να επηρεάσουν τα αποτελέσματα των μοντέλων Τεχνητής Νοημοσύνης. Με την αποκάλυψη των προκαταλήψεων στα δεδομένα και τις εξόδους των μοντέλων, το XAI επιτρέπει στα συστήματα υγειονομικής περίθαλψης να αντιμετωπίσουν ζητήματα που σχετίζονται με τη δικαιοσύνη και την ισότητα. Αυτό διασφαλίζει ότι η υγειονομική περίθαλψη με τεχνητή νοημοσύνη παρεμβάσεις είναι πιο περιεκτικές και παρέχουν ισότιμη θεραπεία για όλες τις ομάδες ασθενών.

Το εξηγήσιμο σύστημα νοημοσύνης (XAI) έχει συμβάλει στην υιοθέτηση της εξατομικευμένης ιατρικής στην υγειονομική περίθαλψη. Μέσω της διαφανούς και ερμηνεύσιμης φύσης των μοντέλων τεχνητής νοημοσύνης, η XAI επιτρέπει στους επαγγελματίες υγείας να προσαρμόσουν τα σχέδια θεραπείας σε μεμονωμένους ασθενείς με βάση το μοναδικό ιατρικό ιστορικό και τα χαρακτηριστικά τους. Αυτή η εξατομικευμένη προσέγγιση έχει οδηγήσει σε βελτιωμένη αποτελεσματικότητα της θεραπείας και αύξηση της ικανοποίησης των ασθενών.

Υπάρχουν τεκμηριωμένες περιπτώσεις και εμπειρικά στοιχεία από διάφορα ιδρύματα υγειονομικής περίθαλψης που δείχνουν ότι η εφαρμογή του εξηγήσιμου συστήματος νοημοσύνης έχει συμβάλει στη μείωση των ιατρικών σφαλμάτων και στην έγκαιρη ανίχνευση ασθενειών. Η ερμηνευσιμότητα των μοντέλων τεχνητής νοημοσύνης επιτρέπει στους παρόχους υγειονομικής περίθαλψης να εντοπίζουν πιθανά σφάλματα και να ανιχνεύουν ανωμαλίες στα ιατρικά δεδομένα, οδηγώντας σε έγκαιρες παρεμβάσεις και καλύτερα αποτελέσματα για τους ασθενείς.

Μελέτες περιπτώσεων και εμπειρικά στοιχεία

Υπάρχει πληθώρα περιπτωσιολογικών μελετών και εμπειρικών στοιχείων, που αποδεικνύουν τον πρακτικό αντίκτυπο του XAI στην υγειονομική περίθαλψη.

Οι πρακτικές επιπτώσεις του XAI στην υγειονομική περίθαλψη είναι εκτεταμένες και έχουν τεράστιες δυνατότητες για τον μετασχηματισμό των ιατρικών πρακτικών. Η διαφάνεια και η ερμηνευσιμότητα των μοντέλων τεχνητής νοημοσύνης που παρέχουν οι τεχνικές XAI έχουν βελτιώσει τα αποτελέσματα των ασθενών, διευκολύνοντας την εξατομικευμένη ιατρική, και μείωσαν τα ιατρικά σφάλματα. Η παρούσα ενότητα υπογραμμίζει τις πραγματικές οφέλη από την ενσωμάτωση της XAI σε περιβάλλοντα υγειονομικής περίθαλψης και υπογραμμίζει την σημασία της για την επανάσταση στις ιατρικές πρακτικές (Ali et al., 2023).

Παραδείγματα χρήσης XAI στην ιατρική επιστήμη

Η ιατρική απεικόνιση καλύπτει ένα ευρύ φάσμα μεθόδων, όπως οι συμβατικές ακτίνες X και οι υπέρηχοι. Η απόκτηση ιατρικών εικόνων συνήθως απαιτεί δαπανηρό εξοπλισμό και μπορεί να απαιτεί αρκετό χρόνο. Η διαθεσιμότητα συνόλων δεδομένων για θωρακική τομογραφία (CT), ακτινογραφίες θώρακα και μαστογραφίες έχει οδηγήσει την έρευνα σε πνευμονικές παθήσεις και καρκίνο του μαστού (Chan et al., 2020). Τα σύνολα δεδομένων, τα μοντέλα DNN και οι τεχνικές XAI καλύπτουν ένα ευρύ φάσμα έρευνας ασθενειών και αφορούν ειδικότητες όπως η ακτινολογία και η οφθαλμολογία. Ωστόσο, η ακτινολογία φαίνεται να έχει μεγαλύτερο μερίδιο και, συνεπώς, μεγαλύτερη ανάγκη για την εφαρμογή τεχνικών XAI (Eldridge, 2021).

Συμβατική απεικόνιση/ακτίνες

Η συμβατική ακτινογραφία CXR είναι ένας φθηνός και άμεσα διαθέσιμος τρόπος απεικόνισης (Singh et al., 2021). Η CXR έχει χρησιμοποιηθεί εκτενώς για τη διάγνωση του COVID-19. Η ανίχνευση του COVID-19 σε εικόνες CXR με ένα καλά ρυθμισμένο EfficientNet-B7 πέτυχε ακρίβεια 95,01% και βαθμολογίες AUC 0,950, οι οποίες ήταν καλύτερες σε σύγκριση με πολλές άλλες προσεγγίσεις DNN για την ανίχνευση του COVID-19 (Chetoui and Akhloufi, 2021). Το Grad-CAM χρησιμοποιήθηκε για την κατανόηση της συμβολής των σημαντικών περιοχών στην πρόβλεψη του μοντέλου (Chetoui and Akhloufi, 2021). Μια ταξινόμηση πολλαπλών κατηγοριών εικόνων του συνόλου δεδομένων NIH CXR σε τρεις κατηγορίες, COVID-19, πνευμονία και φυσιολογική, χρησιμοποίησε την αρχιτεκτονική VGG-16 και VGG-19 CNN για την ταξινόμηση και το Grad-CAM για την ερμηνευσιμότητα (Lee et al., 2020). Παρατηρήθηκε ότι η παρουσία ξένων ουσιών στην πνευμονική κοιλότητα οδήγησε σε εσφαλμένη ταξινόμηση των εικόνων, μερικές φορές λόγω λανθασμένων λόγων (Lee et al., 2020). Ένα πλαίσιο για την πολυταξινόμηση σε τρεις κλάσεις με ένα σύνολο CNN και Naive Bayes ως ταξινομητή χρησιμοποίησε το Grad-CAM για την οπτικοποίηση (Singh et al., 2021). Το XAI χρησιμοποιήθηκε για την ανίχνευση πνευμονικής νόσου και του COVID-19 με εικόνες CXR χρησιμοποιώντας μεταφορά μάθησης με το μοντέλο VGG-16 (Brunese et al., 2020). Χρησιμοποιήθηκαν δύο μοντέλα, το πρώτο για να διάκριση πνευμονικής νόσου και

στη συνέχεια για να διαπιστωθεί εάν εκδηλώνεται το COVID-19 (Brunese et al., 2020). Το Grad-CAM χρησιμοποιήθηκε για τη δημιουργία χαρτών ενεργοποίησης για COVID-19, επισημαίνοντας τις σωστές περιοχές στους χάρτες ενεργοποίησης (Brunese et al., 2020). Για τον εντοπισμό και τη διαβάθμιση της σοβαρότητας του COVID-19, οι χάρτες ενεργοποίησης θα μπορούσαν να είναι ακατάλληλοι για κλινική χρήση, καθώς θα μπορούσαν να παρέχουν ασυνεπή εντοπισμό (Degerli et al., 2021). Για τη διάγνωση του COVID-19 προτάθηκαν χάρτες μόλυνσης από εικόνες CXR για να ξεπεραστούν τα προβλήματα (Degerli et al., 2021). Οι χάρτες μόλυνσης θα μπορούσαν επίσης να χρησιμοποιηθούν για τον προσδιορισμό της εξέλιξης χρησιμοποιώντας τη χρονοσειρά δεδομένων CXR (Degerli et al., 2021).

Το Counterfactual Generative Network (CGN) προτάθηκε για την οπτικοποίηση ιατρικών εικόνων σε ένα σύνολο δεδομένων CXR για το COVID-19 [165]. Οι αντιπραγματικές προβλέψεις βλαβών ενσωματώθηκαν ως προηγούμενες συνθήκες για τη δημιουργία αντιπραγματικών εικόνων βλαβών (Kim et al., 2021). Εκτός από τους οπτικούς χάρτες απόδοσης, το U-Net χρησιμοποιήθηκε ως γεννήτρια που περιλαμβάνει έναν κωδικοποιητή και έναν αποκωδικοποιητή για τη δημιουργία των αντιφατικών εικόνων (Kim et al., 2021). Οκτώ μοντέλα CNN χρησιμοποιήθηκαν για την ταξινόμηση εικόνων αξονικής τομογραφίας COVID-19 με το LIME για να βοηθήσουν στον εντοπισμό συγκεκριμένων διαφοροποιητικών χαρακτηριστικών. Προτάθηκε ένα σύστημα βασισμένο σε ένα σύνολο DNN με δυνατότητα επεξήγησης, το DeepCOVIDExplainer, για την ανίχνευση COVID-19 σε εικόνες CXR από το σύνολο δεδομένων "COVIDx v2.0" (Karim et al., 2020). Η επεξηγηματικότητα των μοντέλων εξερευνήθηκε μέσω των LRP, Grad-CAM και Grad-CAM++ για τον εντοπισμό των περιοχών που είναι σημαντικές για την ταξινόμηση. Το ατομικό DNN με τις καλύτερες επιδόσεις αρχιτεκτονικές ήταν η VGG-19 και η DenseNet-161, και τα heatmap που παράγονται από την Grad-CAM++ ήταν οι πιο αξιόπιστοι σε σύγκριση με τα Grad-CAM και LRP (Karim et al., 2020).

Προτάθηκε ένα επεξηγηματικό πλαίσιο ομαδοποίησης για την ομαδοποίηση των εικόνων CXR με βάση την έκταση της λοίμωξης και το επίπεδο σοβαρότητας για να βοηθηθούν οι κλινικοί γιατροί στις αποφάσεις τους. Το DeepSHAP χρησιμοποιήθηκε για την οπτικοποίηση των περιοχών διάκρισης για τη σοβαρότητα της νόσου (Ravi et al., 2020). Το μοντέλο χρησιμοποιήθηκε για την ταξινόμηση COVID και πνευμονίας και πέτυχε ακρίβεια 95% και 97% αντίστοιχα (Ravi et al., 2020). Το VGG-19 με μια

προσέγγιση μάθησης μεταφοράς προτάθηκε για την ανίχνευση του COVID-19 χρησιμοποιώντας τις εικόνες CXR και CT από τα τρία σύνολα δεδομένων, COVID-chest X-ray, SARS-COV-2 CT-scan και Chest X-ray images (Pneumonia) dataset (Panwar et al., 2020). Η απόδοση του μοντέλου ήταν πιο ακριβής στην αξονική τομογραφία σε σύγκριση με τις εικόνες CXR και είχε ακρίβεια 95,61% στην ανίχνευση του COVID-19. Οι προβλέψεις του μοντέλου απεικονίστηκαν με τη χρήση του Grad-CAM (Panwar et al., 2020). Ένα μοντέλο CNN CovXNet χρησιμοποιήθηκε για δυαδική και πολυταξική ταξινόμηση με συνέλιξη σε βάθος με μεταβαλλόμενους ρυθμούς διαστολής που βοήθησε στην εξαγωγή ποικίλων χαρακτηριστικών από την CXR (Mahmud et al., 2020). Η ταξινόμηση πολλαπλών κατηγοριών (COVID, φυσιολογική, ιογενής ή βακτηριακή πνευμονία) πέτυχε ακρίβεια 90,2 %. Το Grad-CAM χρησιμοποιήθηκε για τον προσδιορισμό των περιοχών της εικόνας CXR που συνέβαλαν στην απόφαση του μοντέλου (Mahmud et al., 2020). Προτάθηκε ένα δίκτυο Bayes, με την ονομασία MulNet, για την ανίχνευση πνευμονίας αξιοποιώντας δεδομένα πολλαπλών πηγών συνδυάζοντας DNN και ένα εξηγήσιμο μοντέλο (Ren et al., 2021).

Ένα δίκτυο καθοδηγούμενο από την προσοχή προτάθηκε για τις θέσεις των βλαβών για την ταξινόμηση θωρακικών ασθενειών με πολλές ετικέτες σε εικόνες CXR (Chen et al., 2019). Η μέθοδος παρείχε ένα μέσο σκορ AUC 0,824 και για τις 14 παθολογίες. Το Grad-CAM χρησιμοποιήθηκε για τον εντοπισμό των βλαβών (Chen et al., 2019). Προτάθηκε ένα δίκτυο βαθιάς σύντηξης βασισμένο στην τμηματοποίηση με δύο μοντέλα ταξινόμησης CNN ως εξολκείων χαρακτηριστικών για την εξαγωγή των περιοχών διάκρισης που χρησιμοποιήθηκαν για την ταξινόμηση θωρακικών ασθενειών σε εικόνες CXR. Η μέθοδος πέτυχε καλύτερες επιδόσεις ταξινόμησης σε σύγκριση με τις άλλες προσεγγίσεις που βασίζονται στο ROC (Liu et al., 2019). Η ανίχνευση των βλαβών αξιολογήθηκε με τη χρήση CAM και βρέθηκε αξιόπιστη (Liu et al., 2019). Η επεξηγηματικότητα θα μπορούσε να ενισχυθεί με την παιχνιδοποίηση. Οι αναδυόμενες γλώσσες περιλαμβάνουν παιχνίδια επικοινωνίας πολλαπλών πρακτόρων με συντεταγμένες αποστολέα/δέκτη για την ολοκλήρωση μιας εργασίας και έχει αποδειχθεί ότι ενισχύουν τις ικανότητες των νευρωνικών δικτύων (Chowdhury et al., 2021). Οι αναδυόμενες συμβολικές γλώσσες αποτελούν τους ακρογωνιαίους λίθους των συστημάτων Good Old-Fashioned AI (GOFAI) τα οποία είναι επεξηγήσιμα (Chowdhury et al., 2021).

Η διάγνωση παιδιατρικής πνευμονίας προτάθηκε με τη χρήση ενός συνόλου CNN (Liz et al., 2021). Η προτεινόμενη τεχνική XAI χρησιμοποιήθηκε για τον συνδυασμό των θερμικών χαρτών από κάθε ένα από τα μοντέλα του συνόλου (Liz et al., 2021). Η μέθοδος χρησιμοποιήθηκε για τον προσδιορισμό της πιθανότητας η CXR να έχει παγίωση ή μη παγίωση (άλλες διηθήσεις), γεγονός που θα μπορούσε να είναι χρήσιμο για τους κλινικούς ιατρούς (Liz et al., 2021). Τα σύνολα ήταν χρήσιμα για τη βελτίωση των αποτελεσμάτων ταξινόμησης τόσο για τη βακτηριακή όσο και για την ιογενή πνευμονία (Liz et al., 2021). Ωστόσο, το σύνολο δεδομένων ήταν μικρό με εικόνες χαμηλής ανάλυσης (Liz et al., 2021). Ένα προ-εκπαιδευμένο μοντέλο VGG16 και ένα προσαρμοσμένο μοντέλο χρησιμοποιήθηκαν για την ταξινόμηση της ιογενούς και βακτηριακής παιδιατρικής πνευμονίας σε εικόνες CXR (Rajaraman et al., 2019). Οι ενεργοποιήσεις του μοντέλου οπτικοποιήθηκαν με τη χρήση των LIME και Grad-CAM και διαπιστώθηκε ότι αναδεικνύουν τις περιοχές διάκρισης στην εικόνα CXR (Rajaraman et al., 2019). Μια άλλη μελέτη για τη διάγνωση παιδιατρικής πνευμονίας χρησιμοποίησε ένα σύνολο πέντε διαφορετικών μοντέλων DNN για την ταξινόμηση της CXR στο σύνολο δεδομένων XrPP (Liz et al., 2021). Η μελέτη έδειξε ότι το σύνολο παρείχε καλύτερα αποτελέσματα σε σύγκριση με τη μάθηση μεταφοράς και πέτυχε AUC 0,92 (Liz et al., 2021). Τα θερμικά διαγράμματα προέκυψαν από το συνδυασμό των μεμονωμένων θερμικών διαγραμμάτων για κάθε από τα μοντέλα του συνόλου (Liz et al., 2021). Οι εξηγήσεις με τη χρήση της διδασκαλίας Bayesian στην CXR ελήφθησαν για τη διάγνωση του πνευμοθώρακα και έδειξαν ότι μπορεί να βοηθήσει τους ειδικούς της ιατρικής να προβλέψουν τις αποφάσεις της AI (Folke et al., 2021). Η μελέτη συνδύασε τους χάρτες αλατότητας και την Bayesian Teaching και διερεύνησε τη σχέση μεταξύ της ταξινόμησης της AI, της διάγνωσης του ακτινολόγου και των προβλέψεων του ακτινολόγου (Folke et al., 2021). Η πρόβλεψη κακοήθειας σε ένα σύνολο δεδομένων για τον καρκίνο του πνεύμονα προέκυψε από τα βάρη των διανυσμάτων κάψουλας που υποδεικνύουν τη συμβολή ενός οπτικού χαρακτηριστικό στην κακοήθεια του όζου (LaLonde et al., 2020). Οι πληροφορίες από το παιδί κάψουλα προς τη γονική κάψουλα δρομολογήθηκαν με τη χρήση ενός νέου σιγμοειδούς δρομολόγησης συνάρτησης (Nazir et al., 2022). Οι εικόνες CXR χρησιμοποιήθηκαν για την πρόβλεψη του καρκίνου του πνεύμονα χρησιμοποιώντας την προτεινόμενη τεχνική X-Caps XAI που παράγει βαθμολογίες κακοήθειας με εξηγήσεις σε ανθρώπινο επίπεδο που ήταν καλύτερα ερμηνεύσιμες από τους ακτινολόγους (LaLonde et al., 2020). Η

καρδιομεγαλία διαγνώστηκε με τη χρήση CNN και επεξηγήσιμων χαρακτηριστικών χάρτη για επεξηγηματικότητα με εικόνες CXR [236]. Χρησιμοποιήθηκαν εικόνες CXR για την ανάγνωση της καρδιακής υπερτροφίας (Yoo et al., 2021). Λειτουργική ακρίβεια του CNN χρησιμοποιήθηκε για την αξιολόγηση του μοντέλου (Yoo et al., 2021). Ο χάρτης χαρακτηριστικών έδειξε θετική τιμές για τα τμήματα με σημαντική επιρροή στη διάγνωση της νόσου (Yoo et al., 2021).

