

DE WAARDE VAN ARTIFICIËLE INTELLIGENTIE BIJ BORSTKANKERSCREENING MET MAMMOGRAFIE

Helena Laseure

Stamnummer: 01808385

Promotoren: Prof. Dr. Koenraad Verstraete en Prof. Dr. Erik Ranschaert

Masterproef voorgelegd in het kader tot het behalen van de graad Master of Medicine in de
Geneeskunde

Academiejaar: 2023 – 2024



**UNIVERSITEIT
GENT**

MAATSCHAPPELIJKE OUTREACH: 'DE WAARDE VAN ARTIFICIËLE INTELLIGENTIE BIJ BORSTKANKERSCREENING MET MAMMOGRAFIE'

Mammografische borstkankerscreening is een onderwerp waarmee veel vrouwen geconfronteerd worden in hun levensloop. Zo krijgen in België alle vrouwen tussen 50 en 69 jaar oud de mogelijkheid om zich eens om de twee jaar te laten screenen op borstkanker. Echter verschijnen er tegenwoordig steeds meer artikelen in de media die de opkomst van AI-algoritmen en hun potentiële impact op radiologen en beeldvorming bespreken. Dat kan soms nogal een overdreven beeld geven. Deze masterproef onderzocht de daadwerkelijke toegevoegde waarde van AI bij borstkankerscreening met mammografie. Uit het uitgevoerde literatuuronderzoek kunnen verschillende conclusies worden getrokken.

Allereerst bevindt AI bij borstkankerscreening zich nog in een vroeg stadium van ontwikkeling. De meeste onderzoeken tot op heden gebeurden niet in een screeningsomgeving. Daarom kunnen de resultaten die daaruit voortkwamen nog niet direct worden toegepast op grootschalige screeningsprogramma's, zoals die in België systematisch voor vrouwen worden uitgevoerd. Verder onderzoek op een echte screeningspopulatie is noodzakelijk. Het zal dan ook nog even duren voordat we AI-algoritmen geïntegreerd zien in het dagelijkse werk van radiologen.

Vervolgens zal AI in de toekomst mogelijk dienen ter ondersteuning van radiologen. Er zouden namelijk vooral voordelen op het gebied van nauwkeurigheid en prestaties van radiologen worden waargenomen, waar ze een verbetering kan betekenen. AI heeft ook de mogelijkheid om bepaalde mammografieën voor te selecteren, dit veroorzaakt niet enkel een daling in de werklust van de radiologen, maar kan er ook voor zorgen dat er meer aandacht wordt besteed aan gevallen met een hoge verdenking van kanker. Verder zou AI vroegtijdig kankers kunnen detecteren die gemist worden door radiologen en zou AI ook in staat zijn tot het geven van persoonlijke screeningsaanbevelingen.

Natuurlijk brengt het gebruik van AI-algoritmen in de praktijk ook enkele uitdagingen met zich mee. Velen hebben ethische en juridische bedenkingen, vooral wanneer deze algoritmen zonder radiologische beoordeling zouden worden toegepast. Bovendien vormt de aanzienlijke investering gepaard gaande met de aanschaf van een AI-systeem een financiële overweging. Verder blijft wegens een tekort aan studies op een echte screeningspopulatie de algemene toepasbaarheid ervan nog onduidelijk.

Concluderend staat het gebruik van AI-algoritmen bij mammografische borstkankerscreening nog in haar kinderschoenen. Er zijn potentiële voordelen, maar er is eveneens behoefte aan toekomstig onderzoek. Het is tot nader order niet realistisch om te denken dat een AI-algoritme radiologen volledig zal vervangen, ondanks soms voorkomende beweringen in de media.

Het bestuderen van de waarde van AI bij borstkankerscreening met mammografie heeft een belangrijke maatschappelijke meerwaarde en impact. Ten eerste kunnen AI-algoritmen door de verhoging van diagnostische nauwkeurigheid en efficiëntie leiden tot een verbeterde patiëntenzorg. Vervolgens kunnen dalingen in het aantal vals-negatieve en vals-positieve resultaten ervoor zorgen dat er respectievelijk minder intervalekankers en minder onterechte kankerdiagnoses zullen optreden. Dit brengt een kostenbesparing voor de gezondheidszorg met zich mee, meer geavanceerde behandelingen en onnodige onderzoeken worden namelijk vermeden. Deze kostenbesparing zet zich eveneens verder in de winst van tijd, werklast en capaciteit van radiologen. Ten slotte kunnen AI-algoritmen ervoor zorgen dat screeningsprogramma's worden uitgebreid naar afgelegen of achtergestelde regio's waar er mogelijk een tekort aan radiologen is.

INHOUDSOPGAVE

INHOUDSOPGAVE	0
1. ABSTRACT	1
2. INLEIDING	3
2.1 Borstkanker	3
2.2 Borstkankerscreening door middel van mammografie	3
2.3 De impact van 'Computer-Aided Detection and Diagnosis' op mammografische screening	4
2.4 De mogelijke rol van artificiële intelligentie bij mammografische screening	5
2.4.1 Nauwkeurigheid van mammografische screening	8
2.4.2 Intervalkanker en de waarde van AI	9
2.4.3 Impact van AI op de werklast en het prestatievermogen van radiologen	11
2.4.4 Personaliseren van borstkankerscreeningsstrategieën	12
2.4.5 Voorkeuren van radiologen, eerstelijnszorgverleners en vrouwen met betrekking tot het gebruik van AI bij borstkankerscreening	12
2.4.6 Bias en generaliseerbaarheid van AI	13
2.4.7 Doelstelling van dit literatuuronderzoek	13
3. METHODOLOGIE	14
4. RESULTATEN	16
4.1 Impact van AI op de nauwkeurigheid, de werklast en het prestatievermogen van radiologen	16
4.1.1 AI-systeem ter ondersteuning van radiologen	16
4.1.2 AI-systeem als autonome lezer	20
4.1.3 AI-systeem als triagemodel	22
4.1.4 Detectie, lokalisatie en classificatie van laesies d.m.v. deep learning	24
4.2 Intervalkanker en de waarde van AI	27
4.3 Personaliseren van borstkankerscreeningsstrategieën	32
4.4 Voorkeuren van radiologen, eerstelijnszorgverleners en vrouwen met betrekking tot het gebruik van AI bij borstkankerscreening	33
4.5 Bias en generaliseerbaarheid van AI	35
5. DISCUSSIE	36
5.1 Optimalisatie van radiologische beoordelingen: de impact van AI-algoritmen op de nauwkeurigheid, de werklast en het prestatievermogen van radiologen	36
5.1.1 De ondersteunende rol van AI-algoritmen voor de radioloog	36
5.1.2 Diagnostische mogelijkheden met een AI-systeem als autonome lezer	39
5.1.3 AI-triagemodellen als efficiënte voorselectiemethode voor mammografieën	40
5.1.4 Vooruitgang in laesiedetectie, lokalisatie en classificatie d.m.v. verschillende deep learning raamwerken	42
5.2 De opmerkelijke rol van AI-algoritmen in intervalkankerdetectie	43
5.3 De potentiële bijdrage van AI-algoritmen aan gepersonaliseerde screeningsstrategieën	45
5.4 Belanghebbendenanalyse: AI-voorkeuren van radiologen, eerstelijnszorgverleners en vrouwen	46
5.5 Kritische blik: bias en generaliseerbaarheid van AI-systemen	47
6. CONCLUSIE	47
7. REFERENTIELIJST	48
8. ADDENDUM	51
8.1 Acroniemenlijst in volgorde van voorkomen in deze masterproef	51
8.2 Tabellen en figuren	52

1. ABSTRACT

Doelstelling: Borstkanker is wereldwijd bij vrouwen de meest voorkomende vorm van kanker en de belangrijkste oorzaak van kankergerelateerde sterfte. Vroegtijdige detectie ervan door middel van mammografie, het standaard screeningsonderzoek, leidt tot een betere prognose en aldus een vermindering van de mortaliteit. Mammografische screening omvat het maken van twee opnames van elke borst, beoordeeld door radiologen, en kan leiden tot verdere evaluatie en diagnostisch onderzoek in geval van verdachte bevindingen. In het verleden werd reeds 'Computer-Aided Detection and Diagnosis' (CAD) software ontwikkeld met als bedoeling de interpretatie van mammografieën te ondersteunen en te verbeteren. Helaas bleek deze niet het nodige prestatieniveau bereikt te hebben om de screeningsprestaties in een echte klinische screeningsomgeving te verbeteren. Intussen is echter een nieuwe beloftevolle vorm van radiologische ondersteuning in opmars, namelijk AI-algoritmen ter verbetering van de efficiëntie van mammografische screening. Dit literatuuronderzoek, van het type gestructureerde narrative review, heeft als doelstelling te achterhalen wat de waarde van AI bij borstkankerscreening met mammografie precies kan betekenen.

Methodologie: De artikelen in deze masterproef werden verzameld uit de databanken 'Google Scholar', 'PubMed' en 'Embase', en werden aangevuld met enkele door de promotoren aangereikte artikelen. Deze publicaties dateren van 2017 tot heden. In totaal werden uiteindelijk 56 artikelen en één online bron in de referentielijst geïnccludeerd.

Resultaten: Deze literatuurstudie bevat een uitgebreid overzicht van de recente ontwikkelingen in AI-algoritmen bij mammografische borstkankerscreening. Ten eerste wordt de impact van AI op de nauwkeurigheid, de werklast en het prestatievermogen van radiologen besproken. Verschillende onderzoeken tonen aan dat radiologen met ondersteuning van AI-algoritmen – al dan niet significant – betere diagnostische prestaties leveren bij het beoordelen van mammografieën. Onderzoeken naar de mogelijkheid van AI als onafhankelijke lezer tonen dan weer vergelijkbare of zelfs betere prestaties dan de gemiddelde radioloog. Verder is het mogelijk om automatisch onderzoeken vooraf te selecteren met behulp van AI, om de efficiëntie van de mammografische borstkankerscreening te verbeteren en de werklast aanzienlijk te verminderen. Een opmerking hierbij is dat verschillende AI-raamwerken andere voor- en nadelen met zich meebrengen. Ten tweede zorgt AI bij mammografische borstkankerscreening, vooral in combinatie met borstdensiteitsmetingen, voor een toegenomen detectie van intervalekankers. Onderzoek naar het personaliseren van screeningsstrategieën toont dan weer aan dat AI mogelijk kan bijdragen aan screeningsaanbevelingen voorspeld uit

patiëntrisicobeoordelingen. Vervolgens is er nood aan toekomstig onderzoek dat zich onder andere zal moeten toespitsen op de voorkeuren van de belanghebbenden (radiologen, eerstelijnszorgverleners en vrouwen). Ten slotte, aangezien de meeste onderzoeken opgenomen in deze literatuurstudie van retrospectieve aard zijn, zal verder prospectief onderzoek moeten gebeuren naar de generaliseerbaarheid en bias van AI-algoritmen in een echte klinische screeningspopulatie en omgeving. Dit exhaustieve literatuuronderzoek onderscheidt zich van reeds bestaande reviews door de uitgebreide integratie van diverse benaderingen waarbij AI van waarde kan zijn bij mammografische borstkankerscreening.

Conclusie: AI-algoritmen hebben toegevoegde waarde bij mammografische borstkankerscreening door het ondersteunen van radiologen, resulterend in o.a. een verbeterde nauwkeurigheid, efficiëntie en verminderde werklust. Echter, prospectief onderzoek in een klinische screeningspopulatie en omgeving blijft essentieel om het volledige potentieel van AI te achterhalen.

2. INLEIDING

2.1 Borstkanker

Wereldwijd is borstkanker de meest voorkomende kanker gediagnosticeerd bij vrouwen (1-8). Bij vrouwen is ze eveneens de belangrijkste oorzaak van kanker gerelateerde mortaliteit (1, 3, 7, 9). Borstkanker is een heterogene ziekte met tumoren die kunnen verschillen in morfologie, moleculair profiel en respons op therapie (10).

Vroegtijdige detectie van borstkanker leidt tot hogere overlevingskansen (4, 8, 11, 12) en dus een reductie van de mortaliteit (13, 14). Vroege detectie resulteert in een betere prognose (12), een betere behandeling en een verminderde ziektelast (5). Deze gunstige factoren hebben ertoe geleid dat er in een aantal landen screeningsprogramma's verricht worden (3, 5, 9, 15, 16).

2.2 Borstkankerscreening door middel van mammografie

Mammografie is het standaard screeningsonderzoek voor borstkanker bij de algemene bevolking (11, 17, 18). Mammografische screening is een nuttige tool voor de vroegtijdige detectie van borstkanker (15, 18-21). Deze screening gaat gepaard met een daling van de aan borstkanker gerelateerde mortaliteit (1, 3, 6, 13, 15, 16, 19, 22-29). Dit werd aangetoond in meerdere gerandomiseerde gecontroleerde studies (1, 3, 19, 23, 26, 27, 29).

Voornamelijk in Europa, maar ook in vele andere landen, is borstkankerscreening een onderdeel van een door de overheid gestuurde gezondheidsstrategie (25, 30). Doorgaans ontvangen vrouwen van de leeftijdscategorie tussen de 50 en 70 jaar een uitnodiging om periodiek een mammografisch screeningsonderzoek te ondergaan. Dit onderzoek vindt meestal om de twee jaar plaats (25). De opgeroepen leeftijdscategorieën en het periodieke interval variëren naargelang het land van screening (30, 31).

Per mammografische screening worden twee opnames van elke borst gerealiseerd, dit zijn de craniocaudale en mediolaterale schuine opnames (7, 25). In vele borstkankerscreeningsprogramma's – in het bijzonder in Europa – worden per patiënte de gemaakte mammografieën afzonderlijk beoordeeld door twee radiologen (25, 30). Als de twee radiologen het oneens zijn over de bevindingen wordt het onderzoek beoordeeld door een derde radioloog (25, 30). Dit proces wordt een 'dubbele lezing' genoemd (25, 29, 30), welke een verbeterde detectie van kanker bij screening (25, 29) en dus een verhoging van de sensitiviteit (3, 20, 32) tot gevolg heeft. Via deze procedure wordt eveneens beslist tot het al

dan niet verder evalueren van een verdachte bevinding. De opnames worden ook vergeleken met eerder gemaakte mammografieën (25). De patiënte ondergaat verder diagnostisch onderzoek wanneer ze een heroproep ('recall') ontvangt (25). Op basis van de methodiek van dubbele lezing wordt aldus beslist of er sprake is van een vals-positieve bevinding of dan wel een reële suspecte laesie waarvoor een aanvullende biopsie nodig is (25). De uitslag van deze biopsie bepaalt uiteindelijk de definitieve diagnose en de daarop aansluitende therapie (25).

2.3 De impact van 'Computer-Aided Detection and Diagnosis' op mammografische screening

In 1998 werd 'Computer-Aided Detection and Diagnosis' (CAD) software ontwikkeld met als doel de interpretatie van mammografieën te ondersteunen en te verbeteren (19, 29, 33, 34). Er is sprake van oude generatie/conventionele CAD-systemen en nieuwe generatie op AI gebaseerde CAD-systemen.

Onder de definitie van conventionele of klassieke CAD-systemen vallen twee categorieën computeralgoritmen, namelijk de computer-aided detection (CAdE) en de computer-aided diagnosis (CAdx) (25). CAdE-algoritmen zijn getraind op het lokaliseren van verdachte laesies, terwijl CAdx-algoritmen getraind zijn op het inschatten van de goedaardigheid of kwaadaardigheid van een al gedetecteerde laesie (25). Om deze taken te kunnen uitvoeren, maakt de conventionele software gebruik van algoritmen die werden getraind om te zoeken naar specifieke morfologische afwijkingen en kenmerken. Deze kenmerken werden voorafgaand door mensen geïdentificeerd als zijnde representatief voor benigne dan wel maligne laesies; dit noemt men geprogrammeerde kenmerken ('programmed-in features') (25). Deze geprogrammeerde kenmerken vormen het belangrijkste onderscheid tussen de conventionele CAD-systemen en de huidige op AI gebaseerde systemen (cfr. infra) (25).

Kort samengevat worden conventionele CAD-systemen gevormd uit algoritmen die verdachte laesies herkennen door middel van door experts geprogrammeerde kenmerken. Ze dienen als hulpmiddel voor de radioloog bij het interpreteren van mammografieën (25, 30).

Ondanks het feit dat de conventionele CAD-systemen al twee decennia lang gebruikt worden, staan de werkzaamheid (29) en doeltreffendheid (1, 22) ervan tot op heden nog onder discussie. Over de hele lijn bestaat er onduidelijkheid over het voordeel van het gebruik van conventionele CAD-systemen bij screening (22).

Eerdere studies gepubliceerd in de jaren 2000-2001 met betrekking tot conventionele CAD-systemen rapporteerden een geoptimaliseerde detectie van kanker (1, 29). Echter toonden op grote schaal uitgevoerde studies een hoog percentage vals-positieve resultaten (1, 24, 25, 29) en m.a.w. een lage specificiteit van deze systemen aan (6, 12, 22, 25, 29, 33, 35). Door de grote hoeveelheid overbodige (vals-positieve) markeringen die de conventionele CAD-systemen weergaven op de mammografische opnames slaagden deze systemen er niet in om de diagnostische prestaties van radiologen te verbeteren (1, 7, 24, 29, 36, 37). De conventionele CAD-systemen hadden dus een negatief effect op de nauwkeurigheid van radiologen (27), wat een toename van onnodige heroproepen (12, 27), onnodige onderzoeken (1, 24) en biopsie-aantallen tot gevolg had (20). Dit alles zorgde eveneens voor een kostentoeename (20).

Niettegenstaande hun vroege hoop, beloften en jarenlang klinisch gebruik, kan geconcludeerd worden dat de conventionele CAD-systemen niet het nodige prestatieniveau bereikt hebben om de screeningsprestaties in een 'real world setting' te verbeteren (25).

2.4 De mogelijke rol van artificiële intelligentie bij mammografische screening

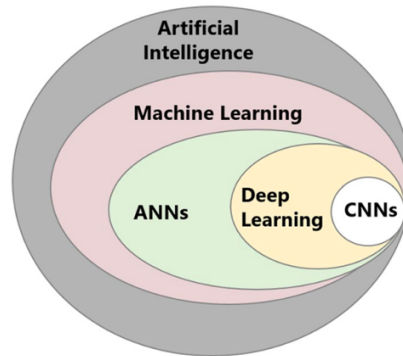
In de gezondheidszorg is medische beeldvorming – en hierbij voornamelijk mammografie – een gebied dat veel aandacht wekt voor het gebruik van kunstmatige/artificiële intelligentie (AI) op basis van machine learning en deep learning (38). AI is een snelgroeiende tak van de computerwetenschap, die reeds voor mijlpalen heeft gezorgd in de ontwikkeling van meerdere toepassingen, zowel binnen als buiten de geneeskunde. AI heeft het potentieel om de bestaande paradigma's te veranderen, ook binnen de beoordeling van mammografieën (29).

In deze literatuurstudie staat AI synoniem voor het niveau van 'Artificial Narrow Intelligence' (ANI). De niveaus 'Artificial General Intelligence' (AGI) en 'Artificial Super Intelligence' (ASI) worden hier niet behandeld. Hoewel er meerdere definities voor AI bestaan, zou men samenvattend kunnen stellen dat AI omschreven kan worden als een vakgebied van de computertechnologie waarin algoritmen en systemen gecreëerd worden om taken uit te voeren die normaal gezien humane intelligentie vereisen (39).

ANI blinkt uit in het herkennen van patronen in grote datasets (40). Het is echter belangrijk hierbij te vermelden dat het algoritme slechts één enkele nauwkeurig gedefinieerde specifieke taak kan oplossen (40, 41). In tegenstelling tot AGI en ASI bevat ANI geen menselijke cognitieve capaciteiten, dergelijke AGI- en ASI-algoritmen zijn op dit moment nog hypothetisch (40, 41). ANI is dus een overkoepelende term voor elk computerprogramma dat één enkele

taak uitvoert die gebaseerd is op analyse van ingevoerde data en niet op basis van een expliciete codering, zoals meestal het geval is bij niet op AI-gebaseerde software (39, 40). AI maakt dus met minimale menselijke interventie een beslissing (42).

AI omvat machine learning (ML) en deep learning (DL) (39, 40, 42, 43). Figuur 1 geeft deze onderverdeling op een duidelijke wijze weer (23: p.2).



Figuur 1: Beschrijvend diagram van de relatie tussen de verschillende technieken op het gebied van AI.

De trainingsmethodieken in ML kunnen onderverdeeld worden in verschillende types (42), namelijk: 'supervised learning', 'unsupervised learning', 'semi-supervised learning', 'reinforcement learning', 'active learning', 'meta learning', ... Om het overzicht te bewaren zullen enkel de twee belangrijkste soorten kort besproken worden, namelijk het 'supervised learning' en het 'unsupervised learning' (44).

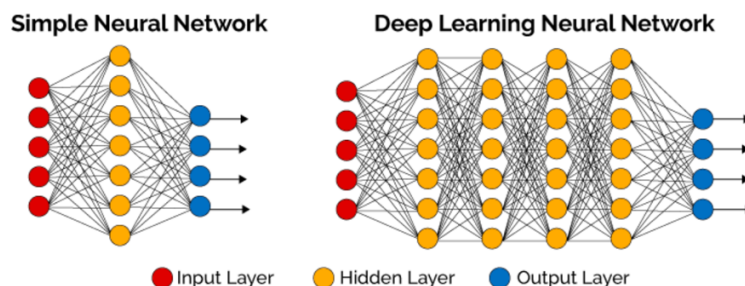
Bij 'supervised learning' bevatten de data die het algoritme nodig heeft om het mathematisch model te construeren zowel de invoer als de gewenste uitvoer ('the ground truth') (42). In deze literatuurstudie zou de input de mammografieën bevatten, de uitvoer is dan bijvoorbeeld het biopsieresultaat (benigne of maligne).

Het basisprincipe van 'supervised learning' is als volgt: machines kunnen leren van een datapool die met zijn relevante kenmerken werd gelabeld door experts of zogenaamde 'annotators' (39, 40, 42). Deze kenmerken zijn specifiek voor eenzelfde soort data input, bijvoorbeeld mammografieën (40). Deze datapool met gelabelde kenmerken noemen we de 'training data set' (43). De labeling is ook gebaseerd op een 'ground truth', bijvoorbeeld de biopsieresultaten van afwijkingen in de borsten zoals die op mammografie zichtbaar zijn (43). In de trainingsfase identificeren deze algoritmen bepaalde patronen in de beelden, gebaseerd op gedetailleerde beeldanalyse (42, 44). Uit de 'training data set' leert de machine dus welke

combinatie van beeldkenmerken geassocieerd wordt met een bepaald onderdeel van het beeld, bijvoorbeeld een borstkanker (40). De bekomen nauwkeurigheid van het getrainde algoritme is ook afhankelijk van de grootte, kwaliteit en representativiteit van de training dataset (43). Met behulp van deze training dataset kan het algoritme ook op nieuwe data uitkomsten maken (39, 42, 44).

Bij 'unsupervised learning' bevatten de data die het algoritme nodig heeft enkel de invoer, de gewenste labels zijn niet aanwezig (42). Het algoritme ontvangt tijdens het leren geen antwoord (44). Het leerproces bestaat uit het groeperen van vergelijkbare zaken voor een bepaalde dataset (44).

Tegenwoordig gaat het als men over AI spreekt vaak over DL (44). DL is een deelverzameling van ML (39, 42). Het woord 'deep' in DL slaat op de algoritmen die uit verschillende tussenlagen bestaan, namelijk een invoerlaag, een uitvoerlaag en verschillende verborgen lagen (39, 42, 44). Het verschil tussen een 'deep learning neural network' en een 'simple neural network' bevindt zich in het aantal verborgen lagen (39). Doordat het 'deep learning neural network' meerdere lagen bevat – in tegenstelling tot het 'simple neural network' dat maar één laag bevat – is het in staat om meer complexe patroonherkenning uit te voeren (39). Dit is overzichtelijk weergegeven in figuur 2 (42: p.4913).



Figuur 2: Verschil tussen een 'simple neural network' en een 'deep learning neural network'.

Deze 'meerlagige algoritmen', ook bekend als 'artificial neural networks' (ANNs), simuleren de neurale netwerken in menselijke hersenen (39, 43). Onze hersenen interpreteren nieuwe informatie door deze te vergelijken met reeds opgeslagen, bekende informatie. Op deze manier worden items gelabeld en toegewezen aan verschillende categorieën (39). DL maakt gebruik van dezelfde methode om een logische structuur te geven aan informatie (39, 43). Een uitdaging hierbij is echter de aanzienlijke behoefte aan computerkracht en trainingsdata, wat leidt tot hoge kosten van hardware en software (39).

Er bestaan verschillende types DL, enkel het convolutional neural network (CNN) zal kort worden besproken vanwege zijn relevantie. CNN wordt voornamelijk gebruikt voor beeldclassificatie (8, 44). Omwille van hun grote efficiëntie bij het automatisch capteren van representatieve beeldkenmerken ontvangen deze CNN-modellen veel aandacht (4, 17). De identificatie van de data, of de patroonherkenning in de beelden, vindt plaats in de verborgen lagen (39). De verborgen laag in een CNN bevat een 'convolutional layer' en een 'pooling layer'. Deze lagen zorgen voor een vermindering/vereenvoudiging van de beeldinformatie, maar met behoud van de karakteristieken van het invoerbeeld (44). Dit alles wordt mogelijk gemaakt door het proces waarbij de 'convolutional layer' – wat het beeld analyseert – de omliggende karakteristieken reduceert in de 'pooling layer' (44). Dit 'gereduceerd beeld' wordt gebruikt om een nieuw beeld waarmee het algoritme gevoed wordt te herkennen of te classificeren (44).

In de mammografische beeldvorming, en met name vooral in de mammografie en de tomosynthese, werd de afgelopen jaren een snelle ontwikkeling van op DL gebaseerde AI-toepassingen waargenomen. Deze algoritmen overtreffen de prestaties en klinische waarde van de conventionele CAD-systemen (15, 45). In tegenstelling tot de conventionele CAD-systemen van de voorgaande decennia, die opereren op basis van door mensen ontworpen 'programmed-in features', opereren DL-modellen door het gebruik van zelf aangeleerde beeldkenmerken (1, 29, 38). De bestaande positieve en negatieve ervaringen met de conventionele CAD-systemen zorgen ervoor dat de op AI-gebaseerde nieuwe systemen sneller te vertalen zijn naar producten en toepassingen binnen de patiëntenzorg, zoals borstkankerscreening (38).

Voor het gebruik van AI in het borstkankerscreeningsproces zijn er verschillende mogelijkheden, echter is er momenteel nog een gebrek aan voldoende onderzoek met betrekking tot deze screeningstoepassingen (16, 28).

Mammografische screening heeft verschillende tekortkomingen (38). In onderstaande secties zullen de mammografische tekortkomingen worden aangehaald en zal ook worden beschreven welke waarde AI zou kunnen bieden bij dit onderdeel van het screeningsproces.

2.4.1 Nauwkeurigheid van mammografische screening

Hoewel de vroegtijdige detectie van borstkanker door middel van grootschalig bevolkingsonderzoek voor een reductie van de mortaliteit kan zorgen (33), wordt in de

wetenschappelijke literatuur de efficiëntie van deze screeningsprogramma's betwist (29, 33, 45). Er zijn immers verschillende nadelen verbonden aan mammografische borstkankerscreening (12), waarvan de meest gerapporteerde het hoge risico op vals-positieve (3, 12, 29, 34, 36-38) en vals-negatieve testresultaten zijn (3, 12, 29, 37, 38).

Vals-positieve onderzoeken kunnen leiden tot onnodige heroproepen van de gescreende patiënten, met daaraan gekoppelde onnodige procedures (19, 33, 46). Voorbeelden hiervan zijn bijkomende beeldvormingsonderzoeken (3, 47) met dus stralingsblootstelling (3, 45) en biopsieën (3, 19, 29, 36, 45), waaraan ook mogelijke complicaties verbonden zijn. Soms geven deze vals-positieve bevindingen aanleiding tot onnodige medische behandelingen (19, 34, 45). Deze zorgen eveneens voor een toename van de werklast in de ziekenhuizen (15, 17, 21). Vervolgens hebben deze vals-positieve testresultaten ook een negatieve impact op de beleving door de patiënt (45, 46): ze zorgen voor angst (3, 17-19, 21, 29, 32, 33, 47) en mogelijk voor terughoudend gedrag om opnieuw deel te nemen aan screeningsonderzoeken (32, 33). Dit alles gaat gepaard met een toename van de kosten (3, 15, 17-19, 38, 46) en dus een reductie van de kosten-effectiviteit bij borstkankerscreening (33).

Vals-negatieve resultaten kunnen leiden tot valse geruststelling en een vertraagde diagnose van kanker (3, 33, 46). Oorzaken van deze vals-negatieve resultaten zijn onder andere de lage prevalentie van borstkanker in een screeningspopulatie waardoor zichtbare kankers over het hoofd worden gezien (21, 33), menselijke interpretatiefouten, verkeerde positionering en dens borstweefsel (29).

Uit deze nadelen kan geconcludeerd worden dat er nood is aan methoden en technieken die de sensitiviteit en de specificiteit van de mammografische screening kunnen verhogen (29). ML- en DL-technieken wekken door hun snelle vooruitgang veel interesse om de nauwkeurigheid van mammografische screening te verhogen (37).

2.4.2 Intervalkanker en de waarde van AI

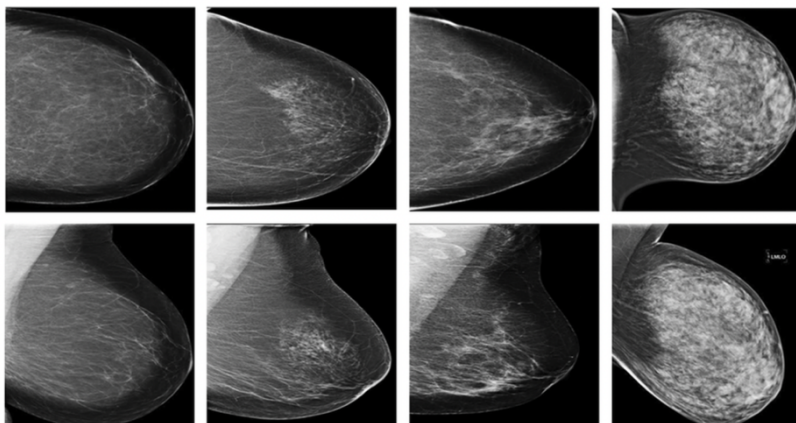
Intervalkankers (ICs) worden gedefinieerd als kankers die worden gediagnosticeerd in het interval tussen twee screeningsrondes, na een negatief/normaal screeningsonderzoek (48, 49). Deze kankers worden doorgaans gediagnosticeerd door de aanvang van klinische symptomen (49). De prognose van ICs is ongunstiger dan deze van bij screening ontdekte kankers (16, 48, 49). De ICs worden meestal ontdekt in een vergevorderd stadium (26, 48) en een groter aandeel ervan is HER2-positief (human epidermal growth factor receptor 2 overexpressie) en 'triple-negatief' (zonder receptoren voor oestrogeen, progesteron en HER2) (48).

Het intervalkankercijfer is een belangrijke indicator voor de kwaliteit en de doeltreffendheid van de mammografische screening (48, 49). Het intervalkankercijfer bevindt zich bij tweejaarlijkse screening tussen de 0.8 en 3.0/1000 gescreeende vrouwen (29, 48). In de Verenigde Staten zijn ongeveer 13% van de gediagnosticeerde borstkankers ICs (11, 26) en in Europa zijn dit er ongeveer 28% (28). Sommige van deze kankers zijn retrospectief te zien op eerdere mammografische screeningsopnames: van alle ICs wordt 20-25% geclassificeerd als 'minimale tekenen' (subtiële afwijking die al dan niet voldoende is om tot heroproep te leiden) en een even groot percentage wordt geclassificeerd als vals-negatief (16, 48, 49). 50% van de ICs zijn niet op voorafgaande mammografieën te zien en worden dus als echt negatief gecategoriseerd (48, 49).

Vanwege het maskerend effect van het fibroglandulair weefsel bij radiologisch dense borsten houdt een verhoogde borstdensiteit een verhoogd risico in voor intervalborstkanker (1, 11, 15, 16, 26). Dense borsten zorgen namelijk voor een daling van de sensitiviteit van de mammografische screening (2, 11, 12, 15, 16, 26, 48). Het is van belang om het aantal gemiste maar detecteerbare kankers en het aantal interpretatiefouten te minimaliseren (22). Zo werd aangetoond dat de mammografische detectiegraad van kanker 80% bedraagt bij vrouwen met overwegend vethoudend borstweefsel in de borst en slechts 30% bedraagt bij vrouwen met zeer dense borsten (26).

Het risico om borstkanker te ontwikkelen ligt eveneens 3 tot 5 maal hoger bij vrouwen met de hoogste mammografische borstdensiteit in vergelijking met vrouwen met de laagste borstdensiteit (2).

Het Breast Imaging-Reporting and Data System (BI-RADS) is de huidige classificatiestandaard die gebruikt wordt bij de beoordeling van mammografische beeldvorming (50). Het scoresysteem bevat 4 categorieën van borstdensiteit (A-D), namelijk 'fatty' (voornamelijk vetweefsel), 'scattered density' (verspreide dichtheid), 'heterogeneously dense' (heterogeen dens) en 'extremely dense' (extreem dens) (50). Dit is weergegeven in figuur 3 (50: p.388).



Figuur 3: Voorbeelden van de mediolaterale schuine (rij 1) en craniocaudale (rij 2) digitale mammografieën, ter illustratie van de vier kwalitatieve BI-RADS categorieën van borstdensiteit, d.w.z. vnl. vetweefsel (kolom 1), verspreide dichtheid (kolom 2), heterogeen dens (kolom 3) en extreem dens (kolom 4).

Het BI-RADS systeem wordt ook gebruikt om de kans op maligniteit van een afwijking in de borsten uit te drukken. Afwijkingen kunnen in 6 verschillende categorieën ondergebracht worden (51).

De introductie van AI en de recente ontwikkelingen van op AI gebaseerde CAD-systemen zouden de mogelijkheid kunnen bieden tot verminderen van het aantal gemiste kankers bij mammografische screening en aldus leiden tot een verbetering van de sensitiviteit (48, 49).

Het bekomen van een meer diepgaand inzicht in de klinische en histopathologische kenmerken van borstkanker middels AI zou kunnen leiden tot een verhoogde detectiegraad en zou dus van belang kunnen zijn (11, 49).

Als laatste zou een van AI gebruikmakend triagemodel om vrouwen op te sporen die een hoog risico lopen om met onopgemerkte kanker de screeningsfaciliteiten te verlaten nuttig kunnen zijn (28).

2.4.3 Impact van AI op de werklust en het prestatievermogen van radiologen

Mammografische screening heeft meerdere voordelen, maar is tegelijk niet eenvoudig uit te voeren (45). Hiervoor zijn er verschillende redenen.

Bij borstkankerscreening worden er een groot aantal mammografische onderzoeken uitgevoerd, wat zorgt voor een verhoogde werklust voor radiologen (14, 22). Ten eerste is borstkankerscreening dus arbeidsintensief (6, 24). Bevolkingsonderzoek zorgt er namelijk voor dat er sprake is van een hoog aantal mammografieën die volledig normaal zijn (13, 15, 21, 28, 30, 52). Dit wil zeggen dat een groot aantal vrouwen moet gescreend worden om in die groep één borstkanker te detecteren (13). Bijkomende moeilijkheid is dat er globaal een tekort is aan radiologen met gekwalificeerde expertise in de interpretatie van borstbeeldvorming (5, 21, 22, 24, 33, 38, 45). Het lezen van een groot aantal mammografische onderzoeken is bovendien tijdrovend (24) en vrij kostenintensief (6, 24).

Een andere uitdaging van de mammografische screening is de variatie in accuratesse van radiologen betreffende de analyse van onderzoeken (1, 7, 38, 46, 47). Niet alleen de nodige opleiding en ervaring, maar ook het volume aan mammografische onderzoeken dat beoordeeld moet worden beïnvloedt de performantie (17). In de VS werd aangetoond dat de sensitiviteit en de specificiteit van radiologen varieerden tussen resp. 66.7%-98.6% en 71.2%-96.9% (46).

Implementatie van AI toepassingen zou een oplossing kunnen zijn voor deze problematiek (5). Er is veel interesse in het gebruik van AI ter vervanging of ter aanvulling van het werk van de menselijke lezers (9). Gedeeltelijke automatisatie van de beoordeling van mammografische onderzoeken, met dus een vermindering aan menselijke lezers (19, 32) tot gevolg, zou verscheidene voordelen met zich meebrengen (30). AI zou het niet alleen mogelijk maken te beantwoorden aan de toenemende nood aan mammografische screening (5), maar ook tijd, inspanningen en kosten kunnen besparen (6, 24).

AI zou kunnen gebruikt worden als een hulpmiddel ter verhoging van de accuratesse van de radioloog bij het lezen van mammografische beelden (5, 22). Men onderzoekt deze mogelijkheid door de diagnostische prestaties van radiologen met en zonder AI-ondersteuning te gaan analyseren (5, 22).

AI zou ook kunnen gebruikt worden als autonome lezer (13). Er is namelijk nood aan een triagemodel dat met grote zekerheid volledig normale mammografische onderzoeken identificeert en filtert, waarbij een beoordeling door de radioloog overbodig wordt (28, 33, 52). Zo zou de radioloog zich kunnen focussen op beelden die door AI gecategoriseerd worden als niet-normaal – m.a.w. een rol als tweede lezer – waarbij de nood tot heroproep kan geëvalueerd worden (13). Dat zou kunnen leiden tot een verbetering van de efficiëntie en specificiteit van de radioloog zonder te raken aan de sensitiviteit (46). Dit model vereist echter wel een extreem hoge nauwkeurigheid van het algoritme; vals-negatieve uitkomsten moeten absoluut vermeden worden (19, 52).

2.4.4 Personaliseren van borstkankerscreeningsstrategieën

Er is een groeiende interesse in zogenaamde gepersonaliseerde screeningsprogramma's (2, 23). Uitgaand van zowel de klinische data van een patiënt, als van data uit mammografieën, zouden het screeningsinterval en het type van onderzoek meer afgestemd kunnen worden op het individuele borstkankerrisico van de betreffende persoon (2). Deze klinische data kunnen bestaan uit een combinatie (indien beschikbaar) van demografische en genetische gegevens (23). Aldus blijft er een grote mogelijkheid tot meer gepersonaliseerde zorg (53). AI-toepassingen die de individuele risicovoorspelling kunnen verbeteren zijn momenteel nog in ontwikkeling (45).

2.4.5 Voorkeuren van radiologen, eerstelijnszorgverleners en vrouwen met betrekking tot het gebruik van AI bij borstkankerscreening

Het is uiteraard van cruciaal belang om het vertrouwen van de belanghebbenden te winnen om de beloofde voordelen van de toepassing van AI bij borstkankerscreening te kunnen

bereiken (34). Deze belanghebbenden zijn onder andere de radiologen, eerstelijnszorgverleners en de vrouwen die het screeningsonderzoek ondergaan (31, 34, 54).

De integratie van AI in de klinische praktijk staat voor meerdere uitdagingen (34). Eén van de bezorgdheden is dat AI-algoritmen uitkomsten kunnen produceren zonder dat te achterhalen is hoe de algoritmen deze resultaten hebben bekomen (34). Niet alleen artsen, maar ook patiënten hebben nood aan transparantie in het bekomen van een diagnose, vooral met als doel het nodige vertrouwen te kunnen genereren (31, 34). Er is dus een groeiende bezorgdheid onder clinici dat het lezen van mammografische beelden door AI-algoritmen onvoldoende gericht is op patiënten (31).

Tevens is het nog onduidelijk welke technologieën geaccepteerd zullen worden door praktiserende radiologen (54). De integratie van AI in het mammografische screeningsproces zal afhangen van meerdere factoren die ook in de bovenstaande secties werden besproken, zoals het bestaan van voldoende evidentie betreffende de diagnostische accuratesse van AI-toepassingen, en de impact ervan op de workflow (54). Gegevens die de houding van radiologen ten opzichte van AI-tools weergeven zijn echter momenteel nog vrij schaars (54).

Er is dus nood aan meer inzicht omtrent de aanvaardbaarheid van AI indien deze zou worden toegepast voor borstkankerscreening (31).

2.4.6 Bias en generaliseerbaarheid van AI

AI-algoritmen die getraind werden met Amerikaanse patiëntendata zouden volgens recent onderzoek op een onevenwichtig grote blanke patiëntenpopulatie gebaseerd zijn (2). Er is stijgende bezorgdheid omtrent de veralgemeende toepasbaarheid of generaliseerbaarheid van deze algoritmen, aangezien ze mogelijk niet altijd representatief zijn voor de bevolkingsgroep waarop ze toegepast zullen worden (2). Er is noodzaak tot verdere validatie van AI-modellen met betrekking tot hun generaliseerbaarheid en het al dan niet aanwezig zijn van een bias (2). Aangezien borstkankerscreening doorgaans op grote schaal plaatsvindt, kan dit de toepasbaarheid van AI voor screening ernstig beïnvloeden.

2.4.7 Doelstelling van dit literatuuronderzoek

Dit literatuuronderzoek – van het type gestructureerde narrative review – heeft als doelstelling te achterhalen wat de waarde van AI bij borstkankerscreening met mammografie precies kan betekenen. Uit de hieruit voortkomende resultaten zou kunnen afgeleid worden of AI al dan niet een oplossing kan bieden voor de aan screening gerelateerde problematiek, zoals die in de inleiding beschreven werd.

Er zal worden nagegaan of AI een meerwaarde biedt bij het verhogen van de nauwkeurigheid van borstkankerdetectie, bij de detectie van intervalekanker, bij het verminderen van de werklast, bij het verhogen van het prestatievermogen van radiologen en bij het personaliseren van borstkankerscreeningsstrategieën.

Tot slot zal worden nagegaan wat de voorkeuren van radiologen, eerstelijnszorgverleners en vrouwen zijn omtrent de implementatie van AI en hoe generaliseerbaar AI-oplossingen zijn bij borstkankerscreening. Al deze factoren kunnen immers leiden tot een mogelijke optimalisatie van het mammografische screeningsproces, dat talrijke voordelen biedt zowel voor het individu als voor de maatschappij.

3. METHODOLOGIE

Voor deze literatuurstudie werd beroep gedaan op verschillende databanken. Geschikte publicaties om de onderzoeksvraag te beantwoorden werden gezocht in de onlinedatabases: 'Google Scholar', 'PubMed' en 'Embase'.

De PRISMA-flow chart te vinden in het addendum biedt een duidelijk overzicht van de zoekstrategie (figuur 4 in bijlage).

Initieel werd een deel van de onderzoeksvraag, namelijk 'Value of AI in breast cancer screening on mammography' ingegeven in de zoekmachine 'Google Scholar', om zo een globaal idee te krijgen van het onderwerp en van verdere geschikte trefwoorden. Van de op deze wijze gevonden publicatie 'Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System' (22) werd vervolgens gebruik gemaakt om geschikte trefwoorden te vinden in de databanken 'PubMed' en 'Embase'. Dit door deze publicatie in beide databases in te geven en de gekoppelde termen te onderzoeken. Er werd voor deze publicatie gekozen vanwege zijn relevante inhoud en zijn vele citaties, daarom werd ze ook opgenomen in de literatuurstudie.

Om het overzicht te behouden en optimale publicaties te vinden, werd de zoekopdracht uitgesplitst voor de verschillende databanken. De uiteindelijke trefwoorden die werden ingegeven via de 'Search Builder' en 'Query Builder' in resp. 'Pubmed' en 'Embase' zijn in onderstaande tabel weergegeven. Met deze zoekstrategie werden er in totaal 64 publicaties verkregen.

	PubMed	Embase
AI bij borstkanker- screening met mammografie	"Artificial Intelligence" [Majr] AND ("Breast Neoplasms/diagnostic imaging" [Majr] OR "Breast Neoplasms/diagnosis" [Majr]) AND "Mammography" [Majr] AND ("Early Detection of Cancer" [Mesh] OR "Mass Screening" [Mesh])	'Artificial intelligence'/exp/mj AND 'Breast cancer'/exp/mj AND 'Mammography'/exp/mj AND ('Mass screening'/exp OR 'Early cancer diagnosis'/exp) (mj = major) (exp = explore)

Verder werden er ook nog 5 publicaties uit de database 'Google Scholar' geïnccludeerd die een algemeen beeld van de werking van artificiële intelligentie schetsen. Hiervoor werden volgende zoektermen ingegeven: 'artificial intelligence machine learning and deep learning' en 'artificial narrow intelligence'. Ook werden er nog 6 andere relevante publicaties geïnccludeerd, deze werden aangereikt door de promotoren of geïnccludeerd op basis van de 'sneeuwbalmethode'.