Το CAM χρησιμοποιήθηκε για τη δημιουργία των θερμικών χαρτών και το σύστημα παρείχε καλά αποτελέσματα ακρίβειας παρά το μικρό σύνολο δεδομένων (Yi et al., 2019).

Υπολογιστική Τομογραφία (CT)

Το LIME χρησιμοποιήθηκε για να βοηθήσει στον εντοπισμό των συγκεκριμένων χαρακτηριστικών που συνέβαλαν στη διαφοροποίηση των εικόνων αξονικής τομογραφίας για το COVID-19 και των φυσιολογικών εικόνων με τη χρήση 8 μοντέλων CNN (Ahsan et al., 2020). Το μοντέλο CNN για την ταξινόμηση των εικόνων αξονικής τομογραφίας COVID-19 χρησιμοποίησε το t-SNE (μείωση διαστάσεων) και το Grad-CAM που έδειξε καλά διαχωρισμένες συστάδες για τις περιπτώσεις COVID-19 και μη COVID-19 (Alshazly et al., 2021). Οι εικόνες αξονικής τομογραφίας COVID-19 ταξινομήθηκαν με το U-Net επιτυγχάνοντας καλύτερες επιδόσεις (ακρίβεια, ακρίβεια, ειδικότητα και AUC) από άλλες σύγχρονες μεθόδους, όπως το COVID-Net και το COVNet (Yang et al., 2022). Το LIME χρησιμοποιήθηκε για την εξήγηση κάθε πρόβλεψης με τη διαίρεση μιας εικόνας σε super-pixel. SHAP μέθοδος χρησιμοποιήθηκε επίσης για την επεξηγηματικότητα για την ποσοτική εκτίμηση της συνεισφορά κάθε υπερ-πίξελ (Yang et al., 2022). Ένα σύστημα ανίχνευσης και κατηγοριοποίησης βλαβών COVID-19 προτάθηκε για αξονικές τομογραφίες συνδυάζοντας δίκτυα τμηματοποίησης και ταξινόμησης (Pennisi et al., 2021). Αποδείχθηκε ότι η τμηματοποίηση των πνευμόνων και των λοβών βελτίωσε τα αποτελέσματα ταξινόμησης κατά περισσότερο από 6% (Pennisi et al., 2021). Μια πνευμονική βλάβη προτάθηκε η ταξινόμηση της κακοήθειας με τη χρήση ενεργοποίησης CycleGAN για την απεικόνιση της απόφασης του ταξινομητή LIDC-IDRI δεδομένων. Ένα σύστημα κοινής ταξινόμησης και τμηματοποίησης (JCS) προτάθηκε για τη διάγνωση και την εξηγησιμότητα του COVID-19 με CT διάγνωση (Wu et al., 2021). Κατασκευάστηκε επίσης ένα μεγάλο σύνολο δεδομένων COVID-CS με σημειώσεις σε επίπεδο ασθενούς και εικονοστοιχείου (Wu et al.,

2021). Χρησιμοποιήθηκαν χάρτες ενεργοποίησης για την εξηγησιμότητα των προβλέψεων και οι βαθμολογίες Dice για την τμηματοποίηση (Wu et al., 2021). Οι εικόνες αξονικής τομογραφίας από πολλαπλές πηγές αναμείχθηκαν για να αποφευχθεί η υπερπροσαρμογή και η μεροληψία (Wu et al., 2021).

Ένα πλαίσιο DNN με τέσσερα προεκπαιδευμένα DNN χρησιμοποιήθηκε για το COVID-19 δυαδική ταξινόμηση σε εικόνες αξονικής τομογραφίας (Ko et al., 2020). Το πλαίσιο δοκιμάστηκε σε εικόνες αξονικής τομογραφίας χαμηλής ποιότητας και το ResNet-50 είχε την καλύτερη απόδοση μεταξύ των επιλεγμένων DNN με ακρίβεια 99,87% και ευαισθησία 99,58% (Ko et al., 2020). Το Grad-CAM χρησιμοποιήθηκε για την οπτικοποίηση των σχετικών περιοχών στις πρόβλεψες του μοντέλου (Ko et al., 2020). Προτάθηκε ένα δίκτυο προσοχής πολλαπλής κλίμακας που αποτελείται από στοιβαγμένες μονάδες προσοχής και μια μονάδα σύντηξης πολλαπλών κλιμάκων για την ταξινόμηση επτά τύπων πνευμονικών υφών σε υψηλή ανάλυση εικόνες αξονικής τομογραφίας (Xu et al., 2019). Το Grad-CAM χρησιμοποιήθηκε για τον προσδιορισμό της λειτουργίας του προτεινόμενου μοντέλου (Xu et al., 2019).

Προτάθηκε ένα πλαίσιο για την κατάτμηση όγκων αξονικής τομογραφίας με χρήση CNN και τη δημιουργία αναφοράς με χρήση γλωσσικού μοντέλου LSTM για όγκο στο ήπαρ. Το δημόσιο σύνολο δεδομένων της πρόκλησης MICCAI 2017 LiTS χρησιμοποιήθηκε για την αξιολόγηση ενός μοντέλου προσοχής με επίβλεψη για τη δημιουργία λεζάντας. Ένα CNN και ένα LSTM χρησιμοποιήθηκαν για την ταξινόμηση του προτύπου εμφυσήματος στο σύνολο δεδομένων αξονικής τομογραφίας COPDGene (Humphries et al., 2020). Οι βαθμολογίες του μοντέλου υπολογίστηκαν συνδυάζοντας τις εξόδους του DNN και συγκρίθηκαν με οπτικοποιήσεις Grad-CAM και κλινικές παραμέτρους (Humphries et al., 2020). Το μοντέλο πέτυχε συμφωνία με τις οπτικές βαθμολογίες εμφυσήματος στη δοκιμαστική ομάδα (Humphries et al., 2020). Εικόνες αξονικής τομογραφίας χρησιμοποιήθηκαν για τον εντοπισμό και την ταξινόμηση όγκων του πνεύμονα. Ένα σύνολο CNN χρησιμοποιήθηκε για την πρόβλεψη κακοήθειας πνευμονικών οζιδίων με τρεις αρχιτεκτονικές CNN, κάθε μία από τις οποίες εκπαιδεύτηκε με επτά διαφορετικούς σπόρους για τα αρχικά βάρη (Paul et al., 2020). Διαπιστώθηκε ότι το σύνολο βελτιώνει την ταξινόμηση του συστήματος και πέτυχε AUC 90,29% στην πρόβλεψη του οζίδιου που πιθανώς θα διαγνωστούν ως καρκινικά σε δύο χρόνια. Το Grad-CAM χρησιμοποιήθηκε για την οπτικοποίηση των ενεργοποιήσεων του μοντέλου για το

εικόνες εισόδου (Paul et al., 2020). Εντοπισμός και ταξινόμηση οζιδίων του πνεύμονα σε εικόνες αξονικής τομογραφίας χαμηλής δόσης (δημόσιο σύνολο δεδομένων LIDC-IDRI) υλοποιήθηκε με τη χρήση χαρτογράφηση μαλακής ενεργοποίησης για ανάλυση λεπτόκοκκων χαρακτηριστικών με ένα CNN (Lei et al., 2020). Οι απεικονίσεις των χαρακτηριστικών των πνευμόνων προέκυψαν με τη χρήση CAM, Grad-CAM και της προτεινόμενης μεθόδου οπτικοποίησης. Τα αποτελέσματα ήταν ενσωματώθηκαν για να αυξήσουν το επίπεδο εμπιστοσύνης των ακτινολόγων για την κλινική διάγνωση (Lei et al., 2020). Προτάθηκε μια στοιβαγμένη ερμηνεύσιμη αρχιτεκτονική για την ταξινόμηση των ανώμαλων πνευμονικών όζων για την πρόβλεψη του καρκίνου του πνεύμονα σε LIDC-IDRI (Kumar et al., 2019). Η προτεινόμενη αρχιτεκτονική παρείχε καλύτερη γνώσεις σχετικά με τις αποφάσεις του μοντέλου και πέτυχε καλύτερες επιδόσεις σε σύγκριση με τις εξεταζόμενες προσεγγίσεις. Το προτεινόμενο ραδιομικροσκοπικό αλληλουχία πέτυχε σωστές προβλέψεις που μπορούν δυνητικά να βελτιώσουν κλινική υιοθέτηση (Kumar et al., 2019).

Μια μέθοδος απόδοσης εννοιών εμπνευσμένη από τη μεγιστοποίηση της ενεργοποίησης (DeepDream) και παρόμοια με την TCAV [88] προτάθηκε για την κατάτμηση όγκων ήπατος σε εικόνες αξονικής τομογραφίας στη βάση δεδομένων LiTS (Couteaux et al., 2019). Η μέθοδος παρείχε πληροφορίες για τη σημασία των χαρακτηριστικών υψηλού επιπέδου στο εκπαιδευμένο μοντέλο CNN (Couteaux et al., 2019). Το CNN χρησιμοποιήθηκε για τον αυτοματοποιημένο εντοπισμό και την ταξινόμηση ηπατικών βλαβών σε φασματικές εικόνες αξονικής τομογραφίας. Τα RoIs από τις εργασίες εντοπισμού χρησιμοποιήθηκαν για την εκπαίδευση ενός άλλου CNN για την ταξινόμηση (υγιής, υπερπυκνή μετάσταση, κύστη) (Shapira et al., 2020). Τα heatmaps χρησιμοποιήθηκαν για οπτικοποιήσεις και απέδωσαν καλά για την ανίχνευση μικρότερων βλαβών (Shapira et al., 2020). Προτάθηκε ένα δίκτυο διπλής προσοχής Dilated Residual Network (DADRN) για την ταξινόμηση των ηπατικών βλαβών σε ένα σύνολο δεδομένων εικόνων αξονικής τομογραφίας και πέτυχε ακρίβεια συγκρίσιμη με τις μεθόδους που βασίζονται σε RoI (Chen et al., 2019), το Grad-CAM χρησιμοποιήθηκε για τις επεξηγήσεις των αποτελεσμάτων ταξινόμησης και διαπιστώθηκε ότι το προτεινόμενο σύστημα αναγνωρίζει τις μικρού μεγέθους βλάβες (Chen et al., 2019).

Ένα τρισδιάστατο DNN, το AppendiXNet, προτάθηκε για την ανίχνευση της σκωληκοειδίτιδας με ένα μικρό σύνολο δεδομένων για εξετάσεις αξονικής τομογραφίας (Rajpurkar et al., 2020). Το σύστημα εκπαιδεύτηκε σε βίντεο του YouTube (Kinetics) και χρησιμοποιήθηκε το Grad-CAM για την ερμηνευσιμότητα των προβλέψεων του μοντέλου (Rajpurkar et al., 2020). Η απόδοση του μοντέλου βρέθηκε συγκρίσιμη με παρόμοια μοντέλα 3D CNN και είχε υψηλή διαγνωστική απόδοση για την ανίχνευση οξείας σκωληκοειδίτιδας (Rajpurkar et al., 2020). Η περίθαλψη του καρκίνου του νεφρού προβλέπεται να κοστίζει περίπου 5,1 δισεκατομμύρια δολάρια στις ΗΠΑ έως το 2020 (Kunapuli et al., 2018). Τα δεδομένα προέρχονταν από μια μελέτη για πολυφασικό υπολογιστικό σύστημα ενισχυμένο με αντίθεση (Contrast-enhanced Computed Τομογραφία (CECT) της κοιλιάς και της πύελου με διαβάθμιση εμπειρογνομόνων. αξιολόγηση της ιστολογίας (Kunapuli et al., 2018).

Μια προσαρμοσμένη αρχιτεκτονική ResNet χρησιμοποιήθηκε στο μητρώο MR CLEAN σύνολο δεδομένων εικόνων αγγειογραφίας αξονικής τομογραφίας και διαπιστώθηκε ότι η αυτοματοποιημένη μέθοδος DNN ήταν καλύτερη από τους βιοδείκτες ακτινολογικών εικόνων για το εγκεφαλικό επεισόδιο πρόβλεψη της έκβασης του εγκεφαλικού επεισοδίου και ήταν χρήσιμη στην επιλογή θεραπείας (Hilbert et al., 2019). Το Grad-CAM++ χρησιμοποιήθηκε για την οπτικοποίηση και το μοντέλο βρέθηκε να είναι χρήσιμο για την πρόβλεψη της έκβασης του εγκεφαλικού επεισοδίου (Hilbert et al., 2019). Αγγειογραφία στεφανιαίας αξονικής τομογραφίας (CCTA) χρησιμοποιήθηκαν εικόνες για την ανίχνευση της αθηροσκλήρωσης των στεφανιαίων αρτηριών και εντοπισμό σε σύνολα δεδομένων CCTA με 3D-CNN (Candemir et al., 2020). Το σύστημα ήταν αξιολογήθηκε και έδειξε AUC 0,91 (Candemir et al., 2020). Ο εντοπισμός που χρησιμοποιήθηκε Grad-CAM για την εξαγωγή οπτικοποιήσεων χαρτών θερμότητας χρήσιμων για τον ιατρό στον εντοπισμό της θέσης και της πιθανότητας αθηροσκλήρωσης (Candemir et al., 2020). Τρισδιάστατη προσοχή Το ταυτόσημο διπλό βαθύ δίκτυο (AID-Net) προτάθηκε για τη στεφανιαία αρτηρία Ασβέστιο (CAC) και ενσωματώθηκαν μηχανισμοί 3D προσοχής. για την ταξινόμηση σε σύνολο δεδομένων ImageVU (Huo et al., 2019). Το 3D Grad-CAM ήταν χρησιμοποιήθηκε για την κατανόηση της πρόβλεψης του DNN (Huo et al., 2019). Το προτεινόμενο σύστημα πέτυχε AUC 0,9627 και ακρίβεια ανίχνευσης 0,9272 (Huo et al., 2019).

Απεικόνιση μαγνητικού συντονισμού (MRI)

Το Αλτσχάιμερ είναι μια νευροεκφυλιστική νόσος και αφορά περίπου 10< εκατομμύρια νέες περιπτώσεις παγκοσμίως (Kamal et al., 2021). Τα δεδομένα εικόνας και γονιδιακής έκφρασης συνδυάστηκαν για μια πολυτροπική ανίχνευση της νόσου Alzheimer. Τα σημαντικά γονίδια εντοπίστηκαν με το LIME για την εξηγησιμότητα (Kamal et al., 2021). Το σύνολο δεδομένων μαγνητικής τομογραφίας ADNI χρησιμοποιήθηκε για την επεξηγηματική ταξινόμηση της νόσου Alzheimer με τη χρήση LRP για την οπτικοποίηση των αποφάσεων του CNN (Bohle et al., 2019). Αποδείχθηκε ότι η LRP ήταν χρήσιμη για την επεξηγηματικότητα των προβλέψεων ταξινόμησης της νόσου Alzheimer και εφαρμόσιμη σε παρόμοιες διαγνώσεις ασθενειών με χρήση δεδομένων MRI (Bohle et al., 2019).

Προτάθηκε μια μέθοδος που αποσκοπεί στην αύξηση της κλινικής ακρίβειας για τα πρώιμα στάδια της νόσου του Alzheimer και έδειξε ότι οι σχέσεις voxel σε διαφορετικές περιοχές μπορούν να παρέχουν διακριτικές πληροφορίες για την κλινική διάγνωση (Lee et al., 2021). Η μέθοδος ελέγχθηκε χρησιμοποιώντας ένα σύνολο δεδομένων μαγνητικής τομογραφίας και έδειξε καλύτερες επιδόσεις σε σχέση με τις μεθόδους Regional Mean Volume (RMV) και Hierarchical Feature Fusion (HFF). Οι επιδόσεις μετρήθηκαν χρησιμοποιώντας τέσσερις μετρικές, δηλαδή ακρίβεια, ευαισθησία, ειδικότητα και AUC, για τη δυαδική και πολυκατηγορική ταξινόμηση (Lee et al., 2021). Η μέθοδος παρείχε επίσης αναπαραστάσεις των περιφερειακών ανωμαλιών για την ερμηνευσιμότητα της πρόβλεψης (Lee et al., 2021).

Η νόσος του Πάρκινσον είναι μια νευροεκφυλιστική πάθηση που επηρεάζει κυρίως τον ηλικιωμένο πληθυσμό. Για τη δημιουργία δεικτών για αυτή τη νόσο χρησιμοποιήθηκε ένα Συνελικτικό Νευρωνικό Δίκτυο (CNN) με είσοδο την ευαίσθητη στη νευρομελανίνη εικόνα μαγνητικού συντονισμού (NMS-MRI). Το CNN πέτυχε καλύτερη ακρίβεια στη δοκιμή συγκριτικά με την ταξινόμηση βασισμένη στην αντίθεση αναλογίας (Shinde et al., 2019). Το CAM χρησιμοποιήθηκε για τον εντοπισμό των σχετικών και πιο διακριτικών περιοχών της εικόνας (Shinde et al., 2019). Το τρισδιάστατο CNN χρησιμοποιήθηκε για την ανίχνευση της νόσου του Πάρκινσον και την ταξινόμηση σαρώσεων μαγνητικής τομογραφίας από το σύνολο δεδομένων PPMI (Chakraborty et al., 2020). Το σύστημα πέτυχε ακρίβεια 95,29% και Receiver Operating Characteristics (ROC)-AUC 0,98 και για τις δύο κλάσεις. Η

μέθοδος CAM χρησιμοποιήθηκε για την επεξηγηματικότητα των προβλέψεων του μοντέλου σε την επιλογή των σχετικών περιοχών (Chakraborty et al., 2020).

Ένα πλαίσιο χρησιμοποίησε ογκομετρική μαγνητική τομογραφία νεογνών για την ταξινόμηση της εγκεφαλικής δυσλειτουργίας με 3D CNN και CAM, αναφέροντας ακρίβεια ταξινόμησης 98,5%. Η γήρανση του εγκεφάλου διερευνήθηκε με τη χρήση συνόλου 3D CNN για τον προσδιορισμό της πρόβλεψης της περιοχής του εγκεφάλου που συμβάλλει (Levakon et al., 2020). Χρησιμοποιήθηκε το SmoothGrad που υλοποιήθηκε με τη χρήση του iNNvestigate και οι εξηγήσεις συγκεντρώθηκαν σε όλα τα δείγματα (Levakon et al., 2020). Από τις συγκεντρωτικές εξηγήσεις προέκυψε το συμπέρασμα ότι οι κυψέλες και οι κοιλίες διέπουν τις προβλέψεις ηλικίας (Levakon et al., 2020). Η πλάκα σκλήρυνσης είναι μια νευρολογική πάθηση που συχνά διαγνώσκεται μέσω εικόνων μαγνητικής τομογραφίας, οι οποίες χρησιμοποιούνται για τον εντοπισμό πιθανών βλαβών στη λευκή ουσία του εγκεφάλου (Lopatina et al., 2020). Το CNN χρησιμοποιήθηκε για την ταξινόμηση των εικόνων και διάφοροι αλγόριθμοι απόδοσης χρησιμοποιήθηκαν για τη δημιουργία heatmaps για την απεικόνιση της συμβολής κάθε voxel στην πρόβλεψη της κλάσης (Lopatina et al., 2020). Συγκρίθηκαν διάφορες μέθοδοι και διαπιστώθηκε ότι η DeepLIFT ήταν προτιμότερη από την LRP (Lopatina et al., 2020). Τρισδιάστατες εικόνες με ζυγισμένη σε T1 αντίσταση, από πολλαπλές ακολουθίες ηχοκυμάτων, προήλθαν από έναν σαρωτή μαγνητικής τομογραφίας. Με τη χρήση ανάλυσης συνάφειας, συμπεραίνεται ότι τα σχετικά voxels βρίσκονταν κοντά στα φλεβικά αγγεία (Lopatina et al., 2020). Τα 3D CNN χρησιμοποιήθηκαν για τη διάγνωση της σκλήρυνσης κατά πλάκας σε σύνολο δεδομένων MRI και η LRP χρησιμοποιήθηκε για την επικύρωση των αποφάσεων του μοντέλου (Eitel et al., 2019). Η μελέτη κατέληξε στο συμπέρασμα ότι το πλαίσιο και το LRP παρείχαν διαφάνεια στην απόφαση του μοντέλου (Eitel et al., 2019).

Προτάθηκε ένα Subtractive Spatial Lightweight CNN (SSLW-CNN) με σκοπό την ανάπτυξη της επεξηγηματικότητας στην ιατρική ανάλυση και την αξιολόγηση της εμπιστοσύνης στην ταξινόμηση όγκων εγκεφάλου σε εικόνες MRI (Kumar et al., 2021). Τα κακοήγη κύτταρα του όγκου χωρίστηκαν σε πέντε κατηγορίες (Kumar et al., 2021). Το μοντέλο χρησιμοποίησε CAM για την εύρεση των σχετικών χαρακτηριστικών και οι τελικές απεικονίσεις έδειξαν πώς το σύστημα κατέληξε στην απόφασή του για την ταξινόμηση των όγκων (Kumar et al., 2021). Το 3D CNN

χρησιμοποιήθηκε για την ταξινόμηση όγκων στο σύνολο δεδομένων BraTS για την επιτάχυνση της αξιολόγησης όγκων για ολόκληρο τον εγκέφαλο και μια καθορισμένη περιοχή όγκου και έδειξε ότι τα αποτελέσματα με καθορισμένο RoI πέτυχαν καλύτερη ακρίβεια (Pereira et al., 2018). Η ερμηνευσιμότητα αξιολογήθηκε με Grad-CAM και Guided Backpropagation για την επικύρωση της αξιοπιστίας (Pereira et al., 2018). Μια προ-εκπαίδευση μοντέλο ResNet50 χρησιμοποιήθηκε στο σύνολο δεδομένων IXI από τομές μαγνητικής τομογραφίας για μια πολυταξική ταξινόμηση σε γλοιοβλάστωμα, αιθουσαίο σβάννωμα ή χωρίς όγκο (Windisch et al., 2020). Το Grad-CAM εφαρμόστηκε για να παρέχει οπτικοποιήσεις που βοήθησαν στον εντοπισμό της τοποθεσίας του όγκου. Η μελέτη κατέληξε στο συμπέρασμα ότι με την ενσωμάτωση μοντέλου εξηγησιμότητας στην πρώιμη εκπαίδευση του μοντέλου μπορεί να εξοικονομήσει χρόνο και να βοηθήσει τους κλινικούς γιατρούς στη χρήση της πρόβλεψης ενός μοντέλου για τη λήψη κλινικών αποφάσεων (Windisch et al., 2020).

Προτάθηκε μια μετρική για τη σημασία των χαρακτηριστικών που είναι ειδικά για κάθε τρόπο (MSFI). για την κωδικοποίηση μοτίβων εξήγησης και κλινικών εικόνων για συγκεκριμένες εντοπισμό χαρακτηριστικών (Jin et al., 2022). Η ταξινόμηση του γλοιώματος πραγματοποιήθηκε με ταξινόμηση σε γλοιώματα χαμηλότερου και υψηλότερου βαθμού με βάση το BraTS 2020. πολυτροπικό σύνολο δεδομένων (Jin et al., 2022). Το MSFI αποδείχθηκε ότι είχε καλύτερες επιδόσεις σε σύγκριση με σε σχέση με τους άλλους επιλεγμένους αλγορίθμους post-hoc heatmap για εργασίες πολυτροπικής απεικόνισης (Jin et al., 2022).

Η κατάτμηση του όγκου του εγκεφάλου διερευνήθηκε για την αβεβαιότητα και την ερμηνευσιμότητα με την εκπαίδευση τριών διαφορετικών μοντέλων στο σύνολο δεδομένων BraTS 2018 (Natekar et al., 2020). Στόχος της μελέτης ήταν η εξαγωγή των οπτικοποιήσεων για το χαρακτηριστικό χάρτες που είναι χρήσιμοι για την ιατρική διάγνωση (Natekar et al., 2020). Η δομή του δικτύου και η μέθοδος Grad-CAM χρησιμοποιήθηκαν για την ανίχνευση των ενεργοποιήσεων και τη χωρική προσοχή, προκειμένου να κατανοήσουν πώς το μοντέλο προβαίνει σε εντοπισμό και τμηματοποίηση (Natekar et al., 2020). Το Comprehensive Attention CNN (CA-Net) χρησιμοποιήθηκε για την τμηματοποίηση πολλαπλών κατηγοριών της εμβρυϊκής μαγνητικής τομογραφίας (Gu et al., 2021). Το σύνολο δεδομένων περιελάμβανε 150 στοίβες από σαρώσεις εμβρυϊκής μαγνητικής τομογραφίας T2-weighted (Gu et al.,

2021). Η αξιολόγηση του προτεινόμενου συστήματος έδειξε σημαντική βελτίωση των επιδόσεων σε σχέση με το U-Net, μέσω της χρήσης ολοκληρωμένης καθοδήγησης της προσοχής με χωρική, κανάλι και κλίμακα προσοχής (Gu et al., 2021).

Η τμηματοποίηση της καρδιακής μαγνητικής τομογραφίας (CMR) πραγματοποιήθηκε για την εξαγωγή συμπερασμάτων σχετικά με τις διάφορες καρδιακές παθήσεις με τη χρήση ενός DNN και την ανακάλυψη και δοκιμή με διανύσματα ενεργοποίησης εννοιών (D-TCAV) για την ερμηνευσιμότητα των σημαντικών χαρακτηριστικών για τη διάγνωση καρδιακών παθήσεων (Janik et al., 2021). Η μελέτη χρησιμοποίησε το σύνολο δεδομένων MICCAI 2017 με την αρχιτεκτονική U-Net (Janik et al., 2021). Προτάθηκε μια προσέγγιση βασισμένη σε συγχώνευση μοντέλων με τη χρήση δεδομένων MRI και υπερήχων. Η επιλογή της κατάλληλης φέτας ή εικόνας MRI αποτελεί σημαντικό βήμα, και οι συγγραφείς πρότειναν να επιλεγεί η βέλτιστη φέτα MRI προκειμένου να βελτιωθεί η απόδοση του μοντέλου. Η χρήση εικόνων μαγνητικής τομογραφίας και αξονικής τομογραφίας για την τμηματοποίηση του υδροκεφάλου οδήγησε σε σημαντική βελτίωση σε εικόνες μαγνητικής τομογραφίας και αξονικής τομογραφίας χαμηλής ανάλυσης (Yang et al., 2022). Η γενίκευση και η αναπαράσταση του μοντέλου βελτιώθηκαν με τη μείωση της αβεβαιότητας του μοντέλου (Yang et al., 2022). Για την αξιολόγηση των αποτελεσμάτων τμηματοποίησης χρησιμοποιήθηκαν βαθμολογίες ζάριας (Yang et al., 2022).

Οι εικόνες υπερήχων δεν έχουν χρησιμοποιηθεί τόσο εκτενώς στην κλινική ανάλυση της τεχνητής νοημοσύνης όσο άλλες απεικονιστικές μέθοδοι, σύμφωνα με τη μελέτη του Komatsu et al., (2021). Αυτό μπορεί να οφείλεται στη δυσκολία ελέγχου της ποιότητας της εικόνας λόγω της χειροκίνητης λειτουργίας, της εξάρτησης από τον χειριστή και των ακουστικών σκιών (Komatsu et al., 2021). Για την ταξινόμηση του καρκίνου του προστάτη, ερευνητές χρησιμοποίησαν εικόνες υπερήχων και μαγνητικής τομογραφίας, συνδυάζοντας δεδομένα από διάφορες προεκπαιδευμένες προσεγγίσεις βαθιάς μάθησης με διάφορες ρηχές προσεγγίσεις μηχανικής μάθησης. Η μέθοδος επεξηγηματικότητας LIME χρησιμοποιήθηκε για να παρέχει ερμηνευσιμότητα, ενώ η προσέγγιση σύντηξης βρέθηκε να βελτιώνει την απόδοση του μοντέλου. Εικόνες υπερήχων χρησιμοποιήθηκαν για την ταξινόμηση των βλαβών του μαστού σε σύνολο δεδομένων BreastMNIST με μεγιστοποίηση της ενεργοποίησης CycleGAN για οπτικοποιήσεις των προβλέψεων του μοντέλου (Katzmann et al., 2021).

Η αξιολόγηση της ποιότητας του ταξινομητή πραγματοποιήθηκε χρησιμοποιώντας διάφορες μετρικές, συμπεριλαμβανομένης της ακρίβειας, της ειδικότητας, του συντελεστή συσχέτισης Matthews και της περιοχής κάτω από το καμπύλο (AUC). Αντίστοιχα, για την ποιοτική αξιολόγηση του ταξινομητή χρησιμοποιήθηκαν διαδικτυακές έρευνες και άλλες μεθόδους. Η μέθοδος είχε καλύτερες επιδόσεις σε σύγκριση με τις DeepTaylor, LRP και DeepSHAP όσον αφορά την ποιότητα της εικόνας και το σημασιολογικό νόημα (Katzmann et al., 2021). Χρησιμοποιήθηκε ένα σύνολο δεδομένων υπερήχων για τον καρκίνο του μαστού για τη διερεύνηση της ευπάθειας σε αντίπαλες επιθέσεις. Χρησιμοποιήθηκε μάθηση πολλαπλών εργασιών (MTL) για τη βελτίωση των ακρίβειας ταξινόμησης. Η διεπιστημονική πρόκληση DREAM ξεκίνησε για τη μείωση των ψευδώς θετικών ποσοστών στη διάγνωση του καρκίνου του μαστού με νικητήριες λύσεις βασισμένες σε CNN (Le et al., 2019).

Η υπερηχογραφία του μαστού και η μαγνητική τομογραφία μπορούν να παρέχουν σημαντικές πρόσθετες πληροφορίες για τη μαστογραφική διάγνωση (Le et al., 2019). Το Grad-CAM χρησιμοποιήθηκε για τη δημιουργία των θερμικών χαρτών για την καθοδήγηση των ανθρώπινων εμπειρογνομόνων (Quan et al., 2021). Οι εγκεφαλικές ανωμαλίες του εμβρύου μπορούν να συμβάλουν στη νευρολογική καθυστέρηση και στη νοητική υστέρηση. Στη μελέτη των Xie και συνεργατών του το 2020, τα επίπεδα του εμβρυϊκού εγκεφάλου ταξινομήθηκαν ως φυσιολογικά ή μη φυσιολογικά χρησιμοποιώντας ένα μοντέλο CNN σε εικόνες υπερήχων. Η μέθοδος Grad-CAM χρησιμοποιήθηκε για τον εντοπισμό και τη διάγνωση των βλαβών στον εγκέφαλο. Η εκτίμηση της περιφέρειας της κεφαλής του εμβρύου αποτελεί βασικό βιομετρικό μέγεθος για τον προσδιορισμό της υγείας του εμβρύου και την εκτίμηση της ηλικίας κύησης. Η εκτίμηση πραγματοποιήθηκε με τη χρήση εικόνων υπερήχων και Regression CNN (Zhang et al., 2022). Η ερμηνεία της ταξινόμησης CNN εξερευνήθηκε με τη χρήση διαφόρων μεθόδων χαρτών αλατότητας, και οι μέθοδοι επεξήγησης αξιολογήθηκαν με ανάλυση διαταραχών σύμφωνα με τη μελέτη των Zhang και συνεργατών του το 2022. Η αρχιτεκτονική Supervised Object Detection with Normal data Only (SONO) προτάθηκε για την ανίχνευση ανωμαλιών σε βίντεο εμβρυϊκών υπερήχων. Τα δεδομένα βίντεο μετατράπηκαν σε χρονογραμμές τύπου γραμμωτού κώδικα για να βοηθηθεί η επεξηγηματικότητα, η οποία διευκόλυνε τον εντοπισμό των υποδομών.