Vervolgens werd er een tijds kader afgebakend, de focus in deze literatuurstudie zal namelijk liggen op publicaties vanaf 2017 tot heden. Deze tijdsafbakening werd gekozen omdat er sinds 2017 een aanzienlijke stijging is van het aantal publicaties betreffende dit onderwerp (55, 56). Eveneens wordt zo een recente kijk op het onderwerp verkregen. Tevens zal er af en toe een minder recente, maar even relevante publicatie opgenomen worden in deze literatuurstudie. Een tijdsafbakening van de ene op de andere dag is namelijk niet wenselijk.

Publicaties waarvan er geen 'free full text' te vinden was via de databank van het Kenniscentrum Gezondheidszorg Gent werden geëxcludeerd.

Er werd ook bewust gekozen om editorials, nieuwsartikelen en voorbereidende studies niet op te nemen in deze literatuurstudie. Dit omwille van lage betrouwbaarheid door subjectiviteit, twijfel aan deskundigheid en een gebrek aan diepgang en complexiteit (te beknopt). Tevens werden publicaties die geen betrekking hebben op borstkankerscreening door middel van mammografie uitgesloten. Het beschouwen van andere beeldvormende technieken zoals echografie, MRI en tomosynthese zou het onmogelijk maken om de resultaten van de verschillende studies met elkaar te vergelijken.

Het occasioneel gebruik van 'Word' en 'ChatGPT' ter controle van de spelling en grammatica in deze literatuurstudie wordt hier tevens vermeld om volledige transparantie te waarborgen.

4. RESULTATEN

In deze resultatensectie komen de verschillende onderwerpen aangehaald in de inleiding opnieuw aan bod met hieromtrent telkens wat duiding en cijfermateriaal.

Van de 33 volgende artikelen zijn 32 uitgevoerd met een retrospectieve opzet en één met een prospectieve opzet.

Een prospectieve studie volgt patiënten longitudinaal op in de tijd, d.w.z. van het heden naar de toekomst toe. Zo worden gegevens verzameld naarmate hun kenmerken of omstandigheden veranderen. Daarnaast vond de geïnccludeerde prospectieve studie plaats in een echte borstkankerscreeningsomgeving, het onderzoek vond aldus plaats in een klinische setting. In tegenstelling hiermee maakt een retrospectieve studie gebruik van reeds beschikbare patiëntengegevens die oorspronkelijk niet werden verzameld met het oog op het specifieke onderzoek. Bij de start van een retrospectief onderzoek hebben de onderzoekers al kennis van de uitkomst voor elke patiënt bij aanvang van het project. Dit duidt op een onderzoeksopzet in een meer gecontroleerde omgeving.

4.1 Impact van AI op de nauwkeurigheid, de werklust en het prestatievermogen van radiologen

Er zijn diverse manieren waarop AI de nauwkeurigheid, de werklust en het prestatievermogen van radiologen beïnvloedt. Zo kan AI worden ingezet ter ondersteuning van radiologen, als een autonome lezer of als triagemodel.

4.1.1 AI-systeem ter ondersteuning van radiologen

Volgende 9 artikelen onderzochten de invloed van een samenwerking tussen AI en radiologen op de nauwkeurigheid en aldus het prestatievermogen van radiologen bij de diagnose van borstkanker op mammografie. Enkele studies beoordeelden ook de invloed van deze samenwerking op de interpretatietijd. Een overzicht is terug te vinden in tabel 1 in bijlage.

Aldus omvatte het retrospectieve onderzoek van Bao et al. (5) 643 mammografieën, waarbij 62.5% Chinese vrouwen kanker vertoonden, gelezen door 71 radiologen met gemiddeld 5 jaar ervaring. De mammografieën en radiologen werden willekeurig verdeeld over twee groepen: één groep van 35 radiologen met hulp van AI-ondersteuning (Yizhun AI v.3.2.3) en één van 36 radiologen zonder. In elke groep werd 75% beoordeeld d.m.v. dubbele lezing door twee

onafhankelijke radiologen, terwijl de overige 25% door een enkele radioloog werd beoordeeld. Interactief klikken op laesies in een systeemgegenereerde lijst onthulde zowel hun omtrek als hun diagnostische details. Het onderzoek hield bij hoe vaak de radiologen het advies van AI accepteerden en hoe vaak ze het negeerden. De resultaten van deze studie werden onder andere weergegeven aan de hand van 'de area under the receiver operating characteristic curve' (AUC), waarbij een hogere AUC duidt op betere prestaties. Met een significante verhoging van de AUC van 0.84 naar 0.91 ($p < 0.01$) en de sensitiviteit van 84.77% naar 95.07% ($p < 0.01$) onderstreepte dit het belang van AI-ondersteuning voor prestatieverbetering van radiologen. Met AI-ondersteuning was er geen significante verandering in specificiteit, maar wel een algemene daling. Verder bedroeg de gemiddelde interpretatietijd 215 seconden (s) en 106 seconden (s) voor resp. de groep zonder en met AI-ondersteuning, dit duidde opnieuw op een significant verschil ($p < 0.01$).

Een gelijkaardig retrospectief onderzoek dat ook gebruik maakt van een interactieve beslissingsondersteunende modus is dat van Rodriguez-Ruiz et al. (22). Hier werden screeningsmammografieën van 240 Amerikaanse en Europese vrouwen geïnccludeerd, bij 100 onder hen was borstkanker aanwezig en bij 40 andere was er sprake van vals-positieve heroproepen. Deze werden beoordeeld door 14 radiologen met gemiddeld 9.5 jaar ervaring, één keer met en één keer zonder AI-ondersteuning (Transpara v.1.3.0 ScreenPoint Medical). Door interactief te klikken op een borstgebied gaf het AI-systeem een verdenkingscore weer, gaande van 1-100. Bovendien werd een 'Transpara score' van 1-10 voor het volledige onderzoek weergegeven. Met AI-ondersteuning steeg de AUC significant van 0.87 naar 0.89, net zoals de sensitiviteit die significant toenam van 83% naar 86%. Deze resultaten zijn gelijkaardig, maar minder uitgesproken dan in het onderzoek van Bao et al. (5) waarbij het AI-ondersteuningssysteem het beter doet. De specificiteit verbeterde met 2% dankzij AI-ondersteuning, maar dit was niet statistisch significant. Interpretatietijd zonder en met AI-ondersteuning waren gelijkaardig en bedroegen resp. 146 s en 149 s. Dit in tegenstelling tot de resultaten van Bao et al. (5). Verder werd een onderscheid gemaakt tussen onderzoeken met een laag en hoog vermoeden van borstkanker. Bij een laag vermoeden daalde de interpretatietijd met 11%, terwijl bij een hoog vermoeden de interpretatietijd steeg met 2%.

Dezelfde opzet werd gehanteerd voor het retrospectieve onderzoek van Lee et al. (24). Er werden in totaal 200 externe validatiegevallen van Aziatische vrouwen opgenomen, waarvan er 100 kanker hadden. De mammografieën werden met en zonder AI-ondersteuning (Lunit INSIGHT MMG, v.1.1.1.0) beoordeeld door 10 radiologen, waaronder vijf borstspecialisten (BS) en vijf algemene radiologen (AR). Dit gebeurde in twee sessies met twee maanden tussentijd. Het AI-systeem gaf een abnormaliteitsscore (0-100) gebaseerd op vier

mammografische opnames. Bij een score van 10 of hoger werd ook een warmtekaart op de locatie van het abnormale gebied getoond. Met AI-ondersteuning vertoonden zowel BS als AR een significante toename in de gemiddelde AUC, resp. van 0.813 naar 0.884 voor BS en van 0.684 naar 0.833 voor AR. Ook de sensitiviteit steeg significant met AI-ondersteuning, van 74.6% voor BS en 52.1% voor AR naar resp. 88.6% en 79.4%. Beide resultaten zijn vergelijkbaar met deze van bovenstaande onderzoeken (5, 22). Verder daalde de specificiteit niet significant voor BS (66.6% naar 66.4%) en AR (70.8% naar 70%) met AI-ondersteuning. Een significante invloed op de interpretatietijd werd wel waargenomen, deze daalde voor BS (van 82.73 s naar 73.04 s) en steeg voor AR (van 35.44 s naar 42.52 s).

In het retrospectieve onderzoek van Dang et al. (45) werden eveneens de AUC-waarden en de interpretatietijd kort besproken voor beoordeling met en zonder AI-ondersteuning (Mammoscreen v.1.2). Een interval van 4 weken tussen beide beoordelingen werd gehanteerd. In totaal werden 314 Franse mammografische onderzoeken van de linker- en rechterborst beoordeeld door 12 radiologen, bestaande uit acht radiologen met meer dan drie jaar ervaring en vier radiologen met maximaal één jaar ervaring. De AI-interface geeft als output een verdenkingsscore gaande van goedaardig tot verdacht (1-10) en de positie van een gedecteerd laesie. Opnieuw steeg de prestatie van radiologen significant met AI-ondersteuning met een stijging van de AUC van 0.739 naar 0.773. De sensitiviteit steeg niet significant (66% naar 70%), evenmin als de specificiteit (79% naar 81%). De interpretatietijd vertoonde tevens geen significant verschil tussen de leescondities, nl. 106 s en 102 s resp. zonder en met AI-ondersteuning.

Vervolgens werd in het retrospectieve onderzoek van Kizildag Yirgin et al. (29) hetzelfde AI-systeem als in Lee et al. (24) gebruikt, namelijk 'Lunit INSIGHT MMG', v.1.1.1.0. Dit onderzoek bevatte de beoordeling door twee radiologen met meer dan vijf jaar ervaring van 211 mammografieën uit een Turks screeningsprogramma. Hiervan vertoonden er 110 mammografieën borstkanker: 74 door screening opgespoord, 27 ICs en 9 gemiste kankers. Vervolgens werd een hypothetische samenwerking tussen de radiologen en het AI-systeem geëvalueerd waarbij deze laatste dienst deed als een derde lezer. De resultaten werden uitgedrukt in het percentage gedecteerd kankers. Door de radiologen werden 67.3% van alle kankergevallen geïdentificeerd (74 van de 110 kankers), terwijl de hypothetische samenwerking met het AI-systeem resulteerde in een identificatiepercentage van 83.6% van de kankergevallen (92 van de 110 kankers, bestaande uit 74 door screening opgespoorde kankers, 12 ICs en 6 gemiste kankers).

Daarentegen werd een eigen AI-algoritme gebaseerd op CNNs met ResNet-34 als basisarchitectuur, d.m.v. de analyse van 170 230 mammografieën, ontwikkeld en gevalideerd in het retrospectieve onderzoek van Kim et al. (1). Deze opnames werden verzameld uit vijf verschillende instituten in Zuid-Korea, de Verenigde Staten en het Verenigd Koninkrijk. Vervolgens verzamelden twee Zuid-Koreaanse instituten 320 mammografieën, inclusief 160 kankerpositieve gevallen, als test dataset. Net zoals in het onderzoek van Lee et al. (24) werden deze opeenvolgend zonder en met AI-ondersteuning beoordeeld door BS en AR (zeven van elk). Een significante toename in de AUC van 0.810 zonder naar 0.881 met AI-ondersteuning werd waargenomen. Met AI-ondersteuning presteerden AR beduidend beter, waardoor hun prestaties evenwaardig werden aan die van BS. Hun AUC steeg nl. van 0.772 naar 0.869. In alle groepen radiologen was met een toename in de AUC van 0.083 de grootste verbetering met AI-ondersteuning te zien bij dense borsten.

Ook Wu et al. (36) trainden en evalueerden net zoals Kim et al. (1) een eigen AI-algoritme gebaseerd op CNNs, maar met een ResNet-22 basisarchitectuur. Dit retrospectieve onderzoek verzamelde gegevens over een hybride AI-radioloogmodel. Een tweestapsarchitectuur maakt gelijktijdig gebruik mogelijk van een patch-level netwerk dat leert van labels op pixelniveau en een netwerk dat leert van labels op macroscopisch borstniveau. Deze aanpak stelde het AI-systeem in staat zowel de aanwezigheid van kanker in de borst te voorspellen als de locatie van verdachte bevindingen te markeren op een warmtekaart. De dataset omvatte 1440 borsten (720 onderzoeken), waarvan er 62 als kwaadaardig werden geclassificeerd. Deze werden gelezen door 12 radiologen, een arts-assistent en een geneeskundestudent. De AUC voor de detectie van kanker bij de humane lezers bedroeg gemiddeld 0.778 terwijl deze voor het hybride model 0.891 bedroeg.

In het retrospectieve onderzoek van Schaffter et al. (19) vond dan weer een internationale uitdaging plaats waarbij 126 teams deelnamen om een AI-algoritme voor mammografieën te ontwikkelen ('DREAM Challenge'). Het uiteindelijke gecombineerde AI-model werd samengesteld door de top 8 teams. Eveneens werd een geïntegreerd AI-model met beoordelingen van radiologen geproduceerd. Voor de evaluatie van het geïntegreerde model werd zowel een Amerikaanse als een Zweedse dataset gebruikt met resp. gegevens van 25 657 en 166 578 vrouwen, waarbij beide een kankerpercentage van 1.1% hadden. Voor het geïntegreerde AI-model werden AUC-waarden van 0.942 bereikt voor zowel de Amerikaanse als de Zweedse dataset. De significant verbeterde specificiteit voor deze Amerikaanse en Zweedse datasets bedroeg resp. 92% en 98.5%, in vergelijking met die van radiologen alleen (resp. 90.5% en 96.7%) bij een vastgezette sensitiviteit (resp. 85.9% en 77.1%).

Ten slotte werd er recent een Zweeds prospectief onderzoek genaamd ScreenTrustCAD uitgevoerd door Dembrower et al. (57). In deze studie werden 55 581 vrouwen geïnccludeerd, 269 hiervan werden gediagnosticeerd met kanker door een initiële positieve lezing. Het hoofddoel van deze studie was om de niet-inferioriteit te beoordelen van een dubbele lezing door een radioloog in combinatie met een AI-systeem (Lunit v.1.1.6), in vergelijking met de dubbele lezing door twee radiologen. De resultaten toonden resp. 261 en 250 gedetecteerde gevallen van kanker, met een relatieve detectieverhouding van 1.04 voor het combinatiemodel van een radioloog plus het AI-systeem ten opzichte van de dubbele lezing door twee radiologen. Hiermee werd niet-inferioriteit van de eerste aanpak bevestigd (95% CI 1.00-1.09; $p < 0.0001$). Aangezien de ondergrens van het 95% betrouwbaarheidsinterval (BI) meer dan 1 bedroeg, beschouwde men de aanpak van een radioloog in combinatie met een AI-systeem superieur aan de huidige standaard van zorg. Bovendien werden twee andere benaderingen onderzocht, namelijk autonome lezing door het AI-systeem en drievoudige lezing door twee radiologen in combinatie met het AI-systeem. Respectievelijk werden 246 en 269 gevallen van kanker gedetecteerd, wat overeenkomt met een relatieve detectieverhouding van resp. 0.98 en 1.08 ten opzichte van de dubbele lezing. Autonome lezing door het AI-systeem bleek niet superieur, maar wel niet-inferieur aan de dubbele beoordeling door twee radiologen. In tegenstelling hiermee was de drievoudige beoordeling zowel niet-inferieur als superieur.

Deze resultaten worden samengevat weergegeven in tabel 2 in bijlage.

4.1.2 AI-systeem als autonome lezer

In een aantal van bovenstaande studies werd ook de AUC berekend voor het AI-systeem als autonome lezer. Voor de onderzoeken van Rodriguez-Ruiz et al. (22), Lee et al. (24), Kizildag Yirgin et al. (29), Kim et al. (1), Wu et al. (36) en Schaffter et al. (19) (in een Amerikaanse en Zweedse dataset) bedroeg deze resp. 0.887, 0.915, 0.853, 0.959, 0.876, 0.858 (VS) en 0.903 (Zweden). De interesse in een AI-systeem als autonome lezer is namelijk groot. Volgende 3 artikelen onderzochten dan ook meer specifiek het gebruik ervan bij het AI-algoritme 'Transpara, ScreenPoint Medical'.

De werking van 'Transpara' is als volgt: een score van 1 tot 100 geeft voor elke verdachte regio het niveau van verdenking op kanker weer. Hoe hoger de score, hoe verdachter de laesie. Verder wordt aan elk onderzoek een score van 1 tot 10 toegekend. Hoe hoger de score, hoe hoger de waarschijnlijkheid dat er kanker aanwezig is in het onderzoek.

Zo onderzochten Rodriguez-Ruiz et al. niet alleen een samenwerking tussen een AI-systeem en radiologen (22), ze vergeleken ook de prestatie van een AI-systeem (Transpara v.1.4.0) als autonome lezer met deze van 101 radiologen bij de detectie van borstkanker op mammografie (6). Dit retrospectieve onderzoek maakte gebruik van negen datasets van verschillende instellingen in Europa en de Verenigde Staten met in totaal 2652 onderzoeken, waarvan er 653 maligne waren. Naast de mammografieën waren ook bijbehorende referentiegegevens beschikbaar. Namelijk de BI-RADS-score (1-5) en de kans op kwaadaardigheid (1-100) toegekend door de radioloog na een enkele lezing, evenals de 'ground truth' voor de aanwezigheid of afwezigheid van kanker. Uit de AUC voor het autonome AI-systeem en de radiologen, die resp. 0.840 en 0.814 bedroeg, werd niet-inferioriteit bevestigd. Het verschil was namelijk 0.026, wat de vereiste van 0 overschreed en het 95% BI varieerde van -0.003 tot 0.055, waarbij de ondergrens boven de vereiste -0.05 lag. Het AI-systeem had verder een hogere AUC dan 61.4% van de radiologen en een sensitiviteit hoger dan 57.9% van de radiologen. Echter was de AI-prestatie altijd lager dan deze van de beste radioloog.

Het retrospectieve onderzoek van Sasaki et al. (35) bestudeerde de mogelijkheid van een eerdere versie van hetzelfde AI-systeem (Transpara v.1.3.0) als autonome lezer. Drie lezers lazen de dataset, die 310 Japanse vrouwen omvatte en waarvan er 69 patiënten kanker hadden. Opnieuw werd voor elke borst een BI-RADS score (0-5) toegekend. In tegenstelling tot het onderzoek van Rodriguez-Ruiz et al. (6), presteerde het autonome AI-systeem significant minder goed dan de humane lezers bij de detectie van borstkanker ($p < 0.001$), met respectievelijke scores voor de AUC van 0.706 en 0.816. De waargenomen sensitiviteit en specificiteit waren resp. 89% en 86% voor de humane lezers, 93% en 45% voor het AI-systeem bij een 'Transpara cutoff score' van 4 en 85% en 67% voor het AI-systeem bij een 'Transpara cutoff score' van 7.

Een recentere versie van 'Transpara' (v.1.7.0) werd dan weer toegepast in het retrospectieve onderzoek van Romero-Martin et al. (13). De data afkomstig uit de 'Tomosynthesis Córdoba Screening Trial' omvatten 15 999 vrouwen, waarvan 113 met kanker, en werden geïnterpreteerd door vijf borstradiologen. Het autonome AI-systeem behaalde een AUC van 0.93, wat een betere detectieprestatie is dan dat van de radiologen. Desondanks kon het AI-systeem 15 van de 113 kankers niet correct lokaliseren. Het autonome AI-systeem behaalde verder een niet-inferieure sensitiviteit van 62.8% en 70.8% vergeleken met resp. enkelvoudige (58.4%) en dubbele (67.3%) menselijke lezing, met lagere heroproeppercentages van resp. -1.4% en -2%. Het gebruik van dit autonoom AI-systeem met niet-inferieure sensitiviteit zou leiden tot een 39% reductie in het aantal vrouwen dat voor vervolgonderzoek opgeroepen moet worden vergeleken met dubbele menselijke beoordeling.

In tabellen 3 en 4 in bijlage worden resp. de karakteristieken en resultaten van bovenstaande onderzoeken gestructureerd weergegeven.

4.1.3 AI-systeem als triagemodel

Werklast, efficiëntie en invloed op de nauwkeurigheid van radiologen zijn andere belangrijke factoren waarbij een AI-systeem mogelijk van nut kan zijn. Onderstaande 7 onderzoeken gingen na in welke mate een AI-triagemodel met diverse werkstromen kan functioneren.

De mogelijkheid tot een daling van het aantal te lezen mammografieën door radiologen werd in een retrospectieve simulatiestudie van Lauritzen et al. (21) geëvalueerd. Het AI-systeem 'Transpara' (v.1.7.0) kende scores van 0 tot 10 toe aan een populatie van 114 421 Deense vrouwen. Binnen elke score viel 10% van deze populatie, onder andere bestaande uit 791 screeningsgedetecteerde kankers, 327 ICs, 1437 kankers gedetecteerd tussen 2 en 5 jaar na screening en 2107 vals-positieve resultaten. De mammografieën werden op basis van twee drempelwaarden in drie categorieën verdeeld. Radiologen lazen geen normale onderzoeken (score < 5) noch verdachte onderzoeken met een score hoger dan de vastgestelde drempelwaarde voor heroproep (score > 9.989). Deze laatste drempelwaarde werd ingesteld om het aantal gemiste borstkankergevallen gelijk te stellen aan het aantal ICs die anders heropgeroepen werden voor aanvullend onderzoek. De overige onderzoeken werden als matig risico beschouwd en door twee radiologen geïnterpreteerd. Radiologen zonder AI-ondersteuning bereikten voor alle mammografieën een sensitiviteit van 70.8% en een specificiteit van 98.1%, terwijl het AI-triagemodel een niet-inferieure sensitiviteit van 69.7% en een significant hogere specificiteit van 98.6% ($p < 0.001$) behaalde. 1.5% van de oorspronkelijk door de radiologen gedetecteerde kankers werd door het triagemodel gemist en als normaal beschouwd, maar 25.1% vals-positieve resultaten konden vermeden worden. Verder werd met behulp van het triagemodel de werklast voor radiologen met 62,6% verminderd. Van de 114 421 mammografieën, werden er nl. 71 585 uitsluitend door het AI-systeem beoordeeld.