Αρχή φόρμας

Οπτική απεικόνιση

Δερματολογία

Η αυτοματοποιημένη κατάτμηση δερματικών βλαβών μπορεί να βοηθήσει στην έγκαιρη διάγνωση δερματικών ασθενειών (Gu et al., 2021). Για την τμηματοποίηση των δερματικών βλαβών με την προτεινόμενη τεχνική χρησιμοποιήθηκε η ISIC 2018, η οποία έχει 2594 εικόνες (Gu et al., 2021). Τα αποτελέσματα συγκρίθηκαν με αλγόριθμους όπως το DenseASPP, το RefineNet και το DeepLabv3+, και έδειξαν συγκρίσιμα αποτελέσματα, σύμφωνα με τη μελέτη των Gu και συνεργατών το 2021. Σημαιολογικά σημαντικές προ-εικόνες προέκυψαν με τη χρήση κλίσεων από τους εκτιμητές απώλειας στο ISIC σύνολο δεδομένων για το δέρμα, το οποίο περιλάμβανε επτά ομάδες ασθενειών, σύμφωνα με τη μελέτη των Narayanaswamy και συνεργατών το 2021. Το προτεινόμενο προσέγγιση αποδείχθηκε αποτελεσματική για τη δημιουργία αντιφατικών εξηγήσεων, οι οποίες αποτελούν σημαντικό εργαλείο για την εξήγηση των προβλέψεων του μοντέλου. Ένα DNN χρησιμοποιήθηκε ως ταξινομητής εικόνων δέρματος, ενώ το LIME σε συνδυασμό με τον κανόνα ABCD χρησιμοποιήθηκε για την εξήγηση του μοντέλου, σύμφωνα με τη μελέτη των Narayanaswamy και συνεργατών το 2021. Ο κανόνας ABCD της δερματοσκοπίας έχει γνωριστεί ότι υπερτερεί έναντι άλλων μεθόδων αξιολόγησης.

Για λόγους επεξηγηματικότητας, το μοντέλο LIME συγχωνεύθηκε με έναν αλγόριθμο ανθρώπινης ιατρικής που είχαν εισαγάγει προηγουμένως οι συγγραφείς (Narayanaswamy et al., 2021). Προτάθηκε ένα πλαίσιο εξήγησης με την ονομασία ExAID για την κακοήθεια των δερματικών βλαβών με τη χρήση διανυσμάτων ενεργοποίησης εννοιών (CAV) Το πλαίσιο χρησιμοποίησε δύο ταξινομητές, έναν ταξινομητή επιπέδου έννοιας για δερματολογικές έννοιες και έναν ταξινομητή επιπέδου νόσου για τη διάγνωση των βλαβών, οι οποίοι αξιολογήθηκαν σε δημόσια σύνολα δεδομένων PH2 και Derm7pt. Η χρήση μεταδεδομένων για τη βελτίωση της διάγνωσης δερματικών ασθενειών προτάθηκε με ένα πλαίσιο σύντηξης δεδομένων με βάση τον πολλαπλασιασμό για την ταξινόμηση εικόνων δέρματος σε μικρότερα σύνολα δεδομένων (Li et al., 2020). Το σύνολο δεδομένων δερματοσκοπικών εικόνων βλαβών ISIC' 2018 χρησιμοποιήθηκε με έξι προ-εκπαιδευμένα μοντέλα CNN για τη διάγνωση δερματικών ασθενειών με συγχώνευση μεταδεδομένων και εικόνων δερματοσκοπίας (Li et al., 2020). Στην ποιοτική αξιολόγηση χρησιμοποιήθηκε η

μέθοδος CAM για τον εντοπισμό των σημαντικών περιοχών της εικόνας (Li et al., 2020). Μια παρόμοια ιδέα χρησιμοποιήθηκε με την αξιοποίηση της ταξινομικής οργάνωσης των δερματικών βλαβών για την ανάπτυξη ενός ιεραρχικού νευρωνικού δικτύου, καθώς και των μονάδων χωρικής και καναλικής προσοχής για οπτικοποιήσεις για διορατικότητα στους δερματολόγους/αποφάσεις (Barata et al., 2021). Χρησιμοποιήθηκαν οι δερματοσκοπικές εικόνες βλαβών από τα σύνολα δεδομένων ISIC 2017 και ISIC 2018. Αποδείχθηκε ότι η ιεραρχική διάγνωση είχε καλύτερες επιδόσεις από τα άλλα μοντέλα ταξινόμησης και αναγνώρισε τις σχετικές περιοχές της εικόνας με χάρτες προσοχής (Barata et al., 2021). Ένα μοντέλο βασισμένο σε Βαθιά Συνελικτικά Νευρωνικά Δίκτυα (Deep CNN) προτάθηκε για την ταυτόχρονη κατάτμηση και ταξινόμηση δερματικών βλαβών, δείχνοντας καλύτερες επιδόσεις σε σχέση με άλλα μοντέλα σε σύνολα δεδομένων όπως τα ISIC 2017 και PH2, όσον αφορά τον δείκτη Jaccard και την AUC. Μια άλλη μελέτη χρησιμοποίησε το σύνολο δεδομένων ISIC για τη δυαδική ταξινόμηση των δερματικών βλαβών. Επιπλέον, οι ενέργειες χαρτών χαρακτηριστικών CNN χρησιμοποιήθηκαν για την κατανόηση των ενεργειών του μοντέλου CNN σε δερματοσκοπικές εικόνες βλαβών με στόχο την αύξηση της εμπιστοσύνης των δερματολόγων στο σύστημα (Molle et al., 2018).

Διαβητική αμφιβληστροειδοπάθεια

Η ΣΔ επηρεάζει εκατομμύρια ανθρώπους σε παγκόσμιο επίπεδο και αποτελεί μία από τις κύριες αιτίες τύφλωσης. Το γλαύκωμα μπορεί να διαγνωστεί με τεχνικές τεχνητής νοημοσύνης που εφαρμόζονται σε εικόνες OCT και φωτογραφίες βυθού (Zhang et al., 2022). Μια ανασκόπηση των τεχνικών AI χρησιμοποιήθηκαν τέσσερα μοντέλα CNN και οι απεικονίσεις με heatmaps προσδιόρισαν τα σημαντικά χαρακτηριστικά τα οποία συγκρίθηκαν με την κρίση του ειδικού (Reguant et al., 2021). Τα σύνολα δεδομένων EyePACS και DIARETDB1 χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση διάφορων μοντέλων CNN για την αυτόματη ταξινόμηση διαβητικής αμφιβληστροειδοπάθειας (DR). Παρόλο που τα διάφορα μοντέλα CNN εμφάνισαν σταθερή απόδοση στην ταξινόμηση, η συμφωνία των χαρακτηριστικών της εικόνας μεταξύ των διαφορετικών αρχιτεκτονικών ήταν υψηλή, φτάνοντας έως και 70% (Reguant et al., 2021). Επιπλέον, προτάθηκε ένα πλαίσιο για την ταξινόμηση πολλαπλών ετικετών εικόνων DR, το οποίο παρείχε επεξηγηματικότητα με την ταξινόμηση κάθε εικονοστοιχείου. (Quellec et al., 2021). Η ταξινόμηση της εικόνας

προήλθε από μια πολυμεταβλητή κατάτμηση βλαβών και κατηγοριοποίηση παθολογικών βλαβών. Τα αποτελέσματα κατέδειξαν υψηλή ακρίβεια ταξινόμησης και επεξηγηματικότητα (Quelleg et al., 2021). Οι συγγραφείς πρότειναν και εφάρμοσαν τον αλγόριθμο ExplAIn για την ταξινόμηση πολλαπλών κατηγοριών εικόνων DR βυθού (Quelleg et al., 2021). Η ακρίβεια ταξινόμησης αξιολογήθηκε τόσο σε επίπεδο εικονοστοιχείου όσο και σε επίπεδο εικόνας (Quelleg et al., 2021). Μια μέθοδος απεικόνισης βασισμένη σε Generative Adversarial Network (GAN), η Patho-GAN, προτάθηκε για την ανίχνευση διαβητικής αμφιβληστροειδοπάθειας και τη δημιουργία εικόνων αμφιβληστροειδούς (Niu et al., 2021). Μια μεθοδολογία παρόμοια με αυτή του Koch Postulates για την Ιατρική βασισμένη σε αποδείξεις (EBM) χρησιμοποιήθηκε για την ερμηνευσιμότητα (Niu et al., 2021). Αυτά τα αξιώματα παρείχαν τη συσχέτιση μεταξύ των ασθένειας και των παθογόνων παραγόντων (Niu et al., 2021).

Προτάθηκε ένα σύστημα διάγνωσης που χρησιμοποιεί ένα CNN για την τμηματοποίηση και ταξινόμηση της ανίχνευσης γλαυκώματος με εικόνες βυθού για ανάπτυξη σε κινητές συσκευές (Martins et al., 2020). Το προτεινόμενο σύστημα είχε τρεις δυνατότητες ερμηνείας μέτρα για τα ενδιάμεσα αποτελέσματα τμηματοποίησης, μορφολογική χαρακτηριστικά και χάρτες ενεργοποίησης με Grad-CAM (Martins et al., 2020). Το σύστημα ήταν θεωρήθηκε χρήσιμο για μαζικό έλεγχο, καθώς μπορούσε να εκτελεστεί χωρίς σύνδεση στο Διαδίκτυο σε μια κινητή συσκευή (Martins et al., 2020). Το CNN χρησιμοποιήθηκε για την ταξινόμηση της πλευρικότητας των εικόνων του βυθού και το Grad-CAM χρησιμοποιήθηκε για τον εντοπισμό των σημαντικών περιοχών της εικόνας για την ταξινόμηση. Η εκτίμηση της αβεβαιότητας και οι οπτικοποιήσεις κρίθηκαν χρήσιμες για την επεξηγηματικότητα του μοντέλου. Το προτεινόμενο μοντέλο CNN επέτυχε υψηλότερη ακρίβεια σε σύγκριση με τα VGG-16 και AlexNet, ενώ βρέθηκε να παρουσιάζει συγκρίσιμες επιδόσεις με αυτές των κλινικών ιατρών. Προτάθηκε ένα σύστημα που λαμβάνει υπόψη την αβεβαιότητα για να διαβαθμίσει τη σοβαρότητα των εικόνων του βυθού του αμφιβληστροειδούς, προσφέροντας έτσι επίγνωση στον εμπειρογνώμονα με την εμπιστοσύνη στην απόφαση του μοντέλου. Επιπλέον, παρείχαν εξηγήσεις και αβεβαιότητα για κάθε πρόβλεψη, κάτι που αναγνωρίστηκε ως χρήσιμο σε κλινικό περιβάλλον (Araújo et al., 2020).

Για την ταξινόμηση και τον εντοπισμό του γλαυκώματος χρησιμοποιήθηκαν CNN και Grad-CAM σε εικόνες βυθού. Πέντε προεκπαιδευμένα μοντέλα CNN εκπαιδεύτηκαν και αξιολογήθηκαν με το καλύτερο μοντέλο να επιτυγχάνει ακρίβεια 96% και ειδικότητα 100% για τον οπτικό δίσκο (Kim et al., 2019). Η μέθοδος εντοπισμού με το Grad-CAM φαίνεται να διαφέρει ανάλογα με τις διαφορετικές αρχιτεκτονικές CNN, ενώ τα εξωτερικά σύνολα δεδομένων φαίνεται να παρέχουν χαμηλότερες ακρίβειες σε σύγκριση με το δικό σύνολο δεδομένων των συγγραφέων (Kim et al., 2019). Περιελήφθη ένα βήμα ανθρώπινης παρέμβασης για τη βελτίωση της ακρίβειας εντοπισμού και πραγματοποιήθηκε μια ποιοτική αξιολόγηση για την επαλήθευση των αποτελεσμάτων ταξινόμησης και εντοπισμού (Meng et al., 2020). Επιτεύχθηκαν καλύτεροι ολοκληρωμένοι χάρτες θερμότητας σε σύγκριση με το Grad-CAM με VGG-16 με τη χρήση προτεινόμενων υπολογισμών ανομοιότητας και διατήρησης της γνώσης (Meng et al., 2020). Τα χαρακτηριστικά πολλαπλών κλιμάκων αποδείχθηκαν χρήσιμα για τη βελτίωση της ακρίβειας στη διάγνωση του γλαυκώματος, ενισχύοντας την τοπική ανίχνευση σε κλινικά απαιτητικά σύνολα δεδομένων εικόνων βυθού ORIGA. Το προτεινόμενο μοντέλο EAMNet επέτυχε μια διάγνωση του γλαυκώματος με AUC 0,88 και ακριβή τμηματοποίηση του οπτικού δίσκου (0,9 Adisc και 0,278 Edisc). Με βάση την εξέλιξη της νόσου, η DR μπορεί να ταξινομηθεί σε έναν από τους τέσσερις βαθμούς (Meng et al., 2020).

Η TCAV χρησιμοποιήθηκε για την πρόβλεψη της διαβητικής αμφιβληστροειδοπάθειας σε εικόνες βυθού αμφιβληστροειδούς σε κλίμακα από το επίπεδο 0 (καμία DR) έως το επίπεδο 4 (πολλαπλασιαστική) και η διαβούλευση με ειδικούς ιατρούς έδειξε ότι η τεχνική μπορεί να βοηθήσει τους ειδικούς να κατανοήσουν τις προβλέψεις του μοντέλου και να βελτιώσουν το μοντέλο. Οι εξηγήσεις των ολοκληρωμένων βαθμίδων χρησιμοποιήθηκαν για τη βελτίωση της διαβάθμισης της σοβαρότητας της DR σε εικόνες βυθού αμφιβληστροειδούς, οι οποίες επίσης βαθμολογήθηκαν από δέκα οφθαλμιάτρους χωρίς βοήθεια, μόνο βαθμούς και βαθμούς συν heatmaps, όπου οι βαθμοί και οι βαθμοί συν heatmaps δημιουργήθηκαν με τη χρήση ενός DNN. Ένα πλαίσιο προτάθηκε για την αξιολόγηση της ποιότητας των εικόνων του βυθού του αμφιβληστροειδούς με ποσοτικές βαθμολογίες και ποιοτική απεικόνιση σε SDRSP και IDRiD (Shen et al., 2020).

Ένα ταχύτερο R-CNN ResNet χρησιμοποιήθηκε για τον εντοπισμό χαρακτηριστικών σε εικόνες βυθού (Kind et al., 2019). Το προτεινόμενο σύστημα παρήγαγε μια έκθεση

για κάθε επεξεργασμένη εικόνα που περιείχε την πρόβλεψη ταξινόμησης, την ανίχνευση βλαβών και σχόλια κειμένου για τον εμπειρογνώμονα (Kind et al., 2019). Το σύστημα χρησιμοποίησε τρία δημόσια σύνολα δεδομένων, Messidor-2, IDRiD και E-Ophtha, συνολικού αριθμού 2314 εικόνων.

Για την οπτική τομογραφία συνοχής (OCT) σε εικόνες αμφιβληστροειδούς χρησιμοποιήθηκαν τα VGG- 16 και SqueezeNet (Marginean et al., 2019). Τα μοντέλα DNN εκπαιδεύτηκαν στο σύνολο δεδομένων OCT-2017 (Marginean et al., 2019). Για τον εντοπισμό των σημαντικών περιοχών της εικόνας για την πρόβλεψη, μια μάσκα απόφραξης μετακινήθηκε κατά μήκος της εικόνας καταγράφοντας τις πιθανότητες ταξινόμησης με τις περιοχές που συνέβαλαν περισσότερο και είχαν ως αποτέλεσμα υψηλή πτώση (Marginean et al., 2019). Προτάθηκε ένα μοντέλο DNN για ταξινόμηση των ασθενειών του αμφιβληστροειδούς που σχετίζονται με τον διαβήτη σε οπτική συνοχή Τομογραφία (OCT) και το μοντέλο βρέθηκε να αποδίδει καλύτερα ως προς την ακρίβεια, την ευαισθησία και την ειδικότητα σε σύγκριση με άλλα παρόμοια μοντέλα, χωρίς να απαιτούνται χειροκίνητα σχόλια (Shorfuzzaman et al., 2021).

Η οφθαλμολογία είναι μια από τις πιο διαδεδομένες εφαρμογές της AI και υπάρχουν προσπάθειες για την εξηγήσιμη χρήση της στον τομέα αυτό, αλλά είναι ακόμη σε πολύ βασικό επίπεδο. Οι επεξηγήσιμοι αλγόριθμοι για την οφθαλμολογία, ειδικά για την DR, πρέπει να περιλαμβάνουν περισσότερες γνώσεις από τον πραγματικό τομέα και θα πρέπει να είναι σε θέση να ενσωματώνουν συμβατικά συστήματα ταξινόμησης, όπως η μελέτη για την πρόωμη θεραπεία της διαβητικής αμφιβληστροειδοπάθειας (ETDRS) και η χαρτογράφηση του πάχους κ.λπ. για λεπτομερέστερες γνώσεις.

Οι οφθαλμίατροι επιθυμούν να κατανοήσουν τις παραμέτρους που οδηγούν σε ένα συγκεκριμένο εύρημα, παρά να απλώς λαμβάνουν την τελική διάγνωση. Στις εικόνες βυθού, οι πραγματικές μεταβολές στα ορόσημα, δηλαδή στα αιμοφόρα αγγεία, στον οπτικό δίσκο και στην περιοχή του φατνίου μαζί με την εμφάνιση οποιασδήποτε νέας βλάβης είναι σημαντικές. Για τις εικόνες OCT, είναι σημαντικό να εξετάζονται οι παραλλαγές των στρωμάτων και οι βλάβες που εμφανίζονται σε αυτά τα στρώματα.