Dezelfde opzet was terug te vinden in het retrospectieve onderzoek van Raya-Povedano et al. (15). Daar werd de 'Transpara'-drempelwaarde (v.1.6.0) op een AI-score van 7 gelegd. Onderzoeken met een score onder deze waarde werden als 'normaal' beschouwd en werden niet door een radioloog beoordeeld. De 2% meest verdachte onderzoeken werden eveneens niet door de radioloog bekeken, maar automatisch heropgeroepen. De overige onderzoeken werden op dezelfde manier beoordeeld als in de oorspronkelijke situatie, door enkele of dubbele lezing. De data voor deze studie waren afkomstig van de 'Córdoba Tomosynthesis

Screening Trial' en omvatten 15 987 mammografieën, inclusief 113 kankers (98 op screening gedetecteerd en 15 ICs). Dankzij een afname van het aantal beoordelingen door de radiologen verminderde het AI-triagemodel de werklast met 71,5%, waarbij van de 31 974 lezingen slechts 9100 nog door de radiologen werden uitgevoerd. Eveneens werd door het AI-triagemodel een significante daling van 16.9% bereikt in het aantal heroproepen in vergelijking met dubbele lezing. De sensitiviteit van het AI-triagemodel (78/113) was niet-inferieur aan dat van de radiologen bij dubbele lezing (76/113), met een relatief verschil van 2.63%.

Een gelijkaardig onderzoek met een eerdere versie van 'Transpara' (v.1.4.0) werd uitgevoerd in de retrospectieve studie van Lång et al. (32). In een cohort van 9581 Zweedse vrouwen, inclusief 68 screeningsgedetecteerde kankers en 187 vals-positieve resultaten, werden door het AI-systeem toegekende scores van 1-5 en van 6-10 beschouwd als resp. lage en hoge risicoscores. Bij verschillende AI-risicodrempels bestudeerde men de impact van het uitsluiten van beelden aan radiologische lezing. De incidentie van kanker bedroeg 0.14% en 1.36% voor mammografieën met resp. lage en hoge risicoscores. Wanneer mammografieën met scores 1 en 2 niet werden beoordeeld door de radioloog, werden 1829 onderzoeken inclusief 10 vals-positieve resultaten, maar geen kankers, uitgesloten. Als alle mammografieën met een lage risicoscore niet werden beoordeeld, zijn dat er 5082, inclusief 57 vals-positieve resultaten en 7 gemiste maar duidelijk zichtbare kankers.

Met eenzelfde versie van 'Transpara' (v.1.4.0) voerden ook Rodriguez-Ruiz et al. (33) een retrospectief onderzoek uit. Dezelfde dataset als in hun eerdere studie, bestaande uit 2652 onderzoeken (6), werd wederom geanalyseerd. Onderzoeken met een laag risico op kanker werden als 'normaal' beschouwd en niet door de radioloog beoordeeld, alleen onderzoeken met een hoog risico op kanker werden geëvalueerd. De werklastvermindering werd bepaald aan de hand van het verschil in het aantal onderzoeken dat door de screeningsradiologen werd beoordeeld. Een AI-drempelwaarde van 2 zou voor radiologen resulteren in een werklastvermindering van 17%. Bij deze score werd 1% echt positieve kankers als normaal beschouwd. Bij een AI-drempelwaarde van 5 was de werklastvermindering dan weer 47%, maar werden 7% echt-positieve kankers gemist. Behalve bij een AI-drempelwaarde van 9, had preselectie van mammografieën voor lezing geen invloed op de prestaties van de radiologen in vergelijking met lezing in de originele populatie. Met een AUC-verschil van minder dan 1% wijst dit aldus op niet-inferioriteit.

Het onderzoek van Dembrower et al. (28) maakte gebruik van hetzelfde principe, maar van een ander AI-algoritme (Lunit v.5.5.0). De studiedataset bestond uit 7364 Zweedse vrouwen, waaronder 347 vrouwen met op screening gedetecteerde borstkanker. Het AI-systeem

fungeerde als de enige lezer, waardoor radiologen het merendeel van de normale mammografieën niet hoefden te beoordelen. 'Lunit' geeft een score tussen 0 en 1 aan ieder onderzoek, waarbij een score van 1 duidt op de hoogste graad van verdachtheid. Wanneer de categorie mammografieën met de 60% laagste AI-scores aan humane lezing onttrokken werd, werden er geen kankers gemist. Dit was wel het geval voor het uitsluiten van mammografieën met scores in de laagste 70%, 80% en 90% met resp. 1 (0.3%), 9 (2.6%) en 14 (4.0%) gemiste kankers.

Verder ontwikkelde men in het retrospectieve onderzoek van Yala et al. (46), met een dataset bestaande uit 7176 vrouwen, een eigen DL-model met ResNet-18 als basisarchitectuur. Vervolgens werden kankervrije mammografieën, vergelijkbaar met bovenstaande onderzoeken, door het AI-triagemodel in een werkstroom zonder radiologen ingedeeld. De radiologen bereikten een sensitiviteit en specificiteit van resp. 90.6% en 93.5% bij het onafhankelijk lezen van alle mammografieën in de dataset. Wanneer radiologen echter alleen de 80.7% van de mammografieën beoordeelden die niet door het AI-triagemodel werden voorgesorteerd als kankervrij, werd een sensitiviteit van 90.1% en een specificiteit van 94.2% bereikt. Sensitiviteit was aldus niet-inferieur met een marge van 5%, maar de stijging in specificiteit was wel statistisch significant.

Ook in het retrospectieve onderzoek van Kyono et al. (52) werd een eigen AI-algoritme gebaseerd op een CNN in combinatie met 'multi-task learning' ontwikkeld. Nogmaals werd onderzocht of het AI-algoritme in staat was om negatieve mammografieën voor te selecteren en zo het aantal onderzoeken dat door de radioloog moest worden gelezen, te verminderen. Het prestatievermogen van het AI-algoritme werd getest op een dataset van 2000 vrouwen uit het Verenigd Koninkrijk. Bij een negatief voorspellende waarde van 99% en bij 15% kankergevallen in de dataset, kon het AI-algoritme 34% van de 'normale' mammografieën uitsluiten aan radiologische beoordeling. Als het aantal vrouwen met borstkanker in de dataset echter werd verlaagd tot 1%, kon het AI-algoritme 91% uitsluiten.

In tabellen 5 en 6 in bijlage worden resp. de karakteristieken en resultaten van bovenstaande onderzoeken gestructureerd weergegeven.

4.1.4 Detectie, lokalisatie en classificatie van laesies d.m.v. deep learning

Er zijn verschillende raamwerken beschikbaar voor het detecteren, lokaliseren en classificeren van borstlaesies op mammografie. In deze sectie wordt hier kort verder op ingegaan.

In de retrospectieve studie van Ribli et al. (7) werd het detectieraamwerk 'Faster R-CNN' onderzocht. Het basis CNN is een VGG16 netwerk en het systeem bevat aanvullende componenten die zijn ontworpen voor het detecteren, lokaliseren en classificeren van objecten in een afbeelding. Zo werd voor de detecterende en lokaliserende functie een tak van CNNs ingezet, genaamd het 'Region Proposal Network'. Zonder enige menselijke tussenkomst genereert het systeem een aanduiding voor elke gedetecteerde laesie, bestaande uit een omlijning en een score die aangeeft hoe zeker de classificatie van de laesie als goedaardig of kwaadaardig is. Het model werd getraind met resp. 2620 en 847 gedigitaliseerde mammografieën uit de 'Digital Database for Screening Mammography' (DDSM) en een dataset van het departement radiologie aan de Semmelweis Universiteit in Budapest, Hongarije. De laesies op de mammografieën uit de DDSM waren voorzien van annotaties op pixelniveau als 'ground truth'. De test dataset 'INbreast' voor de evaluatie van systeemclassificatie bestond dan weer uit 115 mammografieën met eveneens annotaties op pixelniveau en histologisch bewijs voor kanker. Het systeem behaalde in deze dataset, met een AUC van 0.95, een 'state-of-the-art' prestatie. Met een AUC van 0.85 behaalde dezelfde aanpak de tweede plaats in de bovenstaande 'DREAM Challenge' van Schaffter et al. (19). De 'INbreast' test dataset werd ook ingezet om het detectie- en lokaliseringsvermogen van het systeem te beoordelen. Met behulp van een vrije-respons Receiver Operating Characteristic (ROC)-curve werd het gedeelte correct gelokaliseerde laesies, ook bekend als sensitiviteit, in relatie tot het aantal vals-positieve markerings op de afbeelding getoond. Het systeem slaagde erin maligne laesies te detecteren met een sensitiviteit van 90% bij 0.3 vals-positieve markerings per afbeelding. Echter werden alle laesies (100%) correct gedetecteerd bij het verhogen van het toegelaten aantal vals-positieve markerings van 0.3 naar 3.

Frazer et al. (47) onderzochten eveneens retrospectief de borstkankerdetectiemogelijkheden op mammografie van diverse op DL gebaseerde AI-algoritmen, waarbij zij de verschillen in aanpak en architectuur van de gebruikte algoritmen bestudeerden. Zo werden een lokale, een globale en een globale + lokale aanpak bestudeerd. Verder werd in de lokale aanpak ResNet50 gebruikt als basisarchitectuur, terwijl in de globale aanpak zowel Inception-ResNet-V2, EfficientNetB6 en NASNetLarge als basisarchitectuur bestudeerd werden. De globale + lokale aanpak omvatte zowel een handmatig als automatisch model, gebaseerd op respectievelijk VGG16, ResNet50 of een VGG16_ResNet50_hybride alsmede ResNet22 en ResNet18 als basisarchitecturen. De dataset 'BreastScreen Victoria' werd gebruikt, bestaande uit 28 694 mammografieën. Voor het trainen en het testen van de modellen werden resp. 8049 (3919 met kanker) en 930 (434 met kanker) mammografieën uit deze dataset ingezet. Voor de lokale benadering werden in totaal 8949 beeldpatches gegenereerd, hiervan werden er 547 (waarvan 379 met kanker) gereserveerd als test dataset.

Met een AUC van 0.8979 en een nauwkeurigheid van 0.8178 werd het beste resultaat bereikt. Deze resultaten waren gebaseerd op een dataset van 349 testgevallen en 930 testbeelden. Het gebruik van de familie van ResNet-modellen als basisarchitectuur van het AI-algoritme speelde hierbij een cruciale rol. De methode integreert globale eigenschappen van de gehele mammografie met lokale eigenschappen van automatisch geïdentificeerde kankergebieden en niet-kankergebieden in het volledige beeld. Bovendien worden technieken zoals het bijsnijden van de achtergrond, tekstverwijdering, contrastaanpassing en een uitgebreide set trainingsgegevens toegepast.

Verder retrospectief onderzoek naar het classificatievermogen van een op DL gebaseerd AI-algoritme werd uitgevoerd door Aboutalib et al. (17). De hypothese betrof het vermogen van AI om heropgeroepen, maar benigne mammografieën te onderscheiden van negatieve en maligne beelden. Voor het trainen en evalueren van het op CNNs (een aangepaste versie van AlexNet) gebaseerde AI-algoritme, werden twee onafhankelijke datasets gebruikt. Uit de 'Full-field Digital Mammography dataset' (FFDM) en de DDSM werden respectievelijk 5212 en 9648 mammografieën verzameld. Deze datasets werden in verschillende configuraties gebruikt om het AI-model te trainen en te testen. Hierbij werd 95% van de data gebruikt voor het trainen van het model, terwijl de overige 5% als test dataset fungeerde. Belangrijk om te vermelden is dat het AI-model bij alle volgende experimenten vooraf werd getraind op de 'ImageNet'-dataset en werd geoptimaliseerd met de FFDM dataset.

Bij het exclusief gebruik van de FFDM dataset voor het trainen en testen van het AI-model werd, met een AUC variërende tussen 0.70 en 0.81, een duidelijk onderscheid waargenomen tussen bovenstaande drie categorieën van mammografieën. In het geval van exclusief gebruik van de DDSM dataset, varieerde de AUC echter tussen 0.77 en 0.96. Wanneer vervolgens beide datasets werden gecombineerd, strekte de AUC zich uit van 0.76 tot 0.91. Bij al deze bovenstaande datasets waren de geïdentificeerde beeldkenmerken tussen heropgeroepen, maar benigne beelden en negatieve beelden het meest onderscheidend met een resp. AUC van 0.81, 0.96 en 0.91.

In het laatste experiment werden verbeteringen in de AUC waargenomen, variërend van 0.02 tot 0.05. Hierbij werd het op 'ImageNet' voorafgetrainde AI-model verder voorgetraind op de DDSM dataset en geoptimaliseerd met de FFDM dataset. De testen werden uitgevoerd op de FFDM dataset. De AUC steeg in vergelijking met enkel vooraftraining op de niet-medische dataset 'ImageNet'. De grootste toename in prestatie werd waargenomen in de onderscheiding van heropgeroepen, maar benigne beelden met maligne beelden.

4.2 Intervalkanker en de waarde van AI

Om een duidelijk overzicht te bewaren en de resultaten optimaal te interpreteren worden de belangrijkste karakteristieken van de bestudeerde onderzoeken schematisch weergegeven in tabel 7 in bijlage.

Het retrospectieve onderzoek van Lång et al. (48), met een studiepopulatie die enkel intervalkankers (ICs) bevat, onderzocht of AI het aantal ICs bij mammografische screening kan reduceren. Deze studie maakt gebruik van het commercieel beschikbare algoritme 'Transpara v.1.5.0'. Hierbij krijgen regionale bevindingen een abnormaliteitsscore toegekend van 1-100 en krijgt het gehele onderzoek een maligniteitsscore van 1-10 toegekend op basis van de meest verdachte bevinding. De regionale bevindingen worden aangeduid op de mammografieën wanneer ze zich boven een bepaalde vooraf door de gebruiker vastgestelde drempel bevinden (CAD-markering). Ze worden gedetecteerd door een combinatie van waarnemingen omtrent zachte weefsels en verkalkingen. De mammografieën van de ICs (429 in totaal) werden eveneens beoordeeld door twee radiologen en werden zo ingedeeld in 3 categorieën: echt negatief (60.6%), minimale tekenen (26.3%) of vals-negatief (13.1%).

Het AI-systeem categoriseerde 33.3% (143/429) van de screeningsmammografieën met een risicoscore van 10. Minimale tekenen en vals-negatieve ICs maakten hiervan 67.1% (96/143) uit. De mediane risicoscores bedroegen 6.7, 9.0 en 9.7 voor resp. echt negatieve, minimale tekenen en vals-negatieve ICs. Een significante correlatie werd waargenomen tussen de risicoscores en bovenstaande 3 categorieën van IC ($p < 0.0001$). Er bestond eveneens een significant verschil in risicoscore wanneer echt negatieve ICs vergeleken worden met minimale tekenen en vals-negatieve ICs, maar niet wanneer minimale tekenen en vals-negatieve ICs met elkaar vergeleken worden ($p < 0.0001$). Vervolgens werd er door gebruik van het AI-systeem een potentiële reductie van ICs in screeningssetting van 19.3% (83/429) waargenomen bij een drempel voor heroproep die op risicoscore 9.1 ligt. De drempel voor heroproep bij deze risicoscore werd geleverd door het AI-systeem en komt overeen met een heroproeppercentage van ongeveer 10% van de bevolking. Bij een drempel voor heroproep van 4% (risicoscore 9.67) en 1% (risicoscore 9.92) was de reductie in IC resp. 11.2% (48/429) en 4.7% (20/429). Deze ICs waren ofwel minimale tekenen of vals-negatief en werden eveneens correct gelokaliseerd. (48)

Een gelijkaardig retrospectief onderzoek werd uitgevoerd door Byng et al. (49) met behulp van het commercieel beschikbare AI-systeem 'Vara 1.0.7'. Het doel van dit onderzoek was

aantonen dat AI retrospectief zichtbare, maar gemiste en als IC gediagnosticeerde kankers kan detecteren en correct lokaliseren. Het gaat hier om vals-negatieve en minimale tekenen ICs. Ook de prestaties van AI omtrent occulte en echte ICs werden bestudeerd. Een cohort van 2396 vrouwen met ICs werd door radiologen beoordeeld en gecategoriseerd in 4 categorieën: echte (53.4%), minimale tekenen (19.5%), vals-negatieve (15.6%) of occulte (11.4%) ICs. Echte ICs zijn niet zichtbaar op de eerdere normale screeningsmammografie en zijn dus equivalent aan echt negatieve ICs. Ten slotte is een occulte IC symptomatisch, maar niet zichtbaar op screening en diagnostische mammografie.

Uit deze studie bleek dat het AI-systeem 30.2% (113/374) van de vals-negatieve ICs en 13.7% (64/468) van de ICs met minimale tekenen kon detecteren met een specificiteit van 99%. Hierbij zou ongeveer 1% van de bevolking heropgeroepen worden voor verder onderzoek. Eveneens was de accuraatheid van lokalisatie bij deze detectie hoog, met een nauwkeurigheid van resp. 91.2% en 89.1%. Wanneer bovenstaande gegevens gecombineerd worden, komt dit overeen met een potentiële vermindering van vals-negatieve en minimale tekenen ICs met resp. 27.5% (103/374) en 12.2% (57/468). Verder kon het AI-systeem 2.8% (36/1208) van de echte ICs detecteren en lokaliseren op de screeningsmammografie. Van alle retrospectief zichtbare ICs, inclusief minimale tekenen, kon het AI-systeem 21.0% (177/842) detecteren en 90.4% (160/177) hiervan correct lokaliseren met 99% specificiteit. Een voorbeeld van een door AI gedecteerde en correct gelokaliseerde vals-negatieve IC is terug te vinden in figuur 5 in bijlage. Met gedaalde specificiteit en corresponderend een toegenomen sensitiviteit, was AI in staat om grotere hoeveelheden retrospectief zichtbare ICs te detecteren. (49)

Het retrospectieve case-control onderzoek van Wanders et al. (16) maakte net zoals het onderzoek van Lång et al. (48) gebruik van 'Transpara v.1.6.0'. Echter werd Transpara gecombineerd in een neurale netwerk (NN) met borstdensiteit berekend d.m.v. vrij beschikbare software (Laboratory for Individualized Breast Radiodensity Assessment) met als doel het risico op IC in negatieve screeningsmammografieën te gaan voorspellen. De combinatie NN gaf voor elke patiënt een score tussen 0 en 1, wijzend op een toenemend risico op IC. De prestaties geleverd door deze combinatie werden vergeleken met de afzonderlijke prestaties van het AI-systeem en met een risicomodel gebaseerd op borstdensiteit. 2222 vals-negatieve mammografieën werden gematcht met 4661 echt negatieve mammografieën. Resp. 666 en 1398 hiervan werden geïnccludeerd in de test dataset (in totaal 2064 mammografieën). De overige mammografieën werden gebruikt in de training dataset.

De AUC-prestaties in het voorspellen van IC na een negatieve screeningsmammografie bedroegen voor resp. het combinatiemodel AI en borstdensiteit, het AI-systeem afzonderlijk

en een risicomodel gebaseerd op borstdensiteit afzonderlijk 0.79, 0.73 en 0.69. Het aantal geïdentificeerde screeningsmammografieën met IC tot gevolg in de test dataset met een specificiteit van 90% bedroegen 50.9% (339/666), 37.5% (250/666) en 22.4% (149/666) voor resp. het combinatiemodel AI en borstdensiteit, het AI-systeem afzonderlijk en een risicomodel gebaseerd op borstdensiteit afzonderlijk. De combinatie NN (AI en borstdensiteit) doet het significant beter in het identificeren en voorspellen van negatieve screeningsmammografieën die een IC tot gevolg hebben dan het AI-systeem en het op borstdensiteit gebaseerde model afzonderlijk. (16)

Naast het onderzoek van Wanders et al. (16) maakt ook het retrospectieve case-case onderzoek van Hinton et al. (11) gebruik van een ResNet50 model (research algoritme) en beoordeling van borstdensiteit (met BI-RADS), zowel in combinatie als afzonderlijk. Dit onderzoek had als doel te bepalen of door DL gedetecteerde mammografische kenmerken (die traditionele metingen van borstdensiteit overstijgen) kunnen worden toegepast om groepen te identificeren met risico op IC als gevolg van maskering. In totaal werden 182 ICs en 173 op screeningsmammografie gedetecteerde kankers geïnccludeerd.

De AUC-prestaties voor het voorspellen van IC versus op screeningsmammografie gedetecteerde kanker bedroegen 0.82, 0.82 en 0.65. Dit voor resp. het combinatiemodel AI en borstdensiteit, het AI-systeem afzonderlijk en het op borstdensiteit gebaseerde model afzonderlijk. Het AI-systeem (al dan niet in combinatie met borstdensiteit) presteert hier beter dan het op borstdensiteit gebaseerde model. In figuur 6 in bijlage worden de resultaten van dit onderzoek grafisch weergegeven naast de resultaten van het onderzoek van Wanders et al. (16).

75.5% (268/355) van het totale aantal mammografieën in dit onderzoek (11) werd correct gecategoriseerd door het AI-systeem op zichzelf. Er werden 73.6% (134/182) van het totale aantal ICs en 77.5% (134/173) van het totale aantal op screeningsmammografie gedetecteerde kankers correct voorspeld door het AI-systeem op zichzelf. Verder werd aan de hand van 'saliency maps' aangetoond dat lokale informatie de categorisatie sterk zou beïnvloeden, echter is er ook beïnvloeding door bredere regio's.

Het retrospectieve case-case-control onderzoek van Zhu et al. (26) bouwt verder op het onderzoek van Hinton et al. (11) (pilotstudie). Opnieuw wordt gebruik gemaakt van een research algoritme (DenseNet-121 model) en van klinische risicofactoren inclusief borstdensiteit (BI-RADS), zowel in combinatie als afzonderlijk. Het doel van dit onderzoek was nagaan of DL-modellen het risico op IC kunnen inschatten. In plaats van de AUC werd

voornamelijk de 'C-statistic' berekend, dit is een maatstaf ter voorspelling van de nauwkeurigheid. Een 'C-statistic' varieert van 0.5 (geen betere voorspelling dan toeval) tot 1 (perfecte voorspelling). De dataset van Zhu et al. (26) bevat de mammografieën van in totaal 6369 vrouwen, waarvan er 351 IC ontwikkelden.

De 'C-statistic' voor modellen die vrouwen met IC vergeleken met gematchte controles bedroeg 0.72, 0.64 en 0.71, dit voor resp. het combinatiemodel AI en klinische risicofactoren, het AI-systeem afzonderlijk en het op klinische risicofactoren met borstdensiteit gebaseerde model afzonderlijk. Eveneens werd voor elk model de AUC berekend welke vergelijkbaar of identiek bleek te zijn. (26)

Retrospectief onderzoek naar de mogelijkheid om een commercieel beschikbaar AI-systeem (Lunit v.5.5.0) in te zetten als laatste lezer na een negatief onderzoek om zo vrouwen die het hoogste risico lopen op een ongedetecteerde kanker te identificeren voor bijkomende screening, werd uitgevoerd door Dembrower et al. (28). 7364 vrouwen werden geïnccludeerd, 6817 vrouwen als gezonde controles en 547 vrouwen gediagnosticeerd met borstkanker (347 kankers gedetecteerd op screeningsmammografie en 200 klinisch gedetecteerde ICs). Het AI-systeem gaf een score gaande van 0 tot 1 als voorspeller van het risico op tumoraanwezigheid, een score van 1 is het hoogste niveau van verdenking.

Na een negatieve screeningsmammografie, met dubbele lezing door twee radiologen, werden er bij de top 1% (633 vrouwen) van de door AI toegekende scores 12% (24/200) ICs en 14% (48/347) kankers die pas in de volgende screeningsronde zichtbaar waren gedecteerd. Bij de top 5% (1445 vrouwen) van AI-scores waren dit er resp. 27% (53/200) en 35% (121/347). De potentiële detectie door AI van zowel ICs als kankers die pas in de volgende screeningsronde zichtbaar waren, bedroeg 114/1000 vrouwen binnen de top 1% AI-scores en 34/1000 vrouwen binnen de top 5% AI-scores. De AUC voor het voorspellen van IC door middel van het AI-systeem (maximale AI-score) en borstdensiteit bedroeg resp. 0.74 en 0.67. (28)

Vervolgens werd in het retrospectieve simulatie onderzoek van Kizildag Yirgin et al. (29) eveneens gebruik gemaakt van het commercieel beschikbare AI-systeem Lunit. Er werd onderzoek gedaan naar de prestaties en de doeltreffendheid van AI bij het detecteren van ICs en gemiste kankers in screeningssetting. Een risicoscore voor maligniteit tussen 1% en 100% werd door het AI-systeem toegekend. In totaal werden er 211 screeningsmammografieën bestudeerd, waarvan 74 tijdens screening gedetecteerde kankers, 27 ICs, 9 gemiste kankers en 101 negatieve controles.

De drempelwaarde voor de risicoscore bedroeg 34.5%. Hierbij bedroeg de AUC 0.853. Een sensitiviteit en specificiteit van resp. 72.8% en 88.3% voor kankerdetectie door AI werden in de ROC-analyse waargenomen. 44% van de ICs kreeg een risicoscore hoger dan 34.5%, daarnaast bevond 55.5% van de ICs zich onder deze drempelwaarde. Het AI-systeem afzonderlijk was in staat om 72.7% (80/110) van alle kankers te detecteren, 12 van deze 80 mammografiën waren ICs. Wanneer het AI-systeem hypothetisch gecombineerd werd met radiologen, werd een kankerdetectiepercentage van 83.6% waargenomen (92/110), opnieuw waren 12 van deze gedetecteerde kankers ICs. Radiologen waren in staat om 67.3% (74/110) van alle kankers te detecteren. Het AI-systeem was dus in staat om 44.4% (12/27) van de nog niet eerder door de radioloog gedetecteerde ICs te detecteren. Echter categoriseerde het 16.2% van de echt positieve mammografieën als een negatieve mammografie. (29)

In tabel 8 in bijlage worden de resultaten van bovenstaande artikelen samengevat om het overzicht te bewaren.

Verder werden er in het onderzoek van Byng et al. (49) geen statistisch significante verschillen waargenomen tussen de kankers die wel en niet door het AI-systeem gedetecteerd werden, dit met betrekking tot clinico- en histopathologische kenmerken van de retrospectief zichtbare kankers. De enige uitzondering hierop is de lagere detectie van hooggradige tumoren in vergelijking met laaggradige tumoren (OR 0.42). Daarentegen werd in het onderzoek van Wanders et al. (16) geen bewijs van een significant verschil in prestatie tussen de detectie van invasief (AUC 0.79) en ductaal carcinoma in situ (AUC 0.76) waargenomen.

Byng et al. (49) toont eveneens een verhoogde waarschijnlijkheid voor detectie door AI bij carcinomen met betrokken lymfeklieren bij diagnose, in vergelijking met lymfeklier negatieve carcinomen. De 'odds ratio' bedroeg 1.67 voor LN1 versus LN0 en bedroeg 1.16 voor LN2 versus LN0. Echter werd er bij het gecombineerde NN AI en borstdensiteit van Wanders et al. (16) geen bewijs van een verschil in prestatie bij het voorspellen van ICs met positieve (AUC 0.80) of negatieve (AUC 0.80) lymfeklieren waargenomen.

Wanders et al. (16) toonden daarenboven aan dat hun gecombineerde model AI en borstdensiteit beter presteerde bij het detecteren van ICs die gediagnosticeerd werden binnen de 12 maanden na screening (AUC 0.84), in vergelijking met ICs die gediagnosticeerd werden binnen 12-24 (AUC 0.77) of 24-36 maanden (AUC 0.70) na screening ($p < 0.001$).

Ten slotte speelt borstdensiteit duidelijk een belangrijke rol bij de diagnose van ICs. Het is dan ook nuttig om te achterhalen op welke manier radiologen de BI-RADS borstdensiteits-

categorieën toekennen. Voor de lezers kan dit met wat training leiden tot een hogere consistentie en lagere variabiliteit. Dit onbewust en kwalitatief besluitvormingsproces werd onderzocht in het retrospectieve onderzoek van Mohamed et al. (50). Om deze vraag te bestuderen maakte men gebruik van een op DL gebaseerde aanpak. Twee classificatiemodellen met elk twee uitvoerklassen werden ontwikkeld. In het eerste model gaat het om klasse B en C en in het tweede model om niet-dense borsten (klasse A+B) en dense borsten (klasse C+D). De AUC-waarden voor classificatie gebaseerd op enkel MLO- en CC-projecties bedroegen respectievelijk 0.95 en 0.88 voor model 1, en 0.97 en 0.92 voor model 2. MLO-projecties presteerden significant beter in beide modellen ($p \leq 0.001$). De AUC bedroeg resp. 0.92 voor model 1 en 0.95 voor model 2 bij combinatie van beide projecties.

4.3 Personaliseren van borstkankerscreeningsstrategieën

Het retrospectieve onderzoek van Yala et al. (53) introduceert Tempo, dit is een nieuw raamwerk gebaseerd op versterkend leren voor gepersonaliseerde screening. Versterkend leren is een vorm van ML. Zonder gedetailleerde instructies voor elke tussenstap leert het algoritme om in opeenvolgende stappen beslissingen te nemen.

Tempo werd ontwikkeld met behulp van een dataset van het 'Massachusetts General Hospital' (MGH). Voor elk onderzoek werden ook risicobeoordelingen gegenereerd door Mirai en Tyrer-Cuzick versie 8 (TCv8). Mirai en TCv8 zijn resp. een nieuw op AI-gebaseerd mammografisch risicomodel dat risico's op verschillende momenten voorspelt, en een traditioneel model dat diverse risicofactoren zoals leeftijd, familiegeschiedenis en hormonale factoren integreert.

Voor het Tempo-algoritme werd als eerste een NN getraind waarbij toekomstige risicobeoordelingen voorspeld werden op basis van eerdere beoordelingen. Dit model schatte het risico van de patiënt op ongeobserveerde momenten in en maakte een op risico gebaseerd screeningsbeleid mogelijk. Vervolgens werd op retrospectieve gegevens een screeningsbeleid om de beloning te optimaliseren getraind. Dit beleid gebruikte als input zowel de risicobeoordeling van de patiënt als hun voorkeuren voor screeningsintervallen om een suggestie te doen voor het moment van de volgende screening. De aanbeveling kon variëren van 6 maanden tot 3 jaar in de toekomst (in stappen van 6 maanden). De beloning weegt dan weer de kosten van beeldvorming af (gemiddeld aantal mammogrammen per jaar aanbevolen door het beleid) tegen het voordeel van vroege detectie. De beloning voor vroege detectie meet het tijdsverschil in maanden tussen de aanbevolen screeningsdatum van elke patiënt, als deze na hun laatste negatieve mammogram was, en de daadwerkelijke datum waarop een

diagnose gesteld werd.

Het Tempo-algoritme werd vergeleken met bestaande screeningsrichtlijnen, waaronder jaarlijkse screenings, tweejaarlijkse screenings en een hybride screeningsstrategie (gaat over van jaarlijkse naar tweejaarlijkse screening op 55-jarige leeftijd). Een positief getal bij de efficiëntie van screening duidde op een betere prestatie. Vroegtijdige detectie werd uitgedrukt in maanden, waarbij een positief cijfer vroegere detectie aanduidde. Op de MGH-testset behaalden de jaarlijkse, tweejaarlijkse screeningsrichtlijnen en de hybride screeningsstrategie een efficiëntie voor screening (voordeel van vroege detectie per screeningskost) van respectievelijk -10.34, 1.58 en -4.42. Daarentegen behaalden Tempo-Mirai en Tempo-TCv8 efficiënties van respectievelijk 4.29 en 2.16.

Met een hoger voordeel van vroege detectie per screeningskost overtrof Tempo-Mirai significant Tempo-TCv8 en jaarlijkse screening ($p < 0.001$). Aldus resulteerde Tempo-Mirai in een vroegtijdig detectievoordeel van 4.10 maanden bij het aanraden van 0.96 mammografieën per jaar. In vergelijking hiermee behaalden de jaarlijkse en tweejaarlijkse richtlijnen een vroegtijdig detectievoordeel van resp. 1.58 en -5.17 maanden bij het aanraden van resp. 1 en 0.5 mammografieën per jaar.

4.4 Voorkeuren van radiologen, eerstelijnszorgverleners en vrouwen met betrekking tot het gebruik van AI bij borstkankerscreening

Twee gelijkaardige onderzoeken in de Verenigde Staten werden uitgevoerd door Hendrix et al. om zowel de voorkeuren van radiologen (54) als eerstelijnszorgverleners (34) te achterhalen met betrekking tot het gebruik van AI-systemen bij borstkankerscreening. In totaal namen resp. 66 radiologen en 91 eerstelijnszorgverleners deel. 44 van de 66 participerende radiologen gaven aan deel genomen te hebben aan een 'fellowship' voor borstbeeldvorming.

Deze onderzoeken hadden als doel het relatieve belang van verschillende eigenschappen van op AI-gebaseerde systemen te kwantificeren voor beide studiepopulaties. Als onderzoeksmethode werden dan ook discrete keuze-experimenten (DCEs) aangewend. Bij elk thema moest een keuze worden gemaakt uit twee hypothetische AI-alternatieven of een optie voor screening door een radioloog zonder AI. De DCEs werden georganiseerd op een manier dat betere prestaties op bepaalde eigenschappen moesten ingeruild worden voor slechtere prestaties op andere.

Een derde onderzoek werd uitgevoerd door Lennox-Chhugani et al. (31). Bij dit onderzoek ligt

de focus op de houding van vrouwen ten opzichte van het gebruik van AI-systemen bij het lezen van mammografieën. 4096 Engelse vrouwen namen deel, hiervan was 42.7% op screeningsleeftijd (≥ 50 jaar) en waren 57.3% toekomstige gebruikers (<50 jaar) van het programma. Een prospectief 'mixed method' onderzoek werd toegepast om kwantitatieve en kwalitatieve gegevens te verzamelen d.m.v. een korte enquête.

De bevraagde thema's voor deze drie onderzoeken zijn weergegeven in tabel 9 in bijlage.

Het onderzoek van Hendrix et al. (54) met betrekking tot de voorkeuren van radiologen toonde met een kans van 90% een duidelijk voornemen aan tot het gebruik van een willekeurig op AI-gebaseerd systeem voor borstkankerdetectie. Echter was er sprake van grote heterogeniteit, het 95% BI strekt namelijk uit van 39% tot 99% kans. De interesse was over het algemeen het grootst voor AI-systemen met een evenwichtige sensitiviteit en specificiteit, met name 94% sensitiviteit met $<25\%$ van de onderzoeken als positief gemarkeerd. Eveneens bestond de voorkeur uit annotaties die door het AI-systeem werden weergegeven nadat de radiologen hun onafhankelijke beoordeling hadden voltooid. Verder ging de voorkeur uit naar AI-systemen zonder aanpassing van de workflow m.b.t. laag risico onderzoeken en AI-systemen waarbij dubbele beoordeling van verdachte onderzoeken werd aangemoedigd.

Vervolgens werden er 3 categorieën geïdentificeerd. 59% van de radiologen gaf aan één van de gepresenteerde AI-systemen te willen gebruiken, voor hen was sensitiviteit de belangrijkste eigenschap. 27% van de radiologen benaderde de AI-systemen enthousiast, maar veranderde van opinie wanneer er eigenschappen die voor hen als ongewenst werden ervaren verschenen. De overige 14% radiologen vertoonden een algehele weigering om de AI-systemen voor borstkankerscreening toe te passen.

Eveneens werden door Hendrix et al. (34) 3 categorieën bepaald, daarentegen met betrekking tot de voorkeuren van eerstelijnszorgverleners. Net zoals bovenstaande resultaten van Hendrix et al. bij radiologen (54) werd sensitiviteit bij de meerderheid als belangrijkste eigenschap geïdentificeerd. 41% van de zorgverleners gaf aan sensitiviteit twee keer zo belangrijk te vinden in vergelijking met andere eigenschappen van het AI-systeem. Zo wenst deze groep radiologische bevestiging van door AI als positief geclassificeerde beelden. 24% van de zorgverleners wenst bevestiging van alle bevindingen van het AI-systeem door een radioloog. De resterende 35% beschouwde de meeste eigenschappen als gelijkwaardig. Bijkomend accepteerde 76% van de zorgverleners een AI-systeem voor beoordeling van waarschijnlijk negatieve beelden zonder radioloog.

In het onderzoek van Lennox-Chhugani et al. (31) m.b.t. de voorkeur van vrouwen werden

door sentimentanalyse van vrije tekst opnieuw 3 categorieën vastgesteld. Zo had 47.2% van de vrouwen een positieve opinie over het gebruik van AI-systemen bij borstkankerscreening. 35.9% gaf gemengde of onzekere standpunten weer en 17.9% uitte een negatieve mening. Echter gaf 2.7% van de vrouwen geen antwoord op de vraag. Onder andere verhoogde efficiëntie, betrouwbaarheid en veiligheid werden als voordelen benoemd. Als nadelen benoemde men de afwezigheid van een 'human touch', mogelijke discriminatie en een gebrek aan duidelijkheid omtrent het bestuur en de privacy van AI-systemen. Daarnaast werd volgend onderscheid opgemerkt: resp. 49,5% en 45,4% van de vrouwen op screeningsleeftijd en van de toekomstige gebruikers reageerden positief op het gebruik van AI voor het lezen van mammografieën, wat duidde op een positievere houding bij vrouwen op screeningsleeftijd.

4.5 Bias en generaliseerbaarheid van AI

Het onderzoek van Gastounioti et al. (2) valideerde het AI-risicomodel 'ProFound AI® Risk 1.0' (oorspronkelijk ontwikkeld op basis van een Zweedse cohort) in een Amerikaanse screeningscohort voornamelijk bestaande uit zwarte (51%) en blanke (42%) vrouwen die mammografische screening ondergingen. Deze cohortstudie was retrospectief van aard en bestond uit 176 gevallen van borstkanker en 4963 willekeurige controles met een negatieve follow-up van minstens 1 jaar.

Het AI-risicomodel behaalde in deze cohort met een AUC van 0.68 een vergelijkbare discriminerende prestatie aan de overwegend blanke Zweedse ontwikkelingscohort met een AUC van 0.73. Verder was er geen significant verschil ($p = 0.20$) in prestatie van het risicomodel per groep met een AUC van 0.67 en 0.70 voor resp. blanke en zwarte vrouwen.

5. DISCUSSIE

5.1 Optimalisatie van radiologische beoordelingen: de impact van AI-algoritmen op de nauwkeurigheid, de werklast en het prestatievermogen van radiologen

De toegevoegde waarde van AI-algoritmen komt in eerste instantie tot uiting in de verbetering van de nauwkeurigheid, vermindering van de werklast en optimalisatie van het prestatievermogen bij de radiologische interpretatie van mammografieën. Dit wordt benadrukt door resultaten uit een breed scala aan studies en toepassingen.

5.1.1 De ondersteunende rol van AI-algoritmen voor de radioloog

De negen studies die een samenwerking tussen de radioloog en een AI-algoritme onderzochten toonden allen aan dat radiologen met ondersteuning van AI-algoritmen – al dan niet significant – betere diagnostische prestaties leveren bij het beoordelen van mammografieën voor borstkankerscreening, in vergelijking met radiologen zonder deze ondersteuning (1, 5, 19, 22, 24, 28, 29, 36, 45, 57). Dit is voornamelijk weergegeven d.m.v. de AUC.

Zo is de hoogste AUC-waarde (0.942) voor de samenwerking van de radioloog en het AI-algoritme terug te vinden in het onderzoek van Schaffter et al. (19). De overige onderzoeken behaalden gelijkaardige resultaten. Zo behaalden Bao et al. (5), Wu et al. (36), Rodriguez-Ruiz et al. (22), Lee et al. (voor BS en AR) (24) en Kim et al. (1) een respectievelijke AUC van 0.91, 0.891, 0.89, 0.884 (BS) en 0.833 (AR) en 0.881. Het onderzoek van Dang et al. (45) deed het met een AUC van 0.773 net iets slechter. In het onderzoek van Kizildag Yirgin et al. (29) werd de prestatieverbetering dan weer uitgedrukt in een stijging van het kankerdetectiepercentage met 16.3% bij samenwerking. Het is belangrijk te vermelden dat dit retrospectieve onderzoeken bedragen. Dit in tegenstelling tot mogelijk het meest waardevolle onderzoek in deze literatuurstudie, nl. het onderzoek van Dembrower et al. (57). Dit recente prospectieve onderzoek vond als enige plaats in een echte screeningssetting. Met een 4% hoger kankerdetectiepercentage werden zowel niet-inferioriteit als superioriteit aangetoond voor een dubbele lezing waarbij één radioloog door het AI-algoritme werd vervangen.

De diversiteit in AUC-waarden tussen studies vindt zijn verklaring in de verscheidenheid aan opzet en de specifieke AI-algoritmen die zijn toegepast (29).

Uiteraard zijn ook de sensitiviteit en specificiteit in deze context van belang. Er werden duidelijke fluctuaties in de uitkomsten van de verschillende onderzoeken opgemerkt. Sommige studies toonden een significante toename in sensitiviteit aan bij AI-ondersteuning (5, 22, 24), terwijl bij andere studies de stijging niet significant was of de significantie ervan niet werd berekend (29, 45, 57). In één artikel (19) vertoonde de specificiteit een significante toename

met AI-ondersteuning. In twee andere artikelen was er een niet-significante stijging van de specificiteit (22, 45), terwijl in nog twee andere een niet-significante daling werd waargenomen (5, 24). Dankzij de robuuste aard van de software kan deze consistent en nauwkeurig functioneren. Dit suggereert dat AI-ondersteuning kan bijdragen tot het verminderen van de variabiliteit in de diagnostische prestaties van radiologen (1).

Vervolgens leidt een kortere interpretatietijd tot zowel tijds- als kostenbesparing. Vier studies onderzochten de impact van de samenwerking tussen de radioloog en een AI-algoritme op deze tijd. Zo werd in het onderzoek van Bao et al. (5) een significant tijdsvoordeel van 109 s verkregen voor de samenwerking in vergelijking met radiologische lezing zonder AI-ondersteuning. Dang et al. (45) meldden een vermindering in tijd van 4 seconden bij AI-ondersteuning, terwijl Rodriguez-Ruiz et al. (22) juist een toename van 3 seconden observeerden. Deze verschillen waren niet significant en beide leescondities werden aldus als gelijkwaardig beschouwd. Opmerkelijk is dat zowel Rodriguez-Ruiz et al. (22) en Lee et al. (24) deze leestijd verder onderzochten. Bij het eerste onderzoek varieerde de interpretatietijd afhankelijk van de mate van verdenking van borstkanker, met een daling bij lage verdenkingen en een stijging bij hoge verdenkingen. Bij het tweede onderzoek werd een onderscheid gemaakt op basis van ervaring; de interpretatietijd daalde voor BS en steeg voor AR. De afname in leestijd bij BS wordt vermoedelijk veroorzaakt doordat AI de radiologen helpt bij een snellere detectie van laesies. De toename bij AR is daarentegen te verklaren in samenhang met hun verbeterde sensitiviteit, wat ertoe heeft geleid dat radiologen meer tijd besteedden aan gevallen met verdachte bevindingen (24).

De bovenstaande bevindingen laten zien dat AI-ondersteuning op zijn minst vergelijkbare prestaties levert aan radiologen bij de detectie van borstkanker op mammografieën. Daarnaast suggereren de resultaten waarin AI-ondersteuning betere prestaties laat zien voorzichtig het potentieel van AI als hulpmiddel bij het beoordelen van mammografieën, met als doel de effectiviteit van borstkankerscreening te vergroten. Er zijn echter verschillende redenen waarom het raadzaam is deze resultaten met nodige voorzichtigheid te interpreteren.

Ten eerste was de meerderheid van de studies retrospectief van aard. Hierdoor verschilden de leesomstandigheden aanzienlijk van die in een daadwerkelijke klinische setting, wat verschillende vertekeningen tot gevolg heeft. Dit werd onder andere geïllustreerd door het feit dat de meeste datasets verrijkt zijn met kankergevallen in vergelijking met de dagelijkse praktijk (1, 5, 22, 24, 29, 45). Deze retrospectieve met kanker verrijkte onderzoeken bieden op die manier slechts een inschatting van de sensitiviteit en de specificiteit. Nader onderzoek naar hoe deze gegevens zich extrapoleren in een screeningsdataset die representatief is voor de werkelijkheid is dan ook van belang. Een tweede aspect om in overweging te nemen is hoe

het AI-algoritme de nauwkeurigheid van de resultaten mogelijk kan beïnvloeden. In enkele studies hield het AI-algoritme enkel rekening met de geüploade mammografieën en werden klinische factoren zoals symptomen, persoonlijke en familiale voorgeschiedenis niet geïncorporeerd (1, 29). In de praktijk wordt daar wel rekening mee gehouden, en het weglaten ervan zou een grondige analyse kunnen hinderen.

Wereldwijd is er eveneens variatie in screeningspraktijken en heroproeppercentages, dit kan resulteren in verschillende uitkomsten van het AI-systeem, afhankelijk van de locatie en lokale beleidsmaatregelen. De datasets voor het trainen, valideren en evalueren van het algoritme zijn verder vaak te beperkt in omvang en geografische spreiding, waardoor het generaliseren van de resultaten naar een grotere en meer diverse populatie bemoeilijkt wordt. Deze datasets moeten een breed scala aan mammografische beoordelingen bevatten om een zo accuraat mogelijke nabootsing van echte screening te verkrijgen (29). Kim et al. (1) namen dit in overweging door gegevens te gebruiken van vijf verschillende instellingen voor validatie en training van het algoritme met een aanzienlijke geografische spreiding. In de onderzoeken van Schaffter et al. (19) en Dembrower et al. (57) is dan weer sprake van een zeer uitgebreide test dataset.