Καρκίνος

Η τεχνική Cumulative Fuzzy Class Membership Criterion (CFCMC) παρέχει μια

φιλική προς τον άνθρωπο εξήγηση της απόφασης πρόβλεψης. Αυτή η τεχνική προσφέρει τριπλές εξηγήσεις, συμπεριλαμβανομένης μιας σημασιολογικής εξήγησης της πιθανότητας εσφαλμένης ταξινόμησης, ενός δείγματος εκπαίδευσης για την πρόβλεψη, καθώς και δειγμάτων εκπαίδευσης από αντίθετες κλάσεις (Sabol et al., 2020). Η μέθοδος εφαρμόστηκε σε ένα δημόσιο σύνολο δεδομένων που αποτελείται από εικόνες ιστού αιματοξυλίνης και ηωσίνης (H&E) από πλάκες ιστού, οι οποίες προέρχονταν από περιπτώσεις καρκίνου του παχέος εντέρου (Sabol et al., 2020). Ο γαστρικός καρκίνος κατατάσσεται ως ο τέταρτος κοινός τύπος καρκίνου παγκοσμίως. Τα LIME, SHAP και CIU χρησιμοποιήθηκαν για την επεξηγηματικότητα των προβλέψεων του CNN σε εικόνες in-vivo γαστρικών εικόνων Video Capsule Endoscopy (VCE) (Knapic et al., 2021). Οι επεξηγήσεις των τριών μεθόδων XAI αξιολογήθηκαν με άνθρωπο εμπειρογνώμονα και η μελέτη αποσκοπούσε στην ανάπτυξη μιας ανθρώπινης απόφασης συστήματος υποστήριξης ανθρώπου (Knapic et al., 2021). Αρχιτεκτονική βασισμένη στο CNN με βάση το ResNet-34 δίκτυο χρησιμοποιήθηκε για την αναγνώριση και ταξινόμηση έλκους σε κάψουλα ενδοσκόπησης (Wang et al., 2019). Το σύστημα επέδειξε βελτιωμένες επιδόσεις στις μετρικές F1, F2 και ROC-AUC σε σύγκριση με τα άλλα επιλεγμένα δίκτυα CNN. Η προτεινόμενη αρχιτεκτονική επέδειξε υψηλότερες επιδόσεις σε σύγκριση με τις άλλες αρχιτεκτονικές CNN για την ανίχνευση μικρών βλαβών ελκών. Οι απεικονίσεις για τις σημαντικές τμήματα της εικόνας που χρησιμοποιήθηκαν CAM για την επιβεβαίωση του εντοπισμού του έλκους (Wang et al., 2019).

Το LIME χρησιμοποιήθηκε για την επεξηγηματικότητα των ιστολογικών εικόνων για λεμφαδένες μεταστάσεις των κόμβων ταξινόμηση του CNN στο Patch Camelyon (P-CAM) σύνολο δεδομένων (Sousa et al., 2019). Οι εξηγήσεις του LIME ταίριαζαν γενικά με τις σχολιασμούς του κλινικού ιατρού, ενώ τα μοντέλα CNN φάνηκε να εστιάζουν σε συγκεκριμένες χαρακτηριστικά των εικόνων (Sousa et al., 2019). Φάνηκε ότι το μοντέλο CNN είχε μάθει να συσχετίζει τους όγκους με τις δομές της εικόνας στο κέντρο, όπως παρατηρήθηκε στο στους παραγόμενους χάρτες θερμότητας (Sousa et al., 2019). Προτάθηκε μια εξηγήσιμη μέθοδος εκμάθησης χαρακτηριστικών πολλαπλών κλιμάκων για αποφάσεις του μοντέλου CNN (Uehara et al., 2021). Η μέθοδος εφαρμόστηκε για τη διάγνωση του καρκίνου του τραχήλου της μήτρας, χρησιμοποιώντας την ερμηνεία των αποφάσεων μέσω οπτικοποίησης και στατιστικής ανάλυσης. Σύμφωνα με τους Uehara et al., (2021), «αυτή η διαδικασία

εφαρμόστηκε σε παθολογικές εικόνες ενός τραχήλου της μήτρας που παρασχέθηκαν από το Πανεπιστημιακό Νοσοκομείο του Ναγκασάκι, στην Ιαπωνία. Πολλαπλής κλίμακας λεξικά χαρακτηριστικών εισήχθησαν για την καταγραφή των διακριτικών χαρακτηριστικών σε διαφορετικές κλίμακες για την καλύτερη διάγνωση». Τα αποτελέσματα προσδιόρισαν τα σημαντικά και συνεισφέροντα χαρακτηριστικά της εικόνας που μαθαίνονται για την διάγνωση (Uehara et al., 2021). Τέσσερα μοντέλα CNN χρησιμοποιήθηκαν για τον οισοφάγο Barrett μαζί με πέντε τεχνικές απεικόνισης Saliency, Guided Backpropagation, Integrated Gradients, input X gradients και DeepLIFT για τον προσδιορισμό του προβλέψιμου μοντέλου με τις επισημάνσεις των εμπειρογνομόνων (Jr et al., 2019). Η εστίαση ήταν στις πρώτες ανίχνευση και στην αποσαφήνιση των περιοχών της εικόνας που είναι σημαντικές για την ταξινόμηση (Jr et al., 2019).

Κλινικά/παρατηρητικά δεδομένα

Το ΗΚΓ χρησιμοποιείται συχνά για την καρδιαγγειακή απεικόνιση. Προτάθηκε μια λύση ερμηνείας του ΗΚΓ για την ταξινόμηση των καρδιακών παλμών με τη χρήση διαφόρων μοντέλων-διαγνωστικών μεθόδων, PSI, LIME και SHAP (Neves et al., 2021). Η μελέτη επικεντρώθηκε σε ερμηνείες βάσει δειγμάτων και αξιολογήθηκε χρησιμοποιώντας τον δείκτη 1-D Jaccard για τη μέτρηση της ομοιότητας του ταξινομητή και της σχετικής υποακολουθίας από τη μέθοδο XAI (Neves et al., 2021). Τα δυαδικά και πολυταξικά μοντέλα βασίστηκαν σε πέντε ή τέσσερα μπλοκ CNN αντίστοιχα (Neves et al., 2021). Τα πλαίσια CNN, RNN και XGBoost χρησιμοποιήθηκαν για τη διάγνωση του κινδύνου οξέος εμφράγματος του μυοκαρδίου (AMI) χρησιμοποιώντας το σύνολο δεδομένων ECG-ViEW II. Μοντέλα CNN χρησιμοποιήθηκαν με δεδομένα ΗΚΓ για τον εντοπισμό καρδιακών χαρακτηριστικών για την πρόβλεψη συστημικών φαινοτύπων που επηρεάζουν τον καρδιαγγειακό κίνδυνο (Ghorbani et al., 2020). Το Grad-CAM χρησιμοποιήθηκε για οπτικοποιήσεις και εντόπισε RoI για την πρόβλεψη των συστημικών φαινοτύπων, οι οποίοι διαφορετικά θεωρούνται δύσκολοι για τους ανθρώπινους εμπειρογνώμονες (Ghorbani et al., 2020). Εικόνες διατομικής υπερηχοκαρδιογραφίας χρησιμοποιήθηκαν για την αυτόματη λειτουργική ταξινόμηση κατά Carpentier των παθήσεων της μιτροειδούς βαλβίδας με οκτώ βαθιά μοντέλα CNN. Εικόνες CXR για συμφορητική καρδιακή ανεπάρκεια (CHF) χρησιμοποιήθηκαν για την εκπαίδευση

ενός νευρωνικού δικτύου σε ένα σύνολο δεδομένων με ετικέτες και ένα παραγωγικό μοντέλο που εκπαιδεύτηκε σε για την εκτίμηση του νατριουρητικού πεπτιδίου τύπου B BNP ως δείκτη για την CHF (Seah et al., 2019). Παραγωγική οπτική λογική (GVR) με βάση το GAN χρησιμοποιήθηκαν για την οπτικοποίηση και συγκρίθηκαν με άλλες οπτικοποιήσεις τεχνικές όπως η LIME, καταλήγοντας στο συμπέρασμα ότι η GVR μπορεί να εντοπίσει τις συγχυτικές εικόνες (Seah et al., 2019). Ένα αυτοματοποιημένο σύστημα για τον εντοπισμό στένωσης και την ταξινόμηση της σοβαρότητας σε τρεις κατηγορίες στη στεφανογραφία με ακτίνες X (CAG) χρησιμοποίησε μοντέλα CNN και RNN, με CAM και Grad-CAM για την οπτικοποίηση. Τα αποτελέσματα έδειξαν υψηλή ευαισθησία για τη στένωση και υψηλή ακρίβεια για την ταξινόμηση των στενώσεων. Χρησιμοποιήθηκαν ολόκληρες εικόνες αντικειμενοφόρου πλάκας για την αυτοματοποίηση της βαθμολόγησης των τεσσάρων ιστολογικών χαρακτηριστικών στη μη αλκοολική λιπώδη ηπατική νόσο (NAFLD) σε δύο διαφορετικές κλίμακες με τέσσερα μοντέλα CNN (Heinemann et al., 2019).

Ένα μοντέλο CNN χρησιμοποιήθηκε για την αυτόματη ταξινόμηση κακοήθων πνευμονικών κυττάρων με κυτταρολογικές εικόνες και τα αποτελέσματα βρέθηκαν συγκρίσιμα με αυτά ενός κυτταροπαθολόγου (Teramoto et al., 2019). Το Grad-CAM χρησιμοποιήθηκε για την απεικόνιση των σημαντικών περιοχών της εικόνας για την ταξινόμηση και βρέθηκε να αναγνωρίζει τα κακοήθη κύτταρα (Teramoto et al., 2019). Πολυμεταβλητά δεδομένα χρονοσειρών αναλύθηκαν με CNN για την επεξηγηματικότητα με οπτικοποίηση των σχετικών χαρακτηριστικών με την πάροδο του χρόνου (Viton et al., 2020). Οι οπτικοποιήσεις χρησιμοποίησαν heatmaps του ίδιου σχήματος με τα δεδομένα της χρονοσειράς (Viton et al., 2020). Η ενδονοσοκομειακή θνησιμότητα προβλέφθηκε με βάση τα αρχεία της MEΘ (Viton et al., 2020). Προσδιορίστηκαν τα εξέχοντα και διακριτικά σημεία στον χάρτη θερμότητας χρησιμοποιώντας t-test (Viton et al., 2020).

Ανίχνευση καρκίνου

Η ανασκόπηση των προσεγγίσεων της εξηγηματικής τεχνητής νοημοσύνης (XAI) για την ανίχνευση καρκίνου με μαγνητική τομογραφία στα DNNs κατέληξε στο συμπέρασμα ότι η επεξηγηματικότητα μπορεί να βελτιωθεί με αντιπαράδειγματα και μοντέλα που εξηγούνται εγγενώς. Ο καρκίνος του προστάτη είναι μια συχνή μορφή καρκίνου στους άνδρες, και η έγκαιρη ανίχνευση μπορεί να αυξήσει σημαντικά τα

ποσοστά επιβίωσης. Για τη βελτίωση της ακρίβειας ταξινόμησης, χρησιμοποιήθηκε πολυτροπική συγχώνευση για τον συνδυασμό προ-εκπαιδευμένων μοντέλων βαθιάς μάθησης για την ανίχνευση καρκίνου σε εικόνες υπερήχων και μαγνητικής τομογραφίας (Hassan et al., 2022). Τα δεδομένα εικόνων μαγνητικής τομογραφίας και υπερήχων του προστάτη ανοικτού κώδικα χρησιμοποιήθηκαν με το LIME για την επεξήγηση των σημαντικών χαρακτηριστικών για τις ταξινομήσεις καλοήθους και κακοήθους καρκίνου (Hassan et al., 2022). Στη διάγνωση του καρκίνου του ήπατος χρησιμοποιήθηκαν εικόνες MRI και CNN για την ταξινόμηση των βλαβών. Η βαθμολόγηση των χαρακτηριστικών χρησιμοποίησε συναρτήσεις Influence που αντιστοιχούν στα χαρακτηριστικά της εικόνας και επιβεβαίωσε ότι εντοπίστηκαν τα πιο σημαντικά χαρακτηριστικά απεικόνισης. Ένα οπτικό Case Based Reasoning (CBR) χρησιμοποίησε μια βάση δεδομένων προηγούμενων γνωστών περιπτώσεων για να καθορίσει τη λύση για ένα νέο ερώτημα. Η CBR προτάθηκε για οπτική συλλογιστική σε τρία δημόσια σύνολα δεδομένων για τον καρκίνο του μαστού επιτυγχάνοντας ακρίβεια ταξινόμησης παρόμοια με την KNN αλλά παρέχοντας καλύτερη επεξηγηματικότητα (Lamy et al., 2019). Η επεξηγηματικότητα παρέχεται μέσω μιας οπτικής διεπαφής που παρουσίαζε τις ομοιότητες μεταξύ του ερωτήματος και παρόμοιων περιπτώσεων. Αυτό επιτεύχθηκε με τη χρήση διαγραμμάτων διασποράς για ποσοτικές ομοιότητες και κουτιών ουράνιου τόξου για ποιοτικές ομοιότητες (Lamy et al., 2019). Το LRP επεκτάθηκε για εφαρμογή σε Graph-CNN και δοκιμάστηκε ως προς την εξηγησιμότητα για ένα σύνολο δεδομένων για τον καρκίνο του μαστού (Chereda et al., 2021). Η προτεινόμενη μέθοδος GLRP εφαρμόστηκε σε σύνολο δεδομένων γονιδιακής έκφρασης (Chereda et al., 2021). Το Graph CNN μάθαινε χαρακτηριστικά σε σταθμισμένους γράφους με κόμβους και ακμές, σε αντίθεση με το CNN που χρησιμοποιούν μια δομή που μοιάζει με πλέγμα (Chereda et al., 2021).

Μια αρχιτεκτονική CNN με πολλαπλές εισόδους προτάθηκε για την πρόβλεψη της ανταπόκρισης του όγκου του μαστού στη χημειοθεραπεία σε εικόνες DCE-MRI. Οι απεικονίσεις με το Grad-CAM χρησιμοποιήθηκαν για τον εντοπισμό της περιοχής ενδιαφέροντος του όγκου και έδειξαν ότι τα σημαντικά χαρακτηριστικά βρίσκονταν στις περιφερειακές περιοχές. Η αρχιτεκτονική 3D ResNet χρησιμοποιήθηκε για τον εντοπισμό της ταξινόμησης του όγκου με χρήση του Cosine Margin Sigmoid Loss (CMSL) και για τον εντοπισμό των καρκινικών βλαβών με χρήση του COrelation

Attention Map (COAM) σε ένα σύνολο δεδομένων DCE-MRI (Luo et al., 2019). Η εκτίμηση της ογκομετρικής πυκνότητας του μαστού αξιολογήθηκε με τρισδιάστατη παλινδρόμηση CNN σε εικόνες MRI με εξηγησιμότητα με χρήση SHAP (van der Velden et al., 2020). Το κύριο πλεονέκτημα της μεθόδου κατέληξε στο συμπέρασμα ότι δεν απαιτεί τρισδιάστατες τμηματοποιήσεις σε επίπεδο voxel (van der Velden et al., 2020). Το GAN περιλάμβανε ένα ερμηνεύσιμο δίκτυο διάγνωσης και ένα συνθετικό δίκτυο βλαβών (Kim et al., 2018). Το συνθετικό δίκτυο δημιουργίας βλαβών χρησιμοποίησε αντιθετική μάθηση για να βοηθήσει το ερμηνεύσιμο δίκτυο διάγνωσης που σχεδιάστηκε με τη χρήση ενός CNN (Kim et al., 2018). Το προτεινόμενο σύστημα επικυρώθηκε σε ένα δημόσιο σύνολο δεδομένων DDSM (Kim et al., 2018). Η απόδοση του μοντέλου αξιολογήθηκε δείχνοντας ότι το προτεινόμενο σύστημα ερμήνευσε την πρόβλεψη από την άποψη του BI-RADS (Kim et al., 2018). Η έγκαιρη ανίχνευση και η έκταση της εξάπλωσης των καρκινικών κυττάρων είναι υψίστης σημασίας για τη λήψη απόφασης σχετικά με το σχέδιο θεραπείας.

Προκλήσεις και μελλοντικές κατευθύνσεις

Οι τεχνικές Εξήγησης Τεχνητής Νοημοσύνης (XAI) έχουν το δυναμικό να διευκολύνουν την κατανόηση των προβλέψεων ενός μοντέλου βαθιάς μάθησης. Ωστόσο, υπάρχουν προκλήσεις που πρέπει να ξεπεραστούν προτού αυτές οι τεχνικές γίνουν κυρίαρχες για εφαρμογές υγειονομικής περίθαλψης. Σε αυτήν την ενότητα, παρέχουμε μια συζήτηση σχετικά με την εφαρμογή και τις προκλήσεις των μεθόδων XAI για βιοϊατρικές εφαρμογές, καθώς και τις εξελίξεις που βρίσκονται επί του παρόντος σε εξέλιξη και θα μεταμορφώσουν τις εφαρμογές XAI στη βιοϊατρική απεικόνιση.

Η ιατρική απεικόνιση καλύπτει ένα ευρύ φάσμα τεχνικών απόκτησης εικόνων σε χρονικά και χωρικά πεδία, προκειμένου να αποκτηθούν οι πιο σχετικές πληροφορίες που βοηθούν στη διάγνωση της υποκείμενης ιατρικής κατάστασης. Οι μέθοδοι εικόνας και οι λεπτομέρειες εικόνας ποικίλλουν. Ωστόσο, η πιο συνηθισμένη και προτιμώμενη μέθοδος Εξήγησης Τεχνητής Νοημοσύνης (XAI) στις διάφορες μεθόδους ιατρικής απεικόνισης φαίνεται να είναι η Grad-CAM, ακολουθούμενη από τις μεθόδους LRP, LIME και SHAP και άλλες. Η αυξημένη χρήση των τεχνικών XAI παρατηρείται σε νευρολογικές, αναπνευστικές και οφθαλμολογικές μελέτες. Στη

νευρολογία, η επικρατέστερη μέθοδος απεικόνισης είναι η μαγνητική τομογραφία. Οι εικόνες CXR προτιμώνται περισσότερο για τις αναπνευστικές μελέτες λόγω της ευκολίας με την οποία μπορούν να ληφθούν, με τη Grad-CAM ως την επικρατέστερη μέθοδο XAI. Στις οφθαλμολογικές μελέτες, οι εικόνες βυθού φαίνεται να είναι ο κυρίαρχος τρόπος απεικόνισης. Έχουν προταθεί πολλές τεχνικές και βρίσκονται σε εξέλιξη εργασίες για να καταστεί δυνατή η ευρεία υιοθέτηση των τεχνικών Deep Learning και XAI στην κλινική πράξη. Τα DNN επιτρέπουν την αυτοματοποιημένη επιλογή RoI χωρίς τις χρονοβόρες χειροκίνητες επισημειώσεις (Chen et al., 2019). Είναι σημαντικό να συσχετιστεί η αβεβαιότητα πρόβλεψης του μοντέλου για να βοηθηθεί ο κλινικός ιατρός για περισσότερες προσοχή σε αυτές.

Οι τεχνικές επεξηγηματικότητας βασίζονται συνήθως σε οπτικοποιήσεις που παρέχουν μια ποιοτική ένδειξη, τονίζοντας τη σημασία των περιοχών της εικόνας που συμβάλλουν στην απόφαση ενός μοντέλου. Αυτό μπορεί να παρέχει έναν βαθμό εμπιστοσύνης στο μοντέλο, αλλά χωρίς έναν τυποποιημένο τρόπο αξιολόγησης των διαφόρων μοντέλων και των αποτελεσμάτων τους δεν μπορεί να υπάρξει εμπιστοσύνη στην ιατρική απεικόνιση. Η δημιουργία ποσοτικών μετρήσεων για τη σύγκριση μοντέλων για την επεξηγηματικότητα θα οδηγήσει στην επιτυχή εφαρμογή των σχετικών τεχνικών επεξηγηματικότητας σε διαφορετικές μορφές εικόνας. Τέτοιες μετρικές μπορούν επίσης να καλύψουν τις ανάγκες για τους κανονισμούς και τα πρότυπα στην αξιολόγηση και την έγκριση των τεχνικών XAI για την κλινική πράξη. Υπάρχει ανάγκη να αναπτυχθούν μοντέλα DNN που είναι εγγενώς ερμηνεύσιμα και ταυτόχρονα ακριβή. Τέτοια συστήματα XAI πρέπει να ενσωματωθούν στις ροές εργασίας της κλινικής διαδικασίας με διαισθητικές διεπαφές. Οι μοντελοαγνοϊτικές μέθοδοι υπόσχονται περαιτέρω ανάπτυξη και αξιοποίηση για την επεξηγηματικότητα. Παρόμοια με την πρόοδο στην Αυτοματοποιημένη Μηχανική Μάθηση (AutoML) για τη δημιουργία μιας αναπαραγωγίσιμης και βέλτιστης μοντέλο αυτόματα με βάση τους δεδομένους περιορισμούς, υπάρχει μια ανάγκη να αναπτυχθεί παρόμοιο πλαίσιο για αυτοματοποιημένα μοντέλα XAI όπου σε απόκριση στις βασικές απαιτήσεις του χρήστη, ένα κατάλληλο μοντέλο XAI τεχνική μπορεί να επιλεγεί και να εφαρμοστεί αυτόματα. Η επεξηγηματική διαδραστική Μάθηση (XIL) που προσθέτει τον άνθρωπο στον κύκλο εκπαίδευσης του μοντέλου (Schramowski et al., 2020) μπορεί επίσης να συμβάλει στην αύξηση της υιοθέτησης των τεχνικών XAI.

Κατανόηση και βελτίωση της επεξηγηματικότητας

Ο όρος "εξήγηση", όπως χρησιμοποιείται στο XAI, δεν είναι πολύ καλά ορισμένος (Gillmann et al., 2021). Δεν υπάρχει ένας καθολικά αποδεκτός ορισμός των εξηγήσεων και έχουν γίνει κάποιες προσπάθειες προς αυτή την κατεύθυνση με την ανάπτυξη μαθηματικών μεθόδων. Δεν υπάρχει συμφωνία σχετικά με το τι σημαίνει "εξήγηση" (Guidotti et al., 2018). Γίνονται προσπάθειες από ορισμένα ερευνητικά ιδρύματα και ρυθμιστικές αρχές να προδιαγράψουν σαφείς ορισμούς, πεδίο εφαρμογής και δυνατότητα εφαρμογής της εξηγήσιμης TN και άλλων παρόμοιων αλληλεπιδρώντων όρων. Αυτό διαφέρει για τις διαφορετικές οπτικές γωνίες του ασθενούς, του κλινικού ιατρού και του μηχανικού και οι τρέχοντες ορισμοί στερούνται συνοχής με την επεξηγήσιμη TN να αντιπροσωπεύει μια ψεύτικη ελπίδα για την παροχή υποστήριξης αποφάσεων σε επίπεδο ασθενούς (Ghassemi et al., 2021). Τα χαρακτηριστικά που επισημαίνονται από το μοντέλο μπορεί επίσης να διαφέρουν ως προς τη συνάφεια με τα χαρακτηριστικά και τις αποφάσεις που προσδιορίζονται από τον άνθρωπο (Ghassemi et al., 2021). Η υιοθέτηση της τεχνητής νοημοσύνης στην κλινική πρακτική θα απαιτούσε συνεπώς μεγαλύτερη προσπάθεια ώστε οι τεχνικές XAI να είναι πιο κατανοητές. Η ανάπτυξη των DNN και στη συνέχεια η ύπαρξη προσπαθειών για τον προσδιορισμό της εσωτερικής λειτουργίας τους και της διαδικασίας λήψης αποφάσεων έχει οδηγήσει στην ανάπτυξη πολλών χρήσιμων τεχνικών. Ωστόσο, θα ήταν χρήσιμο αν η επεξηγησιμότητα ενσωματωνόταν στο μοντέλο. Οι περισσότερες από τις μελέτες που αφορούν την ανίχνευση καρκίνου χρησιμοποιούν μια μεταγενέστερη τεχνική εξηγήσεων, ενώ η ανάπτυξη τεχνικών εγγενών εξηγήσεων θα ήταν επωφέλης για την υιοθέτηση και την εξηγησιμότητα. Προτάθηκε ένα δίκτυο κάψουλας πολλαπλών καθηκόντων, το X-Caps, για τη βελτίωση της επεξηγησιμότητας με τη χρήση των οπτικών χαρακτηριστικών υψηλού επιπέδου των δικτύων κάψουλας για την παροχή βαθμολογιών κακοήθειας που χρησιμοποιούνται από τους εμπειρογνώμονες (LaLonde et al., 2020). Ενσωματώθηκαν αυτοματοποιημένες αναλύσεις χαρτών θερμότητας στο πλαίσιο του αγωγού MLOps (Machine Learning Operations) για vision AI ως δομικό στοιχείο που χρησιμοποιείται για την ανίχνευση μεροληπτικών μοντέλων από ενεργοποιήσεις με άκυρα εικονοστοιχεία σε εικόνες δοκιμής. Μέθοδοι μοντελοποίησης λευκού κουτιού όπως το XNN, κλιμακούμενες λίστες κανόνων Bayesian, και η μονοτονική GBM, παρέχουν μικρή ποινή ακρίβειας για την ερμηνευσιμότητα (Hall and Gill, 2018). Η ανάπτυξη ενός ερμηνεύσιμου μοντέλου

από την αρχή θα διευκολύνει την εξήγηση και την αξιολόγηση της δικαιοσύνης (Hall and Gill, 2018). Μελλοντικές προσεγγίσεις μπορεί να δουν ένα συνδυασμό ενδογενών, εξαγωγής κανόνων και απόδοσης μεθόδους για την παροχή απαντήσεων σε μια γλώσσα ερμηνεύσιμη από τον άνθρωπο.

ΧΑΙ με επίκεντρο τα δεδομένα

Η ποιότητα και η ποσότητα των δεδομένων εκπαίδευσης αποτελούν κρίσιμο παράγοντα για την απόδοση ενός εκπαιδευμένου μοντέλου, επηρεάζοντας σημαντικά την ερμηνευσιμότητα που μπορεί να επιτευχθεί από αυτό. Τα μεγάλα και υψηλής ποιότητας σύνολα δεδομένων είναι απαραίτητα για τα ΔNN προκειμένου να επιτύχουν καλύτερα αποτελέσματα, αποφεύγοντας τον κίνδυνο της υπερπροσαρμογής, όπου το μοντέλο προσαρμόζεται υπερβολικά σε ένα περιορισμένο σύνολο δεδομένων. Αυτό μπορεί να αποτελέσει πρόκληση σε σύνολα δεδομένων εικόνας, όπου η συλλογή δειγμάτων από τον ασθενή είναι δύσκολη λόγω του περιορισμένου αριθμού δειγμάτων σε σύγκριση με τις φυσιολογικές περιπτώσεις. Η ποιότητα των δεδομένων μπορεί επίσης να επηρεαστεί από το θόρυβο και την εσφαλμένη επισήμανση. Η έλλειψη ποιότητας των δεδομένων μπορεί επίσης να αντιμετωπιστεί ως εμπόδιο στα αποτελέσματα της επεξηγηματικής ανάλυσης μέσω παραδειγμάτων (Kim et al., 2016). Οι μέθοδοι αποθορυβοποίησης μπορούν να βελτιώσουν την ανθεκτικότητα του DNN έναντι μικρών διαταραχών και υποβάθμισης της ποιότητας με τη χρήση μεθόδων αποθορυβοποίησης για την ενίσχυση των ακμών και των περιγραμμάτων (Yim and Sohn, 2017). Η ποιότητα των εικόνων βελτιώνει την απόδοση των αλγορίθμων DNN στην ορθή πρόβλεψη μιας ασθένειας και επομένως μπορεί να μειώσει το κενό εξηγήσεων. Η μείωση του θορύβου στα δεδομένα μπορεί να διευκολύνει την εξήγηση ή την ερμηνεία του μοντέλου. Οι μέθοδοι βαθιάς μάθησης απαιτούν επαρκή ποσότητα και ποιότητα δεδομένων εκπαίδευσης προκειμένου να παράγουν σταθερά και χρήσιμα αποτελέσματα (Singh et al., 2021). Χωρίς αρκετά δεδομένα, τα μοντέλα τεχνητής νοημοσύνης μπορεί να υποφέρουν από υπερπροσαρμογή, ενώ η ποικιλία των πολιτικών και των πρωτοκόλλων συλλογής δεδομένων μεταξύ διαφορετικών νοσοκομείων μπορεί να οδηγήσει σε σημαντικές διαφοροποιήσεις στα δεδομένα. Ορισμένες μελέτες που αξιολογούν τις επιδόσεις των ΔNN στην ιατρική απεικόνιση συχνά βασίζονται σε μικρά σύνολα δεδομένων, κάτι που μπορεί να αποτελέσει σημαντικό περιορισμό. Συλλογή δεδομένων για την κάλυψη του καρκίνου του μαστού όλες τις παραλλαγές

της εικόνας είναι δύσκολο να ληφθούν υπόψη άλλοι παράγοντες όπως η ηλικία, φυλή, εθνικότητα κ.λπ. Τα σύνολα δεδομένων βιοϊατρικής απεικόνισης υποφέρουν κάποτε από θόρυβο, ενώ οι εικόνες αξονικής τομογραφίας συνήθως επηρεάζονται από θόρυβο (Tayaran et al., 2020). Οι προβλέψεις των DNN με την ίδια αρχιτεκτονική μπορεί να είναι διαφορετικές λόγω στην ποιότητα των δεδομένων και στην αρχικοποίηση του μοντέλου (τυχαία) για εικόνες MRI.

Αρχή φόρμας

ΧΑΙ με επίκεντρο το μοντέλο

Η χρήση επαρκών, αντιπροσωπευτικών και ποικίλων τύπων δεδομένων για μοντέλα ΔNN είναι ουσιώδους σημασίας για τη μείωση της προκατάληψης και την ενίσχυση της ικανότητάς τους να γενικεύουν. Συχνά, στις μελέτες που εφαρμόζουν μηχανική μάθηση σε ιατρικές εικόνες, τα εκπαιδευμένα μοντέλα βασίζονται σε ένα μόνο ιατρικό κέντρο ή νοσοκομείο, πράγμα που μπορεί να προκαλέσει προκατάληψη και να συμπεριλάβει σκαναρίσματα χαμηλής ποιότητας (Hampe et al., 2019). Οι αντιπαραβολικοί αλγόριθμοι έχουν την δυνατότητα να αποκαλύπτουν προκαταλήψεις τόσο στα μοντέλα όσο και στα δεδομένα. Η χρήση τεχνικών ΧΑΙ, όπως οι οπτικοποιήσεις, είναι χρήσιμη για τον εντοπισμό του μοντέλου μεροληψίας και ένας έμπειρος κλινικός ιατρός μπορεί να εντοπίσει τυχόν μη υποστηριζόμενα ή κακά αποφάσεις του DNN. Η τρέχουσα επιτυχία της TN οφείλεται σε μοντέλα που παρέχουν καλύτερες επιδόσεις αλλά είναι αδιαφανή ή μαύρα κουτιά (Knight, 2017).

Ένα από τα κύρια πλεονεκτήματα των Δικτύων Νευρώνων Βαθιάς Μάθησης (DNN) είναι η δυνατότητά τους να αυτοματοποιούν την εξαγωγή χαρακτηριστικών από εικόνες και να ξεπερνούν τα προβλήματα που σχετίζονται με τη χειροκίνητη επιλογή χαρακτηριστικών. Αυτόματες διαδικασίες ΧΑΙ από άκρο σε άκρο, όπως η συνδυασμένη τμηματοποίηση και ταξινόμηση, θα συμβάλουν στην ενσωμάτωση της ΧΑΙ στην κλινική πρακτική. Η αρχιτεκτονική ενός μοντέλου έχει πολλές υπερπαραμέτρους που πρέπει να ελεγχθούν για το σχεδιασμό μιας αρχιτεκτονικής με βέλτιστη απόδοση. Το Automated Machine Learning (AutoML) είναι ένα βήμα προς την κατεύθυνση της αυτόματης δημιουργίας μιας βέλτιστης αρχιτεκτονικής και επιτρέπει την αναπαραγωγή και σύγκριση. Τέτοιες αυτοματοποιημένες ροές εργασίας

αναμένεται να ευκολύνουν τη διαδικασία ενσωμάτωσης της XAI στις κλινικές αποφάσεις, επιλέγοντας τις καλύτερες οπτικοποιήσεις ως απάντηση σε απλές ερωτήσεις. Η αυτοματοποίηση όχι μόνο θα καταστήσει αυτές τις τεχνικές XAI ευρέως διαθέσιμες, αλλά θα επιτρέψει στους ιατρούς να λαμβάνουν καλύτερες και τεκμηριωμένες αποφάσεις, αξιοποιώντας τις γνώσεις μέσω της XAI.

Αυτή η προσέγγιση μπορεί να αποτελέσει ένα εργαλείο για τη μελλοντική πιστοποίηση και τυποποίηση μοντέλων, όπως από τον Οργανισμό Τροφίμων και Φαρμάκων (FDA). Η βιβλιοθήκη AI Fairness 360 είναι ένα ανοικτού κώδικα εργαλείο που μπορεί να συμβάλει στον εντοπισμό και την αντιμετώπιση της μεροληψίας σε μοντέλα μηχανικής μάθησης και είναι διαθέσιμη σε Python και R. Ο στόχος της είναι η μετάφραση της έρευνας αλγοριθμικών μοντέλων στην πραγματική πράξη σε τομείς όπως η υγειονομική περίθαλψη. Επίσης, το εργαλείο What-If της Google μπορεί να χρησιμοποιηθεί για τη βελτιστοποίηση της δικαιοσύνης της Τεχνητής Νοημοσύνης του μοντέλου.