Ten slotte is het van belang te benadrukken dat terwijl tijdsbesparing de kosten – zoals de belangrijke financiële impact van de radiologische capaciteit – kan verminderen, men ook de hoge kost van de initiële investering in een AI-ondersteuningssysteem moet overwegen.

Toekomstige studies zijn noodzakelijk om de prestaties van radiologische AI-ondersteuning te onderzoeken in een daadwerkelijke klinische screeningsomgeving met radiologen die voldoende ervaring hebben met AI-algoritmen. Bijkomend zou het AI-algoritme klinische kenmerken in overweging kunnen nemen, voor zover deze beschikbaar zijn tijdens routinematige screening, zoals eerdere mammografieën, voorgeschiedenis, ...

Verder zou het relevant zijn de interactie tussen radiologen en de AI-ondersteuning te bestuderen. Hierbij is het interessant om te analyseren hoe radiologen mogelijk beïnvloed worden door al dan niet foutieve AI-aanduidingen, de rol van de persoonlijke voorkeuren van

KEY POINTS

- Er is een accuratere mammografische interpretatie bij samenwerking tussen AI-algoritmen en radiologen.
- Er is een aangepaste interpretatietijd bij het gebruik van AI-algoritmen met als voordeel besparingen in tijd en kosten.
- AI vermindert de variabiliteit in diagnostische prestaties van radiologen.
- Prospectieve validatieonderzoeken in een klinische screeningsomgeving zijn noodzakelijk.

radiologen in het gebruik van AI-ondersteuning en het vertrouwen van radiologen in zowel hun eigen beoordeling als in de resultaten van het AI-systeem.

5.1.2 Diagnostische mogelijkheden met een AI-systeem als autonome lezer

Een AI-systeem is mogelijk tot meer in staat dan enkel radiologische ondersteuning. Aldus werd door een aantal studies ook de mogelijkheid van een AI-systeem als autonome lezer onderzocht. Dit gebeurde opnieuw voornamelijk a.d.h.v. de AUC. Voor de reeds besproken onderzoeken van Kizildag Yirgin et al. (29), Wu et al. (36), Rodriguez-Ruiz et al. (22), Schaffter et al. (19) (in een Amerikaanse en Zweedse dataset), Lee et al. (24) en Kim et al. (1) bedroeg deze resp. 0.853, 0.876, 0.887, 0.858 (VS) en 0.903 (Zweden), 0.915 en 0.959. In een studie van Romero-Martin et al. (13), een andere studie van Rodriguez-Ruiz et al. (6) en een studie van Sasaki et al. (35) was de analyse van een AI-algoritme (bij allen 'Transpara') als autonome lezer het hoofddoel van het onderzoek. Deze onderzoeken behaalden een AUC van resp. 0.930, 0.840 en 0.706. In vrijwel alle bovenstaande onderzoeken, op één na (35), overtrof de AUC van de autonome AI-lezer die van de gemiddelde radioloog zonder de inzet van AI.

Het onderzoek van Rodriguez-Ruiz et al. (6) zette deze resultaten kracht bij door statistische niet-inferioriteit tussen beide leesmethoden aan te tonen. Echter was de AI-prestatie altijd lager dan deze van de beste radioloog. Een overweging hierbij kan zijn dat radiologen mogelijk beschikken over meer contextuele informatie dan het AI-algoritme. Romero-Martin et al. (13) toonden dan weer een niet-inferieure sensitiviteit aan van 62.8% en 70.8% voor het autonome AI-algoritme in vergelijking met resp. enkelvoudige (58.4%) en dubbele (67.3%) menselijke lezing, met lagere heroproeppercentages van resp. -1.4% en -2%. Verder behaalden Sasaki et al. (35) als enige een minder goede AUC voor het autonome AI-algoritme. Mogelijk is dit te verklaren door het feit dat het Transpara-systeem vergelijkbaar presteert met een gemiddelde individuele radioloog. Aldus moet worden opgemerkt dat dit onderzoek met een panel van 3 lezers is uitgevoerd, wat hun prestaties ten goede kan zijn gekomen.

Zowel Sasaki et al. (35) als Rodriguez-Ruiz et al. (6) komen tot de conclusie dat het geteste Transpara-systeem vergelijkbare prestaties vertoont aan de gemiddelde radioloog bij het opsporen van borstkanker met mammografie. Romero-Martin et al. (13) komen dan weer tot de conclusie dat AI de beoordelingen van radiologen in zijn geheel kan vervangen met een vergelijkbare gevoeligheid, maar een beduidend lager heroproeppercentage.

Bij deze resultaten zijn eveneens enkele kanttekeningen te maken.

Wederom zijn deze drie studies retrospectief van aard. Met als gevolg dat er sprake is van een laboratoriumeffect. Dit wordt geïllustreerd door datasets die met kanker zijn verrijkt (6, 35) en d.m.v. van mammografieën verkregen op één locatie met machines afkomstig van één

leverancier (13, 35). Echter aangezien het AI-systeem primair bedoeld is voor screeningsdoeleinden is het van uiterst belang bovenstaande onderzoeken ook uit te voeren op data equivalent aan die in een screening dataset.

Slechts één AI-algoritme (Transpara) werd als autonome lezer bestudeerd in deze drie studies. Echter heeft ook dit AI-algoritme enkele gebreken. 'Transpara' geeft namelijk geen afzonderlijke score voor de linker- en rechterborst, zoals radiologen dat wel zouden doen (35). Dit kan vragen oproepen omtrent de nauwkeurigheid ervan. Verder overtrof 'Transpara' nooit de beste radioloog in het onderzoek van Rodriguez-Ruiz et al. (6). Diens ervaring blijft echter onbekend, bijgevolg kan niet beoordeeld worden of het autonome AI-systeem het beter of slechter deed in functie van de ervaring van de radioloog. Aldus werd een betere prestatie van het autonome AI-systeem dan de gemiddelde radioloog gerapporteerd.

In de toekomst zou het dus mogelijk zijn om een volledig autonoom AI-systeem te implementeren als eerste en/of tweede lezer bij het proces van dubbele lezing. Dit kan een oplossing bieden voor het tekort aan ervaren borstradiologen, vooral in landen zonder specialisten op dit gebied. Zo kan een werklastereductie van maar liefst 50% tot zelfs 100% bereikt worden. Dit roept wel belangrijke ethische vragen op omtrent medische verantwoordelijkheid, aanvaardbaarheid, opleiding van radiologen en certificering van AI-systemen (57). Een oplossing zou een interactief beslissingsondersteunend hulpmiddel kunnen zijn, zoals hierboven beschreven, om mogelijke laesies aan te tonen (6). Desalniettemin, in landen en regio's waar de beperkte capaciteit van radiologen en de hoge kosten van screening de belangrijkste redenen zijn waarom screeningsprogramma's niet worden uitgevoerd, kan een autonome AI ethischer zijn dan het ontbreken van elke screening.

Alhoewel de resultaten bemoedigend zijn, is het essentieel om verder prospectief onderzoek uit te voeren in een echte screeningscontext om de prestaties en toepassing van AI als autonome lezer te valideren.

5.1.3 AI-triagemodellen als efficiënte voorselectiemethode voor mammografieën

Een alternatieve methode houdt in dat een AI-algoritme wordt gebruikt als een triagemodel voor het voorselecteren van normale mammografieën en die met een duidelijke verdenking van kanker, zonder dat radiologische beoordeling nodig is. Zeven onderzoeken in deze literatuurstudie hebben dit onderzocht. Drempelwaarden werden gebruikt om de mammografieën te categoriseren.

Vier studies toonden een niet-inferieure sensitiviteit aan tussen op AI-gebaseerde screening en screening door radiologen (15, 21, 33, 46). Desalniettemin werden er kankergevallen over het hoofd gezien. Hierbij is opmerkelijk dat bij een hogere drempelwaarde voor radiologische

beoordeling, waarbij meer mammogrammen als normaal werden beschouwd, er meer gevallen van kanker over het hoofd werden gezien (21, 28, 32, 33). Daarnaast toonden twee onderzoeken een significant verhoogde specificiteit aan (21, 46, 53). Dit beïnvloedt tevens het aantal vals-positieve resultaten, waarvan een afname werd opgemerkt (21, 32, 46). Deze daling kan leiden tot een vermindering van de kosten voor onnodig aanvullend onderzoek en het voorkomen van veel ongerustheid bij de patiënt (46). Vooral in regio's met hoge heroproeperpercentages zal dit een grote impact hebben (32). Het is bovendien mogelijk dat de kennis van de mammografische AI-risicobeoordeling de sensitiviteit en/of specificiteit van de radioloog bij de interpretatie van de overige mammografieën kan verbeteren of schaden. Verder onderzoek is dus noodzakelijk.

Het onttrekken van bepaalde mammografieën aan radiologische beoordeling heeft overigens een daling van de werklust tot gevolg. Deze daling wordt uitgedrukt als een percentage, berekend a.d.h.v. het verschil tussen het oorspronkelijke aantal onderzoeken beoordeeld door de screeningsradiologen en het huidige aantal. Dit percentage schommelt afhankelijk van de studie tussen de 17% en 91% (15, 21, 33, 52). Lauritzen et al. (21) behaalden een werklustvermindering van 62,5%. Gezien het grote aantal opgenomen patiënten, wordt deze studie als het meest representatief beschouwd voor de praktijk.

In conclusie is het dus mogelijk om automatisch onderzoeken vooraf te triëren met behulp van AI om de efficiëntie van de mammografische borstkankerscreening te verbeteren en de werklust aanzienlijk te verminderen. De nauwkeurigheid blijft vergelijkbaar met die van radiologen, en er is soms zelfs een toename in specificiteit waargenomen, gevolgd door een afname in het aantal vals-positieve resultaten.

Verder zorgt het voorselecteren ervoor dat er meer tijd beschikbaar is voor de radioloog om moeilijke mammografieën te beoordelen, wat kan leiden tot minder foutieve diagnoses (52). Het zorgt er ook voor dat radiologen zich kunnen richten op kritieke gebieden zoals interventionele radiologie, diagnostiek en interactie met de patiënt, wat hun algehele invloed op de patiëntenzorg versterkt (46).

Ook hier enkele bedenkingen. De werklustreductie vertegenwoordigt in feite het percentage mammografieën dat door de radiologen niet hoeft te worden beoordeeld. De studies tonen echter opvallende discrepanties in de drempelwaarde die als veilig wordt beschouwd om mammografieën te onttrekken aan lezing door de radioloog. Zo tonen studies uitgevoerd met eerdere versies van 'Transpara' een beduidend lagere reductie in de werklust (32, 33). Dit wijst op de mogelijkheid dat de voortdurende verbetering van AI-systemen deze drempelwaarde in de toekomst verder zou kunnen verhogen. Correcte voorselectie zal vermoedelijk ook worden versterkt door de inclusie van meer patiëntinformatie zoals hierboven al aangehaald. De

afname van het aantal te beoordelen screeningsmammografieën hangt vervolgens samen met de aanvaardbaarheid van een lichte afname in gevoeligheid, wat een afweging vraagt tussen werklastvermindering en het risico op het missen van kankergevallen. Aldus kan het protocol aangepast worden door de drempelwaarden te wijzigen.

Indien het gaat over tekortkomingen, zijn de zeven studies nogmaals retrospectief van aard. De datasets zijn maar weer eens met kanker verrijkt (21, 28, 33, 52) en bestaan uit mammografieën verkregen op één locatie met machines en een AI-systeem afkomstig van één leverancier (15, 21, 28, 32, 46). Dit wil zeggen dat er niet kan worden verzekerd dat de screeningsresultaten toepasbaar zullen zijn in klinieken met diverse screeningsregimes en AI-systemen. Bovendien zijn er enkele vraagstukken gerelateerd aan het AI-algoritme. Zo werd in de studies niet ontrafeld waarom het AI-systeem zichtbare kankers over het hoofd zag.

Enkele ethische en legale bekommernissen zijn ook aan de orde. Het vervangen van beide radiologen bij dubbele lezing door voorselectie van normale mammografieën is heel drastisch en roept vragen op omtrent de veiligheid van alle partijen betrokken bij screening. De gevolgen bij een gemiste kanker zijn namelijk enorm groot, daarom is verder uitgebreid onderzoek naar de gevolgen van het aanpassen van de drempelwaarden vereist. Rodriguez-Ruiz (33) stelt een alternatieve mogelijkheid voor om de automatisch gecreëerde groepen te gebruiken om gevallen die dubbele beoordeling behoeven en gevallen waar een enkele beoordeling volstaat te onderscheiden, zodat er altijd ten minste één radioloog betrokken is.

5.1.4 Vooruitgang in laesiedetectie, lokalisatie en classificatie d.m.v. verschillende deep learning raamwerken

Vervolgens is een AI-systeem slechts van waarde indien het in staat is laesies nauwkeurig te detecteren, lokaliseren en classificeren. Drie opgenomen studies verrichtten onderzoek naar welke aanpak en raamwerk dit het beste mogelijk maakten.

Zo hypotheesden Ribli et al. (7) dat een AI-algoritme voor het detecteren van laesies klinisch interessanter is dan een eenvoudige beeldclassificator. Deze laatste is namelijk niet in staat om de kanker te lokaliseren, wat noodzakelijk is voor verdere behandeling. Aldus werd het object detectieraamwerk 'Faster R-CNN' gebruikt (AUC: 0.95). Dit raamwerk lijkt zeer interessant door de behaalde hoge sensitiviteit (90% van de maligne laesies) met zeer weinig vals-positieve markeringen (0.3 per afbeelding). Het onderzoek van Frazer et al. (47) toonde aan dat de beste classificatieresultaten werden behaald door een gecombineerde aanpak van zowel globale als lokale methoden, a.d.h.v. raamwerken ResNet-22 en -18 (AUC: 0.8979). Het

is belangrijk op te merken dat dit onderzoek zich richtte op classificatie en niet op lokalisatie.

KEY POINTS

- AI als autonome lezer is potentieel mogelijk in de toekomst, maar brengt verschillende vragen omtrent medische verantwoordelijkheid en aanvaardbaarheid met zich mee.
- Een AI-triagemodel ter voorselectie van mammografieën met bepaalde eigenschappen kan de werklast van radiologen doen dalen. Ook dit roept enkele ethische en legale vragen op.
- Er bestaan verschillende raamwerken voor de detectie, lokalisatie en classificatie van laesies, elk met hun eigen voordelen en nadelen.
- Belangrijke aandacht is nodig voor het 'black-box' probleem, men is niet altijd in staat te achterhalen hoe AI-algoritmen tot interpretatie of categorisatie van bepaalde gegevens komen.

Bovendien bleek dat door achtergrondbijsnijding, tekstverwijdering en contrastaanpassing de resultaten nog verder werden verbeterd. Voorts toonden ook Aboutalib et al. (17) aan dat CNNs maligne, negatieve en heropgeroepen, maar benigne mammografieën kunnen onderscheiden (AUC: 0.70-0.96). De beste prestatie werd gezien bij het onderscheiden van negatieve en heropgeroepen, maar benigne mammografieën. Aldus is het AI-algoritme in staat het aantal vals-positieve heroproepen te doen dalen.

Ondanks deze prima resultaten, heeft de AUC weinig betekenis vanwege het feit dat het hier retrospectieve data betreft. Toekomstig onderzoek zal zich moeten richten op klinisch relevante resultaten door middel van studies in een werkelijk screeningsprogramma.

Verder wordt een belangrijk punt van kritiek benadrukt door Aboutalib et al. (17). Het onderzoek wijst op aanzienlijke bezwaren tegen het gebruik van AI-algoritmen. DL identificeert namelijk automatisch subtiele kenmerken in mammografieën die niet altijd zichtbaar zijn voor het menselijk oog. Echter is men niet altijd in staat te achterhalen wat deze geïdentificeerde kenmerken zijn en hoe het AI-algoritme intern tot een interpretatie of categorisatie komt. Men refereert hiernaar als het 'black-box' probleem.

5.2 De opmerkelijke rol van AI-algoritmen in intervalekankerdetectie

Een tweede item waarbij AI-algoritmen een aanzienlijk toegevoegde waarde kunnen bieden, is bij de opsporing van intervalekanker. De diagnose van deze vaak meer invasieve kankers, tussen twee screeningsrondes in en na een oorspronkelijk negatieve mammografie, heeft

namelijk een enorme impact op de patiënt. Op verschillende manieren werd in zeven studies onderzocht of AI-algoritmen in staat zijn tot vroegtijdige detectie van ICs. In essentie werden twee overkoepelende bevindingen geregistreerd. Zo heeft het inzetten van AI bij mammografische borstkankerscreening het potentieel om de incidentie van IC te verminderen. Dit komt doordat het algoritme in staat is om meer gevallen van kanker te detecteren en minder vals-negatieve evaluaties over het hoofd te zien (28, 29, 48, 49). Verder leiden combinatiemodellen van AI-algoritmen en borstdensiteitsmetingen, in vergelijking met het afzonderlijke gebruik van deze methoden, tot een grotere identificatie van vrouwen die mogelijk IC zullen ontwikkelen (16, 26). Hieruit kan worden geconcludeerd dat AI-algoritmen in pre-cancereuze mammografieën meer beeldinformatie kunnen detecteren dan alleen borstdensiteit voor het voorspellen van het risico op kanker (11, 26, 28), terwijl borstdensiteit net het risico verhoogt voor de maskering van tumorverschijnselen.

Het ene onderzoek verkreeg al betere resultaten dan het andere. Bij een heroproeppercentage van 1% werden in het onderzoek van Byng et al. (49) en Lång et al. (48) respectievelijk 90.4% en 4.7% van alle retrospectief zichtbare ICs door het AI-systeem correct gedetecteerd en gelokaliseerd. Bij het inzetten van AI als een triagesysteem om vrouwen te selecteren voor aanvullende screening na een negatieve dubbele lezing, kon in de top 1% van de hoogste scores, 12% intervalkankers worden geïdentificeerd (28). In het onderzoek van Kizildag Yirgin et al. werden met behulp van AI 44.4% nog niet eerder door de radioloog gedetecteerde ICs vastgesteld, echter bij een lagere specificiteit van 88.3%. Bij deze studies wordt opgemerkt dat bij een lagere drempel voor heroproep de sensitiviteit toenam met een dalende specificiteit. Bij het gebruik van modellen die AI-algoritmen combineren met borstdensiteit werden eveneens hoopgevende resultaten verkregen. Zo zouden nog meer vrouwen met IC kunnen gedetecteerd worden. Het aantal correct voorspelde mammografieën met intervalkanker bedroeg nl. 50.9% bij een ietwat lagere specificiteit van 90% in het onderzoek van Wanders et al. (16). Echter werd geen informatie over de correcte lokalisatie vermeld. Het onderzoek van Hinton et al. (11) ondersteunt vervolgens de hypothese dat AI-algoritmen meer beeldinformatie kunnen detecteren dan wat wordt gedekt door borstdensiteit. Hier kon het AI-algoritme 12% meer ICs correct classificeren dan een model gebaseerd op BI-RADS borstdensiteit. De specifieke kenmerken van deze beeldinformatie zullen moeten uitgeklaard worden door toekomstig onderzoek, aangezien het onderzoek van Zhu et al. (26) net een minder goed presterend AI-algoritme documenteert dan de classificatie gebaseerd op BI-RADS borstdensiteit. Hoe dan ook zal het een uitdaging worden om subtiele kenmerken te detecteren zonder het aantal heroproepen te verhogen.

Er zijn een paar beperkingen die vermeld moeten worden. Herhalingsgewijs zijn de zeven onderzoeken retrospectief van aard. De gebreken die hiermee gepaard gaan werden reeds vermeld met als belangrijkste tekortkoming dat de onderzochte populaties niet equivalent zijn aan een echte screeningspopulatie. Zo bevatten sommige onderzoeken enkel patiënten met IC (48, 49). Een ander gebrek is dat niet in elke studie weergegeven werd of het AI-algoritme naast het detecteren van de kanker ook in staat was deze correct te lokaliseren (16, 28). Het lokaliseren van de kanker is cruciaal voor vervolgonderzoeken zoals onder andere gerichte echografie.

Helaas werd slechts in enkele van de genoemde studies onderzocht of het AI-systeem verschillende prestaties vertoont op basis van specifieke histopathologische kenmerken van borstkanker (16, 49). Voor meer gedetailleerde informatie over de werking van het AI-systeem zal toekomstig onderzoek zich moeten richten op de histopathologische en radiologische kenmerken van ICs, zoals hun tumorbiologie en stadium. Verder onderzoek moet eveneens gebeuren naar de houding en prestaties van radiologen bij het gebruik van dergelijke AI-algoritmen. Deze kunnen namelijk beïnvloed worden door het feit dat ze op de hoogte zijn van een AI-algoritme dat mogelijk gemiste tekenen van kanker detecteert.

5.3 De potentiële bijdrage van AI-algoritmen aan gepersonaliseerde screeningsstrategieën

Ook onderzoek naar het personaliseren van screeningsstrategieën is volop aan de gang. Uit de bevindingen van Yala et al. (53) blijkt dat AI hier eventueel een bijdrage kan leveren. Men ontwikkelde Tempo, een raamwerk dat vervolgaanbevelingen voor screening uit patiëntrisicobeoordelingen kan voorspellen. Dit beleid laat ruimte voor aanpassing aan diverse screeningsvoorkeuren. Het verdient aandacht dat bij gebruik van nauwkeurigere risicomodellen (Mirai), het AI-algoritme beter zal presteren dan wanneer het beleid wordt afgestemd op minder nauwkeurige modellen (Tyrer-Cuzick). Hieruit volgde de conclusie dat een Tempo-beleid gebaseerd op Mirai risico-inschattingen, met een eerdere detectie per screeningskost, efficiënter was dan jaarlijkse en tweejaarlijkse screening. Het afstemmen van het beleid op Mirai brengt echter ook de beperkingen ervan met zich mee. Het model is enkel gevalideerd met 'Hologic' mammografieën, dus toekomstig onderzoek zal nodig zijn om het risicomodel aan te passen aan andere leveranciers. Toekomstige modellen zouden verder ook kunnen overwegen om andere screeningsmethoden zoals echografie of MRI op te nemen. Verder zou voor elke patiënt zowel de meest kosteneffectieve vroegste als laatste screeningsleeftijd kunnen berekend worden.