Human-in-the-loop

Το εκπαιδευμένο μοντέλο DNN μπορεί να προβάλλει προβλέψεις που είναι συγκρίσιμες με αυτές ενός ανθρώπινου εμπειρογνώμονα σε λιγότερο χρόνο, αλλά αυτό δεν αρκεί για να εμπιστευτεί πλήρως κανείς την ιατρική διάγνωση σε έναν αλγόριθμο. Αντ' αυτού, τα DNN και τα XAI μπορούν να υποστηρίξουν τους κλινικούς γιατρούς να λαμβάνουν καλύτερες και πιο τεκμηριωμένες αποφάσεις. Η προσέγγιση που εμπλέκει τον άνθρωπο στον κύκλο επιτρέπει ιδανικά την αλληλεπίδραση με το μοντέλο για την πραγματοποίηση των απαραίτητων αλλαγών προκειμένου να βελτιωθούν οι προβλέψεις του μοντέλου και να διασφαλιστεί η επαρκής επεξηγηματικότητά του. Τα επεξηγήσιμα συστήματα που εμπλέκουν τον ανθρώπινο χρήστη στη διαδικασία λήψης αποφάσεων του DNN είναι χρήσιμα για την αντίληψη του "γιατί" και "πώς" σχετικά με τις προβλέψεις του μοντέλου (Dong et al., 2017). Ένα σύστημα ανθρώπινης συμμετοχής στον κύκλο προτάθηκε για την αναγνώριση δράσης σε βίντεο, με σκοπό να διευκολυνθεί η κατανόηση των χαρακτηριστικών που μαθαίνονται από μη ειδικούς χρήστες (Dong et al., 2017). Οι αποφάσεις που επηρεάζουν τη ζωή των ασθενών δεν μπορούν να εμπιστευτούν πλήρως σε έναν αλγόριθμο. Εν προκειμένω, προσεγγίσεις οπτικοποίησης που συσχετίζουν τα αρχικά δεδομένα και τις οπτικοποιήσεις είναι πιο σημαντικές για την υποστήριξη της λήψης κλινικών αποφάσεων. Το σύστημα υποστήριξης κλινικών

αποφάσεων πρέπει να είναι αξιόπιστο και να ενισχύει την ανθρώπινη απόφαση. Η χρήση του XAI για μια δυαδική απόφαση σχετικά με την παρουσία ή την απουσία μιας νόσου αποτελεί μεγάλη βοήθεια, επιτρέποντας σε πιο σύνθετες περιπτώσεις που είναι δύσκολο να ταξινομηθούν να καθοριστούν σε ένα κλινικό περιβάλλον από έναν ανθρώπινο εμπειρογνώμονα.

Οι υπογραφές εικόνας CT της ραδιονομικής και οι μέθοδοι DNN χρησιμοποιήθηκαν για να αποκτήσουν εικόνα των χαρακτηριστικών της εικόνας CT που είναι σημαντικά για την πρόβλεψη του COVID-19 (Wang et al., 2021).

Οι ερευνητές έχουν συγκρίνει τα χαρακτηριστικά από Δίκτυα Βαθιάς Μάθησης (DNN) με αυτά που παράγουν οι άνθρωποι ακτινολόγοι, με σκοπό να κατανοήσουν και να βελτιώσουν την ερμηνευσιμότητα των αλγορίθμων Μηχανικής Μάθησης (ML) για την κλινική διάγνωση. Η ενσωμάτωση του Αναλυτικού Εξαιρετικού Νευρωνικού (XAI) στην ιατρική πρακτική παραμένει πρόκληση, καθώς λείπει η συστηματική κατανόηση του ρόλου του ανθρώπου σε αυτό το σύστημα. Η χρήση μιας προσέγγισης όπως η Μπεϋζιανή Διδασκαλία μπορεί να βοηθήσει τους κλινικούς ιατρούς να εκμεταλλευτούν τις εξηγήσεις που παρέχονται από το XAI για καλύτερη κατανόηση των διαγνώσεων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Jin, W., Li, X., Hamarneh, G. (2022). Evaluating explainable AI on a multi-modal medical imaging task: can existing algorithms fulfill clinical requirements?, in: Association for the Advancement of Artificial Intelligence Conference AAAI).
- Natekar, P., Kori, A., Krishnamurthi, G. (2020). Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. *Front. Comput. Neurosci.* 14.
- Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S. (2021). CA-net: comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imag.* 40 (2), 699–711.
- Janik, J. Dodd, G. Ifrim, K. Sankaran, K. Curran, Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset, *Med. Imag. 2021: Image Process.* 11596 (2021) 861–872.

Yang, G. Ye, Q., Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. *Inf. Fusion* 77, 29–52.

Komatsu, M. Sakai, A., Dozen, A., Shozu, K., Yasutomi, S., Machino, H., Asada, K., Kaneko, S., Hamamoto, R. (2021). Towards clinical application of artificial intelligence in ultrasound imaging. *Biomedicines*, 9 (7), 720.

Katzmann, A., Taubmann, O., Ahmad, S., Mühlberg, A., Sühling, M., Groß, H. (2021). Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization. *Neurocomputing*, 141–156.

Le, E.P.V., Wang, Y., Huang, Y., Hickman, S., Gilbert, F.J. (2019). Artificial intelligence in breast imaging. *Clin. Radiol.*, 74 (5), 357–366.

Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., Zheng, W., Sun, Q., Lu, L., Shung, K.K. (2021). Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nat. Biomed. Eng.*, 5 (6), 522–532.

Xie, B., Lei, T., Wang, N., Cai, H., Xian, J., He, M., Zhang, L., Xie, H. (2020). Computer-aided diagnosis for fetal brain ultrasound images using deep convolutional neural networks. *Int. J. Comput. Assist. Radiol. Surg.*, 15 (8), 1303–1312.

Zhang, J., Petitjean, C., Ainouz, S. (2022). Segmentation-based vs. Regression-based biomarker estimation: a case study of fetus head circumference assessment from ultrasound images. *Journal of Imaging*, 8 (2).

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S. (2021). CA-net: comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Trans. Med. Imag.* 40 (2), 699–711.

Narayanaswamy, V., Thiagarajan, J.J., Spanias, A. (2021). Using deep image priors to generate counterfactual explanations, in: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Li, W., Zhuang, J., Wang, R., Zhang, J., Zheng, W.S. (2020). Fusing metadata and dermoscopy images for skin disease diagnosis, in: *IEEE 17th International Symposium on Biomedical Imaging, ISBI*.

Barata, C., Celebi, M.E., Marques, J.S. (2021). Explainable skin lesion diagnosis using taxonomies. *Pattern Recogn.*, 110.

Molle, P.V., Strooper, M.D., Verbelen, T., Vankeirsbilck, B., Simoens, P., Dhoedt, B. (2018). Visualizing convolutional neural networks to improve decision support for skin lesion classification, in: *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*.

- Zhang, J., Petitjean, C., Ainouz, S. (2022). Segmentation-based vs. Regression-based biomarker estimation: a case study of fetus head circumference assessment from ultrasound images. *Journal of Imaging*, 8 (2).
- Kind, A., Azzopardi, G. (2019). An explainable AI-based computer aided detection system for diabetic retinopathy using retinal fundus images, in: *International Conference on Computer Analysis of Images and Patterns*, Springer, Cham., pp. 457–468.
- Marginean, A., Groza, A., Nicoara, S., Muntean, G., Slavescu, R.R, Letia, I. (2019). Towards balancing the complexity of convolutional neural network with the role of optical coherence tomography in retinal conditions, in: 15th International Conference on Intelligent Computer Communication and Processing, ICCP.
- Shorfuzzaman, M., Hossain, M.S., Saddik, A.E. (2021). An explainable deep learning ensemble model for robust diagnosis of diabetic retinopathy grading, *ACM Trans. Multimed Comput. Commun. Appl.*, 17 (3s), 1–24.
- Sabol, P., Sinčák, P., Hartono, P., Kočcan, P., Benetinová, Z., Blichárová, A., Verbóová, L., Štammová, E., Sabolová-Fabianová, A., Jašková, A. (2020). Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *J. Biomed. Inf.*, 109.
- Knapic, S., Malhi, A., Saluja, R., Främling, K. (2021). Explainable artificial intelligence for human decision-support system in medical domain. *Machine Learning and Knowledge Extraction*, 3 (3) 740–770.
- Wang, S., Xing, Y., Zhang, L., Gao, H. Zhang, H. (2019). Deep Convolutional Neural Network for Ulcer Recognition in Wireless Capsule Endoscopy: Experimental Feasibility and Optimization,” *Computational And Mathematical Methods in Medicine*.
- Sousa, I.P.d., Vellasco, M.M.B.R., Silva, E.C.d. (2019). Local interpretable model-agnostic explanations for classification of lymph node metastases. *Sensors* 19.
- Uehara, K., Murakawa, M., Nosato, H. Sakanashi, H. (2020). Multi-scale explainable feature learning for pathological image analysis using convolutional neural networks, in: IEEE International Conference on Image Processing, ICIP.
- Jr, L.A.d.S., Mendel, R., Strasser, S., Ebigbo, A. Probst, A., Messmann, H., Papa, J.P., Palm, C. (2021). Convolutional Neural Networks for the evaluation of cancer in Barrett’s esophagus: explainable AI to lighten up the black-box. *Comput. Biol. Med.*, 135.
- Neves, I., Folgado, D., Santos, S., Barandas, M., Campagner, A., Ronzio, L., Cabitza, F., Gamboa, H. (2021). Interpretable heartbeat classification using local model-agnostic explanations on ECGs. *Comput. Biol. Med.*, 133.

- Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.U. (2020). Deep learning interpretation of echocardiograms. *NPJ Digital Med.*, 3 (1), 1–10.
- Reguant, R., Brunak, S., Saha, S. (2021). Understanding inherent image features in CNN-based assessment of diabetic retinopathy. *Sci. Rep.*, 11 (1).
- Quellec, G., Hajj, H.A., Lamard, M., Conze, P.-H., Massin, P., Cochener, B. (2021). ExplAIIn: explanatory artificial intelligence for diabetic retinopathy diagnosis. *Med. Image Anal.*, 72.
- Niu, Y., Gu, L., Zhao, Y., Lu, F. (2021). Explainable diabetic retinopathy detection and retinal image generation. *IEEE journal of biomedical and health informatics*, 1, 1.
- Martins, J., Cardoso, J.S., Soares, F. (2020). Offline computer-aided diagnosis for Glaucoma detection using fundus images targeted at mobile devices. *Comput. Methods Progr. Biomed.*, 192.
- Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, A., Mendonça, A. M., Campilho, A. (2020). DR|GRADUATE: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. *Med. Image Anal.*, 63.
- Kim, M., Han, J.C., Hyun, S.H., Janssens, O., Van Hoecke, S., Kee, C., De Neve, W. (2019) Medinoid: computer-aided diagnosis and localization of glaucoma using deep learning. *Appl. Sci.*, 9 (15).
- Meng, Q., Hashimoto, Y., Satoh, S.I. (2020). How to extract more information with less burden: fundus image classification and retinal disease localization with ophthalmologist intervention. *IEEE J. Biomed. Health Informat.*, 24 (12), 3351–3361.
- Shen, Y., Sheng, B., Fang, R., Li, H., Dai, L., Stolte, S., Qin, J., Jia, W., Shen, D. (2020). Domain-invariant interpretable fundus image quality assessment. *Med. Image Anal.*, 61.
- Seah, J., Tang, J., Kitchen, A., Gaillard, F., Dixon, A. (2019). Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology*, 290 (2), 514–522.
- Heinemann, F., Birk, G., Stierstorfer, B. (2019). Deep learning enables pathologist-like scoring of NASH models. *Sci. Rep.*, 9 (1).
- Teramoto, A. Yamada, A., Kiriya, Y., Tsukamoto, T., Yan, K., Zhang, L., Imaizumi, K. Saito, K., Fujita, H. (2019). Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. *Inform. Med. Unlocked* 16.
- Viton, F., Elbattah, M., Gu´erin, J.L., Dequen, G. (2020). Heatmaps for visual explainability of CNN-based predictions for multivariate time series with application to healthcare, in: International Conference on Healthcare Informatics, ICHI.

- Hassan, M.R., Islam, M.F., Uddin, M.Z., Ghoshal, G., Hassan, M.M., Huda, S., Fortino, G. (2022). Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence. *Future Generat. Comput. Syst.*, 127, 462–472
- Lamy, J.B., Sekar, B., Guezennec, G., Bouaud, J., S'erotoussi, B. (2019). Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. *Artif. Intell. Med.*, 94, 42–53.
- Chereda, H., Bleckmann, A., Menck, K., Perera-Bel, J., Stegmaier, P., Auer, F. Kramer, F., Leha, A., Beißbarth, T. (2021). Explaining decisions of graph convolutional neural networks: patientspecific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Med.* 13 (1).
- Luo, L., Chen, H., Wang, X., Dou, Q., Lin, H., Zhou, J., Li, G., Heng, P. (2019). Deep angular embedding and feature correlation attention for breast MRI cancer analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention.
- van der Velden, B., Janse, M. Ragusi, M., Loo, C., Gilhuijs, K. (2020). Volumetric breast density estimation on MRI using explainable deep learning regression. *Sci. Rep.* 10 (1), 1–9.
- Kim, S.T., Lee, H., Kim, H.G., Ro a.Y.M (2018). ICADx: interpretable computer aided diagnosis of breast masses. *Medical Imaging 2018: Computer-Aided Diagnosis 10575*, 450–459.
- Chen, X., Lin, L., Liang, D., Hu, H., Zhang, Q., Iwamoto, Y., Han, X.U., Chen, Y.W., Tong, R., Wu, L. (2019). A dual-attention dilated residual network for liver lesion classification and localization on CT images. in: IEEE International Conference on Image Processing, ICIP.
- Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.- G., Mahlein, A.-K., Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.*, 2 (8), 476–486.
- Gillmann, C., Smit, N.N., Groller, E., Preim, B., Vilanova, A., Wischgoll, T. (2021). Ten open challenges in medical visualization. *Comput. Graphics Appl.*, 41 (5), 7–15.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F. Giannotti, F., Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51 (5), 1–42.
- Ghassemi, M., Oakden-Rayner, L., Beam, A.L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3 (11).

- LaLonde, R., Torigian, D., Bagci, U. (2020). Encoding visual attributes in capsules for explainable medical diagnoses. in Medical Image Computing and Computer Assisted Intervention – MICCAI.
- Hall, P., Gill, N. (2018). An Introduction to Machine Learning Interpretability. *O'Reilly Media*.
- Tayarani-N, M.-H. (2020). Applications of Artificial Intelligence in Battling against Covid- 19: A Literature Review,” *Chaos, Solitons & Fractals*.
- Kim, B., Khanna, R., Koyejo, O. (2016). Examples are not enough, learn to criticize! criticism for interpretability, in: Proc. 29th Conf. Neural Inf. Process. Syst, NIPS.
- Yim, J., Sohn, K. (2017). Enhancing the performance of convolutional neural networks on quality degraded datasets, in: International Conference on Digital Image Computing: Techniques and Applications, DICTA.
- Singh, R.K., Pandey, R., Babu, R.N. (2021). COVIDScreen: Explainable Deep Learning Framework for Differential Diagnosis of COVID-19 Using Chest X-Rays,” *Neural Computing And Applications*.
- Hampe, N., Wolterink, J.M., M.v. Velzen, S.G., Leiner, T. (2019). I. Isgum, Machine Learning for Assessment of Coronary Artery Disease in Cardiac CT: A Survey,” *Frontiers In Cardiovascular Medicine*, vol. 6.
- Knight, W. (2017). The U.S. Military Wants its Autonomous Machines to Explain Themselves, MIT Technology Review.
- Dong, Y., Su, H., Zhu, J., Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Wang, H., Wang, L., Lee, E.H., Zheng, J., Zhang, W., Halabi, S., Liu, C., Deng, K., Song, J., Yeom, K.W. (2021). Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. *Eur. J. Nucl. Med. Mol. Imag.*, 48 (5), 1697.
- Folke, T., Yang, S. Anderson, S., Shafto, P. (2021). Explainable AI for medical imaging: explaining pneumothorax diagnoses with Bayesian teaching. *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, 11746, 644–664.
- LaLonde, R., Torigian, D., Bagci, U. (2020). Encoding visual attributes in capsules for explainable medical diagnoses. in Medical Image Computing and Computer Assisted Intervention – MICCAI.
- Nazir, S. Patel, S., Patel, D. (2022). Model optimisation techniques for convolutional neural networks, in: Handbook of Research on New Investigations in Artificial Life, AI, and Machine Learning, IGI Global, pp. 269–298.
- Yoo, H., Han, S., Chung, K. (2021). Diagnosis support model of cardiomegaly based on CNN using ResNet and explainable feature map. *IEEE Access* 9, 55802–55813.

- Yi, P.H., Kim, T.K., Wei, J. Shin, J., Hui, F.K., Sair, H.I., Hager, D.G., Fritz, J. (2019). Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning. *Pediatr. Radiol.*, 49 (8), 1066–1070.
- Ahsan, M.M., Gupta, K.D., Islam, M.M., Sen, S., Rahman, M.L., Hossain, M.S. (2020). COVID- 19 symptoms detection based on NasNetMobile with explainable AI using various imaging modalities. *Machine Learning and Knowledge Extraction*, 2 (4), 490–504.
- Alshazly, H., Linse, C., Barth, E., Martinetz, T. (2021). Explainable COVID-19 detection using chest CT scans and deep learning. *Sensors*, 21 (2).
- Yang, G., Ye, Q., Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. *Inf. Fusion* 77, 29–52.
- Pennisi, M., Kavasidis, I., Spampinato, C., Schinina, V., Palazzo, S., Salanitri, F.P., Bellitto, G., Rundo, F., Aldinucci, M. Cristofaro, M., Campioni, A.P. (2021). An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. *Artif. Intell. Med.*, 118.
- Wu, Y.-H., Gao, S.-H., Mei, J., Xu, J., Fan, D.P., Zhang, R.G., Cheng, M.M. (2021) JCS: an explainable COVID-19 diagnosis system by Joint classification and segmentation. *IEEE Trans. Image Process*, 30, 3113–3126.
- Ko, H., Chung, H., Kang, W.S., Kim, W.K., Shin, Y., Kang, S.J., Lee, L.H., Kim, Y.J., Kim, N.Y., Jung, H., Lee, J. (2020). COVID-19 pneumonia diagnosis using a simple 2D deep learning framework with a single chest CT image: model development and validation. *J. Med. Internet Res.*, 22 (6).
- Xu, R., Cong, Z., Ye, X., Hirano, Y., Kido, S., Gyobu, T., Kawata, Y., Honda, O., Tomiyama, N. (2019). Pulmonary textures classification via a multi-scale attention network. *IEEE journal of biomedical and health informatics* 24 (7), 2041–2052.
- Hilbert, A., Ramos, L.A., van Os, H.J., Olabarriaga, S., Tolhuisen, M., Wermer, M.J., Barros, R.S., van der Schaaf, I., Dippel, D., Roos, Y.W.E.M., van Zwam, W. (2019). Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Comput. Biol. Med.*, 115.
- Kumar, A., Manikandan, R., Kose, U., Gupta, D., Satapathy, S.C. (2021). Doctor’s dilemma: evaluating an explainable subtractive spatial Lightweight convolutional neural network for brain tumor diagnosis, *ACM Trans. Multimed Comput. Commun. Appl.*, 17 (3s), 1–26.
- Lopatina, A., Ropele, S., Sibgatulin, R., Reichenbach, J.R., Güllmar, D. (2020). Investigation of deep-learning-driven identification of multiple sclerosis patients

based on susceptibility-weighted images using relevance analysis. *Front. Neurosci.*, 14.

Kumar, D., Sankar, V., Clausi, D., Taylor, G.W., Wong, A. (2019). SISC: end-to-end interpretable discovery radiomics-driven lung cancer prediction via stacked interpretable sequencing cells. *IEEE Access* 7, 145444–145454.

Lee, E., Choi, J.S., Kim, M., Suk, H.I. (2019). Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning. *Neuroimage* 202.

Eitel, F., Soehler, E., Bellmann-Strobl, J., Brandt, A., Ruprecht, K., Giess, R., Kuchling, J., Asseuer, S., Weygandt, M., Haynes, J., Scheel, M. (2019). Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *Neuroimage, Clinical* 24.

Kunapuli, G., Varghese, A., Ganapathy, P., Desai, B., Cen, S., Aron, M., Gill, I., Duddalwar, V. (2018). A decision-support tool for renal mass classification. *Journal of digital imaging. J. Digit. Imag.*, 31 (6), 929–939.

Levakov, G. Rosenthal, G., Shelef, I., Raviv, T.R., Avidan, G. (2020). From a deep learning model back to the brain—identifying regional predictors and their relation to aging. *Hum. Brain Mapp.*, 41 (12), 3235–3252.

Tian, J., Li, C., Shi, Z., Xu, F. (2020). A diagnostic report generator from CT volumes on liver tumor with semi-supervised attention mechanism, in International Conference on Medical Image Computing and Computer-Assisted Intervention.

Bohle, M., Eitel, F., Weygandt, M., Ritter, K. (2019). Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front. Aging Neurosci.*, 11.

Kamal, M.S., Northcote, A., Chowdhury, L., Dey, N., Crespo, G.R., Herrera- Viedma, E. (2021). Alzheimer's patient analysis using image and gene expression data and explainable-AI to present associated genes. *IEEE Trans. Instrum. Meas.*, 70, 1–7.

Shapira, N., Fokuhl, J., Schultheiß, M., Beck, S., Kopp, F.K., Pfeiffer, D., Dangelmaier, J., Pahn, G., Sauter, A.P., Renger, B., Fingerle, A.A. (2020). Liver lesion localisation and classification with convolutional neural networks: a comparison between conventional and spectral computed tomography. *Biomed. Phys. Eng. Express*, 6 (1).

Rajpurkar, P., Park, A., Irvin, J., Chute, C., Bereket, M., Mastrodicasa, D., Langlotz, M.P. Lungren, A.Y. Ng, B.N. Patel (2020). AppendiXNet: deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Sci. Rep.* 10 (1), 1–7.

Windisch, P. Weber, P., Fürweger, C., Ehret, F., Kufeld, M., Zwahlen, D., Muacevic, A. (2020). Implementation of model explainability for a basic brain tumor detection using convolutional neural networks on MRI slices. *Neuroradiology*, 62 (11).

- Paul, R., Schabath, M., Gillies, R., Hall, L., Goldgof, D. (2020). Convolutional Neural Network ensembles for accurate lung nodule malignancy prediction 2 years in the future. *Comput. Biol. Med.*, 122.
- Candemir, S., White, R.D., Demirer, V. Gupta, M.T. Bigelow, L.M. Prevedello, B. S. Erdal (2020). Automated coronary artery atherosclerosis detection and weakly supervised localization on coronary CT angiography with a deep 3-dimensional convolutional neural network. *Comput. Med. Imag. Graph.*, 83.
- Chakraborty, S. Aich, S., Kim, H.C. (2020). Detection of Parkinson's disease from 3T T1 weighted MRI scans using 3D convolutional neural network. *Diagnostics*, 10 (6).
- Pereira, S., Meier, R., Alves, V., Reyes, M., Silva, C. (2018). Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment, in: *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*.
- Shinde, S. Prasad, S., Saboo, Y., Kaushick, R., Saini, J., Pal, P.K., Ingahalikar, M. (2019). Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI, *Neuroimage: Clinical* 22.
- Humphries, S.M., Notary, A.M., Centeno, J.P., Strand, M.J., Crapo, J.D., Silverman, E. K., Lynch, D.A. (2020). Deep learning enables automatic classification of emphysema pattern at CT. *Radiology*, 294 (2), 434–444.
- Couteaux, V., Nempont, O., Pizaine, G., Bloch, I. (2019). Towards interpretability of segmentation networks by analyzing deepdreams, in: *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*.
- Chen, X., Lin, L., Liang, D., Hu, H., Zhang, Q., Iwamoto, Y., Han, X.H., Chen, Y.W., Tong, R., Wu, J. (2019). A dual-attention dilated residual network for liver lesion classification and localization on CT images, in: *IEEE International Conference on Image Processing, ICIP*.
- Huo, Y., Terry, J.G., Wang, J., Nath, V., Bermudez, C., Bao, S., Parvathaneni, P., Carr, J. J., Landman, B.A. (2019). Coronary calcium detection using 3D attention identical dual deep network based on weakly supervised learning. *Med. Imaging Image Process.* 10949, 308–315.
- Lei, Y. Tian, Y., Shan, H., Zhang, J., Wang, G., Kalra, M.K. (2020). Shape and margin-aware lung nodule classification in low-dose CT images via soft activation mapping. *Med. Image Anal.*, 60.
- Newell, A., Simon, H.A. (1976). Computer science as empirical inquiry: Symbols and search. *Commun. ACM*, 19, 113–126
- Smolensky, P. (2010). On the proper treatment of connectionism. *Behav. Brain Sci.*, 11, 1–23

- Ho, T.K. (1995). Random Decision Forests. In ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition, IEEE Computer Society: Washington, DC, USA, Volume 1, p. 278.
- Breiman, L. (2001). Random Forests. *Mach. Learn.*, 45, 5–32.
- Lötsch, J., Malkusch, S. (2021). Interpretation of cluster structures in pain-related phenotype data using explainable artificial intelligence (XAI). *Eur. J. Pain*, 25, 442–465
- Banerjee, M., Ding, Y., Noone, A.M. (2012). Identifying representative trees from ensembles. *Stat. Med.*, 31, 1601–1616.
- Ribeiro, M.T., Singh, S., Guestrin, C. (2016). Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17, pp. 1135–1144.
- Liaw, A., Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2, 18–22
- Lötsch, J., Malkusch, S. (2021). Interpretation of cluster structures in pain-related phenotype data using explainable artificial intelligence (XAI). *Eur. J. Pain*, 25, 442–465.
- Dasgupta, A. (2014). Reprtree: Representative Trees from Ensembles. Available online: <https://github.com/araastat/reprtree/blob/master/R/ReprTree.R> (accessed on 15 December 2021).
- Holzinger, A. (2020). Explainable AI and Multi-Modal Causability in Medicine. *i-com*, 19, 171–179.
- Bach, S., Binder, A., Müller, K.R., Samek, W. (2016). Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA.
- Montavon, G. (2019). Gradient-Based vs. Propagation-Based Explanations: An Axiomatic Comparison, Springer: Cham, Switzerland, pp. 253–265.
- Linardatos, P., Papastefanopoulos, V., Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23, 18.
- Arrieta, A.B., Díaz-Rodríguez, N., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion*, 58, 82–115.
- Thiebes, S., Lins, S., Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electron. Mark.*, 31, 447–464.

Huang, X., Baker, J., Reddy, R. (2014). A Historical Perspective of Speech Recognition. *Commun. ACM*, 57, 94–103.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J.M., Nguyen, H., Gadde, R.T. (2019). Jasper: An End-to-End Convolutional Neural Acoustic Model., arXiv:1904.03288.

Michalski, R.S. (1983). A theory and methodology of inductive learning. In *Machine Learning*, Michalski, R.S., Carbonell, J.G., Mitchell, T.M., Eds., Morgan Kaufmann: San Francisco, CA, USA, pp. 83–134.

Craven, M.W., Shavlik, J.W. (1996). *Extracting Comprehensible Models from Trained Neural Networks*, Computer Sciences Department, University of Wisconsin-Madison: Madison, WI, USA.

Yanase, J., Triantaphyllou, E. (2019). The seven key challenges for the future of computer-aided diagnosis in medicine. *Int. J. Med. Inf.*, 129, 413–422.

Ultsch, A., Kleine, T., Korus, D., Farsch, S., Guimaraes, G., Pietzuch, W., Simon, J. (2006). Evaluation of Automatic and Manual Knowledge Acquisition for Cerebrospinal Fluid (CSF) Diagnosis. In *Artificial Intelligence in Medicine*, Keravnou, E., Garbay, C., Baud, R., Wyatt, J., Eds., Springer: Berlin/Heidelberg, Germany, Volume 1211.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: Tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet.*, 25, 25–29.

Hodges, W. (2001). *Classical Logic I: First Order Logic*. In *The Blackwell Guide to Philosophical Logic*, Wiley-Blackwell: Hoboken, NJ, USA.

Loh, W.Y. (2011). Classification and regression trees. *WIREs Data Min. Knowl. Discov.*, 1, 14–23.

Cohen, W.W. (1995). Fast Effective Rule Induction. In *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, CA, USA, 9–12 July 1995, Morgan Kaufmann: Burlington, MA, USA, pp. 115–123.

Marewski, J.N., Gigerenzer, G. (2012). Heuristic decision making in medicine. *Dialogues Clin. Neurosci.*, 14, 77–89.

Miller, G.A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.*, 63, 81–97.

Holzinger, A. (2020). Explainable AI and Multi-Modal Causability in Medicine. *i-com*, 19, 171–179

Bach, S., Binder, A., Müller, K.R., Samek, W. (2016). Controlling Explanatory Heatmap Resolution and Semantics via Decomposition Depth. In *Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA.

- Montavon, G. (2019). Gradient-Based vs. Propagation-Based Explanations: An Axiomatic Comparison, Springer: Cham, Switzerland, pp. 253–265.
- Linardatos, P., Papastefanopoulos, V., Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23, 18.
- Thiebes, S., Lins, S., Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electron. Mark.*, 31, 447–464
- Ribeiro, M.T., Singh, S., Guestrin, C. (2016). Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17, pp. 1135–1144.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65, 386–408
- Ho, T.K. (1995). Random Decision Forests. In ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition, IEEE Computer Society: Washington, DC, USA, Volume 1, p. 278.
- Breiman, L. (2001). Random Forests. *Mach. Learn.*, 45, 5–32
- Ultsch, A. (2007). Emergence in Self-Organizing Feature Maps. In Proceedings of the International Workshop on Self-Organizing Maps (WSOM '07), Bielefeld, Germany, 3–6 September 2007, Ritter, H., Haschke, R., Eds., Neuroinformatics Group, Bielefeld University: Bielefeld, Germany.
- Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy.
- Lee, E., Choi, J.S., Kim, M., Suk, H.I. (2019). Toward an interpretable Alzheimer's disease diagnostic model with regional abnormality representation via deep learning. *Neuroimage*, 202, 116113.
- Papadimitroulas, P., Brocki, L., Christopher Chung, N., Marchadour, W., Vermet, F., Gaubert, L., Eleftheriadis, V., Plachouris, D., Visvikis, D., Kagadis, G.C., et al. (2021). Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Phys. Med.*, 83, 108–121.
- Ultsch, A., Thrun, M.C., Hansen-Goos, O., Lötsch, J. (2015). Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox (AdaptGauss). *Int. J. Mol. Sci.*, 16, 25897–25911.
- Malkusch, S., Hahnefeld, L., Gurke, R., Lötsch, J. (2021). Visually guided preprocessing of bioanalytical laboratory data using an interactive R notebook (pguIMP). *CPT Pharmacometrics Syst. Pharmacol.*, 10, 1371–1381.

- Lötsch, J., Sipilä, R., Tasmuth, T., Kringel, D., Estlander, A.M., Meretoja, T., Kalso, E., Ultsch, A. (2018). Machine-learning-derived classifier predicts absence of persistent pain after breast cancer surgery with high accuracy. *Breast Cancer Res Treat.*, 171, 399–411.
- Li, Y., Fujiwara, T., Choi, Y.K., Kim, K.K., Ma, K.L. (2020). A visual analytics system for multi-model comparison on clinical data predictions. *Vis. Inform.*, 4, 122–131.
- Liao, Q.V., Gruen, D., Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA.
- Lötsch, J., Sipilä, R., Dimova, V., Kalso, E. (2018). Machine-learned selection of psychological questionnaire items relevant to the development of persistent pain after breast cancer surgery. *Br. J. Anaesth.*, 121, 1123–1132
- Benioff, M.R., Lazowska, E.D., Bajcsy, R., Beese, J.C., Celis, P., Evans, P.T., Yang, G. (2005). Report to the President: Computational Science: Ensuring America's Competitiveness, President's Information Technology Advisory Committee: Washington, DC, USA.
- Datta, A., Matlock, M.K., Le Dang, N., Moulin, T., Woeltje, K.F., Yanik, E.L., Joshua Swamidass, S. (2021). 'Black Box' to 'Conversational' Machine Learning: Ondansetron Reduces Risk of Hospital-Acquired Venous Thromboembolism. *IEEE J. Biomed. Health Inf.*, 25, 2204–2214.
- Bhattacharya, J., Vogt, W.B. (2012). Do Instrumental Variables Belong in Propensity Scores? *Int. J. Stat. Econ*, 9, A12
- VanderWeele, T.J. (2019). Principles of confounder selection. *Eur. J. Epidemiol.*, 34, 211–219.
- Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Inf.*, 3, 119–131.
- Lötsch, J., Daiker, H., Hähner, A., Ultsch, A., Hummel, T. (2015). Drug-target based cross-sectional analysis of olfactory drug effects. *Eur. J. Clin. Pharmacol.*, 71, 461–471.
- Kovalchik, S. (2021). RISmed: Download Content from NCBI Databases, R Package Version 2.3.0, 2021. Available online: <https://cran.rproject.org/package=RISmed> (accessed on 15 December 2021)
- Fletcher, K.H. (1951). Matter with a mind, a neurological research robot. *Research*, 4, 305–307
- Lanzola, G., Stefanelli, M., Barosi, G., Magnani, L. (1990). NEOANEMIA: A knowledge-based system emulating diagnostic reasoning. *Comput. Biomed. Res.*, 23, 560–582