5.4 Belanghebbendenanalyse: AI-voorkeuren van radiologen, eerstelijnszorgverleners en vrouwen

Vanzelfsprekend vereist de implementatie van een AI-systeem in de praktijk meer dan wat hierboven al is beschreven. Het is vooreerst belangrijk dat men bereid is het AI-systeem te gebruiken. Drie onderzoeken analyseerden de voorkeuren van zowel radiologen, eerstelijnszorgverleners en vrouwen, aangezien zij de voornaamste belanghebbenden zijn.

Samengevat werd tot volgende conclusie gekomen. De meerderheid van de radiologen waren sterk geneigd op AI-gebaseerde hulpmiddelen voor borstkankerdetectie met mammografie te gebruiken, waarbij groot belang werd gehecht aan een evenwicht tussen sensitiviteit en specificiteit (54). Eerstelijnszorgverleners gaven daarentegen aan dat sensitiviteit de voornaamste eigenschap van het AI-systeem moest zijn (34). Ook vrouwen van screeningsleeftijd stonden open voor het gebruik van AI in borstkankerscreening. Toch zijn er velen die twijfelden en overtuigd zullen moeten worden van de betrouwbaarheid ervan (31). In het algemeen zal succesvolle AI-acceptatie in mammografie afstemming op de belanghebbenden vereisen.

Echter zijn er enkele struikelblokken. Ten eerste is er sprake van een groep gemakkelijk ontmoedigde radiologen. Aanpassing van AI-eigenschappen aan deze twijfelaars zal wellicht de praktijkadoptie bevorderen, dit zou gebaseerd kunnen worden op basis van eerdere CAD-ervaring (54). Ten tweede werd aangetoond dat een minderheid aan artsen terughoudend zijn in het gebruik van AI voor onbegeleide of gedeeltelijk begeleide beeldinterpretatie (34, 54). Dit gold ook voor vrouwen die menselijk contact bij de diagnose als essentieel zagen voor hun ervaring van kwalitatieve zorg (31).

Een kritische blik op deze resultaten is nodig. Het is namelijk onduidelijk in hoeverre deze steekproefgegevens generaliseerbaar zijn naar alle belanghebbenden. Verder zijn niet alle potentieel relevante kenmerken opgenomen in de vragenlijsten. Vervolgens hebben de meeste respondenten beperkte ervaring met AI-ondersteunde beslissingstools. Hun houding zal naar verwachting veranderen naarmate zij meer klinische AI-voorbeelden tegenkomen.

Toekomstig onderzoek zal zich onder andere moeten toespitsen op de voorkeuren van de betalers. Tevens verdienen educatie en informatieverstrekking over het gebruik van AI de volle aandacht.

5.5 Kritische blik: bias en generaliseerbaarheid van AI-systemen

Ten slotte onderzochten Gastounioti et al. (2) de bias en generaliseerbaarheid bij gebruik van het AI-risicomodel 'ProFound AI® Risk 1.0'. Men stelde vast dat de prestaties van het op mammografie gebaseerde AI-risicomodel in een Amerikaanse populatie vergelijkbaar waren met eerder gerapporteerde Europese resultaten. Deze prestaties waren bovendien niet significant verschillend bij het vergelijken van blanke en zwarte vrouwen. Toch is oplettendheid geboden aangezien de externe validatie gebeurde met gegevens van slechts één locatie in de VS. Eveneens werd enkel de bias en generaliseerbaarheid van dit ene AI-risicomodel bestudeerd, waarbij de beperkte omvang van de studiegroep een limiterende factor was voor het waarnemen van mogelijk significante verschillen in onderzoeksgroepen, waaronder tumorkenmerken en etniciteit.

KEY POINTS

- AI-algoritmen gebruiken mogelijk subtiële beeldinformatie die niet zichtbaar is voor het menselijk oog om vroegtijdig intervalekankers te detecteren.
- AI is in staat screeningsaanbevelingen te doen op basis van patiëntrisicobeoordelingen.
- Succesvolle implementatie van een AI-systeem in de klinische praktijk zal moeten rekening houden met de voorkeuren van de belanghebbenden (radiologen, eerstelijnszorgverleners en vrouwen).
- Aangezien de meeste onderzoeken retrospectief van aard zijn, zal verder onderzoek naar de generaliseerbaarheid in een screeningspopulatie nodig zijn.
- Verder onderzoek naar bias van AI-algoritmen zal moeten worden verricht.

6. CONCLUSIE

Concluderend kan men stellen dat AI-algoritmen wel degelijk van waarde zijn bij mammografische borstkankerscreening. Ze bieden ondersteuning aan de radiologen met als gevolg een verbeterde nauwkeurigheid, een optimalisatie van de prestaties, een verminderde werklast en een aangepaste interpretatietijd. Verder zouden ze kunnen dienen als triagemodel of onafhankelijke lezer van mammografieën, al brengt dit enkele ethische en juridische vragen met zich mee. Vervolgens zijn AI-algoritmen van nut bij de detectie van intervalekankers, aangezien deze mogelijk subtiële afwijkingen detecteren die over het hoofd worden gezien door de menselijke waarneming. Verder is het personaliseren van borstkankerscreeningsmodaliteiten een ander thema waarbij AI-algoritmen een positieve

bijdrage kunnen leveren voor de patiënten en zelfs een kostendaling kunnen veroorzaken. Toch is er één belangrijke kanttekening: op één na alle onderzoeken geïnccludeerd in deze literatuurstudie waren retrospectief van aard, d.w.z. dat de resultaten niet kunnen gegeneraliseerd worden naar een echte screeningspopulatie. Prospectief onderzoek in een werkelijke klinische screeningsomgeving is aldus van essentieel belang om het volledige potentieel van AI bij borstkankerscreening met mammografie te achterhalen. Ten slotte zal de toekomstige implementatie van AI in de dagelijkse praktijk ook moeten steunen op de voorkeuren van de belanghebbenden en de beschikbare technische en financiële ondersteuning. Een geïntegreerd systeem waarin al deze elementen gekoppeld zijn, dient nog ontwikkeld te worden.

7. REFERENTIELIJST

1. Kim HE, Kim HH, Han BK, Kim KH, Han K, Nam H, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health*. 2020;2(3):e138-e48.
2. Gastounioti A, Eriksson M, Cohen EA, Mankowski W, Pantalone L, Ehsan S, et al. External Validation of a Mammography-Derived AI-Based Risk Model in a U.S. Breast Cancer Screening Cohort of White and Black Women. *Cancers*. 2022;14(19).
3. Yoon JH, Kim EK. Deep learning-based artificial intelligence for mammography. *Korean Journal of Radiology*. 2021;22(8):1225-39.
4. Cai H, Huang Q, Rong W, Song Y, Li J, Wang J, et al. Breast Microcalcification Diagnosis Using Deep Convolutional Neural Network from Digital Mammograms. *Comput Math Methods Med*. 2019;2019:2717454.
5. Bao C, Shen J, Zhang Y, Zhang Y, Wei W, Wang Z, et al. Evaluation of an artificial intelligence support system for breast cancer screening in Chinese people based on mammogram. *Cancer Medicine*. 2022.
6. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *J Natl Cancer Inst*. 2019;111(9):916-22.
7. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. *Sci Rep*. 2018;8(1):4165.
8. Zhang C, Zhao J, Niu J, Li D. New convolutional neural network model for screening and diagnosis of mammograms. *PLoS One*. 2020;15(8):e0237674.
9. Freeman K, Geppert J, Stinton C, Todkill D, Johnson S, Clarke A, Taylor-Phillips S. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *Bmj*. 2021;374:n1872.
10. Bhattarai S, Klimov S, Aleskandarany MA, Burrell H, Wormall A, Green AR, et al. Machine learning-based prediction of breast cancer growth rate in vivo. *Br J Cancer*. 2019;121(6):497-504.
11. Hinton B, Ma L, Mahmoudzadeh AP, Malkov S, Fan B, Greenwood H, et al. Deep learning networks find unique mammographic differences in previous negative mammograms between interval and screen-detected cancers: a case-case study. *Cancer Imaging*. 2019;19(1):41.
12. Chang YW, An JK, Choi N, Ko KH, Kim KH, Han K, Ryu JK. Artificial Intelligence for Breast Cancer Screening in Mammography (AI-STREAM): A Prospective Multicenter Study Design in Korea Using AI-Based CAde/x. *Journal of Breast Cancer*. 2022;25(1):57-68.
13. Romero-Martín S, Elías-Cabot E, Raya-Povedano JL, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. Stand-Alone Use of Artificial Intelligence for Digital Mammography and Digital Breast Tomosynthesis Screening: A Retrospective Evaluation. *Radiology*. 2022;302(3):535-42.
14. Yu X, Pang W, Xu Q, Liang M. Mammographic image classification with deep fusion learning. *Sci Rep*. 2020;10(1):14361.
15. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. *Radiology*. 2021;300(1):57-65.

16. Wanders AJ, Mees W, Bun PA, Janssen N, Rodríguez-Ruiz A, Dalmış MU, et al. Interval Cancer Detection Using a Neural Network and Breast Density in Women with Negative Screening Mammograms. *Radiology*. 2022;303(2):269-75.
17. Aboutalib SS, Mohamed AA, Berg WA, Zuley ML, Sumkin JH, Wu S. Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening. *Clin Cancer Res*. 2018;24(23):5902-9.
18. Chougrad H, Zouaki H, Alheyane O. Deep Convolutional Neural Networks for breast cancer screening. *Comput Methods Programs Biomed*. 2018;157:19-30.
19. Schaffter T, Buist DSM, Lee CI, Nikulin Y, Ribli D, Guan Y, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw Open*. 2020;3(3):e200265.
20. Katzen J, Dodelzon K. A review of computer aided detection in mammography. *Clin Imaging*. 2018;52:305-9.
21. Lauritzen AD, Rodríguez-Ruiz A, von Euler-Chelpin MC, Lynge E, Vejborg I, Nielsen M, et al. An Artificial Intelligence-based Mammography Screening Protocol for Breast Cancer: Outcome and Radiologist Workload. *Radiology*. 2022;304(1):41-9.
22. Rodríguez-Ruiz A, Krupinski E, Mordang JJ, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, Mann RM. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology*. 2019;290(2):305-14.
23. Gastounioti A, Desai S, Ahluwalia VS, Conant EF, Kontos D. Artificial intelligence in mammographic phenotyping of breast cancer risk: a narrative review. *Breast Cancer Research*. 2022;24(1).
24. Lee JH, Kim KH, Lee EH, Ahn JS, Ryu JK, Park YM, et al. Improving the Performance of Radiologists Using Artificial Intelligence-Based Detection Support Software for Mammography: A Multi-Reader Study. *Korean Journal of Radiology*. 2022;23(5):505-16.
25. Sechopoulos I, Teuwen J, Mann R. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. *Seminars in Cancer Biology*. 2021;72:214-25.
26. Zhu X, Wolfgruber TK, Leong L, Jensen M, Scott C, Winham S, et al. Deep Learning Predicts Interval and Screening-detected Cancer from Screening Mammograms: A Case-Case-Control Study in 6369 Women. *Radiology*. 2021;301(3):550-8.
27. Houssami N, Given-Wilson R, Ciatto S. Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. *J Med Imaging Radiat Oncol*. 2009;53(2):171-6.
28. Dembrower K, Wählin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit Health*. 2020;2(9):e468-e74.
29. Kizildag Yirgin I, Koyluoglu YO, Seker ME, Ozkan Gurdal S, Ozaydin AN, Ozcinar B, et al. Diagnostic Performance of AI for Cancers Registered in A Mammography Screening Program: A Retrospective Analysis. *Technol Cancer Res Treat*. 2022;21:15330338221075172.
30. Sechopoulos I, Mann RM. Stand-alone artificial intelligence - The future of breast cancer screening? *Breast (Edinburgh, Scotland)*. 2020;49:254-60.
31. Lennox-Chhugani N, Chen Y, Pearson V, Trzcinski B, James J. Women's attitudes to the use of AI image readers: A case study from a national breast screening programme. *BMJ Health and Care Informatics*. 2021;28(1).
32. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. *Eur Radiol*. 2021;31(3):1687-92.
33. Rodríguez-Ruiz A, Lång K, Gubern-Merida A, Teuwen J, Broeders M, Gennaro G, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *Eur Radiol*. 2019;29(9):4825-32.
34. Hendrix N, Hauber B, Lee CI, Bansal A, Veenstra DL. Artificial intelligence in breast cancer screening: primary care provider preferences. *J Am Med Inform Assoc*. 2021;28(6):1117-24.
35. Sasaki M, Tozaki M, Rodríguez-Ruiz A, Yotsumoto D, Ichiki Y, Terawaki A, et al. Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. *Breast Cancer*. 2020;27(4):642-51.
36. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Trans Med Imaging*. 2020;39(4):1184-94.
37. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Sci Rep*. 2019;9(1):12495.

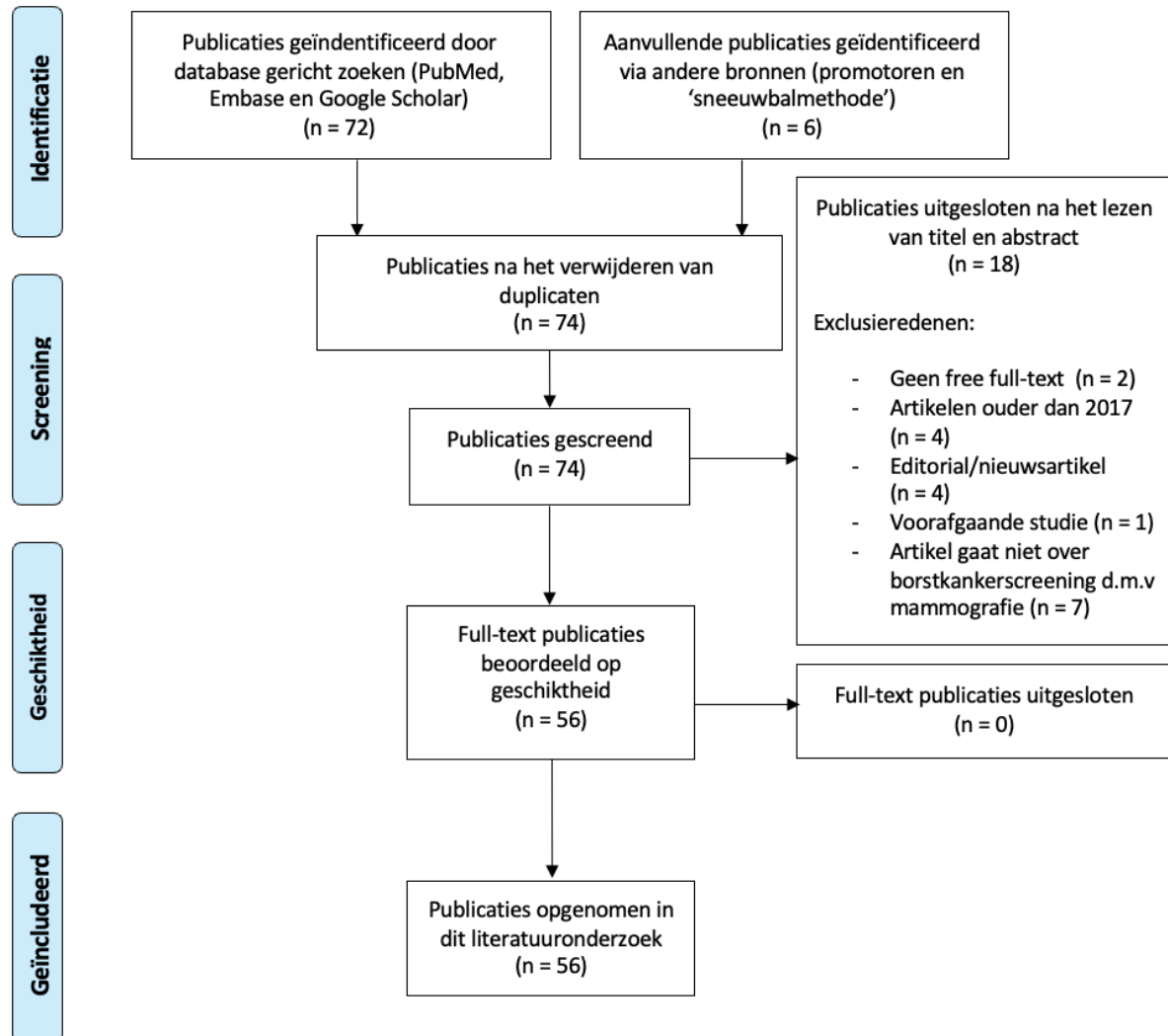
38. Lehman CD, Topol EJ. Readiness for mammography and artificial intelligence. *The Lancet*. 2021;398(10314):1867.
39. Jakhar D, Kaur I. Artificial intelligence, machine learning and deep learning: definitions and differences. *Clinical and Experimental Dermatology*. 2020;45(1):131-2.
40. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. *npj Digital Medicine*. 2020;3(1):126.
41. Kaplan A, Haenlein M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*. 2019;62(1):15-25.
42. Chahal A, Gulia P. Machine learning and deep learning. *International Journal of Innovative Technology and Exploring Engineering*. 2019;8(12):4910-4.
43. Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *The Journal of Arthroplasty*. 2018;33(8):2358-61.
44. Nakaura T, Higaki T, Awai K, Ikeda O, Yamashita Y. A primer for understanding radiology articles about machine learning and deep learning. *Diagnostic and Interventional Imaging*. 2020;101(12):765-70.
45. Dang LA, Chazard E, Poncelet E, Serb T, Rusu A, Pauwels X, et al. Impact of artificial intelligence in breast cancer screening with mammography. *Breast Cancer*. 2022;29(6):967-77.
46. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. *Radiology*. 2019;293(1):38-46.
47. Frazer HML, Qin AK, Pan H, Brothie P. Evaluation of deep learning-based artificial intelligence techniques for breast cancer detection on mammograms: Results from a retrospective study using a BreastScreen Victoria dataset. *J Med Imaging Radiat Oncol*. 2021;65(5):529-37.
48. Lång K, Hofvind S, Rodríguez-Ruiz A, Andersson I. Can artificial intelligence reduce the interval cancer rate in mammography screening? *Eur Radiol*. 2021;31(8):5940-7.
49. Byng D, Strauch B, Gnäs L, Leibig C, Stephan O, Bunk S, Hecht G. AI-based prevention of interval cancers in a national mammography screening program. *European Journal of Radiology*. 2022;152.
50. Mohamed AA, Luo Y, Peng H, Jankowitz RC, Wu S. Understanding Clinical Mammographic Breast Density Assessment: a Deep Learning Perspective. *J Digit Imaging*. 2018;31(4):387-92.
51. American College of Radiology. BI-RADS atlas - Reporting system 2013. ACR 2013. Opgehaald op 26 april 2023, van acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads.
52. Kyono T, Gilbert FJ, van der Schaar M. Improving workflow efficiency for mammography using machine learning. *Journal of the American College of Radiology*. 2020;17(1):56-63.
53. Yala A, Mikhael PG, Lehman C, Lin G, Strand F, Wan YL, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat Med*. 2022;28(1):136-43.
54. Hendrix N, Lowry KP, Elmore JG, Lotter W, Sorensen G, Hsu W, et al. Radiologist preferences for artificial intelligence-based decision support during screening mammography interpretation. *Journal of the American College of Radiology : JACR*. 2022.
55. Tang X. The role of artificial intelligence in medical imaging research. *BJR Open*. 2020;2(1):20190031.
56. Pesapane F, Codari M, Sardanelli F. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*. 2018;2(1):35.
57. Dembrower K, Crippa A, Colón E, Eklund M, Strand F. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *The Lancet Digital Health*. 2023.

8. ADDENDUM

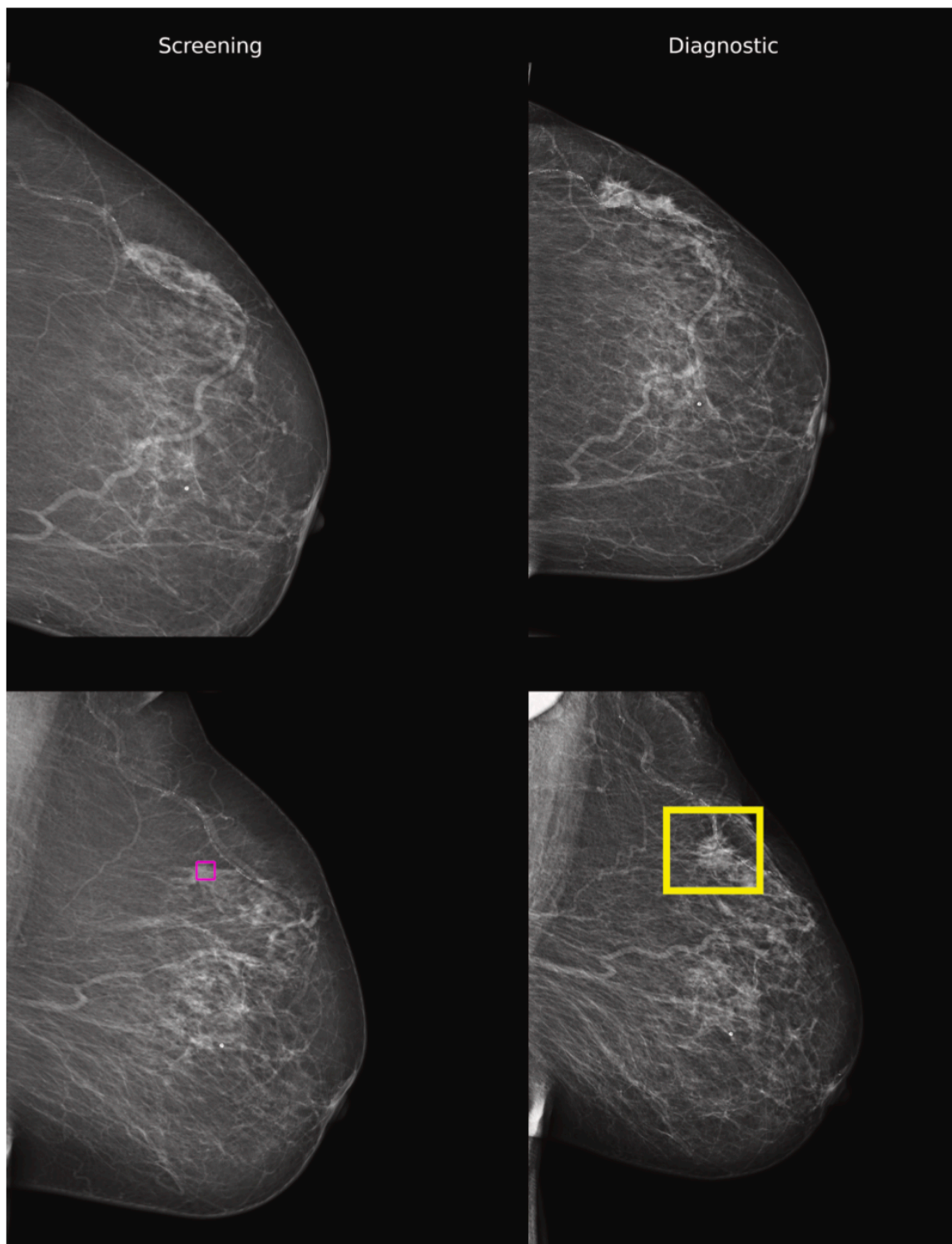
8.1 Acroniemenlijst in volgorde van voorkomen in deze masterproef

Afkorting	Betekenis
CAD	Computer-Aided Detection and Diagnosis
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
AI	Artificial Intelligence/Kunstmatige Intelligentie
ANI	Artificial Narrow Intelligence
AGI	Artificial General Intelligence
ASI	Artificial Super Intelligence
ML	Machine Learning
DL	Deep Learning
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
IC	Intervalcancer/Intervalkanker
BI-RADS	Breast Imaging-Reporting and Data System
AUC	Area Under the receiver operating characteristic Curve
BS	Borstspecialisten
AR	Algemene Radiologen
s	Seconden
BI	Betrouwbaarheidsinterval
DDSM	Digital Database for Screening Mammography
FFDM	Full-Field Digital Mammography dataset
ROC	Receiver Operating Characteristic
NN	Neuraal Netwerk
LN	Lymph Node/Lymfeklier
MGH	Massachusetts General Hospital
TCv8	Tyrer-Cuzick versie 8
DCE	Discrete Choice Experiment/Discrete Keuze-Experiment
KPW	Kaiser Permanente Washington
KI	Karolinska Institute

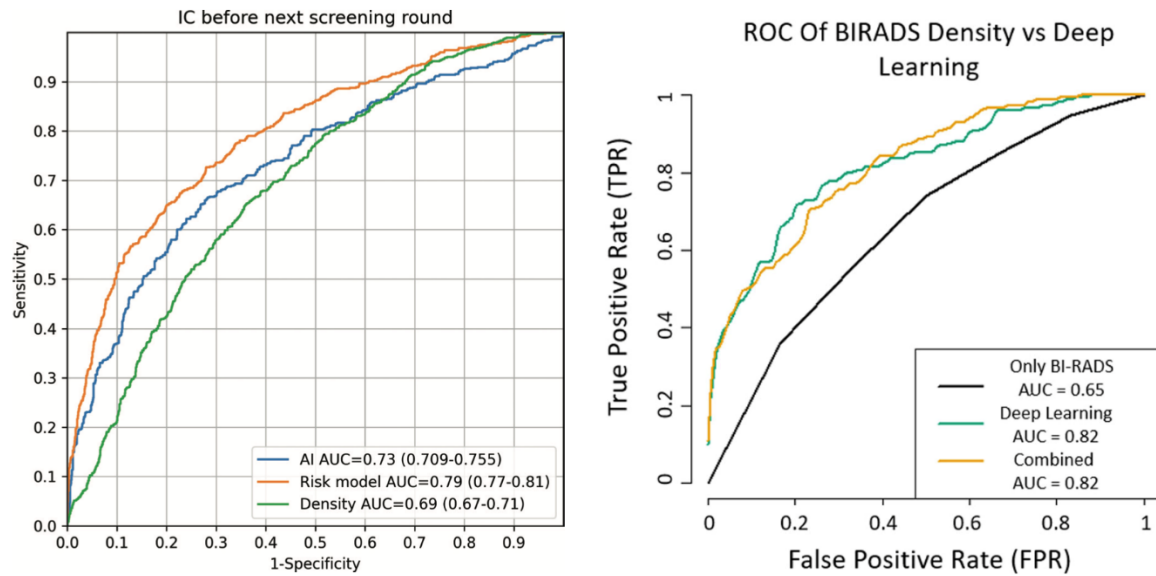
8.2 Tabellen en figuren



Figuur 4: Overzicht van de zoekstrategie zoals weergegeven in de methodologie aan de hand van een PRISMA-flowchart.



Figuur 5: Voorbeeld van een door het AI-systeem gedetecteerde en correct gelokaliseerde vals-negatieve IC op screeningsmammografie. De AI-markering is terug te vinden linksonder op de mediolaterale opname (49: p.7). (Afbeelding gebruikt volgens copyright licentie: <https://creativecommons.org/licenses/by-nc-nd/4.0/>)



Figuur 6: Grafiek links (16: p.272) toont de ROC curve met de prestaties geleverd door het combinatiemodel AI en borstdensiteit, AI afzonderlijk en borstdensiteit afzonderlijk, betreffende de detectie van intervalkanker (IC) bij screeningsonderzoek voorafgaand aan de diagnose. Grafiek rechts (11: p.6) toont de ROC-curve bij het voorspellen van IC versus via screeningsmammografie gedetecteerde kanker met de prestaties geleverd door het combinatiemodel AI en borstdensiteit, AI afzonderlijk en borstdensiteit afzonderlijk. (Afbeeldingen gebruikt volgens copyright licentie: <https://creativecommons.org/licenses/by/4.0/>)

Tabel 1: Karakteristieken van de onderzoeken omtrent AI-ondersteuning van radiologen.

	Bao et al. (5)	Rodriguez-Ruiz et al. (22)	Lee et al. (24)	Dang et al. (45)	Kizildag Yirgin et al. (29)	Kim et al. (1)	Wu et al. (36)	Schaffter et al. (19)	Dembrower et al. (57)
Studieopzet	Retrospectief	Retrospectief	Retrospectief	Retrospectief	Retrospectief	Retrospectief	Retrospectief	Retrospectief	Prospectief
Aantal beoordelingen zonder AI-ondersteuning	564	240	200	628	211	320	720	25 657 KPW 166 578 KI	55 581
Aantal beoordelingen met AI-ondersteuning	562	240	200	628	211	320	720	25 657 KPW 166 578 KI	55 581
Aantal mammografieën met kanker	62.5%	41.7%	50%	20.4%	52%	50%	4.3%	1.1%	0.5%
Gegevensverzameling	Twee 3A ziekenhuizen en één tertiair ziekenhuis in China	Twee instituten, één in de VS en één in Europa	Soon-chunhyang University Hospital Bucheon in Zuid-Korea	Valenciennes Hospital in Frankrijk	Bahcesehir Mammographic Screening Program in Turkije	Yonsei University Severance Hospital, Seoul, South Korea; en Soonchunhyang University Hospital Bucheon, Zuid-Korea	/	Twee instituten: Kaiser Permanente Washington (KPW) in de VS en Karolinska Institute (KI) in Zweden	Capio Sankt Göran Hospital, Zweden
Eigenschappen van de mammografieën	Voor elke vrouw werden ten minste twee projecties verkregen: mediolateraal schuin van de rechter- of linkerborst en craniocaudaal van de rechter- of linkerborst	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Verskillende projecties van het meest recente screeningsonderzoek	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal

AI-algoritme	Yizhun AI v.3.2.3	Transpara v.1.3.0 ScreenPoint Medical (gebaseerd op DL CNNs)	Lunit INSIGHT MMG, v.1.1.1.0 (gebaseerd op CNNs)	Mammoscreen v.1.2, Therapixel	Lunit INSIGHT MMG (gebaseerd op CNNs)	Een eigen AI- algoritme gebaseerd op CNNs met ResNet-34 als basis.	Een eigen AI- algoritme gebaseerd op CNNs met een ResNet-22 basisarchitectuur.	Een AI-model samengesteld door de top 8 teams. Ook werd een geïntegreerd AI- model met beoordelingen van radiologen geproduceerd.	Lunit INSIGHT MMG, v.1.1.6 (gebaseerd op CNNs)
---------------------	----------------------	---	---	-------------------------------------	--	---	---	---	--

Tabel 2: Samenvatting van de resultaten van de onderzoeken omtrent AI-ondersteuning van radiologen.

	Bao et al. (5)	Rodriguez-Ruiz et al. (22)	Lee et al. (24)	Dang et al. (45)	Kizildag Yirgin et al. (29)	Kim et al. (1)	Wu et al. (36)	Schaffter et al. (19)	Dembrower et al. (57)
AUC radiologen zonder AI-ondersteuning	0.84	0.87	0.813 BS 0.684 AR	0.739	/	0.810 BS + AR	0.778	/	/
AUC radiologen met AI-ondersteuning	0.91	0.89	0.884 BS 0.833 AR	0.773	/	0.881 BS + AR	0.891	0.942	/
Sensitiviteit zonder AI-ondersteuning	84.77%	83%	74.6% BS 52.1% AR	66 %	67.3%	/	/	85.9% KPW 77.1% KI	92.9% bij dubbele lezing door twee radiologen
Sensitiviteit met AI-ondersteuning	95.07%	86%	88.6% BS 79.4% AR	70 %	83.6%	/	/	/	91.4% AI autonoom 97.0% AI + één radioloog 100% AI + twee radiologen
Specificiteit zonder AI-ondersteuning	63.55%	77%	66.6% BS 70.8% AR	79 %	/	/	/	90.5% KPW 96.7% KI	/
Specificiteit met AI-ondersteuning	54.88%	79%	66.4% BS 70% AR	81 %	/	/	/	92% KPW 98.5% KI	/
Interpretatietijd zonder AI-ondersteuning	215 s	146 s	82.73 s BS 35.44 s AR	106 s	/	/	/	/	/
Interpretatietijd met AI-ondersteuning	106 s	149 s	73.04 s BS 42.52 s AR	102 s	/	/	/	/	/

Afkortingen:

- BS: Borstspecialisten
- AR: Algemene Radiologen
- KPW: Kaiser Permanente Washington
- KI: Karolinska Institute

Tabel 3: Karakteristieken van een AI-systeem als autonome lezer.

	Rodriguez-Ruiz et al. (6)	Sasaki et al. (35)	Romero-Martin et al. (13)
Studieopzet	Retrospectief	Retrospectief	Retrospectief
Aantal onderzoeken	2652	310	15 999
Aantal mammografieën met kanker	24.6%	22.3%	0.7%
Gegevensverzameling	Negen datasets van verschillende instellingen in Europa en de Verenigde Staten	Sagara Hospital Affiliated Breast Center, Sagara Women's Healthcare Group, Japan	Tomosynthesis Córdoba Screening Trial, Spanje
Eigenschappen van de mammografieën	Twee projecties per borst (mediolateraal schuin en craniocaudaal) die unilateraal of bilateraal kunnen zijn	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	/
AI-algoritme	Transpara v.1.4.0 ScreenPoint Medical (gebaseerd op DL CNNs)	Transpara v.1.3.0 ScreenPoint Medical (gebaseerd op DL CNNs)	Transpara v.1.7.0 ScreenPoint Medical (gebaseerd op DL CNNs)

Tabel 4: Samenvatting van de resultaten van een AI-systeem als autonome lezer.

	Rodriguez-Ruiz et al. (22)	Lee et al. (24)	Kizildag Yirgin et al. (29)	Kim et al. (1)	Wu et al. (36)	Schaffter et al. (19)	Rodriguez-Ruiz et al. (6)	Sasaki et al. (35)	Romero-Martin et al. (13)
AUC radiologen zonder AI- ondersteuning	0.87	0.813 BS 0.684 AR	/	0.810 BS + AR	0.778	/	0.814	0.816	/
AUC AI- systeem als autonome lezer	0.887	0.915	0.853	0.959	0.876	0.858 (VS) 0.903 (Zweden)	0.840	0.706	0.930

Tabel 5: Karakteristieken van een AI-systeem als triagemodel.

	Lauritzen et al. (21)	Raya-Povedano et al. (15)	Lång et al. (32)	Rodriguez-Ruiz et al. (33)	Dembrower et al. (28)	Yala et al. (46)	Kyono et al. (52)
Studieopzet	Retrospectief	Retrospectief	Retrospectief	Retrospectief	Retrospectief	Retrospectief	Retrospectief
Totaal aantal patiënten geïncludeerd	114 421	15 987	9581	2652	7364	7176	2000
Aantal mammografieën met kanker	2.23%	0.71%	0.90%	24.6%	4.7%	/	15%
Gegevensverzameling	Danish Capital Region Borstkanker Screenings-programma, Denemarken	Tomosynthesis Córdoba Screening Trial, Spanje	Malmö Breast Tomosynthesis Screening Trial, Zweden	Negen datasets van verschillende instellingen in Europa en de Verenigde Staten	Karolinska University Hospital, Zweden	Een groot tertiair academisch medisch centrum	De Tommy dataset, verzameld door 6 NHS borstkanker-screenings-programma's, Verenigd Koninkrijk
Eigenschappen van de mammografieën	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Twee projecties per borst (mediolateraal schuin en craniocaudaal) die unilateraal of bilateraal kunnen zijn	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	/	Voor elke vrouw en zowel in de linker- als rechterborst werden twee projecties verkregen: mediolateraal schuin en craniocaudaal
AI-algoritme	Transpara v.1.7.0 ScreenPoint Medical (gebaseerd op DL CNNs)	Transpara v.1.6.0 ScreenPoint Medical (gebaseerd op DL CNNs)	Transpara v.1.4.0 ScreenPoint Medical (gebaseerd op DL CNNs)	Transpara v.1.4.0 ScreenPoint Medical (gebaseerd op DL CNNs)	Lunit, Seoul, South Korea, v.5.5.0 (gebaseerd op CNNs)	Een eigen DL-model met ResNet18 als basisarchitectuur	Een eigen AI-algoritme gebaseerd op een CNN in combinatie met 'multi-task learning'

Tabel 6: Samenvatting van de resultaten van een AI-systeem als triagemodel.

	Lauritzen et al. (21)	Raya-Povedano et al. (15)	Lång et al. (32)	Rodriguez-Ruiz et al. (33)	Dembrower et al. (28)	Yala et al. (46)	Kyono et al. (52)
Sensitiviteit radiologen zonder AI-triagemodel	70.8%	67.3%	/	/	/	90.6%	/
Sensitiviteit met AI-triagemodel	69.7%	69%	/	/	/	93.5%	/
Specificiteit radiologen zonder AI-triagemodel	98.1%	/	/	/	/	90.1%	/
Specificiteit met AI-triagemodel	98.6%	/	/	/	/	94.2%	/
Werklast-vermindering door het AI-triagemodel	62.6%	71.5%	19.1% bij uitsluiting van AI-scores 1 en 2 53.0% bij uitsluiting van AI-scores 1 t.e.m. 5	17% bij een AI-drempelwaarde van 2 47% bij een AI-drempelwaarde van 5	/	19.3%	34% bij 15% gevallen van borstkanker in de dataset 91% bij 1% gevallen van borstkanker in de dataset
Vals-positieven vermeden door het AI-triagemodel	25.1%	16.9%	0.1% bij uitsluiting van AI-scores 1 en 2 0.6% bij uitsluiting van AI-scores 1 t.e.m. 5	/	/	/	/
Gemiste kankers door het AI-triagemodel	1.5%	/	Geen gemiste kankers bij uitsluiting van AI-scores 1 en 2 0.07% bij uitsluiting van AI-scores 1 t.e.m. 5	1% bij een AI-drempelwaarde van 2 7% bij een AI-drempelwaarde van 5	Geen gemiste kankers bij uitsluiting van AI-scores met de laagste 60% 0.3% gemiste kankers bij uitsluiting van AI-scores met de laagste 70% 2.6% gemiste kankers bij uitsluiting van AI-scores met de laagste 80% 4.0% gemiste kankers bij uitsluiting van AI-scores met de laagste 90%	/	/

Tabel 7: Karakteristieken van de onderzoeken omtrent intervalekanker.

	Lång et al. (48)	Byng et al. (49)	Wanders et al. (16)	Hinton et al. (11)	Zhu et al. (26)	Dembrower et al. (28)	Kizildag et al. (29)
Studieopzet	Retrospectieve studie	Retrospectieve studie	Retrospectieve nested case-control studie	Retrospectieve case-case studie	Retrospectieve case-case-control studie	Retrospectieve simulatie studie	Retrospectieve studie
Studiepopulatie							
Totaal aantal patiënten geïncludeerd	429	2396	6883	355	6369	7364	211
Aantal patiënten geïncludeerd met intervalekanker	429	2396	2222	182	351	200	27
Screeningscentrum	Vier verschillende screeningslocaties in Zuid-Zweden (Malmö, Lund, Helsingborg, Kristianstad)	Zeven screeningslocaties in de Duitse deelstaat Nedersaksen	Twee regio's die deelnemen aan het Nederlandse bevolkingsonderzoek naar borstkanker	Vier screeningslocaties in de Verenigde Staten: University of California – San Francisco, California Pacific Medical Center, Marin General Hospital en Novato Community Hospital	Twee screeningslocaties in de Verenigde Staten	Eén locatie in de provincie Stockholm (Karolinska Universitair Ziekenhuis)	Bahcesehir Mammographic Screening Program (BMSP) in Turkije
Eigenschappen van de mammografieën	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal	Voor elke vrouw werden twee projecties verkregen: mediolateraal schuin en craniocaudaal
AI-algoritme	Transpara v.1.5.0, ScreenPoint Medical (gebaseerd op DL CNNs)	Vara 1.0.7	Transpara v.1.6.0 ScreenPoint Medical (gebaseerd op DL CNNs)	ResNet50 (CNN architectuur)	DenseNet-121 (CNN architectuur)	Lunit INSIGHT MMG, Seoul, South Korea, versie 5.5.0 (gebaseerd op DL CNNs)	Lunit INSIGHT MMG, Seoul, South Korea (gebaseerd op DL CNNs)
Gebruikte definitie van intervalekanker	Borstkankers die in het interval (resp. 18 en 24 maanden voor vrouwen onder en boven de 55 jaar) tussen twee screeningsrondes werden gediagnosticeerd, na een eerder negatief screeningsonderzoek.	Borstkankers die klinisch werden gediagnosticeerd in de periode van 24 maanden na een normale borstscreening, vóór de volgende geplande screeningsronde.	Borstkankers waarbij de resultaten van de screening negatief waren, maar waarbij er binnen de volgende 20 maanden een histologisch bewezen gediagnosticeerd.	Borstkankers die binnen 12 maanden na een negatief screeningsonderzoek werden gediagnosticeerd.	Borstkankers die binnen 12 maanden na een negatief screeningsonderzoek werden gediagnosticeerd.	Borstkankers in de periode voor de eerstvolgende geplande tweejaarlijkse screening gediagnosticeerd.	Borstkankers die binnen de 2 jaar na een negatieve mammografische evaluatie werden gediagnosticeerd.

Tabel 8: Samenvatting van de resultaten van de onderzoeken omtrent intervalekanker (IC).

	Lång et al. (48)	Byng et al. (49)	Wanders et al. (16)	Hinton et al. (11)	Zhu et al. (26)	Dembrower et al. (28)	Kizildag et al. (29)
Resultaten met betrekking tot het correct identificeren van een bepaald aantal ICs	AI zou radiologen potentieel kunnen helpen bij het opsporen tot 19.3% van de ICs die bij screening minstens minimale tekenen van maligniteit vertonen.	Van alle retrospectief zichtbare ICs, inclusief minimale tekenen, kon het AI-systeem 21.0% (177/842) detecteren en 90.4% (160/177) hiervan correct lokaliseren met 99% specificiteit.	Het combinatiemodel AI en borstdensiteit maakt een correcte voorspelling mogelijk van de ontwikkeling van IC in 50.9% van de vals-negatieve screenings-mammografieën met een specificiteit van 90%.	75.2% van het totale aantal mammografieën werd correct gecategoriseerd door het AI-systeem afzonderlijk. Hierbij werden 73.6% van het totale aantal ICs correct voorspeld als IC door het AI-systeem afzonderlijk.		Na een negatieve screenings-mammografie, met dubbele lezing door twee radiologen, werden er bij de top 1% (633 vrouwen) van de AI-scores 12% (24/200) ICs gedetecteerd. Bij de top 5% (1445 vrouwen) van de door AI toegekende scores waren dit er 27% (53/200).	Het AI-systeem was in staat om 44.4% (12/27) van de nog niet eerder door de radioloog gedetecteerde ICs te detecteren.
Resultaten met betrekking tot combinatie-modellen van AI en borstdensiteit, AI-systemen afzonderlijk en op borstdensiteit gebaseerde modellen afzonderlijk			De AUC bij het voorspellen van IC na een negatieve screenings-mammografie bedroeg het meest (0.79) voor het combinatiemodel AI en borstdensiteit.	De AUC bij het voorspellen van IC versus op screenings-mammografie gedetecteerde kanker bedroeg het meest (0.82) voor het combinatiemodel AI en borstdensiteit.	De 'C-statistic' (vergelijkbaar of identiek aan de AUC) voor modellen die vrouwen met IC vergeleken met gematchte controles bedroeg het meest (0.72) voor het combinatiemodel AI en klinische risicofactoren.	De AUC voor het voorspellen van IC bedroeg het meest (0.74) voor het AI-systeem (maximale score) afzonderlijk in vergelijking met model gebaseerd op borstdensiteit afzonderlijk.	
Resultaten met betrekking tot combinatie-modellen van AI en radiologen, AI-systemen afzonderlijk en de radiologen afzonderlijk							De totale detectiepercentages voor kanker waren het hoogst (83.6% (92/110)) voor de combinatie AI en radiologen.

Tabel 9: Thema's bevraagd in de enquêtes omtrent de voorkeuren van de belanghebbenden bij het gebruik van AI voor borstkankerscreening.

	Hendrix et al. (54): radiologen	Hendrix et al. (34): eerstelijnszorgverleners	Lennox-Chhugani et al. (31): vrouwen
Bevraagde thema's	Hoe verdenking van maligniteit wordt gecommuniceerd	Sensitiviteit	Persoonlijke attributen, waaronder leeftijd
	Hoe AI presteert	Specificiteit	Ervaring met borstkanker (direct of indirect)
	Hoe de workflow kan worden aangepast	Betrokkenheid van de radioloog	Kennis en ervaring met borstkankerscreening
	Hoe het algoritme werd gevalideerd	Begrijpelijkheid van hoe het AI systeem tot een besluit komt	Gebruik van op AI-gebaseerde technologie in het dagelijks leven
	Hoe AI wordt toegepast op het protocol	Ondersteunend bewijsmateriaal	Houding ten opzichte van op AI-gebaseerde technologie in het algemeen
		Diversiteit van trainings- en validatiegegevens	Houding ten aanzien van het gebruik van AI bij borstkankerscreening